Published in final edited form as:

Genet Epidemiol. 2018 June ; 42(4): 344–353. doi:10.1002/gepi.22124.

A Hierarchical Clustering Method for Dimension Reduction in Joint Analysis of Multiple Phenotypes

Xiaoyu Liang¹, Qiuying Sha¹, Yeonwoo Rho¹, and Shuanglin Zhang¹

¹Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan, United States of America

Abstract

Genome-wide association studies (GWAS) have become a very effective research tool to identify genetic variants of underlying various complex diseases. In spite of the success of GWAS in identifying thousands of reproducible associations between genetic variants and complex disease, in general, the association between genetic variants and a single phenotype is usually weak. It is increasingly recognized that joint analysis of multiple phenotypes can be potentially more powerful than the univariate analysis, and can shed new light on underlying biological mechanisms of complex diseases. In this paper, we develop a novel variable reduction method using hierarchical clustering method (HCM) for joint analysis of multiple phenotypes in association studies. The proposed method involves two steps. The first step applies a dimension reduction technique by using a representative phenotype for each cluster of phenotypes. Then, existing methods are used in the second step to test the association between genetic variants and the representative phenotypes rather than the individual phenotypes. We perform extensive simulation studies to compare the powers of MANOVA, MultiPhen, and TATES using HCM with those of without using HCM. Our simulation studies show that using HCM is more powerful than without using HCM in most scenarios. We also illustrate the usefulness of using HCM by analyzing a whole-genome genotyping data from a lung function study.

Keywords

Association Study; Multiple Phenotypes; Dimension Reduction; Hierarchical Clustering

1 | INTRODUCTION

The successful applications of genome-wide association studies (GWAS) to numerous complex diseases established a large number of genetic associations (Lutz, Fingerlin, Hokanson, & Lange, 2017). Through GWAS, numerous genes have been shown to affect

Correspondence: Shuanglin Zhang, Department of Mathematical Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931. shuzhang@mtu.edu.

AUTHOR CONTRIBUTIONS

S. Z., Q. S., and Y. R. 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.

multiple phenotypes and yet the effect size on each phenotype is small for complex diseases (Yang, Williams, & Buu, 2017). For example, multiple GWAS have found significant signals in the chromosome 15q25 region associated with lung cancer (Chen et al., 2015), chronic obstructive lung disease (COPD) (Cho et al., 2014), emphysema (Cho et al., 2015), and cigarette smoking (Hancock et al., 2015).

Simultaneous testing of multiple phenotypes has been widely recognized as a valuable approach complementary to single phenotype tests. There are two main reasons: one is to increase statistical power, and the other is to shed light on underlying biology to possibly repurpose the use of existing drugs (Deng & Pan, 2017). Therefore, there is an increasing interest in joint analysis of multiple phenotypes with many new tests being recently proposed (Aschard et al., 2014; Casale, Rakitsch, Lippert, & Stegle, 2015; Cole, Maxwell, Arvey, & Salas, 1994; Galesloot, Van Steen, Kiemeny, Janss, & Vermeulen, 2014; Klei, Luca, Devlin, & Roeder, 2008; Korte et al., 2012; Lange et al., 2004; Liang, Wang, Sha, & Zhang, 2016; Marchini, Howie, Myers, McVean, & Donnelly, 2007; O'Reilly et al., 2012; Tang & Ferreira, 2012; Wang, Sha, & Zhang, 2016; Yan, Li, Li, & Zheng, 2013; Zhang, Xu, Shen, Pan, & Alzheimer's Disease Neuroimaging Initiative, 2014; Zhou & Stephens, 2014; Zhou et al., 2015; Zhu, Zhang, & Sha, 2015a).

Existing methods for joint analysis of multiple phenotypes roughly fall into three categories: regression methods, combining test statistics from univariate analyses, and variable reduction methods (Yang & Wang, 2012). In the first category, regression methods, there are three different approaches for analyzing the association of a genetic variant with multiple phenotypes: mixed effects models (Bates & DeBroy, 2004; Yan et al., 2013), generalized estimating equations (Liang & Zeger, 1986), and frailty models (Therneau, Grambsch, & Pankratz, 2003). Tests that fall into the second category, combining test statistics from univariate analyses, conduct a univariate analysis first and then aggregate univariate test statistics. This approach is simple and feasible for meta-analyses (Schaid et al., 2016; Yang, Li, Williams, & Buu, 2016). Recently, many methods of combining test statistics from univariate analyses have been developed 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, Posthuma, & Dolan, 2013; Yang et al., 2016). In the last category, tests based on variable reduction methods are roughly depending on three dimension reduction techniques. The first one is the principal component analysis of phenotypes (PCP) (Aschard et al., 2014). In PCP, the first few principal components (PCs) explaining most of the total phenotype variance are tested for association with a genetic variant, and the remaining components are not analyzed. However, Aschard et al. (2014) showed that considering only the first few PCs often causes low power, whereas considering all PCs can improve the power. The second one is the canonical correlation analysis (CCA) (Tang & Ferreira, 2012). CCA searches for the linear combinations that maximize the correlation between two sets of multidimensional variables. It provides an efficient and powerful approach for both univariate and multivariate tests of association without the need for the permutation test. The last one is the principal component of heritability (PCH) (Klei et al., 2008; Ott & Rabinowitz, 1999; Wang et al., 2016). PCH reduces multiple phenotypes to a linear combination of phenotypes that has the highest heritability among all linear combinations of phenotypes.

In this article, we develop a novel variable reduction method called hierarchical clustering method (HCM) for joint analysis of multiple phenotypes. HCM is a dimension reduction technique by using a representative phenotype for each cluster of phenotypes, then using existing methods for joint analysis of multiple phenotypes to test the association between a genetic variant of interest and the representative phenotypes rather than the individual phenotypes. One way to understand the dimension reduction technique of HCM is that when one cluster consists of highly positively correlated phenotypes, any linear combination of the phenotypes within this cluster can represent the cluster reasonably well (Bühlmann, Rütimann, van de Geer, & Zhang, 2013; Shah & Samworth, 2013). HCM does not require phenotypes themselves, it only requires a dissimilarity matrix of the phenotypes. This dissimilarity matrix can be estimated from the values of summary statistics using all independent single-nucleotide polymorphisms (SNPs) in a GWAS (Zhu et al., 2015b). We use extensive simulation studies to show the validity of the proposed two-step method and to investigate the power. In particular, the performance of three existing methods using HCM, multivariate analysis of variance (MANOVA) (Cole et al., 1994), joint model of multiple phenotypes (MultiPhen) (O'Reilly et al., 2012), and trait-based association test that uses extended simes procedure (TATES) (Van der Sluis et al., 2013), is compared with that of without using HCM. Our simulation studies show that MANOVA, MultiPhen, and TATES using HCM have correct type I error rates and are more powerful than MANOVA, MultiPhen, and TATES without using HCM in most scenarios. We also apply MANOVA, MultiPhen, and TATES with and without using HCM to COPDGene data to further demonstrate the usefulness of HCM.

2 | METHODS

2.1 | Hierarchical clustering method for joint analysis of multiple phenotypes

Consider a sample with n unrelated individuals. Each individual has K phenotypes. Denote $Y_k = (y_{1k}, \dots, y_{nk})^T$ as the k^{th} phenotype of n individuals and $Y = (Y_1, \dots, Y_K)$ as the $n \times K$ phenotype matrix. Denote $X = (x_1, \dots, x_n)^T$ as the genotypic score of n individuals at a genetic variant of interest, where $x_i \in \{0, 1, 2\}$ is the number of minor alleles that the i^{th} individual carries at the genetic variant.

The proposed hierarchical clustering method (HCM) involves two steps. In the first step, we divide the K phenotypes into M clusters and use a representative phenotype for each of the M clusters. In the second step, we apply existing methods to the M representative phenotypes rather than directly to the individual phenotypes to test the association between phenotypes and the variant. In the first step, we need to find a partition $\mathcal{G}$ that partitions K phenotypes into M disjoint clusters $G_1, \dots, G_M$, where $\mathcal{G} = \{G_1, \dots, G_M\}$ with

$$\bigcup_{m=1}^M G_m = \{1, \dots, K\} \text{ and } G_m \cap G_\ell = \emptyset \quad (m \neq \ell). \text{ In this article, we use a hierarchical}$$

clustering strategy to cluster the phenotypes.

Strategies for hierarchical clustering generally fall into two types: agglomerative (bottom-up) and divisive (top-down). The agglomerative method starts with all phenotypes in their own cluster and merges the two clusters that have the smallest dissimilarity in each clustering iteration until there is only one single cluster left. The divisive method starts with all phenotypes in one cluster and splits the cluster into two that have the largest dissimilarity in each clustering iteration until all phenotypes are in their own cluster. Both methods can be described by a dendrogram which is frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering. We need a stopping criterion to cut the dendrogram into several clusters.

In this study, we use the bottom-up hierarchical clustering method based on the dissimilarity matrix of the phenotypes. We define the dissimilarity matrix D with entries $d_{ij} = 1 - P_{ij}^s$, where P_{ij}^s is the ij^{th} entry of $P^s(Y)$ and $P^s(Y)$ is the sample correlation matrix of $Y = (Y_1, \dots, Y_K)$. We choose the average linkage as the dissimilarity between two clusters. Hence, the dissimilarity between clusters G_m and G_ℓ is given by

$$h(G_m, G_\ell) = \frac{1}{|G_m| |G_\ell|} \sum_{i \in G_m, j \in G_\ell} d_{ij} \quad (1)$$

where $|G_m|$ denote the number of phenotypes in G_m . Using the bottom-up hierarchical clustering method, we start with each phenotype as a singleton cluster and then successively merge pairs of clusters with the smallest dissimilarity calculated by equation (1) until all clusters have been merged into a single cluster that contains all phenotypes. We refer the smallest dissimilarity in each iteration as the height of the merged cluster in the dendrogram. We determine the number of clusters in the HCM using a stopping criterion. The stopping criterion is similar to an established principle (Bühlmann et al., 2013). Let h_b denote the smallest dissimilarity between two clusters in iteration b ($b \geq 1$) or the height of iteration b . We define:

$$\hat{b} = \arg \max_{b \geq 1} (h_{b+1} - h_b) \quad (2)$$

Then, we choose the number of clusters identified at the iteration $\hat{b}$.

Before we define the representative phenotype for each cluster, we first scale each phenotype. We define the representative phenotype for the m^{th} cluster as the average phenotype values in the cluster, that is

$$\bar{Y}^{(m)} = \frac{1}{|G_m|} \sum_{k \in G_m} Y_k, \quad m = 1, \dots, M \quad (3)$$

Let $\bar{Y}$ denote the $n \times M$ design matrix whose m^{th} column is given by $\bar{Y}^{(m)}$. Then we apply existing methods to test the association between $\bar{Y}$ and X .

2.2 | Comparison of methods

We compare the performance of MANOVA, MultiPhen, and TATES with using HCM with that of without using HCM. We refer the ones with using HCM as HCMANOVA, HCMultiPhen, and HCTATES, respectively. Since principal component analysis (PCA) is a popular dimension reduction method, we also compare the performance of MANOVA, MultiPhen, and TATES using HCM with that of using the first few PCs of the phenotypes. We choose the number of PCs that explain 95% of the total variance of the phenotypes. We refer MANOVA, MultiPhen, and TATES using PCs as PCMANOVA, PCMultiPhen, and PCTATES, respectively.

3 | RESULTS

3.1 | Simulation studies

To evaluate the type I error rates and powers of HCM, we generate genotypes at a genetic variant according to the minor allele frequency (MAF) under Hardy Weinberg equilibrium. Then, we generate K phenotypes by the factor model (Wang et al., 2016)

$$y = \lambda x + c\gamma f + \sqrt{1 - c^2} \times \varepsilon \quad (4)$$

where $y = (y_1, \dots, y_K)^T$; x is the genotype score at the variant; $\lambda = (\lambda_1, \dots, \lambda_K)^T$ is the vector of effect sizes of the genetic variant on the K phenotypes; f is a vector of factors, $f = (f_1, \dots, f_R)^T \sim MVN(0, \Sigma)$, $\Sigma = (1 - \rho)I + \rho A$, A is a matrix with elements of 1, I is the identity matrix, R is the number of factors, and ρ is the correlation between factors; γ is a K by R matrix; c is a constant number; and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_K)^T$ is a vector of residuals, and $\varepsilon_1, \dots, \varepsilon_K$ are independent, and $\varepsilon_k \sim N(0, 1)$ for $k = 1, \dots, K$. Based on equation (4), we consider the following four models with different number of factors and different number of factors affected by genotypes. In the four models, the within-factor correlation is c^2 and the between-factor correlation is ρc^2 .

Model 1—There is only one factor and genotypes impact on all phenotypes. That is, $R = 1$, $\lambda = \beta(1, 2, \dots, K)^T$, and $\gamma = (1, \dots, 1)^T$.

Model 2—There are two factors and genotypes impact on one factor. That is, $R = 2$,

$$\lambda = \left(0, \dots, 0, \beta, \dots, \beta \right)^T, \text{ and } \gamma = \text{diag}(D_1, D_2), \text{ where } D_i = \begin{pmatrix} 1, \dots, 1 \\ \diamond \\ K/2 \end{pmatrix}^T \text{ for } i = 1, 2.$$

Model 3—There are five factors and genotypes impact on two factors. That is, $R = 5$,

$$\lambda = (\beta_{11}, \dots, \beta_{1k}, \beta_{21}, \dots, \beta_{2k}, \beta_{31}, \dots, \beta_{3k}, \beta_{41}, \dots, \beta_{4k}, \beta_{51}, \dots, \beta_{5k})^T, \text{ and}$$

$$\gamma = \text{diag}(D_1, D_2, D_3, D_4, D_5), \text{ where } D_i = \begin{pmatrix} 1, \dots, 1 \\ \diamond \\ K/5 \end{pmatrix}^T \text{ for } i = 1, \dots, 5, k = \frac{K}{5},$$

$$\beta_{11} = \dots \beta_{1k} = \beta_{21} = \dots = \beta_{2k} = \beta_{31} = \dots = \beta_{3k} = 0, \beta_{41} = \dots = \beta_{4k} = -\beta, \text{ and}$$

$$(\beta_{51}, \dots, \beta_{5k}) = \frac{2\beta}{k+1}(1, \dots, k).$$

Model 4—There are five factors and genotypes impact on four factors. That is, $R = 5$,

$$\lambda = (\beta_{11}, \dots, \beta_{1k}, \beta_{21}, \dots, \beta_{2k}, \beta_{31}, \dots, \beta_{3k}, \beta_{41}, \dots, \beta_{4k}, \beta_{51}, \dots, \beta_{5k})^T, \text{ and}$$

$$\gamma = \text{diag}(D_1, D_2, D_3, D_4, D_5), \text{ where } D_i = \begin{pmatrix} 1, \dots, 1 \\ \diamond \\ K/5 \end{pmatrix}^T \text{ for } i = 1, \dots, 5, k = \frac{K}{5}, \beta_{11} = \dots \beta_{1k} = 0,$$

$$\beta_{21} = \dots = \beta_{2k} = \beta, \beta_{31} = \dots = \beta_{3k} = -\beta, \beta_{41} = \dots = \beta_{4k} = -\frac{2\beta}{k+1}(1, \dots, k), \text{ and}$$

$$(\beta_{51}, \dots, \beta_{5k}) = \frac{2\beta}{k+1}(1, \dots, k).$$

To evaluate type I error rates, we let $\beta = 0$. To evaluate powers, we let $\beta > 0$. In the simulation studies for evaluation of type I error rates and powers, we set MAF = 0.3, $c = \sqrt{0.5}$, and $\rho = 0.2$.

3.2 | Simulation results

For each model, we estimate the p-values of all test statistics using their asymptotic distributions.

For type I error evaluation, we consider different numbers of phenotypes, different significance levels, different sample sizes, and different models. For 10,000 replicated samples, the 95% confidence intervals (CIs) for type I error rates at the nominal levels 0.05, 0.01, and 0.001 are (0.0457, 0.0543), (0.008, 0.012), and (0.0004, 0.0016), respectively. The estimated type I error rates of HCMANOVA, HCMultiPhen, and HCTATES are summarized in Tables 1 to 3. The estimated type I error rates of PCMANOVA, PCMultiPhen, and PCTATES are summarized in Tables S1 to S3. From these tables, we can see that most of the estimated type I error rates are within the 95% CIs. In addition, the type I error rates outside of the 95% CIs are very close to the bounds of the corresponding 95% CI, which indicates that HCMANOVA, HCMultiPhen, HCTATES, PCMANOVA, PCMultiPhen, and PCTATES are valid tests.

For power comparisons, we consider different numbers of phenotypes and different models. The powers of all tests are evaluated based on 1000 replications and 5000 subjects at 5% significance level. Figure 1 and Figure 2 provide the power comparisons of the six tests (HCMANOVA, MANOVA, HCMultiPhen, MultiPhen, HCTATES, and TATES) for the power as a function of the effect size under the four models. We consider 20 phenotypes and 40 phenotypes in Figure 1 and Figure 2, respectively.

These two figures show that (1) when the effect sizes of the genetic variant on phenotypes show no groups (Model 1), HCMANOVA, HCMultiPhen, and HCTATES are slightly less powerful than MANOVA, MultiPhen, and TATES, respectively, because in most replications, HCM clusters each phenotype in a singleton cluster; (2) when the effect sizes show some groups and have the same direction (Model 2), HCMANOVA, HCMultiPhen, and HCTATES are much more powerful than MANOVA, MultiPhen, and TATES, respectively; (3) when the effect sizes show some groups and have different directions (Models 3 and 4), HCMANOVA and HCMultiPhen are more powerful than MANOVA and MultiPhen, respectively, but HCTATES is less powerful than TATES; (4) HCMANOVA and HCMultiPhen have similar power; MANOVA and MultiPhen have similar power; (5) HCTATES and TATES are much less powerful than other methods when genotypes directly impact on all phenotypes (Model 1). Figures S1 and S2 provide the power comparisons of HCM with those of using PCs of phenotypes that explain 95% of the total variance. These figures show that using HCM as a dimension reduction method is more powerful than using PCs that explain 95% of the total variance. We also set up an additional simulation model (Model S1) to compare the powers of MANOVA, MultiPhen, and TATES with those of HCMANOVA, HCMultiPhen, and HCTATES (Figure S3). Figure S3 shows that under Model S1, HCMANOVA, HCMultiPhen, and HCTATES are more powerful than MANOVA, MultiPhen, and TATES, respectively.

In summary, the existing methods using HCM have correct type I error rates and are more powerful than or comparable with those without using HCM, and the existing methods using HCM are also more powerful than those using PCs of phenotypes as a dimension reduction method.

3.3 | Real data analysis

Chronic obstructive pulmonary disease (COPD) is a progressive respiratory disease including chronic bronchitis, emphysema, non-reversible asthma, and some forms of bronchiectasis. This disease is characterized by reduced maximum expiratory flow and slow forced emptying of the lungs (Siafakas et al., 1995). Despite being a treatable and preventable disease, it is still a major cause of morbidity and mortality. The prevalence continues to rise because of the worldwide epidemic of smoking (Nazir & Erbland, 2009). In this article, we demonstrated the application of the proposed method by conducting analysis on the data from the Genetic Epidemiology of COPD (COPDGene). The COPDGene Study was designed to investigate the underlying genetic factors of COPD, to define and characterize disease-related phenotypes, and to assess the association of disease-related phenotypes with the identified susceptibility genes (Regan et al., 2011).

To evaluate the performance of our proposed method on a real data set, we applied the six methods to the COPDGene of non-Hispanic Whites population to carry out GWAS of COPD-related phenotypes. Similar to Liang et al., 2016, we selected 7 quantitative COPD-related phenotypes, including % predicted FEV1 (FEV1), Emphysema (Emph), Emphysema Distribution (EmphDist), Gas Trapping (GasTrap), Airway Wall Area (Pi10), Exacerbation Frequency (ExacerFreq), Six-minute Walk Distance (6MWD), and 4 covariates, including BMI, Age, Pack-Years (PackYear) and Sex. In our analysis, EmphDist is the ratio of emphysema at -950 HU in the upper 1/3 of lung fields compared to the lower 1/3 of lung fields. Followed by Chu et al. (Chu et al., 2014), we did a log transformation on EmphDist in the following analysis. We excluded participants with missing data in any of the 11 variables. There were total 5,430 individuals across 630,860 SNPs used in the analyses. We first adjusted each of the 7 phenotypes for the 4 covariates using linear models. Then, we performed the analysis based on the adjusted phenotypes. The detailed information can be found in Liang et al. (Liang et al., 2016).

Based on the correlation structure of the 7 COPD-related phenotypes given in Figure 4 in Liang et al. 2016, we changed the signs for phenotypes of FEV1 and 6MWD because the correlations of FEV1 and 6MWD with other 5 phenotypes are all negative. After changing the signs for the phenotypes of FEV1 and 6MWD, the pair-wise correlations among the 7 phenotypes are all positive.

To identify SNPs associated with the 7 COPD-related phenotypes, we adopted the commonly used genome-wide significance level 5×10^{-8} to account for multiple testing. HCM divided the 7 phenotypes into 5 clusters (Figure 3). The first cluster contains three phenotypes including FEV1, Emph, and GasTrap. Each of the other four clusters contains only one phenotype. Table 4 summarized the significant SNPs identified by at least one method. There are 14 SNPs in Table 4. All of the 14 SNPs had previously been reported to be in association with COPD (Brehm et al., 2011; Cho et al., 2010 & 2014; Cui, Ge, & Ma, 2014; Du, Xue, & Xiao, 2016; Hancock et al., 2010; Li et al., 2011; Lutz et al., 2015; Pillai et al., 2009; Wilk et al., 2009 & 2012; Young et al., 2010; Zhang, Summah, Zhu, & Qu, 2012; Zhu et al., 2014). From Table 4, we can see that HCMANOVA identifies 13 SNPs which are same as MANOVA; HCMultiPhen identifies 13 SNPs which are one less than MultiPhen; and HCTATES identifies 10 SNPs which are one more than TATES. The results of the real data analysis are consistent with our simulation results, that is, the existing methods using HCM are more powerful than or comparable with those without using HCM.

4 | DISCUSSION

In this paper, we developed a HCM for joint analysis of multiple phenotypes in association studies. The proposed method is a dimension reduction technique by using a representative phenotype for each cluster of phenotypes. Applying HCM, we used existing methods to test the association between genetic variants and the representative phenotypes rather than the individual phenotypes.

HCM has several important advantages over other dimension reduction techniques. First, it can produce a dendrogram of the phenotypes, which may provide more information on the

structure of phenotypes. Second, it is computationally fast and easy to implement. Third, it has the distinct advantage that any valid measure of distance can be used in the hierarchical clustering procedure. In fact, HCM does not require phenotypes themselves, it only requires a dissimilarity matrix of phenotypes. This dissimilarity matrix of phenotypes can be estimated from the values of summary statistics using independent SNPs in a GWAS (Zhu et al., 2015b). Last, any linear combination of the phenotypes within each cluster can represent the cluster reasonably well when the cluster consists of highly positively correlated phenotypes (Bühlmann et al., 2013; Shah & Samworth, 2013).

We used extensive simulation studies as well as real data application to compare the performance of MANOVA, MultiPhen, and TATES with using HCM with that of without using HCM. Our simulation results showed that the three methods using HCM have correct type I error rates and are more powerful than or comparable with those without using HCM under a variety of simulation scenarios. Additionally, the real data analysis results demonstrated that HCM has great potential in GWAS with multiple phenotypes such as COPD. We also compared the proposed method with a popular dimension reduction method, PCA of phenotypes. Our simulation results showed that the three methods using HCM are more powerful than those using PCs of phenotypes.

In this study, we use the average phenotype in each cluster as the representative phenotype of the cluster. We can also use the first PC of the phenotypes in each cluster as a representative phenotype of the cluster. However, our simulation studies (Figure S4 and Figure S5) show that using the average as the representative has very similar performance as using the first PC as the representative. As we pointed out in the introduction section, any linear combination of the phenotypes within one cluster can represent the cluster reasonably well when the cluster consists of highly positively correlated phenotypes. The proposed method is more suitable for quantitative phenotypes. After scaling the phenotypes, the proposed method can be applied to binary or mixed traits. However, the performance of this approach for applying to binary or mixed traits needs further investigation.

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.

This research used data generated by the COPDGene study, which was supported by National Institutes of Health (NIH) grants U01HL089856 and U01HL089897. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health. The COPDGene project is also supported by the COPD Foundation through contributions made by an Industry Advisory Board comprised of Pfizer, AstraZeneca, Boehringer Ingelheim, Novartis, and Sunovion.

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. *The American Journal of Human Genetics*. 2014; 94:662–676. [PubMed: 24746957]
- Bates DM, DebRoy S. Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*. 2004; 91:1–17.
- Brehm JM, Hagiwara K, Tesfaigzi Y, Bruse S, Mariani TJ, Bhattacharya S, Cho MH. Identification of FGF7 as a novel susceptibility locus for chronic obstructive pulmonary disease. *Thorax*. 2011; 66:1085–1090. [PubMed: 21921092]
- Bühlmann P, Rütimann P, van de Geer S, Zhang CH. Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*. 2013; 143:1835–1858.
- Casale FP, Rakitsch B, Lippert C, Stegle O. Efficient set tests for the genetic analysis of correlated traits. *Nature Methods*. 2015; 12:755–758. [PubMed: 26076425]
- Chen LS, Hung RJ, Baker T, Horton A, Culverhouse R, Saccone N, Horsman J. CHRNA5 risk variant predicts delayed smoking cessation and earlier lung cancer diagnosis—a meta-analysis. *Journal of the National Cancer Institute*. 2015; 107:djv100. [PubMed: 25873736]
- Cho MH, Boutaoui N, Klanderman BJ, Sylvia JS, Ziniti JP, Hersh CP, Lange C. Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nature Genetics*. 2010; 42:200–202. [PubMed: 20173748]
- Cho MH, McDonald MLN, Zhou X, Mattheisen M, Castaldi PJ, Hersh CP, Lange C. Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *The Lancet Respiratory Medicine*. 2014; 2:214–225. [PubMed: 24621683]
- Cho MH, Castaldi PJ, Hersh CP, Hobbs BD, Barr RG, Tal-Singer R, Coxson HO. A genome-wide association study of emphysema and airway quantitative imaging phenotypes. *American Journal of Respiratory and Critical Care Medicine*. 2015; 192:559–569. [PubMed: 26030696]
- Chu JH, Hersh CP, Castaldi PJ, Cho MH, Raby BA, Laird N, Silverman EK. Analyzing networks of phenotypes in complex diseases: methodology and applications in COPD. *BMC Systems Biology*. 2014; 8:78. [PubMed: 24964944]
- 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. *Psychological Bulletin*. 1994; 115:465.
- Cui K, Ge X, Ma H. Four SNPs in the CHRNA3/5 alpha-neuronal nicotinic acetylcholine receptor subunit locus are associated with COPD risk based on meta-analyses. *PloS One*. 2014; 9:e102324. [PubMed: 25051068]
- Deng Y, Pan W. Conditional analysis of multiple quantitative traits based on marginal GWAS summary statistics. *Genetic Epidemiology*. 2017
- Du Y, Xue Y, Xiao W. Association of IREB2 gene rs2568494 polymorphism with risk of chronic obstructive pulmonary disease: a meta-analysis. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*. 2016; 22:177. [PubMed: 26775557]
- 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]
- Hancock DB, Eijgelsheim M, Wilk JB, Gharib SA, Loehr LR, Marcianti KD, Schabath MB. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nature Genetics*. 2010; 42:45–52. [PubMed: 20010835]
- Hancock DB, Reginsson GW, Gaddis NC, Chen X, Saccone NL, Lutz SM, Stacey SN. Genome-wide meta-analysis reveals common splice site acceptor variant in CHRNA4 associated with nicotine dependence. *Translational Psychiatry*. 2015; 5:e651. [PubMed: 26440539]
- Klei L, Luca D, Devlin B, Roeder K. Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic Epidemiology*. 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. *Nature Genetics*. 2012; 44:1066–1071. [PubMed: 22902788]

- Kwak IY, Pan W. Gene- and pathway-based association tests for multiple traits with GWAS summary statistics. *Bioinformatics*. 2016; 33:64–71. [PubMed: 27592708]
- Lange C, Van Steen K, Andrew T, Lyon H, DeMeo DL, Raby B, Laird NM. A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. *Statistical Applications in Genetics and Molecular Biology*. 2004; 3:1–27.
- Li X, Howard TD, Moore WC, Ampleford EJ, Li H, Busse WW, Fitzpatrick AM. Importance of hedgehog interacting protein and other lung function genes in asthma. *Journal of Allergy and Clinical Immunology*. 2011; 127:1457–1465. [PubMed: 21397937]
- Liang X, Wang Z, Sha Q, Zhang S. An Adaptive Fisher's Combination Method for Joint Analysis of Multiple Phenotypes in Association Studies. *Scientific Reports*. 2016; 6
- Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–22.
- Lutz SM, Cho MH, Young K, Hersh CP, Castaldi PJ, McDonald ML, Foreman M. A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC Genetics*. 2015; 16:138. [PubMed: 26634245]
- Lutz SM, Fingerlin TE, Hokanson JE, Lange C. A general approach to testing for pleiotropy with rare and common variants. *Genetic Epidemiology*. 2017; 41:163–170. [PubMed: 27900789]
- Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*. 2007; 39:906. [PubMed: 17572673]
- Nazir SA, Erbland ML. Chronic obstructive pulmonary disease. *Drugs & Aging*. 2009; 26:813–831. [PubMed: 19761275]
- 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]
- Ott J, Rabinowitz D. A principal-components approach based on heritability for combining phenotype information. *Human Heredity*. 1999; 49:106–111. [PubMed: 10077732]
- Pillai SG, Ge D, Zhu G, Kong X, Shianna KV, Need AC, Ruppert A. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genetics*. 2009; 5:e1000421. [PubMed: 19300482]
- Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, Crapo JD. Genetic epidemiology of COPD (COPDGene) study design. COPD: Journal of Chronic Obstructive Pulmonary Disease. 2011; 7:32–43.
- Schaid DJ, Tong X, Larrabee B, Kennedy RB, Poland GA, Sinnwell JP. Statistical methods for testing genetic pleiotropy. *Genetics*. 2016; 204:483–497. [PubMed: 27527515]
- Shah RD, Samworth RJ. Discussion of 'Correlated variables in regression: clustering and sparse estimation' by Peter Bühlmann, Philipp Rütimann, Sara van de Geer and Cun-Hui Zhang. *Journal of Statistical Planning and Inference*. 2013; 143:1866–1868.
- Siafakas NM, Vermeire P, Pride NA, Paoletti P, Gibson J, Howard P, Postma DS. Optimal assessment and management of chronic obstructive pulmonary disease (COPD). The European Respiratory Society Task Force. *European Respiratory Journal*. 1995; 8:1398–1420. [PubMed: 7489808]
- Tang CS, Ferreira MA. A gene-based test of association using canonical correlation analysis. *Bioinformatics*. 2012; 28:845–850. [PubMed: 22296789]
- Therneau TM, Grambsch PM, Pankratz VS. Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*. 2003; 12:156–175.
- Van der Sluis S, Posthuma D, Dolan CV. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genetics*. 2013; 9:e1003235. [PubMed: 23359524]
- Wang Z, Sha Q, Zhang S. Joint analysis of multiple traits using “optimal” maximum heritability test. *PloS One*. 2016; 11:e0150975. [PubMed: 26950849]
- Wilk JB, Chen TH, Gottlieb DJ, Walter RE, Nagle MW, Brandler BJ, O'Connor GT. A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genetics*. 2009; 5:e1000429. [PubMed: 19300500]

- Wilk JB, Shrine NR, Loehr LR, Zhao JH, Manichaikul A, Lopez LM, Loth DW. Genome-wide association studies identify CHRNA5/3 and HTR4 in the development of airflow obstruction. *American Journal of Respiratory and Critical Care Medicine*. 2012; 186:622–632. [PubMed: 22837378]
- Yan T, Li Q, Li Y, Li Z, Zheng G. Genetic association with multiple traits in the presence of population stratification. *Genetic Epidemiology*. 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 Bioinformatics*. 2016; 17:19. [PubMed: 26729364]
- Yang JJ, Williams LK, Buu A. Identifying Pleiotropic Genes in Genome-Wide Association Studies for Multivariate Phenotypes with Mixed Measurement Scales. *PloS One*. 2017; 12:e0169893. [PubMed: 28081206]
- Yang Q, Wang Y. Methods for analyzing multivariate phenotypes in genetic association studies. *Journal of Probability and Statistics*. 2012; 2012
- Young RP, Whittington CF, Hopkins RJ, Hay BA, Epton MJ, Black PN, Gamble GD. Chromosome 4q31 locus in COPD is also associated with lung cancer. *European Respiratory Journal*. 2010; 36:1375–1382. [PubMed: 21119205]
- Zhang J, Summah H, Zhu YG, Qu JM. Nicotinic acetylcholine receptor variants associated with susceptibility to chronic obstructive pulmonary disease: a meta-analysis. *Respiratory Research*. 2012; 12:158.
- 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. *Human Heredity*. 2015; 79:93–104. [PubMed: 26111731]
- Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*. 2014; 11:407–409. [PubMed: 24531419]
- Zhu AZ, Zhou Q, Cox LS, David SP, Ahluwalia JS, Benowitz NL, Tyndale RF. Association of CHRNA5-A3-B4 SNP rs2036527 With Smoking Cessation Therapy Response in African-American Smokers. *Clinical Pharmacology & Therapeutics*. 2014; 96:256–265. [PubMed: 24733007]
- Zhu H, Zhang S, Sha Q. Power comparisons of methods for joint association analysis of multiple phenotypes. *Human Heredity*. 2015a; 80:144–152. [PubMed: 27344597]
- Zhu X, Feng T, Tayo BO, Liang J, Young JH, Franceschini N, Chen W. Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *The American Journal of Human Genetics*. 2015b; 96:21–36. [PubMed: 25500260]

Figure 1. Power comparisons of the six tests (HCMANOVA, MANOVA, HCMultiPhen, MultiPhen, HCTATES, and TATES) for the power as a function of effect size β for 20 quantitative phenotypes. MAF is 0.3. The sample size is 5000. The number of replication is 1000. The within-factor correlation is 0.5 ($c^2 = 0.5$) and the between-factor correlation is 0.1 ($\rho c^2 = 0.1$). The powers are evaluated at 5% significance level.

Figure 2.

Power comparisons of the six tests (HCMANOVA, MANOVA, HCMultiPhen, MultiPhen, HCTATES, and TATES) for the power as a function of effect size β for 40 quantitative phenotypes. MAF is 0.3. The sample size is 5000. The number of replication is 1000. The within-factor correlation is 0.5 ($c^2 = 0.5$) and the between-factor correlation is 0.1 ($\rho c^2 = 0.1$). The powers are evaluated at 5% significance level.

Figure 3.
The dendrogram of the seven phenotypes in the COPDGene study.

Table 1

The estimated type I error rates of HCMANOVA. MAF is 0.3. K is the number of phenotypes. α is the significance level. 10,000 replicates are used in the simulations. The type I error rates in bold indicate the values out of the bounds of the 95% CIs.

K	α	Sample Size	Model			
			1	2	3	4
20	0.050	2000	0.0491	0.0479	0.0491	0.0512
		5000	0.0488	0.0507	0.0482	0.0491
	0.010	2000	0.0092	0.0084	0.0102	0.0108
		5000	0.0113	0.0107	0.0077	0.0089
	0.001	2000	0.0015	0.0002	0.0014	0.0009
		5000	0.0010	0.0013	0.0006	0.0009
40	0.050	2000	0.0482	0.0473	0.0495	0.0505
		5000	0.0509	0.0487	0.0500	0.0509
	0.010	2000	0.0091	0.0080	0.0098	0.0104
		5000	0.0120	0.0104	0.0100	0.0094
	0.001	2000	0.0013	0.0005	0.0009	0.0008
		5000	0.0012	0.0009	0.0005	0.0005

Table 2

The estimated type I error rates of HCMultiPhen. MAF is 0.3. K is the number of phenotypes. α is the significance level. 10,000 replicates are used in the simulations. The type I error rates in bold indicate the values out of the bounds of the 95% CIs.

K	α	Sample Size	Model			
			1	2	3	4
20	0.050	2000	0.0519	0.0474	0.0515	0.0501
		5000	0.0482	0.0479	0.0484	0.0513
	0.010	2000	0.0100	0.0082	0.0109	0.0104
		5000	0.0112	0.0111	0.0085	0.0100
	0.001	2000	0.0018	0.0006	0.0011	0.0008
		5000	0.0013	0.0013	0.0008	0.0006
40	0.050	2000	0.0513	0.0464	0.0512	0.0502
		5000	0.0539	0.0484	0.0496	0.0490
	0.010	2000	0.0112	0.0078	0.0104	0.0091
		5000	0.0127	0.0111	0.0096	0.0102
	0.001	2000	0.0013	0.0007	0.0009	0.0010
		5000	0.0013	0.0012	0.0002	0.0004

Table 3

The estimated type I error rates of HCTATES. MAF is 0.3. K is the number of phenotypes. α is the significance level. 10,000 replicates are used in the simulations. The type I error rates in bold indicate the values out of the bounds of the 95% CIs.

K	α	Sample Size	Model			
			1	2	3	4
20	0.050	2000	0.0445	0.0469	0.0509	0.0491
		5000	0.0452	0.0516	0.0498	0.0478
	0.010	2000	0.0105	0.0088	0.0119	0.0091
		5000	0.0096	0.0102	0.0087	0.0088
	0.001	2000	0.0016	0.0006	0.0013	0.001
		5000	0.0007	0.0012	0.0007	0.0013
40	0.050	2000	0.0409	0.0464	0.0519	0.0495
		5000	0.0413	0.0486	0.0488	0.0499
	0.010	2000	0.0088	0.0098	0.0095	0.0096
		5000	0.0097	0.011	0.0096	0.0094
	0.001	2000	0.0008	0.0005	0.001	0.0009
		5000	0.0012	0.0009	0.0013	0.0008

Significant SNPs and the corresponding p-values in the analysis of COPDGene. The p-values of six tests are evaluated using asymptotic distributions. The bold p-values indicate the p-values $> 5 \times 10^{-8}$.

Table 4

Chr	Position	Variant identifier	MANOVA	HCMANOVA	MultiPhen	HCMultiPhen	TATES	HCTATES
4	145431497	rs1512282	1.69×10^{-9}	1.61×10^{-10}	1.03×10^{-9}	9.68×10^{-11}	5.77×10^{-9}	4.76×10^{-9}
4	145434744	rs1032297	6.52×10^{-14}	5.16×10^{-15}	7.69×10^{-14}	6.67×10^{-15}	6.22×10^{-13}	5.52×10^{-13}
4	145474473	rs1489759	1.11×10^{-16}	7.46×10^{-18}	1.22×10^{-16}	9.32×10^{-18}	2.52×10^{-16}	0
4	145485738	rs1980057	6.68×10^{-17}	4.52×10^{-18}	8.14×10^{-17}	6.30×10^{-18}	9.35×10^{-17}	0
4	145485915	rs7655625	7.12×10^{-17}	4.98×10^{-18}	9.13×10^{-17}	7.33×10^{-18}	1.64×10^{-16}	0
15	78882925	rs16969968	1.32×10^{-11}	1.98×10^{-10}	7.84×10^{-12}	1.46×10^{-10}	2.98×10^{-8}	2.44×10^{-8}
15	78894339	rs1051730	1.41×10^{-11}	1.51×10^{-10}	8.16×10^{-12}	1.17×10^{-10}	2.63×10^{-8}	2.15×10^{-8}
15	78898723	rs12914385	1.76×10^{-12}	4.33×10^{-12}	1.48×10^{-12}	4.84×10^{-12}	5.14×10^{-10}	4.31×10^{-10}
15	78911181	rs8040868	2.74×10^{-12}	1.74×10^{-11}	2.59×10^{-12}	2.08×10^{-11}	2.40×10^{-9}	1.99×10^{-9}
15	78878541	rs951266	1.77×10^{-11}	3.04×10^{-10}	1.02×10^{-11}	2.24×10^{-10}	5.17×10^{-8}	4.21×10^{-8}
15	78806023	rs8034191	2.14×10^{-10}	2.95×10^{-9}	7.74×10^{-11}	1.70×10^{-9}	1.02×10^{-7}	8.24×10^{-8}
15	78851615	rs2036527	3.99×10^{-10}	6.62×10^{-9}	1.77×10^{-10}	4.05×10^{-9}	1.56×10^{-7}	1.26×10^{-7}
15	78826180	rs931794	2.35×10^{-10}	1.68×10^{-8}	9.09×10^{-11}	1.19×10^{-8}	1.18×10^{-7}	9.58×10^{-8}
15	78740964	rs2568494	1.05×10^{-7}	6.35×10^{-6}	4.23×10^{-8}	5.54×10^{-6}	2.88×10^{-5}	7.44×10^{-5}