SCIENTIFIC REPORTS

OPEN

Joint Analysis of Multiple Phenotypes in Association Studies based on Cross-Validation Prediction Error

Received: 5 April 2018
Accepted: 19 November 2018
Published online: 31 January 2019

Xinlan Yang, Shuanglin Zhang & Qiuying Sha

In genome-wide association studies (GWAS), joint analysis of multiple phenotypes could have increased statistical power over analyzing each phenotype individually to identify genetic variants that are associated with complex diseases. With this motivation, several statistical methods that jointly analyze multiple phenotypes have been developed, such as O'Brien's method, Trait-based Association Test that uses Extended Simes procedure (TATES), multivariate analysis of variance (MANOVA), and joint model of multiple phenotypes (MultiPhen). However, the performance of these methods under a wide range of scenarios is not consistent: one test may be powerful in some situations, but not in the others. Thus, one challenge in joint analysis of multiple phenotypes is to construct a test that could maintain good performance across different scenarios. In this article, we develop a novel statistical method to test associations between a genetic variant and Multiple Phenotypes based on cross-validation Prediction Error (MultP-PE). Extensive simulations are conducted to evaluate the type I error rates and to compare the power performance of MultP-PE with various existing methods. The simulation studies show that MultP-PE controls type I error rates very well and has consistently higher power than the tests we compared in all simulation scenarios. We conclude with the recommendation for the use of MultP-PE for its good performance in association studies with multiple phenotypes.

Traditionally, genome-wide association studies (GWAS) have performed on individual phenotype. In spite of the success of GWAS in identifying thousands of associations between genetic variants and complex diseases, these identified variants only contribute to a small proportion of the phenotypic variation. In the study of a complex disease, several correlated phenotypes are usually measured for a disorder or its risk factors[1], therefore, by jointly analyzing multiple correlated phenotypes, we may increase statistical power to detect causal variants with weak genetic effects on complex diseases.

One method to use multiple phenotypes in association studies is to analyze each phenotype separately as the standard univariate association test and then aggregate the results. This approach will have a loss in power due to the penalties from the multiple testing[1,2] and the ignorance of the correlation structure among phenotypes[3,4]. Thus, multiple-phenotype association study that uses multiple phenotypes simultaneously has become popular.

Several methods to detect association using multiple phenotypes simultaneously have been introduced in recent years. For example, O'Brien method (OB) is proposed to combines test statistics obtained from association test for each individual phenotype[5]. OB is the most powerful test when the genetic effects are homogeneous and loses power when genetic effects are heterogeneous, especially when genetic effects have opposite directions[1,6]. van der Sluis *et al.*[7] proposed a trait-based association test using an extended Simes procedure (TATES) that conducts association test for each phenotype and then combines the univariate p-values while correcting for the correlation between p-values. The canonical correlation analysis (CCA) conducts the linear combination of phenotypes that explain the largest amount of correlation between a genetic variant and phenotypes[8]. One could also use multivariate analysis of variance (MANOVA) in regression to study multiple phenotypes[9]. MANOVA is equivalent to CCA when canonical correlation analysis is applied to a single variant[10]. MultiPhen proposed by O'Reilly *et al.*[2] can be used to detect the association between one variant and multiple phenotypes by reversing response and predictors via a proportional odds regression model. When a small number of phenotypes are

Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan, United States of America. Correspondence and requests for materials should be addressed to Q.S. (email: qsha@mtu.edu)

included, MultiPhen and MANOVA lead to similar performance[6,11]. MANOVA and CCA require the assumption of normality of multiple phenotypes, while MultiPhen has no inflated type I error rates on non-normal phenotypes[2]. Some other variable reduction methods have also been proposed to test for the association between a genetic variant and the linear combination of multiple phenotypes rather than the original phenotypes[12–14]. For example, principal component of phenotypes (PCP) that maximizes the phenotype variation is the most popular dimension reduction method[13]. Based on PCP, Klei *et al.*[12] developed principal component of heritability (PCH) by maximizing the heritability among all linear combination of phenotypes. Recently, Turley *et al.*[15] introduced the Multi-Trait Analysis of GWAS (MTAG) for joint analysis of multiple phenotypes. MTAG can be applied to GWAS summary statistics from an arbitrary number of phenotypes without access to individual-level data.

Although there are many proposed methods for joint analysis of multiple phenotypes, the performance of these methods under a wide range of scenarios is not consistent[6]: one test may be powerful in some situations, but not in the others. Thus, one challenge in multiple phenotype analysis is to construct a test that could maintain good performance across different scenarios. In this article, we develop a novel statistical method to test the association between a genetic variant and Multiple Phenotypes based on cross-validation Prediction Error (MultP-PE). Extensive simulation studies are conducted to evaluate the type I error rates and to compare the power performance of MultP-PE with various existing methods. Our simulation studies show that MultP-PE controls the type I error rates very well and has consistently higher power than other methods we compared in all simulation scenarios.

Method

We consider a sample with n unrelated individuals. Each individual has K (potentially correlated) phenotypes and has been genotyped at a variant of interest. Let y_{ik} denote the k^{th} phenotype value of the i^{th} individual and x_i denote the genotype score of the i^{th} individual, where $x_i \in \{0, 1, 2\}$ is the number of minor alleles that the i^{th} individual carries. We model the relationship between the multiple phenotypes and the genetic variant using an inverse linear regression model, in which the genotype at the variant of interest is the response variable and the multiple phenotypes are predictors. That is,

$$x_i = \beta_0 + \beta_1 y_{i1} + \ldots + \beta_K y_{iK} + \varepsilon_i. \quad (1)$$

We are not the first using an ordinal variable as response variable in a linear model. To correct for population stratification, Price *et al.*[16] used a qualitative phenotype or genotypes as response variables in linear models. To adjust the effects of covariates in rare variant association studies, Sha *et al.*[17] also used a qualitative phenotype or genotypes as response variables in linear models. To test the association between the K multiple phenotypes and the variant, we test the null hypothesis $H_0: \beta_1 = \cdots = \beta_K = 0$ under model (1).

Let $y_i = (1, y_{i1}, \ldots, y_{iK})^T$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_K)^T$, then the regression model in equation (1) can be written as $x_i = y_i^T \beta + \varepsilon_i$, $i = 1, 2, \ldots, n$. The ordinary linear square estimate of β is $\hat{\beta} = (Y^T Y)^{-1} Y^T x$, where $Y = (y_1, \ldots, y_n)^T$ and $x = (x_1, \ldots, x_n)^T$. When multiple phenotypes are highly correlated, the rank of matrix Y may be less than K, then the inverse of $Y^T Y$ may not exist, which results in that the ordinary linear square estimate of β may not be unique[18]. Since multiple phenotypes in a GWAS are usually highly correlated, we propose to use Ridge regression[19–24]. Ridge regression penalizes the size of the regression coefficients. The Ridge regression estimator of β is defined as the value of β that minimizes

$$\sum_i (x_i - y_i^T \beta)^2 + \lambda \sum_j \beta_j^2,$$

where λ ($\lambda \geq 0$) is a tuning parameter. The solution to the Ridge regression is given by $\hat{\beta}_\lambda = (Y^T Y + \lambda I)^{-1} Y^T x$. Here the estimator of β depends on λ and we use the subscript λ to indicate that the estimator of β is a function of λ.

Based on Ridge regression, we propose to use the leave-one-out cross validation (LOOCV) prediction error under model (1) as a test statistic. Let $\hat{x}_{-i}^{\lambda}$ denote the LOOCV predicted value (leave the i^{th} individual out) of x_i under model (1) with parameter λ in Ridge regression. Then, the statistic can be written as $T_\lambda = \sum_{i=1}^n (x_i - \hat{x}_{-i}^{\lambda})^2$. Note that T_λ is the LOOCV prediction error, thus low values of T_λ would imply significance. Let p_λ denote the p-value of T_λ (see next paragraph on how to calculate p_λ). We define the test statistic of Multiple Phenotypes based on Prediction Error (MultP-PE) as

$$T_{MultP-PE} = \min_\lambda p_\lambda. \quad (2)$$

We propose to use a grid search method in equation (2) to evaluate the minimization. We divide the interval $[0, \infty)$ into subintervals $0 \leq \lambda_1 < \cdots < \lambda_{M-1} < \lambda_M < \infty$. Then, $T_{MultP-PE} = \min_\lambda p_\lambda = \min_{1 \leq m \leq M} p_{\lambda_m}$. We use a permutation procedure to evaluate the p-value of $T_{MultP-PE}$. Intuitively, we need to use two layers of permutations to estimate p_{λ_m} and the overall p-value for the test statistic $T_{MultP-PE}$. For microarray data analysis, Ge *et al.*[25] proposed that one layer of permutation can be used to estimate p-values. We use the permutation procedure of Ge *et al.* to estimate p_{λ_m} and the overall p-value for the test statistic $T_{MultP-PE}$. In each permutation, we randomly shuffle the genotypes at the variant. Suppose that we perform B times of permutations. Let $T_{\lambda_m}^{(b)}$ denote the value of T_{λ_m} based on the b^{th} permuted data for $b = 0, 1, \ldots, B$ and $m = 1, \ldots, M$, and $p_{\lambda_m}^{(b)}$ denote the p-value of $T_{\lambda_m}^{(b)}$, where $b = 0$ represents the original data. Then, we can estimate $p_{\lambda_m}^{(b)}$ using $p_{\lambda_m}^{(b)} = \frac{\#\{d : T_{\lambda_m}^{(d)} < T_{\lambda_m}^{(b)} \text{ for } d = 1, \ldots, B\}}{B}$. Let

Sample Size	Number of Phenotypes	Significance Level	Model 1	Model 2	Model 3	Model 4
500	10	$\alpha = 0.01$	0.0103	0.0109	0.0112	0.0094
		$\alpha = 0.05$	0.0480	0.0512	0.0523	0.0532
	20	$\alpha = 0.01$	0.0116	0.0107	0.0114	0.0112
		$\alpha = 0.05$	0.0503	0.0499	0.0473	0.0515
	40	$\alpha = 0.01$	0.0112	0.0118	*0.0121*	0.0103
		$\alpha = 0.05$	0.0524	0.0515	0.0518	0.0541
1000	10	$\alpha = 0.01$	0.0108	0.099	0.0104	0.0116
		$\alpha = 0.05$	0.0535	0.0532	0.0514	0.0492
	20	$\alpha = 0.01$	0.0101	0.0095	0.0112	0.0083
		$\alpha = 0.05$	0.0500	0.0501	0.0524	0.0469
	40	$\alpha = 0.01$	0.0094	0.0116	0.0117	0.0105
		$\alpha = 0.05$	0.0472	0.0512	0.0514	0.0508
2000	10	$\alpha = 0.01$	0.0111	0.0094	0.0118	0.0094
		$\alpha = 0.05$	0.0489	0.0491	0.0508	0.0465
	20	$\alpha = 0.01$	0.0113	0.0107	0.0098	0.0108
		$\alpha = 0.05$	0.0513	0.0491	0.0516	0.0523
	40	$\alpha = 0.01$	0.0099	0.0091	0.0107	0.0110
		$\alpha = 0.05$	0.0498	0.0480	0.0492	0.0476

Table 1. Estimated type I error rates for the MultP-PE method under four models. The type I error rates are evaluated using 10,000 replicated samples. *P*-values of MultP-PE are estimated by 1,000 permutations. α is the significance level. The number of replications is 10,000. The type I error rate in italics indicates the value out of the bounds of the 95% CI.

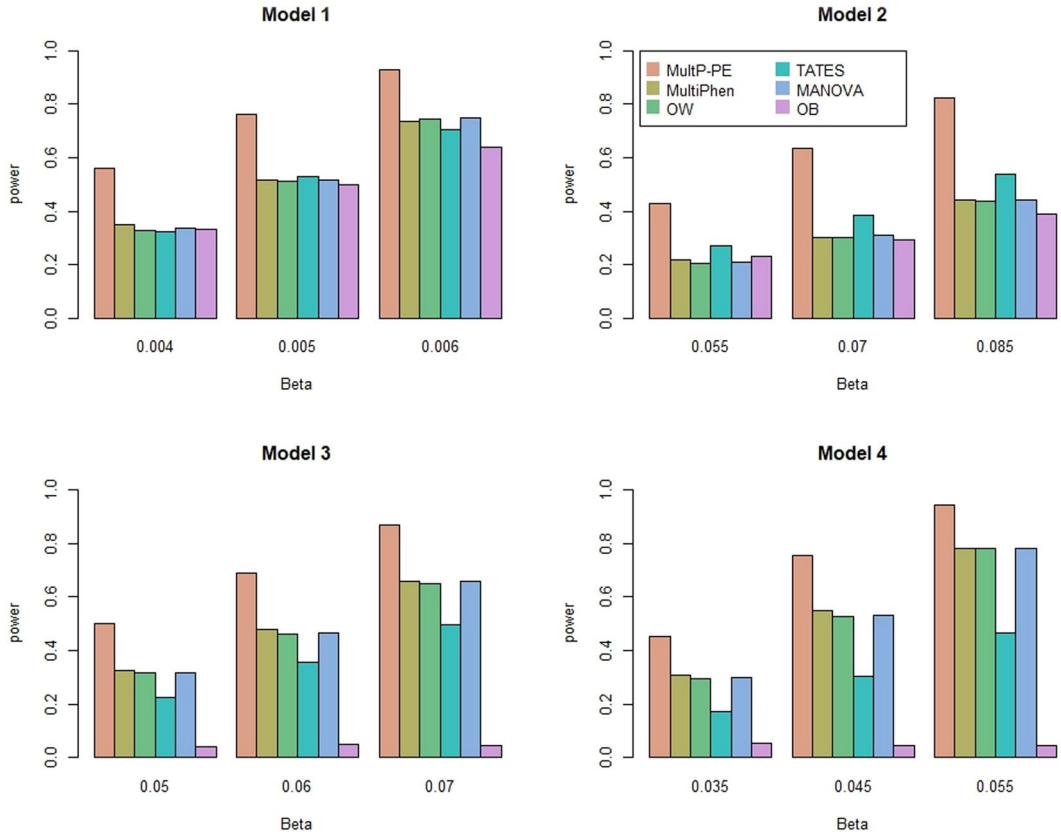

Figure 1. Power comparisons of the six methods as a function of effect size β. The total number of phenotypes is $K = 20$, sample size is 1000, MAF is 0.3, the between-factor correlation is 0.15, and the within-factor correlation is 0.25. Significance is assessed at the 5% level.

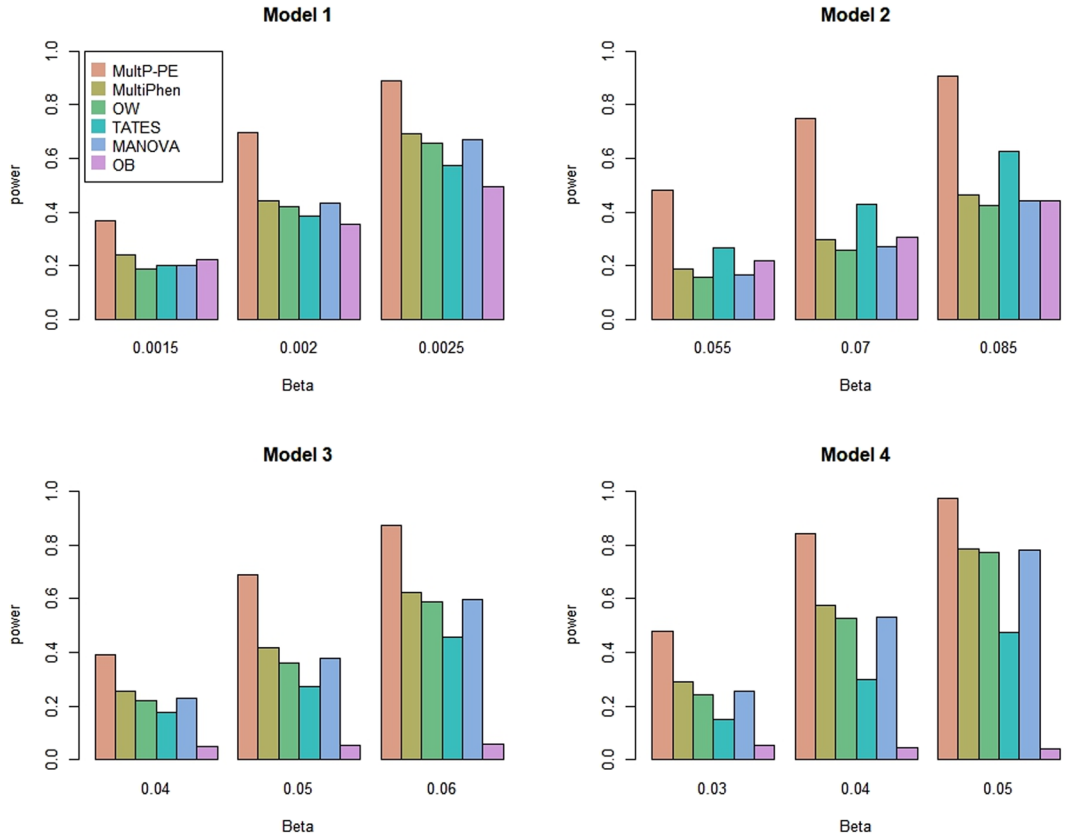

Figure 2. Power comparisons of the six methods as a function of effect size β. The total number of phenotypes is $K = 40$, sample size is 1000, MAF is 0.3, the between-factor correlation is 0.15, and the within-factor correlation is 0.25. Significance is assessed at the 5% level.

$T^{(b)}_{MultP-PE} = \min_{1\le m\le M} p^{(b)}_{\lambda_m}$ denote the test statistic of $T_{MultP-PE}$ based on the b^{th} permuted data, then the p-value of $T_{MultP-PE}$ is given by

$$\frac{\#\{T^{(b)}_{MultP-PE} : T^{(b)}_{MultP-PE} < T^{(0)}_{MultP-PE} \text{ for } b = 1, 2, \ldots, B\}}{B} \tag{3}$$

To apply MultP-PE to GWAS with hundreds of thousands of SNPs, we also propose an algorithm that can perform the permutation procedure described above more efficiently in the following section.

A Fast Algorithm for the Permutation Procedure. We use the notations in the above section and let $A_\lambda = (Y^TY + \lambda I)^{-1}$, $h_i^\lambda = y_i^T A_\lambda y_i$, $h_\lambda = (h_1^\lambda, \ldots, h_n^\lambda)$, and $\hat{\beta}_\lambda = A_\lambda Y^T x$. Then, the Ridge predicted value of x_i is $\hat{x}_i^\lambda = y_i^T \hat{\beta}_\lambda$ and $\hat{x}_\lambda = (\hat{x}_1^\lambda, \ldots, \hat{x}_n^\lambda)^T = Y(Y^TY + \lambda I)^{-1}Y^Tx$. We can show that the LOOCV prediction error in Ridge regression has a closed-form formula[24,26], that is, $x_i - \hat{x}_{-i}^\lambda = (x_i - \hat{x}_i^\lambda)/(1 - h_i^\lambda)$. Note that for two matrices or vectors A and B, we use $A*B$ and $\frac{A}{B}$ to denote the element-wise operations; for a matrix C, we use $colSum(C)$ to denote the sums of the columns of matrix C. We assume $n \ge K+1$. We perform singular value decomposition of Y, that is, $Y = UDV$, where U is an $n \times (K+1)$ matrix with orthonormal columns, D is $(K+1) \times (K+1)$ diagonal matrix with non-negative real numbers on the diagonal, and V is an $(K+1) \times (K+1)$ orthogonal matrix. Let $D = diag(d_1, \ldots, d_{K+1})$. Then, $\hat{x}_\lambda = UC_\lambda U^Tx$, where $C_\lambda = diag(c_{\lambda,1}, \ldots, c_{\lambda,K+1})$ and $c_{\lambda,j} = d_j^2/(d_j^2 + \lambda)$ for $j = 1, \ldots, K+1$. Let $c_\lambda = (c_{\lambda,1}, \ldots, c_{\lambda,K+1})^T$ and $x^{(K)} = U^Tx$ be a $K+1$ dimensional vector. Then, $\hat{x}_\lambda = UC_\lambda x^{(K)} = U(c_\lambda * x^{(K)})$ and $h_\lambda = diag(UC_\lambda U^T)$. For $0 \le \lambda_1 < \ldots < \lambda_M < \infty$, let $C = (c_{\lambda_1}, \ldots, c_{\lambda_M})$ and $H = (h_{\lambda_1}, \ldots, h_{\lambda_M})$. Then, $(\hat{x}_{\lambda_1}, \ldots, \hat{x}_{\lambda_M}) = U(C * x^{(K)}) = U(c_{\lambda_1} * x^{(K)}, \ldots, c_{\lambda_M} * x^{(K)})$. If we denote $Q = \frac{(x - \hat{x}_{\lambda_1}, \ldots, x - \hat{x}_{\lambda_M})}{1 - H}$, then $(T_{\lambda_1}, \ldots, T_{\lambda_M}) = colSum(Q * Q)$. Note that C, U, and H only depend on phenotypes and $\lambda_1, \ldots, \lambda_M$. Thus, C, U, and H do not change in each permutation. For a GWAS, C, U, and H also do not change at different SNPs. For 1,000 permutations on one SNP, our fast algorithm is about 20 times faster than the original algorithm (the original algorithm calculates T_λ by $T_\lambda = \sum_{i=1}^n (x_i - \hat{x}_{-i}^\lambda)^2$). To perform a GWAS with hundreds of thousands of SNPs, we can use the same approach as was suggested in Zhu *et al.*[14], that is, we can first select SNPs that show evidence of association based on a small number of permutations (e.g. 1,000), then use a large number of permutations to test the selected SNPs. For example, in our real data analysis with 630,860 SNPs, we first performed 1,000

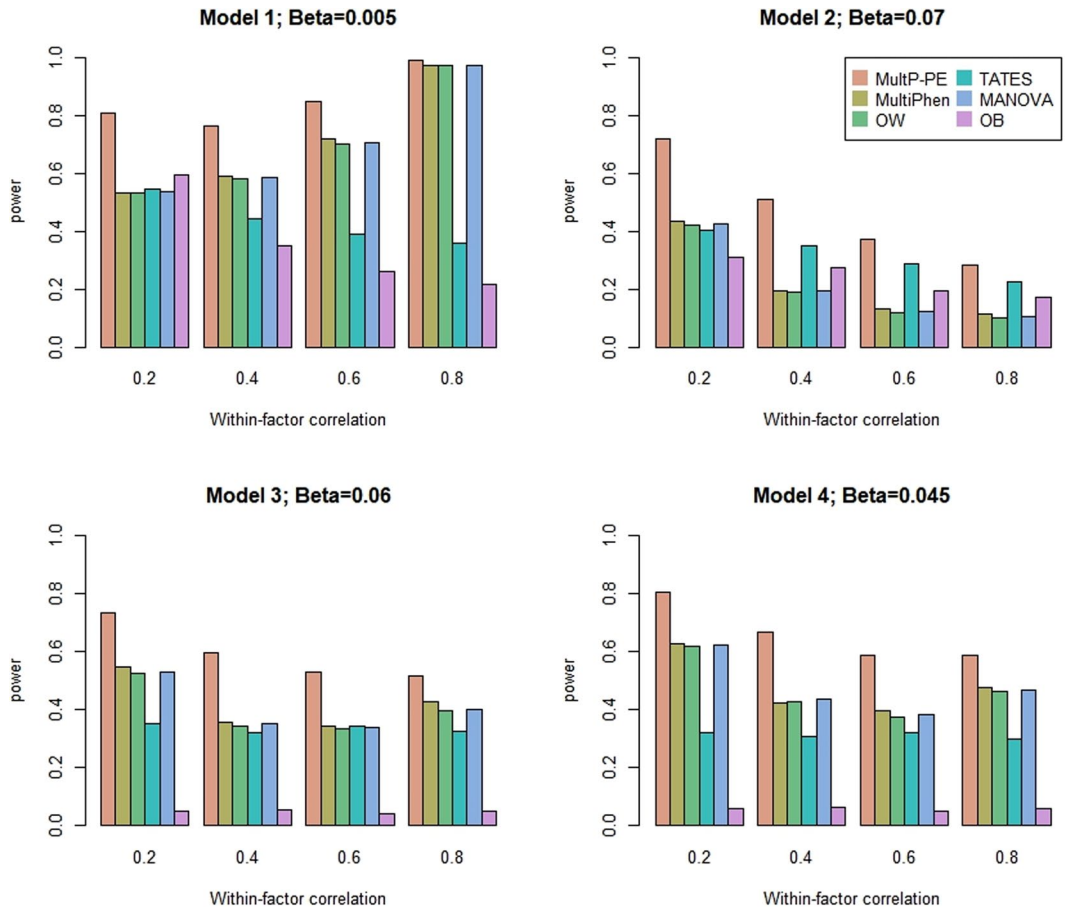

Figure 3. Power comparisons of the six methods as a function of within-factor correlation c^2. The total number of phenotypes is $K = 20$, sample size is 1000, MAF is 0.3, and the between-factor correlation is 0.15. Significance is assessed at the 5% level.

permutations and selected SNPs with $p\text{-}value \leq 0.005$, then we performed 10^8 permutations on the selected SNPs because SNPs with $p\text{-}value > 0.005$ are not significantly associated with phenotypes.

Although we use a permutation procedure to calculate the p-value of MultP-PE, by using our fast algorithm, we can use less than one day to perform a typical GWAS. In our read data analysis on COPD in the following section, we performed a GWAS with 5,430 individuals across 630,860 SNPs and seven phenotypes. We completed the analysis in 10 hours on Intel Xeon E5-2680v3 by using a single node.

In the above section, we describe MulP-PE without considering covariates. If covariates are needed to be considered, we can incorporate covariates using the following approach in MultP-PE. Suppose that there are total G covariates we would like to consider and let $(z_{i1}, \ldots, z_{iG})^T$ denote the covariates for the i^{th} individual. We can adjust esch of the phenotypes by the covariates by applying the linear regression model $y_{ik} = a_{0k} + a_{1k}z_{i1} + \ldots + a_{Mk}z_{iG} + \varepsilon_{ik}$, for $i = 1, 2, \ldots, n$, $k = 1, 2, \ldots, K$, and use the residual of y_{ik} to replace y_{ik} in MultP-PE. In our real data analysis, we used this approach to incorporate four covariates. This approach has been used in the literature. For example, Sha *et al.*[16] and Zhu *et al.*[14] also used the same approach to adjust phenotypes for the covariates.

In association studies for unrelated individuals, it has been well known that population stratification can seriously confound association results[27]. There are several methods that have been developed to control for population stratification. For example, Genomic Control (GC) approach[28,29], Principal Component (PC) approach[16,30–32], and Mixed Linear Model (MLM) approach[33,34]. Similar to most association tests for unrelated individuals, MulP-PE subjects to bias due to population stratification. To make MultP-PE robust to population stratification, we can use the PC approach. Let $c_{i1}, \ldots, c_{iL}$ denote the top L PCs of the genotypes at a set of genomic markers for the i^{th} individual. We can use the residuals of the regression model $x_i = \alpha + \beta_1 c_{i1} + \cdots + \beta_L c_{iL} + \varepsilon_i$ to replace x_i and use the residuals of the regression model $y_{ik} = \alpha_k + \beta_{1k} c_{i1} + \cdots + \beta_{Lk} c_{iL} + \varepsilon_{ik}$ to replace y_{ik} for $k = 1, 2, \ldots, K$ in MultP-PE to adjust for population stratification.

Comparison of Methods. We evaluate the performance of the proposed test MultP-PE by comparing it with five most commonly used methods for association studies using multiple phenotypes. These five methods include the O'Brien's method (OB)[5], Trait-based Association Test that uses Extended Simes procedure (TATES)[7], Optimal weight method (OW)[6], Multivariate analysis of variance (MANOVA)[9], and Joint model of multiple phenotypes (MultiPhen)[2].

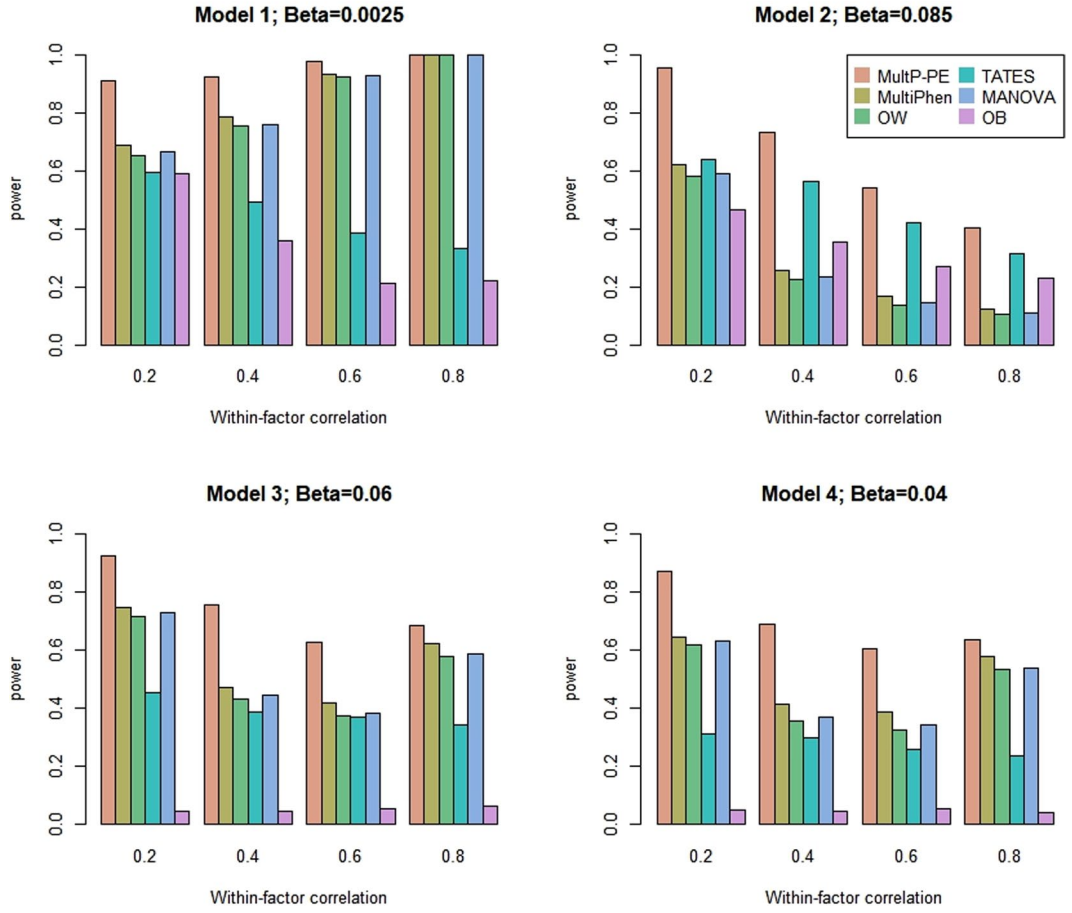

Figure 4. Power comparisons of the six methods as a function of within-factor correlation c^2. The total number of phenotypes is $K = 40$, sample size is 1000, MAF is 0.3, and the between-factor correlation is 0.15. Significance is assessed at the 5% level.

Simulation Study

In simulation studies, we evaluate type I error rates of MultP-PE by generating data sets with three different sample sizes, 500, 1,000 and 2,000. For power comparison, we compare the powers of different methods by simulation data sets with 1,000 unrelated individuals.

For genotype data, we generate genotype at a genetic variant according to minor allele frequency (MAF) and assume Hardy-Weinberg Equilibrium (HWE). For each individual, we generate K phenotypes using models similar to the models used in Zhu *et al.*[14] and Wang *et al.*[35]. The K phenotypes are generated from the following model

$$y = \phi x + c\gamma\omega + \sqrt{1 - c^2} \times \varepsilon \tag{4}$$

where $y = (y_1, \ldots, y_K)^T$; $\phi = (\phi_1, \ldots, \phi_K)$ are the genetic effects of the variant on the K phenotypes; x is the genotypic score at the variant; c is a constant number; γ is a $K \times R$ matrix; $\omega = (\omega_1, \ldots, \omega_R)^T$ is a vector of factors with R elements and $\omega = (\omega_1, \ldots, \omega_R)^T \sim MVN(0, \Sigma)$, $\Sigma = \rho A + (1 - \rho)I$, ρ is the correlation between factors, A is a matrix with elements of 1, and I is the identity matrix; $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_K)^T$ is a vector of residuals, $\varepsilon_1, \ldots, \varepsilon_K$ are independent, and $\varepsilon_k \sim N(0, 1)$ for $k = 1, \ldots, K$. Based on equation (4), we consider the following four models in which the within-factor correlation is c^2 and the between-factor correlation is ρc^2.

Model 1. There is only one factor and genotypes impact on all phenotypes with different effect sizes. That is, $R = 1$, $\phi = \beta(1, 2, \ldots, K)^T$, and $\gamma = (1, \ldots, 1)^T$.

Model 2. There are two factors and genotypes impact on one factor. That is, $R = 2$, $\phi = \left(0, \ldots, 0, \underbrace{\beta, \ldots, \beta}_{K/2}\right)^T$, and $\gamma = Bdiag(D_1, D_2)$, where $D_i = \left(\underbrace{1, \ldots, 1}_{K/2}\right)^T$ for $i = 1, 2$ and $Bdiag$ means block diagonal.

Chr	Position	Variant identifier	OB	TATES	OW	MANOVA	MultiPhen	MultP-PE
4	145431497	rs1512282	0.46	7.09×10^{-13}	8.10×10^{-14}	6.52×10^{-14}	1.03×10^{-9}	$<1\times10^{-8}$
4	145434744	rs1032297	0.49	6.22×10^{-13}	1.11×10^{-16}	1.11×10^{-16}	7.69×10^{-14}	$<1\times10^{-8}$
4	145474473	rs1489759	0.42	2.49×10^{-16}	1.11×10^{-16}	6.68×10^{-17}	1.22×10^{-16}	1.00×10^{-8}
4	145485738	rs1980057	0.49	8.35×10^{-17}	1.11×10^{-16}	7.12×10^{-17}	8.14×10^{-17}	1.00×10^{-8}
4	145485915	rs7655625	0.34	6.11×10^{-9}	1.87×10^{-9}	1.69×10^{-9}	9.13×10^{-17}	5.00×10^{-8}
15	78882925	rs16969968	0.96	$\mathbf{5.40\times10^{-8}}$	2.05×10^{-11}	1.77×10^{-11}	7.84×10^{-12}	$<1\times10^{-8}$
15	78894339	rs1051730	0.99	3.13×10^{-8}	1.54×10^{-11}	1.32×10^{-11}	8.16×10^{-12}	$<1\times10^{-8}$
15	78898723	rs12914385	0.99	2.76×10^{-8}	1.64×10^{-11}	1.41×10^{-11}	1.48×10^{-12}	$<1\times10^{-8}$
15	78911181	rs8040868	0.99	5.53×10^{-10}	2.09×10^{-12}	1.76×10^{-12}	2.59×10^{-12}	$<1\times10^{-8}$
15	78878541	rs951266	0.77	2.55×10^{-9}	3.24×10^{-12}	2.74×10^{-12}	1.02×10^{-11}	$<1\times10^{-8}$
15	78806023	rs8034191	0.87	$\mathbf{1.06\times10^{-7}}$	2.42×10^{-10}	2.14×10^{-10}	7.74×10^{-11}	$<1\times10^{-8}$
15	78851615	rs2036527	0.88	$\mathbf{1.62\times10^{-7}}$	4.47×10^{-10}	3.99×10^{-10}	1.77×10^{-10}	$<1\times10^{-8}$
15	78826180	rs931794	0.91	$\mathbf{1.23\times10^{-7}}$	2.64×10^{-10}	2.35×10^{-10}	9.09×10^{-11}	$<1\times10^{-8}$
15	78740964	rs2568494	0.27	$\mathbf{2.93\times10^{-5}}$	$\mathbf{1.12\times10^{-7}}$	$\mathbf{1.05\times10^{-7}}$	4.23×10^{-8}	$\mathbf{1.50\times10^{-7}}$

Table 2. Significant SNPs and the corresponding p-values in the analysis of COPDGene. The p-values of MultP-PE are evaluated using 10^8 permutations. The p-values of OB, TATES, OW, MANOVA, and MultiPhen are evaluated using their asymptotic distributions. The bold out p-values indicate the p-values $> 5\times10^{-8}$.

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

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

For the type I error rates, we set $\beta=0$ to indicate that the genetic variant has no effect on all phenotypes. For power comparisons, we consider different values of β. To evaluate type I error rate and power, we set MAF = 0.3, the between-factor correlation is 0.14, and the within-factor correlation is 0.25. In the following simulation studies and real data analysis, we use eight different values of λ ($M=8$) and set $\log\lambda=0,\ 1,\ 2,\ 3,\ 3.5,\ 3.8,\ 4,\ 4.5$.

The R codes for implementation of MultP-PE and for simulation of data under the four models are available at Dr. Shuanglin Zhang's homepage http://www.math.mtu.edu/shuzhang/software.html.

Results

To evaluate the type I error rates of MultP-PE, we consider different significance levels (0.01 and 0.05), different sample sizes (500, 1000 and 2000), and different number of phenotypes (10, 20 and 40). We use 1,000 permutations to calculate the p-values of MultP-PE and use 10,000 replicated samples to evaluate type I error rates of MultP-PE. For 10,000 replicated samples, the 95% confidence intervals (CIs) for the estimated type I error rates with nominal levels 0.05 and 0.01 are (0.04562, 0.05438) and (0.00804, 0.01196), respectively. We summarize the estimated type I error rates of the proposed test in Table 1. This table shows that only one type I error rate is not in the corresponding 95% CI (it is very close to the upper-bound of the CI), which indicates that the proposed method is valid.

In power comparisons, we calculate the p-values of MultP-PE using 1,000 permutations and the p-values of MultiPhen, OW, TATES, MANOVA, OB 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. Figures 1 and 2 show the powers of the six methods as a function of the effect size β with $K=20$ and 40, respectively. As shown in these two figures: (1) MultP-PE is the most powerful test. The power of MultP-PE is much higher than the second most powerful test; (2) as the effect size β increases, the powers of all tests increase as well; as the number of phenotypes K increases from 20 to 40, MultP-PE presents more ascendancy than the other five tests; (3) MultiPhen, OW, and MANOVA have similar powers under all four models. A similar conclusion has been reached in some published papers[2,6,7]; (4) OB is comparable to MultiPhen, OW, and MANOVA in models 1 and 2, but has almost no power when the genetic effects have different directions (models 3 and 4); (5) TATES is more powerful than MultiPhen, OW, and MANOVA in model 2, but is less powerful than MultiPhen, OW, and MANOVA in models 3 and 4.

Power comparisons of the six methods as a function of the within-factor correlation, c^2, with $K=20$ and 40 are given in Figs 3 and 4, respectively. As shown in these two figures: (1) the patterns of the power performance are similar to those in Figs 1 and 2; (2) when the within-factor correlation is increasing, the powers of all six tests

have increasing trend or decreasing trend depending on different model settings. This pattern has been confirmed in Zhu's paper[6]; (3) OB is the least powerful test except under model 2 with the within-factor correlation > 0.2.

Power comparisons of the six methods as a function of the between-factor correlation, $c^2\rho$, with $K = 20$ and 40 are given in Figs S1 and S2, respectively. As shown in these two figures: (1) the patterns of the power performance are similar to those in Figs 1 and 2; (2) when the between-factor correlation is increasing, the powers of all six tests have increasing trend except for these under model 1; (3) MultP-PE is the most powerful test, while OB is the least powerful test except under model 2 with the between-factor correlation $= 0.1$.

In summary, MultP-PE is consistently the most powerful test among the tests we compared under all simulation scenarios.

Real Data Analysis

Chronic obstructive pulmonary disease (COPD) is a terminology to describe progressive life-threatening lung diseases that causes breathlessness and serious illness, including emphysema, chronic bronchitis, refractory asthma, and some forms of bronchiectasis. A global prevalence of 251 million cases of COPD is reported in 2016 and it is estimated that COPD caused 3.17 million deaths in 2015[36]. The COPDGene aims to find inherited or genetic factors that associated with COPD. The COPDGene dataset includes 10,192 participants, 3,408 of them are African-Americans (AA), and 6,784 of them are Non-Hispanic Whites (NHW). Same as Liang *et al.*[37], we considered Age, Sex, BMI, and Pack-Years as four covariates and selected seven quantitative COPD-related phenotypes (FEV1, Emphysema, Emphysema Distribution, Gas Trapping, Airway Wall Area, Exacerbation frequency, and Six-minute walk distance) in the following data analysis.

We deleted individuals and genotypes with missing data. After excluding missing data, a set of 5,430 NHW across 630,860 SNPs was used in the analysis. Then we adjusted the phenotypes for the covariates by applying a linear regression[14,17]. We regressed each phenotype on the four covariates, replaced original phenotypes with the residuals of the regression, and applied each of the six tests to detect the association between the covariates-adjusted phenotypes (residuals) and each SNP.

We used genome-wide significance level 5×10^{-8} to identify SNPs that are significantly associated with the seven COPD-related phenotypes. There were total 14 SNPs identified by at least one method (Table 2). All of the 14 SNPs had been reported to be associated with COPD by previous studies[38–50]. As shown in Table 2, MultiPhen identified 14 SNPs; OW, MANOVA, and MultP-PE identified 13 SNPs; TATES identified 9 SNPs; and OB did not identify any SNPs. The number of SNPs identified by MultP-PE was comparable to the largest number of SNPs identified by other tests and the COPD analysis results were consistent with our simulation results. We also performed individual phenotype analysis on each of the seven phenotypes. Table S1 gives the adjusted p-values (Bonferroni correction for multiple testing) to test each of the seven phenotypes on the 14 significant SNPs. We can see from Table S1, among the 14 SNPs, only nine SNPs are significantly associated with Emphysema Distribution at the genome-wide significance level. The number of SNPs identified by individual phenotype is the same as TATES and is less than the number of SNPs identified by four multiple phenotype analyses (OW, MANOVA, Multiphen, and MultP-PE), which showed that the simultaneous analysis of multiple phenotypes can increase power comparing with single phenotype analysis.

Discussion

For complex diseases in GWAS, the association between a genetic variant and each phenotype is usually weak. Analyzing multiple disease-related phenotypes could increase statistical power to identify the association between genetic variants and complex diseases. In this article, we developed a novel statistical method, MultP-PE, to test the association between a genetic variant and multiple phenotypes based on cross-validation prediction error. We showed that MultP-PE controls type I error rates very well and has consistently higher power than other methods we compared among all the simulation scenarios. Overall, MultP-PE is the most powerful test and has much higher power than the second most powerful test; OW, MANOVA, and MultiPhen have very similar performance; OB loses power dramatically when genetic effects have opposite directions on phenotypes; TATES is more powerful when the genetic effect only works on a portion of phenotypes. In real data analysis, MultP-PE identified 13 out of 14 significant SNPs, which is comparable to MultiPhen (14 out of 14).

References

1. Yang, Q., Wu, H., Guo, C. Y. & Fox, C. S. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genet Epidemiol* **34**(5), 444–454 (2010).
2. O'Reilly, P. F. *et al.* MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One* **7**(5), e34861 (2012).
3. Wang, Y. *et al.* Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. *Genet Epidemiol* **39**(4), 259–275 (2015).
4. Yang, J. J., Li, J., Williams, L. K. & Buu, A. An efficient genome-wide association test for multivariate phenotypes based on the Fisher combination function. *BMC Bioinform* **17**(1), 1 (2016).
5. O'Brien, P. C. Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–1087 (1984).
6. Zhu, H., Zhang, S. & Sha, Q. Power Comparisons of Methods for Joint Association Analysis of Multiple Phenotypes. *Hum Hered* **80**(3), 144–52 (2016).
7. van der Sluis, S., Posthuma, D. & Dolan, C. V. TATES: Efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet* **9**(1), e1003235 (2013).
8. Ferreira, M. A. & Purcell, S. M. A multivariate test of association. *Bioinformatics* **25**(1), 132–133 (2009).
9. Cole, D. A., Maxwell, S. E., Avrey, R. & Salas, E. How the power of MANOVA can both increase and decrease as a funcion of the intercorrelations among the dependent variables. *Psychol Bull* **115**(3), 465 (1994).
10. Galesloot, T. E., van Steen, K., Kiemeney, L. A. L. M., Janss, L. L. & Vermeulen, S. H. A comparison of multivariate genome-wide association methods. *PLoS One* **9**, e95923 (2014).
11. Aschard, H. *et al.* Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am J Hum Genet* **94**(5), 662–676 (2014).

12. Klei, L., Luca, D., Devlin, B. & Roeder, K. Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genet Epidemiol* **32**(1), 9–19 (2008).
13. Wang, K. & Abbott, D. A principal components regression approach to multilocus genetic association studies. *Genet Epidemiol* **32**, 108–118 (2008).
14. Zhu, H., Zhang, S. & Sha, Q. 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 (2018).
15. Turley, P. *et al*. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet* **50**, 229–37 (2018).
16. Price, A. L. *et al*. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904–9 (2006).
17. Sha, Q., Wang, X., Wang, X. & Zhang, S. Detecting association of rare and common variants by testing an optimally weighted combination of variants. *Genet Epidemiol* **36**, 561–571 (2012).
18. Draper, N. R. & Smith, H. *Applied Regression Analysis*, (John Wiley & Sons, 2014).
19. Cule, E. & De Iorio, M. Ridge regression in prediction problems: automatic choice of the ridge parameter. *Genet Epidemiol* **37**(7), 704–14, https://doi.org/10.1002/gepi.21750. PubMed PMID: 23893343; PMCID: PMC4377081 (2013).
20. Cule, E., Vineis, P. & De Iorio, M. Significance testing in ridge regression for genetic data. *BMC Bioinformatics* **12**, 372, https://doi.org/10.1186/1471-2105-12-372. PubMed PMID: 21929786; PMCID: PMC3228544 (2011).
21. Halawa, A. & El Bassiouni, M. Tests of regression coefficients under ridge regression models. *J Stat Comput and Simul* **65**(1–4), 341–56 (2000).
22. Hoerl, A. E., Kannard, R. W. & Baldwin, K. F. Ridge regression: some simulations. *Commun Stat Theory Methods* **4**(2), 105–23 (1975).
23. Malo, N., Libiger, O. & Schork, N.J. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *Am J Hum Genet* **82**(**2**), 375–85, https://doi.org/10.1016/j.ajhg.2007.10.012. PubMed PMID: 18252218; PMCID: PMC2427310 (2008).
24. Yang, X., Wang, S., Zhang, S. & Sha, Q. Detecting association of rare and common variants based on cross-validation prediction error. *Genet Epidemiol* **41**(3), 233–243 (2017).
25. Ge, Y., Dudoit, S. & Speed, T. P. Resampling-based multiple testing for microarray data analysis. *Test* **12**(1), 1–77 (2003).
26. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An introduction to statistical learning*, (Springer, 2013).
27. Lander, E. S. & Schork, N. J. Genetic dissection of complex traits. *Science* **265**, 2037–48 (1994).
28. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
29. Reich, D. E. & Goldstein, D. B. Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* **20**, 4–16 (2001).
30. Chen, H. S., Zhu, X., Zhao, H. & Zhang, S. Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. *Ann Hum Genet* **67**, 250–64 (2003).
31. Zhang, S., Zhu, X. & Zhao, H. On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genet Epidemiol* **24**, 44–56 (2003).
32. Zhu, X., Zhang, S., Zhao, H. & Cooper, R. S. Association mapping, using a mixture model for complex traits. *Genet Epidemiol* **23**, 181–96 (2002).
33. Zhang, Z. *et al*. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* **42**, 355–60 (2010).
34. Kang, H. M. *et al*. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348–54 (2010).
35. Wang, Z., Sha, Q. & Zhang, S. Joint Analysis of Multiple Traits Using "Optimal" Maximum Heritability Test. *PloS one* **11**(3), e0150975 (2016).
36. Chronic obstructive pulmonary disease (COPD). *WHO*. Retrieved from, http://www.who.int/mediacentre/factsheets/fs315/en/ (Nov. 2017).
37. Liang, X. *et al*. An Adaptive Fisher's Combination Method for Joint Analysis of Multiple Phenotypes in Association Studies. *Sci Rep* **6**, 34323, https://doi.org/10.1038/srep34323 (2016).
38. Brehm, J. M. *et al*. Identification of FGF7 as a novel susceptibility locus for chronic obstructive pulmonary disease. *Thorax* **66**(12), 1085–1090 (2011).
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* **9**(7), e102324 (2014).
40. Du, Y., Xue, Y. & Xiao, W. Association of IREB2 gene rs2568494 polymorphism with risk of chronic obstructive pulmonary disease: a meta-analysis. *Med Sci Monit* **22**, 177 (2016).
41. Cho, M. H. *et al*. Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat Genet* **42**(3), 200–202 (2010).
42. Hancock, D. B. *et al*. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat Genet* **42**(1), 45–52 (2010).
43. Lutz, S. M. *et al*. A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC Genet* **16**(1), 1 (2015).
44. Li, X. *et al*. Importance of hedgehog interacting protein and other lung function genes in asthma. *J Allergy Clin Immunol* **127**(6), 1457–1465 (2011).
45. Pillai, S. G. *et al*. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet* **5**(3), e1000421 (2009).
46. Wilk, J. B. *et al*. A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genet* **5**(3), e1000429 (2009).
47. Wilk, J. B. *et al*. Genome-wide association studies identify CHRNA5/3 and HTR4 in the development of airflow obstruction. *Am J Respir Crit Care Med* **186**(7), 622–632 (2012).
48. Young, R. P. *et al*. Chromosome 4q31 locus in COPD is also associated with lung cancer. *Eur Respir J* **36**(6), 1375–1382 (2010).
49. Zhang, J., Summah, H., Zhu, Y. G. & Qu, J. M. Nicotinic acetylcholine receptor variants associated with susceptibility to chronic obstructive pulmonary disease: a meta-analysis. *Respir Res* **12**(1), 1 (2011).
50. Zhu, A. Z. *et al*. Association of CHRNA5-A3-B4 SNP rs2036527 with smoking cessation therapy response in African-American smokers. *Clin Pharmacol Ther* **96**(2), 256–265 (2014).

Acknowledgements

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number R15HG008209. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research used data generated by the COPDGene study, which was supported by National Institutes of Health (NIH) grants U01HL089856 and U01HL089897. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health. The COPDGene project is also supported by the COPD Foundation through contributions made by an Industry Advisory Board comprised of Pfizer, AstraZeneca, Boehringer Ingelheim, Novartis, and Sunovion. Superior, a

high-performance computing infrastructure at Michigan Technological University, was used in obtaining results presented in this publication.

Author Contributions

S.Z. and Q.S. designed research, X.Y. and S.Z. performed statistical analysis, and X.Y., S.Z. and Q.S. wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-37538-y.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.