PLOS ONE

RESEARCH ARTICLE

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

Meida Wang, Shuanglin Zhang, Qiuying Sha*

Mathematical Sciences, Michigan Technological University, Houghton, MI, United States of America

* qsha@mtu.edu

Citation: Wang M, Zhang S, Sha Q (2022) A computationally efficient clustering linear combination approach to jointly analyze multiple phenotypes for GWAS. PLoS ONE 17(4): e0260911. https://doi.org/10.1371/journal.pone.0260911

Editor: Zhaozhong Zhu, Massachusetts General Hospital/Harvard Medical School, UNITED STATES

Received: November 17, 2021

Accepted: April 13, 2022

Published: April 28, 2022

Data Availability Statement: The COPDGene data upon which these findings are based are available through the dbGaP study page for COPDGene: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000179.v3.p2. There is a link from this page (Authorized Access Section) to dbGaP's controlled access system that allows someone to request the data. The accession numbers for this data are phs000179/HMB and phs000179/DS-CS-RD.

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 needs to 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

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

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. Furthermore, with more and more GWAS summary statistics from univariate phenotype analysis in the traditional GWAS being publicly available, many approaches, such as MTAG [30], CPASSOC [31], and MPATs [32] that are only based on the GWAS summary statistics, were proposed.

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 [33] 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 [33]. 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 needs to 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 [34], 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 combined p-values [35], 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 (binary), and each individual has been genotyped at a genetic variant of interest. Let $Y_i = (Y_{i1},\cdots,Y_{iK})^T$ represent K correlated phenotypes for the ith individual (1 for cases and 0 for controls for a qualitative trait) with $i = 1,2,\cdots,N$. Let G_i denote the genotype for the ith individual at the variant of interest, 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 $z_{i1},\ldots,z_{ip}$, we adjust both genotypes and phenotypes for the covariates [36, 37] using linear models $G_i = \alpha_0 + \alpha_1 z_{i1} + \cdots + \alpha_p z_{ip} + \varepsilon_i$ and $Y_{ik} = \alpha_{0k} + \alpha_{1k} z_{i1} + \cdots + \alpha_{pk} z_{ip} + \tau_{ik}$, 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 [38]:

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

where β_{1k} is the genetic effect of the variant on the kth 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 a univariate test to test H_0: $\beta_{1k} = 0$ for each phenotype, $k = 1,2,\cdots,K$, using the score test statistic [39]

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

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 $T = (T_1,\cdots,T_K)^T$ approximately follows a multivariate normal distribution with mean vector $\mu = (\mu_1,\cdots,\mu_K)^T$ 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: $\beta_{11} = \cdots = \beta_{1K} = 0$. Sha et al. [33] showed that under the null hypothesis, Σ converges to $P(Y)$ almost surely, where $P(Y)$ is the correlation matrix of $Y = (Y_1,\cdots,Y_K)^T$. Therefore, we can use the sample correlation matrix of Y, $P^s(Y)$, to estimate Σ.

Based on the CLC [33] and ACAT methods [34], we propose a computational efficient CLC (ceCLC) method in this paper. Same as the CLC method [33], 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 $L = 1,\cdots,K$, and B is a $K\times L$ matrix with the $(k, l)^{th}$ element equals 1 if the kth phenotype belongs to the lth cluster, otherwise it equals 0. The CLC test statistic [33] 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 $L = 1,\cdots,K$. Since for a given data set, the number of clusters of the phenotypes is unknown, in the last step of the CLC method [33], $T_{CLC} = \min_{1\le L\le K} p_L$ 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 [35] 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, we know that p_L is uniformly distributed between 0 and 1, therefore tan $\{(0.5-p_L)\pi\}$ follows a standard Cauchy distribution. If $p_1,\cdots,p_K$ are independent, the test statistic T_{ceCLC} follows a standard Cauchy distribution under the null hypothesis. However, most likely there exists a correlated structure between $p_1,\cdots,p_K$. Liu. et. al [35] 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 $0.5-\{\arctan(T_{ceCLC})/\pi\}$. The R code for the implementation of ceCLC is available at github https://github.com/MeidaWang/ceCLC.

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, 33]

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

where $Y = (Y_1,\cdots,Y_K)^T$, G is the genotype at the variant of interest, $\lambda = (\lambda_1,\cdots,\lambda_K)^T$ is the vector of genetic effect sizes on K phenotypes, c is a constant number, f is a vector of factors, and $f = (f_1,\cdots,f_R)^T \sim MVN(0,\Sigma)$, 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, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1,\cdots,\boldsymbol{\varepsilon}_K)^T$ is a vector of residuals, and $\boldsymbol{\varepsilon}_1,\cdots,\boldsymbol{\varepsilon}_K$~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, $R = 1$, $\lambda = \beta(1,2,\cdots,K)^T$ and $\gamma = (1,\cdots,1)^T$.

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

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

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 (binary). 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, that is, $\lambda = (0,\cdots,0)^T$, and consider three different sample sizes, $N = 1000$, 2000, and 3000, under four different models. The type I error rates are evaluated by 10^6 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 ceCLC with other multiple-traits analysis methods: CLC [33], MANOVA [40], MultiPhen [19], TATES [23], O'Brien [20], and Omnibus. Moreover, we also compare ceCLC with CPASSOC [31], which is an approach that is based on GWAS summary statistics and contains two different tests (Het and Hom). Based on our simulation setting on individual-level data, we can obtain the corresponding summary statistics using linear model for quantitative traits and logistic regression model for binary traits. Notably, the empirical distribution of the Het test statistic is approximated by a gamma distribution, whereas the gamma distribution may not work well when the number of traits is large, in this case, a simulation procedure needs to be used to construct the empirical distribution under the null hypothesis [31]. Since CLC and Het need a simulation procedure to obtain the final p-values, we use 10^5 replications to evaluate Type I error rates for both of the methods.

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 eight methods (CLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus, Het, Hom) 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 eight methods for $K = 40$ are summarized in S5 and S6 Tables. For 10^6 replications, the 95% confidence intervals of Type I error rates divided by nominal significance levels of 0.001 and 0.0001 are (0.9381, 1.0619) and (0.8040, 1.1960), respectively; for 10^5 replications, the corresponding confidence intervals are (0.8041, 1.1959) and (0.3802, 1.6198), respectively.

From Tables 1 and 2 (S3 and S4 Tables), we can see that ceCLC can control the Type I error rate very well, therefore we can conclude that the ceCLC method is a valid test. From S1, S2 and S5, S6 Tables, in summary, we observe that CLC, MANOVA, TATES, O'Brien, Het, and Hom can control type I error rates well, but some of the type I error rates of MultiPhen are slightly inflated.

Table 1. The estimated type I error rates divided by the nominal significance levels of the ceCLC method for 20 quantitative phenotypes with 10^6 replications.

α	Sample	Model1	Model2	Model3	Model4
	1000	0.97	0.97	0.92	0.96
0.001	2000	1.05	1.04	1.02	1.05
	3000	0.99	1.03	1.06	0.99
	1000	0.94	0.77	0.71	0.75
0.0001	2000	0.89	1.10	0.97	0.95
	3000	0.78	0.86	0.97	0.81

https://doi.org/10.1371/journal.pone.0260911.t001

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 with 10^6 replications.

α	Sample	Model1	Model2	Model3	Model4
	1000	0.99	0.95	0.93	0.98
0.001	2000	1.05	0.97	1.05	0.99
	3000	1.05	1.06	1.03	1.06
	1000	1.02	0.90	0.83	0.58
0.0001	2000	1.06	0.91	1.09	1.08
	3000	1.10	0.95	1.08	1.04

https://doi.org/10.1371/journal.pone.0260911.t002

(b) Assessment of powers. Fig 1 shows the results of power comparisons for all the nine tests with 20 quantitative phenotypes when the sample size is 5000. From Fig 1, we find that 1) when the variant of interest affects phenotypes with groups (Models 2–4), the ceCLC and CLC methods are more powerful than other methods; 2) the O'Brien and Hom methods are very sensitive to the direction of the genetic effect on the phenotypes. Their powers will decrease dramatically with different directions of the genetic effect on the phenotypes (Models 3 and 4); 3) MANOVA, Omnibus, and MultiPhen show the similar powers in most scenarios. 4) When the effect is homogeneous (Models 1 and 2), Hom is more powerful than Het; when heterogeneity is present (Models 3 and 4), Het performs better than Hom. Fig 2 shows the results of power comparisons for all the nine 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, MultiPhen, and Het are higher than those in Fig 1 for Models 3 and 4.

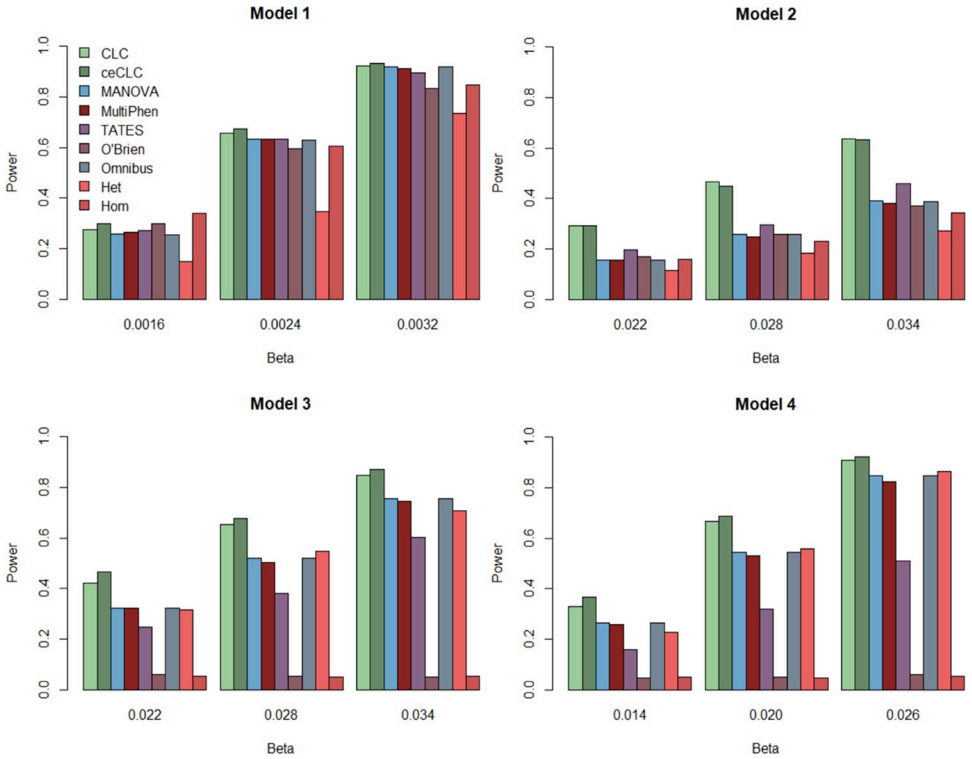

Fig 1. Power comparisons of the nine tests, CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus, Het, and Hom with 20 quantitative phenotypes for the sample size of 5000.

https://doi.org/10.1371/journal.pone.0260911.g001

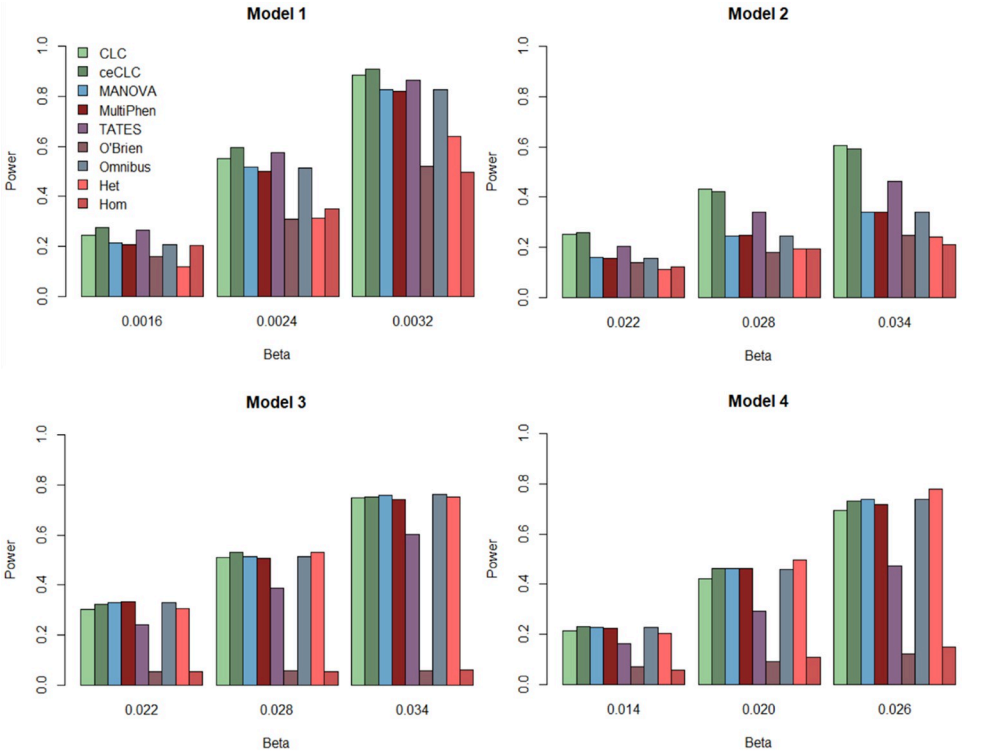

Fig 2. Power comparisons of the nine tests, CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus, Het, and Hom with 10 quantitative and 10 qualitative phenotypes for the sample size of 5000.

https://doi.org/10.1371/journal.pone.0260911.g002

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, CLC and ceCLC are more powerful than the other methods under most scenarios, and ceCLC is much more computationally efficient than CLC.

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 [41–43]. COPD causes a high mortality and has been reported to be potentially affected by genetic factors [44, 45]. The COPDGene study is a representative multicenter research to detect hereditary factors of this disease [46]. The corresponding dataset of this study was introduced in our previous papers [22, 33], and we use the same processed data as described in Sha et al. [33] for the COPDGene data analysis.

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 removing the missing data, there are 5,430 subjects across 630,860 SNPs left for the analysis. Same with the analysis in [22, 33], the signs of six-minute walk distance and FEV1 were changed, so that the correlations between the 7 phenotypes are all positive. MANOVA, MultiPhen, TATES, and Omnibus are not affected by the sign alignment in phenotypes. CLC and ceCLC are not affected much by the sign alignment. However, O'Brien and Hom are affected very much by the sign alignment [33].

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	Het	Hom
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}	7.98×10^{-10}	7.38×10^{-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}	2.34×10^{-13}	2.95×10^{-10}
4	14547447	rs1489759	10^{-9}	3.30×10^{-17}	1.11×10^{-16}	1.22×10^{-16}	2.52×10^{-16}	2.61×10^{-11}	1.11×10^{-16}	1.51×10^{-15}	2.24×10^{-11}
4	14548573	rs1980057	10^{-9}	3.29×10^{-17}	6.68×10^{-17}	8.14×10^{-17}	9.35×10^{-17}	3.04×10^{-11}	1.11×10^{-16}	7.52×10^{-16}	2.61×10^{-11}
4	14548591	rs7655625	10^{-9}	3.30×10^{-17}	7.12×10^{-17}	9.13×10^{-17}	1.64×10^{-16}	3.08×10^{-11}	1.11×10^{-16}	1.38×10^{-15}	2.64×10^{-11}
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}	1.37×10^{-11}	9.40×10^{-6}
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}	1.14×10^{-11}	8.67×10^{-6}
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}	6.26×10^{-14}	5.80×10^{-8}
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}	1.90×10^{-13}	1.46×10^{-7}
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}	2.80×10^{-11}	1.49×10^{-5}
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}	3.41×10^{-10}	2.06×10^{-5}
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}	5.06×10^{-10}	2.58×10^{-5}
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}	1.07×10^{-9}	2.27×10^{-5}
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}	1.26×10^{-6}	2.36×10^{-3}

The p-values of CLC are evaluated using 10^9 simulations. The p-values of ceCLC, O'Brien, Omnibus, TATES, MANOVA, MultiPhen, Hom, and Het are evaluated using their asymptotic distributions. The graying out p-values indicate the p-values $> 5\times10^{-8}$.

https://doi.org/10.1371/journal.pone.0260911.t003

In our analysis, we choose the commonly used genome-wide significant level $\alpha = 5\times10^{-8}$ to identify SNPs significantly associated with the 7 COPDrelated phenotypes, 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 [47–50]. From Table 3, we can see that MultiPhen detected 14 SNPs; ceCLC, CLC, MANOVA, Omnibus and Het detected 13 SNPs; TATES detected 9 SNPs; O'Brien and Hom only detected 5 SNPs. In Sha et al. [33], single-trait analysis was also performed between each of the seven phenotypes and each of the 14 SNPs. There are four SNPs rs951266, rs8034191, rs2036527, and rs931794, identified by ceCLC, but not identified by any of the single-trait tests. Therefore, these four SNPs are more likely to have pleiotropic effects. Even though we performed the sign alignment, O'Brien and Hom only identified five SNPs. TATES detected 9 SNPs because it mainly depends on the smallest P-value of the seven univariate tests. In summary, the number of SNPs identified by ceCLC is comparable to the largest number of SNPs identified by other tests, which is consistent with our simulation results.

Discussion

In the medical field, many human complex diseases are often accompanied by multiple correlated phenotypes which are usually measured simultaneously, 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 [33] and ACAT [34] 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. We compared the running time between ceCLC and CLC in the power comparison. Both methods consider one genetic variant and 20 quantitative phenotypes for 5000 individuals. The running time of ceCLC with 1000 replications on a computer with 4 Intel Cores @3.60 GHz and 16GB memory is about 25s, whereas CLC with 1000 replications and 1000 permutations is about 3min30s. 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 computationally efficient.

In this paper, we apply ceCLC to the COPDGene with seven quantitative COPD-related phenotypes. Recent studies indicate that the pleiotropic effects and genetic heterogeneity are common in the COPD comorbid traits and other immune diseases. For example, Zhu et al. [45] showed evidence of significant positive genetic correlations between COPD and cardiovascular disease-related traits (CVD); Zhu Z et al. [51–53] identified the shared genetic architecture between asthma and allergic diseases [51, 52] and between asthma and mental health disorders [53]. Moreover, pleiotropic effects were found between eight psychiatric disorders [54]. Therefore, ceCLC can also be applied to jointly analyze those phenotypes with shared genetic architecture, thus making it possible to boost statistical power to identify SNPs that were missed by the single-trait genome-wide association analysis. The SNP is more likely to have pleiotropic effect if it was identified by the multiple-trait test but missed by the single-trait test. The detection of SNPs with pleiotropic effects is helpful to promote understanding of the molecular mechanism between co-morbid diseases.

Recent phenome-wide association studies (PheWAS) require more powerful and efficient methods to identify significantly associated SNPs as a large number of phenotypes are collected, the ceCLC method developed in this paper can be applied to PheWAS. However, one limitation of the ceCLC method is that it requires individual-level phenotype data and GWAS summary statistics, where the individual-level phenotypes are used to estimate the trait correlation matrix. Because the individual-level data is often not easily accessible as a result of privacy concerns, we are currently considering a new strategy to extend the ceCLC method applicable to GWAS summary statistics without the requirement for individual-level phenotype data.

Supporting information

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

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

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

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.
(DOCX)

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

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

S1 Fig. Power comparisons of the nine tests, CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus, Hom, and Het with 40 quantitative phenotypes for the sample size of 5000.
(PDF)

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

S3 Fig. Power comparisons of the nine tests, CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus, Het, and Hom with 20 quantitative phenotypes for the sample size of 3000.
(PDF)

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

S5 Fig. Power comparisons of the nine tests, CLC, ceCLC, MANOVA, MultiPhen, TATES, O'Brien, Omnibus, Het, and Hom with 40 quantitative phenotypes for the sample size of 3000.
(PDF)

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

Acknowledgments

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. 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.

Author Contributions

Conceptualization: Shuanglin Zhang, Qiuying Sha.

Formal analysis: Meida Wang.

Methodology: Shuanglin Zhang, Qiuying Sha.

Resources: Shuanglin Zhang.

Supervision: Shuanglin Zhang, Qiuying Sha.

Validation: Shuanglin Zhang.

Writing – original draft: Meida Wang.

Writing – review & editing: Meida Wang, Shuanglin Zhang, Qiuying Sha.

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. https://doi.org/10.1038/nature08494 PMID: 19812666
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. https://doi.org/10.1016/j.ajhg.2011.11.029 PMID: 22243964
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. https://doi.org/10.1093/nar/gkt1229 PMID: 24316577
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. https://doi.org/10.1002/gepi.22011 PMID: 27900789
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. https://doi.org/10.1371/journal.pone.0169893 PMID: 28081206
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. https://doi.org/10.1016/j.ajhg.2011.10.004 PMID: 22077970
7. Gratten J, Visscher PM. Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. Genome medicine. 2016; 8(1):1–3. https://doi.org/10.1186/s13073-015-0257-9 PMID: 26750923
8. Wang Z, Wang X, Sha Q, Zhang S. Joint analysis of multiple traits in rare variant association studies. Annals of human genetics. 2016; 80(3):162–71. https://doi.org/10.1111/ahg.12149 PMID: 26990300
9. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. Nature Reviews Genetics. 2013; 14(7):483–95. https://doi.org/10.1038/nrg3461 PMID: 23752797
10. Schifano ED, Li L, Christiani DC, Lin X. Genome-wide association analysis for multiple continuous secondary phenotypes. The American Journal of Human Genetics. 2013; 92(5):744–59. https://doi.org/10.1016/j.ajhg.2013.04.004 PMID: 23643383
11. Deng Y, Pan W. Conditional analysis of multiple quantitative traits based on marginal GWAS summary statistics. Genetic epidemiology. 2017; 41(5):427–36. https://doi.org/10.1002/gepi.22046 PMID: 28464407
12. Liang X, Sha Q, Rho Y, Zhang S. A hierarchical clustering method for dimension reduction in joint analysis of multiple phenotypes. Genetic epidemiology. 2018; 42(4):344–53. https://doi.org/10.1002/gepi.22124 PMID: 29682782
13. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature methods. 2014; 11(4):407–9. https://doi.org/10.1038/nmeth.2848 PMID: 24531419
14. Jiang C, Zeng ZB. Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics. 1995; 140(3):1111–27. https://doi.org/10.1093/genetics/140.3.1111 PMID: 7672582
15. Stephens M. A unified framework for association analysis with multiple related phenotypes. PloS one. 2013; 8(7):e65245. https://doi.org/10.1371/journal.pone.0065245 PMID: 23861737
16. Bates DM, DebRoy S. Linear mixed models and penalized least squares. Journal of Multivariate Analysis. 2004; 91(1):1–7.
17. 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(6):571–80. https://doi.org/10.1002/gepi.21738 PMID: 23740720
18. 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; 96:309–25. https://doi.org/10.1016/j.neuroimage.2014.03.061 PMID: 24704269

19. O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FC, Elliott P, Jarvelin MR, et al. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PloS one. 2012; 7(5):e34861. https://doi.org/10.1371/journal.pone.0034861 PMID: 22567092

20. O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics. 1984:1079–87. PMID: 6534410

21. Yang Q, Wu H, Guo CY, Fox CS. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. Genetic epidemiology. 2010; 34(5):444–54. https://doi.org/10.1002/gepi.20497 PMID: 20583287

22. 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(1):1–0. https://doi.org/10.1038/s41598-016-0001-8 PMID: 28442746

23. Van der Sluis S, Posthuma D, Dolan CV. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. PLoS genetics. 2013; 9(1):e1003235. https://doi.org/10.1371/journal.pgen.1003235 PMID: 23359524

24. 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(5):662–76. https://doi.org/10.1016/j.ajhg.2014.03.016 PMID: 24746957

25. Klei L, Luca D, Devlin B, Roeder K. Pleiotropy and principal components of heritability combine to increase power for association analysis. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society. 2008; 32(1):9–19. https://doi.org/10.1002/gepi.20257 PMID: 17922480

26. Wang Z, Sha Q, Zhang S. Joint analysis of multiple traits using" optimal" maximum heritability test. PloS one. 2016; 11(3):e0150975. https://doi.org/10.1371/journal.pone.0150975 PMID: 26950849

27. Tang CS, Ferreira MA. A gene-based test of association using canonical correlation analysis. Bioinformatics. 2012; 28(6):845–50. https://doi.org/10.1093/bioinformatics/bts051 PMID: 22296789

28. Zhu H, Zhang S, Sha Q. A novel method to test associations between a weighted combination of phenotypes and genetic variants. PloS one. 2018; 13(1):e0190788. https://doi.org/10.1371/journal.pone.0190788 PMID: 29329304

29. Chung J, Jun GR, Dupuis J, Farrer LA. Comparison of methods for multivariate gene-based association tests for complex diseases using common variants. European Journal of Human Genetics. 2019; 27(5):811–23. https://doi.org/10.1038/s41431-018-0327-8 PMID: 30683923

30. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nature genetics. 2018; 50(2):229–37. https://doi.org/10.1038/s41588-017-0009-4 PMID: 29292387

31. Zhu X, Feng T, Tayo BO, Liang J, Young JH, Franceschini N, et al. Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. The American Journal of Human Genetics. 2015; 96(1):21–36. https://doi.org/10.1016/j.ajhg.2014.11.011 PMID: 25500260

32. Liu Z, Lin X. Multiple phenotype association tests using summary statistics in genome-wide association studies. Biometrics. 2018; 74(1):165–75. https://doi.org/10.1111/biom.12735 PMID: 28653391

33. Sha Q, Wang Z, Zhang X, Zhang S. A clustering linear combination approach to jointly analyze multiple phenotypes for GWAS. Bioinformatics. 2019; 35(8):1373–9. https://doi.org/10.1093/bioinformatics/bty810 PMID: 30239574

34. 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. https://doi.org/10.1016/j.ajhg.2019.01.002 PMID: 30849328

35. 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. https://doi.org/10.1080/01621459.2018.1554485 PMID: 33012899

36. 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. https://doi.org/10.1038/ng1847 PMID: 16862161

37. 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. https://doi.org/10.1002/gepi.21649 PMID: 22714994

38. Nelder JA, Wedderburn RW. Generalized linear models. Journal of the Royal Statistical Society: Series A (General). 1972; 135(3):370–84. https://doi.org/10.1042/bj1280389 PMID: 5084797

39. Sha Q, Zhang Z, Zhang S. Joint analysis for genome-wide association studies in family-based designs. PloS One. 2011; 6(7):e21957. https://doi.org/10.1371/journal.pone.0021957 PMID: 21799758

40. 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.

41. 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. https://doi.org/10.1056/NEJMoa032158 PMID: 15215480

42. Barnes PJ. Chronic obstructive pulmonary disease: effects beyond the lungs. PLoS medicine. 2010; 7 (3):e1000220. https://doi.org/10.1371/journal.pmed.1000220 PMID: 20305715

43. Agusti AG, Noguera A, Sauleda J, Sala E, Pons J, Busquets X. Systemic effects of chronic obstructive pulmonary disease. European Respiratory Journal. 2003; 21(2):347–60. https://doi.org/10.1183/09031936.03.00405703 PMID: 12608452

44. Sandford AJ, Weir TD, Pare PD. Genetic risk factors for chronic obstructive pulmonary disease. European Respiratory Journal. 1997; 10(6):1380–91.

45. Zhu Z, Wang X, Li X, Lin Y, Shen S, Liu CL, et al. Genetic overlap of chronic obstructive pulmonary disease and cardiovascular disease-related traits: a large-scale genome-wide cross-trait analysis. Respiratory research. 2019; 20(1):1–4. https://doi.org/10.1186/s12931-018-0967-9 PMID: 30606211

46. 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.

47. 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. https://doi.org/10.1038/ng.535 PMID: 20173748

48. 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. https://doi.org/10.1183/09031936.00033310 PMID: 21119205

49. 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. https://doi.org/10.1164/rccm.201202-0366OC PMID: 22837378

50. 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. https://doi.org/10.1186/1465-9921-12-158 PMID: 22176972

51. Zhu Z, Lee PH, Chaffin MD, Chung W, Loh PR, Lu Q, et al. A genome-wide cross-trait analysis from UK Biobank highlights the shared genetic architecture of asthma and allergic diseases. Nature genetics. 2018; 50(6):857–64. https://doi.org/10.1038/s41588-018-0121-0 PMID: 29785011

52. Zhu Z, Hasegawa K, Camargo CA Jr, Liang L. Investigating asthma heterogeneity through shared and distinct genetics: Insights from genome-wide cross-trait analysis. Journal of Allergy and Clinical Immunology. 2021; 147(3):796–807. https://doi.org/10.1016/j.jaci.2020.07.004 PMID: 32693092

53. Zhu Z, Zhu X, Liu CL, Shi H, Shen S, Yang Y, et al. Shared genetics of asthma and mental health disorders: a large-scale genome-wide cross-trait analysis. European Respiratory Journal. 2019; 54(6). https://doi.org/10.1183/13993003.01507-2019 PMID: 31619474

54. Lee PH, Anttila V, Won H, Feng YC, Rosenthal J, Zhu Z, et al. Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. Cell. 2019; 179(7):1469–82. https://doi.org/10.1016/j.cell.2019.11.020 PMID: 31835028