RESEARCH ARTICLE

A novel method to test associations between a weighted combination of phenotypes and genetic variants

Huanhuan Zhu, Shuanglin Zhang, Qiuying Sha*

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

* qsha@mtu.edu

Abstract

Many complex diseases like diabetes, hypertension, metabolic syndrome, et cetera, are measured by multiple correlated phenotypes. However, most genome-wide association studies (GWAS) focus on one phenotype of interest or study multiple phenotypes separately for identifying genetic variants associated with complex diseases. Analyzing one phenotype or the related phenotypes separately may lose power due to ignoring the information obtained by combining phenotypes, such as the correlation between phenotypes. In order to increase statistical power to detect genetic variants associated with complex diseases, we develop a novel method to test a weighted combination of multiple phenotypes (WCmulP). We perform extensive simulation studies as well as real data (COPDGene) analysis to evaluate the performance of the proposed method. Our simulation results show that WCmulP has correct type I error rates and is either the most powerful test or comparable to the most powerful test among the methods we compared. WCmulP also has an outstanding performance for identifying single-nucleotide polymorphisms (SNPs) associated with COPD-related phenotypes.

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

Editor: Zhongxue Chen, Indiana University Bloomington, UNITED STATES

Received: June 18, 2017

Accepted: December 20, 2017

Published: January 12, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This research was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number R15HG008209 to QS and SZ. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the

Introduction

Genome-wide association studies (GWAS) aim to discover genetic variants associated with complex diseases [1, 2]. In GWAS, researchers often collect data on multiple correlated phenotypes to get a better understanding of the complex disease [3]. Here are some examples of what diseases are measured by multiple phenotypes. In type 2 diabetes (T2D) studies data are usually collected on a number of risk factors and diabetes-related quantitative phenotypes. Hypertension is measured by systolic blood pressures (SBP) and diastolic blood pressures (DBP) [2], and the correlation coefficient between SBP and DBP was greater than 0.5 in 95% of patients [4]. The metabolic syndrome refers to the co-occurrence of insulin resistance, obesity, atherogenic dyslipidemia and hypertension, and these factors are associated and share underlying mediators, pathway and mechanisms [5]. The correlations between multiple phenotypes can be leveraged to improve the power of genetic association tests to identify markers associated

manuscript. This research used data generated by the COPDGene study, which was supported by NIH grants U01HL089856 and U01HL089897. 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.

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

with one or more of the phenotypes [6]. The standard approach to analyze these multiple correlated phenotypes is to perform single-phenotype analyses separately and report the findings for each phenotype [1]. However, analyzing one phenotype at a time will suffer penalties from the multiple testing and result in a reduced power especially for GWAS [3]. Recently, the joint analysis of multiple phenotypes has become popular because it can increase statistical power over analyzing phenotypes separately in detecting genetic variants [3, 6].

There are three commonly used strategies to detect genetic associations between a genetic variant and multiple correlated phenotypes. The first one is combining test statistics (or p-values) from univariate analysis. This strategy first tests an association between each phenotype and a genetic variant individually and then combines the univariate analysis results, i.e. test statistics or p-values, by using different approaches. The O'Brien's method [7], sample splitting and cross-validation method [3], Trait-based Association Test that uses Extended Simes procedure (TATES) [8], Unified Score-Based Association Test (USAT) [9], Fisher's Combination [10], and Adaptive Fisher's Combination (AFC) [11] belong to this strategy. The advantage of this strategy is its simplicity and is especially useful for analyzing different types of phenotypes such as continuous, dichotomous and survival [2]. The second one is data reduction. This strategy derives a single or a few new phenotypes that are linear combinations of the original phenotypes. Existing methods include projection-based techniques and canonical correlation analysis (CCA). Projection-based approaches include principal components analysis (PCA) and principal component of heritability (PCH), where principal components (PCs) are built to maximize either the phenotypic variance or heritability [2, 6, 12, 13]. Canonical correlation analysis (CCA) finds the linear combination of phenotypes that explain the largest possible amount of the correlation between the genetic variant and all multiple phenotypes [14]. Data reduction approaches are in general only applicable to multiple phenotypes consisting of all continuous phenotypes that are approximately normally distributed [2]. The third strategy is regression models which include mixed effect models [15–17], the generalized estimating equation (GEE) [18, 19], and reverse regression methods [1, 20, 21]. The linear mixed effects model (LME) and generalized linear mixed effects model (GLMM) are two commonly used mixed effects models, where the fixed effects are used for the genetic variant and random effects are used to account for phenotypic correlations. The GEE methods collapse the random effects and random residual errors in marginal regression models which are a class of models different from mixed effect models. The reverse regression methods take genotypes as the response variable and multiple phenotypes as predictors, such as the proportional odds logistic regression for joint model of multiple phenotypes (MultiPhen) [1]. Regression approaches are able to deal with a mixture of continuous, dichotomous, and survival phenotypes, but they are complicated and few available software were developed to implement these methods [2].

In this article, we developed a novel allele-based method for testing association between multiple phenotypes and a genetic variant. First, we take the allele at the genetic variant as the response variable and the multiple phenotypes as predictors. Then, we present a new multivariate method that we refer to as WCmulP (Weighted Combination of multiple Phenotypes), inspired by TOW (Test for testing the effect of an Optimally Weighted combination of variants) procedure proposed by Sha et al. [22] for rare variant association studies and allele-based aproach proposed by Majumdar et al. [23]. For each of the independent individuals, WCmulP linearly combines the multiple phenotypes to "one phenotype" by using the optimal weights proposed by Sha et al. [22]. Then we use the score test based on the logistic model to test the association between the genetic variant and the linear combination of phenotypes. Using extensive simulation studies, we compare the performance of WCmulP with some of the existing methods, MultiPhen[1], O'Brien's method [7], TATES [8], CCA [14], and SHet [24]. Our results show that, in all of the simulation scenarios, WCmulP is either the most powerful test or comparable to the most powerful tests among the methods we compared. Finally, we evaluate the performance of our proposed method using a real data set, the COPDGene study from dbGaP.

Methods

We consider a sample of n unrelated individuals. Each individual has K possibly correlated phenotypes. Let $Y_{i,k}$ denote the k^{th} phenotype of the i^{th} individual. We propose to use an allele-based logistic regression model to test the association between a variant of interest and multiple phenotypes. For a genetic variant with two alleles, we use x_{2i-1} and x_{2i} to denote the coding of the two alleles of the i^{th} individual such that we use x_1 and x_2 to code the two alleles of the first individual, use x_3 and x_4 to code the two alleles of the second individual, and so on. For a variant with two alleles A and a, if the genotype of the i^{th} individual is AA, we define $x_{2i-1} = x_{2i} = 1$; if the genotype is aa, we define $x_{2i-1} = x_{2i} = 0$; and if the genotype is Aa, we define $x_{2i-1} = 1$; and $x_{2i} = 0$. We define the k^{th} phenotype corresponding to the two alleles x_{2i-1} and x_{2i} of the i^{th} individual as $y_{2i-1,k}$ and $y_{2i,k}$, where $y_{2i-1,k} = y_{2i,k} = Y_{i,k}$. Hence, the total number of observations in the allele-based data is $2n$. We model the relationship between alleles and multiple phenotypes using the inverse logistic regression model

$$\text{logit}(\pi_j) = \alpha + y_{j,1}\beta_1 + y_{j,2}\beta_2 + \cdots + y_{j,K}\beta_K, \quad j = 1, 2, \ldots, 2n, \tag{1}$$

where $\pi_j = \Pr(x_j = 1|\boldsymbol{Y}_j = (y_{j,1},\ldots,y_{j,K})^T)$, α is the intercept, and $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_K)^T$ is a K-dimention vector of parameters. To test the association between multiple phenotypes and the variant is equivalent to test the null hypothesis H_0: $\boldsymbol{\beta} = \mathbf{0}$ under Eq (1). We use the score test statistic given by Sha et al. [25] to test H_0: $\boldsymbol{\beta} = \mathbf{0}$ under Eq (1). The test statistic is

$$S = \boldsymbol{U}^T\boldsymbol{V}^{-1}\boldsymbol{U}, \tag{2}$$

where $\boldsymbol{U} = \sum_{j=1}^{2n}(x_j - \bar{x})\boldsymbol{Y}_j$, $\boldsymbol{V} = (1-\bar{x})\bar{x}\sum_{j=1}^{2n}(\boldsymbol{Y}_j - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_j - \bar{\boldsymbol{Y}})^T, \bar{x} = \frac{1}{2n}\sum_{j=1}^{2n}x_j$, $\bar{\boldsymbol{Y}} = (\bar{y}_1, \ldots, \bar{y}_K)^T$ and $\bar{y}_k = \frac{1}{2n}\sum_{j=1}^{2n}y_{j,k}$ for $k = 1,\ldots,K$. The test statistic S asymptotically follows a chi-square distribution with K degrees of freedom.

When K is large, the score test may lose power due to the large degrees of freedom. To overcome this problem, we combine the K phenotypes to one variable by using a linear combination of phenotypes, $y_j = \sum_{k=1}^{K} w_k y_{j,k}$, where $w_1,\ldots, w_K$ are the weights. With the linear combination of phenotypes $y_j = \sum_{k=1}^{K} w_k y_{j,k}$, the score test statistic in Eq (2) becomes

$$S(w_1, \ldots, w_K) = 2n\frac{\left(\sum_{j=1}^{2n}(x_j - \bar{x})y_j\right)^2}{\sum_{j=1}^{2n}(x_j - \bar{x})^2\sum_{j=1}^{2n}(y_j - \bar{y})^2}. \tag{3}$$

We propose to use the optimal weights proposed by Sha et al. [22], that is, $w_k^o = \frac{\sum_{j=1}^{2n}(x_j-\bar{x})(y_{j,k}-\bar{y}_k)}{\sum_{j=1}^{2n}(y_{j,k}-\bar{y}_k)^2}$ for $k = 1,2, \ldots, K$. Actually, the optimal weights $w_1^o, \ldots, w_K^o$ maximize $S(w_1,\ldots, w_K)$ in Eq (3). With this optimally weighted combination of phenotypes $y_j^o = \sum_{k=1}^{K} w_k^o y_{j,k}$, the test statistic given in Eq (3) becomes

$$S(w_1^o, \ldots, w_K^o) = 2n \cdot \frac{\sum_{j=1}^{2n}(x_j - \bar{x})(y_j^o - \bar{y}^o)}{\sum_{j=1}^{2n}(x_j - \bar{x})^2}, \tag{4}$$

where $\bar{y}^o = \frac{1}{2n}\sum_{j=1}^{2n}y_j^o$. From Eq (2)–Eq (4), we reduced the dimension of the phenotypes from multivariate ($y_{j,k}$, $k = 1, \ldots, K$) to univariate (y_j^o) with optimal weights w_k^o such that Eq (4) is the maximum of Eq (3). Since $w_1^o, \ldots, w_K^o$ are data-driven weights, $S(w_1^o, \ldots, w_K^o)$ does not follow a chi-square distribution. We use a permutation procedure to evaluate the p-value of $S(w_1^o, \ldots, w_K^o)$. In each permutation, we randomly shuffle the genotypes and keep the phenotypes unchanged. Since $\sum_{j=1}^{2n}(x_j - \bar{x})^2$ does not change under each permutation, the test statistic $S(w_1^o, \ldots, w_K^o)$ is equivalent to

$$T = \sum_{j=1}^{2n}(x_j - \bar{x})(y_j^o - \bar{y}^o). \tag{5}$$

This test statistic T is our proposed test statistic to test the effect of the Weighted Combination of multiple Phenotypes (WCmulP).

The WCmulP method can also be extended to incorporate covariates. Suppose that there are p covariates. Let $Z_{i,l}$ denote the l^{th} covariate of the i^{th} individual. We define the l^{th} covariate corresponding to the two alleles x_{2i-1} and x_{2i} of the i^{th} individual as $z_{2i-1,l}$ and $z_{2i,l}$, where $z_{2i-1,l} = z_{2i,l} = Z_{i,l}$. We then adjust the phenotype value $y_{j,k}$ for the covariates by applying linear regressions. That is,

$$y_{j,k} = \alpha_{0,k} + \alpha_{1,k} z_{j,1} + \cdots + \alpha_{p,k} z_{j,p} + \tau_{j,k}.$$

Let $\tilde{y}_{j,k}$ denote the residuals of $y_{j,k}$ in the linear regression. We incorporate the covariate effects in WCmulP by replacing $y_{j,k}$ in Eq (5) by $\tilde{y}_{j,k}$. With covariates, the statistic of WCmulP is defined as

$$T_{\text{WCmulP}} = T|_{y_{j,k}=\tilde{y}_{j,k}}.$$

Comparison of methods

We compare the power of the proposed WCmulP with that of the following methods:

Score (Score test): the test statistic of Score is given by Eq (2).

OB (O'Brien's method) [7]: the test statistic of OB, $\boldsymbol{e}^T\Sigma^{-1}\boldsymbol{T}_{\text{uni}}$, is a linear combination of univariate test statistics, and it is the most powerful test among a class of test statistics that are linear combination of $\boldsymbol{T}_{\text{uni}}$, where $\boldsymbol{T}_{\text{uni}}$ is the vector of the univariate test statistics, Σ is the covariance matrix of $\boldsymbol{T}_{\text{uni}}$, and $\boldsymbol{e} = (1,1\ldots,1)^T$ is a 1's vector with length K (the number of phenotypes).

MultiPhen (Joint model of Multiple Phenotypes) [1]: it uses the proportional odds logistic regression to model the genotype data as ordinal response and phenotypes as predictors. A likelihood ratio test is used to test the null hypothesis.

TATES (Trait-based Association Test that uses Extended Simes procedure) [8]: it combines univariate p-values to acquire one phenotype-based p-value, while correcting for correlations between phenotypes. The TATES p-value is given by $Min\left(\frac{m_e p_{(k)}}{m_{e(k)}}\right)$, where $p_{(k)}$ is the k^{th} ($k = 1,\ldots,K$) sorted p-value in ascending order, m_e and $m_{e(k)}$ are the effective numbers of independent p-values of all K phenotypes and k specified phenotypes, respectively. The effective numbers can be calculated from the correlation matrix of p-values.

CCA (Canonical Correlation Analysis) [14]: it extracts the linear combination of phenotypes that maximizes the correlations between linear combinations of phenotypes and genotypes at the variant of interest. The test is based on Wilks' lambda and the corresponding F-approximation.

SHet (Test for Heterogeneous genetic effects) [24]: The test statistic of SHet, S_{Het}, is based on S_{Hom}, which is the most powerful test statistic when the genetic effect is homogeneous. Both S_{Hom} and S_{Het} are quadratic combinations of the univariate test statistics. The test statistic of S_{Hom} is $S_{Hom} = \frac{e^T(RW)^{-1}\,\boldsymbol{T}_{\text{uni}}(e^T(RW)^{-1}\boldsymbol{T}_{\text{uni}})^T}{e^T(WRW)^{-1}e}$, where R is the correlation matrix of $\boldsymbol{T}_{\text{uni}}$, W is a diagonal matrix of weights for the univariate test statistics, and e is a 1's vector with length K (number of phenotypes). S_{Het} can be viewed as the maximum of S_{Hom}'s satisfying different thresholds. More specifically, given a threshold, only test statistics with absolute values that are greater than the threshold are used, R and W are therefore partially used corresponding to the selected test statistics. The p-values of S_{Het} can be evaluated by simulation.

Simulation studies

Our simulations are similar to that of Wang et al. [13]. To evaluate the type I error rates and powers of our method, we simulate genotype-phenotype data sets for n unrelated individuals with total K phenotypes according to a variety of simulation scenarios. Specifically, genotype data at a genetic variant are simulated according to the minor allele frequency (MAF) under the assumption of Hardy-Weinberg equilibrium. We generate K phenotypes by the factor model

$$y = \lambda x + c\boldsymbol{\gamma} f + \sqrt{1-c^2} \times \boldsymbol{\varepsilon}, \tag{6}$$

where $y = (y_1,\ldots,y_K)^T$; x is the genotype score at the variant of interest; $\lambda = (\lambda_1,\ldots\lambda_K)$ is the vector of effect sizes of the genetic variant on the K phenotypes; $f = (f_1,\ldots,f_R)^T \sim MVN(0,\Sigma)$, $\Sigma = (1-\rho)I + \rho A$, R is the number of factors, A is a matrix with elements of 1, I is the identity matrix, and ρ is the correlation between f_i and f_j for $i \neq j$; $\boldsymbol{\gamma}$ is a K by R matrix; c is a constant number; and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_K)^T$ is a vector of residuals, $\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_K$ are independent, and $\boldsymbol{\varepsilon}_k \sim N(0,1)$ for $k = 1,\ldots,K$. Based on Eq (6), we consider the following six models.

Model 1: There is only one factor and genotype has an impact on all traits with the same effect size. That is, $R = 1$, $\lambda = (\beta,\ldots,\beta)^T$, and $\boldsymbol{\gamma} = (1,\ldots,1)^T$.

Model 2: There are two factors and genotype has an impact on two factors with opposite effects. That is, $R = 2$, $\lambda = \left(\underbrace{-\beta,\ldots,-\beta}_{K/2},\underbrace{\beta,\ldots,\beta}_{K/2}\right)^T$, and $\gamma = bdiag(D_1,D_2)$, where $D_i = \left(\underbrace{1,\ldots,1}_{K/2}\right)^T$ for $i = 1,2$, "*bdiag*" indicates the block diagonal matrix.

Model 3: There are two factors and genotype has an impact on one factor. That is, $R = 2$, $\lambda = \left(0,\ldots,0,\underbrace{\beta,\ldots,\beta}_{K/2}\right)^T$, and $\gamma = bdiag(D_1,D_2)$, where $D_i = \left(\underbrace{1,\ldots,1}_{K/2}\right)^T$ for $i = 1,2$.

Model 4: There are four factors and genotype has an impact on one factor. That is, $R = 4$, $\lambda = \left(0,\ldots,0,\underbrace{\beta,\ldots,\beta}_{K/4}\right)^T$, and $\gamma = bdiag(D_1,D_2,D_3,D_4)$, where $D_i = \left(\underbrace{1,\ldots,1}_{K/4}\right)^T$ for $i = 1,\ldots,4$.

Model 5: There are four factors and genotype has an impact on two factors with opposite effects. That is, $R = 4$, $\lambda = \left(0,\ldots,0,\underbrace{-\beta,\ldots,-\beta}_{K/4},\underbrace{\beta,\ldots,\beta}_{K/4}\right)^T$, and $\gamma = bdiag(D_1,D_2,D_3,D_4)$, where $D_i = \left(\underbrace{1,\ldots,1}_{K/4}\right)^T$ for $i = 1,\ldots,4$.

Model 6: There are four factors and genotype has an impact on three factors with effects of different directions. That is, $R = 4$,

$\lambda = \left(0,\ldots,0,\frac{2\beta}{K/4+1}\times 1,\frac{2\beta}{K/4+1}\times 2,\ldots,\frac{2\beta}{K/4+1}\times\frac{K}{4},\underbrace{-\beta,\ldots,-\beta}_{K/4},\underbrace{\beta,\ldots,\beta}_{K/4}\right)^T$, and $\gamma = bdiag(D_1,$ $D_2,D_3,D_4)$, where $D_i = \left(\underbrace{1,\ldots,1}_{K/4}\right)^T$ for $i = 1,\ldots,4$.

In the six models, the within-factor correlation is c^2 and the between-factor correlation is ρc^2. Table A in S1 File gives the structures of γ and cov($y|x$) for different numbers of factors (R = 1,2, and 4) when the number of phenotypes is 8.

We also generate phenotypes with covariates effects. We refer to Sha et al. [22] and Sun et al. [26] by adding two covariates in Eq (6) as $y = (0.5z_1 + 0.5z_2)e + \lambda x + c\gamma f + \sqrt{1-c^2} \times \varepsilon$, where z_1 is a continuous random variable generated from a standard normal distribution, z_2 is a binary random variable taking values of 0 and 1 with a probability of 0.5, and e is a K-dimensional vector with all elements being 1's. To evaluate type I error rates and powers, we consider n = 1,000 unrelated individuals, MAF = 0.3, and different numbers of phenotypes K = 8,16. To evaluate the type I error rates of all methods, we generate all phenotypes independent of genotypes by setting $\beta = 0$. We evaluate type I error rates at significance levels α = 0.001 and 0.01 for all methods. To evaluate powers, we vary the values of β (within-factor correlation c^2 = 0.5 and between-factor correlation ρc^2 = 0.1) and vary the values of within-factor correlation c^2 (0.3,0.5,. . .,0.9) (between-factor correlation ρc^2 = 0.1 and β = 0.1,).

Simulation results

To evaluate the type I error rates of WCmulP and other six methods, we consider different numbers of phenotypes, different significance levels, and different numbers of factors. In each simulation scenario, the p-values of WCmulP and SHet are estimated using 10,000 permutations, and the p-values of Score, MultiPhen, TATES, CCA and OB are estimated using their asymptotic distributions. The type I error rates of the seven methods are evaluated using 10,000 replicated samples. For 10,000 replicated samples, the 95% confidence intervals (CIs) for type I error rates of nominal levels 0.001 and 0.01 are (0.00038,0.00162) and (0.008,0.012), respectively. The estimated type I error rates of WCmulP and other six methods are summarized in Table 1 (K = 8) and Table 2 (K = 16). From these tables, we can see that all estimated type I error rates of WCmulP are within 95% CIs, which indicates that the proposed WCmulP is a valid test. The estimated type I error rates of SHet, Score, MultiPhen, TATES, CCA and OB are not significantly different from the nominal levels.

For power comparisons, we consider power as a function of genetic effect β (Figs 1 and 2) and power as a function of within-factor correlation c^2 (Figs 3 and 4). In each of the simulation scenario, the p-values of WCmulP and SHet are estimated using 1,000 permutations and the p-values of Score, MultiPhen, TATES, CCA and OB are estimated using their asymptotic distributions. The powers of the seven methods are evaluated using 1,000 replicated samples at a significance level of 0.01.

Table 1. Estimated type I error rates for the seven methods under three simulation settings. The number of phenotypes is K = 8, c^2 = 0.5, ρc^2 = 0.1, and MAF = 0.3. The p-values of WCmulp and SHet are evaluated using 10,000 permutations. The type I error rate of all of the seven methods is evaluated using 10,000 replicated samples at a significance level of α. R is the number of factors.

	Type I error rates					
	R = 1		R = 2		R = 4	
Methods	**α = 0.001**	**α = 0.01**	**α = 0.001**	**α = 0.01**	**α = 0.001**	**α = 0.01**
WCmulP	0.0008	0.0097	0.0011	0.0091	0.0011	0.0104
SHet	0.0008	0.0106	0.0009	0.0093	0.0008	0.0104
Score	0.0006	0.0102	0.0008	0.0103	0.0004	0.0105
MultiPhen	0.0011	0.0106	0.0011	0.0105	0.0005	0.0107
TATES	0.0012	0.0094	0.0007	0.0121	0.0004	0.0106
CCA	0.0008	0.0107	0.0010	0.0099	0.0008	0.0107
OB	0.0007	0.0095	0.0016	0.0092	0.0013	0.0105

https://doi.org/10.1371/journal.pone.0190788.t001

Table 2. Estimated type I error rates for the seven methods under three simulation settings. The number of phenotypes is $K = 16$, $c^2 = 0.5$, $\rho c^2 = 0.1$, and $MAF = 0.3$. The p-values of WCmulp and SHet are evaluated using 10,000 permutations. The type I error rate of all of the seven methods is evaluated using 10,000 replicated samples at a significance level of α.

	Type I error rates					
	***R* = 1**		***R* = 2**		***R* = 4**	
Methods	**α = 0.001**	**α = 0.01**	**α = 0.001**	**α = 0.01**	**α = 0.001**	**α = 0.01**
WCmulP	0.0011	0.0089	0.0006	0.0094	0.0008	0.0098
SHet	0.0009	0.0098	0.0009	0.0126	0.0008	0.0088
Score	0.0010	0.0096	0.0011	0.0098	0.0010	0.0086
MultiPhen	0.0011	0.0096	0.0011	0.0121	0.0013	0.0103
TATES	0.0013	0.0110	0.0012	0.0102	0.0008	0.0104
CCA	0.0012	0.0097	0.0009	0.0111	0.0011	0.0089
OB	0.0011	0.0085	0.0006	0.0092	0.0007	0.0097

https://doi.org/10.1371/journal.pone.0190788.t002

Our simulation results show that:

1. As expected, the powers of all methods increase as the genetic effect β increases in each model (Figs 1 and 2).

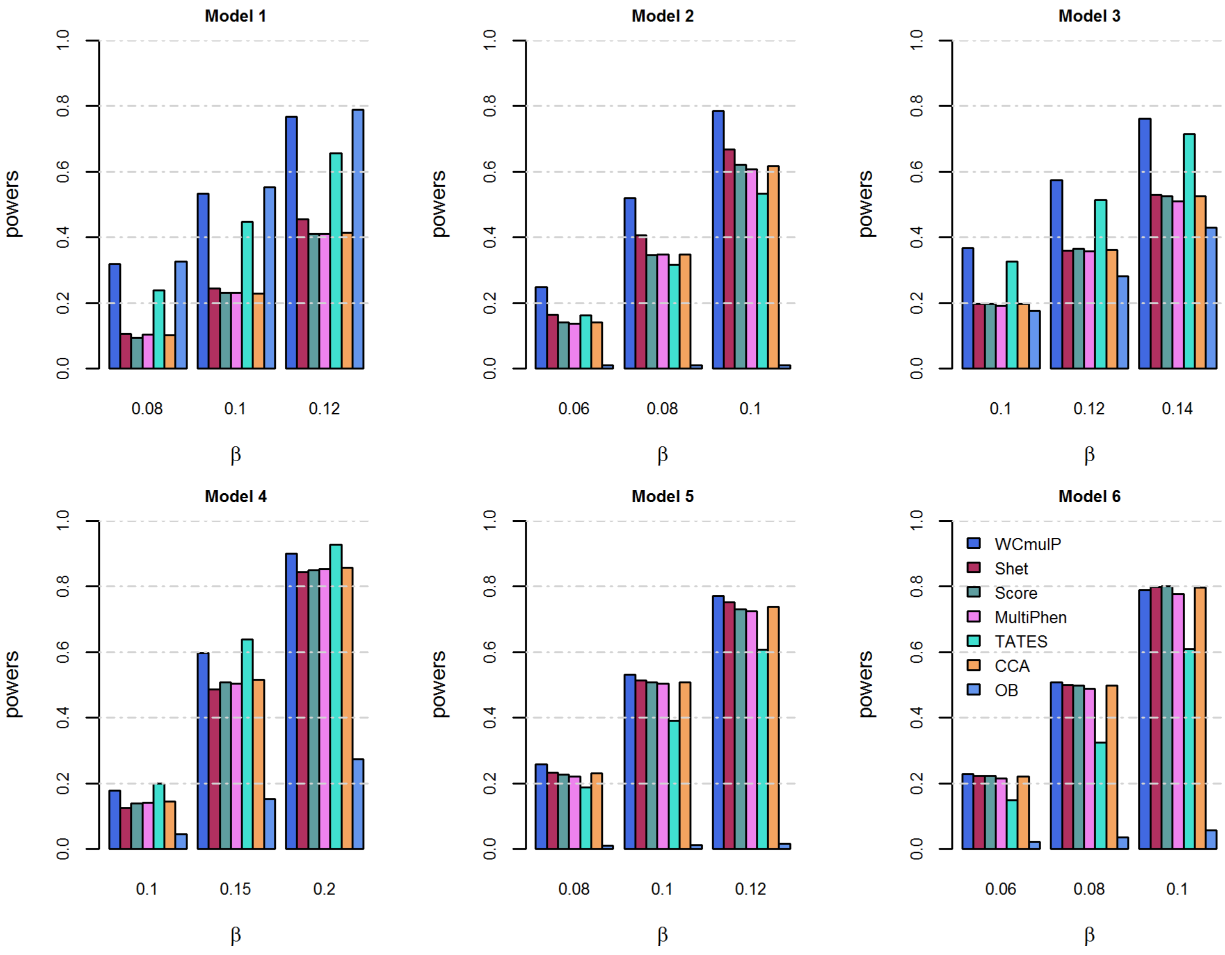

Fig 1. Power comparisons of the seven methods as a function of *β* for the six models. The total number of phenotypes is $K = 8$, $c^2 = 0.5$, $\rho c^2 = 0.1$, and $MAF = 0.3$. The p-values of WCmulP and SHet are evaluated using 1,000 permutations. The power of all of the seven methods is evaluated using 1,000 replicated samples at a significance level of 0.01.

https://doi.org/10.1371/journal.pone.0190788.g001

2. WCmulP is either the most powerful test or comparable to the most powerful tests in all six models (Figs 1–4).

3. As number of phenotypes increases from $K = 8$ to $K = 16$, WCmulP presents more obvious ascendancy than other methods.

4. SHet, Score, MultiPhen, and CCA have similar performance in all six models; we call these four tests as group 1.

5. OB is the most powerful test when the genetic effects are homogeneous (model 1). However, OB reduces power significantly when genetic effects are heterogeneous, especially when opposite directions of the genetic effects exist (models 2, 5–6) or when the genetic variant impacts only a small portion of phenotypes (model 4). This phenomenon was also observed by Zhu et al. [27].

6. Power comparisons of TATES with tests in group 1 depend on the models. In general, TATES is more powerful than tests in group 1 when the genetic variant impacts on a portion of phenotypes (models 3 and 4).

7. In general, as the within-factor correlation c^2 increases, the powers of all methods decrease (Figs 3 and 4). TATES is relatively robust to c^2 because it essentially only depends on the

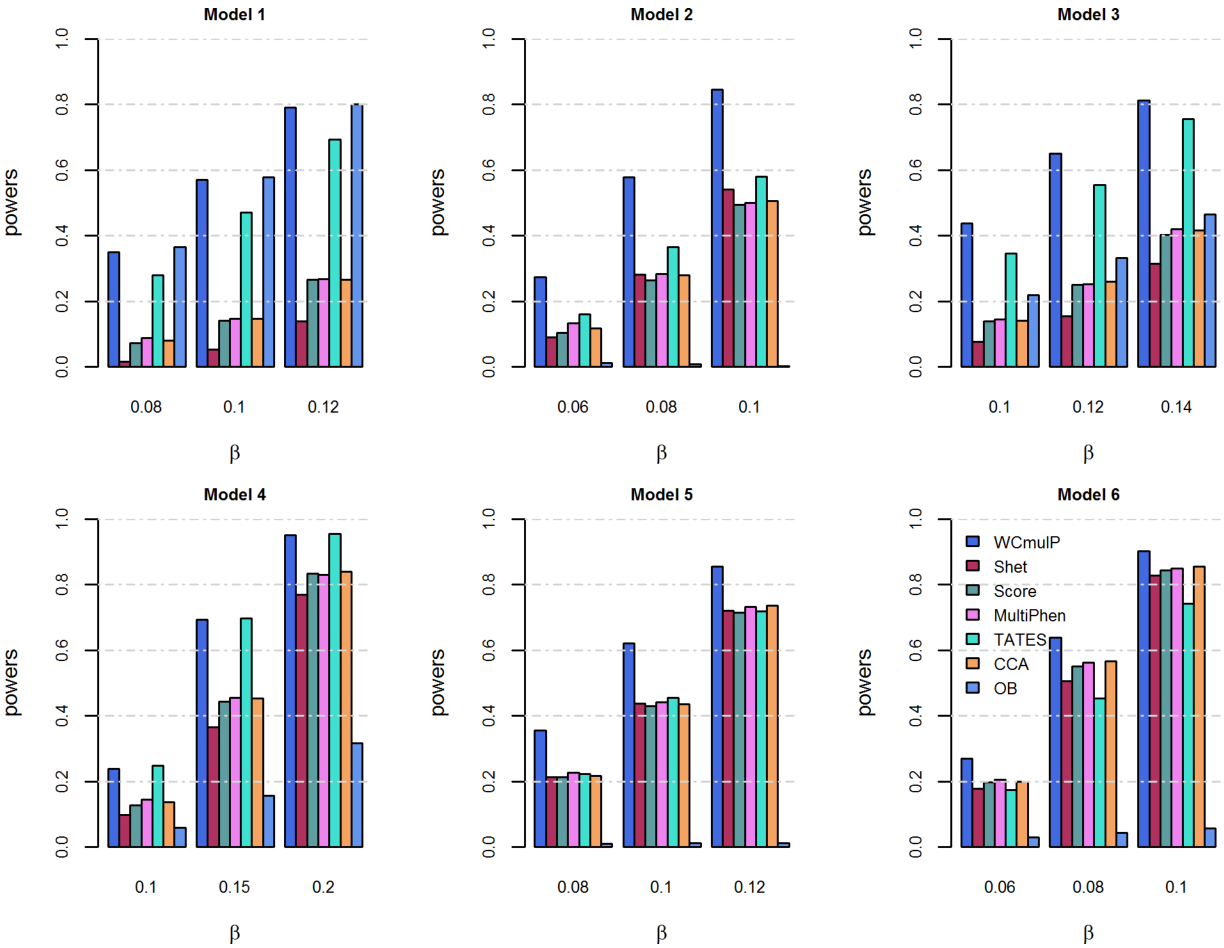

Fig 2. Power comparisons of the seven methods as a function of *β* for the six models. The total number of phenotypes is $K = 16$, $c^2 = 0.5$, $\rho c^2 = 0.1$, and $MAF = 0.3$. The p-values of WCmulP and SHet are evaluated using 1,000 permutations. The power of all of the seven methods is evaluated using 1,000 replicated samples at a significance level of 0.01.

https://doi.org/10.1371/journal.pone.0190788.g002

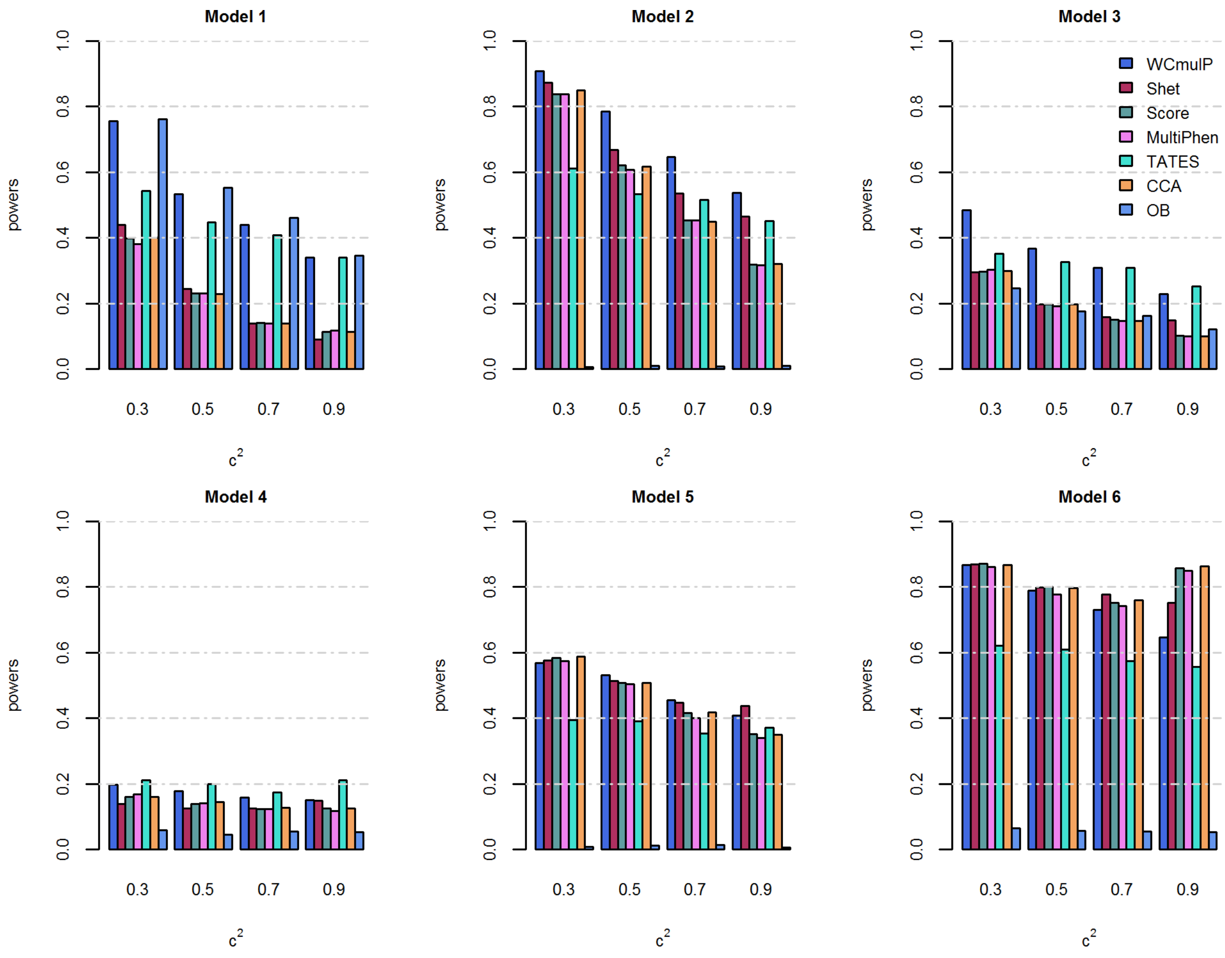

Fig 3. Power comparisons of the seven methods as a function of c^2 for the six models. The total number of phenotypes is $K = 8$, $\rho c^2 = 0.1$, $\beta = 0.1$, and $MAF = 0.3$. The p-values of WCmulP and SHet are evaluated using 1,000 permutations, the p-values of other methods are evaluated using asymptotic distribution. The power of all of the seven methods is evaluated using 1,000 replicated samples at a significance level of 0.01.

https://doi.org/10.1371/journal.pone.0190788.g003

phenotype that has the strongest association with the genetic variant, as explained in Zhu et al. [27].

We also considered using principal components (PCs) of the phenotypes instead of the original phenotypes to do power comparisons and the results are given in Figures A-D in S1 File. We exclude PCs that explain less than 10^{-6} of the total variation. Using PCs of the phenotypes, we observe that: (1) WCmulP, Score, MultiPhen, and CCA have very similar powers in all six models (Figures A-D in S1 File). We call these tests as group s1. The tests in group s1 are either the most powerful tests or comparable to the most powerful one; (2) SHet is less powerful than the tests in group s1; (3) OB is the least powerful method in all six models because PCs likely have effects with different directions; (4) TATES becomes the most powerful method when the genetic variant has effects on all phenotypes with the same absolute value of effect sizes (models 1 and 2) because in this case, one of the PCs may capture the most of association information.

We also compared the powers using a lower significance level 5×10^{-5} (Figure E in S1 File). Figure E in S1 File shows that the pattern of the power comparisons by using significance level 5×10^{-5} is similar to that by using significance level 0.01 (Fig 1).

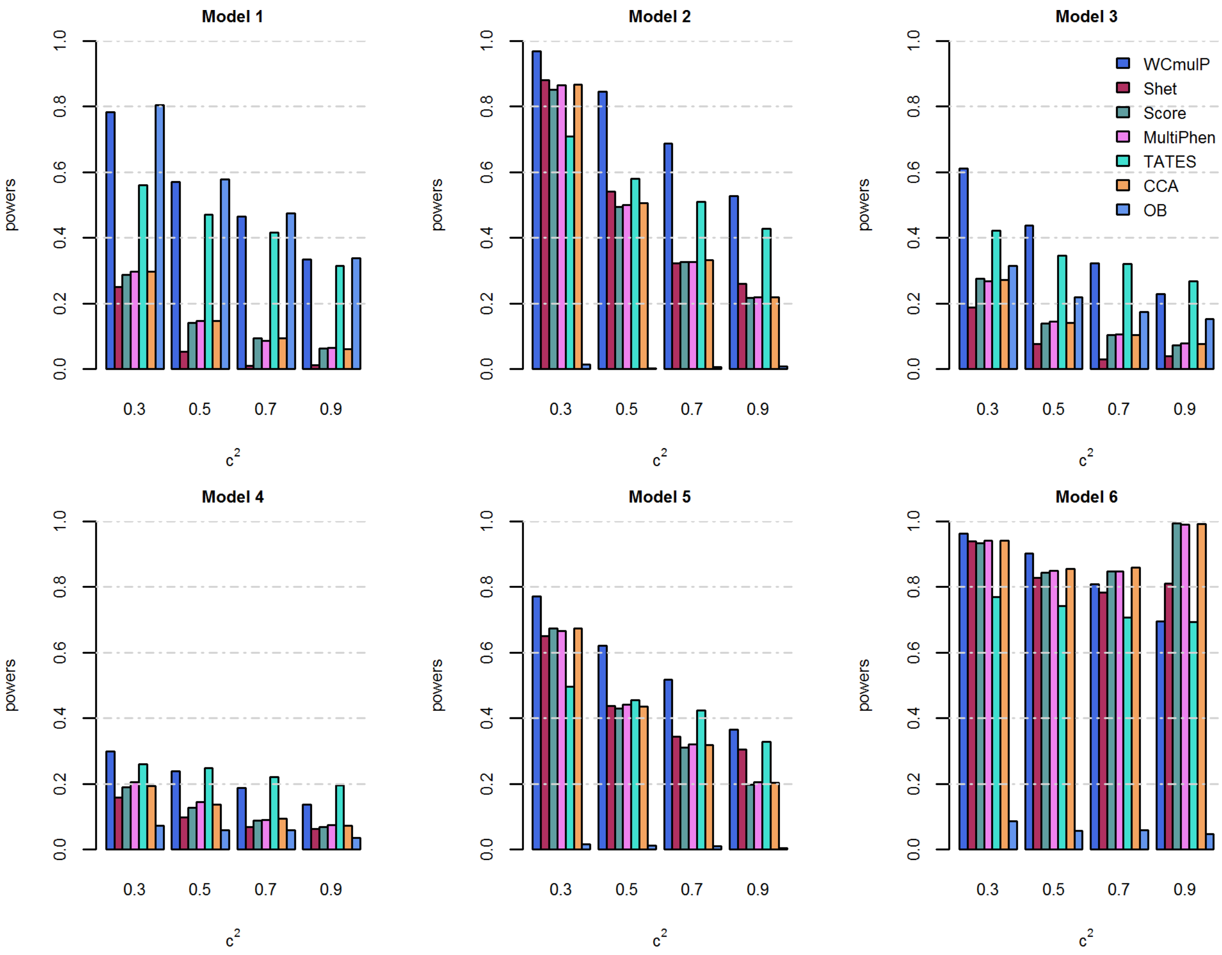

Fig 4. Power comparisons of the seven methods as a function of c^2 for the six models. The total number of phenotypes is $K = 16$, $\rho c^2 = 0.1$, $\beta = 0.1$, and $MAF = 0.3$. The p-values of WCmulP and SHet are evaluated using 1,000 permutations, the p-values of other methods are evaluated using asymptotic distribution. The power of all of the seven methods is evaluated using 1,000 replicated samples at a significance level of 0.01.

https://doi.org/10.1371/journal.pone.0190788.g004

Real data analysis

Chronic obstructive pulmonary disease (COPD) refers to a group of diseases that cause airflow blockage and breathing-related problems. The Genetic Epidemiology of COPD Study (COPD-Gene) is a multicenter observational study designed to identify genetic factors associated with

Table 3. Description of COPD-related phenotypes.

Phenotypes	Descriptions
Gas Trapping (GasTrap)	Air trapping at -856 Hounsfield units (HU) on expiratory chest CT scan
Exacerbation Frequency (ExacerFreq)	Number of COPD exacerbations during the year before study enrollment
Emphysema (Emph)	% Emphysema at -950 HU
Airway Wall Area (Pi10)	Square root of the wall area of a hypothetical 10 mm internal perimeter airway
Emphysema Distribution (EmphDist)	Log ratio of emphysema at -950 HU in the upper 1/3 of lung fields compared to the lower 1/3 of lung fields
Six Minute Walk Distance (6MWD)	Measure of exercise capacity
FEV1	Observed FEV1 (liters)/predicted FEV1 (liters), with predicted values from Hankinson reference equations

https://doi.org/10.1371/journal.pone.0190788.t003

Table 4. Significant SNPs and the corresponding p-values in the analysis of COPDGene. The p-values of WCmulP are evaluated using 10^9 permutations; the p-values of SHet are evaluated using 10^8 permutations. The p-values of Score, MultiPhen, CCA, TATES, and OB are evaluated using asymptotic distributions. The grayed-out p-values indicate the p-values $> 5 \times 10^{-8}$.

Chr	Position	Variant identifier	WCmulP	SHet	Score	MultiPhen	CCA	TATES	OB
4	145431497	rs1512282	0	1.0E-08	1.90E-09	1.03E-09	1.69E-09	5.77E-09	0.339
4	145434744	rs1032297	0	0	5.55E-14	7.69E-14	6.52E-14	6.22E-13	0.452
4	145474473	rs1489759	0	0	1.11E-16	1.22E-16	1.11E-16	2.52E-16	0.483
4	145485738	rs1980057	0	0	1.11E-16	8.14E-17	0	9.35E-17	0.411
4	145485915	rs7655625	0	0	1.11E-16	9.13E-17	0	1.64E-16	0.478
15	78882925	rs16969968	0	0	1.91E-11	7.84E-12	1.32E-11	2.98E-08	0.986
15	78894339	rs1051730	1.00E-08	0	2.05E-11	8.16E-12	1.41E-11	2.63E-08	0.992
15	78898723	rs12914385	0	0	1.78E-12	1.48E-12	1.76E-12	5.14E-10	0.999
15	78911181	rs8040868	0	0	2.21E-12	2.59E-12	2.74E-12	2.40E-09	0.768
15	78878541	rs951266	2.00E-08	0	2.42E-11	1.02E-11	1.77E-11	5.17E-08	0.956
15	78806023	rs8034191	4.00E-08	1.0E-08	2.95E-10	7.74E-11	2.14E-10	1.02E-07	0.868
15	78851615	rs2036527	4.00E-08	1.0E-08	5.58E-10	1.77E-10	3.99E-10	1.56E-07	0.880
15	78826180	rs931794	4.80E-08	3.0E-08	3.13E-10	9.09E-11	2.35E-10	1.18E-07	0.913
15	78740964	rs2568494	7.18E-06	1.93E-06	1.22E-07	4.23E-08	1.05E-07	2.88E-05	0.269
15	78733731	rs17483721	8.12E-06	2.29E-06	2.26E-07	9.87E-08	2.11E-07	3.57E-05	0.308
15	78742376	rs17483929	8.15E-06	2.13E-06	1.65E-07	6.53E-08	1.50E-07	2.82E-05	0.347

https://doi.org/10.1371/journal.pone.0190788.t004

COPD, to define and characterize disease-related phenotypes, and to assess the association of disease-related phenotypes with the identified susceptibility genes [28]. 10,192 participants (including 6,784 non-Hispanic Whites (NHW) and 3,408 African-Americans (AA)) are included in COPDGene. We selected 7 key quantitative COPD-related phenotypes and 4 covariates that are the same as those in Liang et al. [11]. The detailed description of these 7 phenotypes is in Table 3, and their correlation structure is given in Figure F in S1 File. The four covariates include Body Mass Index, Age, Pack-Years (one pack-year is defined as smoking one pack per day for one year), and gender. A set of 5,430 NHW across 630,860 SNPs were used in the analysis after excluding subjects with missing data in any of the 11 variables.

We apply WCmulP and other six methods to both original 7 phenotypes (Table 4) and the principal components (PCs) of the phenotypes (Table B in S1 File). PCs that explain less than 10^{-6} of the total variation are excluded. In this way, one PC is excluded and there are 6 PCs left. Using the first few PCs is also a dimension reduction method. Thus, using PCs of the phenotypes, WCmulP uses two dimension reduction methods: using the first few PCs and the weighted combination of those PCs. To identify SNPs significantly associated with the 7 COPD-related phenotypes and the top 6 PCs of the phenotypes, we use the genome-wide significance threshold of 5×10^{-8}. There are total 16 SNPs that are significant under at least one method (Table 4 and Table B in S1 File). Those 16 SNPs have been reported being associated with the COPD-related phenotypes by previous studies [29–42]. From Table 4, we can see that MultiPhen identified the largest number of SNPs, 14 SNPs; WCmulP, SHet, Score, and CCA identified 13 SNPs; TATES identified 9 SNPs; and OB didn't identify any SNPs, that's likely because the true genetic effects of each SNP are heterogeneous for all phenotypes. From Table B in S1 File, we can see that using PCs of the phenotypes, WCmulP identified all of the 16 SNPs; MultiPhen identified 15 SNPs; SHet, Score, and CCA identified 13 SNPs; TATES identified 4 SNPs; and OB identified 3 SNPs. In summary, the number of SNPs identified by WCmulP is comparable to the largest number of SNPs identified by other tests; and using PCs of phenotypes, WCmulP is the only method that identified all 16 SNPs. The results of the real data analysis are consistent with our simulation results.

Discussion

In this article, we developed WCmulP to perform multivariate analysis of multiple phenotypes in association studies based on the following reasons: (1) complex diseases are usually measured by multiple correlated phenotypes in genetic association studies; and (2) there is increasing evidence showing that studying multiple correlated phenotypes jointly may increase powers for detecting genetic variants that are associated with complex diseases. Our results show that WCmulP has correct type I error rates and is either the most powerful test or comparable to the most powerful tests among the seven tests we considered. None of the other methods showed consistent good performances under the simulation scenarios. OB is the most powerful test when the genetic effects are homogeneous, while it loses power dramatically when genetic effects are heterogeneous; especially when opposite directions of the genetic effects exist. SHet, Score, MultiPhen, and CCA have similar powers and they are less powerful than WCmulP in most scenarios. TATES is more powerful only when the genetic variant affects a portion of phenotypes. In addition, in the real data analysis, WCmulP identified 13 (out of 16) significant SNPs, 1 SNP less than the largest number of identified SNPs; using PCs of phenotypes, WCmulP is the only method that identified all 16 SNPs. The real data analysis results show that WCmulP has excellent performance in identifying SNPs associated with complex disease with multiple correlated phenotypes such as COPD.

In the context of association studies, it is important to correct for population stratification (PS). PS refers to allele frequency differences between populations unrelated to the outcome of interest, but due to systematic ancestry differences. PS can cause seriously confounded associations if not adjusted properly [43, 44]. The principal component analysis (PCA) method [45–49] and linear mixed model (LMM) approach [50–52] have been used to adjust for population stratification. There are also other methods such as multidimensional scaling (MDS) [53], the robust PCA based on resampling by half means (RPCA-RHM) [54], and the robust PCA based on the projection pursuit (RPCA-PP) [54], which are extension methods of the PCA approach. PCA identifies several top principal components of the genotype data matrix and uses them as covariates in the association analysis. We propose to use PCA to control for PS in our proposed method when samples from different populations are involved. However, the performance needs further investigations.

One disadvantage of WCmulP is that the test statistic does not have an asymptotic distribution and a permutation procedure is needed to calculate its p-value, which is time consuming compared to the methods whose test statistics have asymptotic distributions. The running time of WCmulP with 1,000 permutations on a data set with 5,000 individuals and 20 phenotypes on a laptop with 4 Intel(R) Cores(TM) i7-4790 CPU @ 3.6GHz and 4 GB memory is no more than 0.15s. To perform GWAS, we can first select genetic variants that show evidence of association based on a small number of permutations (e.g. 1,000), and then a large number of permutations are used to test the selected significant genetic variants [21]. Furthermore, WCmulP cannot be used for rare variant association studies, although recent studies have shown that complex diseases are caused by both common and rare variants [50, 55–58]. How to extend WCmulP to rare variant association studies is our future work.

In our simulation studies, the numbers of phenotypes varied from 8 to 16 and the methods rely on all observations having fully observed phenotypes. However, in real data analysis, as the number of phenotypes increases the chance that missing at least one observation increases exponentially, especially in epidemiological and clinical research [59, 60]. There are several approaches to handle missing phenotypes: deletion-based methods, simple replacement methods, and imputation methods [59]. The most commonly used method for dealing with missing data is deletion-based method, in which observations with missing values are removed from the analysis [59]. However, removal of observations with missing values will reduce sample size, thus resulting in power losses [60]. The simple replacement methods replace the missing values with plausible values for the variable with missing values, such as the sample mean [8, 59]. It is a simple, unconditional method that does not depend on other variables. However, mean substitution approach may result in biased estimates where data are not missing completely at random [59]. Imputation is a more sophisticated approach that fills in missing values with predicted values using model-based methods or conditional imputation, including multiple imputation (MI), multivariate normal imputation (MVNI), and fully conditional specification (FCS) [59, 61–66]. In MI, the incomplete dataset is generated multiple times and missing values are replaced by values drawn from a posterior distribution according to a suitable imputation model that utilizes the rest of the data [59, 61]. MVNI fits a joint imputation model to all the variables containing missing values under the assumption that the variables follow a multivariate normal distribution [62, 63]. For each variable with missing values, FCS fits separate univariate regression models and iteratively cycles through the univariate regression models [64–66]. In our real data analysis, we removed 1354 observations with missing either phenotypes or covariates from 6784 samples. An alternative approach is to use mean substitution or imputation approaches to fill in the missing values.

Supporting information

S1 File.
(PDF)

Acknowledgments

Research reported in this article 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 NIH grants U01HL089856 and U01HL089897. 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.

The authors have no conflict of interests to declare.

Author Contributions

Formal analysis: Huanhuan Zhu, Shuanglin Zhang.

Methodology: Huanhuan Zhu, Shuanglin Zhang, Qiuying Sha.

Project administration: Qiuying Sha.

Writing – original draft: Huanhuan Zhu, Shuanglin Zhang.

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

References

1. 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; PubMed Central PMCID: PMC3342314.

2. Yang Q, Wang Y. Methods for analyzing multivariate phenotypes in genetic association studies. J Probab Stat. 2012; 2012:652569. https://doi.org/10.1155/2012/652569 PMID: 24748889; PubMed Central PMCID: PMCPMC3989935.
3. 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; PubMed Central PMCID: PMC3090041.
4. Gavish B, Ben-Dov IZ, Bursztyn M. Linear relationship between systolic and diastolic blood pressure monitored over 24 h: assessment and correlates. Journal of hypertension. 2008; 26(2):199–209. https://doi.org/10.1097/HJH.0b013e3282f25b5a PMID: 18192832.
5. Huang PL. A comprehensive definition for metabolic syndrome. Disease models and mechanisms. 2009; 2(5–6):231–7. https://doi.org/10.1242/dmm.001180 PMID: 19407331
6. Aschard H, Vilhjalmsson BJ, Greliche N, Morange PE, Tregouet DA, Kraft P. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. American journal of human genetics. 2014; 94(5):662–76. https://doi.org/10.1016/j.ajhg.2014.03.016 PMID: 24746957; PubMed Central PMCID: PMC4067564.
7. O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics. 1984; 40(4):1079–87. PMID: 6534410.
8. 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; PubMed Central PMCID: PMC3554627.
9. Ray D, Pankow JS, Basu S. USAT: A Unified Score-Based Association Test for Multiple Phenotype-Genotype Analysis. Genetic epidemiology. 2016; 40(1):20–34. https://doi.org/10.1002/gepi.21937 PMID: 26638693; PubMed Central PMCID: PMCPMC4785800.
10. 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. https://doi.org/10.1186/s12859-015-0868-6 PMID: 26729364; PubMed Central PMCID: PMCPMC4704475.
11. 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; 6:34323. https://doi.org/10.1038/srep34323 PMID: 27694844
12. 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(1):9–19. https://doi.org/10.1002/gepi.20257 PMID: 17922480.
13. 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
14. Ferreira MA, Purcell SM. A multivariate test of association. Bioinformatics. 2009; 25(1):132–3. https://doi.org/10.1093/bioinformatics/btn563 PMID: 19019849.
15. 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; PubMed Central PMCID: PMC4211878.
16. Korte A, Vilhjalmsson 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(9):1066–71. https://doi.org/10.1038/ng.2376 PMID: 22902788; PubMed Central PMCID: PMC3432668.
17. Casale FP, Rakitsch B, Lippert C, Stegle O. Efficient set tests for the genetic analysis of correlated traits. Nature methods. 2015; 12(8):755–8. https://doi.org/10.1038/nmeth.3439 PMID: 26076425.
18. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986; 42(1):121–30. PMID: 3719049.
19. Zhang Y, Xu Z, Shen X, Pan W, Alzheimer's Disease Neuroimaging I. 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; PubMed Central PMCID: PMCPMC4043944.
20. 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.
21. 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
22. 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.

23. Majumdar A, Witte JS, Ghosh S. Semiparametric Allelic Tests for Mapping Multiple Phenotypes: Binomial Regression and Mahalanobis Distance. Genetic epidemiology. 2015; 39(8):635–50. https://doi.org/10.1002/gepi.21930 PMID: 26493781; PubMed Central PMCID: PMCPMC4958458.

24. 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. American journal of human genetics. 2015; 96(1):21–36. https://doi.org/10.1016/j.ajhg.2014.11.011 PMID: 25500260; PubMed Central PMCID: PMCPMC4289691.

25. Sha Q, Zhang Z, Zhang S. An improved score test for genetic association studies. Genetic epidemiology. 2011; 35(5):350–9. https://doi.org/10.1002/gepi.20583 PMID: 21484862.

26. Sun J, Oualkacha K, Forgetta V, Zheng H-F, Brent Richards J, Ciampi A, et al. A method for analyzing multiple continuous phenotypes in rare variant association studies allowing for flexible correlations in variant effects. European journal of human genetics. 2016; 24(9):1344–51. https://doi.org/10.1038/ejhg.2016.8 PMID: 26860061

27. Zhu H, Zhang S, Sha Q. Power Comparisons of Methods for Joint Association Analysis of Multiple Phenotypes. Human heredity. 2015; 80(3):144–52. https://doi.org/10.1159/000446239 PMID: 27344597.

28. Regan EA, Hokanson JE, Murphy JR, Make B, Lynch DA, Beaty TH, et al. Genetic Epidemiology of COPD (COPDGene) Study Design. COPD. 2010; 7(1):32–43. https://doi.org/10.3109/15412550903499522 PMID: 20214461

29. Pillai SG, Ge D, Zhu G, Kong X, Shianna KV, Need AC, et al. A Genome-Wide Association Study in Chronic Obstructive Pulmonary Disease (COPD): Identification of Two Major Susceptibility Loci. PLoS genetics. 2009; 5(3):e1000421. https://doi.org/10.1371/journal.pgen.1000421 PMID: 19300482

30. Wilk JB, Chen TH, Gottlieb DJ, Walter RE, Nagle MW, Brandler BJ, et al. A genome-wide association study of pulmonary function measures in the Framingham Heart Study. PLoS genetics. 2009; 5(3):e1000429. https://doi.org/10.1371/journal.pgen.1000429 PMID: 19300500; PubMed Central PMCID: PMCPMC2652834.

31. Wilk JB, Shrine NRG, 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

32. Cho MH, Boutaoui N, Klanderman BJ, Sylvia JS, Ziniti JP, Hersh CP, 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

33. Cho MH, Castaldi PJ, Wan ES, Siedlinski M, Hersh CP, Demeo DL, et al. A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. Human molecular genetics. 2012; 21(4):947–57. https://doi.org/10.1093/hmg/ddr524 PMID: 22080838

34. Cho MH, McDonald M-LN, Zhou X, Mattheisen M, Castaldi PJ, Hersh CP, et al. Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. The lancet respiratory medicine. 2014; 2(3):214–25. https://doi.org/10.1016/S2213-2600(14)70002-5 PMID: 24621683

35. Hancock DB, Eijgelsheim M, Wilk JB, Gharib SA, Loehr LR, Marciante KD, et al. Meta-analyses of genome-wide association studies identify multiple novel loci associated with pulmonary function. Nature genetics. 2010; 42(1):45–52. https://doi.org/10.1038/ng.500 PMID: 20010835

36. 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. The European respiratory journal. 2010; 36(6):1375–82. https://doi.org/10.1183/09031936.00033310 PMID: 21119205.

37. Li X, Howard TD, Moore WC, Ampleford EJ, Li H, Busse WW, et al. Importance of hedgehog interacting protein and other lung function genes in asthma. Journal of allergy and clinical immunology. 2011; 127(6):1457–65. https://doi.org/10.1016/j.jaci.2011.01.056 PMID: 21397937

38. 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; 12:158. https://doi.org/10.1186/1465-9921-12-158 PMID: 22176972; PubMed Central PMCID: PMCPMC3283485.

39. 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(7):e102324. https://doi.org/10.1371/journal.pone.0102324 PMID: 25051068

40. Zhu AZX, Zhou Q, Cox LS, David SP, Ahluwalia JS, Benowitz NL, et al. Association of CHRNA5-A3-B4 SNP rs2036527 with smoking cessation therapy response in African American smokers. Clinical pharmacology and therapeutics. 2014; 96(2):256–65. https://doi.org/10.1038/clpt.2014.88 PMID: 24733007

41. Lutz SM, Cho MH, Young K, Hersh CP, Castaldi PJ, McDonald M-L, et al. A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. BMC genetics. 2015; 16:138. https://doi.org/10.1186/s12863-015-0299-4 PMID: 26634245

42. Lee JH, Cho MH, Hersh CP, McDonald M-LN, Wells JM, Dransfield MT, et al. IREB2 and GALC Are Associated with Pulmonary Artery Enlargement in Chronic Obstructive Pulmonary Disease. American journal of respiratory cell and molecular biology. 2015; 52(3):365–76. https://doi.org/10.1165/rcmb.2014-0210OC PMID: 25101718

43. Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. American journal of human genetics. 1988; 43 (4):520–6. PMID: 3177389; PubMed Central PMCID: PMC1715499.

44. Lander ES, Schork NJ. Genetic dissection of complex traits. Science. 1994; 265(5181):2037–48. PMID: 8091226.

45. Chen HS, Zhu X, Zhao H, Zhang S. Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. Annals of human genetics. 2003; 67(Pt 3):250–64. PMID: 12914577.

46. Zhang S, Zhu X, Zhao H. On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. Genetic epidemiology. 2003; 24(1):44–56. https://doi.org/10.1002/gepi.10196 PMID: 12508255.

47. Zhu X, Zhang S, Zhao H, Cooper RS. Association mapping, using a mixture model for complex traits. Genetic epidemiology. 2002; 23(2):181–96. https://doi.org/10.1002/gepi.210 PMID: 12214310.

48. 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.

49. Bauchet M, McEvoy B, Pearson LN, Quillen EE, Sarkisian T, Hovhannesyan K, et al. Measuring European population stratification with microarray genotype data. American journal of human genetics. 2007; 80(5):948–56. https://doi.org/10.1086/513477 PMID: 17436249; PubMed Central PMCID: PMC1852743.

50. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nature genetics. 2010; 42(4):348–54. https://doi.org/10.1038/ng.548 PMID: 20208533; PubMed Central PMCID: PMC3092069.

51. Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, et al. Mixed linear model approach adapted for genome-wide association studies. Nature genetics. 2010; 42(4):355–60. https://doi.org/10.1038/ng.546 PMID: 20208535; PubMed Central PMCID: PMC2931336.

52. Hoffman GE. Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions. PloS one. 2013; 8(10):e75707. https://doi.org/10.1371/journal.pone.0075707 PMID: 24204578

53. Li Q, Yu K. Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. Genetic epidemiology. 2008; 32(3):215–26. https://doi.org/10.1002/gepi.20296 PMID: 18161052.

54. Liu L, Zhang D, Liu H, Arendt C. Robust methods for population stratification in genome wide association studies. BMC bioinformatics. 2013; 14:132. https://doi.org/10.1186/1471-2105-14-132 PMID: 23601181; PubMed Central PMCID: PMCPMC3637636.

55. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nature genetics. 2008; 40(6):695–701. https://doi.org/10.1038/ng.f.136 PMID: 18509313; PubMed Central PMCID: PMC2527050.

56. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant. . .or not? Hum Mol Genet. 2002; 11(20):2417–23. PMID: 12351577.

57. Teer JK, Mullikin JC. Exome sequencing: the sweet spot before whole genomes. Hum Mol Genet. 2010; 19(R2):R145–51. https://doi.org/10.1093/hmg/ddq333 PMID: 20705737; PubMed Central PMCID: PMC2953745.

58. Walsh T, King MC. Ten genes for inherited breast cancer. Cancer Cell. 2007; 11(2):103–5. https://doi.org/10.1016/j.ccr.2007.01.010 PMID: 17292821.

59. Ali AM, Dawson SJ, Blows FM, Provenzano E, Ellis IO, Baglietto L, et al. Comparison of methods for handling missing data on immunohistochemical markers in survival analysis of breast cancer. British journal of cancer. 2011; 104(4):693–9. https://doi.org/10.1038/sj.bjc.6606078 PMID: 21266980; PubMed Central PMCID: PMCPMC3049587.

60. Dahl A, Iotchkova V, Baud A, Johansson A, Gyllensten U, Soranzo N, et al. A multiple-phenotype imputation method for genetic studies. Nat Genet. 2016; 48(4):466–72. https://doi.org/10.1038/ng.3513 PMID: 26901065; PubMed Central PMCID: PMCPMC4817234.

61. De Silva AP, Moreno-Betancur M, De Livera AM, Lee KJ, Simpson JA. A comparison of multiple imputation methods for handling missing values in longitudinal data in the presence of a time-varying covariate with a non-linear association with time: a simulation study. BMC medical research methodology. 2017;

17(1):114. https://doi.org/10.1186/s12874-017-0372-y PMID: 28743256; PubMed Central PMCID: PMCPMC5526258.

62. Schafer JL. Analysis of incomplete multivariate data: CRC press; 1997.

63. Carlin J. Multiple imputation: a perspective and historical overview. Handbook of Missing Data. 2015.

64. Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey methodology. 2001; 27 (1):85–96.

65. Van Buuren S, Brand JP, Groothuis-Oudshoorn C, Rubin DB. Fully conditional specification in multivariate imputation. Journal of statistical computation and simulation. 2006; 76(12):1049–64.

66. Carpenter J, Kenward M. Multiple imputation and its application: John Wiley & Sons; 2012.