Published in final edited form as:

Ann Hum Genet. 2018 November ; 82(6): 389–395. doi:10.1111/ahg.12260.

Joint Analysis of Multiple Phenotypes in Association Studies Using Allele-based Clustering Approach for Non-Normal Distributions

Xiaoyu Liang¹, Qiuying Sha¹, and Shuanglin Zhang^{1,*}

¹Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan

Summary

In the study of complex diseases, several correlated phenotypes are usually measured. There is also increasing evidence showing that testing the association between a single-nucleotide polymorphism (SNP) and multiple-dependent phenotypes jointly is often more powerful than analyzing only one phenotype at a time. Therefore, developing statistical methods to test for genetic association with multiple phenotypes has become increasingly important. In this paper, we develop an Allele-based Clustering Approach (ACA) for the joint analysis of multiple non-normal phenotypes in association studies. In ACA, we consider the alleles at a SNP of interest as a dependent variable with two classes, and the correlated phenotypes as predictors to predict the alleles at the SNP of interest. We perform extensive simulation studies to evaluate the performance of ACA and compare the power of ACA with the powers of Adaptive Fisher's Combination test (AFC), Trait-based Association Test that uses Extended Simes procedure (TATES), Fisher's Combination test (FC), the standard MANOVA, and the joint model of Multiple Phenotypes (MultiPhen). Our simulation studies show that the proposed method has correct type I error rates and is much more powerful than other methods for some non-normal distributions.

Keywords

nonparametric method; clustering approach; non-normal distribution; multiple phenotypes

Introduction

In the study of a complex disease, data on multiple phenotypes are often collected to have a better understanding of the disease. For example, cardiovascular disease (CVD) is characterized by high levels of both low-density serum lipoprotein levels (LDL) and systolic blood pressure (SBP) (Newman III et al., 1986; Majumdar et al., 2015); Chronic obstructive pulmonary disease (COPD) is one of the most common lung diseases characterized by

*Corresponding author: SHUANGLIN ZHANG, Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan 49931, USA. shuzhang@mtu.edu.

Author contributions

S. Z. and Q. S. designed research, X. L. and S. Z. performed statistical analysis, and X. L., Q. S., and S. Z. wrote the manuscript.

Competing financial interests

The authors declare no conflict of interest.

reduced expiratory airflow, symptoms of cough, sputum production, and dyspnea (Casaburi et al., 2002); Thromboembolic disease is characterized by the intermediate correlated phenotypes such as Factor VII, VIII, IX, XI, XII, and von Willebrand factor (Germain et al., 2011; Ray et al. 2016; Souto et al. 2000).

Statistical methods for detecting association using multivariate phenotypes have been developed. These methods can be roughly divided into two groups: univariate methods and multivariate methods. Univariate methods are based on univariate analyses and combine the results of univariate analyses. Multivariate methods are based on models that jointly model multiple phenotypes. Univariate methods are popular in practice because it is much easier to construct a univariate association test statistic than a multivariate association test statistic. Recently, several univariate methods have appeared to explore the genetic association with multiple phenotypes by considering the correlation structure among phenotypes (Kwak & Pan, 2016; Liang et al., 2016; van der Sluis et al., 2013; Yang et al., 2016). Kwak and Pan (2016) proposed an adaptive gene-based test and a pathway-based test for association analysis of multiple phenotypes with summary statistics from genome-wide association studies (GWAS). The Trait-based Association Test that uses Extended Simes procedure (TATES) combines p-values obtained in standard univariate GWAS while correcting for the correlation between p-values (van der Sluis et al., 2013). Adaptive Fisher's Combination test (AFC) combines p-values obtained in standard univariate GWAS by using the optimal number of p-values which is determined by the data (Liang et al., 2016).

Comparing univariate methods with multivariate methods, univariate methods may lose power since univariate methods ignore the extra information by analyzing multiple phenotypes in one unified analysis. Recently, several multivariate methods have been developed (Marchini et al., 2007; Galesloot et al., 2014; Zhou and Stephens, 2014; Korte et al., 2012; Casale et al., 2015; Zhang et al., 2014; Yan et al., 2013; O'Reilly et al., 2012; Tang and Ferreira, 2012; Aschard et al., 2014; Lange et al., 2004; Klei et al., 2008; Zhou et al., 2015). However, most of the multivariate methods are based on the assumption of normality for phenotypes. In this article, we develop a novel multivariate method, named an Allele-based Clustering Approach (ACA), for the joint analysis of multiple non-normal phenotypes. ACA is nonparametric in that it does not assume any particular genetic models. Using extensive simulation studies, we evaluate the performance of the proposed method and compare the power of the proposed method with the powers of AFC (Liang et al., 2016), TATES (van der Sluis et al., 2013), Fisher's Combination test (FC) (Yang et al., 2016), MANOVA (Cole et al., 1994), and MultiPhen (O'Reilly et al., 2012). Our simulation studies show that the proposed method has correct type I error rates and is either the most powerful test or comparable with the most powerful tests for all simulation scenarios.

Method

Consider a sample of N unrelated individuals. Each individual has M phenotypes. We propose an allele-based method to test the null hypothesis H_0 : none of the M phenotypes are associated with a SNP of interest. Denote $Y_{i,m}$ as the M^{th} phenotype of the i^{th} individual. For the SNP of interest, we use A (minor allele) and a to denote the two alleles of this SNP.

Because each individual has two alleles at the SNP of interest, we use x_{2i-1} and x_{2i} to code the two alleles of the i^{th} individual. If the genotype of the i^{th} individual is AA , we define $x_{2i-1} = x_{2i} = 1$; if the genotype is aa , we define $x_{2i-1} = x_{2i} = 0$; and if the genotype is Aa , we define $x_{2i-1} = 1$ and $x_{2i} = 0$. We define that each allele of each individual has M phenotypes and define the m^{th} phenotype corresponding to the two alleles x_{2i-1} and x_{2i} of the i^{th} individual as $y_{2i-1,m}$ and $y_{2i,m}$, where $y_{2i-1,m} = y_{2i,m} = Y_{i,m}$. Hence, the total number of observations in the allele-based data is $2N$. We reindex the $2N$ allele-based data and use index j to denote the j^{th} allele-based data. Therefore, $y_{j,m}$ and x_j denote the j^{th} allele-based phenotypic and genotypic data.

We develop an Allele-based Clustering Approach (ACA) based on k -fold cross-validation for joint analysis of multiple phenotypes in association studies. In ACA, we consider the alleles at the SNP of interest as a dependent variable with two classes, and the correlated phenotypes as predictors to predict the alleles at the SNP of interest. In the k -fold cross-validation (in this study, $k = 10$), the $2N$ allele-based phenotypic and genotypic data are divided into k mutually exclusive groups. We use each of the k groups as the testing set and the other $k - 1$ groups as the training set. For each pair of the testing and training sets, we use the training set to calculate the centers of the phenotypes corresponding to allele a and A , respectively, namely, $\bar{y}^{(0)} = (\bar{y}_1^{(0)}, \dots, \bar{y}_M^{(0)})^T$ and $\bar{y}^{(1)} = (\bar{y}_1^{(1)}, \dots, \bar{y}_M^{(1)})^T$ and use these centers to predict the alleles in the corresponding testing set, where $\bar{y}_m^{(0)}$ and $\bar{y}_m^{(1)}$ denote the mean values of the M^{th} phenotype for subsets $\{j: x_j = 0\}$ and $\{j: x_j = 1\}$ in the training set, $\{j: x_j = 0\}$ and $\{j: x_j = 1\}$ denote the subsets of index j in which $x_j = 0$ and $x_j = 1$, respectively. In the testing set, we calculate the Euclidean distance between y_j and $\bar{y}^{(0)}$, and y_j and $\bar{y}^{(1)}$, respectively. If y_j is closer to $\bar{y}^{(0)}$ than to $\bar{y}^{(1)}$, we predict the corresponding x_j as $\hat{x}_j = 0$; otherwise, we predict the corresponding x_j as $\hat{x}_j = 1$. In this way, each x_j has a predicted value $\hat{x}_j$ for $j = 1, \dots, 2N$. For the SNP of interest, let q_0 and q_1 denote the allele frequency of a and A , respectively. For a given allele at the SNP of interest, the probability that we predict this allele by chance as allele a is q_0 and the probability that we predict this allele by chance as allele A is q_1 . Therefore, we define the prediction accuracy as

$$T_{AC} = \frac{\#\{j: \hat{x}_j = x_j = 1\}}{q_1} + \frac{\#\{j: \hat{x}_j = x_j = 0\}}{q_0}, \text{ where for a set } D, \#D \text{ denotes the number of}$$

elements in set D . We use T_{AC} as the test statistic to test the association between the M phenotypes and the SNP.

First, we delete those phenotypes that have weak associations with the SNP of interest based on the univariate analysis. For a given threshold t , $0 < t < 1$, we delete phenotypes with $p_m > t$ that means we select phenotypes with $p_m \leq t$, where p_m is the p -value to test the association between the M^{th} phenotype and the SNP. To test the association between the M^{th}

phenotype and the SNP, we use the score test statistic under the logistic model $\text{logit}(\pi_j) = \alpha_m + y_{j,m}\beta_m$, $j = 1, 2, \dots, 2N$, where $\pi_j = \Pr(x_j = 1|y_{j,m})$. The score test statistic to test the null hypothesis $H_0: \beta_m = 0$ is given by $S_m = U_m^2/V_m$, where $U_m = \sum_{j=1}^{2N} y_{j,m}(x_j - \bar{x})$, $V_m = \frac{1}{2N} \sum_{j=1}^{2N} (y_{j,m} - \bar{y}_m)^2 \sum_{j=1}^{2N} (x_j - \bar{x})^2$, $\bar{x} = \frac{1}{2N} \sum_{j=1}^{2N} x_j$, and $\bar{y}_m = \frac{1}{2N} \sum_{j=1}^{2N} y_{j,m}$. Under the null hypothesis, S_m follows a chi-square distribution with 1 degree of freedom. Let p_m denote the p-value of S_m to test the association between the M^{th} phenotype and the SNP of interest. For a given threshold t , $0 < t < 1$, we denote the prediction accuracy using the selected phenotypes as T_{AC_t} . Let P_t denote the p-value of the statistic T_{AC_t} . Our test statistic of ACA is given by

$$T_{ABC} = \min_t P_t.$$

T_{ABC} can be obtained by a simple grid search across a range of t . We choose grids $t_1, \dots, t_S$ such that $0 \leq t_1 < \dots < t_{S-1} < t_S \leq 1$. In this study, we use $(t_1, \dots, t_S) = (0.001, 0.01, 0.05, 1)$. By a grid search,

$$T_{ABC} = \min_t P_t = \min \{P_{t_1}, \dots, P_{t_S}\}.$$

We use the following permutation procedure to evaluate the p-values of T_{ABC} .

1. In each permutation, we randomly shuffle the genotypes and recalculate $T_{AC_{t_1}}, \dots, T_{AC_{t_S}}$. Suppose that we perform B times of permutations. Let $T_{AC_{t_s}}^{(b)}$ ($b = 0, 1, \dots, B$) denote the value of $T_{AC_{t_s}}$ based on the b^{th} permuted data, where $b = 0$ represents the original data.
2. We transfer $T_{AC_{t_s}}^{(b)}$ to $P_{t_s}^{(b)}$ by

$$P_{t_s}^{(b)} = \frac{\# \left\{ d: T_{AC_{t_s}}^{(d)} > T_{AC_{t_s}}^{(b)} \text{ for } d = 0, 1, \dots, B \right\}}{B} \quad (1)$$

3. Let $T_{ABC}^{(b)} = \min \{P_{t_1}^{(b)}, \dots, P_{t_S}^{(b)}\}$. Then, the p-value of T_{ABC} is given by

$$\frac{\# \left\{ b: T_{ABC}^{(b)} < T_{ABC}^{(0)} \text{ for } b = 1, 2, \dots, B \right\}}{B} \quad (2)$$

Comparison of Methods

We compare the performance of ACA with those of AFC (Liang et al., 2016), TATES (van der Sluis et al., 2013), FC (Yang et al., 2016), MANOVA (Cole et al., 1994), and MultiPhen (O'Reilly et al., 2012). Let p_m denote the p-value of score test to test the association between the M^{th} phenotype and the SNP of interest and $p_{(m)}$ denote the M^{th} smallest p-value. We briefly introduce each of those methods.

AFC

Let $T_m = -\sum_{h=1}^m (\log p_{(h)})$ ($m = 1, 2, \dots, M$) and p_{T_m} denote the p-value of T_m . The statistic of AFC to test the association between the M phenotypes and a SNP is given by $T_{AFC} = \min_{1 \leq m \leq M} p_{T_m}$. We use the same permutation procedure as we mentioned in the method section to evaluate the p-values of T_m and T_{AFC} .

TATES

TATES combines the M phenotype-specific p-values, $p_1, p_2, \dots, p_M$, obtained in standard univariate GWAS to acquire one overall p-value, $\min_{1 \leq m \leq M} \left(\frac{m_e p_{(m)}}{m_{e(m)}} \right)$, where m_e denotes the effective number of independent p-values among all M phenotypes, and $m_{e(m)}$ denotes the effective number of p-values among the top m p-values, where $m = 1, 2, \dots, M$.

FC

The Fisher's combination test statistic is defined as $T_{FC} = \sum_{m=1}^M -2\log(p_m)$. Yang et al. (Yang et al., 2016) adopted three different approaches to calculate the p-value for correlated phenotypes. In this article, we calculate the p-value using a permutation procedure.

MANOVA

Denote $Y_m = (Y_{1,m}, Y_{2,m}, \dots, Y_{N,m})^T$ as the M^{th} phenotype of N individuals, $Y = (Y_1, Y_2, \dots, Y_M)$ as the $N \times M$ matrix of phenotypes, and $G = (g_1, g_2, \dots, g_N)^T$ as the genotypic score of the N individuals at a SNP. Consider a multivariate multiple linear regression model: $Y = G\beta^T + \mathcal{E}$, where $\beta = (\beta_1, \dots, \beta_M)^T$ is a vector of coefficients corresponding to the M phenotypes, and $\mathcal{E}$ is the $N \times M$ matrix of random errors with each row of $\mathcal{E}$ to be i.i.d. $MVN(0, \Sigma)$, where Σ is the covariance matrix of $\mathcal{E}$. To test $H_0: \beta = 0$, the likelihood ratio test is equivalent to the Wilk's Lambda test statistic of MANOVA (Rencher, 2003), that is, $-2\log\Lambda = 2(l(\hat{\beta}, \hat{\Sigma}) - l(0, \hat{\Sigma}_0)) = N\log \frac{|\hat{\Sigma}_0|}{|\hat{\Sigma}|} = -N\log \left(\frac{|E|}{|E+H|} \right)$. Here Λ denotes the ratio of the likelihood function under H_0 to the likelihood function under H_1 , $l(\beta, \Sigma)$ is the

log-likelihood function, $H = \hat{\beta}(G^T G)\hat{\beta}^T$ and $E = Y^T Y - \hat{\beta}(G^T G)\hat{\beta}^T$, where $\hat{\beta} = Y^T G(G^T G)^{-1}$ is the maximum likelihood estimator (MLE) of β , and $|\cdot|$ denotes the determinant of a matrix. Then the test statistic has an asymptotic χ_M^2 distribution (Ray et al., 2016).

MultiPhen

By performing ordinal regression, MultiPhen develops a reversed analysis for joint analysis of multiple phenotypes by considering a genetic variant of interest $X = (x_1, \dots, x_n)^T$ as a dependent variable, and the correlated phenotypes $Y_k = (y_{1k}, \dots, y_{nk})^T$ as predictors.

Simulation

We generate genotype data at a SNP according to a minor allele frequency (MAF) (MAF=0.3 in all simulation studies) under Hardy-Weinberg equilibrium. Based on the genotype data, phenotypes are generated similar to that of Liang et al. (Liang et al., 2016) and van der Sluis et al. (van der Sluis et al., 2013). The phenotypic correlation structures (the phenotypes are divided into several groups (factors) and the within-group correlation is larger than the between-group correlation) are also similar to that of UK10K (UK10K Consortium, 2015).

In the following six scenarios, we use the following notations. Denote $Y_m = (Y_{1,m}, Y_{2,m}, \dots, Y_{N,m})^T$ as the M^{th} phenotype of N individuals and $G = (g_1, g_2, \dots, g_N)^T$ as the genotypic score of N individuals at the SNP. Let R denote the number of factors and $(f_1, \dots, f_R)^T \sim MVN(0, \Sigma)$, where Σ is a correlation matrix with all off-diagonal elements equal to ρ_{fa} and ρ_{fa} is the correlation between any two factors ($\rho_{fa} = 0.1$ in all simulation studies). β is the effect size; a is a factor loading ($a = 0.75$ in all simulation studies); $\epsilon_m = (\epsilon_{1m}, \dots, \epsilon_{Nm})^T$ and ϵ_{im} ($i = 1, \dots, N; m = 1, \dots, M$) are random errors.

Scenario 1

We simulate M phenotypes by

$$Y_m = a\beta G + \epsilon_m \text{ for } m = 1, 2, \dots, M \quad (3)$$

Scenario 2

Consider a 4-factor model ($R = 4$) with the SNP effect on the fourth factor. We simulate M phenotypes using

$$Y_m = \begin{cases} af_1 + \varepsilon_m & \text{for } m = 1, \dots, \frac{M}{4} \\ af_2 + \varepsilon_m & \text{for } m = \frac{M}{4} + 1, \dots, \frac{M}{2} \\ af_3 + \varepsilon_m & \text{for } m = \frac{M}{2} + 1, \dots, \frac{3M}{4} \\ a(f_4 + \beta G) + \varepsilon_m & \text{for } m = \frac{3M}{4} + 1, \dots, M \end{cases} \quad (4)$$

Scenario 3

Consider a 2-factor model ($R = 2$) with the SNP effect on the second factor. We simulate M phenotypes using

$$Y_m = \begin{cases} af_1 + \varepsilon_m & \text{for } m = 1, \dots, \frac{M}{2} \\ a(f_2 + \beta G) + \varepsilon_m & \text{for } m = \frac{M}{2} + 1, \dots, M \end{cases} \quad (5)$$

Scenario 4

Consider a 4-factor model ($R = 4$) with the SNP effect on the M^{th} phenotype. We simulate M phenotypes using

$$Y_m = \begin{cases} af_1 + \varepsilon_m & \text{for } m = 1, \dots, \frac{M}{4} \\ af_2 + \varepsilon_m & \text{for } m = \frac{M}{4} + 1, \dots, \frac{M}{2} \\ af_3 + \varepsilon_m & \text{for } m = \frac{M}{2} + 1, \dots, \frac{3M}{4} \\ af_4 + \varepsilon_m & \text{for } m = \frac{3M}{4} + 1, \dots, M-1 \\ a(f_4 + \beta G) + \varepsilon_m & \text{for } m = M \end{cases} \quad (6)$$

Scenario 5

Consider a 2-factor model ($R = 2$) with the SNP effect on the M^{th} phenotype. We simulate M phenotypes using

$$Y_m = \begin{cases} af_1 + \varepsilon_m & \text{for } m = 1, \dots, \frac{M}{2} \\ af_2 + \varepsilon_m & \text{for } m = \frac{M}{2} + 1, \dots, M-1 \\ a(f_2 + \beta G) + \varepsilon_m & \text{for } m = M \end{cases} \quad (7)$$

Scenario 6

Consider an M -factor model ($R = M$) with the SNP effect on all phenotypes. We simulate M phenotypes using

$$Y_m = a(f_m + \beta G) + \varepsilon_m \text{ for } m = 1, 2, \dots, M \quad (8)$$

In scenarios 1-6, we consider random errors having the following four non-normal distributions: two log-normal distributions ($\log\text{-norm}(0, 1)$ and $\sqrt{5}\log\text{-norm}(0, 1)$), Student's t -distribution ($t(2)$), and Inverse-gamma distribution ($\text{Inverse-gamma}(2, 1)$). We also consider random errors following the standard normal distribution. To evaluate type I error rates, we generate phenotypic values independent of genotypes by assigning $\beta = 0$. To evaluate power, we generate phenotypic values according to the six scenarios described above. In type I error evaluations and power comparisons, MAF is 0.3, the total number of phenotypes (M) is 8, the sample size (N) is 1,000, the factor loading (a) is 0.75, and the factorial correlation (ρ_{fa}) conditional on genotypes is 0.1.

Simulation Results

In each simulation scenario, the p-values of the proposed test (ACA), AFC, and FC are estimated using 1,000 permutations and the p-values of TATES, MANOVA, and MultiPhen are estimated using asymptotic distributions.

For type I error evaluation, we consider different distributions of random errors, different correlation structures of phenotypes, and different significance levels. In each simulation scenario, the type I error rates of all of the six tests are evaluated using 1,000 replicated samples. For 1,000 replicated samples, the 95% confidence intervals (CIs) for the type I error rates at the nominal levels 0.01 and 0.05 are (0.0038, 0.0162) and (0.0365, 0.0635), respectively. The estimated type I error rates for random errors following Student's $t(2)$ distribution, Inverse-gamma(1,2) distribution, $\log\text{-norm}(0, 1)$ distribution, $\sqrt{5}\log\text{-norm}(0, 1)$ distribution, and the standard normal distribution are summarized in Table 1 and Tables S1-S4, respectively. From these tables, we can see that most of the estimated type I error rates of ACA, AFC, FC, TATES, and MANOVA are within the 95% CIs and those type I error rates not within 95% CIs are very close to the bound of the corresponding 95% CI, which indicates that ACA, AFC, FC, TATES, and MANOVA are valid. MultiPhen for Inverse-gamma(1,2) distribution has little inflated type I error rates.

The powers of all tests are evaluated using 1,000 replicated samples at 1% significance level. Figure 1 and Figures S1-S4 provide the power comparisons of the six tests (ACA, AFC, TATES, FC, MANOVA, and MultiPhen) for the power as a function of the effect size based on the six scenarios for 8 phenotypes. Figure 1 shows that when random errors follow a Student's t distribution, ACA is the most powerful test in all the scenarios and is much more powerful than other methods. Figure S1 shows that when random errors follow an Inverse-gamma distribution, ACA is the most powerful test in all the scenarios and is much more powerful than the other methods when genotypes directly impact on a portion of phenotypes

(scenarios 2-5). Figure S2 shows that when random errors follow a log-normal distribution, ACA is the most powerful test when genotypes directly impact on a portion of phenotypes (scenarios 2-5) and is comparable to the most powerful one when genotypes directly impact on all phenotypes (scenarios 1 and 6). To show the robustness of ACA to phenotype outliers, we generate random errors following a log-normal distribution with a variance 5 times as that in Figure S2. As shown in Figure S3, increasing variance 5 times, the pattern of power comparisons is still the same as that in Figure S2. We also provide power comparisons when random errors follow a standard normal distribution (Figure S4). As expected, ACA loses some power comparing with that when random errors follow non-normal distributions, but is still more powerful than TATES for scenarios 1 and 6, more powerful than FC for scenarios 4 and 5, and more powerful than MANOVA and MultiPhen for scenario 3.

Discussion

In GWAS for complex diseases, we focus on common variants ($MAF \geq 0.05$). However, the association between a SNP and each phenotype is usually weak. Combining multiple related phenotypes can increase the power to identify causal SNPs, thus is a practically important area that requires methodology work. So far, existing methods for associations of multiple phenotypes primarily focus on the phenotypes with a normal distribution. However, when the normal assumption is violated, these methods may not control type I error rates or may lose power (Majumdar et al., 2015).

In this paper, we introduced an Allele-based Clustering Approach (ACA) for joint analysis of multiple non-normal distributed phenotypes in association studies. ACA is nonparametric in that it does not assume any particular genetic models and there are no parameters need to be estimated. We use a data-driven approach to select phenotypes that are associated with a SNP and use these phenotypes to predict the alleles at the SNP by using Euclidean distance. Besides Euclidean distance, we can also use other similarity or dissimilarity measures to predict the alleles at a SNP, such as Minkowski distance, Mahalanobis distance, correlation coefficient, and cosine measure.

We use extensive simulations to assess the performance of ACA. We are able to demonstrate that ACA has correct type I error rates and is much more powerful than other methods for some non-normal distributions. Therefore, our simulations show that ACA is especially useful for non-normal data and it truly adds to the geneticist's toolbox. Moreover, ACA does not estimate any parameters and does not assume any particular models. It provides another way for analyzing multiple phenotypes simultaneously. In the method section, we describe our methods without considering covariates. If there are covariates need to be considered, we can use the following approach to incorporate covariates in ACA. Suppose that we have

p covariates. Let $(z_{i1}, \dots, z_{ip})^T$ denote the covariates of the i^{th} individual. We can adjust the effects of covariates by applying the linear regression model $Y_{i,m} = a_0 + a_{1m}z_{i1} + \dots + a_{pm}z_{ip} + \varepsilon_{im}$ and using the residual of $Y_{i,m}$ to replace $Y_{i,m}$ in ACA. Although using the regular linear regression to regress effects of covariates out is more suited for normally distributed data, it can also be applied to non-normally distributed data. For examples, to correct for population stratification, Price et al. (2006) used the regular

linear regression to regress effects of covariates out for a qualitative phenotype and genotypes (both a qualitative phenotype and genotypes are non-normally distributed data); to adjust effects of covariates in rare variant association studies, Sha et al. (2012) also used the regular linear regression to regress effects of covariates out for a qualitative phenotype and genotypes.

In the definition of T_{AC} , we adjusted the prediction accuracy by dividing the corresponding allele frequency, therefore, our proposed method is only suitable for GWAS and is not suitable for whole genome sequencing studies with rare variants. In GWAS, all the variants are common ($MAF \geq 0.05$). Although in our simulation studies, we used $MAF=0.3$, we also performed power comparisons for $MAF=0.05$ (results not shown). The patterns of the power comparisons for $MAF=0.05$ are same as those of $MAF=0.3$. Another condition for using the proposed method is that it is suitable to analyze multiple phenotypes with non-normal distributions. The proposed method is more powerful than others for some non-normal distributions. For normal distributions, our proposed method is less powerful than some methods.

The computation time required for running ACA depends on the sample size, the number of phenotypes, the number of folds in cross-validation, and the number of permutations. The running time of ACA with 1,000 permutations on the data set with 10,000 individuals, 8 phenotypes, and 10-fold cross-validation on a laptop with 4 Intel Cores @ 2.00GHz and 4 GB memory is no more than 2.4s. To perform genome-wide studies, we can first select SNPs that show evidence of association based on a small number of permutations (e.g. 1,000), and then a large number of permutations are used to test the selected variants.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number R15HG008209. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Superior, a high-performance computing infrastructure at Michigan Technological University, was used in obtaining results presented in this publication.

References

- Aschard H, Vilhjálmsson BJ, Greliche N, Morange PE, Trégouët DA, Kraft P. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am J Hum Genet.* 2014; 94:662–676. [PubMed: 24746957]
- Casaburi R, Mahler DA, Jones PW, Wanner A, San Pedro G, ZuWallack RL, Menjoge SS, Serby CW, Witek T. A long-term evaluation of once-daily inhaled tiotropium in chronic obstructive pulmonary disease. *Eur Respir J.* 2002; 19:217–224. [PubMed: 11866001]
- Casale FP, Rakitsch B, Lippert C, Stegle O. Efficient set tests for the genetic analysis of correlated traits. *Nat Methods.* 2015; 12:755–758. [PubMed: 26076425]
- Cole DA, Maxwell SE, Arvey R, Salas E. How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychol Bull.* 1994; 115:465.

- Galesloot TE, Van Steen K, Kiemeny LA, Janss LL, Vermeulen SH. A comparison of multivariate genome-wide association methods. *PloS one*. 2014; 9:e95923. [PubMed: 24763738]
- Germain M, Saut N, Greliche N, Dina C, Lambert JC, Perret C, Cohen W, Oudot-Mellakh T, Antoni G, Alessi MC, Zelenika D. Genetics of venous thrombosis: insights from a new genome wide association study. *PloS one*. 2011; 6:e25581. [PubMed: 21980494]
- Klei L, Luca D, Devlin B, Roeder K. Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genet Epidemiol*. 2008; 32:9–19. [PubMed: 17922480]
- Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat Genet*. 2012; 44:1066–1071. [PubMed: 22902788]
- Kwak IY, Pan W. Gene-and pathway-based association tests for multiple phenotypes with GWAS summary statistics. *Bioinformatics*. 2016:btw577.
- Lange C, Van Steen K, Andrew T, Lyon H, DeMeo DL, Raby B, Murphy A, Silverman EK, MacGregor A, Weiss ST, Laird NM. A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. *Stat Appl Genet Molec Biol*. 2004; 3:1–27.
- Liang X, Wang Z, Sha Q, Zhang S. An adaptive Fisher's combination method for joint analysis of multiple phenotypes in association studies. *Sci Rep*. 2016; 6:34323. [PubMed: 27694844]
- Majumdar A, Witte JS, Ghosh S. Semiparametric allelic tests for mapping multiple phenotypes: Binomial regression and mahalanobis distance. *Genet Epidemiol*. 2015; 39:635–650. [PubMed: 26493781]
- Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*. 2007; 39:906–913. [PubMed: 17572673]
- Newman WP III, Freedman DS, Voors AW, Gard PD, Srinivasan SR, Cresanta JL, Williamson GD, Webber LS, Berenson GS. Relation of serum lipoprotein levels and systolic blood pressure to early atherosclerosis. *N Engl J Med*. 1986; 314:138–144. [PubMed: 3455748]
- O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FC, Elliott P, Jarvelin MR, Coin LJ. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PloS one*. 2012; 7:e34861. [PubMed: 22567092]
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006; 38:904–909. [PubMed: 16862161]
- Ray D, Pankow JS, Basu S. USAT: a unified score-based association test for multiple phenotype-genotype analysis. *Genet Epidemiol*. 2016; 40:20–34. [PubMed: 26638693]
- Rencher AC. *Methods of multivariate analysis*. Vol. 492. John Wiley & Sons; 2003.
- Sha Q, Wang X, Wang X, Zhang S. Detecting association of rare and common variants by testing an optimally weighted combination of variants. *Genet Epidemiol*. 2012; 36:561–571. [PubMed: 22714994]
- Souto JC, Almasry L, Borrell M, Blanco-Vaca F, Mateo J, Soria JM, Coll I, Felices R, Stone W, Fontcuberta J, Blangero J. Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. *Am J Hum Genet*. 2000; 67:1452–1459. [PubMed: 11038326]
- Tang CS, Ferreira MA. A gene-based test of association using canonical correlation analysis. *Bioinformatics*. 2012; 28:845–850. [PubMed: 22296789]
- UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature*. 2015; 526:82–90. [PubMed: 26367797]
- van der Sluis S, Posthuma D, Dolan CV. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet*. 2013; 9:e1003235. [PubMed: 23359524]
- Yan T, Li Q, Li Y, Li Z, Zheng G. Genetic association with multiple traits in the presence of population stratification. *Genet Epidemiol*. 2013; 37:571–580. [PubMed: 23740720]
- Yang JJ, Li J, Williams LK, Buu A. An efficient genome-wide association test for multivariate phenotypes based on the Fisher combination function. *BMC Bioinform*. 2016; 17:1.

- Zhang Y, Xu Z, Shen X, Pan W, Alzheimer's Disease Neuroimaging Initiative. Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. *Neuroimage*. 2014; 96:309–325. [PubMed: 24704269]
- Zhou JJ, Cho MH, Lange C, Lutz S, Silverman EK, Laird NM. Integrating multiple correlated phenotypes for genetic association analysis by maximizing heritability. *Hum Hered*. 2015; 79:93–104. [PubMed: 26111731]
- Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods*. 2014; 11:407–409. [PubMed: 24531419]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure. 1.

Table 1
The estimated type I error rates of the 6 tests under Student’s $t(2)$ distribution. MAF is 0.3. The total number of phenotypes (M) is 8. The sample size (N) is 1,000. The factor loading (λ) is 0.75. The factorial correlation (ρ_{fd}) conditional on genotypes is 0.1. The number of replications is 1,000. The type I error rates in italics indicate the values out of the bounds of the 95% CIs.

Significance Level	Method	Scenario					
		1	2	3	4	5	6
0.01	ACA	0.009	0.010	0.015	0.012	0.008	0.011
	AFC	0.009	0.010	0.015	0.013	0.015	0.015
	TATES	0.007	0.007	0.011	0.007	0.007	0.007
	FC	0.009	0.011	0.011	0.008	0.011	<i>0.018</i>
	MANOVA	0.008	0.010	0.009	0.008	0.008	0.013
	MultiPhen	0.014	0.007	0.014	0.008	0.012	<i>0.019</i>
0.05	ACA	0.039	0.042	0.055	0.048	0.045	0.057
	AFC	0.050	0.058	0.051	0.050	0.049	0.049
	TATES	0.042	0.045	0.043	0.041	0.053	0.044
	FC	0.053	0.054	0.056	0.046	0.045	0.048
	MANOVA	0.050	0.051	0.046	0.040	0.044	0.046
	MultiPhen	0.062	0.052	0.063	0.042	0.058	<i>0.069</i>