PLOS ONE

RESEARCH ARTICLE

Testing an optimally weighted combination of common and/or rare variants with multiple traits

Zhenchuan Wang[1], Qiuying Sha[1], Shurong Fang[2], Kui Zhang[1], Shuanglin Zhang[1]*

1 Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan, United States of America, **2** Department of Mathematics and Computer Science, John Carroll University, University Heights, Ohio, United States of America

* shuzhang@mtu.edu

OPEN ACCESS

Citation: Wang Z, Sha Q, Fang S, Zhang K, Zhang S (2018) Testing an optimally weighted combination of common and/or rare variants with multiple traits. PLoS ONE 13(7): e0201186. https://doi.org/10.1371/journal.pone.0201186

Editor: Qizhai Li, University of the Chinese Academy of Sciences, CHINA

Received: March 25, 2018

Accepted: July 10, 2018

Published: July 26, 2018

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

Funding: Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number R15HG008209 to QS. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The Genetic Analysis workshops are supported by NIH grant R01 GM031575 from the National

Abstract

Recently, joint analysis of multiple traits has become popular because it can increase statistical power to identify genetic variants associated with complex diseases. In addition, there is increasing evidence indicating that pleiotropy is a widespread phenomenon in complex diseases. Currently, most of existing methods test the association between multiple traits and a single genetic variant. However, these methods by analyzing one variant at a time may not be ideal for rare variant association studies because of the allelic heterogeneity as well as the extreme rarity of rare variants. In this article, we developed a statistical method by testing an optimally weighted combination of variants with multiple traits (TOWmuT) to test the association between multiple traits and a weighted combination of variants (rare and/or common) in a genomic region. TOWmuT is robust to the directions of effects of causal variants and is applicable to different types of traits. Using extensive simulation studies, we compared the performance of TOWmuT with the following five existing methods: gene association with multiple traits (GAMuT), multiple sequence kernel association test (MSKAT), adaptive weighting reverse regression (AWRR), single-TOW, and MANOVA. Our results showed that, in all of the simulation scenarios, TOWmuT has correct type I error rates and is consistently more powerful than the other five tests. We also illustrated the usefulness of TOWmuT by analyzing a whole-genome genotyping data from a lung function study.

Introductions

Many large cohort studies collected many correlated traits that can reflect underlying mechanism of complex diseases. For example, the UK10K cohort study collected 64 correlated phenotypic traits [1]. Usually, complex diseases are characterized by multiple endophenotypes. For example, hypertension can be characterized by systolic and diastolic blood pressure [2]; metabolic syndrome is evaluated by four component traits: high-density lipoprotein (HDL) cholesterol, plasma glucose and Type 2 diabetes, abdominal obesity, and diastolic blood

Institute of General Medical Sciences. Preparation of the Genetic Analysis Workshop 17 Simulated Exome Data Set was supported in part by NIH R01 MH059490 and used sequencing data from the 1000 Genomes Project (www.1000genomes.org). 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. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

pressure [3]; and schizophrenia can be diagnosed by eight neurocognitive domains [4]. Multiple correlated traits can be influenced by a gene simultaneously. Therefore, by joint analysis of multiple traits, we can not only gain more statistical power to detect pleiotropic variants [5–12], but also can better understand the genetic architecture of the disease of interest [13].

Several statistical methods have been developed to test the association between multiple traits and a single common variant. These methods can be roughly divided into three groups: dimension reduction methods [10, 13–15], regression methods [16–18], and combining test statistics from univariate analysis [9, 19–23]. However, due to the allelic heterogeneity and the extreme rarity of rare variants, the methods by analyzing one variant at a time for common variant association studies may not be ideal for rare variant association studies [24]. Recent genetic association studies show that complex diseases are affected by both common and rare variants [25–31]. Next-generation sequencing technology allows sequencing of the whole genome of large number of individuals, and makes rare variant association studies viable [32, 33]. Currently, statistical methods for rare variant association studies with a single trait have been developed. These methods summarize genotype information from multiple rare variants and can be divided into three groups: burden tests [24, 34–37], quadratic tests [38–41], and combined tests [42–45].

As we pointed out above, it is essential to develop statistical methods to test the association between multiple traits and multiple variants (common and/or rare variants). Very recently, a few statistical methods for this purpose are appeared [11, 46–50]. Casale et al. [47] proposed a set-based association test based on the linear mixed-model. This method enables jointly analyzing multiple correlated traits in rare variant association studies while accounting for population structure and relatedness. Wang et al. [11] proposed a multivariate functional linear model approach to test association between multiple traits and rare variants in a genomic region. In this approach, the genetic effects of variants are treated as smooth functions of genomic positions of these variants. Gene association with multiple traits (GAMuT) proposed by Broadaway et al. [46] provide a nonparametric test of independence between a set of traits and a set of genetic variants. This method compares the similarities of multiple traits with the similarities of genotypes at variants in a genomic region. Multivariate Rare-Variant Association Test (MURAT) proposed by Sun et al. [48] tests association between multiple correlated quantitative traits and a set of rare variants based on a linear mixed model. This method assumes that the effects of the variants follow a multivariate normal distribution with a zero mean and a specific covariance structure. Wu and Pankow [50] extended the commonly used sequence kernel association test (SKAT) [40] for a single trait to multiple traits and proposed multiple sequence kernel association test (MSKAT). Wang et al. [11] proposed an adaptive weighting reverse regression (AWRR) method. This method uses the score test based on the reverse regression, in which the summation of adaptively weighted genotypes is treated as the response variable and multiple traits are treated as independent variables.

In this article, we developed a new statistical method by testing an optimally weighted combination of variants with multiple traits (TOWmuT) to test the association between multiple traits and a weighted combination of variants (rare and/or common) in a genomic region. TOWmuT is based on the score test under a linear model, in which the weighted combination of variants is treated as the response variable and multiple traits including covariates are treated as independent variables. The statistic of TOWmuT is the maximum of the score test statistic over weights. The weights at which the score test statistic reaches its maximum are called the optimal weights. TOWmuT is applicable to different types of traits and can include covariates. Using extensive simulation studies, we compared the performance of TOWmuT with single-TOW [39], GAMuT [46], MSKAT [50], AWRR [11] and MANOVA [7]. Our results showed that, in all the simulation scenarios, TOWmuT is either the most powerful test

or comparable to the most powerful test among the six tests. We also illustrated the usefulness of TOWmuT by analyzing a real whole-genome genotyping data from a lung function study.

Methods

We consider a sample with n unrelated individuals. Each individual has K potentially correlated quantitative or qualitative traits (1 for cases and 0 for controls for a qualitative trait) and has been genotyped at M variants in a genomic region. Let y^*_{ik} denote the k^{th} trait value of the i^{th} individual and x^*_{im} denote the genotype score of the i^{th} individual at the m^{th} variant, where x^*_{im} is the number of minor alleles that the i^{th} individual carries at the m^{th} variant. We first centralize y^*_{ik} and x^*_{im} as $y_{ik} = y^*_{ik} - \bar{y}_k$ and $x_{im} = x^*_{im} - \bar{x}_m$, where $\bar{y}_k = \frac{1}{n}\sum_{i=1}^{n} y^*_{ik}$ and $\bar{x}_m = \frac{1}{n}\sum_{i=1}^{n} x^*_{im}$. Let $Y_i = (y_{i1},\ldots,y_{iK})^T$, $X_i = (x_{i1},\ldots,x_{iM})^T$, $Y = (Y_1,\ldots,Y_n)^T$, and $X = (X_1,\ldots,X_n)^T$. For the i^{th} individual, we consider a linear combination of the variants $x_i = \sum_{m=1}^{M} w_m x_{im}$, where $w = (w_1,\ldots,w_M)^T$ are weights and their values will be decided later.

Without covariates

We first describe our method without covariates. Consider the linear model

$$x_i = \beta_1 y_{i1} + \cdots + \beta_K y_{iK} + \varepsilon_i. \quad (1)$$

The score test statistic to test the null hypothesis $H_0: \beta_1 = \cdots = \beta_K = 0$ is given by

$$T_{score} = U^T V^{-1} U/\sigma^2, \quad (2)$$

where $U = \sum_{i=1}^{n} x_i Y_i = Y^T X w$, $V = \sum_{i=1}^{n} Y_i Y_i^T = Y^T Y$, and $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{n} w^T X^T X w$.

To simplify the computation of Eq (2), we replace $X^T X/n$ with the diagonal of $X^T X/n$ and let $A = diag(X^T X/n)$. This simplification was also used in the past by Pan [51] and Sha et al. [39]. Then σ^2 becomes $\sigma_0^2 = w^T A w$ and T_{score} becomes $T^0_{score}(w) = \frac{w^T X^T Y (Y^T Y)^{-1} Y^T X w}{w^T A w}$. We define the test statistic of TOWmuT as

$$T_{TOWmuT} = \max_w T^0_{score}(w). \quad (3)$$

Let $W = A^{1/2} w$, then $T_{TOWmuT} = \max_w T^0_{score}(w) = \lambda_{max}(A^{-1/2} X^T Y (Y^T Y)^{-1} Y^T X A^{-1/2})$, where $\lambda_{\max}(\bullet)$ indicates the largest eigenvalue of a matrix. Let W^0 denote the eigenvector of $A^{-1/2} X^T Y (Y^T Y)^{-1} Y^T X A^{-1/2}$ corresponding to the largest eigenvalue, then $w^0 = A^{-1/2} W^0$ is the optimal weights. Actually, we do not need to calculate w^0 in order to calculate T_{TOWmuT}. If we let $C = XA^{-1}X^T$, then

$$T_{TOWmuT} = \lambda_{max}(A^{-1/2} X^T Y (Y^T Y)^{-1} Y^T X A^{-1/2}) = \lambda_{max}((Y^T Y)^{-1} Y^T C Y). \quad (4)$$

We use a permutation test to evaluate the p-value of T_{TOWmuT}. In details, we randomly shuffle the traits in each permutation. Note that C and $(Y^T Y)^{-1}$ do not change in each permutation. Suppose that we perform B times of permutations. Let $T^{(b)}_{TOWmuT}$ denote the value of T_{TOWmuT} based on the b^{th} permuted data, where $b = 0$ represents the original data. Then, the p-value of T_{TOWmuT} is given by

$$\frac{\#\{b : T^{(b)}_{TOWmuT} \geq T^{(0)}_{TOWmuT} \text{ for } b = 1,\ldots,B\}}{B}. \quad (5)$$

With covariates

Assume that there are p covariates and $z_{i1},\ldots z_{ip}$ denote the p covariates of the i^{th} individual. Consider the linear model

$$x_i = \alpha_0 + \alpha_1 z_{i1} + \cdots + \alpha_p z_{ip} + \beta_1 y_{i1} + \cdots + \beta_K y_{iK} + \varepsilon_i. \quad (6)$$

In the appendix, we showed that under model (6), the score test statistic with covariates to test the null hypothesis H_0:$\beta_1 = \cdots = \beta_K = 0$ is given by

$$T^c_{score} = \tilde{U}^T \tilde{V}^{-1} \tilde{U}/\tilde{\sigma}^2, \quad (7)$$

where $\tilde{U} = \tilde{Y}^T \tilde{X} w$, $\tilde{V} = \tilde{Y}^T \tilde{Y}$, $\tilde{\sigma}^2 = \frac{1}{n} w^T \tilde{X}^T \tilde{X} w$, $\tilde{X} = (\tilde{x}_{im})$, $\tilde{Y} = (\tilde{y}_{ik})$, $\tilde{y}_{ik}$ and $\tilde{x}_{im}$ denote the residuals of y_{ik} and x_{im} under

$$y_{ik} = \alpha_{0k} + \alpha_{1k} z_{i1} + \ldots + \alpha_{pk} z_{ip} + \varepsilon_{ik} \text{ and } x_{im} = \alpha_{0m} + \alpha_{1m} z_{i1} + \ldots + \alpha_{pm} z_{ip} + \tau_{im}. \quad (8)$$

We can see the score test statistic with covariates

$$T^c_{score} = T_{score}|_{y_{ik}=\tilde{y}_{ik}, x_{im}=\tilde{x}_{im}}. \quad (9)$$

That is, replacing y_{ik} and x_{im} by their residuals $\tilde{y}_{ik}$ and $\tilde{x}_{im}$ in the score test statistic without covariates T_{score}, it becomes the score test statistic with covariates T^c_{score}.

Therefore, we define TOWmuT statistic with covariates as

$$T^c_{TOWmuT} = T_{TOWmuT}|_{y_{ik}=\tilde{y}_{ik}, x_{im}=\tilde{x}_{im}}. \quad (10)$$

In summary, to apply TOWmuT with covariates, we adjust both trait value y_{ik} and genotypic score x_{im} for the covariates by applying linear regressions in (8) and apply TOWmuT without covariates to the residuals $\tilde{y}_{ik}$ and $\tilde{x}_{im}$.

Comparison of methods

We compare the performance of our proposed method with the following methods: Multivariate Analysis of Variance (MANOVA) [9], MSKAT [50], GAMuT [46], AWRR [11] and single-TOW [39]. In the following, we briefly introduce each of those methods using the notations in the method section.

MANOVA: Consider a multivariate multiple linear regression model: $Y = X\beta + \varepsilon$, where Y denotes the $n \times K$ matrix of phenotypes; X denotes the $n \times M$ matrix of genotypes; β is a $M \times K$ matrix of coefficients; ε is the $n \times K$ matrix of random errors with each row of ε to be i.i.d. $MVN(0,\Sigma)$, where Σ is the covariance matrix of ε. To test H_0:$\beta = 0$, the likelihood ratio test is equivalent to the Wilk's Lambda test statistic of MANOVA, that is, $-2\log\Lambda = 2(l(\hat{\beta}, \hat{\Sigma}) - l(0, \hat{\Sigma}_0)) = n\log\frac{|\hat{\Sigma}_0|}{|\hat{\Sigma}|} = -n\log\left(\frac{|E|}{|E+H|}\right)$. Here Λ denote the ratio of the likelihood function under H_0 to the likelihood function under H_1, $l(\beta,\Sigma)$ is the log-likelihood function, $H = \hat{\beta}^T (X^T X)\hat{\beta}$ and $E = Y^T Y - \hat{\beta}^T (X^T X)\hat{\beta}$, where $\hat{\beta} = (X^T X)^{-1} X^T Y$ is the maximum likelihood estimator (MLE) of β, and $|\bullet|$ denotes the determinant of a matrix. The test statistic has an asymptotic χ^2_K distribution.

MSKAT: MSKAT extends the commonly used SKAT [40] for single trait analysis to test for the joint association of rare variant set with multiple continuous traits.

GAMuT: GAMuT compares the similarity in multivariate phenotypes to the similarity in rare-variant genotypes in a genomic region by a machine-learning framework called kernel distance covariance.

AWRR: by collapsing genotypes using adaptive weights, AWRR uses the score test to test association based on the reverse regression, in which collapsed genotypes are treated as the response variable and multiple traits are treated as independent variables.

Single-TOW: Let T^k_{TOW} denote the test statistic of TOW to test the association between the kth trait and the genotypes at the variants in a genomic region. The test statistic of single-TOW is given by $T_{single-TOW} = \min_{1\le k\le K} p_k$, where p_k is the p-value of T^k_{TOW} for $k = 1,\ldots,K$. The p-value of $T_{single-TOW}$ is estimated using a permutation procedure.

Simulations

In our simulation studies, we use the empirical Mini-Exome genotype data provided by the genetic analysis workshop 17 (GAW17) to generate genotypes. This dataset contains genotypes of 697 unrelated individuals on 3205 genes. Same as the simulation studies in Sha et al. [39] and Fang et al. [52], we choose four genes in the empirical Mini-Exome genotype data. These four genes are ELAVL4 (gene1), MSH4 (gene2), PDE4B (gene3), and ADAMTS4 (gene4). Each gene contains 10, 20, 30, and 40 variants, respectively. Then, we merge the four genes to form a super gene (Sgene) with 100 variants. We generate genotypes based on the genotypes of 697 individuals in the Sgene since the distribution of the minor allele frequencies (MAFs) in the Sgene are similar to the distribution of MAFs in all of the 3205 genes (Figure A in S1 File). To generate a qualitative trait, we use a liability threshold model based on a quantitative trait [44]. An individual is classified as affected if the individual's trait is at least one standard deviation larger than the mean of the trait. This leads to a prevalence of 16% for the simulated disease in the general population. In the following, we only describe how to generate a quantitative trait.

We assume that all causal variants are rare (MAF < 0.01). We randomly choose n_c rare variants as causal variants, where n_c is determined by the percentage of causal variants among rare variants. We use n_r and n_p to denote the number of risk rare variants and protective rare variants, respectively, where $n_r + n_p = n_c$. Let x^r_{qi} and x^p_{ji} denote the genotypic scores of the q^{th} risk rare variant and the j^{th} protective rare variant for the i^{th} individual, respectively. We assume that genotypes impact on L traits. Let h and h_l denote the heritability of all the n_c rare causal variants for the L traits and the l^{th} trait among the L traits, respectively. We generate L random numbers $t_1,\ldots,t_L$ from a uniform distribution between 0 and 1. Then, the heritability of l^{th} trait among the L traits is $h_l = ht_l \Big/ \sum_{l=1}^{L} t_l$. Given the heritability of the l^{th} trait h_l, we generate n_c random numbers $r_1,\ldots,r_{n_c}$ from a uniform distribution between 0 and 1. The heritability of the m^{th} causal variant for the l^{th} trait is given by $h_l^{(m)} = h_l r_m \Big/ \sum_{j=1}^{n_c} r_j$.

In our simulation studies, we consider two covariates Z_1 and Z_2, where Z_1 is a continuous covariate generated from a standard normal distribution, and Z_2 is a binary covariate taking values 0 and 1 with a probability of 0.5. We generate K traits by considering the factor model [10, 13, 21]

$$y = (0.5Z_1 + 0.5Z_2)e + (\lambda_1,\ldots,\lambda_K)^T + c\gamma f + \sqrt{1-c^2}\times\varepsilon, \tag{11}$$

where $y = (y_1,\ldots,y_K)^T$; $e = (1,\ldots,1)^T$, $\lambda = (\lambda_1,\ldots,\lambda_K)$ is the vector involved genotypes; $f = (f_1,\ldots, f_R)^T \sim MVN(0,\Sigma)$, $\Sigma = (1-\rho)I + \rho A$, A is a matrix with elements of 1, I is the identity matrix, and ρ is the correlation between f_i and f_j; R is the number of factors; γ is a K by R matrix; c is a constant number; $\varepsilon = (\varepsilon_1,\ldots,\varepsilon_K)^T$ is a vector of residuals; and $\varepsilon_1,\ldots,\varepsilon_K$ are independent, $\varepsilon_k \sim N(0,1)$ for $k = 1,\ldots,K$.

As in Wang et al. [10], we consider the following six models with different number of factors and different number of traits affected by genotypes. In these models, the within-factor correlation is c^2 and the between-factor correlation is $\rho_1 = \rho c^2$.

Model 1: There is only one factor and genotypes impact on 6 traits with the same effect size. This is equivalent to set $R = 1$ and $\boldsymbol{\gamma} = (1,\ldots,1)^T$. In details,

$$y_k = \begin{cases} 0.5Z_1 + 0.5Z_2 + \sum_{q=1}^{n_r}\beta_{kq}^r x_q^r - \sum_{j=1}^{n_p}\beta_{kj}^p x_j^p + cf_1 + \sqrt{1-c^2}\times\varepsilon_k, & 1 \le k \le 6 \\ 0.5Z_1 + 0.5Z_2 + cf_1 + \sqrt{1-c^2}\times\varepsilon_k, & k > 6 \end{cases}. \quad (12)$$

Model 2: There are five factors and genotypes impact on 6 traits. We set $R = 5$ and $\boldsymbol{\gamma} = diag(D_1, D_2, D_3, D_4, D_5)$, where $D_i = \left(\underbrace{1,\ldots,1}_{K/5}\right)^T$ for $i = 1,\ldots,5$. In details,

$$y_k = \begin{cases} 0.5Z_1 + 0.5Z_2 + \sum_{q=1}^{n_r}\beta_{kq}^r x_q^r - \sum_{j=1}^{n_p}\beta_{kj}^p x_j^p + cf_{[(k-1)/2]+1} + \sqrt{1-c^2}\times\varepsilon_k, & 1 \le k \le 6 \\ 0.5Z_1 + 0.5Z_2 + cf_{[(k-1)/2]+1} + \sqrt{1-c^2}\times\varepsilon_k, & k > 6 \end{cases}. \quad (13)$$

Model 3: There are two factors and genotypes impact on 6 traits. That is, $R = 2$ and $\boldsymbol{\gamma} = diag(D_1, D_2)$, where $D_i = \left(\underbrace{1,\ldots,1}_{K/2}\right)^T$ for $i = 1,2$. In details,

$$y_k = \begin{cases} 0.5Z_1 + 0.5Z_2 + \sum_{q=1}^{n_r}\beta_{kq}^r x_q^r - \sum_{j=1}^{n_p}\beta_{kj}^p x_j^p + cf_{[(k-1)/5]+1} + \sqrt{1-c^2}\times\varepsilon_k, & 1 \le k \le 6 \\ 0.5Z_1 + 0.5Z_2 + cf_{[(k-1)/5]+1} + \sqrt{1-c^2}\times\varepsilon_k, & k > 6 \end{cases}. \quad (14)$$

Model 4: There are five factors and genotypes impact on one trait. That is, $R = 5$ and $\boldsymbol{\gamma} = diag(D_1, D_2, D_3, D_4, D_5)$, where $D_i = \left(\underbrace{1,\ldots,1}_{K/5}\right)^T$ for $i = 1,\ldots,5$. In details,

$$y_k = \begin{cases} 0.5Z_1 + 0.5Z_2 + \sum_{q=1}^{n_r}\beta_{kq}^r x_q^r - \sum_{j=1}^{n_p}\beta_{kj}^p x_j^p + cf_{[(k-1)/2]+1} + \sqrt{1-c^2}\times\varepsilon_k, & k = 1 \\ 0.5Z_1 + 0.5Z_2 + cf_{[(k-1)/2]+1} + \sqrt{1-c^2}\times\varepsilon_k, & k > 1 \end{cases}. \quad (15)$$

Model 5: There are only two factors and genotypes impact on one trait. That is, $R = 2$ and $\boldsymbol{\gamma} = diag(D_1, D_2)$, where $D_i = \left(\underbrace{1,\ldots,1}_{K/2}\right)^T$ for $i = 1,2$. In details,

$$y_k = \begin{cases} 0.5Z_1 + 0.5Z_2 + \sum_{q=1}^{n_r}\beta_{kq}^r x_q^r - \sum_{j=1}^{n_p}\beta_{kj}^p x_j^p + cf_{[(k-1)/5]+1} + \sqrt{1-c^2}\times\varepsilon_k, & k = 1 \\ 0.5Z_1 + 0.5Z_2 + cf_{[(k-1)/5]+1} + \sqrt{1-c^2}\times\varepsilon_k, & k > 1 \end{cases}. \quad (16)$$

Model 6: There is K factors and genotypes impact on 6 traits. That is, $R = K$, $\boldsymbol{\gamma} = I$, and $c = 1$. In details,

$$y_k = \begin{cases} 0.5Z_1 + 0.5Z_2 + \sum_{q=1}^{n_r} \beta_{kq}^r x_q^r - \sum_{j=1}^{n_p} \beta_{kj}^p x_j^p + cf_k + \sqrt{1-c^2} \times \varepsilon_k, & 1 \le k \le 6 \\ 0.5Z_1 + 0.5Z_2 + cf_k + \sqrt{1-c^2} \times \varepsilon_k, & k > 6 \end{cases}. \quad (17)$$

Results

To evaluate the type I error rates of the proposed test TOWmuT, we set $\lambda_k = 0$ for $k = 1,\ldots,K$ in all of the 6 models. We consider different models, different sample sizes, different significance levels, and different types of traits. In our simulations we consider 10 traits (K = 10). In each simulation scenario, we estimate the p-values of TOWmuT using 1000 permutations and evaluate the type I error rates of TOWmuT using 10,000 replicated samples. For 10,000 replicated samples, the 95% confidence interval (CI) for the estimated type I error rates of nominal level 0.05 is (0.046, 0.054) and the 95% CI at the nominal level of 0.01 is (0.008, 0.012). Tables 1 and 2 summarize the estimated type I error rates of TOWmuT. From these two tables, we can see that 70 out of 72 (greater than 95%) estimated type I error rates are within the 95% CIs and the two estimated type I error rates not within the 95% CIs (0.05555 and 0.01295) are very close to the bound of the corresponding 95% CI, which indicates that TOWmuT is valid.

For power comparisons, we consider different models, different types of traits, different percentages of protective variants, different values of heritability, different values of between-factor correlation, and different values of within-factor correlation. In each of the simulation scenarios, we estimate the p-values of TOWmuT, AWRR and single-TOW using 1,000 permutations and we estimate the p-values of MANOVA, GAMuT, and MSKAT using their asymptotic distributions. We evaluate the powers of all of the six tests using 1,000 replicated samples at a significance level of 0.05.

Fig 1 gives the power comparisons of the six tests (Single-TOW, MSKAT, AWRR, MANOVA, GAMuT, and TOWmuT) for the power as a function of the total heritability based on

Table 1. The estimated type I error rates of TOWmuT for 10 quantitative traits under each model with covariates.

	Sample Size			
	Model	**500**	**1000**	**2000**
α = 0.05	1	0.05365	0.0515	0.0515
	2	0.0521	0.0528	0.0504
	3	0.0513	0.0540	0.0503
	4	0.0514	0.0511	0.05
	5	0.05381	0.04825	0.05
	6	0.0482	0.0508	0.05325
α = 0.01	1	0.01165	0.0098	0.0117
	2	0.012	0.01015	0.0102
	3	0.01175	0.01075	0.0113
	4	0.01145	0.01075	0.0118
	5	0.01141	0.01095	0.0117
	6	0.0097	0.0105	0.01185

https://doi.org/10.1371/journal.pone.0201186.t001

Table 2. The estimated type I error rates of TOWmuT for the mixture of five quantitative traits and five qualitative traits under each model with covariates.

	Sample Size			
	Model	**500**	**1000**	**2000**
$\alpha = 0.05$	1	0.05365	0.05385	0.05005
	2	0.0511	0.0483	0.05115
	3	0.0508	0.05375	0.052
	4	0.0529	0.04915	0.0536
	5	0.054	0.05355	0.04825
	6	0.05555	0.0493	0.0529
$\alpha = 0.01$	1	0.0105	0.01295	0.00995
	2	0.0105	0.009	0.0097
	3	0.01145	0.0104	0.0101
	4	0.01065	0.00945	0.01165
	5	0.0118	0.0105	0.00875
	6	0.01195	0.00935	0.01105

https://doi.org/10.1371/journal.pone.0201186.t002

the six models for 10 quantitative traits. This figure shows that (1) TOWmuT is consistently the most powerful one among the six tests; (2) MANOVA is the second most powerful when genotypes impact on multiple traits (models 1–3 and 6) while AWRR is the second most powerful when genotypes impact on a single trait (models 4–5); (3) MSKAT is consistently less powerful than other multivariate tests probably because SKAT gives larger weights than that of TOW to only those variants with MAF in the range (0.01,0.035) and there are only 8% variants with MAF in the range (0.01,0.035) in Sgene which our simulations are based on; and (4) MSKAT and GAMuT have similar powers in all six models.

Fig 2 gives the power comparisons of the five tests (Single-TOW, AWRR, MSKAT, GAMuT, and TOWmuT) for the power as a function of the total heritability for the mixture of 5 quantitative traits and 5 qualitative traits. We only compare the powers of five tests because MANOVA has inflated type I error rate in this case. This figure shows that (1) TOWmuT is consistently the most powerful one among the five tests; (2) AWRR is second most powerful when genotypes impact on multiple traits (models 1–3 and 6) while MSKAT and GAMuT are second most powerful when genotypes impact on a single trait (models 4–5); (3) MSKAT and GAMuT have similar powers in all six models; and (4) single-TOW is consistently less powerful than other four multivariate tests because we keep correlations between traits similar to that in Fig 1 such that correlations between original quantitative traits are larger than that in Fig 1.

We also compare the powers of the six tests for the power as a function of the within-factor correlation for models 1–5 and between-factor correlation for model 6 for 10 quantitative traits (Figure B in S1 File). As shown in this figure, the power of single-TOW is robust to the between-factor correlation or the within-factor correlation since the minimum p-value-based approach is largely unaffected by the trait correlation [50]. However, with the increasing of the between-factor correlation or within-factor correlation, the power of other five tests essentially increases. Other patterns of the power comparisons are similar to those of in Fig 1.

Power comparisons of the six tests for the power as a function of the percentage of protective variants for 10 quantitative traits are given by Figure C in S1 File. This figure shows that the power of all six tests are robust to the percentage of protective variants, therefore, all of these methods are robust to the directions of the genetic effects. Other patterns of the power comparisons are similar to those of in Fig 1.

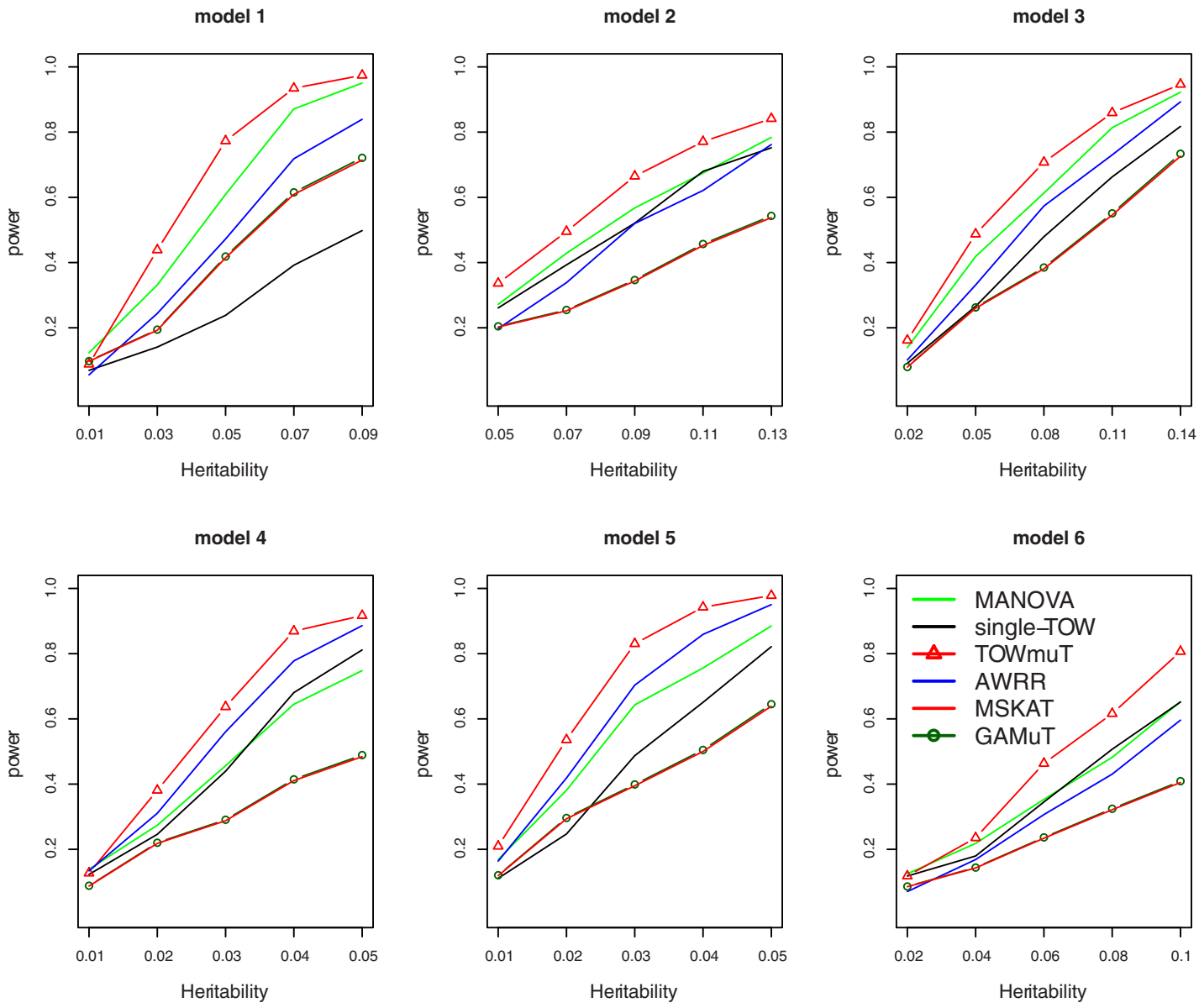

Fig 1. Power comparisons of the six tests (Single-TOW, MSKAT, AWRR, MANOVA, GAMuT and TOWmuT) for the power as a function of total heritability for 10 quantitative traits with covariates. The sample size is 1000. The between-factor correlation is 0.3 and the within-factor correlation is 0.7. The percentage of the causal variants is 0.2. All causal variants are risk variants.

https://doi.org/10.1371/journal.pone.0201186.g001

Application to the COPDGene

Chronic obstructive pulmonary disease (COPD) is a common disease in elderly patients that causes significant morbidity and mortality [53]. The Genetic Epidemiology of COPD Study (COPDGene) [54] was designed to identify genetic factors associated with COPD. In this COPDGene study, a total of more than 10,000 subjects have been enrolled including 2/3 non-Hispanic Whites (NHW) and 1/3 African-Americans (AA). In this analysis, we only include 5,430 NHW with no missing phenotypes. Each of the 5,430 NHW has been genotyped at 630,860 SNPs. Based on the literature studies of COPD [9, 55, 56], we chose BMI, Age, Pack-Years (PackYear) and Sex as covariates and selected seven quantitative COPD-related

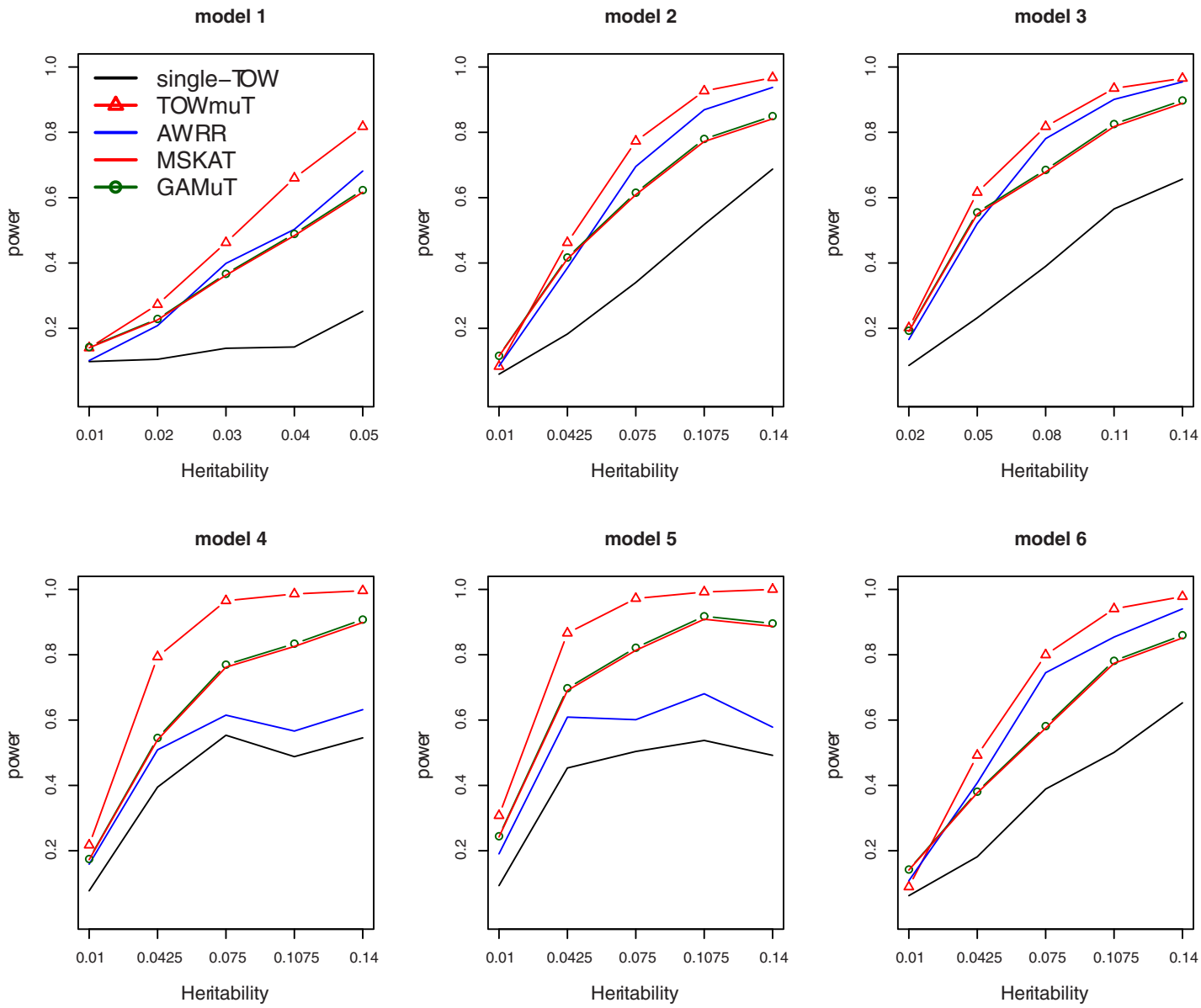

Fig 2. Power comparisons of the five tests (Single-TOW, AWRR, GAMuT, MSKAT and TOWmuT) for the power as a function of heritability for the mixture of half quantitative traits and half qualitative traits with covariates. The sample size is 1000. Covariance matrix of 10 traits is similar to that of 10 quantitative traits with between-factor correlation being 0.3 and the within-factor correlation being 0.7. The percentage of the causal variants is 0.2. All causal variants are risk variants.

https://doi.org/10.1371/journal.pone.0201186.g002

phenotypes. These seven phenotypes are FEV1 (% predicted FEV1), Emphysema (Emph), Emphysema Distribution (EmphDist), Gas Trapping (GasTrap), Airway Wall Area (Pi10), Exacerbation frequency (ExacerFreq), and Six-minute walk distance (6MWD) [9]. The correlation structure of the seven COPD-related phenotypes is given in Figure D in S1 File.

To evaluate the performance of our proposed method on a real data set, we applied six methods (TOWmuT, MANOVA, MSKAT, GAMuT, AWRR, and single-TOW) to the COPDGene of NHW population to test the association between each of 50-SNP blocks and the seven quantitative COPD-related phenotypes. To identify significant 50-SNP blocks associated with the phenotypes, we used Bonferroni correction to decide the significance level. The total number of 50-SNP blocks is 12617, therefore, the Bonferroni corrected significance level is 0.05/

Table 3. Significant blocks identified by at least one method (p-values less than 4×10^{-6}) and the corresponding p-values in the analysis of COPDGene.

CHR	POS1	POS2	Genes	TOWmuT	MANOVA	MSKAT	GAMuT	AWRR	Single-TOW
2	178000985	178419117	NFE2L2	0.20883	2.62E-06	0.02508	0.02505	0.25796	0.15468
4	145278837	145697040	HHIP	1.00E-07	7.71E-06	0.03992	0.03984	0	0.00085
10	26908475	27150093	PDSS1, ABI1	4.00E-06	0.04050	0.01242	0.01247	1.6E-05	0.02845
15	78593362	78825917	IREB2, AGPHD1	1.00E-07	0.00191	0.70349	0.70357	5.6E-06	0.23484
15	78826180	79006442	PSMA4, CHRNA5, CHRNA3, CHRNB4	2.90E-06	0.00037	0.06255	0.06252	0	0.37643
15	79006582	79267817	ADAMTS7	9.01E-05	4.78E-05	2.25E-06	6.42E-07	0.04849	0.01953

https://doi.org/10.1371/journal.pone.0201186.t003

$12617 \approx 4\times10^{-6}$. Table 3 summarized the significant blocks identified by at least one method. There were total six significant blocks in Table 3. All of the six blocks have been previously reported to be in association with COPD or lung functions [57–60]. PDSS1 and ABI1 are located between LOC107984176 and LOC105376467, which are Intergenic regions and contain the SNPs associated with pulmonary function [60, 61]. From Table 3, we can see that TOWmuT identified four blocks; AWRR identified two blocks; MANOVA, MSKAT and GAMuT identified one block; single-TOW did not identify any blocks. From these results, we can see that TOWmuT identified the most of significant 50-SNP blocks among the six methods, which is consistent with the results of our simulation studies.

Discussion

In this article, we developed TOWmuT to perform joint analysis of multiple traits in gene-based association studies. The motivations to develop this method are based on the following: (1) for complex diseases, multiple correlated traits are usually measured in genetic association studies; (2) there is increasing evidence demonstrating that pleiotropy is a widespread phenomenon in complex diseases [5]; and (3) there is a shortage of gene-based approaches for multiple traits. We used extensive simulation studies to compare the performance of TOWmuT with MANOVA, MSKAT, AWRR, GAMuT and Single-TOW. Our simulation results showed that TOWmuT has correct type I error rates and is consistently more powerful than other five methods we compared. Furthermore, the results from real data analysis showed that the proposed method has great potential in gene-based association study for complex diseases with multiple phenotypes such as COPD.

Recently, it has become a major focus of investigation to identify a small number of rare causal variants that contribute to complex diseases [62]. Several methods to pinpoint the causal variants have been developed for testing the association with a single trait. These methods include backward elimination (BE) method [63], hierarchical model method [63], and adaptive combination of p-values method [64]. To extend the TOWmuT method to identify a small number of causal variants which are associated with multiple traits, we can use the BE method. In each step, we remove one variant that has the smallest contribution to the association between multiple traits and the set of variants and then we evaluate the p-value for testing association between multiple traits and the remaining variants by TOWmuT. Causal variants are the set of variants corresponding to the smallest p-value.

The computation time required for running TOWmuT depends on the number of traits, the sample size, the number of permutations, and the number of variants in a genomic region. The running time of TOWmuT with 1000 permutations on a data set with 5000 individuals, seven traits, and 10 variants in a genomic region on a laptop with 4 Intel Cores @ 3.30GHz and 4 GB memory is about 0.14s. To perform real data analysis at a genome-wide level, we can first select genomic regions that show evidence of association based on a small number of permutations (e.g. 1,000), and then use a large number of permutations to test the selected regions.

Appendix

We use the same notations in the method section. Let $Y = (Y_1,\ldots,Y_n)^T$, $Z_i = (1z_{i1},\ldots,z_{ip})^T$, $Z = (Z_1,\ldots,Z_n)^T$, and $x = (x_1,\ldots,x_n)^T$. Under the linear model

$$x_i = \alpha^T Z_i + \beta^T Y_i + \varepsilon_i, \tag{18}$$

the log-likelihood (up to a constant) is given by

$$\log l = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(x - Z\alpha - Y\beta)^T(x - Z\alpha - Y\beta), \tag{19}$$

where $\alpha = (\alpha_0,\ldots,\alpha_p)^T$, $\beta = (\beta_1,\ldots,\beta_K)^T$, and $\varepsilon_1,\ldots,\varepsilon_n$ are independent and $\varepsilon_i \sim N(0,\sigma^2)$. Then,

$$\frac{\partial \log l}{\partial \beta} = \frac{1}{\sigma^2}(x - Z\alpha - Y\beta)^T Y,\ \ \frac{\partial \log l}{\partial \alpha} = \frac{1}{\sigma^2}(x - Z\alpha - Y\beta)^T Z, \tag{20}$$

$$\frac{\partial^2 \log l}{\partial \beta \beta^T} = -\frac{1}{\sigma^2}Y^T Y,\ \ \frac{\partial^2 \log l}{\partial \alpha \alpha^T} = -\frac{1}{\sigma^2}Z^T Z,\ \text{and}\ \ \frac{\partial^2 \log l}{\partial \alpha \beta^T} = -\frac{1}{\sigma^2}Z^T Y. \tag{21}$$

Let $\hat{\alpha}$ and $\hat{\sigma}^2$ denote the maximum likelihood estimates of α and σ^2 under null hypothesis H_0:$\beta = 0$. Then, $\hat{\alpha} = (Z^T Z)^{-1} Z^T x$ and $\hat{\sigma}^2 = \frac{1}{n}x^T(I - P)x = \frac{1}{n}w^T X^T(I - P)Xw$, where $P = Z(Z^T Z)^{-1}Z^T$. Let $\theta = (\alpha^T, \beta^T)^T$. The score and information matrix are $S = \frac{\partial \log l}{\partial \theta}|_{\alpha=\hat{\alpha},\beta=0} = \frac{1}{\hat{\sigma}^2}(0, U^T)^T$ and $I = -E\frac{\partial^2 \log l}{\partial \theta \theta^T}|_{\alpha=\hat{\alpha},\beta=0} = \frac{1}{\hat{\sigma}^2}\begin{pmatrix} Z^T Z & Z^T Y \\ Y^T Z & Y^T Y \end{pmatrix}$, where $U = Y^T(I–P)x = Y^T(I–P)Xw$. The score test statistic is given by

$$T^c_{score} = \frac{1}{\hat{\sigma}^2}U^T V^{-1} U, \tag{22}$$

where $V = Y^T(I–P)Y$. Note that $(I–P)^2 = I–P$. We have $U = Y^T(I - P)Xw = \tilde{Y}^T\tilde{X}w$, $X^T(I - P)X = \tilde{X}^T\tilde{X}$, $\hat{\sigma}^2 = \frac{1}{n}w^T X^T(I - P)Xw = \frac{1}{n}w^T\tilde{X}^T\tilde{X}w$, and $V = Y^T(I - P)Y = \tilde{Y}^T\tilde{Y}$, where $\tilde{X} = (\tilde{x}_{im})$ and $\tilde{x}_{im}$ is the residual of x_{im} under the linear regression model (8); $\tilde{Y} = (\tilde{y}_{ik})$ and $\tilde{y}_{ik}$ is the residual of y_{ik} under the linear regression model (8). Therefore,

$$T^c_{score} = T_{score}|_{y_{ik}=\tilde{y}_{ik}, x_{im}=\tilde{x}_{im}}. \tag{23}$$

Supporting information

S1 File. Supplementary information.
(PDF)

Acknowledgments

The Genetic Analysis workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences. Preparation of the Genetic Analysis Workshop 17 Simulated Exome Data Set was supported in part by NIH R01 MH059490 and used sequencing data from the 1000 Genomes Project (www.1000genomes.org).

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.

Author Contributions

Formal analysis: Zhenchuan Wang, Shurong Fang, Shuanglin Zhang.

Methodology: Qiuying Sha, Shuanglin Zhang.

Visualization: Shurong Fang.

Writing – original draft: Zhenchuan Wang, Shuanglin Zhang.

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

References

1. The UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015; 526(7571):82–90. https://doi.org/10.1038/nature14962 PMID: 26367797.
2. Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, Coin L, et al. Genome-wide association study identifies eight loci associated with blood pressure. Nature genetics. 2009; 41(6):666–76. https://doi.org/10.1038/ng.361 PMID: 19430483.
3. Zabaneh D, Balding DJ. A genome-wide association study of the metabolic syndrome in Indian Asian men. PloS one. 2010; 5(8):e11961. https://doi.org/10.1371/journal.pone.0011961 PMID: 20694148.
4. Gur RE, Nimgaonkar VL, Almasy L, Calkins ME, Ragland JD, Pogue-Geile MF, et al. Neurocognitive endophenotypes in a multiplex multigenerational family study of schizophrenia. Am J Psychiatry. 2007; 164(5):813–9. https://doi.org/10.1176/ajp.2007.164.5.813 PMID: 17475741.
5. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. Nat Rev Genet. 2013; 14(7):483–95. https://doi.org/10.1038/nrg3461 PMID: 23752797.
6. 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.
7. 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.
8. 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.
9. Liang X, Wang Z, Sha Q, Zhang S. An Adaptive Fisher's Combination Method for Joint Analysis of Multiple Phenotypes in Association Studies. Sci Rep. 2016; 6:34323. https://doi.org/10.1038/srep34323 PMID: 27694844.
10. 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.
11. 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.
12. Zhu H, Zhang S, Sha Q. Power Comparisons of Methods for Joint Association Analysis of Multiple Phenotypes. Hum Hered. 2015; 80(3):144–52. https://doi.org/10.1159/000446239 PMID: 27344597.
13. 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. Am J Hum Genet. 2014; 94(5):662–76. https://doi.org/10.1016/j.ajhg.2014.03.016 PMID: 24746957.
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. 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.
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.
17. 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.
18. 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.

19. O'Brien PC. Procedures for comparing samples with multiple endpoints. Biometrics. 1984; 40(4):1079–87. PMID: 6534410.

20. 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.

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

22. Kim J, Bai Y, Pan W. An Adaptive Association Test for Multiple Phenotypes with GWAS Summary Statistics. Genetic epidemiology. 2015. https://doi.org/10.1002/gepi.21931 PMID: 26493956.

23. 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. Am J Hum Genet. 2015; 96(1):21–36. https://doi.org/10.1016/j.ajhg.2014.11.011 PMID: 25500260.

24. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83(3):311–21. https://doi.org/10.1016/j.ajhg.2008.06.024 PMID: 18691683.

25. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 2008; 40(6):695–701. https://doi.org/10.1038/ng.f.136 PMID: 18509313.

26. 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.

27. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet. 2001; 69(1):124–37. https://doi.org/10.1086/321272 PMID: 11404818.

28. 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.

29. Stratton MR, Rahman N. The emerging landscape of breast cancer susceptibility. Nat Genet. 2008; 40(1):17–22. https://doi.org/10.1038/ng.2007.53 PMID: 18163131.

30. 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.

31. 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.

32. Andres AM, Clark AG, Shimmin L, Boerwinkle E, Sing CF, Hixson JE. Understanding the accuracy of statistical haplotype inference with sequence data of known phase. Genetic epidemiology. 2007; 31(7):659–71. https://doi.org/10.1002/gepi.20185 PMID: 17922479.

33. Metzker ML. Sequencing technologies—the next generation. Nat Rev Genet. 2010; 11(1):31–46. https://doi.org/10.1038/nrg2626 PMID: 19997069.

34. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009; 5(2):e1000384. https://doi.org/10.1371/journal.pgen.1000384 PMID: 19214210.

35. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res. 2007; 615(1–2):28–56. https://doi.org/10.1016/j.mrfmmm.2006.09.003 PMID: 17101154.

36. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, et al. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010; 86(6):832–8. https://doi.org/10.1016/j.ajhg.2010.04.005 PMID: 20471002.

37. Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zollner S. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. Am J Hum Genet. 2010; 87(5):604–17. https://doi.org/10.1016/j.ajhg.2010.10.012 PMID: 21070896.

38. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an unusual distribution of rare variants. PLoS Genet. 2011; 7(3):e1001322. https://doi.org/10.1371/journal.pgen.1001322 PMID: 21408211.

39. 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.

40. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89(1):82–93. https://doi.org/10.1016/j.ajhg.2011.05.029 PMID: 21737059.

41. Yang X, Wang S, Zhang S, Sha Q. Detecting association of rare and common variants based on cross-validation prediction error. Genetic epidemiology. 2017; 41(3):233–43. https://doi.org/10.1002/gepi.22034 PMID: 28176359.

42. Derkach A, Lawless JF, Sun L. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. Genetic epidemiology. 2013; 37(1):110–21. https://doi.org/10.1002/gepi.21689 PMID: 23032573.

43. Lee S, Teslovich TM, Boehnke M, Lin X. General framework for meta-analysis of rare variants in sequencing association studies. Am J Hum Genet. 2013; 93(1):42–53. https://doi.org/10.1016/j.ajhg.2013.05.010 PMID: 23768515.

44. Sha Q, Zhang S. A rare variant association test based on combinations of single-variant tests. Genetic epidemiology. 2014; 38(6):494–501. https://doi.org/10.1002/gepi.21834 PMID: 25065727.

45. Greco B, Hainline A, Arbet J, Grinde K, Benitez A, Tintle N. A general approach for combining diverse rare variant association tests provides improved robustness across a wider range of genetic architectures. Eur J Hum Genet. 2015. https://doi.org/10.1038/ejhg.2015.194 PMID: 26508571

46. Broadaway KA, Cutler DJ, Duncan R, Moore JL, Ware EB, Jhun MA, et al. A Statistical Approach for Testing Cross-Phenotype Effects of Rare Variants. Am J Hum Genet. 2016; 98(3):525–40. https://doi.org/10.1016/j.ajhg.2016.01.017 PMID: 26942286.

47. 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.

48. Sun J, Oualkacha K, Forgetta V, Zheng HF, 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. Eur J Hum Genet. 2016. https://doi.org/10.1038/ejhg.2016.8 PMID: 26860061.

49. Wang Y, Liu A, Mills JL, Boehnke M, Wilson AF, Bailey-Wilson JE, et al. Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. Genetic epidemiology. 2015; 39(4):259–75. https://doi.org/10.1002/gepi.21895 PMID: 25809955.

50. Wu B, Pankow JS. Sequence Kernel Association Test of Multiple Continuous Phenotypes. Genetic epidemiology. 2016; 40(2):91–100. https://doi.org/10.1002/gepi.21945 PMID: 26782911.

51. Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genetic epidemiology. 2009; 33(6):497–507. https://doi.org/10.1002/gepi.20402 PMID: 19170135.

52. Fang S, Zhang S, Sha Q. Detecting association of rare variants by testing an optimally weighted combination of variants for quantitative traits in general families. Annals of human genetics. 2013; 77(6):524–34. Epub 2013/08/24. https://doi.org/10.1111/ahg.12038 PMID: 23968488.

53. Nazir SA, Erbland ML. Chronic obstructive pulmonary disease: an update on diagnosis and management issues in older adults. Drugs Aging. 2009; 26(10):813–31. https://doi.org/10.2165/11316760-000000000-00000 PMID: 19761275.

54. 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.

55. Chu JH, Hersh CP, Castaldi PJ, Cho MH, Raby BA, Laird N, et al. Analyzing networks of phenotypes in complex diseases: methodology and applications in COPD. BMC Syst Biol. 2014; 8:78. https://doi.org/10.1186/1752-0509-8-78 PMID: 24964944.

56. Han MK, Kazerooni EA, Lynch DA, Liu LX, Murray S, Curtis JL, et al. Chronic obstructive pulmonary disease exacerbations in the COPDGene study: associated radiologic phenotypes. Radiology. 2011; 261(1):274–82. https://doi.org/10.1148/radiol.11110173 PMID: 21788524.

57. 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.

58. 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 Genet. 2009; 5(3):e1000421. https://doi.org/10.1371/journal.pgen.1000421 PMID: 19300482.

59. Figarska SM, Vonk JM, Boezen HM. NFE2L2 polymorphisms, mortality, and metabolism in the general population. Physiol Genomics. 2014; 46(12):411–7. https://doi.org/10.1152/physiolgenomics.00178.2013 PMID: 24790085.

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

61. Imboden M, Bouzigon E, Curjuric I, Ramasamy A, Kumar A, Hancock DB, et al. Genome-wide association study of lung function decline in adults with and without asthma. Journal of allergy and clinical immunology. 2012; 129(5):1218–28. https://doi.org/10.1016/j.jaci.2012.01.074 PMID: 22424883

62. Capanu M, Ionita-Laza I. Integrative analysis of functional genomic annotations and sequencing data to identify rare causal variants via hierarchical modeling. Front Genet. 2015; 6:17. https://doi.org/10.3389/fgene.2015.00176 PMID: 26005447.

63. Ionita-Laza I, Capanu M, De Rubeis S, McCallum K, Buxbaum JD. Identification of rare causal variants in sequence-based studies: methods and applications to VPS13B, a gene involved in Cohen syndrome and autism. PLoS Genet. 2014; 10(12):e1004729. https://doi.org/10.1371/journal.pgen.1004729 PMID: 25502226.

64. Lin WY. Beyond Rare-Variant Association Testing: Pinpointing Rare Causal Variants in Case-Control Sequencing Study. Sci Rep. 2016; 6:21824. https://doi.org/10.1038/srep21824 PMID: 26903168.