



Genome-Wide Gene-Based Multi-Trait Analysis

Yamin Deng¹, Tao He², Ruling Fang¹, Shaoyu Li³, Hongyan Cao¹ and Yuehua Cui^{4*}

¹ Division of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan, China, ² Department of Mathematics, San Francisco State University, San Francisco, CA, United States, ³ Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC, United States, ⁴ Department of Statistics and Probability, Michigan State University, East Lansing, MI, United States

OPEN ACCESS

Edited by:

Guolian Kang,
St. Jude Children's Research
Hospital, United States

Reviewed by:

Ming Li,
Indiana University, United States
Qing Lu,
University of Florida, United States

*Correspondence:

Yuehua Cui
cuiy@msu.edu

Specialty section:

This article was submitted to
Statistical Genetics and Methodology,
a section of the journal
Frontiers in Genetics

Received: 24 February 2020

Accepted: 08 April 2020

Published: 19 May 2020

Citation:

Deng Y, He T, Fang R, Li S, Cao H
and Cui Y (2020) Genome-Wide
Gene-Based Multi-Trait Analysis.
Front. Genet. 11:437.
doi: 10.3389/fgene.2020.00437

Genome-wide association studies focusing on a single phenotype have been broadly conducted to identify genetic variants associated with a complex disease. The commonly applied single variant analysis is limited by failing to consider the complex interactions between variants, which motivated the development of association analyses focusing on genes or gene sets. Moreover, when multiple correlated phenotypes are available, methods based on a multi-trait analysis can improve the association power. However, most currently available multi-trait analyses are single variant-based analyses; thus have limited power when disease variants function as a group in a gene or a gene set. In this work, we propose a genome-wide gene-based multi-trait analysis method by considering genes as testing units. For a given phenotype, we adopt a rapid and powerful kernel-based testing method which can evaluate the joint effect of multiple variants within a gene. The joint effect, either linear or nonlinear, is captured through kernel functions. Given a series of candidate kernel functions, we propose an omnibus test strategy to integrate the test results based on different candidate kernels. A p -value combination method is then applied to integrate dependent p -values to assess the association between a gene and multiple correlated phenotypes. Simulation studies show a reasonable type I error control and an excellent power of the proposed method compared to its counterparts. We further show the utility of the method by applying it to two data sets: the Human Liver Cohort and the Alzheimer Disease Neuroimaging Initiative data set, and novel genes are identified. Our method has broad applications in other fields in which the interest is to evaluate the joint effect (linear or nonlinear) of a set of variants.

Keywords: gene-based association, kernel function, multi-trait, nonlinear effect, p -value combination

INTRODUCTION

Methods on genome-wide association studies (GWAS) are mostly focused on single variant (e.g., single nucleotide polymorphism, SNP) analysis with a single phenotype, the so-called single-variant single-trait analysis. Increasing evidence shows that pleiotropy, the effect of one gene on multiple phenotypes (often correlated), plays a pivotal role in many complex traits (Stearns, 2010; Schifano et al., 2013). For example, cognitive ability is often assessed in many domains such as memory, intelligence, language, and visual-spatial function (Yang and Wang, 2012). Instead of analyzing one trait at a time, we can take the correlated structure of multiple phenotypes into account and analyze them in a multi-trait analysis. As a complementary approach, such type of analysis can not only

gain association power by aggregating multiple weak signals (He et al., 2013; Schifano et al., 2013; Wang, 2014) but also lead to better understanding of disease etiology by detecting genetic variants with pleiotropic effects (Amos and Laing, 1993; Jiang and Zeng, 1995; Schifano et al., 2013).

For a multi-trait analysis, one commonly applied method is the one-way multivariate analysis of variance (MANOVA) (Bilodeau, 2013). Unfortunately, most multi-trait data do not satisfy the multivariate normal assumption for MANOVA, hence greatly limiting its applicability. Other methods are developed based on the idea of dimension reduction. For example, a multivariate response can be summarized into a univariate score using principal component (PC) analysis, based on which traditional univariate association methods can be applied (e.g., Zhang et al., 2012). As the first PC contains the most information about multiple phenotypes, this can change the test between a SNP and multiple phenotypes into a univariate test of association between a SNP and the first PC. The downside for this analysis is the lack of interpretability. Methods focusing on summary statistics have gained much popularity recently since the individual-level data are typically unavailable (e.g., Kim et al., 2015; Turley et al., 2018). However, such methods are largely undermined if the published GWAS summary statistics have limited accuracy. In addition, the marginal SNP effect is usually quite small in many complex diseases, and many identified SNPs have limited biological interpretation, for example, SNPs identified in non-coding regions.

These limitations motivated the development of gene- or pathway-based association analysis aimed at improving the statistical power and gaining novel insight into disease etiology (Wang et al., 2007; Cui et al., 2008; Liu et al., 2010). Firstly, the gene- or pathway-based analysis can largely alleviate the multiple testing burden by more than 10 or 100 folds. Secondly, due to allelic heterogeneity, most diseases are associated with a set of SNPs at different loci, making it hard to replicate the results based on a single-SNP analysis (Neale and Sham, 2004). In this case, a gene- or pathway-based analysis may provide additional insight to reveal the functional mechanism of complex diseases (Wang et al., 2010). Unlike the heterogeneity of a single locus, the biological function of genes is more consistent across populations, which enhances the likelihood of replication (Neale and Sham, 2004; Wang et al., 2010).

Most reports in the literature on multi-trait analysis are focused on a single-variant analysis, which shares the same limitation as described for the single-trait GWAS. Although methods for gene-based analysis focusing on a single trait have been developed, multi-trait analysis focusing on genes or gene sets is largely under-developed. There is a pressing need to develop a gene-based method for a multi-trait analysis.

In a gene-based single-trait analysis, the kernel-based testing (KBT) method is gaining much popularity recently due to its power and flexibility in capturing potential nonlinear effects (Kwee et al., 2008; Mukhopadhyay et al., 2010; Wu et al., 2010; Li and Cui, 2012; Lin et al., 2013; Marceau et al., 2015; Wei and Lu, 2017). The power of the KBT methods depends on the choice of kernel functions which measure the similarity between individuals across multiple genetic variants in a gene.

When the underlying true disease function is unknown, this limits the applicability of the KBT methods since the choice of the kernel function needs to be determined. Given a series of candidate kernel functions under the KBT framework, a common method is to choose the kernel function leading to the smallest p -value. This idea, however, could inflate the type I error rate due to the greedy process of kernel selection. We recently proposed a nonparametric KBT testing procedure which relaxes the distributional assumption required in most KBT methods (He et al., 2019). The asymptotic distribution of the test statistics approximately follows a normal distribution when the number of SNP variants in a gene set, p , is large. In fact, the normal approximation works well under a large p setting. Given a series of candidate kernel functions, we provided an analytical procedure to evaluate the p -value of the maximum statistics.

Based on empirical studies, the approximation method could be underperformed when p is relatively small. In this work, we borrowed the same idea but relaxed the large p assumption required for the normal approximation and proposed an omnibus testing procedure when multiple candidate kernels are available. Obtaining a p -value needs almost negligible computation and can be extremely fast. When extending the method to a multi-trait analysis, we adopted a Fisher p -value combination (FPC) method with correlated dependent variables, as proposed by Yang et al. (2016). The FPC provides an alternative approach for multi-trait analysis by integrating the single-trait analysis results. The proposed Omnibus Multi-trait Gene-based Association (OMGA) analysis can capture linear or nonlinear effects without kernel selection and is computationally efficient.

We conduct extensive simulation studies to evaluate the type I error control and power and further compare it with its counterparts. We demonstrate the performance of our proposed method through two real data applications of the Human Liver Cohort (HLC) study and the Alzheimer Disease Neuroimaging Initiative (ADNI) study. The results tell which genes are specific to a single phenotype or contributed to a common genetic construction of multiple phenotypes. Our OMGA method enriches the literature of genome-wide gene-based multi-trait association analysis and has broad applications in other fields where the interest is to evaluate the joint effect (linear or nonlinear) of a set of variants.

STATISTICAL METHODS

Gene-Based Association Test Based on a Single Trait The Model

To model the association between a gene and a quantitative trait, we consider the following semiparametric model (He et al., 2019),

$$Y_i = \mu + \alpha^T W_i + h(x_i) + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (1)$$

where Y_i is the response variable for the i -th individual, n is the sample size, α is the effect corresponding to $W_i = (W_{i1}, W_{i2}, \dots, W_{iH})^T$, a vector of H -dimensional covariates containing variables such as age and gender,

$x_i = (x_{i1}, \dots, x_{ip})^T$ is a vector of a p -dimensional SNP set in a given gene where p can be large, $h(\cdot)$ is an unknown function that captures the joint effect of multiple variants in a given SNP set, and ε_i is the random error with mean 0 and variance σ^2 . Here, we relax the error distribution assumption for the error term which does not have to follow a normal distribution.

Following model (1), assessing the effect of multiple variants in a given SNP set (e.g., a gene) is equivalent to test the hypotheses $H_0: h(\cdot) = 0$, while adjusting for the effects of covariates. Wu et al. (2011) proposed a kernel-based test by considering the joint effect of multiple SNPs in a given set and showed great power compared to a multiple-regression approach. In Wu et al. (2011), the function $h(\cdot)$ is modeled as a random effect and $h(\cdot) \sim N(0, \tau^2 K)$ where τ^2 is the variance and K is a kernel matrix which measures the similarity between individuals across multiple SNP variants. However, the normality assumption on $h(\cdot)$ limits its power when this assumption is violated. To relax this assumption, He et al. (2019) proposed a U-statistic defined as:

$$T_n = \frac{1}{n(n-1)} \sum_{i \neq j} K(X_i, X_j) (Y_i - \hat{Y}_i) (Y_j - \hat{Y}_j) / \hat{\sigma}^2,$$

where \hat{Y} and $\hat{\sigma}^2$ are sample estimates under the null model $Y_i = \mu + \alpha^T W_i + \varepsilon_i$; $K(X_i, X_j) = \frac{K_\theta(X_i, X_j)}{\sqrt{E\{K_\theta(X_i, X_i)\}E\{K_\theta(X_j, X_j)\}}}$ is the normalized kernel for kernel $K_\theta(X_i, X_j)$. In practice, the choice of kernel function for $K_\theta(X_i, X_j)$ depends on the underlying relationship between SNPs and the disease response. For example, a linear kernel is applied if the relationship between multiple SNP variants and the disease response is linear, and a Gaussian or polynomial kernel can be applied if a nonlinear relationship between multiple SNPs and the disease response is assumed. Several widely used kernel functions include the linear kernel $K_\theta(X_i, X_j) = X_i^T X_j / \theta$, IBS kernel for discrete SNP genotype data $K_\theta(X_i, X_j) = \sum_{k=1}^p (2 - |X_{ik} - X_{jk}|) / 2p$, and Gaussian kernel $K_\theta(X_i, X_j) = \exp(-\|X_i - X_j\|^2 / \theta)$. These kernels will be our candidate kernels in the simulation and real data analysis.

Let $\tilde{W}_{n \times (L+1)} = [1_n, W_{n \times L}]$ and $A = \tilde{W} (\tilde{W}^T \tilde{W})^{-1} \tilde{W}^T$. Then, we have $\hat{\sigma}^2 = Y^T (I - A) Y / (n - L - 1)$ and $\hat{Y} = AY$. Following the Eigen-decomposition, $K(X_i, X_j) = \sum_{m=1}^\infty \lambda_m \phi_m(X_i) \phi_m(X_j)$ where λ_m is the eigenvalues and $\phi_m(\cdot)$ is the orthonormal eigenvectors of the kernel K . For any positive integer k , let $V_k = \sum_{m=1}^\infty \lambda_m^k$. Then, under the null hypothesis of no association, the asymptotic distribution of the test statistic T_n follows a chi-square distribution, i.e.:

$$n T_n / V_1 \xrightarrow{d} \sum_{m=1}^\infty \lambda_{k,m} (x_{1,m}^2 - 1),$$

where $x_{1,m}^2$ are independent chi-square distributions with one degree of freedom. Then, we can apply a Satterthwaite approximation to the mixture of chi-squares by a scaled chi-square distribution $\hat{\alpha} \chi_{\hat{g}}^2 / \hat{V}_1 - 1$, where $\hat{g} = \hat{V}_1 / \hat{\alpha}$, $\hat{\alpha} = \hat{\sigma}_{T_n}^2 / 2 \hat{V}_1$,

and $\hat{V}_1 = n^{-1} \text{tr}(\mathbf{H}\mathbf{K})$ is a consistent estimator of V_1 with $\mathbf{H} = \mathbf{I} - n^{-1} \mathbf{J}$ as a projection matrix. Then, an asymptotic α -level test rejects the null if

$$(nT_n + \hat{V}_1) / \hat{\alpha} > \chi_{\hat{g}, 1-\alpha}^2,$$

where $\chi_{\hat{g}, 1-\alpha}^2$ is the $(1 - \alpha)$ th quantile of a chi-square distribution with \hat{g} degrees of freedom. Following He et al. (2019), $\sigma_{T_n}^2$ can be estimated by

$$\begin{aligned} & 1/n^2 \left(2 - \frac{12}{n^2} + \frac{6\hat{\Delta}}{n} \right) \text{tr}(\mathbf{H}\mathbf{K}\mathbf{H}\mathbf{K}) \\ & - \left(\frac{2}{n} + \frac{\hat{\Delta}}{n} \right) \text{tr}^2(\mathbf{H}\mathbf{K}) + \hat{\Delta} \text{tr}(\mathbf{B} \circ \mathbf{B}), \end{aligned}$$

where \circ represents the Hadamard product, $\hat{\Delta} = n^{-1} \sum_{i=1}^n \left[\frac{(Y_i - \hat{Y}_i)^4}{\hat{\sigma}^4} \right] - 3$, and $\mathbf{B} = \mathbf{H}\mathbf{K}\mathbf{H}$. Then, the p -value of T_n can be obtained.

An Omnibus Test With Multiple Candidate Kernels

The method described above works for a given kernel function. There are various kernel functions available to use. For example, if a linear relationship is assumed, then one can apply a linear kernel, while a Gaussian kernel can be applied when potential nonlinear relationship exists. Thus, the power of the proposed test statistic largely depends on the choice of the kernel function. If the optimal kernel function that captures the underlying true relationship cannot be determined, the testing power will suffer. In practice, the true relationship is generally unknown, so does the choice of the kernel function.

To overcome the issue of selecting the optional kernel function, we propose an omnibus test strategy in this work. Given a set of L candidate kernels denoted by $K_1(\cdot, \cdot), K_2(\cdot, \cdot), \dots, K_L(\cdot, \cdot)$, we can apply the proposed method and get the corresponding p -value denoted by p_1, p_2, \dots, p_L . These L kernel functions can come from a wide range of choices, such as the linear kernel, the Gaussian kernel, and the polynomial kernel. Then, we can transform the L p -values by a Cauchy transformation and combine the transformed p -values to form a new statistic (Liu et al., 2019),

$$T_O = \frac{1}{L} \sum_{j=1}^L \tan \{ (0.5 - p_j) \pi \}.$$

If p_j comes from the null hypothesis, the transformation $\tan \{ (0.5 - p_i) \pi \}$ follows a Cauchy distribution. Then, the p -value of T_O can be approximated by

$$p\text{-value} \approx 0.5 - \{\arctan(T_O)\} / \pi$$

This Cauchy combination method performs similarly as the minimum p -value method. In addition, it works well under different correlation structures. Thus, when the underlying true relationship is unknown, if the choice of the kernel function is rich enough, we can always achieve good power regardless

of the underlying disease gene action mode. More importantly, this method is computationally fast and robust to different dependence structures between p -values (Liu and Xie, 2019).

Gene-Based Association Test With Multiple Traits

When multiple correlated traits are available, it is more powerful to analyze them together to find the disease-gene association. One way to do so is to perform a multivariate analysis by treating multiple traits as a multivariate response. Generally speaking, it is much easier to conduct a univariate association test than a multivariate association test. Suppose there are a total of d quantitative traits. For a given gene, we can get d gene-level p -values, denoted by p_1, p_2, \dots, p_d . Since these d traits are generally correlated and the p -values are obtained based on the same gene, these p -values are typically correlated. To obtain a gene-based p -value for multiple traits, one simple way is to do a p -value combination. Unfortunately, the aforementioned Cauchy combination method does not work well in many cases since it functions like a minimum p -value approach, and this is not the intention for multi-trait analysis.

When the d p -values are independent, the Fisher combination method defined as $T = -2 \sum_{j=1}^d \log(p_j)$ follows a chi-square distribution with $2d$ degrees of freedom (Littell and Folks, 1971). For correlated traits, this method cannot be directly applied to find the association between one gene and multiple traits. In fact, the statistic T is a sum of correlated chi-square statistics which can be approximated by a scaled chi-square distribution δx_τ^2 or a gamma distribution with a scale parameter of 2δ and a shape parameter of $\tau/2$ under the null hypothesis (Yang et al., 2016). Let $E(T) = \mu$ and $\text{Var}(T) = \sigma^2$. Then, δ and τ can be computed as $\delta = \sigma^2/2\mu$ and $\tau = 2\mu^2/\sigma^2$. Here we adopt the method proposed by Yang et al. (2016) to combine the d -dependent p -values. The variance σ^2 can be calculated as

$$\begin{aligned} \sigma^2 &= \text{Var}[T] = \text{Var} \left\{ -2 \sum_{j=1}^d \log(p_j) \right\} \\ &= \sum_{j=1}^d \text{Var} \{-2 \log(p_j)\} + \sum_{j \neq k} \text{cov}(-2 \log(p_j), -2 \log(p_k)) \\ &= 4d + \sum_{j \neq k} \text{cov}(-2 \log(p_j), -2 \log(p_k)) \end{aligned}$$

Let $\delta_{jk} = \text{cov} \{-2 \log(p_j), -2 \log(p_k)\}$. Yang et al. (2016) proposed a method to estimate δ_{jk} based on which we can estimate σ^2 [please refer to Yang et al. (2016) for the technical details of estimating σ^2 and μ]. An R package implementing the method can be found at <https://github.com/jjyang2019/FisherCombinationStat>. Then, based on the estimators of μ and σ^2 for the gamma distribution parameters, the overall testing p -value of T can be calculated as

$$p\text{-value} = 1 - \Gamma(\mu^2/\sigma^2, \sigma^2/\mu).$$

The number of the gene-level test is much smaller than the number of the SNP-level test. After obtaining the gene-level

p -values, multiple testing adjustment such as FDR can be applied to claim the significance of a gene.

SIMULATION STUDIES

Simulation Design

To evaluate the statistical power and the type 1 error rate of the proposed method, we conducted extensive simulation studies to compare the proposed method (OMGA) with some existing methods. Specifically, we compared with the method of multivariate multiple linear regression (RMMLR) proposed by Basu et al. (2013) and the MANOVA method. RMMLR was developed based on multivariate regression and transformed the phenotype and genotype data to achieve a rapid gene-based genome-wide association test for multiple traits. The R package that implements the method, termed as RMMLR, is available at GitHub: <https://github.com/SAONLIB/RMMLR>. For the MANOVA analysis, the association between each SNP in a gene and multi-trait is implemented with the MANOVA function in R. The minimum p -value in a gene is recorded as the gene-level p -value.

The genetic data were simulated to mimic the real structure of a gene through the software EpiSIM (Shang et al., 2013). The software package of EpiSIM can be downloaded at <https://sourceforge.net/projects/episimsimulator/>. We simulated correlated quantitative phenotypes with the following model:

$$Y_i = 0.02Z_{i1} + 0.6Z_{i2} + h(X_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \dots, \epsilon_{id})^T$ is a d -dim random error vector generated from a multivariate normal distribution with mean 0 and covariance Σ ; $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{id})^T$ is a d -dim-dependent trait vector; $Z_{i1} \sim N(2, 1)$ and $Z_{i2} \sim \text{Ber}(0.6)$ are two independent covariates; $X_i = (X_{i1}, X_{i2}, \dots, X_{ip})^T$ is a p -dim SNP genotype vector in a gene. Under all scenarios, we simulated genes with different dimensions, i.e., $p = 50$ and $p = 100$, and with different sample sizes, namely, $n = 100, 200$, and 400 . For the number of traits, we assumed $d = 5$. The correlation between traits was assumed to be $\rho = 0.3$ and 0.8 , with the purpose to evaluate the impact of correlation on the testing power. In each scenario, we applied 1,000 simulation replications.

We assessed the type 1 error rates under the null hypothesis [i.e., $h(\cdot) = 0$] by the proportion of results that incorrectly rejected the null hypothesis. To evaluate the power, we set up four different scenarios for the $h(\cdot)$ function and recorded the proportion of results that rejected the null hypothesis. Under scenario A, we assumed that $h(x) = 0.2(x_1 - x_6) + \cos(x_6) \exp(-x_6^2/4)$, where the 1st and 6th SNP have a main effect with different directions and the 6th SNP also has a nonlinear effect on the five response traits. Under scenario B, we assumed that $h(x) = 0.3x_2 + 0.6x_4 - 0.07x_8$.

To mimic the situation where a large number of SNPs influence the traits, we assumed the following model:

$$h(x) = c_M \sum_{k \in S_M} \alpha_k x_k + c_N \sum_{k, k' \in S_N} \beta_{kk'} x_k x_{k'}$$

where S_M consists of a predefined set of 10 SNPs with main effect, and S_N contains a set of 30 SNP pairs with interactions. Both $\{\alpha_k, k \in S_M\}$ and $\{\beta_{kk'}, (k, k') \in S_N\}$ were generated from a uniform distribution with $\text{Unif}(0, 0.02)$, and were fixed for all simulation replicates once generated. Under scenario C, we set $C_M = 0.02$ and $C_N = 1.8$, which gave a combination of weak main effect and relatively strong interaction effect. Under scenario D, we set $C_M = 3.8$, and $C_N = 0$, with a pure main effect model. The four scenarios with their corresponding mean functions are summarized here:

Scenario A: $h(x) = 0.2(x_1 - x_6) + \cos(x_6) \exp(-x_6^2/4)$
 Nonlinear effect

Scenario B: $h(x) = 0.3x_2 + 0.6x_4 - 0.07x_8$
 Linear effect

Scenario C: $h(x) = 0.02 \sum_{k \in S_M} \alpha_k x_k + 1.8 \sum_{k, k' \in S_N} \beta_{kk'} x_k x_{k'}$
 Weak main but strong interaction effects

Scenario D: $h(x) = 3.8 \sum_{k \in S_M} \alpha_k x_k$
 Pure main effects

Simulation Results

Table 1 displays the empirical type 1 error rate of different methods under different settings, from which we conclude that the three methods maintained reasonable type 1 error rate control in most settings.

The power simulation results for the case with $p = 0.3$ are shown in Figure 1. Under different scenarios, the power of the three methods all increases as the sample size increases. Among the three methods, MANOVA performs the worst in most cases. Although the power decreases as the SNP dimension increases for all the three methods, the power decrease is more dramatic for RMMLR and MANOVA compared to that for OMGA. This indicates the relative advantage of the proposed method against the other two when the data dimension is high. The result clearly shows that the proposed omnibus test outperforms the other two methods under different scenarios since it can better capture the potential nonlinear effect of variants within a gene by applying a nonparametric KBT procedure with different kernel choices.

TABLE 1 | The type 1 error rate of different methods under different settings.

Data dimension	Sample size (n)	Correlation (ρ)	OMGA	RMMLR	MANONA
ρ = 50	100	0.3	0.059	0.037	0.052
		0.8	0.045	0.052	0.041
	200	0.3	0.050	0.061	0.038
		0.8	0.048	0.049	0.032
	400	0.3	0.048	0.064	0.052
		0.8	0.051	0.061	0.061
ρ = 100	100	0.3	0.044	0.052	0.046
		0.8	0.049	0.038	0.044
	200	0.3	0.061	0.041	0.046
		0.8	0.041	0.067	0.043
	400	0.3	0.051	0.057	0.035
		0.8	0.047	0.050	0.037

Figure 2 shows the empirical testing power of the three methods with $p = 0.8$. Compared with the $p = 0.3$ case, the power of RMMLR and MANONA decreased, while our proposed method can still maintain a comparable power as the $p = 0.3$ case. Note that the MANOVA method implemented here uses a minimum p -value approach among multiple SNPs to denote a gene-level p -value. The simulation result echoes the work of Basu and Pan (2011), in which the minimum p -value method performs the worst among the three methods that the authors compared in their simulation study.

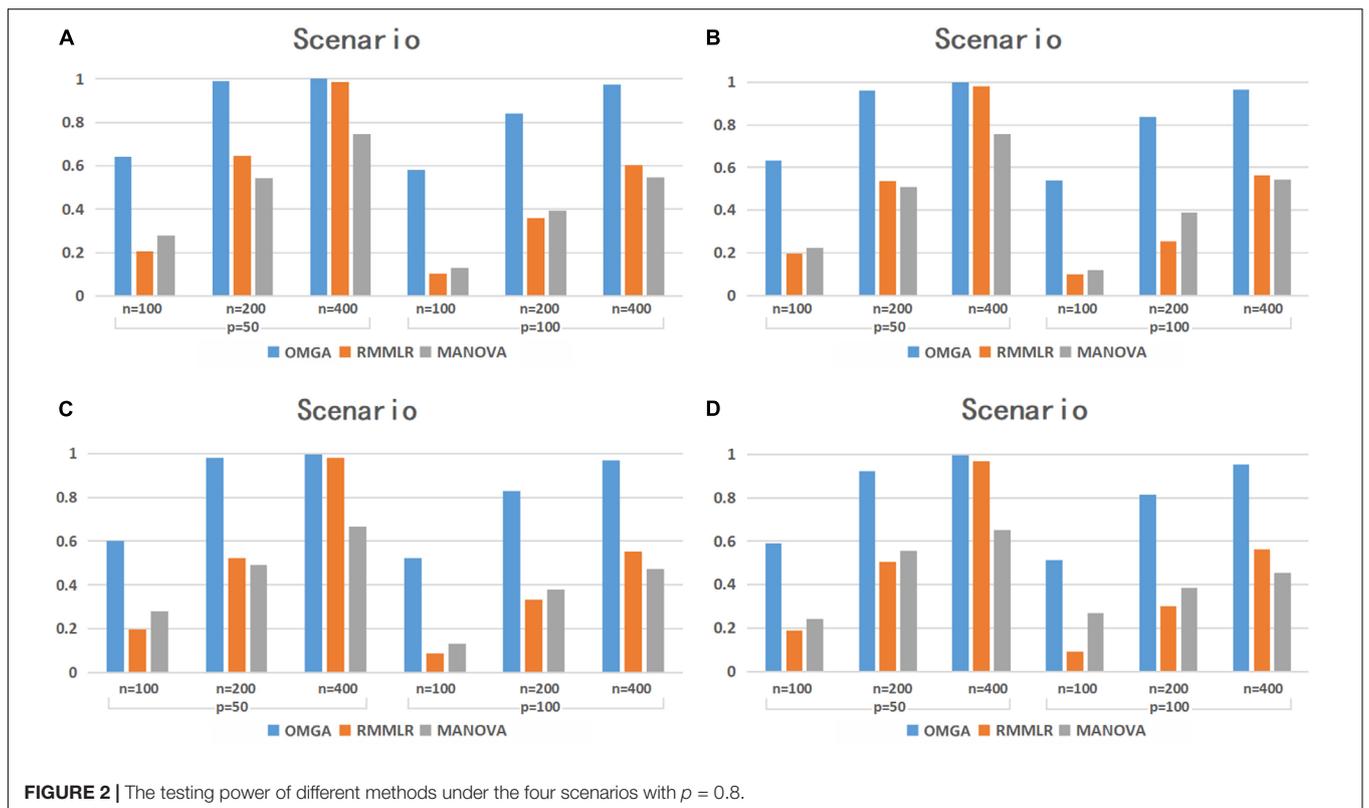
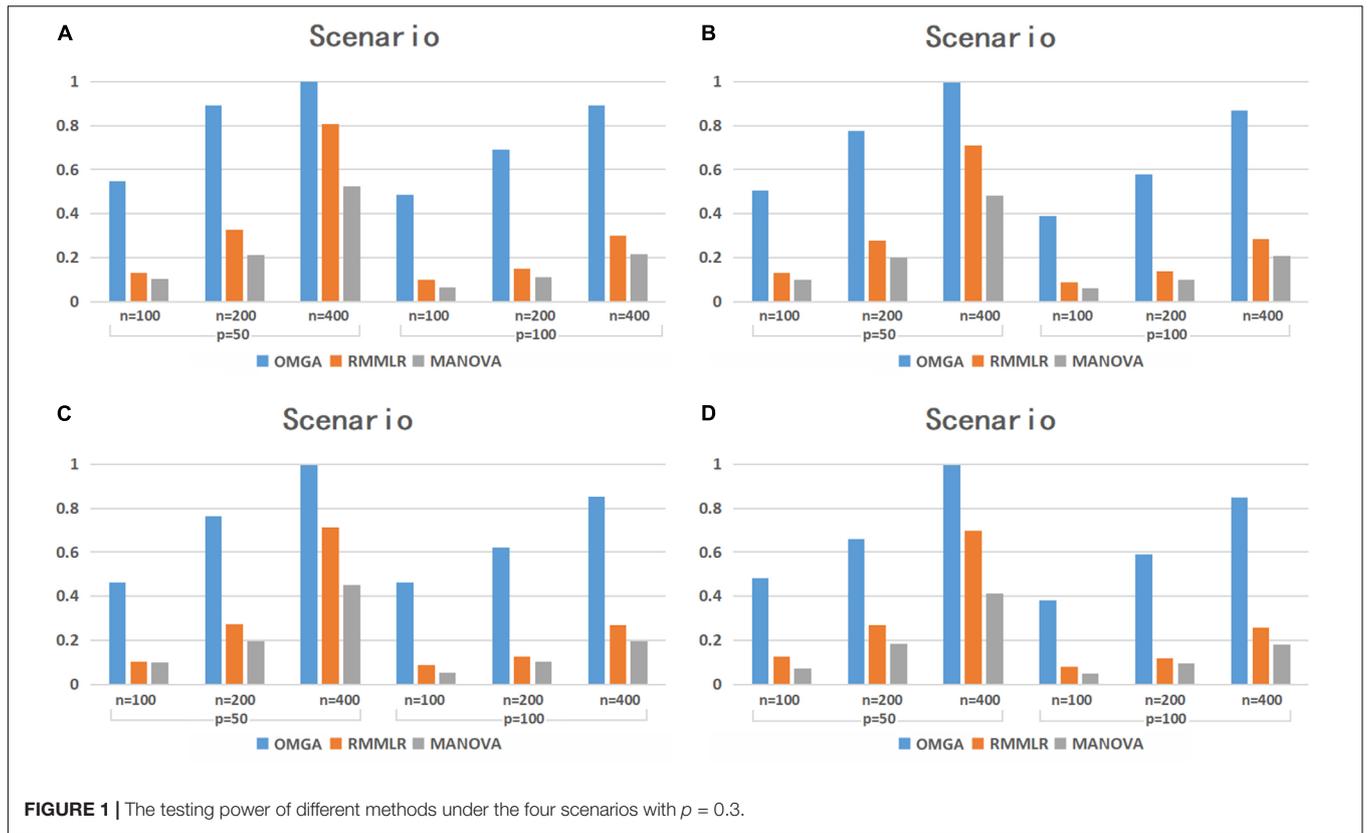
In summary, the simulation results clearly demonstrate that the proposed omnibus test method can maintain a reasonable type I error control while having better power than the other two methods under different scenarios. This is because the proposed omnibus testing method can efficiently capture a linear or a nonlinear relationship between multiple variants in a gene and multiple phenotypes. In practice, the underlying true disease-gene relationship is never known. This makes our proposed omnibus test method particularly attractive in real application since it does not put any model assumption. As long as the choice of kernel functions is rich enough, the omnibus test can achieve its power advantage against the other methods which only function well under the desired model assumption.

REAL DATA ANALYSIS

Case One: The Human Liver Cohort Data Analysis

To demonstrate the power and the applicability of our approach, we applied the proposed method OMGA together with RMMLR and MANONA to a HLC study data set, which can be downloaded from <https://www.synapse.org/#!/Synapse:syn4499>. The HLC study aims to explore the genetic architecture of gene expressions in human liver. There are a total nine phenotypes of P450 enzymes (CYP1A2, 2B6, 2C8, 2A6, 2C9, 2D6, 2C19, 2E1, and 3A4) from unrelated liver samples of Caucasian individuals. The samples were removed if their genotype and phenotype information were missed, and the final data included in our study contained 170 individuals. DNAs were genotyped by the Illumina 650Y SNP and Affymetrix 500K SNP genotyping arrays. SNPs with a minor allele frequency (MAF) less than 5% were removed. The total number of SNPs that remained was 312,082, which were further mapped into 11,579 genes using tools from the NCBI website <ftp://ftp.ncbi.nih.gov/snp/>.

The cytochrome P450s compose a superfamily of monooxygenases which are critical for anabolic and catabolic metabolism in almost all living organisms (Nelson et al., 1996; Aguiar et al., 2005; Plant, 2007). With its importance in physiology and drug metabolism in human, the regulatory mechanisms and genetic variations of P450 enzyme have been extensively studied. As there is a relatively close relationship among the CYP family enzymes, a joint analysis of multiple P450 enzyme traits and gene association can potentially lead to the identification of novel genes. Based on a hierarchical clustering analysis, we focused on six enzyme activity traits,



namely, CYP1A2, CYP3A4T, CYP2C8, CYP2B6, CYP2C9, and CYP2A6, as the response variables since they show a moderate correlation (see **Supplementary Figure S1**). We included age and gender as covariates in the analysis and log-transformed the six response variables.

For each individual trait, we first conducted a marginal gene-based single-trait analysis with the omnibus KBT. Then, we integrated the p -values for the six traits and applied the p -value combination method to get a gene-based multi-trait p -value. In the multi-trait analysis, we also applied the RMMLR and the MANONA methods. The Q–Q plot of the single-trait analysis is shown in **Supplementary Figure S2** and no p -value inflation was observed. **Figure 3** shows the Q–Q plot of the multi-trait analysis.

If we use the genome-wide gene-level Bonferroni correction, the threshold to claim a significant gene level significance is 4.3×10^{-6} . This leads to no significant genes in our analysis. Here, we only listed a few top genes with p -value less than 6×10^{-5} as suggestive significance. In the single-trait analysis, the top genes for each trait are *HAUS8* and *IRS12* for CYP1A2, *TRAPPC10* for CYP3A4T, *TARID* and *FUNDC2* for CYP2C9, and *PAPLN* for CYP2A6. No genes pass the suggested threshold for trait CYP2B6 and CYP2C8 (see **Supplementary Table S1** for a detailed list of associated genes for each trait and the corresponding p -values). For the multi-trait analysis, we listed in **Table 2** the results of the top genes along with the results by RMMLR and MANOVA. Among the four genes, *TARID*, *TRAPPC10*, and *HAUS8* were also in the list of single-trait analysis. Gene *ATAD3C* is not shown in the top list of the single-trait analysis. This may be due to the low power of the single-trait analysis. If we ignore the correlation information among the six enzyme traits and only focus on a single-trait analysis, we may miss some discoveries.

TABLE 2 | List of top genes and the p -values with different methods in the Human Liver Cohort study.

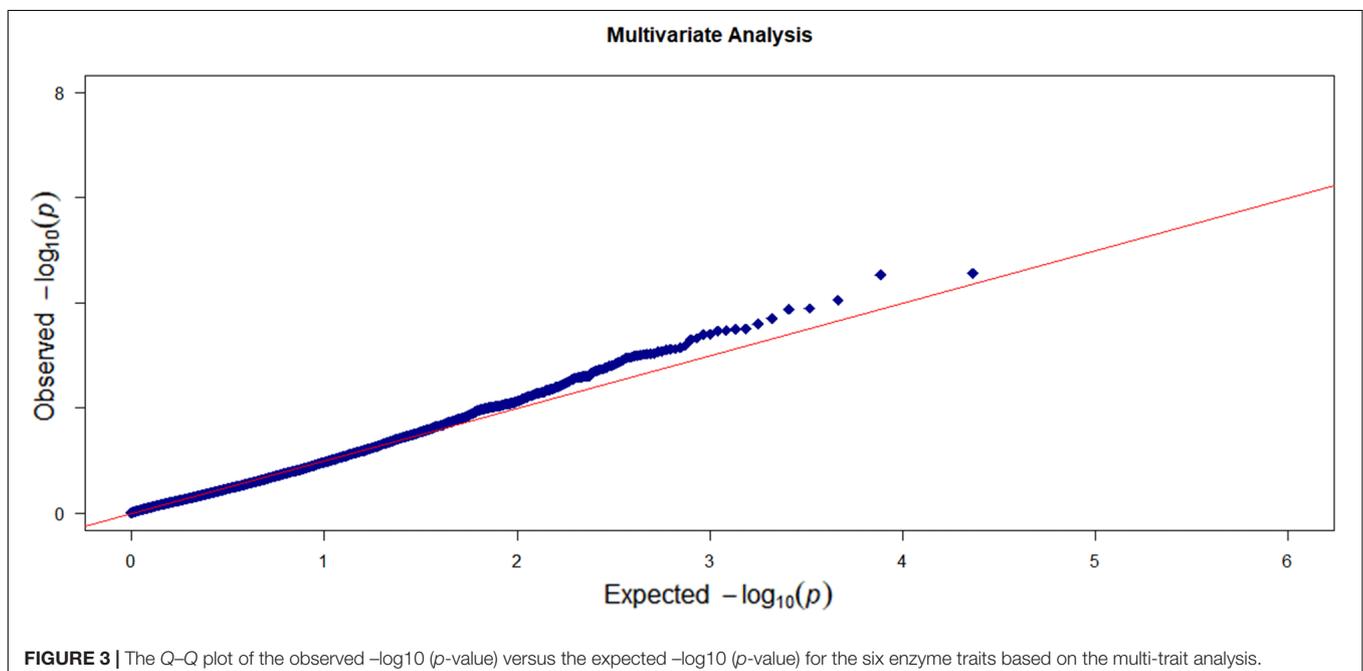
Gene name	Number of single nucleotide polymorphisms	Chr	OMGA	RMMLR	MANONA
<i>TARID</i>	80	6	1.11E–05	0.1227	0.1048
<i>TRAPPC10</i>	58	21	1.29E–05	0.0072	0.1003
<i>HAUS8</i>	42	19	4.22E–05	0.0425	0.1022
<i>ATAD3C</i>	150	1	5.53E–05	0.0789	0.0926

For the top four genes by OMGA, the p -values by RMMLR and MANOVA are all quite large. This could be due to the potential complex functional relationship between the genes and the traits. RMMLR and MANOVA were not designed to capture those complex relationships.

Empirical evidence supports some of the identified genes. For example, gene *ATAD3C* has been reported in literature to be associated with aldosterone metabolism and P450 enzyme (Chu et al., 2017). Gene *TARID* participates in liver cell metabolism (Yuan et al., 2016). Gene *TRAPPC10* is associated with the toxic effect of octylphenol on the expression of genes in the liver (Li et al., 2014).

Case Two: The Alzheimer Disease Neuroimaging Initiative Data Analysis

We also applied the developed OMGA method to the ADNI data set which can be accessed at <http://adni.loni.usc.edu/>. From the ADNI1 and ADNI2 studies, we selected 490 samples with complete genetic and phenotypic information. We deleted SNPs with MAF < 0.05 or those that could not pass the Hardy–Weinberg equilibrium test. This ended up with 620,901 SNPs.



We included SNPs within 20 kb upstream and downstream of each gene and mapped them to 22,890 genes according to human genome version GRCh38.

Alzheimer's disease (AD) is a central nervous system degenerative disease with insidious onset and chronic progress and has affected over 5.5 million Americans, especially among the elderly over the age of 65 years. ADNI provides pre-calculated volumes of five cortical regions including entorhinal, hippocampus, ventricles, midtemp, and fusiform. Brain atrophy is a typical clinical symptom among AD patients (Ferrarini et al., 2006). Studies have pointed out that the volumes in the different cortical regions show different rates of decline and are functionally related to AD. For example, the hippocampus region helps humans to deal with memory sounds, long-term learning, and taste and is a sensitive early indicator of AD (Mu and Gage, 2011). The loss in the entorhinal region is highly correlated with the severity of AD and the loss is obvious even in mild AD patients (Juottonen et al., 1998). Similarly, the volumes in the regions fusiform and midtemp also slightly decrease in AD patients (Thambisetty et al., 2011). This motivates us to take the volumes of the five cortical regions as a multi-trait response and to identify which genes are associated with the volume variation in the different brain regions.

We first conducted the marginal single-trait analysis with the proposed gene-based omnibus kernel testing approach. We log-transformed the volumes of the five cortical regions and took the age, education level, gender, and *APOE4* alleles as the covariates. The Q-Q plot of the gene-based single-trait analysis is shown in **Supplementary Figure S3**. No sign of *p*-value inflation was observed. Also, there is no strong indication of significant signals either. Then, we carried out the multi-trait analysis which can more accurately reflect the brain atrophy in AD patients.

We also applied MANOVA and RMMLR methods for multi-trait analysis. The Q-Q plot of the multi-trait analysis results by OMGA is shown in **Figure 4**. There is no significant indication of *p*-value inflation.

Again no significant genes were identified based on the genome-wide gene-level Bonferroni threshold. Here, we listed the top 12 genes based on a suggestive threshold of 5×10^{-5} in **Table 3**. From the single-trait analysis, we found eight, 10, 10, five, and six genes associated with the regions entorhinal, ventricles, hippocampus, fusiform, and midtemp, respectively (see **Supplementary Table S2** for a detailed list of the genes). Two genes (*SNORA30* and *TLR4*) that were not in the single-trait analysis list but showed up in the multi-trait analysis list are highlighted in bold font in **Table 3**. Compared to RMMLR and MANOVA analyses, the *p*-values by OMGA are uniformly smaller, indicating the power of OMGA by taking both linear and nonlinear effect into consideration.

For the 12 genes associated with multi-trait of brain atrophy in AD patients, some of them have been reported in the literature. For example, gene *RBM45*, known as the RNA-binding motif protein 45 or developmentally regulated RNA-binding protein-1 (*Drbp1*), has been shown to be associated with the degenerative neurological changes in AD patients (Eck et al., 2018). Gene *UPK1B* has been shown to be cooperated with *CD9* and *CD81* and is directly involved in the pathological process of AD (De Strooper and Wakabayashi, 2011; Orre et al., 2014; Wężyk and Żekanowski, 2017). Mutation in gene *TLR4* reduces microglial activation, increases A β deposits, and exacerbates cognitive deficits in a mouse model of AD (Song et al., 2011). A study showed that polymorphisms in gene *TLR4* and *CD14* were closely related to AD (Balistreri et al., 2008). Others reported the increasing expressions of *TLR2* and *TLR4* on the peripheral blood mononuclear cells of AD patients

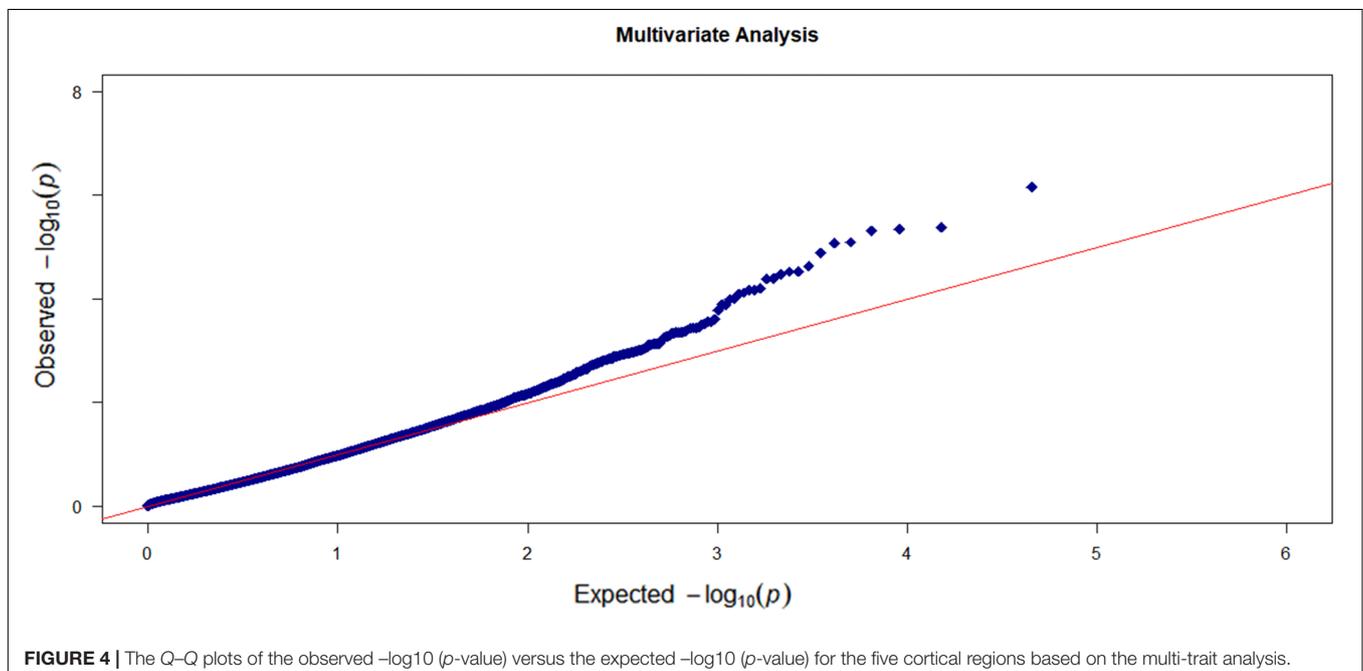


TABLE 3 | List of top genes and the p -values with different methods in the Alzheimer Disease Neuroimaging Initiative study.

Gene name	Number of single nucleotide polymorphisms	Chr	OMGA	RMMLR	MANOVA
<i>TMEM26-AS1</i>	731	10	3.45E-06	0.0004	0.2572
<i>TPRG1-AS2</i>	320	3	6.60E-06	0.0238	0.4595
<i>ST3GAL4</i>	2,457	11	8.37E-06	0.1373	0.0165
<i>LMNTD1</i>	89	12	9.64E-06	0.6580	0.1698
<i>OR4F5</i>	2,234	1	1.03E-05	0.1887	0.1364
<i>MIR6723</i>	170	14	1.83E-05	0.5421	0.2648
<i>RBM45</i>	468	2	2.25E-05	0.0017	0.0077
<i>ADAMTS7P1</i>	1,444	15	2.29E-05	0.0003	0.3606
<i>SNORA30</i>	200	16	2.30E-05	0.0213	0.0093
<i>TLR4</i>	153	9	3.45E-05	0.0015	0.1364
<i>C5orf46</i>	663	5	3.69E-05	0.1254	0.0232
<i>UPK1B</i>	772	3	4.10E-05	0.1855	0.0036

(Zhang et al., 2012). These empirical evidences support the results of the analysis.

DISCUSSION

Increasing evidence has shown that, for correlated phenotypes, multi-trait analysis can significantly increase the power of association analysis (e.g., He et al., 2013; Schifano et al., 2013; Wang, 2014). Given that genes are functional units in most living organisms, we proposed a rapid and powerful gene-based multi-trait analysis method. Our method is developed under the KBT framework without specific error distribution assumptions. It possesses a few advantages over existing methods. First, the method achieves fast calculation speed and decreases the computational burden for high-dimensional data. A testing p -value can be quickly computed with the asymptotic results, making the method computationally attractive. Second, it can capture a potential nonlinear effect within genes by using a nonparametric KBT procedure. By incorporating different kernel functions, potential linear or nonlinear genetic effects can be captured and tested. When a given series of candidate kernel functions is available, the omnibus testing procedure is robust against misspecification of kernel functions. Moreover, it is built upon the Cauchy transformation and is computationally fast (Liu and Xie, 2019). Thus, the proposed method enjoys both theoretical rigor and computational efficiency and can be widely used in gene-based analysis.

We conducted extensive simulation studies to evaluate the type I error control and the power of the proposed method. The results show that the proposed OMGA method can maintain a reasonable type 1 error rate and achieve great power compared to other popular methods such as MANOVA and RMMLR. Furthermore, the omnibus testing procedure incorporating different kernels performs as well as if the underlying true genetic function is correctly specified. Thus, the method is safe to apply in real applications regardless of the underlying disease function, making the method practically attractive.

For multi-trait analysis, there are two different frameworks proposed. One is to jointly model multiple traits as a multivariate response and further assess their association with SNP variants. This framework can directly take correlation information into consideration. Methods for such type of multi-trait analysis include the RMMLR and the MANOVA methods as discussed in this work and many others (e.g., Maity et al., 2012). Another framework is to conduct a single-trait disease-gene association test and then combine p -values to assess the joint association. The method developed by Yang et al. (2016) falls into this category. Nevertheless, methods to combining p -values have to take the correlation information into consideration. Otherwise, the results can be biased. Ideally, the first framework should be preferable since it models multiple traits simultaneously in one joint model. On the other hand, the second framework has its advantages. For example, it can be computationally less expensive and ease theoretical evaluations. Especially with the proposed method in this work, the second framework can be a better choice since the asymptotic evaluation of the joint association statistics can be theoretically challenging or may not even be feasible.

Our method can be easily applied to a genome-wide pathway-based multi-trait analysis. It is known that genes usually do not work alone. For example, cellular pathways and complex molecular networks are often more directly involved in the progression and the susceptibility of diseases. Thus, a pathway-based analysis can shed light on the mechanics of complex diseases. On the other hand, the current study only focused on quantitative multivariate phenotypes. It can be extended to qualitative response variables or a combination of qualitative and quantitative phenotypes. However, the extension is non-trivial and will be studied in our future investigation. The R code that implements the method can be found in GitHub at <https://github.com/yamin-19/OMGA>.

DATA AVAILABILITY STATEMENT

The HLC dataset can be downloaded at <https://www.synapse.org/#Synapse:syn4499>. The ADNI dataset can be accessed through <http://adni.loni.usc.edu/>.

AUTHOR CONTRIBUTIONS

YD implemented the method and drafted the manuscript. TH derived the kernel testing method. RF, SL, and HC were involved in the simulation and data analysis. YC conceived the idea, designed the study, and drafted the manuscript. All the authors read and approved the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.00437/full#supplementary-material>

REFERENCES

- Aguar, M., Masse, R., and Gibbs, B. F. (2005). Regulation of cytochrome P450 by posttranslational modification. *Drug Metab. Rev.* 37, 379–404. doi: 10.1081/dmr-200046136
- Amos, C., and Laing, A. J. G. E. (1993). A comparison of univariate and multivariate tests for genetic linkage. *Genetic Epidemiol.* 10, 671–676. doi: 10.1002/gepi.1370100657
- Balistreri, C., Grimaldi, M., Chiappelli, M., Licastro, F., Castiglia, L., Listi, F., et al. (2008). Association between the polymorphisms of TLR4 and CD14 genes and Alzheimer's disease. *Curr. Pharm. Design* 14, 2672–2677. doi: 10.2174/138161208786264089
- Basu, S., and Pan, W. J. G. E. (2011). Comparison of statistical tests for disease association with rare variants. *Gen. Epidemiol.* 35, 606–619. doi: 10.1002/gepi.20609
- Basu, S., Zhang, Y., Ray, D., Miller, M. B., Iacono, W. G., and McGue, M. J. H. H. (2013). A rapid gene-based genome-wide association test with multivariate traits. *Hum. Hered.* 76, 53–63. doi: 10.1159/000356016
- Bilodeau, M. (2013). *Analysis of Variance, Multivariate (MANOVA)*. Hoboken, NJ: John Wiley & Sons.
- Chu, C., Zhao, C., Zhang, Z., Wang, M., Zhang, Z., Yang, A., et al. (2017). Transcriptome analysis of primary aldosteronism in adrenal glands and controls. *PLoS One* 12, e0170018. doi: 10.1371/journal.pone.0170018
- Cui, Y., Kang, G., Sun, K., Qian, M., Romero, R., and Fu, W. J. G. (2008). Gene-centric genomewide association study via entropy. *Genetics* 179, 637–650. doi: 10.1534/genetics.107.082370
- De Strooper, B., and Wakabayashi, T. (2011). *Extracellular Targets for Alzheimer's Disease*. Google Patents. US20110008350A1.
- Eck, A. G., Lopez, K. J., and Henderson, J. O. J. J. (2018). RNA-binding motif protein 45 (Rbm45)/developmentally regulated RNA-binding protein-1 (Drbp1): association with neurodegenerative disorders. *J. Stud. Res.* 7, 33–37.
- Ferrarini, L., Palm, W. M., Olofson, H., van Buchem, M. A., Reiber, J. H., and Admiraal-Behloul, F. (2006). Shape differences of the brain ventricles in Alzheimer's disease. *Neuroimage* 32, 1060–1069. doi: 10.1016/j.neuroimage.2006.05.048
- He, Q., Avery, C. L., and Lin, D. Y. J. G. E. (2013). A general framework for association tests with multivariate traits in large-scale genomics studies. *Genetic Epidemiol.* 37, 759–767. doi: 10.1002/gepi.21759
- He, T., Li, S., Zhong, P. S., and Cui, Y. (2019). An optimal kernel-based U-statistic method for quantitative gene-set association analysis. *Gen. Epidemiol.* 43, 137–149. doi: 10.1002/gepi.22170
- Jiang, C., and Zeng, Z.-B. J. G. (1995). Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 140, 1111–1127.
- Juottonen, K., Laakso, M., Insausti, R., Lehtovirta, M., Pitkänen, A., Partanen, K., et al. (1998). Volumes of the entorhinal and perirhinal cortices in Alzheimer's disease. *Neurobiol. Aging* 19, 15–22. doi: 10.1016/s0197-4580(98)00007-4
- Kim, J., Bai, Y., and Pan, W. J. G. E. (2015). An adaptive association test for multiple phenotypes with GWAS summary statistics. *Gen. Epidemiol.* 39, 651–663. doi: 10.1002/gepi.21931
- Kwee, L. C., Liu, D., Lin, X., Ghosh, D., and Epstein, M. P. J. (2008). A powerful and flexible multilocus association test for quantitative traits. *Am. J. Hum. Gen.* 82, 386–397. doi: 10.1016/j.ajhg.2007.10.010
- Li, S., and Cui, Y. (2012). Gene-centric gene-gene interactions: a model-based kernel machine method. *Ann. Appl. Stat.* 6, 1134–1161. doi: 10.1214/12-aos454
- Li, X.-Y., Xiao, N., and Zhang, Y.-H. J. E. (2014). Toxic effects of octylphenol on the expression of genes in liver identified by suppression subtractive hybridization of *Rana chensinensis*. *Ecotoxicology* 23, 1–10. doi: 10.1007/s10646-013-1144-z
- Lin, X., Lee, S., Christiani, D. C., and Lin, X. J. B. (2013). Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* 14, 667–681. doi: 10.1093/biostatistics/kxt006
- Littell, R. C., and Folks, J. L. (1971). Asymptotic optimality of Fisher's method of combining independent tests. *J. Am. Stat. Assoc.* 66, 802–806. doi: 10.1080/01621459.1971.10482347
- Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., et al. (2010). A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Gen.* 87, 139–145. doi: 10.1016/j.ajhg.2010.06.009
- Liu, W., and Xie, J. (2019). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *J. Am. Stat. Assoc.* 114, 384–392.
- Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E., and Lin, X. (2019). ACAT: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am. J. Hum. Genet.* 104, 410–421. doi: 10.1016/j.ajhg.2019.01.002
- Maitly, A., Sullivan, P. F., and Tzeng, J. Y. (2012). Multivariate phenotype association analysis by marker-set kernel machine regression. *Genet. Epidemiol.* 36, 686–695. doi: 10.1002/gepi.21663
- Marceau, R., Lu, W., Holloway, S., Sale, M. M., Worrall, B. B., Williams, S. R., et al. (2015). A fast multiple-kernel method with applications to detect gene-environment interaction. *Gen. Epidemiol.* 39, 456–468. doi: 10.1002/gepi.21909
- Mu, Y., and Gage, F. H. (2011). Adult hippocampal neurogenesis and its role in Alzheimer's disease. *Mol. Neurodegrad.* 6:85. doi: 10.1186/1750-1326-6-85
- Mukhopadhyay, I., Feingold, E., Weeks, D. E., and Thalamuthu, A. J. G. E. T. (2010). Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Gen. Epidemiol.* 34, 213–221.
- Neale, B. M., and Sham, P. C. J. T. (2004). The future of association studies: gene-based analysis and replication. *Am. J. Hum. Gen.* 75, 353–362. doi: 10.1086/423901
- Nelson, D. R., Koymans, L., Kamataki, T., Stegeman, J. J., Feyereisen, R., Waxman, D. J., et al. (1996). P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. *Eur. PMC* 6, 1–42. doi: 10.1097/00008571-199602000-00002
- Orre, M., Kamphuis, W., Osborn, L. M., Jansen, A. H., Kooijman, L., Bossers, K., et al. (2014). Isolation of glia from Alzheimer's mice reveals inflammation and dysfunction. *Neurobiol. Aging* 35, 2746–2760. doi: 10.1016/j.neurobiolaging.2014.06.004
- Plant, N. (2007). The human cytochrome P450 sub-family: transcriptional regulation, inter-individual variation and interaction networks. *Biochim. Biophys. Acta Gen. Sub.* 1770, 478–488. doi: 10.1016/j.bbagen.2006.09.024
- Schifano, E. D., Li, L., Christiani, D. C., and Lin, X. (2013). Genome-wide association analysis for multiple continuous secondary phenotypes. *Am. J. Hum. Gen.* 92, 744–759. doi: 10.1016/j.ajhg.2013.04.004
- Shang, J., Zhang, J., Lei, X., Zhao, W., and Dong, Y. J. G. (2013). EpiSIM: simulation of multiple epistasis, linkage disequilibrium patterns and haplotype blocks for genome-wide interaction analysis. *Genes Genomics* 35, 305–316. doi: 10.1007/s13258-013-0081-9
- Song, M., Jin, J., Lim, J.-E., Kou, J., Pattanayak, A., Rehman, J. A., et al. (2011). TLR4 mutation reduces microglial activation, increases A β deposits and exacerbates cognitive deficits in a mouse model of Alzheimer's disease. *J. Neuroinflamm.* 8:92. doi: 10.1186/1742-2094-8-92
- Stearns, F. W. J. G. (2010). One hundred years of pleiotropy: a retrospective. *Genetics* 186, 767–773. doi: 10.1534/genetics.110.122549
- Thambisetty, M., Simmons, A., Hye, A., Campbell, J., Westman, E., Zhang, Y., et al. (2011). Plasma biomarkers of brain atrophy in Alzheimer's disease. *PLoS One* 6:e28527.
- Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., et al. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Gen.* 50:229. doi: 10.1038/s41588-017-0009-4
- Wang, K. (2014). Testing genetic association by regressing genotype over multiple phenotypes. *PLoS One* 9:e106918. doi: 10.1371/journal.pone.0106918
- Wang, K., Li, M., and Bucan, M. J. (2007). Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Gen.* 81, 1278–1283. doi: 10.1086/522374
- Wang, K., Li, M., and Hakonarson, H. J. (2010). Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* 11:843. doi: 10.1038/nrg2884
- Wei, C., and Lu, Q. (2017). A generalized association test based on U statistics. *Bioinformatics* 33, 1963–1971. doi: 10.1093/bioinformatics/btx103
- Wężyk, M. M., and Żekanowski, C. J. (2017). "Presenilins interactome in Alzheimer disease and pathological ageing," in *Senescence: Physiology or Pathology*, Vol. 95, eds J. Dorszewska and W. Kozubski (London: InTechOpen).
- Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., et al. (2010). Powerful SNP set analysis for case-control genome wide association studies. *Am. J. Hum. Genet.* 86, 929–942. doi: 10.1016/j.ajhg.2010.05.002
- Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. J. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89, 82–93. doi: 10.1016/j.ajhg.2011.05.029
- Yang, J. J., Li, J., Williams, L. K., and Buu, A. (2016). An efficient genome-wide association test for multivariate phenotypes based on the Fisher combination function. *BMC Bioinform.* 17:19.

- Yang, Q., and Wang, Y. J. J. (2012). Methods for analyzing multivariate phenotypes in genetic association studies. *J. Probab. Stat.* 2012: 652569.
- Yuan, S.-X., Zhang, J., Xu, Q.-G., Yang, Y., and Zhou, W. (2016). Long noncoding RNA, the methylation of genomic elements and their emerging crosstalk in hepatocellular carcinoma. *Cancer Lett.* 379, 239–244. doi: 10.1016/j.canlet.2015.08.008
- Zhang, F., Guo, X., Wu, S., Han, J., Liu, Y., Shen, H., et al. (2012). Genome-wide pathway association studies of multiple correlated quantitative phenotypes using principle component analyses. *PLoS One* 7:e53320. doi: 10.1371/journal.pone.0053320
- Zhang, W., Wang, L.-Z., Yu, J.-T., Chi, Z.-F., and Tan, L. J. (2012). Increased expressions of TLR2 and TLR4 on peripheral blood mononuclear cells from patients with Alzheimer's disease. *J. Neurol. Sci.* 315, 67–71. doi: 10.1016/j.jns.2011.11.032

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Deng, He, Fang, Li, Cao and Cui. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.