frontiers in Genetics

METHODS
published: 22 March 2022
doi: 10.3389/fgene.2022.791920

A Novel Hierarchical Clustering Approach for Joint Analysis of Multiple Phenotypes Uncovers Obesity Variants Based on ARIC

Liwan Fu[1,2], *Yuquan Wang*[2], *Tingting Li*[2], *Siqian Yang*[2] *and Yue-Qing Hu*[2,3]*

[1]Center for Non-communicable Disease Management, National Center for Children's Health, Beijing Children's Hospital, Capital Medical University, Beijing, China, [2]State Key Laboratory of Genetic Engineering, Human Phenome Institute, Institute of Biostatistics, School of Life Sciences, Fudan University, Shanghai, China, [3]Shanghai Center for Mathematical Sciences, Fudan University, Shanghai, China

Genome-wide association studies (GWASs) have successfully discovered numerous variants underlying various diseases. Generally, one-phenotype one-variant association study in GWASs is not efficient in identifying variants with weak effects, indicating that more signals have not been identified yet. Nowadays, jointly analyzing multiple phenotypes has been recognized as an important approach to elevate the statistical power for identifying weak genetic variants on complex diseases, shedding new light on potential biological mechanisms. Therefore, hierarchical clustering based on different methods for calculating correlation coefficients (HCDC) is developed to synchronously analyze multiple phenotypes in association studies. There are two steps involved in HCDC. First, a clustering approach based on the similarity matrix between two groups of phenotypes is applied to choose a representative phenotype in each cluster. Then, we use existing methods to estimate the genetic associations with the representative phenotypes rather than the individual phenotypes in every cluster. A variety of simulations are conducted to demonstrate the capacity of HCDC for boosting power. As a consequence, existing methods embedding HCDC are either more powerful or comparable with those of without embedding HCDC in most scenarios. Additionally, the application of obesity-related phenotypes from Atherosclerosis Risk in Communities *via* existing methods with HCDC uncovered several associated variants. Among these, *UQCC1*-rs1570004 is reported as a significant obesity signal for the first time, whose differential expression in subcutaneous fat, visceral fat, and muscle tissue is worthy of further functional studies.

Keywords: GWAS, hierarchical clustering, multiple phenotypes, obesity, bioinformatics

OPEN ACCESS

Edited by:
Can Yang,
Hong Kong University of Science and Technology, Hong Kong SAR, China

Reviewed by:
Limin Li,
Xi'an Jiaotong University, China
Yiming Hu,
Yale University, United States

***Correspondence:**
Yue-Qing Hu
yuehu@fudan.edu.cn

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

Received: *21 October 2021*
Accepted: *27 January 2022*
Published: *22 March 2022*

Citation:
Fu L, Wang Y, Li T, Yang S and Hu Y-Q (2022) A Novel Hierarchical Clustering Approach for Joint Analysis of Multiple Phenotypes Uncovers Obesity Variants Based on ARIC. Front. Genet. 13:791920. doi: 10.3389/fgene.2022.791920

INTRODUCTION

The applications of genome-wide association studies (GWASs) have successfully established a large number of genetic variants associated with numerous complex diseases (Lutz et al., 2017), contributing to the understanding of the mechanisms of complex diseases such as obesity (Locke et al., 2015; Shungin et al., 2015). Notably, GWASs usually apply the univariate analysis to examine the association between genetic variants and a single phenotype, and in general, multiple phenotypes related to diseases are typically collected together for better understanding the

physiological process of diseases (Yang et al., 2010). For example, information about individual status of obesity, insulin resistance, hypertension, and atherosclerotic dyslipidemia is required jointly to explore metabolic syndrome (Sattar et al., 2008). A research of hypertension inevitably takes account of the magnitude of systolic blood pressure (SBP) and diastolic blood pressure (DBP) (Yang and Wang, 2012), From the aspect of pleiotropy, namely, some genes could simultaneously affect multiple related phenotypes, the significance of biological process emphasizes the importance of multiple phenotypes analyses. Univariate analysis means conducting single phenotype separately and showing the outcomes for each phenotype (O'Reilly et al., 2012). However, analyzing one phenotype at each time will absolutely suffer multiple testing corrections, which results in a power loss in GWASs (Yang et al., 2010). Recently, jointly analyzing multiple phenotypes together has become popular due to its increased statistical power of identifying genetic variants compared to analyzing each phenotype separately, enhancing the magnitude of explanation for the biological progress of relevant diseases, and elevating the credibility of the results (Yang et al., 2010; Aschard et al., 2014; Fu et al., 2021).

In the past decade, joint analysis of multiple phenotypes has developed rapidly, which may roughly be classified into three categories: regression approaches, integrating testing statistics from univariate analyses, and variable reduction approaches (Yang and Wang, 2012). Tests that fall into the first category, regression approaches, mainly encompass three different methods to analyze the association of multiple phenotypes with a genetic variant: mixed effect models (Bates and DebRoy, 2004), frailty models (Therneau et al., 2003), and generalized estimating equations (Zeger and Liang, 1986). In the second category, integrating testing statistics from univariate analyses, as the name suggests, integrates different test statistics or *p*-values from univariate association analyses via various strategies (Schaid et al., 2016; Yang et al., 2016). Nowadays, various approaches of integrating test statistics or *p*-values from univariate analyses have been established to investigate the association between genetic variants and multiple phenotypes concerning the correlation structure among phenotypes (van der Sluis et al., 2013; Kwak and Pan, 2016; Liang et al., 2016; Yang et al., 2016). In the last category, tests on the basis of variable reduction approaches roughly adopt three dimension reduction techniques. The first one is the principal component analysis (PCA) (Aschard et al., 2014). In PCA, the first few principal components (PCs) with regard to majority of the total phenotype variance are selected for evaluating their association with a genetic variant. The second one is the canonical correlation analysis (CCA) (Tang and Ferreira, 2012). CCA supplies an efficient and powerful method for both univariate and multivariate analyses ignoring the need for permutation test in association studies by searching for linear combinations that maximize the association between two classes of multidimensional variables. The last one is the principal component of heritability (PCH) (Ott and Rabinowitz, 1999; Klei et al., 2008; Wang et al., 2016). PCH adopts a linear combination of phenotypes that represents the highest heritability among all linear combinations of phenotypes for reducing multiple phenotypes.

In this study, we develop a novel variable reduction approach called hierarchical clustering based on different methods for calculating correlation coefficients (HCDC) aiming at jointly analyzing multiple phenotypes. By means of a dimension reduction technique, HCDC constructs a typical phenotype from each cluster of phenotypes, then applies the existing approaches for jointly analyzing multiple phenotypes to estimate the genetic associations with the typical phenotypes instead of the individual phenotypes. The vital significance in dimension reduction technique of HCDC is that when one cluster is composed of positively highly correlated phenotypes, every linear combination of phenotypes is a representative of the cluster reasonably (Bien and Wegkamp, 2013; Bühlmann et al., 2013). One specific advantage of HCDC is that it does not need to know individual phenotypes, and it actually requires a similarity matrix about the phenotypes. In real data analysis, the similarity matrix of phenotypes can be evaluated from the summary statistical values with regard to the usage of independent single nucleotide polymorphisms (SNPs) in a GWAS (Zhu et al., 2015). Previously, hierarchical clustering method (HCM) is also a clustering approach (Liang et al., 2018). However, when calculating the correlation coefficients between distinct clusters, HCM adopts the uniform expression of correlation coefficients, not concerning the number of phenotypes in each cluster. As a result, HCM obtains lower statistical power in some scenarios. On the contrary, we propose HCDC by virtual extensive simulations to reveal the validity of the improved two-step approach and to explore its power. Notably, the performance of three existing approaches employing HCDC or HCM, namely, multivariate analysis of variance (MANOVA) (Cole and MaxwellScott, 1994), joint model of multiple phenotypes (MultiPhen) (O'Reilly et al., 2012), trait-based association test that uses extended Simes procedure (TATES) (van der Sluis et al., 2013), is compared with that of without employing HCDC or HCM. In this way, scientific issues about whether there exists an advantage of clustering (MANOVA, MultiPhen, and TATES using HCDC or HCM are compared with these approaches without using HCDC or HCM) and which clustering approach has more obviously outstanding performance (MANOVA, MultiPhen, and TATES using HCDC are compared with these approaches using HCM) can be solved. Our simulations reveal that MANOVA, MultiPhen, and TATES employing HCDC have correct type I error rates and possess more power than MANOVA, MultiPhen, and TATES without employing HCDC in most simulation scenarios. Finally, we emphatically explore the performance of HCDC approach by utilizing the obesity-related phenotypes from a real dataset, Atherosclerosis Risk in Communities (ARIC) Study (Author Anonymous, 1989) from dbGaP. Consequently, a total of eight significant SNPs are detected, and subsequent bioinformatics analysis is carried out for better understanding the results. From another point of view, the interesting results indicate the effective performance of HCDC in real data application.

METHODS

Proposed HCDC

Assume a sample with N individuals, and M phenotypes $Y_1, Y_2, \ldots, Y_M$. Meanwhile, let $X = (x_1, \ldots, x_N)^{\mathrm{T}}$ denote the genotypic score of N individuals at a genetic variant of interest, where $x_i \in \{0, 1, 2\}$ represents the number of minor alleles that i th subject carries at that variant.

Note that the key issue in the hierarchical clustering is to specify a measure of similarity between disjoint groups of phenotypes. Now let us take two disjoint clusters G_1 and G_2 of phenotypes as an example to demonstrate the calculation of similarity between these two groups. Denote M_1 and M_2 as the numbers of phenotypes in G_1 and G_2, respectively.

1. If $M_1 = M_2 = 1$, Pearson correlation coefficient (Jin and Lin, 2019) between two phenotypes is calculated to represent the similarity between G_1 and G_2.
2. If $M_1 = 1$ and $M_2 > 1$, or $M_1 > 1$ and $M_2 = 1$ multiple correlation coefficient (Cohen and Cohen, 1983; Jin and Lin, 2019) is employed based on the phenotypes involved in G_1 and G_2, respectively, to reveal the similarity between a pair of clusters.
3. If $M_1 > 1$ and $M_2 > 1$, canonical correlation coefficient (Ferreira and Purcell, 2009) is applied according to the phenotypes involved in G_1 and G_2 respectively to show the similarity between two clusters.

Once we have the similarity measure between two clusters of phenotypes, we apply a hierarchical clustering approach to cluster the phenotypes. Specifically, following the agglomerative (bottom–up) procedure, we start at the bottom (i.e., the lowest level) where each phenotype is a cluster and then recursively merge a selected pair of clusters with the biggest intergroup similarity at the next lower level into a single cluster. This produces a grouping at the next higher level with one less cluster until all phenotypes are grouped as one cluster at the highest level. Finally, there are $M - 1$ levels in the hierarchy.

For any b, $1 \le b \le M - 1$, let h_b denote the height at the level b in the dendrogram, which is the biggest intergroup similarity at the level $b - 1$. Similar to a proposed principle (Bühlmann et al., 2013), a stopping criterion is adopted to determine the optimal number K of clusters,

$$K = \arg \min_{1 \le b \le M-2} (h_{b+1} - h_b).$$

Without loss of generality, the corresponding K clusters are denoted as $G_1, G_2, \ldots, G_K$.

The established HCDC encompasses the following two steps. First, M phenotypes are grouped into K clusters as aforementioned, and each of the K clusters singles out a representative phenotype. Second, existing approaches to the K representative phenotypes instead of the original M phenotypes are employed to evaluate the genetic association of multiple phenotypes with a genetic variant.

Notice that each phenotype should be scaled first before constructing the representative phenotype for each other. We define the representative phenotype for the kth cluster as the mean phenotype values in the cluster, namely

$$\bar{Y}^{(k)} = \frac{1}{M_k} \sum_{m \in G_k} Y_m, k = 1, \ldots, K,$$

where M_k is the number of phenotypes in the cluster $G_k, k = 1, 2, \ldots, K$. Denote $\bar{Y}$ as the $N \times K$ design matrix whose kth column is given by $\bar{Y}^{(k)}$. Then, existing approaches are employed to evaluate the association between $\bar{Y}$ and X.

The source code for HCDC approach can be found in https://github.com/YQHuFD/HCDC.

Comparison of Methods

For convenience, let $\mathbf{1}_n$ denote the ones vector of length n and $\mathbf{0}_n$ represent the all zeroes vector of length n, where n is a positive integer. First, we need to introduce one of the potential competitors, HCM (Liang et al., 2018). Same as the process of HCDC, HCM also adopts the bottom–up hierarchical clustering method on the basis of the similarity. But unlike HCDC, HCM defines the similarity matrix with S_{ij}, where S_{ij} is the $i\,j$th entry of the sample correlation matrix of M phenotypes $Y_1, Y_2, \ldots, Y_M$. The average linkage is employed as the similarity between two clusters in HCM. To be precise, the similarity between clusters G_k and G_l (which are two disjoint subsets of $\{1, 2, \ldots, M\}$) is given by

$$\frac{1}{M_k \cdot M_l} \sum_{i \in G_k, j \in G_l} S_{ij},$$

where M_k and M_l are the numbers of phenotypes in the respective clusters G_k and G_l, $1 \le k, l \le K$.

Except the different definition of similarity between pairs of clusters, the remaining processes of HCM are exactly the same as the HCDC. Second, the performance of MANOVA (Cole and MaxwellScott, 1994), MultiPhen (O'Reilly et al., 2012), and TATES (van der Sluis et al., 2013) with using HCDC is compared with that of with using HCM and that of without using HCDC/HCM approaches. The ones with employing HCDC and HCM are referred as HCDCMANOVA, HCMANOVA, HCDCMultiPhen, HCMultiPhen, HCDCTATES, and HCTATES, respectively. In the following, we briefly review the existing approaches for easy reference.

MANOVA (multivariate analysis of variance) (Cole and MaxwellScott, 1994): A total of M phenotypes are involved in the standard MANOVA and the background variance–covariance matrix Σ including $M \times M$ symmetrical elements is unconstrained. There are $((M + 1) \times M)/2$ freely evaluated elements in the covariances and variances. Standard MANOVA tests the null hypothesis that the M regression coefficients are all zeroes, which asymptotically follows F distribution.

MutiPhen (joint model of multiple phenotypes) (O'Reilly et al., 2012): In the MultiPhen model, the genotypes and phenotypes are treated as ordinal response and predictors, respectively. Likelihood ratio test is performed to test the null hypothesis in the proportional odds logistic regression.

TATES (trait-based association test that uses extended Simes procedure) (van der Sluis et al., 2013): The p-values from

univariate analysis is integrated to get a comprehensive p-value, and simultaneously, correlation between phenotypes is considered for adjustment. Denote $\min(M_e p_{(j)}/M_{e(j)})$ as the p-value of TATES, where $p_{(j)}$ represents the j^{th} $(j = 1, \ldots, M)$ ascending sorted p-value; M_e and $M_{e(j)}$ are the effective number of independent p-values among all involved M phenotypes and j specific phenotypes, respectively. The correlation matrix of p-values is derived to obtain the effective numbers.

RESULTS

Simulation Studies

Suppose that a population is in Hardy–Weinberg equilibrium (HWE), and we generate the genotypes of the genetic variants following the binomial distribution with parameter two and the minor allele frequency (MAF). This simulation study sets MAF = 0.3 in most scenarios. We generate multiple phenotypes by means of the following factor model (van der Sluis et al., 2013):

$$y = \lambda x + dc\gamma f + d\sqrt{1 - c^2} \times \varepsilon,$$

where $y = (y_1, \ldots, y_M)^T$ denotes the M phenotypes; x is the genotype; $\lambda = (\lambda_1, \ldots, \lambda_M)^T$ represents the vector of values suggesting the effects of genetic variant on the M phenotypes; f shows the vector of factors; $f = (f_1, \ldots, f_R)^T \sim MVN(0, \Sigma)$, $\Sigma = (1 - \rho)I + \rho \mathbf{1}_R \mathbf{1}_R^T$; I is the identity matrix; R represents the number of factors, and ρ is the correlation between factors; γ is an $M \times R$ matrix; d is a diagonal matrix for correcting the variance of phenotypes; c denotes a constant; $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_M)^T$ represents a vector of random errors, and $\varepsilon_1, \ldots, \varepsilon_M$ are mutually independent and follow the standard normal distributions. Consider the following four models with different numbers of factors affected by genotypes.

Model 1: There is only one factor, and the genotype has an effect on all phenotypes with the same effect size. That is, $R = 1$, $\lambda = \beta \mathbf{1}_M$, $d = diag(\mathbf{1}_M)$, and $\gamma = \mathbf{1}_M$.

Model 2: There are two factors and the genotype impacts on one factor with the same effect. Namely, $R = 2$, $\lambda = (\mathbf{0}_{M/2}^T, \beta \mathbf{1}_{M/2}^T)^T$, $d = diag(\mathbf{1}_M)$, and $\gamma = bdiag(\mathbf{1}_{M/2}, \mathbf{1}_{M/2})$, which is the block diagonal matrix of $\mathbf{1}_{M/2}$ and $\mathbf{1}_{M/2}$.

Model 3: There are four factors, and the genotype has an effect on the last two factors with varied effect directions. That is, $R = 4$, $\lambda = (\mathbf{0}_{M/2}^T, -\beta \mathbf{1}_{3M/16}^T, \beta \mathbf{1}_{M/4}^T, -\beta \mathbf{1}_{M/16}^T)^T$,

$$\gamma = bdiag\left((\mathbf{1}_{3M/16}^T, -\mathbf{1}_{M/16}^T)^T, (\mathbf{1}_{3M/16}^T, -\mathbf{1}_{M/16}^T)^T, (\mathbf{1}_{3M/16}^T, -\mathbf{1}_{M/16}^T)^T, (\mathbf{1}_{3M/16}^T, -\mathbf{1}_{M/16}^T)^T \right),$$

and

$$d = diag\left(\left(\frac{8}{M}[1{:}\ M/4]^T, \frac{8}{M}[1{:}\ M/4]^T, \frac{8}{M}[1{:}\ M/4]^T, \frac{8}{M}[1{:}\ M/4]^T \right)^T \right)$$

where $[1{:}\ M/4]$ denotes the vectors of components $1, 2, \ldots, M/4$.

Model 4: There are four factors, and the genotype has an influence on the last three factors with different sizes. Namely, $R = 4$,

$$\lambda = \left(\mathbf{0}_{M/4}^T, \frac{2\beta}{M/4 + 1}[1{:}\ M/4]^T, -\beta \mathbf{1}_{3M/16}^T, \beta \mathbf{1}_{M/4}^T, -\beta \mathbf{1}_{M/16}^T \right)^T$$

$$\gamma = bdiag\left((\mathbf{1}_{3M/16}^T, -\mathbf{1}_{M/16}^T)^T, \mathbf{1}_{M/4}^T, (\mathbf{1}_{3M/16}^T, -\mathbf{1}_{M/16}^T)^T, (\mathbf{1}_{3M/16}^T, -\mathbf{1}_{M/16}^T)^T \right),$$

and

$$d = diag\left(\left(\frac{8}{M}[1{:}\ M/4]^T, \frac{8}{M}[1{:}\ M/4]^T, \frac{8}{M}[1{:}\ M/4]^T, \frac{8}{M}[1{:}\ M/4]^T \right)^T \right)$$

For the all models, the within-factor correlation is c^2, and the between-factor correlation is ρc^2. For evaluating type I error rates and powers, this study sets $N = 2{,}000$ unrelated individuals, and the number of phenotypes $M = 16, 32$. According to $\beta = 0$, all phenotypes independent of genotypes are generated to estimate the type I error rates of all investigated approaches, encompassing MANOVA, MultiPhen, TATES, HCMANOVA, HCMultiPhen, HCTATES, HCDCMANOVA, HCDCMultiPhen, and HCDCTATES. The corresponding Q–Q plots of type I error rates in varied approaches are shown in **Supplementary Figures S1–8**. Notably, for assessing powers, we do not only alter the values of β (meanwhile, the within-factor correlation $c^2 = 0.5$ and between-factor correlation $\rho c^2 = 0.1$) but also vary the values of within-factor correlation $c^2 = 0.3, 0.5, 0.7$, and 0.9 (meanwhile, the between-factor correlation $\rho c^2 = 0.1$).

Simulation Results

We establish varied nominal significance levels, distinct number of phenotypes, and different number of factors to assess the type I error rates of all the nine methods. In each simulation model, the p-values of all these evaluated methods are estimated by their asymptotic distributions. The type I error rates of MANOVA, MultiPhen, TATES, HCMANOVA, HCMultiPhen, HCTATES, HCDCMANOVA, HCDCMultiPhen, and HCDCTATES are evaluated by 10,000 replicated samples. For 10,000 replicated samples, we calculate that the 95% confidence intervals (CIs) for type I error rates in the nominal levels of 0.01 and 0.05 are about (0.008, 0.012) and (0.0457, 0.0543), respectively. The estimated type I error rates of all these tested methods are shown in **Table 1** ($M = 16$) and **Table 2** ($M = 32$). We observe that the majority of the type I error rates of HCDCMANOVA, HCDCMultiPhen, and HCDCTATES are within 95%CIs, which reflects the validity of the established HCDC applied to existing methods. Additionally, the type I error rates of MANOVA, MultiPhen, TATES, HCMANOVA, HCMultiPhen, and HCTATES are not obviously deviated from the nominal levels. For more information, please see the Q–Q plots in **Supplementary Figures S1–8**.

For power comparison for these nine methods, we alter distinct numbers of phenotypes and different models. The powers of all tests are estimated on the basis of 1,000 replications and 2,000 subjects at a significance level of 0.05. From the plots of power against genetic effect β (**Figure 1**), the following are observed and can be shown:

1. When the genetic variant has the same effect on all the phenotypes (Model 1), HCDCMANOVA, HCDCMultiPhen, and HCDCTATES are powerful than HCMANOVA, HCMultiPhen, and HCTATES, respectively. Meanwhile, HCMANOVA,

TABLE 1 | Evaluations of type I error rates of the nine methods in four simulation models.

Type I error rates								
Methods	**Model 1**		**Model 2**		**Model 3**		**Model 4**	
	α = 0.01	**α = 0.05**	**α = 0.01**	**α = 0.05**	**α = 0.01**	**α = 0.05**	**α = 0.01**	**α = 0.05**
HCDCMANOVA	0.0102	0.0523	0.0113	0.0522	0.0086	0.0532	0.0095	0.05
HCMANOVA	0.01	0.0517	0.0113	0.0524	0.0094	0.0478	0.0101	0.0509
MANOVA	0.0108	0.0505	0.0112	**0.0547**	0.0089	0.0514	0.0103	0.0519
HCDCMultiPhen	0.0101	0.0538	0.012	0.0527	0.0089	0.0532	0.0102	0.0483
HCMultiPhen	0.0091	0.0528	**0.0121**	0.0526	0.0102	0.0519	0.0101	0.0494
MultiPhen	0.0107	0.0523	0.0116	0.052	0.0094	0.0517	0.011	0.0537
HCDCTATES	0.0108	0.0502	0.0112	0.0511	0.0099	0.0466	0.0112	0.0506
HCTATES	**0.0122**	0.051	0.0114	0.0512	0.0109	0.0488	0.0103	0.05
TATES	0.0111	0.0473	0.0119	0.0512	0.0112	0.0514	0.0121	0.0535

Sample size N = *2,000, the number of phenotypes* M = *16,* c^2 = *0.5,* ρc^2 = *0.1, and minor allele frequency (MAF) = 0.3. The type* I *error rates of all nine methods are evaluated by 10,000 replicated samples at the significance of* α. *The values in bold indicate that the type* I *error rates are out of 95% CI of the nominal significance level.*

TABLE 2 | Evaluations of type I error rates of the nine methods in four simulation models.

Type I error rates								
Methods	**Model 1**		**Model 2**		**Model 3**		**Model 4**	
	α = 0.01	**α = 0.05**	**α = 0.01**	**α = 0.05**	**α = 0.01**	**α = 0.05**	**α = 0.01**	**α = 0.05**
HCDCMANOVA	0.01	0.0515	0.0118	0.0543	0.01	0.0498	0.0099	0.048
HCMANOVA	0.0111	0.0502	0.0118	**0.0544**	0.0111	0.0503	0.0102	0.0506
MANOVA	0.0101	0.051	0.0106	**0.0582**	0.0115	**0.0545**	0.0102	0.0515
HCDCMultiPhen	0.0099	0.0502	0.0117	**0.0545**	0.0098	0.05	0.0091	0.0497
HCMultiPhen	0.011	0.0516	0.0119	0.0543	0.0102	0.0503	0.0099	0.0512
MultiPhen	0.0102	0.0495	0.011	**0.0589**	0.0115	**0.0573**	0.0106	0.0511
HCDCTATES	0.0112	0.0514	0.0119	0.0539	0.0097	0.0483	0.0086	0.0463
HCTATES	0.0093	**0.045**	0.012	0.0538	0.0111	**0.0546**	0.0106	0.0516
TATES	**0.0078**	**0.041**	0.0105	0.0465	**0.0128**	0.0524	0.0101	0.0496

Sample size N = *2,000, the number of phenotypes* M = *32,* c^2 = *0.5,* ρc^2 = *0.1, and minor allele frequency (MAF) = 0.3. The type* I *error rates of all nine methods are evaluated by 10,000 replicated samples at the significance of* α. *The values in bold indicate that the type* I *error rates are out of 95% CI of the nominal significance level.*

HCMultiPhen, and HCTATES are powerful than MANOVA, MultiPhen, and TATES, respectively. In most replications, HCDC and HCM cluster various phenotypes into one or several categories to reduce the number of phenotypes to be analyzed for enhancing the power of test. Obviously, HCDC is slightly powerful than HCM in this scenario.

2. When the genetic effects on phenotypes reveal some groups and possess the same direction (Model 2), the power of HCDCMANOVA, HCDCMultiPhen, and HCDCTATES is equal to that of HCMANOVA, HCMultiPhen, and HCTATES, respectively. However, MANOVA, MultiPhen, and TATES with HCDC or HCM are much more powerful than MANOVA, MultiPhen, and TATES, respectively. These results indicate that clustering can definitely increase the power of test.

3. When the genetic effects on phenotypes appear in some groups and show different directions (Models 3 and 4), MANOVA, MultiPhen, and TATES are powerful than MANOVA, MultiPhen, and TATES with HCDC or HCM, respectively.

4. No matter altering of genetic effects β or changes in correlation coefficients between varied phenotypes, HCDCMANOVA and HCDCMultiPhen, HCMANOVA and HCMultiPhen, MANOVA and MutiPhen have similar performance in all four models, respectively.

5. When the genetic effects on phenotypes show obvious same direction within a group (Models 1 and 2), HCDCTATES, HCTATES, and TATES have better performance than other approaches.

From the within-factor correlation c^2 (**Supplementary Figures 9, 10**), we can observe the following:

6. When the genetic variant has the same effect on the phenotypes within a group, and there exists the same variance among phenotypes within this group, the powers of all estimated methods decrease as the within-factor c^2 increases (Models 1 and 2). However, our proposed MANOVA, MultiPhen, and TATES with using HCDC have obvious advantage over MANOVA, MultiPhen, and TATES without using HCDC, respectively.

7. When the genetic variant has the distinct effects on the phenotypes within a group, and there are different variances among phenotypes within this group (Models 3 and 4), MANOVA, MultiPhen, and TATES with using HCDC have more power than MANOVA, MultiPhen, and TATES without using HCDC as the within-factor c^2 is <0.5, but MANOVA and MultiPhen get more advantage as c^2 is >0.5, which reveal that

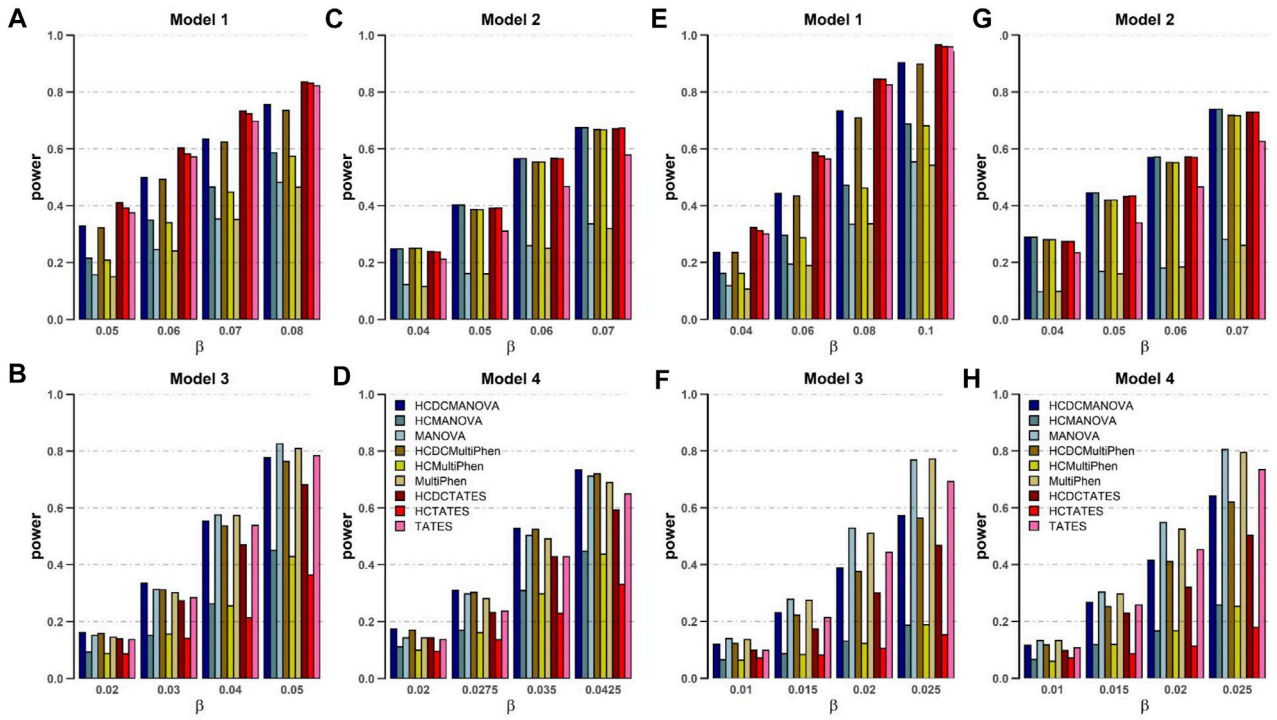

FIGURE 1 | Power comparisons of the nine methods as a function of β in the four models. Sample size $N = 2{,}000$, the number of phenotypes $M = 16$ **(A–D)** and M = 32 **(E–H)**, $c^2 = 0.5$, $\rho c^2 = 0.1$, and MAF = 0.3. The power of all the methods is evaluated by 1,000 replicated samples at a significance level of 0.05.

MANOVA and MultiPhen take heteroscedasticity between different phenotypes into account when calculating genetic associations.

In summary, the existing approaches employing HCDC has controlled type I error rates and have more advantage over or are comparable with those without employing HCDC. Therefore, we could draw that our established HCDC could give more power than HCM or original approaches without using HCDC, and in some scenarios, the advantage is more obvious. In other scenarios, the existing methods using HCDC is comparable with the most powerful test.

Real Data Analysis

We use our established approach, HCDC, together with other existing methods to the real data analysis in ARIC study (Author Anonymous, 1989). Briefly, ARIC is a prospective cohort study supported by the National Heart, Lung, and Blood Institute (NHLBI), aiming at assessing atherosclerosis risk in community. It keeps track of the altering of the occurrence of atherosclerosis-relevant diseases and cardiovascular risk factors in different regions, races, genders, and periods of time, in order to explore the natural process of atherosclerosis (Morrison et al., 2013). We acquire the clinical phenotypic and genotyped data of ARIC from dbGaP server of the United States National Center for Biotechnology Information (accession number: phs000090.v4.p1).

To evaluate the performance of HCDC together with other existing methods in analyzing real data, we evaluate the nine approaches to explore obesity-related phenotypes in ARIC. We choose nine continuous phenotypes concerning obesity comprising body weight, body mass index (BMI), mean skinfold thickness of the triceps brachii, average subscapular skinfold thickness, hip girth, waist, waist-to-hip ratio (WHR), calf girth, and wrist breadth and three covariates of age, gender, and race. The description of these variables is shown in **Table 3** in detail, and the correlation matrix of obesity-related phenotypes is displayed in **Supplementary Figure S11**. A total of 12,701 individuals across 272,027 SNPs are left to be analyzed subsequently after removing subjects with missing data under any of these 12 variables together with the genetic variants concerning missing rate more than 0.2 or HWE $<10^{-4}$. Each phenotype is adjusted for those three covariates by conducting the linear regression model.

According to the scaled phenotypes with respect to obesity, we use these nine methods to identify associated genetic variants. Due to multiple testing correction, we apply the genome-wide significance threshold of 5×10^{-8}. HCDC clusters nine phenotypes into two groups in this real data analysis, one only containing wrist breadth, while the other includes the rest. As comparisons, three groups are obtained after clustering by HCM, one only containing wrist breadth, and another encompasses WHR phenotype, while the other contains the remaining phenotypes. The dendrogram of clustering process for HCM and HCDC in ARIC data are presented in **Figure 2**. From these graphs, we can observe that there are significant differences between the HCM method and the HCDC method

TABLE 3 | The descriptions of involved obesity-related phenotypes and covariates in ARIC.

Index	All	Gender			Race		
		Male	Female	*p* Value	White	Black	*p* value
N	12771	5,704	7067	—	9,633	3,138	—
Male, %	44.66	—	—	—	47.02	37.44	$\mathbf{9.11 \times 10^{-21}}$
Age, years	54.09 ± 5.73	54.450 ± 5.75	53.76 ± 5.69	$\mathbf{6.76 \times 10^{-13}}$	54.34 ± 5.68	53.34 ± 5.80	$\mathbf{5.51 \times 10^{-17}}$
Weight, lb	173.13 ± 36.85	188.27 ± 31.46	160.92 ± 36.36	$\mathbf{<2.2 \times 10^{-16}}$	169.61 ± 35.69	183.99 ± 38.25	$\mathbf{1.90 \times 10^{-74}}$
Weight missing, %	0.149	0.158	0.142	0.995	0.083	0.351	**0.002**
BMI, kg/m^2	27.66 ± 5.30	27.54 ± 4.18	27.75 ± 6.05	**0.020**	27.01 ± 4.86	29.65 ± 6.05	$\mathbf{9.98 \times 10^{-104}}$
BMI missing, %	0.149	0.158	0.142	0.995	0.083	0.351	**0.002**
Triceps, mm	25.26 ± 10.02	19.34 ± 7.87	30.04 ± 8.97	$\mathbf{<2.2 \times 10^{-16}}$	24.54 ± 9.08	27.48 ± 12.23	$\mathbf{1.73 \times 10^{-34}}$
Triceps missing, %	0.157	0.175	0.142	0.798	0.093	0.351	**0.004**
Scapular, mm	24.48 ± 11.59	22.22 ± 9.19	26.31 ± 12.92	$\mathbf{1.13 \times 10^{-94}}$	21.85 ± 9.33	32.59 ± 13.89	$\mathbf{1.60 \times 10^{-299}}$
Scapular missing, %	0.446	0.561	0.354	0.107	0.353	0.733	**0.009**
WC, cm	96.94 ± 13.83	99.23 ± 10.93	95.09 ± 15.54	$\mathbf{1.25 \times 10^{-68}}$	96.19 ± 13.33	99.25 ± 15.02	$\mathbf{5.34 \times 10^{-24}}$
WC missing, %	0.141	0.123	0.156	0.798	0.104	0.255	0.092
HC, cm	104.55 ± 10.31	102.85 ± 8.09	105.93 ± 11.63	$\mathbf{2.81 \times 10^{-68}}$	103.50 ± 9.478	107.79 ± 11.98	$\mathbf{7.52 \times 10^{-72}}$
HC missing, %	0.141	0.140	0.142	0.999	0.104	0.255	0.092
WHtR	0.926 ± 0.078	0.963 ± 0.054	0.895 ± 0.081	$\mathbf{<2.2 \times 10^{-16}}$	0.928 ± 0.079	0.920 ± 0.076	$\mathbf{4.66 \times 10^{-8}}$
WHtR missing, %	0.149	0.140	0.156	0.999	0.114	0.255	0.131
Calf, cm	37.44 ± 3.67	38.06 ± 3.17	36.95 ± 3.95	$\mathbf{1.48 \times 10^{-68}}$	37.39 ± 3.58	37.60 ± 3.93	**0.006**
Calf missing, %	0.157	0.210	0.113	0.248	0.114	0.287	0.062
Wrist, mm	53.62 ± 5.18	57.78 ± 3.66	50.27 ± 3.53	$\mathbf{<2.2 \times 10^{-16}}$	53.59 ± 5.26	53.74 ± 4.91	0.137
Wrist missing, %	0.117	0.123	0.113	0.999	0.073	0.255	**0.022**

N *is the number of subjects; BMI, is body mass index; Triceps is average skinfold thickness of triceps brachii; Scapular is mean subscapular skinfold thickness; WC, is waist; HC, is hip girth; WHtR is waist-to-hip ratio; Calf is calf girth; and Wrist is wrist breadth. The distributions of normal index are described by mean ± standard deviation; the distributions of non-normal indicators are described by means (25% quantile, 75% quantile). For normal distribution indicators, the differences between groups are estimated using the* t*-test (the variances of two groups are homogeneous) or the approximate* t*-test (the variances of two groups are heterogeneous). For non-normally indicators, Wilcoxon signed-rank test is used to test the differences between indicators to get the* p*-values of differences. For discrete indicators, the chi-square test is used for hypothesis testing and then deriving* p*-values. Bold number indicates* $p < 0.05$. *ARIC, atherosclerosis risk in communities.*

we proposed in the clustering process. Specifically, when the correlation coefficients between different clusters are calculated, the correlation coefficients increase with the increase in clustering times in HCM (*h* is gradually increasing), while in HCDC, the correlation coefficients may increase, or they may decrease compared to the last clustering result. These differences can be explained by the distinct ways to calculate the correlation coefficients. HCM uses a uniform formula to evaluate the similarity between pairs of clusters. However, pairs of clusters generally include varied situations, comprising single phenotype versus single phenotype, single phenotype versus multiple phenotypes, or multiple phenotypes versus multiple phenotypes. Nevertheless, HCDC takes those into account fully to deal with complex and changeable situations; as a result, such clustering result may be more convincing for most of circumstances.

A total of eight SNPs are identified as significant signals for at least one method (**Table 4**). Previously, a large amount of studies (Frayling et al., 2007; Heard-Costa et al., 2009; Lindgren et al., 2009; Meyre et al., 2009; Thorleifsson et al., 2009; Willer et al., 2009; Heid et al., 2010; Speliotes et al., 2010; Bradfield et al., 2012; Wen et al., 2012; Berndt et al., 2013; Monda et al., 2013; Locke et al., 2015; Shungin et al., 2015) have covered that *FTO* contributes to the risk of obesity due to the population-based studies and the relevant experiments elaborating specific mechanisms. Among the eight associated SNPs, rs9939609 and rs8050136 are located in *FTO* gene. In addition, *UQCC* region is covered to be associated with height (Sanna et al., 2008). Few other SNPs have been explored to assess the association with obesity or obesity-related phenotypes. From **Table 4**, we can observe that HCDCMANOVA identified three SNPs; HCMANOVA identified two SNPs; MANOVA identified four SNPs; HCDCMultiPhen identified three SNPs, more than the number of SNPs identified by HCMultiPhen (twoSNPs) and MultiPhen (one SNP); HCDCTATES identified three SNPs; TATES identified four SNPs; while no SNP was identified by HCTATES. Overall, the results in real data analysis are highly consistent with the simulation performance. The number of SNPs identified by existing methods with HCDC is comparable with the largest number of SNPs identified by existing methods without HCDC. In order to make the overall performance clearer in real data results, we draw Q–Q plots and Manhattan plots after the application of these nine different methods in ARIC data (**Supplementary Figures S12–18**). From these plots, we can intuitively observe the SNPs identified by distinct methods, and their *p*-values in the same plot to compare their sizes.

Characteristics of the Significant Variants

We searched the annotations of the associated SNPs on the basis of the Ensemble website (https://asia.ensembl.org) and SCAN website (http://scandb.org), which are shown in **Table 5**. From **Table 5**, it can be observed that these significant SNPs are located in intergenic or intron region, and some of them have been covered to be associated with BMI, type 2 diabetes, or height. In general, the first or large-scale GWASs have reported some of

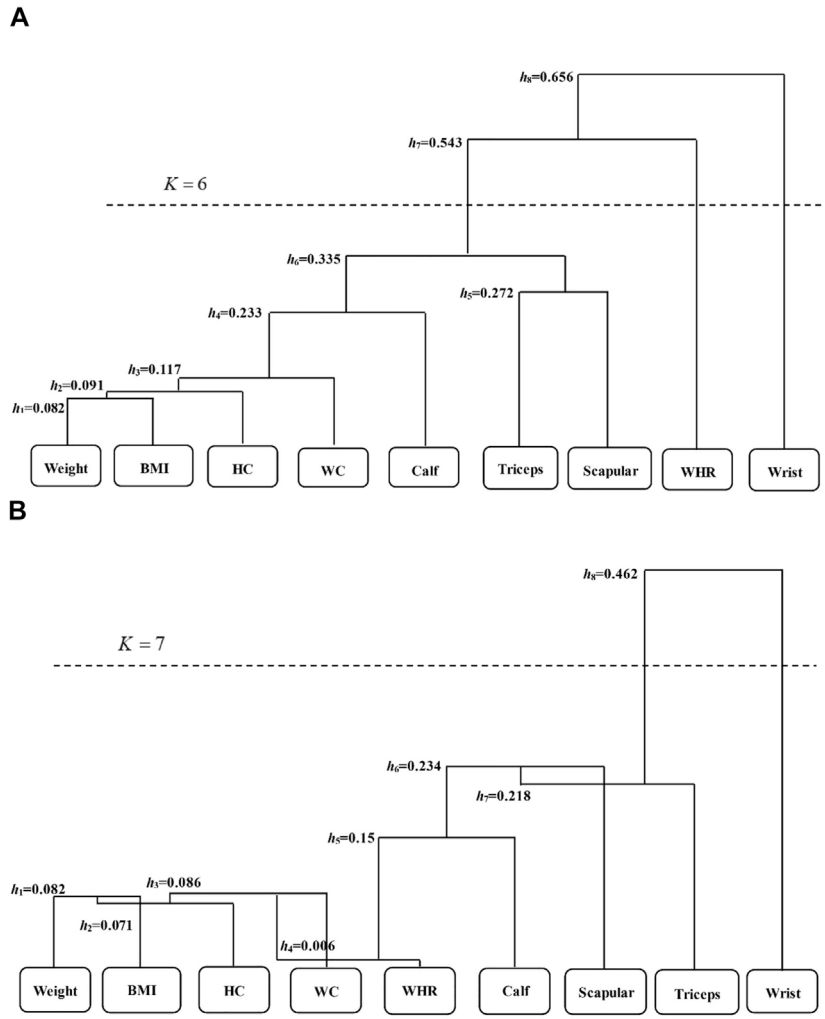

FIGURE 2 | The dendrogram of the nine phenotypes in the ARIC study *via* HCM **(A)** and HCDC **(B)**. h represents the maximum value of correlation coefficient in each clustering process, which is taken as the "branch length" of the clustering tree. K reveals the final clustering times according to the stopping criteria. BMI is body mass index; Triceps is average skinfold thickness of triceps brachii; Scapular is mean subscapular skinfold thickness; WC is waist; HC is hip girth; WHR is waist-to-hip ratio; Calf is calf girth; and Wrist is wrist breadth.

TABLE 4 | Display of significant SNPs and the corresponding *p*-values in the analysis of ARIC.

Chr	SNP	HCDCMANOVA	HCMANOVA	MANOVA	HCDCMultiphe	HCMultiphen	Multiphen	HCDCTATES	HCTATES	TATES
3	rs17017947	0.873	0.184	**1.02×10^{-11}**	NA	NA	NA	0.803	0.690	0.314
10	rs41470552	0.102	0.004	**6.25×10^{-9}**	NA	NA	NA	0.285	0.748	0.0358
11	rs7927943	1.72×10^{-7}	1.88×10^{-7}	5.57×10^{-6}	1.88×10^{-7}	1.21×10^{-7}	3.33×10^{-6}	9.18×10^{-8}	0.513	**1.16×10^{-8}**
11	rs1945647	5.83×10^{-7}	2.49×10^{-7}	1.19×10^{-5}	4.26×10^{-7}	1.12×10^{-7}	6.27×10^{-6}	2.31×10^{-7}	0.554	**1.77×10^{-8}**
16	rs9939609	**1.67×10^{-11}**	**9.53×10^{-11}**	**1.85×10^{-8}**	**2.98×10^{-11}**	**1.67×10^{-10}**	**3.39×10^{-8}**	**1.68×10^{-11}**	0.331	**2.97×10^{-10}**
16	rs8050136	**3.83×10^{-11}**	**2.10×10^{-10}**	**4.29×10^{-8}**	**8.07×10^{-11}**	**4.33×10^{-10}**	8.66×10^{-8}	**1.11×10^{-10}**	0.277	**2.86×10^{-9}**
20	rs201561	**1.06×10^{-8}**	5.18×10^{-8}	2.48×10^{-6}	**1.11×10^{-8}**	5.45×10^{-8}	2.91×10^{-6}	2.57×10^{-7}	0.861	7.99×10^{-7}
20	rs1570004	1.07×10^{-7}	4.86×10^{-7}	5.28×10^{-5}	1.54×10^{-7}	7.06×10^{-7}	7.77×10^{-5}	**1.97×10^{-8}**	0.864	6.12×10^{-8}

The p-values of nine methods are calculated based on asymptotic distribution. p-Value <5×10^{-8} are in bold. "NA" reveals MultiPhen is not available because the genotype at the specified SNP does not take all three values of 0, 1, and 2 in these data. SNP, single-nucleotide polymorphism; ARIC, atherosclerosis risk in communities.

TABLE 5 | The annotations of the significant identified SNPs.

SNPs	Chr	Position (GRCh38)	Alleles (alt/Ref)	Gene (nearest)	Feature	Expression genes	Reported (yes/No)	Reported phenotypes	GWAS references
rs17017947	3	276171	A/C	*CHL1*	Intron	—	No	—	—
rs41470552	10	102222133	T/G	*PITX3*	Intergenic	—	No	—	—
rs1945647	11	81602715	C/T	*MTND6P25*	Intergenic	*GNAI2,STK40,LIMK1,LIG4, HLTF,ZNF511,CBLL1,NUDT17, POLR3C,DAGLB,KDELR2,NUP93, PRCC,C16orf80,RAB33B,LRP8*	No	—	—
rs7927943	11	81637194	C/T	*MTND6P25*	Intergenic	*WSCD2,GNAI2,ZFHX3,NUP93, FAM60A,LIMK1,MAP4, FLJ31958,LIG4,HLTF*	No	—	—
rs8050136	16	53782363	C/A	*FTO*	Intron	*HES7,LATS2*	Yes	BMI, T2D, Adiposity	PMID: 18372903 PMID: 31217584 PMID: 19079260
rs9939609	16	53786615	T/A	*FTO*	Intron	*CR1,CR1L,ZNRF1,ANKRD50, LATS2,TSPYL4,HES7*	Yes	BMI, T2D	PMID: 17434869 PMID: 31217584 PMID: 17554300
rs1570004	20	35370450	A/T	*UQCC*	Intron	—	Yes	Height	PMID: 18193045
rs201561	20	22018575	G/C	*RPL41P1*	Intergenic	*P2RX3,EHD4*	Yes	Balding type 1	PMID: 30595370

Annotations are from Ensemble website (https://asia.ensembl.org) and SCAN website (http://scandb.org); intron denotes the SNP is located between exons; intergenic denotes the SNP is located between genes. Expression genes denote annotations added after analysis of transcriptional levels of eQTL in cell lines from HapMap CEU and YRI samples using Affymetrix human exon 1.0 ST array; GWAS references indicate the identifications of PubMed. SNP, single-nucleotide polymorphism; GWAS, genome-wide association study; eQTL, expression quantitative trait locus.

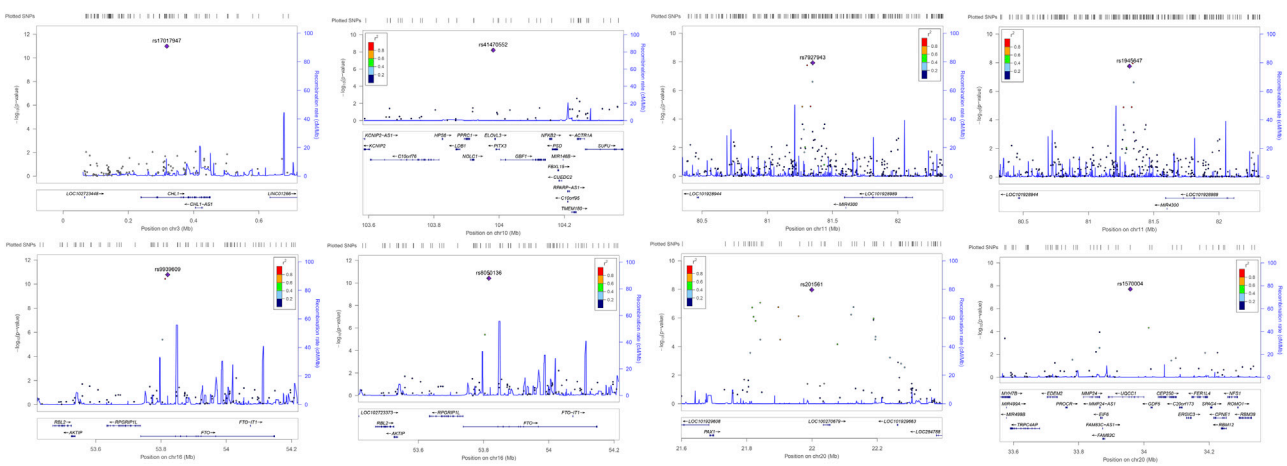

FIGURE 3 | The regional association plots of the significant SNPs identified in ARIC. The *p*-values of rs17017947 and rs41470552 are evaluated by MANOVA method. The *p*-values of 7927943 and rs1945647 are estimated by TATES method. The *p*-values of rs9939609, rs8050136, and rs201561 are assessed by HCDCMANOVA method. The *p*-values of rs1570004 is evaluated by HCDCTATES. LD is constructed using the hg19 version of the 1000 Genome (American). The plots where rs7927943 and rs1945647 are located show the 1,000-kb range around these most significant SNPs. The plots where the rest SNPs (rs7927943, rs1945647, rs9939609, rs8050136, rs201561, and rs1570004) are located present the 400-kb range around these identified significantly SNPs. SNP, single-nucleotide polymorphism.

these associated signals. The ID of PubMed could be inquired to retrieve the relevant progress of these SNPs. Additionally, there is no influence for us to explore the expressions of the genes that the significant SNPs are associated with, although most of them are located in the intron or intergenic region. Moreover, most of these significant SNPs reveal that their possible effects on the expressions of corresponding genes based on the cell lines of HapMap CEU and YRI (**Table 5**).

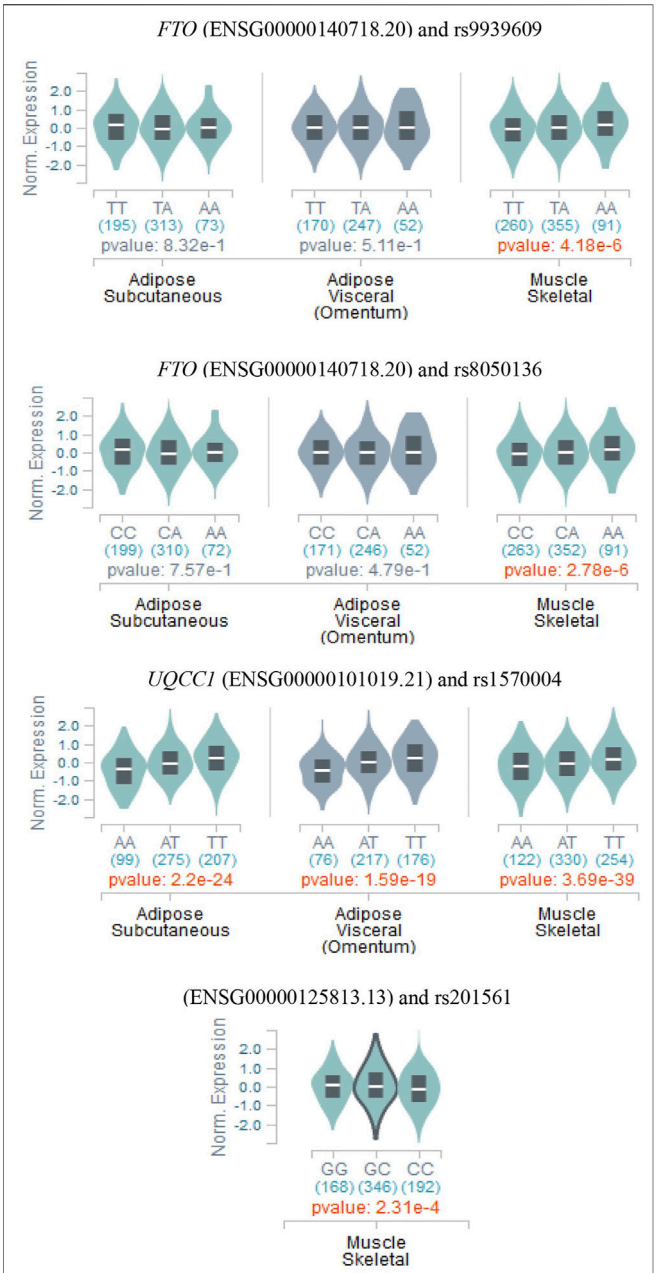

FIGURE 4 | The relationship between the genotypes of the significant SNPs discovered by HCDC method and eQTL in subcutaneous adipose tissue, visceral adipose tissue, and muscle tissue (data are from GTEx website).

For more extensive investigation of whether the significant SNPs identified in ARIC have LD with the other nearby loci, that is, to detect the correlations between these eight associated significantly SNPs in this study with the undetectable surrounding loci, we produced regional plots presented in **Figure 3**. From **Figure 3**, it is clear that rs7927943 is physically close to rs1945647, and their LD is quite robust, which reflects that their r^2 is >0.8. What is more, both of them are located near the *LOC101928989* gene, regulating the expressions of certain genes (*LIMK1*, *GNAI2*, etc.). Since both rs7927943 and rs1945647 manipulate corresponding expressions of genes, subsequently, the relationship between these SNPs and obesity can be studied from the perspective of gene expression. Notice that both SNPs rs9939609 and rs8050136 are located in *FTO* gene attaching to chromosome 16, and their physical regions are close to each other with a high correlation r^2 >0.8 (**Figure 3**). It is well known that rs9939609 acts as an obese variant (Frayling et al., 2007). Because of the strong LD between rs9939609 and rs8050136, it is reasonable to speculate that rs8050136 is also associated with obesity-related phenotypes. Three SNPs, namely, rs17017947, rs1570004, and rs41470552, are located in the intron region of genes *CHL1*, *UQCC*, and *NOLC1*, respectively. None of them possesses relatively strong LD with the surrounding loci, so these SNPs probably have an effect on corresponding phenotypic characteristics independently. The rs201561 around *LOC100270679* has a profound LD with the surrounding loci (**Figure 3**), combined with the fact that the association result of *p*-value for rs201561 is the smallest among all the nearby variants, revealing that the surrounding loci have an impact on the phenotypes because of the high LD with rs201561.

With the purpose of exploring the SNPs associated with obesity-related phenotypes, and the expressions of those identified by all the methods employed in this study in different adipose tissues, we retrieved the relevant content of GTEx website (https://www.gtexportal.org/home/). Consequently, the significant SNPs (rs17017947, rs41470552, rs7927943, and rs1945647) not identified by existing methods with HCDC have not been detected to be expressed in relevant tissues *via* GTEx query, while these distinct genotypes of significant SNPs (rs9939609, rs8050136, rs201561, and rs1570004) identified by existing methods with HCDC present differential expressions in adipose tissue or muscle tissue (**Figure 4**). In other words, the proposed HCDC has certain significance for biological research from the perspective of gene expression. Furthermore, it is noteworthy that the different genotypes of *UQCC1*-rs1570004 are differentially expressed in subcutaneous adipose, visceral adipose, or muscle tissue ($p < 1.59 \times 10^{-19}$). Moreover, the phenotypes adopted in real data analysis denote various measurement phenotypes about obesity, so the differentially expressed tissues are highly consistent with the phenotypes adopted in this study. Thus, *UQCC1*-rs1570004, as a SNP that has not been reported to be associated with obesity-related phenotypes in other studies so far, is worthy of further functional experimental studies in the future to confirm its impressive value.

DISCUSSION

In this article, HCDC is proposed to jointly analyze multiple phenotypes in association analyses. The established approach employs the similarity measure to cluster multiple phenotypes. Using HCDC, we apply the existing methods to detect the genetic associations with the combined phenotypes rather than the individual phenotypes. HCDC owns several obvious advantages

compared to other dimension reduction approaches. First, a dendrogram involved in the multiple phenotypes can be produced by HCDC (see **Figure 2**), which could supply more information about the structure of phenotypes. Second, not limited to the correlation coefficients, any proper measurements of distance can be used for the hierarchical clustering procedure, although the specific effects are worth further consideration. Third, HCDC is computationally fast, so it is easy to implement. Notably, HCDC does not need to acquire the individual phenotypes, and on the contrary, it only acquires the similarity matrix of phenotypes. This similarity of matrix can be evaluated from the test statistics of summary data employing the independent SNPs in a GWAS (Zhu et al., 2015). This is a major advantage of HCDC clustering using correlation coefficients between phenotypes.

We performed extensive simulations together with the real data analysis to assess the performance of MANOVA, MultiPhen, and TATES combined with applying HCDC and compared these with their original versions. The simulation results reveal that these three methods applying HCDC not only possess correct type I error rates but also own more advantage over these without applying HCDC under a series of simulation scenarios. For more realistic simulation settings, GCTA software is the first choice. Thus, further tests should be evaluated in the future (Yang et al., 2011). More importantly, the real data analysis results elucidate that HCDC shows great potential in multiple phenotypes analysis of ARIC *via* GWAS about obesity, and the bioinformatics analysis for these results also supports them. In addition, we also use another clustering method, HCM, as a major competitor to compare its performance with that of HCDC. We suggest that the most important thing for HCM to be improved is that when calculating the correlation coefficient between two clusters, it should take the imbalanced numbers of phenotypes in two clusters into account, and it may not be appropriate to use a unified calculation formula of correlation coefficient. In real data analysis, the fact that the performance of HCDC is better than HCM confirms our point of view. Presently, HCDC is more suitable for continuous phenotypes. After the transformation of phenotypes, it can also be applied to dichotomous or mixed traits. However, its performance in dichotomous or mixed traits situation still needs to be further investigated.

Then, we use HCDC to analyze ARIC data and discovered that *UQCC1*-rs1570004 has a significant correlation with multiple phenotypes about obesity traits. Bioinformatics exploration shows that varied genotypes of *UQCC1*-rs1570004 are differentially expressed in subcutaneous fat, visceral fat, and muscle tissue ($p < 1.59 \times 10^{-19}$). The differentially expressed tissues are consistent with the phenotypes studied in this work. Therefore, *UQCC1*-rs1570004, as an SNP that has not been reported to be associated to obesity-related phenotypes in the literature, is worthy of further functional experiments in the future to confirm its potential value. From the perspective of application in real data, HCDC owns certain value and significance for further association studies.

In summary, HCDC is an effective approach for the association study between multiple phenotypes and genetic variants in varied research fields. In medical research, many research disciplines have strong intersection. Generally, different disciplines carry out the association study between phenotypes and genetic variants separately. Interdisciplinary research on multiple phenotypes, such as phenotypes across multiple tissues, including various indicators with behavior, morphology, and physiology, will be likely extended to phenome research (Houle et al., 2010), which would be very meaningful. Because there is no assumption for HCDC in the aspect of genetic effect model, clustering multiple phenotypes into different categories according to similarity measure between phenotypes in HCDC is very useful for phenome research. Moreover, in a large number of phenotypes, HCDC does not need to understand the specific model for generating data, while only understanding the correlation matrix between phenotypes is undoubtedly another decent feature. In reality, it is common that the genetic structure among different phenotypes is complex and usually unknown. HCDC provides an effective and novel research strategy for exploring high-dimensional phenotypic data in the coming era of phenome as shown in simulations.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: The datasets ARIC for this study can be found in the dbGaP https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study id = phs000090.v4.p1.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The ARIC Investigators. The datasets ARIC for this study can be found in the dbGaP https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?studyid=phs000090.v4.p1. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

Study concept and design: LF and Y-QH; acquisition of data: LF, Y-QH, and YW; methodology and interpretation of data: LF; drafting of the manuscript: LF and Y-QH; critical revision of the manuscript for important intellectual content: LF, YW, TL, SY, and Y-QH; all authors have read and approved the final version of manuscript.

FUNDING

This study was supported by grants to Y-QH from the National Natural Science Foundation of China (grant nos. 11971117 and 11571082).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.791920/full#supplementary-material

REFERENCES

Aschard, H., Vilhjálmsson, B. J., Greliche, N., Morange, P. E., Trégouët, D. A., and Kraft, P. (2014). Maximizing the Power of Principal-Component Analysis of Correlated Phenotypes in Genome-wide Association Studies. *Am. J. Hum. Genet.* 94, 662–676. doi:10.1016/j.ajhg.2014.03.016

Author Anonymous (1989). The Atherosclerosis Risk in Communities (ARIC) Study: Design and Objectives. The ARIC Investigators. *Am. J. Epidemiol.* 129, 687–702.

Bates, D. M., and DebRoy, S. (2004). Linear Mixed Models and Penalized Least Squares. *J. Multivariate Anal.* 91, 1–17. doi:10.1016/j.jmva.2004.04.013

Berndt, S. I., Gustafsson, S., Mägi, R., Ganna, A., Wheeler, E., Feitosa, M. F., et al. (2013). Genome-wide Meta-Analysis Identifies 11 New Loci for Anthropometric Traits and Provides Insights into Genetic Architecture. *Nat. Genet.* 45, 501–512. doi:10.1038/ng.2606

Bien, J., and Wegkamp, M. (2013). Discussion of "Correlated Variables in Regression: Clustering and Sparse Estimation". *J. Stat. Plan. Infer* 143, 1859–1862. doi:10.1016/j.jspi.2013.05.020

Bradfield, J. P., Taal, H. R., Timpson, N. J., Scherag, A., Lecoeur, C., Warrington, N. M., et al. (2012). A Genome-wide Association Meta-Analysis Identifies New Childhood Obesity Loci. *Nat. Genet.* 44, 526–531. doi:10.1038/ng.2247

Bühlmann, P., Rütimann, P., van de Geer, S., and Zhang, C-H. (2013). Correlated Variables in Regression: Clustering and Sparse Estimation. *J. Stat. Plan. Infer* 143, 1835–1858. doi:10.1016/j.jspi.2013.05.019

Cohen, J., and Cohen, P. (1983). *Applied Multiple Regression/correlation Analysis for the Behavioral Science*. 2nd edition. Hillsdale (NJ): Erlbaum.

Cole, David, A., and MaxwellScott, E. (1994). How the Power of MANOVA Can Both Increase and Decrease as a Function of the Inter Correlations Among the Dependent Variables. *Psychol. Bull.* 115, 465–474. doi:10.1037/0033-2909.115.3.465

Ferreira, M. A., and Purcell, S. M. (2009). A Multivariate Test of Association. *Bioinformatics* 25, 132–133. doi:10.1093/bioinformatics/btn563

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., et al. (2007). A Common Variant in the FTO Gene Is Associated with Body Mass index and Predisposes to Childhood and Adult Obesity. *Science* 316, 889–894. doi:10.1126/science.1141634

Fu, L., Wang, Y., Li, T., and Hu, Y. Q. (2021). A Novel Approach Integrating Hierarchical Clustering and Weighted Combination for Association Study of Multiple Phenotypes and a Genetic Variant. *Front. Genet.* 12, 654804. doi:10.3389/fgene.2021.654804

Heard-Costa, N. L., Zillikens, M. C., Monda, K. L., Johansson, A., Harris, T. B., Fu, M., et al. (2009). NRXN3 Is a Novel Locus for Waist Circumference: a Genome-wide Association Study from the CHARGE Consortium. *Plos Genet.* 5, e1000539. doi:10.1371/journal.pgen.1000539

Heid, I. M., Jackson, A. U., Randall, J. C., Winkler, T. W., Qi, L., Steinthorsdottir, V., et al. (2010). Meta-analysis Identifies 13 New Loci Associated with Waist-Hip Ratio and Reveals Sexual Dimorphism in the Genetic Basis of Fat Distribution. *Nat. Genet.* 42, 949–960. doi:10.1038/ng.685

Houle, D., Govindaraju, D. R., and Omholt, S. (2010). Phenomics: the Next challenge. *Nat. Rev. Genet.* 11, 855–866. doi:10.1038/nrg2897

Jin, L., and Lin, Y. (2019). Discrimination of Several Correlation Coefficients and Their Implementation in R Software. *Stat. Inf. Forum* 34, 3–11. (in Chinese). doi:10.3969/j.issn.1007-3116.2019.04.001

Klei, L., Luca, D., Devlin, B., and Roeder, K. (2008). Pleiotropy and Principal Components of Heritability Combine to Increase Power for Association Analysis. *Genet. Epidemiol.* 32, 9–19. doi:10.1002/gepi.20257

Kwak, I. Y., and Pan, W. (2016). Adaptive Gene- and Pathway-Trait Association Testing with GWAS Summary Statistics. *Bioinformatics* 32, 1178–1184. doi:10.1093/bioinformatics/btv719

Liang, X., Sha, Q., Rho, Y., and Zhang, S. (2018). A Hierarchical Clustering Method for Dimension Reduction in Joint Analysis of Multiple Phenotypes. *Genet. Epidemiol.* 42, 344–353. doi:10.1002/gepi.22124

Liang, X., Wang, Z., Sha, Q., and Zhang, S. (2016). An Adaptive Fisher's Combination Method for Joint Analysis of Multiple Phenotypes in Association Studies. *Sci. Rep.* 6, 34323. doi:10.1038/srep34323

Lindgren, C. M., Heid, I. M., Randall, J. C., Lamina, C., Steinthorsdottir, V., Qi, L., et al. (2009). Genome-wide Association Scan Meta-Analysis Identifies Three Loci Influencing Adiposity and Fat Distribution. *Plos Genet.* 5, e1000508. doi:10.1371/journal.pgen.1000508

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., et al. (2015). Genetic Studies of Body Mass index Yield New Insights for Obesity Biology. *Nature* 518, 197–206. doi:10.1038/nature14177

Lutz, S. M., Fingerlin, T. E., Hokanson, J. E., and Lange, C. (2017). A General Approach to Testing for Pleiotropy with Rare and Common Variants. *Genet. Epidemiol.* 41, 163–170. doi:10.1002/gepi.22011

Meyre, D., Delplanque, J., Chèvre, J. C., Lecoeur, C., Lobbens, S., Gallina, S., et al. (2009). Genome-wide Association Study for Early-Onset and Morbid Adult Obesity Identifies Three New Risk Loci in European Populations. *Nat. Genet.* 41, 157–159. doi:10.1038/ng.301

Monda, K. L., Chen, G. K., Taylor, K. C., Palmer, C., Edwards, T. L., Lange, L. A., et al. (2013). A Meta-Analysis Identifies New Loci Associated with Body Mass index in Individuals of African Ancestry. *Nat. Genet.* 45, 690–696. doi:10.1038/ng.2608

Morrison, A. C., Voorman, A., Johnson, A. D., Liu, X., Yu, J., Li, A., et al. (2013). Whole-genome Sequence-Based Analysis of High-Density Lipoprotein Cholesterol. *Nat. Genet.* 45, 899–901. doi:10.1038/ng.2671

O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C., Elliott, P., Jarvelin, M. R., et al. (2012). MultiPhen: Joint Model of Multiple Phenotypes Can Increase Discovery in GWAS. *PLoS One* 7, e34861. doi:10.1371/journal.pone.0034861

Ott, J., and Rabinowitz, D. (1999). A Principal-Components Approach Based on Heritability for Combining Phenotype Information. *Hum. Hered.* 49, 106–111. doi:10.1159/000022854

Sanna, S., Jackson, A. U., Nagaraja, R., Willer, C. J., Chen, W. M., Bonnycastle, L. L., et al. (2008). Common Variants in the GDF5-UQCC Region Are Associated with Variation in Human Height. *Nat. Genet.* 40, 198–203. doi:10.1038/ng.74

Sattar, N., McConnachie, A., Shaper, A. G., Blauw, G. J., Buckley, B. M., de Craen, A. J., et al. (2008). Can Metabolic Syndrome Usefully Predict Cardiovascular Disease and Diabetes? Outcome Data from Two Prospective Studies. *Lancet* 371, 1927–1935. doi:10.1016/s0140-6736(08)60602-9

Schaid, D. J., Tong, X., Larrabee, B., Kennedy, R. B., Poland, G. A., and Sinnwell, J. P. (2016). Statistical Methods for Testing Genetic Pleiotropy. *Genetics* 204, 483–497. doi:10.1534/genetics.116.189308

Shungin, D., Winkler, T. W., Croteau-Chonka, D. C., Ferreira, T., Locke, A. E., Mägi, R., et al. (2015). New Genetic Loci Link Adipose and Insulin Biology to Body Fat Distribution. *Nature* 518, 187–196. doi:10.1038/nature14132

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., et al. (2010). Association Analyses of 249,796 Individuals Reveal 18 New Loci Associated with Body Mass index. *Nat. Genet.* 42, 937–948. doi:10.1038/ng.686

Tang, C. S., and Ferreira, M. A. (2012). A Gene-Based Test of Association Using Canonical Correlation Analysis. *Bioinformatics* 28, 845–850. doi:10.1093/bioinformatics/bts051

Therneau, T. M., Grambsch, P. M., and Pankratz, V. S. (2003). Penalized Survival Models and Frailty. *J. Comput. Graph Sta* 12, 156–175. doi:10.1198/1061860031365

Thorleifsson, G., Walters, G. B., Gudbjartsson, D. F., Steinthorsdottir, V., Sulem, P., Helgadottir, A., et al. (2009). Genome-wide Association Yields New Sequence Variants at Seven Loci that Associate with Measures of Obesity. *Nat. Genet.* 41, 18–24. doi:10.1038/ng.274

van der Sluis, S., Posthuma, D., and Dolan, C. V. (2013). TATES: Efficient Multivariate Genotype-Phenotype Analysis for Genome-wide Association Studies. *Plos Genet.* 9, e1003235. doi:10.1371/journal.pgen.1003235

Wang, Z., Sha, Q., and Zhang, S. (2016). Joint Analysis of Multiple Traits Using "Optimal" Maximum Heritability Test. *PLoS One* 11, e0150975. doi:10.1371/journal.pone.0150975

Wen, W., Cho, Y. S., Zheng, W., Dorajoo, R., Kato, N., Qi, L., et al. (2012). Meta-analysis Identifies Common Variants Associated with Body Mass index in East Asians. *Nat. Genet.* 44, 307–311. doi:10.1038/ng.1087

Willer, C. J., Speliotes, E. K., Loos, R. J., Li, S., Lindgren, C. M., Heid, I. M., et al. (2009). Six New Loci Associated with Body Mass index Highlight a Neuronal Influence on Body Weight Regulation. *Nat. Genet.* 41, 25–34. doi:10.1038/ng.287

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* 88, 76–82. doi:10.1016/j.ajhg.2010.11.011

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 Bioinformatics* 17, 19. doi:10.1186/s12859-015-0868-6

Yang, Q., and Wang, Y. (2012). Methods for Analyzing Multivariate Phenotypes in Genetic Association Studies. *J. Probab. Stat.* 2012, 652569. doi:10.1155/2012/652569

Yang, Q., Wu, H., Guo, C. Y., and Fox, C. S. (2010). Analyze Multivariate Phenotypes in Genetic Association Studies by Combining Univariate Association Tests. *Genet. Epidemiol.* 34, 444–454. doi:10.1002/gepi.20497

Zeger, S. L., and Liang, K. Y. (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics* 42, 121–130. doi:10.2307/2531248

Zhu, X., Feng, T., Tayo, B. O., Liang, J., Young, J. H., Franceschini, N., et al. (2015). Meta-analysis of Correlated Traits via Summary Statistics from GWASs with an Application in Hypertension. *Am. J. Hum. Genet.* 96, 21–36. doi:10.1016/j.ajhg.2014.11.011

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.