A computationally efficient clustering linear combination approach to jointly analyze multiple phenotypes for GWAS

Meida Wang¹, Shuanglin Zhang¹, Qiuying Sha^{1*}

¹ Mathematical Sciences, Michigan Technological University, Houghton, MI, USA.

* Corresponding author

E-mail: qsha@mtu.edu

Abstract

There has been an increasing interest in joint analysis of multiple phenotypes in genome-wide association studies (GWAS) because jointly analyzing multiple phenotypes may increase statistical power to detect genetic variants associated with complex diseases or traits. Recently, many statistical methods have been developed for joint analysis of multiple phenotypes in genetic association studies, including the Clustering Linear Combination (CLC) method. The CLC method works particularly well with phenotypes that have natural groupings, but due to the unknown number of clusters for a given data, the final test statistic of CLC method is the minimum p-value among all p-values of the CLC test statistics obtained from each possible number of clusters. Therefore, a simulation procedure must be used to evaluate the p-value of the final test statistic. This makes the CLC method computationally demanding. We develop a new method called computationally efficient CLC (ceCLC) to test the association between multiple phenotypes and a genetic variant. Instead of using the minimum p-value as the test statistic in the CLC method, ceCLC uses the Cauchy combination test to combine all p-values of the CLC test statistics obtained from each possible number of clusters. The test statistic of ceCLC approximately follows a standard Cauchy distribution, so the p-value can be obtained from the cumulative density function without the need for the simulation procedure. Through extensive simulation studies and application on the COPDGene data, the results demonstrate that the type I error rates of ceCLC are effectively controlled in different simulation settings and ceCLC either outperforms all other methods or has statistical power that is very close to the most powerful method with which it has been compared.

Introduction

Genome-wide association study (GWAS) has successfully identified a large number of genetic variants that are associated with human complex diseases or phenotypes [1-4]. Among these results, a phenomenon in which a genetic variant affects multiple phenotypes often occurs [5], which is significant evidence to show that pleiotropic effects on human complex diseases are universal [6-9]. Moreover, several disease-related phenotypes are usually measured simultaneously as a disorder or risk factors of a complex disease in GWAS. Therefore, considering the correlated structure of multiple phenotypes in genetic association studies can aggregate multiple effects and increase the statistical power [10-15].

At present, a variety of approaches that focus on jointly analyzing multiple phenotypes have been proposed. These statistical methods can be roughly divided into three categories, including approaches based on regression models [16-19], combining the univariate analysis results [20-23], and variable reduction techniques [24-27]. For example, MultiPhen [19] performs an ordinal regression model, which uses an inverted model whereby the phenotypes are the predictor variables and the genotype is the dependent variable [28-29]. In terms of the second category, combining the univariate test statistics or integrating the p-values of univariate tests are two basic methods. For instance, the O'Brien [20-21] method constructs a test statistic for pleiotropic effect by combining univariate test statistics of multiple phenotypes; the Trait-based Association

Test that uses the Extended Simes procedure (TATES) [23] integrates the p-values from univariate tests to obtain an overall trait-based p-value. In addition, principal components analysis of phenotypes (PCP) [24], principal component of heritability (PCH) [25-26], and canonical correlation analysis (CCA) [27] are three variable reduction methods in the third category.

In practice, multiple phenotypes considered may be in different clusters, but most methods for detecting the association between multiple phenotypes and genetic variants either treat all phenotypes as a group or treat each phenotype as one group and combine the results of univariate analysis. Unlike these methods, the clustering linear combination (CLC) method [30] works particularly well with phenotypes that have natural clusters. In the CLC method, individual statistics from the association tests for each phenotype are clustered into positively correlated clusters using the hierarchical clustering method, then the CLC test statistic is used to combine the individual test statistics linearly within each cluster and combine the between-cluster terms in a quadratic form. It was theoretically proved that if the individual statistics can be clustered correctly, the CLC test statistic is the most powerful test among all tests with certain quadratic forms [30]. Due to the unknown number of clusters for a given data, the final test statistic of CLC method is the minimum p-value among all p-values of the CLC test statistics obtained from each possible number of clusters. Therefore, a simulation procedure must be used to evaluate the p-value of the final test statistic because it does not have an asymptotic distribution, and that makes the CLC method computationally demanding. If we can construct a test statistic with an approximate distribution, the computational efficiency will be greatly improved. In this paper, based on the Aggregated Cauchy Association Test (ACAT)

method [31], we develop a new method named computationally efficient CLC (ceCLC). In ceCLC, the p-values of the CLC test statistics with L clusters are transformed to follow a standard Cauchy distribution, then the transformed p-values are combined linearly with equal treatment to obtain the ceCLC test statistic. This test statistic of ceCLC has an approximately standard Cauchy distribution even though there is a correlated structure between multiple phenotypes [32], so the p-value of the ceCLC test statistic can be calculated based on the cumulative density function of standard Cauchy distribution. We perform extensive simulation studies and apply ceCLC to the COPDGene real dataset. The results show that the ceCLC method has correct type I error rates and either outperforms all other methods or has statistical power that is very close to the most powerful method with which it has been compared.

Materials and Methods

Assume we consider N unrelated individuals with K correlated phenotypes, which can be quantitative or qualitative, and each individual has been genotyped at a genetic variant of interest. Let $Y_i = (Y_{i1}, \dots, Y_{iK})^T$ represent K correlated phenotypes for the i th individual (1 for cases and 0 for controls for a qualitative trait) with $i = 1, 2, \dots, N$. Let G_i denote the genotypic score for the i th individual at the variant, where $G_i \in \{0, 1, 2\}$ corresponds to the number of minor alleles. We suppose that there are no covariates. If there are p covariates $z_{i1}, \dots, z_{ip}$, we adjust both genotypes and phenotypes for the covariates [33-34] using linear models $G_i = \alpha_0 + \alpha_1 z_{i1} + \dots + \alpha_p z_{ip} + \varepsilon_i$ and $Y_{ik} = \alpha_{0k} +$

$\alpha_{1k}Z_{i1} + \dots + \alpha_{pk}Z_{ip} + \tau_{ik}$, and use the residuals of the respective linear models to replace the original genotypes and phenotypes.

For each phenotype, we consider the following generalized linear model [35]:

$$g(E(Y_{ik} | G_i)) = \beta_{0k} + \beta_{1k}G_i,$$

where β_{1k} is the genetic effect of the variant on the k th phenotype and $g(\cdot)$ is a monotone “link” function. Two types of generalized linear model are commonly used: 1) linear model with an identity link for quantitative phenotypes and 2) logistic regression model with a logit link for qualitative phenotypes. We first conduct an univariate test to test $H_0: \beta_{1k} = 0$ for each phenotype, $k = 1, 2, \dots, K$, using the score test statistic [36]

$$T_k = U_k / \sqrt{V_k},$$

where $U_k = \sum_{i=1}^N Y_{ik}(G_i - \bar{G})$ and $V_k = \frac{1}{N} \sum_{i=1}^N (Y_{ik} - \bar{Y}_k)^2 \sum_{i=1}^N (G_i - \bar{G})^2$. Since the test statistic T_k has an approximate normal distribution with mean $\mu_k = E(T_k)$ and variance 1, we can assume that $T = (T_1, \dots, T_K)^T$ approximately follows a multivariate normal distribution with mean vector $\mu = (\mu_1, \dots, \mu_K)^T$ and covariance matrix Σ . Our objective is to test the association between multiple phenotypes and a genetic variant, so the null hypothesis is $H_0: \beta_{11} = \dots = \beta_{1K} = 0$. Sha et al. [30] showed that under the null hypothesis, Σ converges to $P(Y)$ almost surely, where $P(Y)$ is the correlation matrix of $Y = (Y_1, \dots, Y_K)^T$. Therefore, we can use the sample correlation matrix of Y , $P^s(Y)$, to estimate Σ [30].

Based on the CLC [30] and ACAT methods [31], we propose a computational efficient CLC (ceCLC) method in this paper. Same as the CLC method [30], we use the

hierarchical clustering method with similarity matrix $\hat{\Sigma} = P^s(Y)$ and dissimilarity matrix $1 - P^s(Y)$ to cluster K phenotypes. Suppose that the phenotypes are clustered into L clusters, considering $L = 1, \dots, K$, and B is a $K \times L$ matrix with the $(k, l)^{th}$ element equals 1 if the k th phenotype belongs to the l th cluster, otherwise it equals 0. The CLC test statistic [30] with L clusters is given by

$$T_{CLC}^L = (WT)^T(W\Sigma W^T)^{-1}(WT),$$

where $W = B^T\Sigma^{-1}$. T_{CLC}^L follows a χ_L^2 distribution under the null hypothesis, therefore we can obtain the p-value of T_{CLC}^L , represented by p_L , for $L = 1, \dots, K$. Since for a given data set, the number of clusters of the phenotypes is unknown, in the last step of the CLC method [30], $T_{CLC} = \min_{1 \leq L \leq K} p_L$ is used as the final test statistic. Because T_{CLC} does not have an asymptotic distribution, a simulation procedure is needed to evaluate the p-value of T_{CLC} . This makes the CLC method computationally demanding. In this paper, instead of using the minimum p-value as the test statistic in the CLC method, we use the Cauchy combination test [32] to combine all p-values of the CLC test statistics obtained from each possible number of clusters. We define the ceCLC test statistic as the linear combination of the transformed p-values over the number of K clusters, which is given by

$$T_{ceCLC} = \frac{1}{K} \sum_{L=1}^K \tan\{(0.5 - p_L)\pi\}$$

Under the null hypothesis, $\tan\{(0.5 - p_L)\pi\}$ follows the standard Cauchy distribution, for $L = 1, \dots, K$. Although there exists a correlated structure between $p_1, \dots, p_K$, Liu. et. al [32] has proved that a weighted sum of “correlated” standard Cauchy variables still has

an approximately Cauchy tail, and the influence of correlated structure on the tail is quite limited because of the heaviness of the Cauchy tail. Therefore, T_{ceCLC} can be well approximated by a standard Cauchy distribution. According to the cumulative density distribution of standard Cauchy distribution, the p-value of T_{ceCLC} can be approximated by $0.5 - \{\arctan(T_{ceCLC})/\pi\}$.

Results

Simulation Design

In our simulation studies, we generate one common variant and $K = 20$ and 40 correlated phenotypes for N individuals. Firstly, we generate the genotypes of the genetic variant according to the minor allele frequency (MAF = 0.3) under Hardy Weinberg equilibrium. Secondly, the K quantitative phenotypes are generated by the following factor model [22, 26, 28, 30]

$$Y = \lambda G + c\gamma f + \sqrt{1 - c^2} \times \varepsilon.$$

where $Y = (Y_1, \dots, Y_K)^T$, G is the genotypic score at the variant, $\lambda = (\lambda_1, \dots, \lambda_K)^T$ is the vector of genetic effect sizes on K phenotypes, c is a constant number, f is a vector of factors, and $f = (f_1, \dots, f_R)^T \sim MVN(0, \Sigma)$, where R is the number of factors, $\Sigma = (1 - \rho)I + \rho A$, all elements of matrix A equals 1, I is an identity matrix, ρ is the correlation between factors; γ is a $K \times R$ matrix, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_K)^T$ is a vector of residuals, and $\varepsilon_1, \dots, \varepsilon_K \sim \text{i.i.d. } N(0, 1)$.

According to different number of factors affected by the genotypes and different effect sizes, we consider the following four models. In each model, the within-factor correlation is c^2 and the between-factor correlation is ρc^2 . We set $c = 0.5$ and $\rho = 0.6$.

Model 1: There is only one factor and genotypes influence 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 influence one factor. That is, $R = 2$, $\lambda = (\underbrace{0, 0, \dots, 0}_{K/2}, \underbrace{\beta, \beta, \dots, \beta}_{K/2})^T$, and $\gamma = Bdiag(D_1, D_2)$, where $D_i = 1_{K/2}$ for $i = 1, 2$.

Model 3: There are five factors and genotypes influence two factors. That is, $R = 5$, $\lambda = (\beta_{11}, \dots, \beta_{1k}, \beta_{21}, \dots, \beta_{2k}, \beta_{31}, \dots, \beta_{3k}, \beta_{41}, \dots, \beta_{4k}, \beta_{51}, \dots, \beta_{5k})^T$, and $\gamma = Bdiag(D_1, D_2, D_3, D_4, D_5)$, where $D_i = 1_{K/5}$ for $i = 1, \dots, 5$, $k = K/5$, $\beta_{11} = \dots = \beta_{1k} = \beta_{21} = \dots = \beta_{2k} = \beta_{31} = \dots = \beta_{3k} = 0$, $\beta_{41} = \dots = \beta_{4k} = -\beta$ and $(\beta_{51}, \dots, \beta_{5k}) = \frac{2\beta}{k+1}(1, \dots, k)$.

Model 4: There are five factors and genotypes influence four factors. That is, $R = 5$, $\lambda = (\beta_{11}, \dots, \beta_{1k}, \beta_{21}, \dots, \beta_{2k}, \beta_{31}, \dots, \beta_{3k}, \beta_{41}, \dots, \beta_{4k}, \beta_{51}, \dots, \beta_{5k})^T$, and $\gamma = Bdiag(D_1, D_2, D_3, D_4, D_5)$, where $D_i = 1_{K/5}$ for $i = 1, \dots, 5$, $k = K/5$. $\beta_{11} = \dots = \beta_{1k} = 0$, $\beta_{21} = \dots = \beta_{2k} = \beta$, $\beta_{31} = \dots = \beta_{3k} = -\beta$, $(\beta_{41}, \dots, \beta_{4k}) = -\frac{2\beta}{k+1}(1, \dots, k)$, and $(\beta_{51}, \dots, \beta_{5k}) = \frac{2\beta}{k+1}(1, \dots, k)$.

We consider two types of multiple phenotypes. The first one is that all K phenotypes are quantitative and the second one is that half phenotypes are quantitative and the other half are qualitative. To generate a qualitative phenotype, we use a liability threshold model based on a quantitative phenotype. A qualitative phenotype is defined to be affected if the corresponding quantitative phenotype is at least one standard deviation larger (smaller) than the phenotypic mean.

In order to ensure the validity of the ceCLC method, we first evaluate the type I error rates of this method. We simulate data under the null hypothesis and consider three different sample sizes, $N = 1000, 2000$, and 3000 , under four different models. The type I error rates are evaluated by 10^5 replications and at the nominal significance levels of 0.001 and 0.0001 , respectively. To evaluate power, we simulate data under the alternative hypothesis and consider two different sample sizes, $N = 3000$ and 5000 . The powers are evaluated by 1000 replications at the nominal significance levels of 0.05 . To better demonstrate the advantages of the ceCLC method, we compare the performance of ceCLC method with other six existing association tests using multiple phenotypes: CLC [30], MANOVA [37], MultiPhen [19], TATES [23], O'Brien [20], and Omnibus.

Simulation Results

(a) Evaluation of type I error rates

Table 1 presents the type I error rates of the ceCLC method for $K = 20$ quantitative phenotypes, and the type I error rates of the other six methods (CLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus) are summarized in S1 Table. The corresponding type I error rates for the case of half quantitative traits and half qualitative phenotypes are recorded in Table 2 and S2 Table. In addition, The type I error rate of the ceCLC method for $K = 40$ are listed in S3 and S4 Tables, and the type I error rates of the other six methods for $K = 40$ are summarized in S5 and S6 Tables. For 10^5 replications, the 95% confidence intervals of Type I error rates divided by nominal significance levels of

0.001 and 0.0001 are (0.8041, 1.1959) and (0.3802, 1.6198), respectively. In the tables, we use gray color to indicate that the type I error rate is not in the 95% CI.

Table 1. The estimated type I error rates divided by the nominal significance levels of the ceCLC method for 20 quantitative phenotypes.

α	Sample	Model1	Model2	Model3	Model4
0.001	1000	0.97	0.95	0.79	0.92
	2000	1.06	1.01	1.14	1.02
	3000	1.13	1.14	1.17	1.11
0.0001	1000	0.60	1.20	0.50	1.20
	2000	0.80	0.70	1.00	1.40
	3000	0.90	1.20	1.20	1.20

From Tables 1 and 2 (S3 and S4 Tables), we can see that all the estimated type I error rates are within the corresponding 95% CIs except 0.79 in Table 1 and 0.76 in S3 Table. However, these two values are very close to the lower bound of its CI (0.8041), therefore we can conclude that the ceCLC method is a valid test. From S1-S2 and S5-S6 Tables, we observe that CLC, MANOVA, TATES, and O'Brien can control type I error rates well, but some of the type I error rates of MultiPhen are slightly inflated and some of the type I error rates of Omnibus are conservative.

Table 2. The estimated type I error rates divided by the nominal significance levels of the ceCLC method for 10 quantitative and 10 qualitative phenotypes.

α	Sample	Model1	Model2	Model3	Model4
0.001	1000	0.98	0.93	0.87	0.91
	2000	1.00	1.09	1.06	1.03

	3000	0.97	1.08	1.02	1.04
0.0001	1000	0.90	1.10	1.10	1.00
	2000	0.90	0.60	1.20	1.30
	3000	0.60	0.80	1.10	1.20

(b) Assessment of powers

Fig 1 shows the results of power comparisons for all seven tests with 20 quantitative phenotypes when the sample size is 5000. From Fig 1, we find that 1) the ceCLC and CLC methods are more powerful than other methods; 2) the O'Brien method is very sensitive to the direction of the genetic effect on the phenotypes. Its power will decrease dramatically with different directions of the genetic effect on the phenotypes (Models 3 and 4); 3) MANOVA, Omnibus, and MultiPhen have similar powers. Fig 2 shows the results of power comparisons for all seven tests with 10 quantitative and 10 qualitative phenotypes when the sample size is 5000. The general trend of Fig 2 is similar to Fig 1, but the powers of MANOVA, Omnibus, and MultiPhen are higher than those in Fig 1 for Models 3 and 4. S1 and S2 Figs present the results of power comparisons with 40 phenotypes for the sample size of 5000, and all the results of power comparisons for the sample size of 3000 are showed in S3-S6 Figs. In summary, ceCLC and CLC methods are more powerful than other methods under most scenarios ceCLC is much more computationally efficient than CLC.

Fig 1. Power comparisons of the seven tests, CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, and Omnibus, with 20 quantitative phenotypes for the sample size of 5000.

Fig 2. Power comparisons of the seven tests, CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, and Omnibus, with 10 quantitative and 10 qualitative phenotypes for the sample size of 5000.

Application to the COPDGene Study

Chronic obstructive pulmonary disease (COPD) is a common disease characterized by the presence of expiratory dyspnea due to the excessive inflammatory reaction of harmful gases and particles [38-40]. COPD causes a high mortality and has been reported to be potentially affected by genetic factors [41]. The COPDGene study is a representative multicenter research to detect hereditary factors of this disease [42]. The corresponding dataset of this study has been introduced in our previous papers [22, 30].

Same with the analysis in [22, 30], we consider seven quantitative COPD-related phenotypes, containing FEV1, Emphysema, Emphysema Distribution, Gas Trapping, Airway Wall Area, Exacerbation frequency, and Six-minute walk distance. We also consider four covariates which include BMI, Age, Pack-Years and Sex. After deleting the missing data, there are 5,430 subjects across 630,860 SNPs left for the analysis. We change the sign of six-minute walk distance and FEV1, so that the correlations between the 7 phenotypes are all positive. We apply all seven tests to the processed COPD data.

In our analysis, we choose the commonly used genome-wide significant level $\alpha = 5 \times 10^{-8}$, Table 3 presents 14 SNPs that are detected by at least one method. All of these 14 SNPs have been reported to be associated with COPD before [43-46]. From Table 3, we can see that MultiPhen detected 14 SNPs; ceCLC, CLC, MANOVA and Omnibus detected 13 SNPs; TATES detected 9 SNPs; and O'Brien only detected 5 SNPs.

Table 3. Significant SNPs and the corresponding p-values in the analysis of COPDGene study.

Chr	Position	Variant identifier	CLC	ceCLC	MANOVA	MultiPhen	TATES	O'Brien	Omnibus
4	14543149	rs1512282	10^{-9}	5.70×10^{-11}	1.69×10^{-9}	1.03×10^{-9}	5.77×10^{-9}	7.69×10^{-9}	1.82×10^{-9}
4	14543474	rs1032297	10^{-9}	2.39×10^{-15}	6.52×10^{-14}	7.69×10^{-14}	6.22×10^{-13}	3.35×10^{-10}	7.73×10^{-14}
4	14547447	rs1489759	10^{-9}	0	1.11×10^{-16}	1.22×10^{-16}	2.52×10^{-16}	2.61×10^{-11}	1.11×10^{-16}
4	14548573	rs1980057	10^{-9}	0	6.68×10^{-17}	8.14×10^{-17}	9.35×10^{-17}	3.04×10^{-11}	1.11×10^{-16}
4	14548591	rs7655625	10^{-9}	0	7.12×10^{-17}	9.13×10^{-17}	1.64×10^{-16}	3.08×10^{-11}	1.11×10^{-16}
15	78882925	rs16969968	10^{-9}	4.91×10^{-11}	1.32×10^{-11}	7.84×10^{-12}	2.98×10^{-8}	9.75×10^{-6}	1.26×10^{-11}
15	78894339	rs1051730	10^{-9}	4.74×10^{-11}	1.41×10^{-11}	8.16×10^{-12}	2.63×10^{-8}	8.99×10^{-6}	1.35×10^{-11}
15	78898723	rs12914385	10^{-9}	2.57×10^{-12}	1.76×10^{-12}	1.48×10^{-12}	5.14×10^{-10}	6.12×10^{-8}	1.66×10^{-12}
15	78911181	rs8040868	10^{-9}	5.08×10^{-12}	2.74×10^{-12}	2.59×10^{-12}	2.40×10^{-9}	1.53×10^{-7}	2.50×10^{-16}
15	78878541	rs951266	10^{-9}	7.03×10^{-11}	1.77×10^{-11}	1.02×10^{-11}	5.17×10^{-8}	1.50×10^{-5}	1.69×10^{-11}
15	78806023	rs8034191	10^{-9}	8.03×10^{-10}	2.14×10^{-10}	7.74×10^{-11}	1.02×10^{-7}	2.13×10^{-5}	1.99×10^{-10}
15	78851615	rs2036527	8.33×10^{-10}	1.52×10^{-9}	3.99×10^{-10}	1.77×10^{-10}	1.56×10^{-7}	2.65×10^{-5}	3.76×10^{-10}
15	78826180	rs931794	10^{-9}	1.18×10^{-9}	2.35×10^{-10}	9.09×10^{-11}	1.18×10^{-7}	2.33×10^{-5}	2.19×10^{-10}
15	78740964	rs2568494	3.98×10^{-7}	5.02×10^{-7}	1.05×10^{-7}	4.23×10^{-8}	2.88×10^{-5}	2.38×10^{-3}	9.73×10^{-8}

The p-values of CLC are evaluated using 10^9 simulations. The p-values of ceCLC, O'Brien, Omnibus, TATES, MANOVA, and MultiPhen are evaluated using their asymptotic distributions. The graying out p-values indicate the p-values $> 5 \times 10^{-8}$.

Discussion

In the medical field, many human complex diseases are often accompanied by multiple correlated phenotypes which are usually measured simultaneously, and pleiotropy is a universal phenomenon, so jointly analyzing multiple phenotypes in genetic association studies will very likely increase the statistical power to identify genetic variants that are associated with complex diseases. In this paper, based on the existing CLC method [30] and ACAT [31] strategy, we develop the ceCLC method to test association between multiple phenotypes and a genetic variant. We perform a variety of simulation studies, as well as an application to the COPDGene study to evaluate our new method. The results suggest that the ceCLC method not only has the advantages of the CLC method

but is also computationally efficient. The test statistic of the ceCLC method can be well approximated by a standard Cauchy distribution, so the p-value can be obtained from the cumulative density function without the need for the simulation procedure. Therefore, the ceCLC method is computational efficient.

In recent phenome-wide association studies (PheWAS), a great number of phenotypes are collected, which requires powerful and efficient statistical methods. The ceCLC method developed here can be applied to PheWAS. One limitation of the ceCLC method is that it requires individual-level phenotype data and GWAS summary statistics. We know that individual-level phenotype data are often not easily accessible as a result of privacy concerns. Therefore, we are currently considering a new strategy to extend the ceCLC method so that it can be applied to GWAS summary statistics without the requirement for individual-level phenotype data.

Acknowledgements

This research used data generated by the COPDGene study (phs000179/HMB and phs000179/DS-CS-RD), 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.

References

1. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009; 461(7265):747-53.
2. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *The American Journal of Human Genetics*. 2012; 90(1):7-24.
3. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*. 2014; 42(D1):D1001-6.
4. Lutz SM, Fingerlin TE, Hokanson JE, Lange C. A general approach to testing for pleiotropy with rare and common variants. *Genetic epidemiology*. 2017; 41(2):163-70.
5. 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(1):e0169893.
6. Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, et al. Abundant pleiotropy in human complex diseases and traits. *The American Journal of Human Genetics*. 2011; 89(5):607-18.
7. Gratten J, Visscher PM. Genetic pleiotropy in complex traits and diseases: implications for

325 genomic medicine. *Genome medicine*. 2016; 8(1):1-3.

326 8. Wang Z, Wang X, Sha Q, Zhang S. Joint analysis of multiple traits in rare variant association
327 studies. *Annals of human genetics*. 2016; 80(3):162-71.

328 9. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits:
329 challenges and strategies. *Nature Reviews Genetics*. 2013; 14(7):483-95.

330 10. Schifano ED, Li L, Christiani DC, Lin X. Genome-wide association analysis for multiple
331 continuous secondary phenotypes. *The American Journal of Human Genetics*. 2013;
332 92(5):744-59.

333 11. Deng Y, Pan W. Conditional analysis of multiple quantitative traits based on marginal
334 GWAS summary statistics. *Genetic epidemiology*. 2017; 41(5):427-36.

335 12. Liang X, Sha Q, Rho Y, Zhang S. A hierarchical clustering method for dimension reduction
336 in joint analysis of multiple phenotypes. *Genetic epidemiology*. 2018; 42(4):344-53.

337 13. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide
338 association studies. *Nature methods*. 2014; 11(4):407-9.

339 14. Jiang C, Zeng ZB. Multiple trait analysis of genetic mapping for quantitative trait loci.
340 *Genetics*. 1995; 140(3):1111-27.

341 15. Stephens M. A unified framework for association analysis with multiple related phenotypes.
342 *PloS one*. 2013; 8(7):e65245.

343 16. Bates DM, DebRoy S. Linear mixed models and penalized least squares. Journal of
344 Multivariate Analysis. 2004; 91(1):1-7.

345 17. Yan T, Li Q, Li Y, Li Z, Zheng G. Genetic association with multiple traits in the presence of
346 population stratification. Genetic epidemiology. 2013; 37(6):571-80.

347 18. Zhang Y, Xu Z, Shen X, Pan W, Alzheimer's Disease Neuroimaging Initiative. Testing for
348 association with multiple traits in generalized estimation equations, with application to
349 neuroimaging data. NeuroImage. 2014; 96:309-25.

350 19. O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FC, Elliott P, Jarvelin MR, et al. MultiPhen:
351 joint model of multiple phenotypes can increase discovery in GWAS. PloS one. 2012;
352 7(5):e34861.

353 20. O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics.
354 1984:1079-87.

355 21. Yang Q, Wu H, Guo CY, Fox CS. Analyze multivariate phenotypes in genetic association
356 studies by combining univariate association tests. Genetic epidemiology. 2010; 34(5):444-54.

357 22. Liang X, Wang Z, Sha Q, Zhang S. An adaptive Fisher's combination method for joint
358 analysis of multiple phenotypes in association studies. Scientific reports. 2016; 6(1):1-0.

359 23. Van der Sluis S, Posthuma D, Dolan CV. TATES: efficient multivariate genotype-phenotype
360 analysis for genome-wide association studies. PLoS genetics. 2013; 9(1):e1003235.

- 361 24. Aschard H, Vilhjálmsdóttir BJ, Greliche N, Morange PE, Trégouët DA, Kraft P. Maximizing
362 the power of principal-component analysis of correlated phenotypes in genome-wide
363 association studies. *The American Journal of Human Genetics*. 2014; 94(5):662-76.
- 364 25. Klei L, Luca D, Devlin B, Roeder K. Pleiotropy and principal components of heritability
365 combine to increase power for association analysis. *Genetic Epidemiology: The Official*
366 *Publication of the International Genetic Epidemiology Society*. 2008; 32(1):9-19.
- 367 26. Wang Z, Sha Q, Zhang S. Joint analysis of multiple traits using "optimal" maximum
368 heritability test. *PloS one*. 2016; 11(3):e0150975.
- 369 27. Tang CS, Ferreira MA. A gene-based test of association using canonical correlation analysis.
370 *Bioinformatics*. 2012; 28(6):845-50.
- 371 28. Zhu H, Zhang S, Sha Q. A novel method to test associations between a weighted
372 combination of phenotypes and genetic variants. *PloS one*. 2018; 13(1):e0190788.
- 373 29. Chung J, Jun GR, Dupuis J, Farrer LA. Comparison of methods for multivariate gene-based
374 association tests for complex diseases using common variants. *European Journal of Human*
375 *Genetics*. 2019; 27(5):811-23.
- 376 30. Sha Q, Wang Z, Zhang X, Zhang S. A clustering linear combination approach to jointly
377 analyze multiple phenotypes for GWAS. *Bioinformatics*. 2019; 35(8):1373-9.
- 378 31. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: a fast and powerful p value

combination method for rare-variant analysis in sequencing studies. The American Journal of Human Genetics. 2019; 104(3):410-21.

32. Liu Y, Xie J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. Journal of the American Statistical Association. 2020; 115(529):393-402.

33. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006; 38(8):904-9.

34. Sha Q, Wang X, Wang X, Zhang S. Detecting association of rare and common variants by testing an optimally weighted combination of variants. Genetic epidemiology. 2012; 36(6):561-71.

35. Nelder JA, Wedderburn RW. Generalized linear models. Journal of the Royal Statistical Society: Series A (General). 1972; 135(3):370-84.

36. Sha Q, Zhang Z, Zhang S. Joint analysis for genome-wide association studies in family-based designs. PloS One. 2011; 6(7):e21957.

37. 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(3):465.

38. Hogg JC, Chu F, Utokaparch S, Woods R, Elliott WM, Buzatu L, et al. The nature of small-airway obstruction in chronic obstructive pulmonary disease. *New England Journal of Medicine*. 2004; 350(26):2645-53.
39. Barnes PJ. Chronic obstructive pulmonary disease: effects beyond the lungs. *PLoS medicine*. 2010; 7(3):e1000220.
40. Agusti AG, Noguera A, Saucedo J, Sala E, Pons J, Busquets X. Systemic effects of chronic obstructive pulmonary disease. *European Respiratory Journal*. 2003; 21(2):347-60.
41. Sandford AJ, Weir TD, Pare PD. Genetic risk factors for chronic obstructive pulmonary disease. *European Respiratory Journal*. 1997; 10(6):1380-91.
42. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, et al. Genetic epidemiology of COPD (COPDGene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*. 2011; 7(1):32-43.
43. Cho MH, Boutaoui N, Klanderman BJ, Sylvia JS, Ziniti JP, et al. Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nature genetics*. 2010; 42(3):200-2.
44. Young RP, Whittington CF, Hopkins RJ, Hay BA, Epton MJ, Black PN, et al. Chromosome 4q31 locus in COPD is also associated with lung cancer. *European Respiratory Journal*. 2010; 36(6):1375-82.
45. Wilk JB, Shrine NR, Loehr LR, Zhao JH, Manichaikul A, Lopez LM, et al. 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(7):622-32.

46. 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. 2011; 12(1):1-9.

Supporting information

S1 Table. The estimated type I error rates divided by nominal significance levels of the other six methods, CLC, MANOVA, MultiPhen, TATES, O'Brien, and Omnibus, for 20 quantitative phenotypes.

S2 Table. The estimated type I error rates divided by nominal significance levels of the other six methods, CLC, MANOVA, MultiPhen, TATES, O'Brien, and Omnibus for 10 quantitative and 10 qualitative phenotypes.

S3 Table. The estimated type I error rates divided by the nominal significance levels of the ceCLC method for 40 quantitative phenotypes.

S4 Table. The estimated type I error rates divided by the nominal significance levels of the ceCLC method for 20 quantitative and 20 qualitative phenotypes.

S5 Table. The estimated type I error rates divided by nominal significance levels of the other six methods, CLC, MANOVA, MultiPhen, TATES, O'Brien, and Omnibus, for 40 quantitative phenotypes.

S6 Table. The estimated type I error rates divided by nominal significance levels of the other six methods, CLC, MANOVA, MultiPhen, TATES, O'Brien, and Omnibus for 20 quantitative and 20 qualitative phenotypes.

S1 Fig. Power comparisons of the seven tests, CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, and Omnibus, with 40 quantitative phenotypes for the sample size of 5000.

S2 Fig. Power comparisons of the seven tests, CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, and Omnibus, with 20 quantitative and 20 qualitative phenotypes for the sample size of 5000.

S3 Fig. Power comparisons of the seven tests, CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, and Omnibus, with 20 quantitative phenotypes for the sample size of 3000.

S4 Fig. Power comparisons of the seven tests, CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, and Omnibus, with 10 quantitative and 10 qualitative phenotypes for the sample size of 3000.

S5 Fig. Power comparisons of the seven tests, CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, and Omnibus, with 40 quantitative phenotypes for the sample size of 3000.

S6 Fig. Power comparisons of the seven tests, CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, and Omnibus, with 20 quantitative and 20 qualitative phenotypes for the sample size of 3000.

Figure 1

Figure 2