



**Michigan  
Technological  
University**

Michigan Technological University  
**Digital Commons @ Michigan Tech**

---

Dissertations, Master's Theses and Master's Reports

---

2019

## STATISTICAL METHODS FOR JOINT ANALYSIS OF MULTIPLE PHENOTYPES AND THEIR APPLICATIONS FOR PHEWAS

Xueling Li

*Michigan Technological University, [xuelingl@mtu.edu](mailto:xuelingl@mtu.edu)*


Copyright 2019 Xueling Li

---

### Recommended Citation

Li, Xueling, "STATISTICAL METHODS FOR JOINT ANALYSIS OF MULTIPLE PHENOTYPES AND THEIR APPLICATIONS FOR PHEWAS", Open Access Dissertation, Michigan Technological University, 2019. <https://digitalcommons.mtu.edu/etdr/813>

Follow this and additional works at: <https://digitalcommons.mtu.edu/etdr>

 Part of the [Applied Statistics Commons](#), [Bioinformatics Commons](#), [Biostatistics Commons](#), [Health Information Technology Commons](#), [Respiratory Tract Diseases Commons](#), and the [Statistical Methodology Commons](#)

**STATISTICAL METHODS FOR JOINT ANALYSIS  
OF MULTIPLE PHENOTYPES AND THEIR  
APPLICATIONS FOR PHEWAS**

By

Xueling Li

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Statistics

MICHIGAN TECHNOLOGICAL UNIVERSITY

2019

© 2019 Xueling Li

This dissertation has been approved in partial fulfillment of the requirements for the Degree of DOCTOR OF PHILOSOPHY in Statistics.

Department of Mathematical Sciences

Dissertation Advisor: *Qiuying Sha*

Committee Member: *Shuanglin Zhang*

Committee Member: *Kui Zhang*

Committee Member: *Jingfeng Jiang*

Department Chair: *Mark S. Gockenbach*

# Contents

<b>Contents</b> .....	iii
<b>Preface</b> .....	v
<b>Acknowledgements</b> .....	vi
<b>List of Abbreviations</b> .....	viii
<b>Abstract</b> .....	x
<b>Chapter 1: HC-CLC</b> .....	1
1.1 Introduction .....	2
1.2 Methods .....	6
1.3 Real Data Analysis .....	15
1.3.1 COPD Data Set .....	15
1.3.2 Real Data Analysis Results .....	16
1.4 Discussion .....	17
1.5 Tables and Figures .....	19
<b>Chapter 2: UKB-PheCLC</b> .....	25
2.1 Introduction .....	27
2.2 UK Biobank Data Pre-processing .....	31
2.2.1 Introduction of UK Biobank Phenotypes .....	31
2.2.2 Introduction of UK Biobank Genotypes .....	33
2.2.3 Introduction of Covariates of UKB-PheCLC .....	34
2.2.4 Quality Controls .....	34
2.2.5 Hierarchical Groups of UK Biobank ICD-10 Codes .....	36
2.3 Methods .....	37
2.4 Methods Comparison .....	40
2.5 Results .....	42
2.6 Discussion .....	44
2.7 Tables and Figures .....	47

<b>Reference List</b> .....	49
<b>Appendix A: Supplementary Tables</b> .....	55
<b>Appendix B: Supplementary Figures</b> .....	56

# Preface

This dissertation presents my research work in pursuing the Doctor of Philosophy degree in Statistic at Michigan Technological University. The research presented here was conducted under the supervision of Dr. Qiuying Sha and Dr. Shuanglin Zhang from the Department of Mathematical Sciences. The work is to the best of my knowledge and belief original, except where due reference is made in the text of the dissertation.

Chapter 1 entitled Joint analysis of multiple phenotypes using a clustering linear combination method based on hierarchical clustering is a manuscript ready for submission. Qiuying Sha and Shuanglin Zhang developed the methodologies. Xueling Li and Qiuying Sha performed the statistical analyses. Zhenchuan Wang preprocessed the COPDGene real data. Xueling Li, Shuanglin Zhang, and Qiuying Sha drafted the manuscript.

Chapter 2 entitled Application of UKBiobank data for phenome-wide association study is a continuation of the collaborative work of Huanhuan Zhu, Shuanglin Zhang, and Qiuying Sha. Huanhuan Zhu, Shuanglin Zhang, and Qiuying Sha focused their research on method development. The focus of this dissertation is refining the proposed method and applying it to the UK Biobank data, a large cohort study across the United Kingdom, to test the validity and understand the limitations of the proposed method. Shuanglin Zhang and Qiuying Sha continued their contribution to methodology development. Xueling Li pre-processed the UK Biobank data and performed all subsequent statistical analysis.

## Acknowledgements

My sincere and deep gratitude should first and foremost be given to my advisor, Dr. Qiuying Sha, who serves not only as an incredible academic supervisor for my Ph.D. research, but also a fabulous mentor for my career development. Dr. Sha's devotedness to the Statistical Genetics profession and her positive attitude and energy towards everything really impress me. I feel extremely lucky and honored to have the opportunity to work with Dr. Sha.

Furthermore, I would like to extend my deep gratitude to Dr. Shuanglin Zhang for his valuable insights and suggestions for my Ph.D. research. This dissertation could be finished without his expertise and help. I would also like to thank Dr. Kui Zhang and Dr. Jingfeng Jiang for serving as my dissertation committee members and providing me with valuable feedbacks about my research. I feel humbled by the intelligence of these outstanding professionals.

In addition, I would like to acknowledge Ms Hua Huang and Dr. Xuexia Wang for sharing their ideas and offering advice in pre-processing the UK Biobank data.

Moreover, I really appreciate the opportunity that the math department had provided me to teach. This really helped improve my communication skills and I truly enjoyed working with every single one of my students. It's such a pleasure and honor to work as an instructor while I am still in school pursuing my own degree.

Finally, I would like to express my special thanks to my family. I am very grateful for my parents Mr. Lianming Li and Mrs. Chaoquan Wang for their unconditional love and

support. I also want to thank my younger brother Qucheng Li for his understanding and willingness to take care of the family while I am far away from home. Lastly, I want to give my biggest “thank-Yu” to my husband Yu Wang for his boundless love and encouragement.



## List of Abbreviations

---

6MWD	Six-Minute Walk Distance
AFC	Adaptive Fisher's Combination
BMI	Body Mass Index
CCA	Canonical Correlation Analysis
Cis	Confidence Intervals
CLC	Clustering Linear Combination
COPD	Chronic Obstructive Pulmonary Disease
HER	Electronic Health Record
Emph	Emphysema
EmphDist	Emphysema Distribution
ExacerFreq	Exacerbation Frequency
FEV1	% predicted FEV1
GasTrap	Gas Trapping
GEE	Generalized Estimating Equations
GWAS	Genome-Wide Association Studies
HC	Hierarchical Clustering
HC-CLC	HC approach followed by CLC method
HWE	Hardy Weinberg Equilibrium
ICD	International Classification of Disease
MAF	Minor Allele Frequencies
OB	O'Brian's method (1984)

---

---

PackYear	Pack-Years
PCA	Principle Component Analysis
PCH	Principle Component of Heritability
PCP	Principle Component of Phenotypes
PCs	Principle Components
PheCLC	Phenome-wide association study using CLC method
PheWAS	Phenome-Wide Association Studies
Pi10	Airway Wall Area
SKAT	Sequence Kernel Association Test
SNP	Single Nucleotide Polymorphism
UDI	Unique Data Identifier

---

# Abstract

Genome-wide association studies (GWAS) have successfully detected tens of thousands of robust SNP-trait associations. Earlier researches have primarily focused on association studies of genetic variants and some well-defined functions or phenotypic traits. Emerging evidence suggests that pleiotropy, the phenomenon of one genetic variant affects multiple phenotypes, is widespread, especially in complex human diseases. Therefore, individual phenotype analyses may lose statistical power to identify the underlying genetic mechanism. Contrasting with single phenotype analyses, joint analysis of multiple phenotypes exploits the correlations between phenotypes and aggregates multiple weak marginal effects and is therefore likely to provide new insights into the functional consequences of genetic variations. This dissertation includes two papers, corresponding to two primary research projects I have done during my Ph.D. study, with each distributed in one chapter.

Chapter 1 proposed an innovative method, which referred to as HC-CLC, for joint analysis of multiple phenotypes using a Hierarchical Clustering (HC) approach followed by a Clustering Linear Combination (CLC) method. The HC step partitions phenotypes into clusters. The CLC method is then used to test the association between the genetic variant and all phenotypes, which is done by combining individual test statistics while taking full advantage of the clustering information in the HC step. Extensive simulations together with the COPDGene data analysis have been used to assess the Type I error rates and the power of our proposed method. Our simulation results demonstrate that the Type I error rates of HC-CLC are effectively controlled in different realistic settings. HC-CLC

either outperforms all other methods or has statistical power that is very close to the most powerful alternative method with which it has been compared. In addition, our real data analysis shows that HC-CLC is an appropriate method for GWAS.

Chapter 2 redesigned the PheCLC (Phenome-wide association study that uses the CLC method) which was previously developed by our research group. The refined method is then applied on the UKBiobank data, a large cohort study across the United Kingdom, to test the validity and understand the limitations of the proposed method. We have named our new method UKB-PheCLC. The UKB-PheCLC method is an EHR-based PheWAS. In the first step, it classifies the whole phenome into different phenotypic categories according to the UK Biobank ICD codes. In the second step, the CLC method is applied to each phenotypic category to derive a CLC-based p-value for testing the association between the genetic variant of interest and all phenotypes in that category. In the third step, the CLC-based p-values of all categories are combined by using a strategy resemble that of the Adaptive Fisher's Combination (AFC) method. Overall, UKB-PheCLC harnesses the powerful resource of the UK Biobank and considers the possibility that phenotypes can be grouped into different phenotypic categories, which is very common in EHR-based PheWAS. Moreover, UKB-PheCLC can handle both qualitative and quantitative phenotypes, and it also doesn't require raw phenotype information. The real data analysis results confirm that UKB-PheCLC is more powerful than the existing methods we have it compared with. Thus, UKB-PheCLC can serve as a compelling method for phenome-wide association study.

# Chapter 1: HC-CLC

## **Joint analysis of multiple phenotypes using a clustering linear combination method based on hierarchical clustering**

Emerging evidence suggests that a genetic variant can affect multiple phenotypes, especially in complex human diseases. Individual phenotype analyses are generally less informative and less powerful for uncovering the genetic variants underlying complex traits and diseases. The joint analysis of multiple phenotypes may offer new insights into disease etiology. In this paper, we develop an innovative method for joint analysis of multiple phenotypes using a hierarchical clustering approach followed by a clustering linear combination method. We have named our method HC-CLC. The proposed method consists of two consecutive steps: a Hierarchical Clustering (HC) step and a testing step using Clustering Linear Combination (CLC). The HC step partitions the original phenotypes into a small number of clusters; phenotypes within each cluster strongly correlate with each other while phenotypes between clusters are less likely to be correlated. The CLC method is then adopted to test the association between a genetic variant of interest and multiple phenotypes, which is done by combining individual test statistics while taking full advantage of the clustering information in the HC step. Extensive simulations together with the COPDGene data analysis have been used to assess the Type I error rates and the power of our proposed method. Our simulation results demonstrate that the Type I error rates of HC-CLC are effectively controlled in different realistic settings. HC-CLC either outperforms all other methods or has statistical power that is very close to the most powerful alternative method with which it has been compared. In addition, our real data

analysis shows that HC-CLC is an appropriate method for genome-wide association studies (GWAS).

## **1.1 Introduction**

Pleiotropy is a well-established phenomenon in which a single locus affects more than one distinct, but possibly correlated, phenotypic traits (Gratten et al., 2016). Pleiotropy has had many important implications on physiological and medical genetics and evolutionary biology (Stearns, 2010). Substantial evidence has shown that pleiotropy is ubiquitous in complex human diseases (Sivakumaran et al., 2011).

Genome-wide association studies (GWAS) have been very successful in detecting genetic variants that are responsible for complex human diseases. To date, the GWAS catalog contains more than 3,600 publications and roughly 90,000 unique SNP-trait associations. Traditional genotype-phenotype association studies focus on the pairwise relationship between phenotypes and genotypes. However, single phenotype analyses ignore the pleiotropic effect and suffer from multiple testing penalties, therefore these analyses may be considerably less powerful for detecting causal variants of weak effects (Sivakumaran et al., 2011).

Contrasting with single phenotype analyses, joint analysis of multiple phenotypes exploits the correlations between phenotypes and aggregates multiple weak marginal effects and is therefore likely to provide important insights into the functional consequences of genetic variations. In addition, multiple correlated disease attributes (also termed disease phenotypes) that relate to clinically meaningful outcomes, such as

symptoms, exacerbations, responses to therapy, rate of disease progression, or death, are often collected and frequently encountered in genetic association studies (Han et al., 2010).

In recent years, there has been increasing interest in jointly testing the association between a single genetic variant and multiple correlated phenotypes, with the null hypothesis that there is no association between the genetic variant of interest and any of the phenotypes while the alternative hypothesis is that the genetic variant of interest is associated with at least one of the phenotypes. The most widely used strategies for those research efforts are combining univariate analysis results, dimension reduction, and regression models. Methods involving the first strategy combine either the univariate test statistics (Kim et al., 2015; Peter C O'Brien, 1984a; Wei et al., 1985) or p-values (Liang et al., 2016; van der Sluis et al., 2013; J. J. Yang et al., 2016). They are generally very easy to implement and can cope with a mixture of different types of phenotypes; however, the statistical power of those methods might heavily rely on the homogeneity of univariate test statistics (H. Zhu et al., 2015a, 2018). The most popular methods in this category include O'Brien's method (Peter C. O'Brien, 1984b; Wei et al., 1985), Trait-based Association Test that uses Extended Simes procedure (TATES) (van der Sluis et al., 2013), Fisher's Combination (J. J. Yang et al., 2016), and Adaptive Fisher's Combination (AFC) (Liang et al., 2016). For the strategy of dimension reduction, instead of testing one phenotype at a time, one first constructs a small number of latent variables, which are linear combinations of the observed phenotypes, and then tests the associations between the latent variables and the genetic variant of interest. Dimension reduction methods are in general suitable only when all phenotypes are normally distributed (Q. Yang et al., 2012a); in addition, the newly

derived latent variables are usually difficult to interpret in the real-world applications. The most popular methods in this category include Principal Component of Phenotypes (PCP) (Aschard et al., 2014), Principal Component of Heritability (PCH) (Klei et al., 2008; Wang et al., 2016; J. J. Zhou et al., 2015), and Canonical Correlation Analysis (CCA) (Ferreira et al., 2008; Tang et al., 2012). Compared with the other two strategies, the strategy involving regression models seems relatively complicated to implement but there have been a lot of R or SAS packages readily available for use. Regression models are able to handle a wide variety of single phenotype data types such as continuous, categorical, or survival, but not a mixture of them (Q. Yang et al., 2012b). Common models in this category include linear and generalized mixed effects models (Korte et al., 2012; Z. Zhang et al., 2009; X. Zhou et al., 2014), frailty models (Wienke, 2010; Q. Yang et al., 2012a), and Generalized Estimating Equations (GEE) (Zeger et al., 1986; Y. Zhang et al., 2014).

Hierarchical Clustering (HC) is a cluster analysis approach that builds a hierarchy of clusters (Ding et al., 2002; Johnson, 1967; Karypis et al., 1999; Rokach et al., 2005). There are two main types of hierarchical clustering: agglomerative (bottom-up) and divisive (top-down). An agglomerative method starts with each phenotype as a single cluster and merges the two clusters that have the smallest distance at each step of the clustering iteration until a given stopping criterion is met or there is only one cluster left. A divisive method starts with all phenotypes belonging to the same cluster and repeatedly partitions a cluster into two such that the distance between the two new clusters are maximized in each step of the iteration until a given stopping criterion is met or each phenotype is in its own singleton cluster. Traditional hierarchical clustering algorithms



often merge or split one cluster at a time. In addition, there are many ways of defining distance between clusters; these distances are often referred to as link functions. The most popular distances used for hierarchical clustering are single-link distance, complete-link distance, and average-link distance (Ding et al., 2002). The arrangement of the clusters generated by hierarchical clustering can be easily visualized in a dendrogram, a tree-like hierarchical taxonomy that records the sequences of merges or partitions. In the world of data mining and statistics, the smallest distance in each step of the iteration is usually referred to as the height of the merged cluster in the dendrogram. For both types of hierarchical clustering approaches, the number of clusters does not need to be specified in advance, however, a termination condition of the clustering process is required. In practice, the clustering iteration stops at the step that provides maximum cluster separation.

Clustering Linear Combination (CLC) is a recently developed approach which combines individual test statistics for joint analysis of multiple phenotypes in association analyses (Sha et al., 2018b). CLC works particularly well with phenotypes that have natural groupings. In the CLC step, individual test statistics are combined linearly within each cluster and cluster-specific effects are then combined in a quadratic form. CLC has shown to be not only robust to different signs of the means of individual statistics, but also reduces the degrees of freedom of the test statistic. In addition, CLC can be theoretically proven to be the most powerful test among all tests that have certain quadratic forms.

In this paper, we provide an innovative method for joint analysis of multiple phenotypes using an HC approach followed by a CLC method. We have named our proposed method HC-CLC. Extensive simulation studies have been conducted to evaluate

the performance of HC-CLC. Five competitive methods for joint analysis of multiple phenotypes, i.e., MANOVA (Cole et al., 1994), MultiPhen (Guo et al., 2015; O’Reilly et al., 2012), TATES (van der Sluis et al., 2013), AFC (Liang et al., 2016), and CLC (Sha et al., 2018b), have been applied for comparison with our proposed method. Our results indicate that HC-CLC can control Type I error rates very well in all simulation scenarios and is either the most powerful method or has statistical power that is very similar to the most powerful method among the five existing methods we have compared it with. We also validate our proposed method by applying it to a COPDGene real dataset.

## 1.2 Methods

Consider a sample of  $n$  unrelated individuals, where each individual has been genotyped at a genetic variant and has  $K$  potentially correlated phenotypes. Let  $\mathbf{x} = (x_1, \dots, x_n)^T$  denote the genotypic score of the  $n$  individuals at the genetic variant of interest, where  $x_i \in \{0, 1, 2\}$  is the number of minor alleles that the  $i^{\text{th}}$  individual carries at the genetic variant. Let  $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_K)$  denote the  $n \times K$  phenotype matrix, where  $\mathbf{Y}_k = (y_{1k}, \dots, y_{nk})^T$  is the  $k^{\text{th}}$  phenotype of the  $n$  individuals.

In this study, we first apply a HC method (Liang et al., 2018) to partition the original  $K$  phenotypes into  $M$  disjoint clusters; the phenotypes within each cluster highly correlate with each other, while the phenotypes between clusters are much less likely to be correlated. More specifically, in the HC step, an agglomerative hierarchical clustering is directly applied on a phenotypic distance matrix  $D$ , whose  $(i, j)^{\text{th}}$  entry  $d_{ij}$  is the distance between the  $i^{\text{th}}$  phenotype and the  $j^{\text{th}}$  phenotype. We use correlation between the two

phenotypes to define the distance,  $d_{ij} = 1 - \text{corr}(\mathbf{Y}_i, \mathbf{Y}_j)$ , where  $\text{corr}$  denotes correlation. We estimate the phenotypic distance matrix  $D$  through the sample correlation matrix of the  $K$  phenotypes. That is,  $\hat{D} = J_K - C^S(\mathbf{Y})$ , where  $J_K$  is a  $K \times K$  matrix and each entry equals 1, and then  $C^S(\mathbf{Y})$  is the sample correlation matrix of the phenotypes.

We define the distance between any two clusters to be the average linkage of the two clusters. For example, the distance between cluster  $C_m$  and cluster  $C_\ell$  is calculated by the following equation:

$$D(C_m, C_\ell) = \frac{1}{|C_m| \cdot |C_\ell|} \sum_{i \in C_m, j \in C_\ell} d_{ij}, \quad (1)$$

where  $|C_m|$  denotes the number of phenotypes in cluster  $C_m$  and where  $|C_\ell|$  denotes the number of phenotypes in cluster  $C_\ell$ . The agglomerative hierarchical clustering starts with each phenotype as a singleton cluster, and then successively merges pairs of clusters that have the smallest distance until a given stopping criterion is met or all clusters have been merged into a single cluster that contains all phenotypes.

We determine the total number of clusters (i.e., the value of  $M$ ) in the HC step by using a stopping criterion that maximizes cluster separation (Bühlmann et al., 2013; Liang et al., 2018). Let  $d_b$  denote the smallest distance between any two clusters in the  $b^{\text{th}}$  step of iteration ( $b \geq 1$ ). If we let

$$\hat{b} = \arg \max_{b \geq 1} (d_{b+1} - d_b), \quad (2)$$

then the total number of clusters yielded at the step  $\hat{b}$  is our desired value for  $M$ .

Next, we incorporate the hierarchical clustering information from the HC step to the CLC method. We refer to the combination of the HC method and the CLC method as HC-CLC. First, we give a brief introduction of the CLC method recently developed by our group (Sha et al., 2018b).

In Sha et al. (Sha et al., 2018b), we developed a statistical method for jointly analysis of multiple phenotypes in association studies. First, we cluster  $K$  phenotypes into  $L$  clusters ( $L = 1, \dots, K$ ) using the hierarchical clustering method with the same distance as we described above. Then we use  $T_{CLC}^L = (WT)^T(W\Sigma W^T)^{-1}(WT)$  to test the association between a genetic variant and the  $K$  phenotypes with  $L$  clusters, where  $T = (T_1, \dots, T_K)^T$  and  $T_k$  is the score test statistic to test the association between the genetic variant and the  $k^{\text{th}}$  phenotype ( $k = 1, \dots, K$ ) under the generalized linear model (Nelder et al., 1972),  $g(E(y_{ik}|x_i)) = \beta_{0k} + \beta_{1k}x_i$ ;  $W = B^T\Sigma^{-1}$ , where  $B$  is a  $K \times L$  matrix with the  $(k, l)^{\text{th}}$  entry denoted by  $b_{kl}$  with  $b_{kl} = 1$  if the  $k^{\text{th}}$  phenotype belongs to the  $l^{\text{th}}$  cluster or otherwise  $b_{kl} = 0$ , and  $\Sigma$  is the variance-covariance matrix of  $T$  and can be estimated by the sample correlation matrix of the  $K$  phenotypes. Under the null hypothesis that none of the phenotypes are associated with the genetic variant,  $T_{CLC}^L$  follows a chi-square distribution with degrees of freedom equal to  $L$ . We use  $T_{CLC} = \min_{1 \leq L \leq K} p_L$  as the final test statistic of CLC, where  $p_L$  denotes the p-value of  $T_{CLC}^L$  for  $L = 1, \dots, K$ . Since  $T_{CLC}$  does not have an asymptotic distribution, we use a simulation procedure to evaluate the p-value of  $T_{CLC}$ .

In this paper, instead of considering all possible number of clusters in HC, we use a stopping criterion to determine the number of clusters. Suppose that the number of clusters using the stopping criterion in HC is  $M$ . Then we can use  $T_{CLC}^M = (WT)^T(W\Sigma W^T)^{-1}(WT)$  to test the association between multiple phenotypes and the genetic variant. Therefore, our HC-CLC test statistics is given by

$$T_{HCCLC} = (WT)^T(W\Sigma W^T)^{-1}(WT), \quad (3)$$

Under the null hypothesis that none of the  $K$  phenotypes are associated with the genetic variant of interest,  $T_{HCCLC}$  follows a chi-square distribution with degrees of freedom equal to  $M$ . Comparing with CLC, HC-CLC does not need to use a simulation procedure to evaluate the p-value, so it is computationally more efficient than CLC.

We compare the performance of our HC-CLC method with those of the other five existing methods: MANOVA, MultiPhen, TATES, AFC, and CLC. Since we have previewed CLC approach in the method section, here, we briefly introduce the other four methods as they apply to the current study.

**MANOVA** (Cole et al., 1994): Consider a multivariate simple linear regression model:  $Y = \mathbf{j}\boldsymbol{\beta}_0^T + \mathbf{x}\boldsymbol{\beta}^T + \boldsymbol{\varepsilon}$ , where  $\mathbf{j}$  is an  $n$ -dimensional vector of all 1's;  $\boldsymbol{\beta}_0$  is a  $K$ -dimensional intercept vector and  $\boldsymbol{\beta}_0 = (\beta_{01}, \dots, \beta_{0K})^T$ ;  $\boldsymbol{\beta}$  is a  $K$ -dimensional vector of coefficients with the  $K$  elements corresponding to the  $K$  phenotypes; while  $\boldsymbol{\varepsilon}$  is an  $n \times K$  residual matrix, with each row following an independent identically distributed (i.i.d.) multivariate normal distribution with mean  $\mathbf{0}$  and a constant variance-covariance matrix. To test  $H_0: \boldsymbol{\beta} = \mathbf{0}$ , the Wilk's Lambda test statistic is commonly used, which is equivalent

to the test statistic of the likelihood ratio test. Under the null hypothesis, the MANOVA test statistic has an asymptotic  $\chi_K^2$  distribution.

**MultiPhen** (O'Reilly et al., 2012): MultiPhen uses the ordinal regression (also known as proportional odds logistic regression) and inverts the general linear regression model of a single phenotype on multiple genotypes. That is, MultiPhen treats the genetic variant of interest as an ordinal response variable and the multiple correlated phenotypes as regressors. A likelihood ratio test is then used to test association between the genetic variant and the phenotypes. The resulting test statistic asymptotically follows a chi-square distribution with degrees of freedom equals to the number of phenotypes ( $K$ ).

**TATES** (van der Sluis et al., 2013): TATES combines phenotype-specific p-values obtained from standard univariate GWAS while considering the correlations between components. Denote  $p_k$  the p-value of the test statistic to test the association between the  $k^{\text{th}}$  phenotype and the genetic variant,  $p_{(k)}$  the  $k^{\text{th}}$  smallest p-value among all  $p_k$ 's, where  $k = 1, 2, \dots, K$ . Then, the p-value of TATES is given by  $\min_{1 \leq k \leq K} \left( \frac{m_e p_{(k)}}{m_{e(k)}} \right)$ , where  $m_e$  represents the effective number of independent p-values among all  $K$  p-values, and  $m_{e(k)}$  represents the effective number of independent p-values among the first smallest  $k$  p-values.

**AFC** (Liang et al., 2016): The AFC method combines p-values obtained in standard univariate GWAS by using the optimal number of p-values which is determined by the data. Using *the* same notations in TATES, let  $p_{(k)}$  denote the  $k^{\text{th}}$  smallest p-value among all  $p_k$ 's, where  $k = 1, 2, \dots, K$ , and let  $p_{T_k}$  denote the p-value of  $T_k$ . The statistic of AFC

is given by  $T_{AFC} = \min_{1 \leq k \leq K} p_{T_k}$ . A permutation procedure is then used to evaluate the p-values of  $T_{AFC}$ .

## SIMULATION STUDIES

### Simulation settings

To evaluate the Type I error rates and the statistical power of our proposed method, we simulate genotype and phenotype data for  $n$  unrelated individuals. The genotype of each individual at a variant of interest is generated based on minor allele frequency (MAF) assuming Hardy Weinberg equilibrium. The phenotypes of each individual are generated according to the following factor model (Wang et al., 2016)

$$\mathbf{y} = \boldsymbol{\lambda}x + c\boldsymbol{\gamma}\mathbf{f} + \sqrt{1 - c^2} \times \boldsymbol{\varepsilon} \quad (4)$$

where  $\mathbf{y}$  is a vector of phenotypes and  $\mathbf{y} = (y_1, \dots, y_K)^T$ ;  $x$  is the genotypic score at the genetic variant;  $\boldsymbol{\lambda}$  is the effect sizes of the genetic variant on the  $K$  phenotypes and  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)^T$ ;  $c$  is a constant;  $\boldsymbol{\gamma}$  is a  $K \times R$  block diagonal matrix used for setting up various simulation scenarios;  $\mathbf{f}$  is a vector of factors and  $\mathbf{f} = (f_1, \dots, f_R)^T \sim MVN_R(\mathbf{0}, \boldsymbol{\Sigma})$ , where  $R$  is the number of factors,  $\boldsymbol{\Sigma} = (1 - \rho)\mathbf{I} + \rho\mathbf{J}$ ,  $\mathbf{I}$  is an identity matrix,  $\mathbf{J}$  is a matrix with elements of all 1's, and  $\rho$  is the correlation between factors;  $\boldsymbol{\varepsilon}$  is a vector of residuals and  $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_K)^T$ , where  $\varepsilon_1, \dots, \varepsilon_K$  are pairwise independent and each  $\varepsilon_k \sim N(0, 1)$ ,  $k = 1, \dots, K$ . Based on equation (4), we consider the following four models for which the within-factor correlation equals  $c^2$  and the between-factor correlation equals  $\rho c^2$ .

**Model 1:** There is only one factor and the genetic variant impacts all of the  $K$  phenotypes but with different effect sizes. That is,  $R = 1$ ,  $\lambda = \beta(1, 2, \dots, K)^T$ , and  $\gamma = (1, \dots, 1)^T$ .

**Model 2:** There are two factors and the genetic variant impacts one of the factors. That is,

$$R = 2, \lambda = \left( 0, \dots, 0, \underbrace{\beta, \dots, \beta}_{K/2} \right)^T, \text{ and } \gamma = \text{bdiag}(D_1, D_2), \text{ where "bdiag" indicates the}$$

$$\text{block diagonal matrix, and } D_i = \left( \underbrace{1, \dots, 1}_{K/2} \right)^T \text{ for } i = 1, 2.$$

**Model 3:** There are five factors and the genetic variant impacts two of the factors. That is,

$$R = 5, \lambda = (\beta_{11}, \dots, \beta_{1k}, \beta_{21}, \dots, \beta_{2k}, \beta_{31}, \dots, \beta_{3k}, \beta_{41}, \dots, \beta_{4k}, \beta_{51}, \dots, \beta_{5k})^T, \text{ and } \gamma =$$

$$\text{bdiag}(D_1, D_2, D_3, D_4, D_5), \text{ where } D_i = \left( \underbrace{1, \dots, 1}_{K/5} \right)^T \text{ for } i = 1, \dots, 5, k = \frac{K}{5}, \beta_{11} = \dots \beta_{1k} =$$

$$\beta_{21} = \dots = \beta_{2k} = \beta_{31} = \dots = \beta_{3k} = 0, \beta_{41} = \dots = \beta_{4k} = -\beta, \text{ and } (\beta_{51}, \dots, \beta_{5k}) = \frac{2\beta}{k+1} * (1, \dots, k).$$

**Model 4:** There are five factors and the genetic variant impacts four of the factors. That is,

$$R = 5, \lambda = (\beta_{11}, \dots, \beta_{1k}, \beta_{21}, \dots, \beta_{2k}, \beta_{31}, \dots, \beta_{3k}, \beta_{41}, \dots, \beta_{4k}, \beta_{51}, \dots, \beta_{5k})^T, \text{ and } \gamma =$$

$$\text{bdiag}(D_1, D_2, D_3, D_4, D_5), \text{ where } D_i = \left( \underbrace{1, \dots, 1}_{K/5} \right)^T \text{ for } i = 1, \dots, 5, k = \frac{K}{5}, \beta_{11} = \dots \beta_{1k} =$$

$$0, \beta_{21} = \dots = \beta_{2k} = \beta, \beta_{31} = \dots = \beta_{3k} = -\beta, \beta_{41} = \dots = \beta_{4k} = -\frac{2\beta}{k+1} * (1, \dots, k), \text{ and}$$

$$(\beta_{51}, \dots, \beta_{5k}) = \frac{2\beta}{k+1} * (1, \dots, k).$$

To evaluate Type I error rates, we let  $\beta = 0$ ,  $MAF = 0.3$ , and vary the significance level  $\alpha$ , the total number of phenotypes  $K$ , and sample size  $n$ . To evaluate statistical power,



we set sample size  $n = 5,000$ ,  $MAF = 0.3$ , and vary the values of  $\beta$ , the total number of phenotypes  $K$ , the within-factor correlation  $c^2$ , and the between-factor correlation  $\rho c^2$ .

### **Simulation results**

In each simulation scenario, the p-values of the test statistics of MANOVA, MultiPhen, TATES, and HC-CLC are estimated based on their asymptotic distributions. The p-values of AFC and CLC are estimated using 10,000 permutations.

We use 10,000 replicated samples to evaluate the Type I error. For 10,000 replicated samples, the 95% confidence intervals (CIs) for Type I error rates at the nominal levels 0.05, 0.01, and 0.001 are (0.0457, 0.0543), (0.0080, 0.0120), and (0.0004, 0.0016), respectively. Table 1.1 summarizes the estimated Type I error rates of HC-CLC. The results indicate that all of the estimated Type I error rates are within the 95% CIs, which confirms that HC-CLC is a valid method.

For power comparisons, we use a 5% significance level. The power of each method under four different models is estimated using 1,000 replicated samples. Figure 1.1 and 1.2 provide the power comparisons of the six methods (MANOVA, MultiPhen, TATES, AFC, CLC, and HC-CLC) as a function of genetic effect size  $\beta$ , with 20 phenotypes and 40 phenotypes, respectively. From these two figures, we can see that HC-CLC is the most powerful test among all six methods we compared with under the four models. CLC is the second most powerful method among the six methods. MANOVA and MultiPhen have similar statistical power under all four models; these two methods have power close to the most powerful test (HC-CLC) under model 1 but are the least powerful methods under

model 2 where the genetic variant has effect on only half of the phenotypes. TATES and AFC have similar power under models 1 to 3. Under model 4, where the genetic variant has effect on part of the phenotypes with different the effect sizes and different directions of the effects, AFC performs better than TATES. TATES and AFC are the least powerful methods under model 1 and perform better than MANOVA and MultiPhen under model 2. As anticipated, when the effect size increases, the statistical powers of all six methods increase.

Figure 1.3 and 1.4 provide the power comparisons of the six methods as a function of the within-factor correlation  $c^2$ , with 20 phenotypes and 40 phenotypes, respectively. The pattern of the powers of the six models are similar to these observed in Figure 1.1 and 1.2 except that under model 1, when the within-factor correlation is small ( $< 0.2$ ), AFC and CLC are more powerful than the other four methods. In general, the powers of all six methods decreases as the within-factor correlation increases.

We also provide the power comparisons of the six methods as a function of the between-factor correlation  $\rho c^2$ , with 20 phenotypes and 40 phenotypes, respectively (Figures B.1.1 and B.1.2). The pattern of the powers of the six models are similar to these observed in Figure 1.1 and 1.2. When the between-factor correlation  $\rho c^2$  increases, the statistical power of the six methods stay at almost the same levels, indicating that these methods are not sensitive to the changes in between-factor correlation, as they are to the changes due to within-factor correlation.

## 1.3 Real Data Analysis

### 1.3.1 COPD Data Set

Chronic Obstructive Pulmonary Disease (COPD) is a chronic inflammatory lung disease that obstructs airflow (Chu et al., 2014b). Possible signs and symptoms of COPD include breathing difficulty, cough, mucus (sputum) production, and wheezing (Chung et al., 2008). People with COPD have a higher risk of developing heart disease, lung cancer, and numerous other afflictions. Though cigarette smoking is commonly recognized as a trigger for COPD, genetic risk factors also seem to play an important role in the development of the disease (Mannino et al., 2007; Pillai et al., 2009; Regan et al., 2011; Sandford et al., 1997; Schellenberg et al., 1998; Silverman et al., 1998; Silverman et al., 2004). The COPDGene is one of the largest studies to uncover the underlying genetic factors of COPD and other smoking-related diseases; important information that is routinely applied to develop new therapeutic approaches to cure those diseases. There was a total of 10,192 smokers who were potentially affected by COPD, of which 6,784 were non-Hispanic white and 3,408 were African-American, recruited for the COPDGene study (Chu et al., 2014b).

In this paper, we apply six methods, MANOVA, MultiPhen, TATES, AFC, CLC, and HC-CLC, to the non-Hispanic white cohort of the COPDGene study to discover genetic variants associated with COPD-related phenotypes. Following a similar study (Liang et al., 2016) and the literature on which it was based, we select seven quantitative COPD-related phenotypes and four covariates. The seven phenotypes are % predicted FEV1 (FEV1), Emphysema (Emph), Emphysema Distribution (EmphDist), Gas Trapping (GasTrap), Airway Wall Area (Pi10), Exacerbation Frequency (ExacerFreq), and Six-

minute Walk Distance (6MWD). The details of these seven phenotypes are shown in Figure A.1.1 (Chu et al., 2014a). With reference to a previous study (Chu et al., 2014a), we perform a log transformation on the phenotype EmphDist. The correlation plot of the seven phenotypes is given in Figure B.1.3. We also change the signs of phenotypes FEV1 and 6MWD because their correlations with the other five phenotypes are negative. After this modification, the pair-wise correlations between phenotypes are all positive. The four covariates considered in this study are BMI, Age, Pack-Years (PackYear), and Sex. We eliminate participants with missing SNPs and missing values in any of the 11 variables. After the data preprocessing steps, there remains a total of 5,430 subjects and 630,860 SNPs. For each of the seven phenotypes, we adjust the phenotype values for the four covariates through a linear regression (Sha et al., 2018a). All the subsequent analyses are based on the adjusted phenotypes. To identify SNPs associated with the seven COPD-related phenotypes, we use the standard genome-wide significance p-value threshold of  $5 \times 10^{-8}$  to account for multiple testing.

### **1.3.2 Real Data Analysis Results**

In the COPDGene real data analysis, HC divides the seven phenotypes into five clusters, that is,  $M = 5$ . The dendrogram of HC on the seven phenotypes is given in Figure B.1.4. One of the clusters contains phenotypes FEV1, Emph, and GasTrap, while the other four clusters each contain only a single phenotype. This finding is aligned with a previous study, which showed that GasTrap is a “hub” in the phenotypic network; the pairings of GasTrap with FEV1 and GasTrap with Emph are both highly correlated in the race-specific networks (Chu et al., 2014b). Table 1.2 summarizes the 14 significant SNPs that have been identified

by at least one of the six methods. All of these 14 SNPs were previously reported to be associated with COPD (Brehm et al., 2011; Cho et al., 2010; Cho et al., 2014; Cui et al., 2014; Du et al., 2016; Hancock et al., 2010; Li et al., 2011; Liang et al., 2018; Lutz et al., 2015; Pillai et al., 2009; Sha et al., 2018a; Sha et al., 2018b; Wilk et al., 2009; Wilk et al., 2012; Young et al., 2010; J. Zhang et al., 2011; A. Z. Zhu et al., 2014). As seen in Table 2, MultiPhen identifies 14 SNPs; HC-CLC, CLC, and MANOVA each identifies 13 SNPs; AFC identifies 12 SNPs; and TATES identifies nine SNPs. The reason that TATES only identifies nine SNPs may be due to the fact that this method depends heavily on the smallest among the seven p-values from the univariate analysis results. The results of the real data analysis are consistent with the findings from our simulations, corroborating that the HC-CLC has similar or superior performance to that of the other five methods.

#### **1.4 Discussion**

In this paper, we develop a novel method, HC-CLC, for joint analysis of multiple phenotypes in genetic association studies. HC-CLC has several important advantages over existing methods. First of all, HC-CLC takes advantage of the natural grouping information of the phenotypes, which can be easily obtained from hierarchical clustering. In addition, HC-CLC is easy to implement and computationally efficient for GWAS. HC-CLC avoids the computational burden of AFC and CLC, which use permutation to evaluate the p-values of their test statistics; instead, HC-CLC has an asymptotic distribution. Furthermore, HC-CLC does not require access to individual phenotypes themselves; it only requires a distance matrix of phenotypes. When individual phenotype data is not available, this

distance matrix of phenotypes can be estimated from the summary statistics of univariate GWAS (X. Zhu et al., 2015b).

Our simulation results demonstrate that HC-CLC has controlled Type I error rates effectively and almost always exceeds the other five competing methods in terms of statistical power within various simulation scenarios. Additionally, the real data analysis results indicate that HC-CLC has great potential for GWAS.

In this study, we use the bottom-up hierarchical clustering method to cluster phenotypes in the HC step. For future studies, we can explore other clustering approaches and incorporate the corresponding clustering information into the CLC step.

In addition, we choose the average linkage as the distance between two clusters in this study. In fact, we can also choose other linkages; for example, single linkage, complete linkage, or average linkage. However, the outcome of applying other linkages within our proposed method still needs further investigation.

## 1.5 Tables and Figures

**Table 1.1.** The estimated Type I error rates of HC-CLC, where MAF is 0.3, and the number of replications is 10,000.  $K$  is the number of phenotypes, and  $\alpha$  is the significance level. All estimated Type I error rates are within the corresponding 95% CIs.

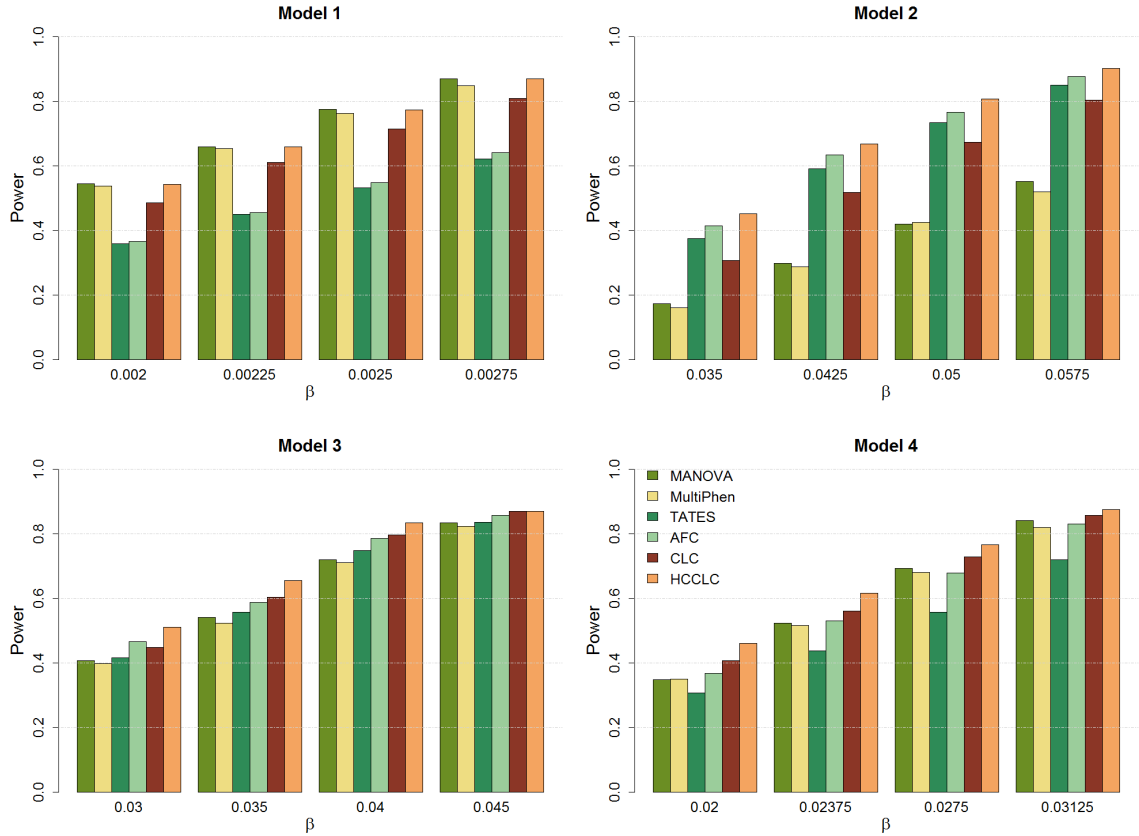
$K$	$\alpha$	Sample Size	Model			
			1	2	3	4
20	0.050	2000	0.0473	0.0528	0.0516	0.0477
		5000	0.0505	0.0482	0.0520	0.0493
	0.010	2000	0.0099	0.0114	0.0104	0.0100
		5000	0.0111	0.0099	0.0104	0.0106
	0.001	2000	0.0008	0.0007	0.0011	0.0011
		5000	0.0014	0.0011	0.0009	0.0010
40	0.050	2000	0.0497	0.0481	0.0489	0.0498
		5000	0.0484	0.0495	0.0520	0.0470
	0.010	2000	0.0096	0.0083	0.0098	0.0113
		5000	0.0090	0.0114	0.0110	0.0099
	0.001	2000	0.0004	0.0010	0.0005	0.0011
		5000	0.0008	0.0012	0.0004	0.0009

**Table 1.2.** Significant SNPs and the corresponding p-values in the COPDGene real data analysis. The p-values of MANOVA, MultiPhen, TATES, and HC-CLC are estimated based on their asymptotic distributions and the p-values of AFC and CLC are estimated using 10,000 permutations. The graying out p-values indicate values greater than  $5 \times 10^{-8}$ .

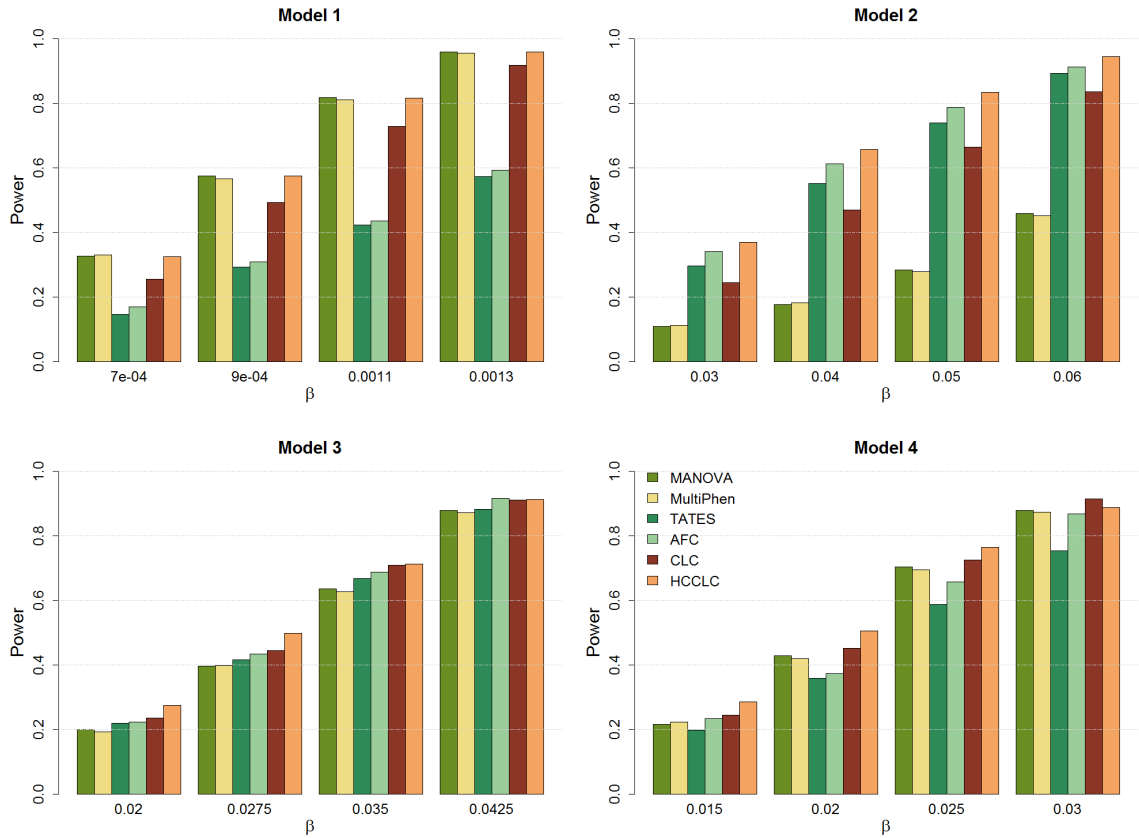
Chr	Position	Variant identifier	MANOVA	MultiPhen	TATES	AFC	CLC	HC-CLC
4	145431497	rs1512282	1.69E-09	1.03E-09	5.77E-09	1.10E-08	0	5.28E-10
4	145434744	rs1032297	6.52E-14	7.69E-14	6.22E-13	0	0	1.17E-13
4	145474473	rs1489759	1.11E-16	1.22E-16	2.52E-16	0	0	0
4	145485738	rs1980057	6.68E-17	8.14E-17	9.35E-17	0	0	0
4	145485915	rs7655625	7.12E-17	9.13E-17	1.64E-16	0	0	0
15	78882925	rs16969968	1.32E-11	7.84E-12	2.98E-08	0	0	6.18E-11
15	78894339	rs1051730	1.41E-11	8.16E-12	2.63E-08	0	0	3.18E-11
15	78898723	rs12914385	1.76E-12	1.48E-12	5.14E-10	0	0	1.09E-12
15	78911181	rs8040868	2.74E-12	2.59E-12	2.40E-09	0	0	2.96E-12
15	78878541	rs951266	1.77E-11	1.02E-11	5.17E-08	0	0	6.36E-11
15	78806023	rs8034191	2.14E-10	7.74E-11	1.02E-07	1.40E-08	0	8.08E-10
15	78851615	rs2036527	3.99E-10	1.77E-10	1.56E-07	2.90E-08	8.33E-10	1.21E-09
15	78826180	rs931794	2.35E-10	9.09E-11	1.18E-07	6.30E-08	0	3.63E-09
15	78740964	rs2568494	1.05E-07	4.23E-08	2.88E-05	5.00E-06	3.98E-07	1.23E-06



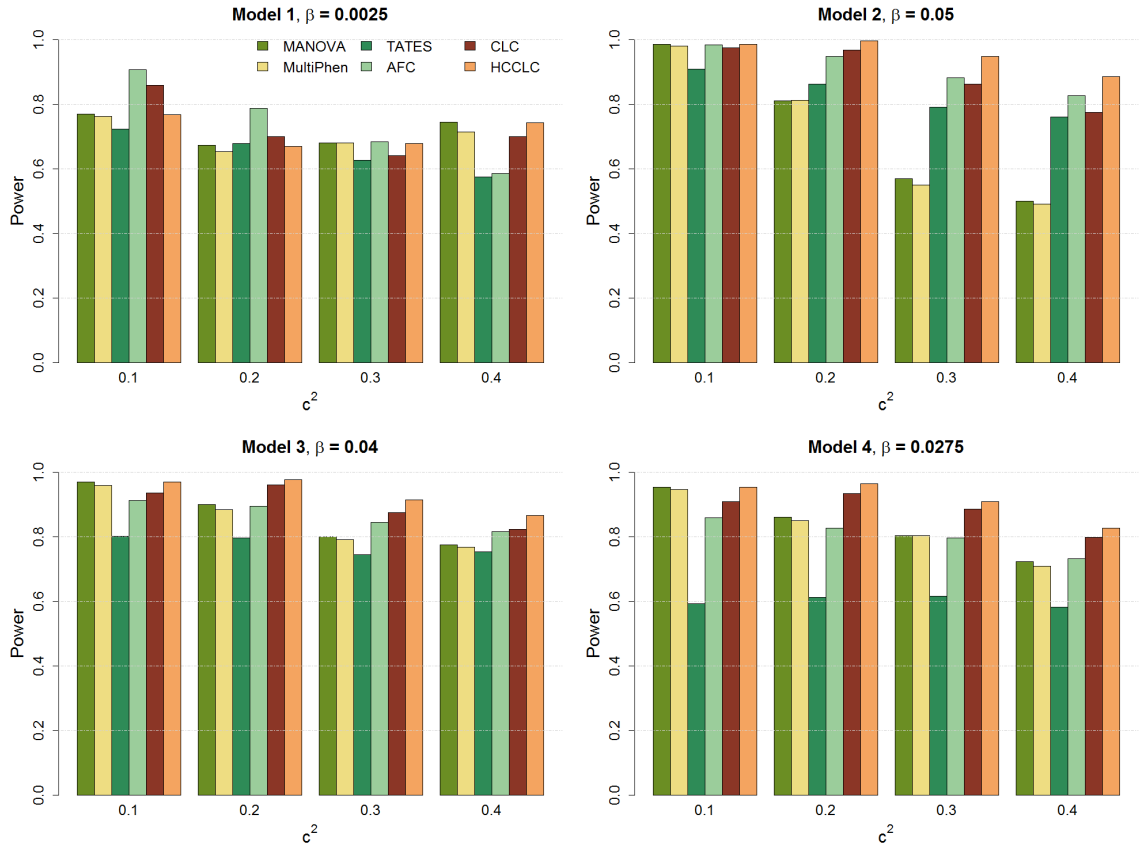
**Figure 1.1.** The power comparisons of the six methods for 20 quantitative phenotypes assessed at a 5% significance level. Statistical power varies with the effect size  $\beta$ , where MAF is 0.3, the sample size is 5,000, the number of replications is 1,000, the within-factor correlation is 0.5 ( $c^2 = 0.5$ ), and the between-factor correlation is 0.1 ( $\rho c^2 = 0.1$ ).



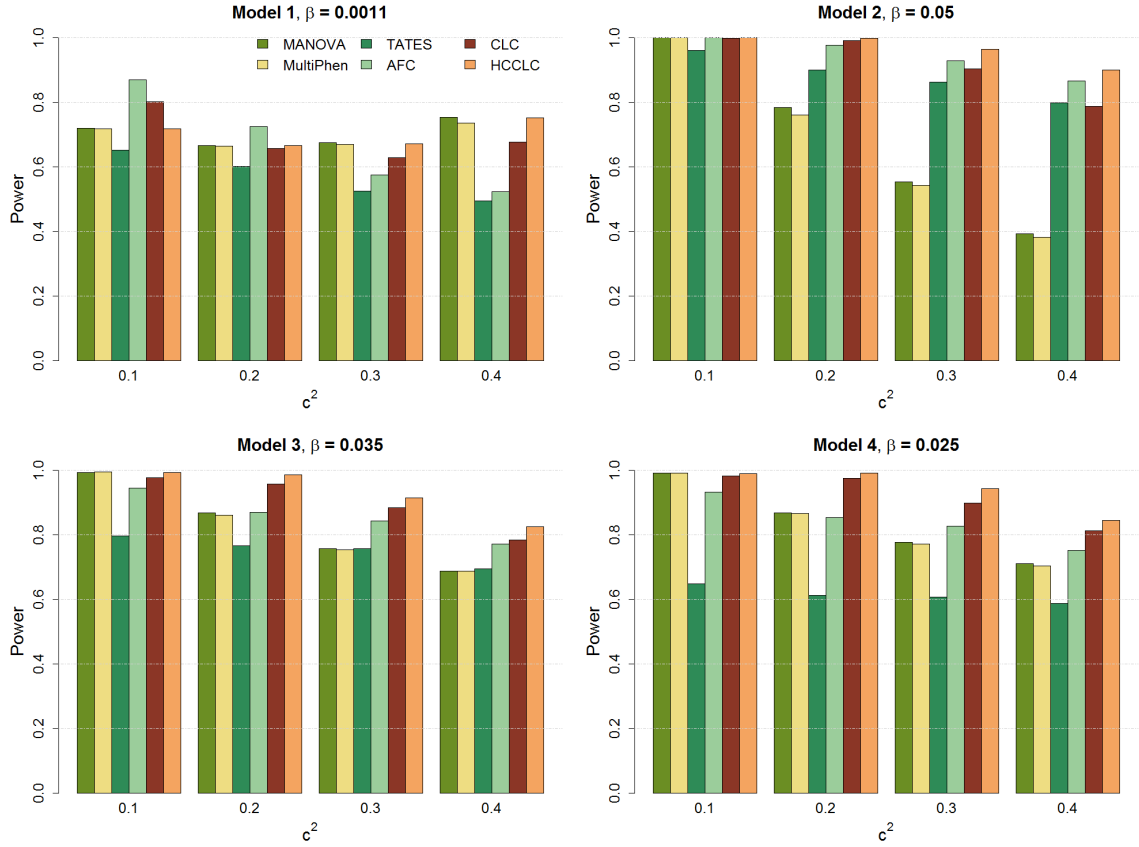
**Figure 1.2.** The power comparisons of the six methods for 40 quantitative phenotypes assessed at a 5% significance level. Statistical power varies with the effect size  $\beta$ , where MAF is 0.3, the sample size is 5,000, the number of replications is 1,000, the within-factor correlation is 0.5 ( $c^2 = 0.5$ ), and the between-factor correlation is 0.1 ( $\rho c^2 = 0.1$ ).



**Figure 1.3.** The power comparisons of the six methods for 20 quantitative phenotypes assessed at a 5% significance level. Statistical power varies with the within-factor correlation  $c^2$ , where MAF is 0.3, the sample size is 5,000, the number of replications is 1,000, and the between-factor correlation is 0.1 ( $\rho c^2 = 0.1$ ).



**Figure 1.4.** The power comparisons of the six methods for 40 quantitative phenotypes assessed at a 5% significance level. Statistical power varies with the within-factor correlation  $c^2$ , where MAF is 0.3, the sample size is 5,000, the number of replications is 1,000, the between-factor correlation is 0.1 ( $\rho c^2 = 0.1$ ).



## Chapter 2: UKB-PheCLC

### Application of UKBiobank Data for Phenome-Wide Association Studies

With the development of high throughput, massively parallel sequencing technologies, GWAS has become a very effective tool in identifying genetic components underlying complex diseases. To date, the GWAS catalog contains more than 3,600 publications and roughly 90,000 unique SNP-trait associations. The greatest advantage of GWAS is that it can discover novel genes and pathways involved in disease pathogenesis. The greatest limitation of GWAS is that it primarily focuses on a pre-defined and limited phenotypic domain. A complementary approach to GWAS is the PheWAS, in which the association between genomic markers and a diverse range of phenotypes are investigated. PheWAS has recently become feasible due to the wide availability of the electronic health records (EHR), which usually involves using the International Classification of Disease (ICD) codes, a standardized coding system for defining disease status as well as for billing purpose. The UK Biobank is a population-based cohort study with a wide variety of genetic and phenotypic information collected on ~ 500K participants from multiple sites across the United Kingdom, aged between 40 and 69 years when recruited in 2006–2010 (Sudlow et al., 2015). In this manuscript, we have redesigned the PheCLC (Phenome-wide association study that uses Clustering Linear Combination) method which was previously developed by our research group. The refined method is then applied on the UKBiobank data to test the validity and understand the limitations of the proposed method. We have named our new method UKB-PheCLC. The UKB-PheCLC method is an EHR-based PheWAS. In the first step, it classifies all phenotypes (the whole phenome) into numerous phenotypic

categories according to the UK Biobank ICD-10 level 2 code. In the second step, the Clustering Linear Combination (CLC) method is applied to each phenotypic category to derive a CLC-based p-value for testing the association between the genetic variant of interest and all phenotypes in that category. In the third step, the CLC-based p-values of all categories are combined by using a strategy resemble that of the Adaptive Fisher's Combination (AFC) method. The biggest advantage of UKB-PheCLC is that it takes into account the possibility that phenotypes are from different phenotypic categories, which is very common and readily available in EHR-based PheWAS. Moreover, UKB-PheCLC can handle both qualitative and quantitative phenotypes since we only need to classify the univariate test statistics. Following the same logic, UKB-PheCLC doesn't require raw phenotype information and it can work on summary test statistics from other studies. Furthermore, the permutation procedure that UKB-PheCLC adopted to generate the empirical null distribution of the final test-statistic only needs to be done once for different genetic variants. The real data analysis results confirm that the proposed method is more powerful than the existing methods we have it compared with. Thus, UKB-PheCLC can serve as a compelling method for phenome-wide association study.

## 2.1 Introduction

A full understanding of the impact of genetics on phenotypic and disease variation, and its potential interactions with other factors is very crucial and is in urgent need in the scientific community as it helps us understand the etiology behind complex diseases and provides important information on precise medicine development. Technological advances have made genome sequencing a reality and open up many new possibilities for identifying genetic variants associated with complex diseases. To date, the GWAS catalog collects more than 3,600 publications and roughly 90,000 unique SNP-trait associations (<https://www.ebi.ac.uk/gwas/>). GWAS has enjoyed its popularity for its capability of discovering novel genes and pathways involved in disease pathogenesis. GWAS commonly starts with a single phenotype and tests the genetic association between the phenotype of interest and a broad spectrum of genetic variants across the genome. In contrast to GWAS, PheWAS starts with a single genetic variant and test the genetic association between the genetic variant of interest and a wide range of phenotypes across the phenome.

PheWAS has recently become feasible due to the wide availability of the electronic health records (EHR), which usually involves using the International Classification of Disease (ICD) codes. ICD coding system is an international standard for reporting diseases and health conditions. In addition to indicate disease status, ICD codes are also widely used by hospitals and insurance companies for billing purpose. Currently, there are two versions of ICD codes, i.e., ICD-9 and ICD-10, where the ICD-10 is an updated version of ICD-9 and can be used to include more diseases types. The two versions of codes are not one-to-

one exchangeable. As the emerging of EHR data, researchers have come to the realization of the importance of mining the information contained in the ICD codes to aid in searching for robust SNP-disease associations.

UK Biobank is a large cohort study with deep genetic and phenotypic information collected on ~500K participants between age 40-69 (Bycroft et al., 2018). At recruitment, participants have provided detailed information about themselves (e.g., socio-demographics, lifestyle, health-related factors), undergone a wide range of physical measures, donated blood, urine and saliva samples for analysis, and signed consent to have their health information followed (i.e., allow follow-up through linkage to their health-related records). The primary goal of UK Biobank is to help improve the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses, for example, cancer, heart diseases, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia.

There are two general approaches for PheWAS, i.e., univariate approach and multivariate approach. Univariate approach tests the association between the genetic variant of interest and each phenotype individually and use Bonferroni correction to adjust multiple testing. Numerous studies in GWAS have shown that univariate tests have some intrinsic drawbacks and are not as powerful as multivariate tests. Moreover, emerging evidence suggests that pleiotropy (Gratten et al., 2016), the phenomenon of one genetic variant affect multiple phenotypes, is widespread, especially in complex human diseases. Therefore, analyzing one phenotype at a time may lose statistical power to identify the underlying genetic mechanism, especially when causal variants have weak effects



(Sivakumaran et al., 2011). By contrast, multivariate approaches test the association between the genetic variant of interest and all the phenotypes across the phenome jointly, which is likely to boost the power performance of association testing.

In this manuscript, we have redesigned the PheCLC (Phenome-wide association study that uses Clustering Linear Combination) method which was previously developed by our research group. The refined method is then applied on the UKBiobank data to test the validity and understand the limitations of the proposed method. We have named our new method UKB-PheCLC. The UKB-PheCLC method is an EHR-based PheWAS. In the first step, it classifies all phenotypes (the whole phenome) into numerous phenotypic categories according to the UK Biobank ICD-10 level 2 code (<http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=41202&nl=1>). In the second step, the Clustering Linear Combination (CLC) method (Sha et al., 2018b) is applied to each phenotypic category to derive a CLC-based p-value for testing the association between the genetic variant of interest and all phenotypes in that category. In the final step, the CLC-based p-values of all categories are combined by using a strategy resemble that of the Adaptive Fisher's Combination (AFC) method (Liang et al., 2016).

We have compared the performance of UKB-PheCLC with that of two popular multivariate analysis methods commonly used in GWAS. That is, the Trait-based Association Test that uses Extended Simes procedure (TATES) (van der Sluis et al., 2013) and the O'Brien's method (OB) (O'Brien, 1984). The two methods mentioned above either combine univariate p-values or univariate test statistics. They are very similar to our proposed UKB-PheCLC method. Our results confirm that the proposed method is more

powerful than the existing methods we have it compared with. Thus, UKB-PheCLC can serve as an alternative multivariate method for joint analysis of multiple phenotypes. Furthermore, we have demonstrated the feasibility of UKB-PheCLC in phenome-wide association study.

In summary, we have showed the following advantage of UKB-PheCLC in this study.

1. UKB-PheCLC can harness the powerful data sources, e.g., a wide of range of electronic health records, of the UK Biobank.
2. UKB-PheCLC considers the possibility that phenotypes can be grouped into different phenotypic categories, which is very common in EHR-based PheWAS.
3. Moreover, UKB-PheCLC can handle both qualitative and quantitative phenotypes because we only need to classify the univariate test statistics. Following the same logic, UKB-PheCLC doesn't require raw phenotype information and it can work on summary test statistics from other studies.
4. In addition, the permutation procedure UKB-PheCLC adopted to generate the empirical null distribution of its final test-statistic only needs to be done once for different genetic variants.

## 2.2 UK Biobank Data Pre-processing

### 2.2.1 Introduction of UK Biobank Phenotypes

Phenotype data can be downloaded from the UK Biobank Showcase system, it needs decryption before use. The data can be converted into different formats (e.g., csv, docs, sas, stata, or r). File size is ~12GB.

The phenotype data is stored in a matrix-like form, with each row corresponding to an individual, each column corresponding to a feature/property of individuals. Each individual has an eid number and 7197 data fields. The eid numbers are anonymous identities of participants and the data fields are the Unique Data Identifiers (UDIs) of the participants' phenotypic information. Each data field is composed of three components: field ID, instant index, and array index. All data fields are labelled with the format "FieldID-InstantIndex.ArrayIndex". Instant index indicates assessment instance (or visit). Array index indicates multiple answers to the same question. It is worth noting that all indices start from the value of 0, rather than 1.

Let's use data field 41202-0.0 as an example to demonstrate the components and the meaning of each component of a data field. The first component 41202 is the field ID, represents "Diagnoses - main ICD10". The second component 0 is the instant index, represents "baseline measurement". The third component 0 is the array index, represents "first measurement taken". If we extend this example to a more general situation, an instant index of 1 can be used to indicate repeated measurements, an instant index of 2 can be used to indicate imagine measurements. In UK Biobank, data fields of phenotypes always

appear in instance and array order. i.e., all the measurements taken on the first instance appear first, followed by the second, third, fourth instance and so on. Moreover, UK Biobank releases its phenotypic and genotypic information separately. But an individual's phenotypic information and genotypic information can be linked through eid number in phenotype data, and FID (family ID) and IID (individual ID) in genotype data. For instance, a patient with eid 1000018 in the phenotype will have an FID 1000018 and IID 1000018 in the genotype data. Here, FID and IID are the same for this individual because we assume all the participants in the study are unrelated.

There are a total of 972 distinct fields, covering 502,591 participants. In this study, we define phenotypes based on all data fields in the field 41202 (Diagnoses - main ICD10) and field 41204 (Diagnoses - secondary ICD10). The array indices for field 41202 run from 0 to 379. The array indices for field 41204 run from 0 to 434. That is, we consider data fields 41202-0.0, 41202-0.1, 41202-0.2, ..., 41202-0.379 and 41204-0.0, 41204-0.1, 41204-0.2, ..., 41204-0.434. A sample of phenotype data in the field 41202 is shown in table B.1.6, and a sample of phenotype data in the field 41204 is shown in table B.1.7.

Next, we convert the selected data fields which are in the format of ICD-10 codes to case-control phenotypes.

**Step 1:** Trunk each full ICD-10 code to UK Biobank ICD-10 level 2 code. For example, we convert Z36.3, K50, D25.92 to Z36, K50, D25, respectively.

**Step 2:** Let each unique truncated ICD code be a column name of phenotype. In our study, we have a total of 1869 unique truncated ICD codes, thus we will have 1869 unique

phenotypes, with the column names of these phenotypes being the unique truncated ICD codes.

**Step 3:** For each individual, if a certain truncated ICD code ever appears, we denote the disease status for that individual as “1” for that phenotype, otherwise, we denote the disease status for that individual as “0” for that phenotype.

### **2.2.2 Introduction of UK Biobank Genotypes**

There are two types of genotype data available in UK Biobank. The regular genotype data (in Binary PED format) and the imputed genotype data (in Oxford format). Both types of genotypes are segmented into different chromosomes. 488,377 participants have regular genotype data. 487,327 participants have imputed genotype data.

Registered researchers can download the genotype information either using the ukbgene utility or from the EGA website. The size of the regular genotype ranges from 1.3-7.3GB per chromosome. The size of the imputed genotype ranges from 36.4-188GB per chromosome. In this study, we only consider the regular genotyped data and the SNPs located in autosomal chromosomes. To avoid the heavy computation burden, we further restrict the SNPs of interest to GWAS Catalog significant SNPs. That is SNPs with p-values less than  $5 \times 10^{-8}$ . As of Oct. 21<sup>st</sup>, 2018, GWAS catalog contains 3640 publications, 62099 SNPs, and 78161 unique SNP-trait associations. Digging further into the GWAS Catalog, we have a total of 90428 data entries, covering 61613 unique SNPs. Among all entries, 49451 of them with p-values less than  $5 \times 10^{-8}$ , including 29297 unique SNPs. For the rest of data process and analysis, we only consider the 29297 significant SNPs.

### 2.2.3 Introduction of Covariates of UKB-PheCLC

In this study, we consider age, sex, genotyping array, and the first 10 genetic principal components (PCs) as covariates. UK Biobank genetic data were assayed using two different genotyping arrays, the UK BiLEVE Axiom Array and UK Biobank Axiom Array (Bycroft et al., 2018). Participants assayed by UK BiLEVE Axiom Array was primarily recruited to study lung diseases. Marker contents of UK Biobank Axiom Array was designed to capture genome-wide genetic variation (SNPs) and short insertions and deletions (indels). Thus, it is important to adjust the variations in samples when performing association studies.

### 2.2.4 Quality Controls

Next, we performed quality controls (QCs) on both markers and samples. The detailed steps are shown below.

**Step 1:** Preprocess genotype data using the following criteria.

**--geno 0.05:** filters out variants with missingness exceeding 0.05. (70,551 SNPs removed)

**--hwe 1e-6:** filters out variants which have Hardy-Weinberg equilibrium exact test p-value below  $10^{-6}$ . (182,847 SNPs removed)

**--maf 0.05:** filters out variants with minor allele frequency below 0.05. (280,008 SNPs removed)

**--mind 0.05:** filters out samples with missingness exceeding 0.05. (21,797 samples removed)

**--nosex:** remove individuals with ambiguous sex. (81 individuals removed)

250,850 SNPs and 466,501 individuals are kept after the first step of QC. It is worth noting that some SNPs violate multiple QC criteria. Thus, the total number of SNPs we start with minus the sum of SNPs need to be removed doesn't necessarily equal to the number of SNPs we keep.

**Step 2:** Restrict genetic variants of interest to GWAS Catalog significant SNPs.

Among the 250,850 SNPs left in the first step of QC, 3267 of them also were reported as significant SNPs in GWAS Catalog. Thus, we will only consider those SNPs in the rest of the analysis.

**Step 3:** Preprocess the phenotype data using the following criteria.

**in\_white\_British\_ancestry\_subset:** restrict samples to participants who self-report themselves from a white British ancestry. (92,919 samples removed)

**used\_in\_pca\_calcuation:** restrict samples to individuals who have very similar ancestry based on a principle component analysis of the genotypes (Bycroft et al., 2018). (95,405 samples removed)

**het\_missing\_outliers:** only consider individuals who are not marked as outliers for heterozygosity or missing rates. (968 samples removed)

**excess\_relative:** exclude participants who have been identified to have ten or more third-degree relatives. (14,440 samples removed)

**recommend\_removal:** remove individuals that recommended for removal by the UK Biobank. (480 samples need to be removed)

After the third step of QC, 337,285 individuals are kept. Again, it is worth noting that some individuals violate multiple QC criteria. Thus, the total number of individuals we start with minus the sum of individuals need to be removed doesn't necessarily equal to the number of individuals we keep.

**Step 4:** Keep individuals who have both genotype and phenotype information. (322,607 participants satisfy this condition)

**Step 5:** Sort the genotype data such that each person's phenotype and genotype are matched and linked through the eid number in phenotype data, and FID and IID in genotype data.

Table 2.1. shows the number of autosomal SNPs in each chromosome before and after QCs.

### **2.2.5 Hierarchical Groups of UK Biobank ICD-10 Codes**

As UKB-PheCLC plans to take advantages of the clustering information of UK Biobank ICD codes, here we briefly discuss how we classify the UK Biobank phenotypes into different groups.

ICD coding system is an international standard for reporting diseases and health conditions. ICD-10 codes are in hierarchical order, with five levels in total. The top level



has 22 chapters. Level 1 has 263 disease blocks. Level 2 covers 2070 different diseases. Level 3 can denote ~12384 different diseases. Level 4 is the most bottom level. Not every disease has such detailed subcategory information. Table 2.3 is a demonstration of the hierarchical ICD-10 coding system.

It has been noted that when case-control ratio is too small, normal approximation of score test statistics will have inflated type I error (Dey et al., 2017). Therefore, we remove phenotypes with number of cases less than 50 to avoid the potential problems. After removing phenotypes with number of cases less than 50 in the UKB Biobank data, we have a total of 1101 phenotypes, distributed in 223 disease blocks.

In summary, after performing all data pre-process procedures, we have matched genotype, phenotype, and covariates information for a total of 322,607 individuals, where each individual has 3267 SNPs, 1101 case-control phenotypes, and 13 covariates.

## 2.3 Methods

Consider a sample of  $n$  unrelated individuals, indexed by  $i = 1, 2, \dots, n$ . Each individual has a total of  $K$  phenotypes and a genetic variant of interest. Suppose the  $K$  phenotypes can be divided into  $M$  phenotypic categories in which the effects of genetic variant might be different. Suppose that there are  $K_m$  phenotypes in the  $m^{\text{th}}$  category, where  $m = 1, 2, \dots, M = 223$  and  $K_1 + \dots + K_M = K = 1101$ . Let  $y_{imk}$  denote the  $k^{\text{th}}$  phenotype in the  $m^{\text{th}}$  category of the  $i^{\text{th}}$  individual and  $x_i$  denote the genotype at the variant of interest for the  $i^{\text{th}}$  individual. We incorporate covariates adjustment into our analyses according to

Price et al. (2006) and Sha et al. (2012). Suppose that there are  $p$  covariates,  $z_{i1}, \dots, z_{ip}$ , for the  $i^{th}$  individual, we regress both the genotypes and phenotypes on the covariates through the following linear models

$$y_{imk} = \alpha_{0mk} + \alpha_{1mk} z_{i1} + \dots + \alpha_{pmk} z_{ip} + \varepsilon_{imk} \quad \text{and} \quad x_i = \gamma_0 + \gamma_1 z_{i1} + \dots + \gamma_p z_{ip} + \tau_i$$

For easier demonstration, we assume all the phenotypes and genotypes have been covariates-adjusted. Then the score statistics to test for association between the  $k^{th}$  phenotype in the  $m^{th}$  category and the variant of interest under the generalized linear model  $g(E(y_{imk} | x_i)) = \beta_{mk}^0 + \beta_{mk}^1 x_i$  is given by

$$T_{mk} = U_{mk} / \sqrt{V_{mk}} ,$$

where  $U_{mk} = \sum_{i=1}^n (y_{imk} - \bar{y}_{mk})(x_i - \bar{x})$  and  $V_{mk} = \sum_{i=1}^n (y_{imk} - \bar{y}_{mk})^2 \sum_{i=1}^n (x_i - \bar{x})^2 / n$ . Under the null hypothesis that there is no association between the genetic variant and the  $k^{th}$  phenotype in the  $m^{th}$  category (i.e.  $\beta_{mk}^1 = 0$ ),  $T_{mk}$  asymptotically follows the standard normal distribution. Following the same univariate association testing procedure, we obtain  $K_m$  such score test statistics in the  $m^{th}$  category. Next, we define an overall test statistic for each category by combining the univariate test statistics in each category through the CLC (Clustering Linear Combination) method.

For the  $m^{th}$  category, we apply CLC method to combine  $T_{m_1}, T_{m_2}, \dots, T_{m_{K_m}}$  and obtain a CLC test statistic. In general, CLC can cluster  $T_{m_1}, T_{m_2}, \dots, T_{m_{K_m}}$  to  $S$  clusters, where  $s = 1, \dots, K_m$ . Let  $CLC_s$  and  $p_{ms}$  denote the CLC test statistic and p-value when CLC method clusters  $T_{m_1}, T_{m_2}, \dots, T_{m_{K_m}}$  into  $S$  clusters. Then, the CLC test statistic for the

$m^{\text{th}}$  category is given by  $T_m = \min_{1 \leq s \leq K_m} p_{ms}$  ( $m = 1, 2, \dots, M$ ). Let  $p_1, \dots, p_M$  be the p-values of  $T_1, \dots, T_M$  and  $p_{(1)}, p_{(2)}, \dots, p_{(M)}$  be the order statistics of  $p_1, \dots, p_M$  such that  $p_{(1)} \leq \dots \leq p_{(M)}$ . For a predefined integer  $L$ , we define the summation of negative  $\log p_{(m)}$  at cut-off point  $l$  as

$$w_l = -\sum_{m=1}^l \log p_{(m)}, l = 1, \dots, L.$$

Let  $P_l$  denote the p-value of  $w_l$ . Then, our proposed test statistic of PheCLC for testing the association between the genetic variant and all phenotypes across the phenome is given by

$$T = \min_{1 \leq l \leq L} P_l.$$

To calculate the p-value of  $T$ , we adopt the ‘‘one-layer’’ permutation procedure previously developed by our research group (Liang et al. 2016). Here, we briefly review the permutation steps. Suppose that we perform  $B$  times of permutations.

Step 1. In each permutation, randomly shuffle the genotypes and recalculate  $T_1, \dots, T_M$ .

Let  $T_m^{(b)}$  and  $w_l^{(b)}$  ( $b = 0, 1, \dots, B$ ) denote the value of  $T_m$  and  $w_l$  based on the  $b^{\text{th}}$  permuted data, where  $b = 0$  indicates calculating  $T_m$  and  $w_l$  using the original data.

Step 2. Transfer  $T_m^{(b)}$  to  $p_m^{(b)}$  by

$$p_m^{(b)} = \frac{\#\{d: T_m^{(d)} < T_m^{(b)} \text{ for } d=0, 1, \dots, B\} + 1}{B+1}.$$

Step 3. Let  $p_{(1)}^{(b)}, p_{(2)}^{(b)}, \dots, p_{(M)}^{(b)}$  be order statistics of  $p_1^{(b)}, \dots, p_M^{(b)}$  such that  $p_{(1)}^{(b)} \leq \dots \leq$

$p_{(M)}^{(b)}$ . Define  $w_l^{(b)} = -\sum_{m=1}^l \log p_{(m)}^{(b)}$ . We transfer  $w_l^{(b)}$  to  $P_l^{(b)}$  by

$$P_l^{(b)} = \frac{\#\{d: w_l^{(d)} > w_l^{(b)} \text{ for } d=0,1,\dots,B\}+1}{B+1}.$$

Step 4. Let  $T^{(b)} = \min_{1 \leq l \leq L} P_l^{(b)}$ . Then, the p-value of  $T$  is given by

$$\frac{\#\{b: T^{(b)} < T^{(0)} \text{ for } b=1,2,\dots,B\}+1}{B+1}.$$

## 2.4 Methods Comparison

We compare the power of the proposed method UKB-PheCLC with that of the following methods: Trait-based Association Test that uses Extended Simes procedure (TATES) (van der Sluis et al., 2013) and O'Brien's method (OB) (O'Brien, 1984), whose test statistic is either a linear combination of univariate test statistics or p-values.

Here, we review the TATES and OB methods. Assume there are  $K$  phenotypes in a phenotypic category. Denote  $p_k$  the p-value of the test statistic to test the association between the  $k^{\text{th}}$  phenotype and the genetic variant,  $p_{(k)}$  the  $k^{\text{th}}$  smallest p-value among all  $p_k$ 's, where  $k = 1, 2, \dots, K$ .

**TATES** (trait-based association test that uses Extended Simes procedure):

Calculate the univariate p-values  $p = (p_1, p_2, \dots, p_K)^T$  and order the univariate p-values

such that  $p_{(1)} \leq p_{(2)} \leq p_{(K)}$ . The TATES p-value is given by  $\min\left(\frac{m_e p_{(k)}}{m_{e(k)}}\right)$ , where  $m_e$  and

$m_{e(k)}$  are effective numbers of independent p-values among all  $K$  and the top  $k$  phenotypes, respectively.

**OB** (O'Brien's Method): Calculate the univariate test statistics  $T = (T_1, T_2, \dots, T_K)^T$  for the  $K$  phenotypes. Then test statistic  $T_{OB} = e^T \Sigma^{-1} T \sim N(0, e^T \Sigma^{-1} e)$ , where  $e = (1, \dots, 1)^T$ ,  $\Sigma$  is the covariance matrix of  $T$ .

In this study, TATES and OB were first adopted to test association between the genetic variant and phenotypes in each category. That is, for each phenotypic category, we obtain a TATES (or OB) p-value corresponding to it. Let  $p_m$  denote the p-value of the  $m^{\text{th}}$  category. Let  $p_{(1)}, \dots, p_{(M)}$  be the order statistics of  $p_1, \dots, p_M$  such that  $p_{(1)} \leq \dots \leq p_{(M)}$ . For any predefined integer  $L$  (in this study, we let  $L = 10$ ), we define the summation of negative  $\log p_{(m)}$  at cut point  $l$  as

$$w_l = -\sum_{m=1}^l \log p_{(m)}, \quad l = 1, \dots, L = 10.$$

Let  $P_l$  denote the p-value of  $w_l$ . Then, the test statistic of TATES and OB for testing the association between the genetic variant and all phenotypes across the phenotype is given by  $T = \min_{1 \leq l \leq L} P_l$ .

To calculate the p-value of  $T$ , we use a slightly different permutation procedure from the one we used for UKB-PheCLC. But the essence of the two permutation procedures are the same, that is, using the AFC method to obtain the overall p-value for testing the association between the genetic variant and all phenotypes across the phenotype. The reason we vary the details of permutation is that we try to avoid the computational

burden of permuting twice for UKB-PheCLC because the p-value of OB for each phenotypic category can be estimated using its asymptotic distributions and TATES has its own way to compute its p-values without the need of permutation. List below are the details of the permutation procedures of TATES and OB.

Step 1. In each permutation, we randomly shuffle the genotypes and recalculate  $p_{(1)}, \dots, p_{(M)}$  and  $w_1, \dots, w_L$ . Suppose that we perform  $B$  times of permutations. Let  $w_l^{(b)}$  ( $b = 0, 1, \dots, B$ ) denote the value of  $w_l$  based on the  $b^{th}$  permuted data, where  $b = 0$  represents the original data.

Step 2. Transfer  $w_l^{(b)}$  to  $P_l^{(b)}$  by

$$P_l^{(b)} = \frac{\#\{d: w_l^{(d)} > w_l^{(b)} \text{ for } d=0,1,\dots,B\}}{B}.$$

Step 3. Let  $T^{(b)} = \min_{1 \leq l \leq L} P_l^{(b)}$ . Then, the p-value of  $T$  is given by

$$\frac{\#\{b: T^{(b)} < T^{(0)} \text{ for } b=1,2,\dots,B\}}{B}.$$

## 2.5 Results

To evaluate UKB-PheCLC, we use a fast but efficient way to compare its performance with that of other two competing methods. For each method, we do the follows.

Step1. Permutation 100 times, we select SNPs with p-value less or equal 0.02.

Step2. Permutation 1,000 times, we select SNPs with p-value less or equal 0.002.

Step3. Permutation 10,000 times, we select SNPs with p-value less or equal 0.0002.

Step4. Permutation 100,000 times, we select SNPs with p-value less than  $\frac{0.05}{3267} \approx 1.53 \times 10^{-5}$  (Note: 3267 is the total number of SNPs we considered).

We summarize the results as follows.

1. When B=100, UKB-PheCLC identifies 671 significant SNPs. OB identifies 798. TATES identifies 1,125.
2. When B=1,000, UKB-PheCLC identifies 435 significant SNPs. OB identifies 505. TATES identifies 836.
3. When B=10,000, UKB-PheCLC identifies 381 significant SNPs out of the 435 significant SNPs from the previous step. OB identifies 391 significant SNPs out of 490 significant SNPs from the previous step. TATES identifies 526 significant SNPs out of 636 significant SNPs from the previous step.

Based on the results, we can draw the following conclusion that even though UKB-PheCLC method fail to identify as many SNPs as the other two competing methods when number of permutation B is small, however, as B increases from 10,000, UKB-PheCLC gradually shows its better performance over TATAS and OB. Due to computational consideration, we skip the case when B=100,000. But we firmly believe the pattern of UKB-PheCLC method's superior performance over other methods will continue because we have theoretically proved that CLC is the most powerful methods among all tests that have certain quadratic forms (Sha et al., 2018).

## 2.6 Discussion

With the advancement of next-generation sequencing (NGS), GWAS has become a very popular tool for detecting genetic elements underlying complex diseases. Up to now, the GWAS catalog contains more than 3,600 publications and roughly 90,000 unique SNP-trait associations. A huge advantage of GWAS is that it can detect new genes and pathways involved in disease pathogenesis. However, a big limitation of GWAS is that it only focuses on a pre-defined phenotypic domain. As a complementary approach to GWAS, PheWAS investigates the association between genomic markers and a diverse range of phenotypes. PheWAS has recently become possible due to the emerging use of electronic health records (EHR), which commonly use the International Classification of Disease (ICD) codes, a standardized coding system for defining disease status as well as for billing purpose for hospitals and insurance agencies. The UK Biobank is a population-based cohort study with deep genetic and phenotypic information collected on ~ 500K participants from multiple sites across the United Kingdom, aged between 40 and 69 years when recruited in 2006–2010 (Sudlow et al., 2015). In this manuscript, we have redesigned the PheCLC (Phenome-wide association study that uses Clustering Linear Combination) method which was previously developed by our research group. The refined method is then applied on the UKBiobank data to test the validity and understand the limitations of the proposed method. We have denoted our new method UKB-PheCLC. The UKB-PheCLC method is a typical example of EHR-based PheWAS. In the first step, it classifies all phenotypes across the whole phenome into numerous phenotypic categories according to the UK Biobank ICD-10 level 2 code. In the second step, the Clustering Linear



Combination (CLC) method is applied to each phenotypic category to derive a CLC-based p-value for testing the association between the genetic variant of interest and all phenotypes in that category. In the third step, the CLC-based p-values of all categories are combined by using a strategy resemble that of the Adaptive Fisher's Combination (AFC) method. The biggest advantage of UKB-PheCLC is that it considers the possibility that phenotypes are from different phenotypic categories, which is very common and readily available in EHR-based PheWAS. Moreover, UKB-PheCLC can handle both qualitative and quantitative phenotypes since we only need to classify the univariate test statistics. By the same token, UKB-PheCLC doesn't require raw phenotype information and it can work on summary test statistics from other studies. Furthermore, the permutation procedure that UKB-PheCLC adopted to generate the empirical null distribution of the final test-statistic only needs to be done once for different genetic variants. The real data analysis results confirm that UKB-PheCLC is more powerful than TATES and OB. Thus, UKB-PheCLC can serve as a new method for PheWAS and an alternative method for joint analysis of multiple phenotypes.

Even though UKB-PheCLC has been very successful in discovering new disease-SNP associations, it still faces interpretation challenges. When we detect a strong association between a genetic variant and the phenome, we cannot point out which disease or diseases the genetic variant has impact on.

As a future study, we can consider more phenotypic information rather than just the ICD-10 codes. For example, it has been noticed that verbal interview answers can provide

additional information about disease diagnosis and status and likely to boost the performance of robust SNP-trait association (Cassidy et al., 2016, Howard et al., 2017).

In addition, in this study, we only consider the direct genotyped data of UK Biobank, which contains roughly 800K SNPs. However, the full release of the UK Biobank imputed genotype data has roughly 90 million SNPs, which is very likely to contain more useful information. However, the performance of the proposed method on the UK Biobank imputed genotype data still needs further investigation.

## 2.7 Tables and Figures

**Table 2.1.** The number of autosomal SNPs in each chromosome for the 488,377 genotyped participants.

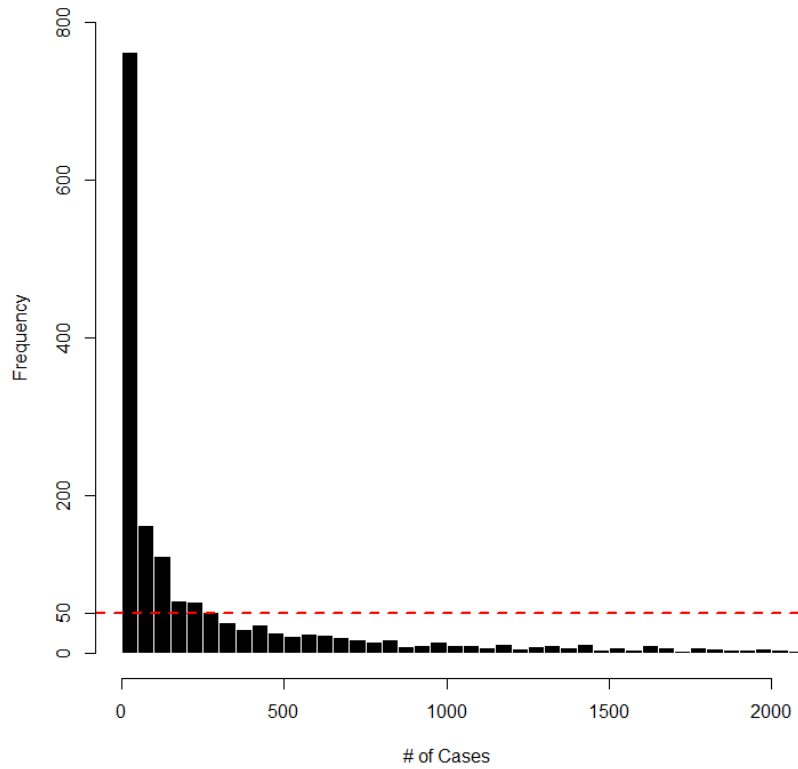
Chr	Count	
	Before QC	After QC
1	63,487	43,805
2	61,966	42,764
3	52,300	36,409
4	47,443	33,519
5	46,314	32,541
6	53,695	36,671
7	42,722	29,349
8	38,591	27,534
9	34,310	23,672
10	38,308	26,535
11	40,824	27,416
12	37,302	25,764
13	26,806	18,053
14	25,509	17,405
15	24,467	16,628
16	28,960	19,257
17	28,835	18,229
18	21,962	15,628
19	26,186	15,776
20	19,959	14,049
21	11,342	7,932
22	12,968	8,841
Total	784,256	537,777

**Note.** the applied QC filters include --geno 0.05, --hwe 1e-10, --keep founders, --maf 0.0001, --mind 0.1.

**Table 2.3.** Hierarchical structure of ICD-10 coding system

ICD-10 Hierarchical Levels	# of Categories	Examples
Top level	22	Chapter I, Chapter II, Chapter XXII
Level 1	263	Block A00-A09, Block A15-A19
Level 2	2070	A00, A01, A09
Level 3	~12,384	S06.2, S06.3, S06.4
Level 4	~4425	S06.20, S06.21, S06.30, S06.40

**Figure 2.1.** Histogram of # of cases of the 1869 phenotypes in the UKBiobank data.



## Reference List

- Aschard, H., Vilhjalmsson, B. J., Greliche, N., Morange, P. E., Tregouet, D. A., & Kraft, P. (2014). Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am J Hum Genet*, 94(5), 662-676. doi:10.1016/j.ajhg.2014.03.016
- Biobank, U. K. (2015). Genotyping and quality control of UK Biobank, a large-scale, extensively phenotyped prospective resource. Available at [biobank.ctsu.ox.ac.uk/crystal/docs/genotyping\\_qc.pdf](http://biobank.ctsu.ox.ac.uk/crystal/docs/genotyping_qc.pdf). Accessed April, 1, 2016.
- Brehm, J. M., Hagiwara, K., Tesfaigzi, Y., Bruse, S., Mariani, T. J., Bhattacharya, S., . . . Avila, L. (2011). Identification of FGF7 as a novel susceptibility locus for chronic obstructive pulmonary disease. *Thorax*, 66(12), 1085-1090.
- Bühlmann, P., Rütimann, P., van de Geer, S., & Zhang, C.-H. (2013). Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143(11), 1835-1858.
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., ... & Cortes, A. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726), 203.
- Cassidy, S., Chau, J. Y., Catt, M., Bauman, A., & Trenell, M. I. (2016). Cross-sectional study of diet, physical activity, television viewing and sleep duration in 233 110 adults from the UK Biobank; the behavioural phenotype of cardiovascular disease and type 2 diabetes. *BMJ open*, 6(3), e010038.
- Cho, M. H., Boutaoui, N., Klanderman, B. J., Sylvia, J. S., Ziniti, J. P., Hersh, C. P., . . . Sparrow, D. (2010). Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nature genetics*, 42(3), 200.
- Cho, M. H., McDonald, M.-L. N., Zhou, X., Mattheisen, M., Castaldi, P. J., Hersh, C. P., . . . Laird, N. M. (2014). Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *The lancet Respiratory medicine*, 2(3), 214-225.
- Chu, J.-h., Hersh, C. P., Castaldi, P. J., Cho, M. H., Raby, B. A., Laird, N., . . . Quackenbush, J. (2014a). Analyzing networks of phenotypes in complex diseases: methodology and applications in COPD. *BMC Systems Biology*, 8(1), 78.
- Chu, J.-h., Hersh, C. P., Castaldi, P. J., Cho, M. H., Raby, B. A., Laird, N., . . . Silverman, E. K. (2014b). Analyzing networks of phenotypes in complex diseases: methodology and applications in COPD. *BMC Systems Biology*, 8, 78-78. doi:10.1186/1752-0509-8-78

- Chung, K. F., & Pavord, I. D. (2008). Prevalence, pathogenesis, and causes of chronic cough. *The Lancet*, 371(9621), 1364-1374.
- Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychological Bulletin*, 115(3), 465.
- Cui, K., Ge, X., & Ma, H. (2014). Four SNPs in the CHR3/5 alpha-neuronal nicotinic acetylcholine receptor subunit locus are associated with COPD risk based on meta-analyses. *PLoS ONE*, 9(7), e102324.
- Dey, R., Schmidt, E. M., Abecasis, G. R., & Lee, S. (2017). A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *The American Journal of Human Genetics*, 101(1), 37-49.
- Ding, C., & He, X. (2002). Cluster merging and splitting in hierarchical clustering algorithms. Paper presented at the Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on.
- Du, Y., Xue, Y., & Xiao, W. (2016). Association of IREB2 gene rs2568494 polymorphism with risk of chronic obstructive pulmonary disease: a meta-analysis. *Medical science monitor: international medical journal of experimental and clinical research*, 22, 177.
- Ferreira, M. A., & Purcell, S. M. (2008). A multivariate test of association. *Bioinformatics*, 25(1), 132-133.
- Gratten, J., & Visscher, P. M. (2016). Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Medicine*, 8, 78. doi:10.1186/s13073-016-0332-x
- Guo, X., Li, Y., Ding, X., He, M., Wang, X., & Zhang, H. (2015). Association tests of multiple phenotypes: ATeMP. *PLoS ONE*, 10(10), e0140348.
- Han, M. K., Agustí, A., Calverley, P. M., Celli, B. R., Criner, G., Curtis, J. L., . . . MacNee, W. (2010). Chronic obstructive pulmonary disease phenotypes: the future of COPD. *American journal of respiratory and critical care medicine*, 182(5), 598-604.
- Hancock, D. B., Eijgelsheim, M., Wilk, J. B., Gharib, S. A., Loehr, L. R., Marcante, K. D., . . . Barr, R. G. (2010). Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nature genetics*, 42(1), 45.
- Howard, D. M., Adams, M. J., Shirali, M., Clarke, T. K., Marioni, R. E., Davies, G., ... & Wigmore, E. M. (2017). Genome-wide association study of depression phenotypes in UK Biobank (n= 322,580) identifies the enrichment of variants in excitatory synaptic pathways. *bioRxiv*, 168732.

- Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3), 241-254.
- Karypis, G., Han, E.-H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8), 68-75.
- Kim, J., Bai, Y., & Pan, W. (2015). An adaptive association test for multiple phenotypes with GWAS summary statistics. *Genetic epidemiology*, 39(8), 651-663.
- Klei, L., Luca, D., Devlin, B., & Roeder, K. (2008). Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 32(1), 9-19.
- Korte, A., Vilhjálmsson, B. J., Segura, V., Platt, A., Long, Q., & Nordborg, M. (2012). A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature genetics*, 44(9), 1066.
- Li, X., Howard, T. D., Moore, W. C., Ampleford, E. J., Li, H., Busse, W. W., . . . Erzurum, S. C. (2011). Importance of hedgehog interacting protein and other lung function genes in asthma. *Journal of Allergy and Clinical Immunology*, 127(6), 1457-1465.
- Liang, X., Sha, Q., Rho, Y., & Zhang, S. (2018). A hierarchical clustering method for dimension reduction in joint analysis of multiple phenotypes. *Genetic epidemiology*, 42(4), 344-353.
- Liang, X., Wang, Z., Sha, Q., & Zhang, S. (2016). An Adaptive Fisher's Combination Method for Joint Analysis of Multiple Phenotypes in Association Studies. *Scientific reports*, 6, 34323.
- Lutz, S. M., Cho, M. H., Young, K., Hersh, C. P., Castaldi, P. J., McDonald, M.-L., . . . Parker, M. (2015). A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC genetics*, 16(1), 138.
- Mannino, D. M., & Buist, A. S. (2007). Global burden of COPD: risk factors, prevalence, and future trends. *The Lancet*, 370(9589), 765-773.
- Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3), 370-384. doi:10.2307/2344614
- O'Brien, P. C. (1984a). Procedures for comparing samples with multiple endpoints. *Biometrics*, 1079-1087.
- O'Brien, P. C. (1984b). Procedures for Comparing Samples with Multiple Endpoints. *Biometrics*, 40(4), 1079-1087. doi:10.2307/2531158
- O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C. F., Elliott, P., Jarvelin, M.-R., & Coin, L. J. M. (2012). MultiPhen: Joint Model of Multiple Phenotypes Can Increase Discovery in GWAS. *PLoS ONE*, 7(5), e34861. doi:10.1371/journal.pone.0034861

- Pillai, S. G., Ge, D., Zhu, G., Kong, X., Shianna, K. V., Need, A. C., . . . Gulsvik, A. (2009). A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet*, 5(3), e1000421.
- Regan, E. A., Hokanson, J. E., Murphy, J. R., Make, B., Lynch, D. A., Beaty, T. H., . . . Crapo, J. D. (2011). Genetic epidemiology of COPD (COPDGene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 7(1), 32-43.
- Rokach, L., & Maimon, O. (2005). Clustering methods. In *Data mining and knowledge discovery handbook* (pp. 321-352): Springer.
- Sandford, A., Weir, T., & Pare, P. D. (1997). Genetic risk factors for chronic obstructive pulmonary disease. *European Respiratory Journal*, 10(6), 1380-1391.
- Schellenberg, D., Pare, P. D., Weir, T. D., Spinelli, J. J., Walker, B. A., & Sandford, A. J. (1998). Vitamin D binding protein variants and the risk of COPD. *American journal of respiratory and critical care medicine*, 157(3), 957-961.
- Sha, Q., Wang, Z., Li, X., & Zhang, S. (2018a). A Novel Association Test for Joint Analysis of Multiple Phenotypes Using Conditional CUR. Manuscript Submitted for Publication.
- Sha, Q., Wang, Z., Zhang, X., & Zhang, S. (2018b). A Clustering Linear Combination Approach to Jointly Analyze Multiple Phenotypes for GWAS. *Bioinformatics*, bty810, <https://doi.org/10.1093/bioinformatics/bty810>.
- Silverman, E. K., Chapman, H. A., Drazen, J. M., Weiss, S. T., Rosner, B., Campbell, E. J., . . . Mentzer, S. (1998). Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease: risk to relatives for airflow obstruction and chronic bronchitis. *American journal of respiratory and critical care medicine*, 157(6), 1770-1778.
- Silverman, E. K., & Weiss, S. T. (2004). Risk Factors for the Development of COPD. In *Long-Term Intervention in Chronic Obstructive Pulmonary Disease* (pp. 81-98): CRC Press.
- Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T., . . . Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. *Am J Hum Genet*, 89(5), 607-618. doi:10.1016/j.ajhg.2011.10.004
- Stearns, F. W. (2010). One Hundred Years of Pleiotropy: A Retrospective. *Genetics*, 186(3), 767-773. doi:10.1534/genetics.110.122549
- Tang, C. S., & Ferreira, M. A. (2012). A gene-based test of association using canonical correlation analysis. *Bioinformatics*, 28(6), 845-850. doi:10.1093/bioinformatics/bts051



- van der Sluis, S., Posthuma, D., & Dolan, C. V. (2013). TATES: Efficient Multivariate Genotype-Phenotype Analysis for Genome-Wide Association Studies. *PLoS Genet*, 9(1), e1003235. doi:10.1371/journal.pgen.1003235
- Wang, Z., Sha, Q., & Zhang, S. (2016). Joint analysis of multiple traits using "optimal" maximum heritability test. *PLoS ONE*, 11(3), e0150975.
- Wei, L., & Johnson, W. E. (1985). Combining dependent tests with incomplete repeated measurements. *Biometrika*, 72(2), 359-364.
- Wienke, A. (2010). *Frailty models in survival analysis*: CRC Press.
- Wilk, J. B., Chen, T.-h., Gottlieb, D. J., Walter, R. E., Nagle, M. W., Brandler, B. J., . . . Weiss, S. T. (2009). A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genet*, 5(3), e1000429.
- Wilk, J. B., Shrine, N. R., Loehr, L. R., Zhao, J. H., Manichaikul, A., Lopez, L. M., . . . Tang, W. (2012). Genome-wide association studies identify CHRNA5/3 and HTR4 in the development of airflow obstruction. *American journal of respiratory and critical care medicine*, 186(7), 622-632.
- Yang, J. J., Li, J., Williams, L. K., & Buu, A. (2016). An efficient genome-wide association test for multivariate phenotypes based on the Fisher combination function. *BMC bioinformatics*, 17(1), 19.
- Yang, Q., & Wang, Y. (2012a). Methods for Analyzing Multivariate Phenotypes in Genetic Association Studies. *Journal of Probability and Statistics*, 2012, 13. doi:10.1155/2012/652569
- Yang, Q., & Wang, Y. (2012b). Methods for analyzing multivariate phenotypes in genetic association studies. *Journal of Probability and Statistics*, 2012.
- Young, R., Whittington, C., Hopkins, R., Hay, B., Epton, M., Black, P., & Gamble, G. (2010). Chromosome 4q31 locus in COPD is also associated with lung cancer. *European Respiratory Journal*, 36(6), 1375-1382.
- Zeger, S. L., & Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 121-130.
- Zhang, J., Summah, H., Zhu, Y.-g., & Qu, J.-M. (2011). Nicotinic acetylcholine receptor variants associated with susceptibility to chronic obstructive pulmonary disease: a meta-analysis. *Respiratory research*, 12(1), 158.
- Zhang, Y., Xu, Z., Shen, X., Pan, W., & Initiative, A. s. D. N. (2014). Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. *NeuroImage*, 96, 309-325.

- Zhang, Z., Buckler, E. S., Casstevens, T. M., & Bradbury, P. J. (2009). Software engineering the mixed model for genome-wide association studies on large samples. *Briefings in Bioinformatics*, 10(6), 664-675. doi:10.1093/bib/bbp050
- Zhou, J. J., Cho, M. H., Lange, C., Lutz, S., Silverman, E. K., & Laird, N. M. (2015). Integrating multiple correlated phenotypes for genetic association analysis by maximizing heritability. *Human heredity*, 79(2), 93-104.
- Zhou, X., & Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods*, 11(4), 407.
- Zhu, A. Z., Zhou, Q., Cox, L. S., David, S. P., Ahluwalia, J. S., Benowitz, N. L., & Tyndale, R. F. (2014). Association of CHRNA5-A3-B4 SNP rs2036527 With Smoking Cessation Therapy Response in African-American Smokers. *Clinical Pharmacology & Therapeutics*, 96(2), 256-265.
- Zhu, H., Zhang, S., & Sha, Q. (2015a). Power comparisons of methods for joint association analysis of multiple phenotypes. *Human heredity*, 80(3), 144-152.
- Zhu, H., Zhang, S., & Sha, Q. (2018). A novel method to test associations between a weighted combination of phenotypes and genetic variants. *PLoS ONE*, 13(1), e0190788.
- Zhu, X., Feng, T., Tayo, B. O., Liang, J., Young, J. H., Franceschini, N., . . . Edwards, T. L. (2015b). Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *The American Journal of Human Genetics*, 96(1), 21-36.

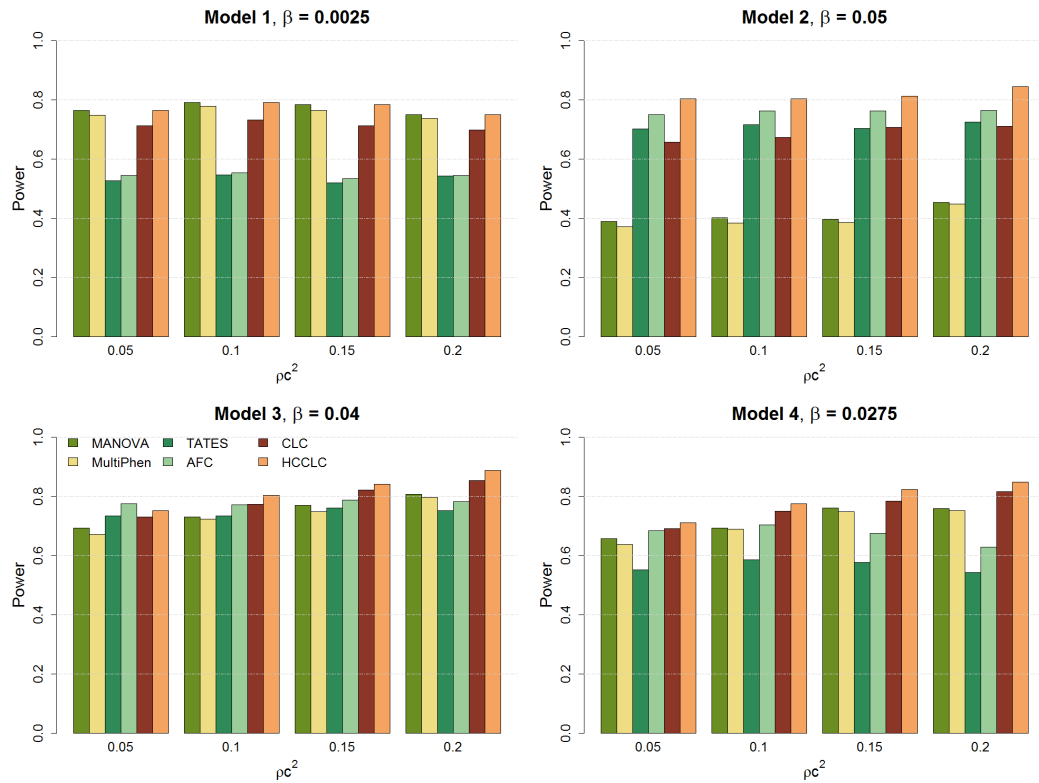
## Appendix A: Supplementary Tables

**Table A.1.1.** Description of COPD-related phenotypes

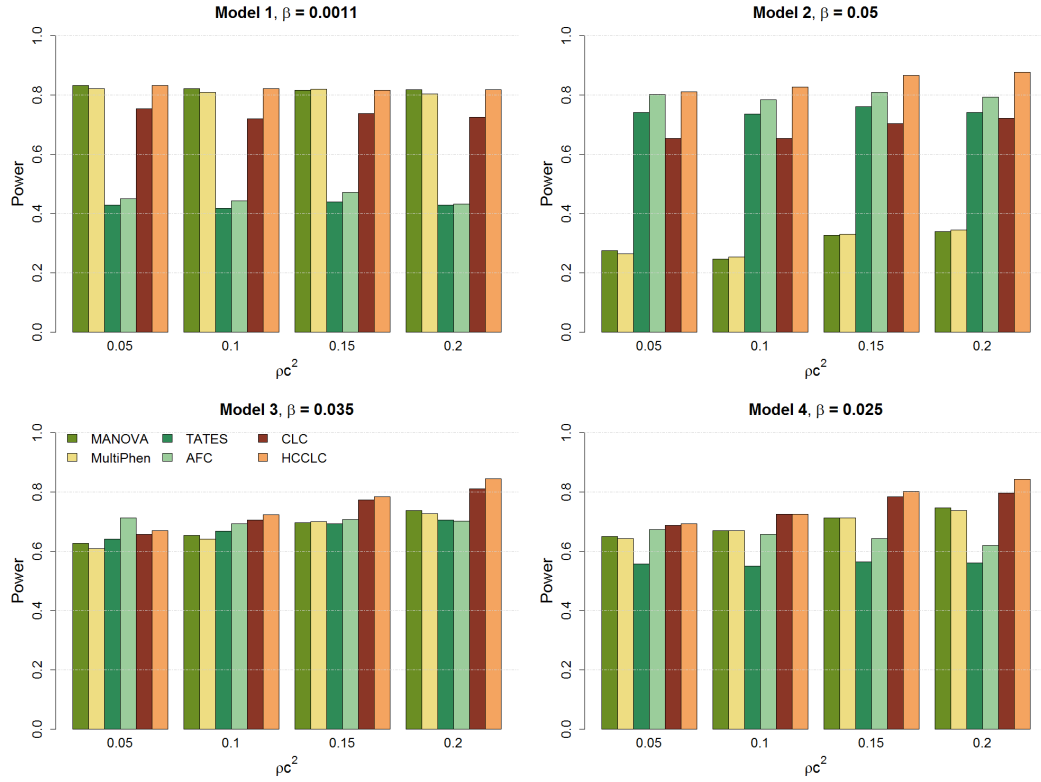
Phenotypes	Descriptions
Gas Trapping (GasTrap)	Air trapping at -856 Hounsfield units (HU) on expiratory chest CT scan
Exacerbation Frequency (ExacerFreq)	Number of COPD exacerbations during the year before study enrollment
Emphysema (Emph)	% Emphysema at -950 HU
Airway Wall Area (Pi10)	Square root of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

## Appendix B: Supplementary Figures

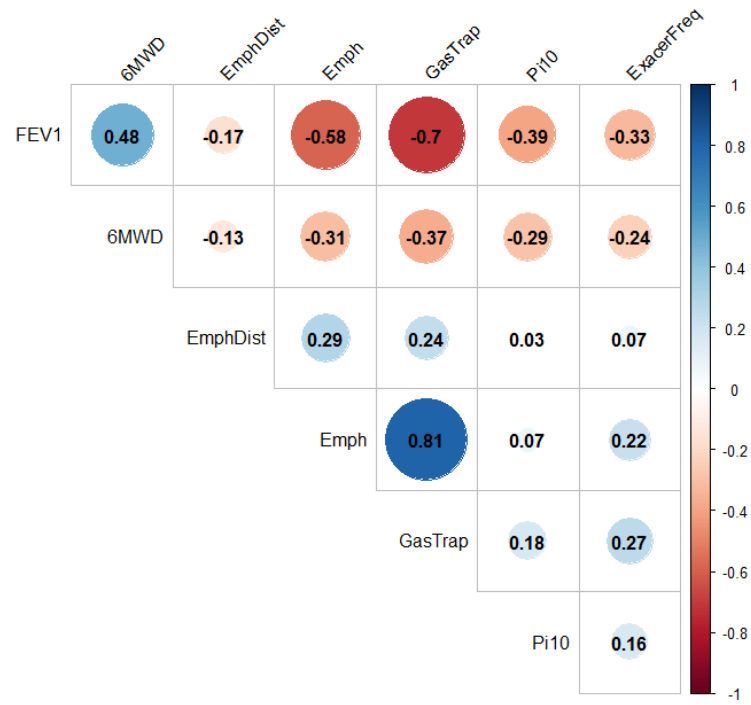
**Figure B.1.1.** The power comparisons of the six methods for 20 quantitative phenotypes assessed at a 5% significance level. Statistical power varies with the between-factor correlation  $\rho c^2$ , where MAF is 0.3, the sample size is 5,000, the number of replications is 1,000, and the within-factor correlation is 0.5 ( $c^2 = 0.5$ ).



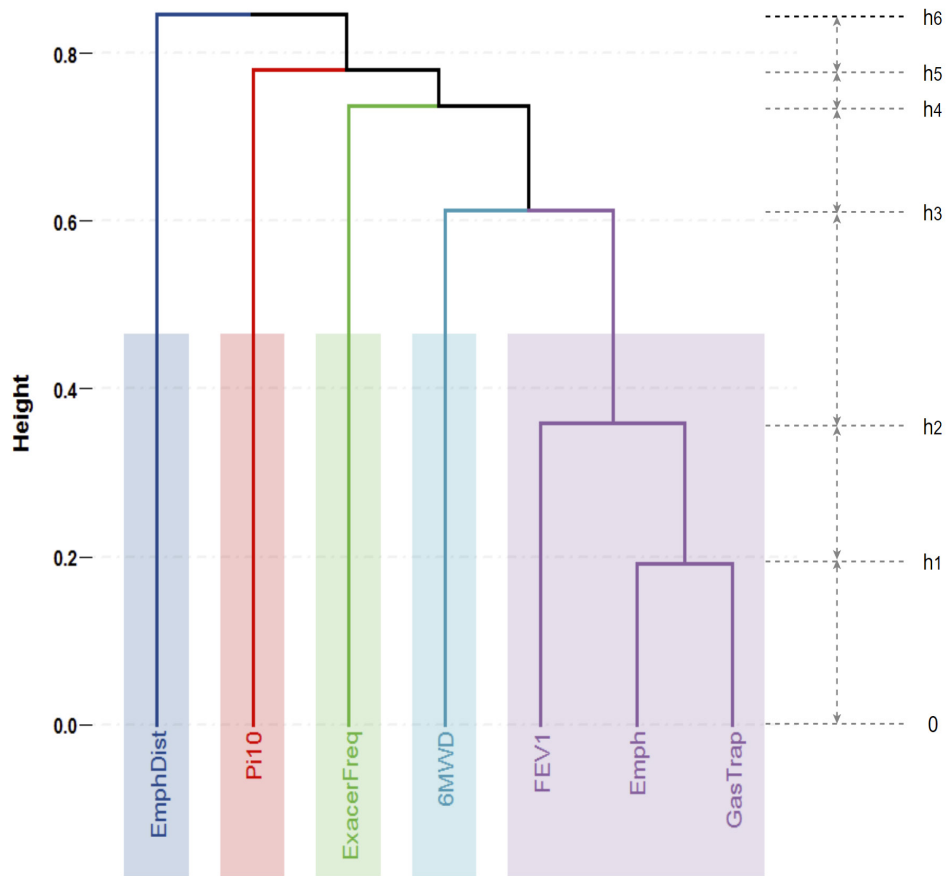
**Figure B.1.2.** The power comparisons of the six methods for 40 quantitative phenotypes assessed at a 5% significance level. Statistical power varies with the between-factor correlation  $\rho c^2$ , where MAF is 0.3, the sample size is 5,000, the number of replications is 1,000, and the within-factor correlation is 0.5 ( $c^2 = 0.5$ ).



**Figure B.1.3.** The correlation plot of the seven COPD-related phenotypes.



**Figure B.1.4.** The dendrogram based on the agglomerative hierarchical clustering of the seven COPD-related phenotypes.



**Figure B.1.5.** A Sample of UK Biobank ICD-10 billing code

---

C301	C30.1 Middle ear
C31	C31 Malignant neoplasm of accessory sinuses
C310	C31.0 Maxillary sinus
C311	C31.1 Ethmoidal sinus
C312	C31.2 Frontal sinus
C318	C31.8 Overlapping lesion of accessory sinuses
C319	C31.9 Accessory sinus, unspecified
C32	C32 Malignant neoplasm of larynx
C320	C32.0 Glottis
C321	C32.1 Supraglottis
C322	C32.2 Subglottis
C323	C32.3 Laryngeal cartilage
C328	C32.8 Overlapping lesion of larynx
C329	C32.9 Larynx, unspecified
C33	C33 Malignant neoplasm of trachea
C34	C34 Malignant neoplasm of bronchus and lung
C340	C34.0 Main bronchus
C341	C34.1 Upper lobe, bronchus or lung
C342	C34.2 Middle lobe, bronchus or lung
C343	C34.3 Lower lobe, bronchus or lung
C348	C34.8 Overlapping lesion of bronchus and lung
C349	C34.9 Bronchus or lung, unspecified
C37	C37 Malignant neoplasm of thymus
C38	C38 Malignant neoplasm of heart, mediastinum and pleura
C380	C38.0 Heart
C381	C38.1 Anterior mediastinum

---



**Figure B.1.6.** Example of UKB ICD-10 main diagnosis

ID	41202-0.0	41202-0.1	41202-0.2	41202-0.3	41202-0.4	...	41202-0.379
1	I841	H55	H251				
2	R002						
3	Z038	R500	R074	N390			
4							
5	Z098	K811	C509				
6	Z368	Z363	O0469	N871	D122	...	C61
7							
8	G473	R195	M159	D120			
9	R55	R074	K210				
10							
11	M751	C189	T842	D693			
12	N47	K802	J154				
13	R69	J348					
14	T812	K922	K861	K85	R194		

**Figure B.1.7.** Example of UKB ICD-10 secondary diagnosis

ID	41204-0.0	41204-0.1	41204-0.2	41204-0.3	41204-0.4	...	41204-0.379
1	Z961	Z836	H353				
2	Z824	Z035	R072	R074	F419		
3	Z886	Z034	R31	M6099	K219	...	C509
4	R55						
5	I10						
6	Z864	M199	N950	N816			
7	K573						
8							
9	E669	Z880					
10	N210	Z921	Z867				
11	E109	I10					
12	K297						
13	Y428	T388	N840				
14	Z922	Z871	Z720	G409	E780	...	E86

**Figure B.1.8.** Ethnic background of UK Biobank participants (UK Biobank, 2015).

Self-reported ethnicity	Representation (%)
White	94.06
British	88.07
Irish	2.63
Any other white background	3.36
Asian	2.28
Indian	1.18
Pakistani	0.37
Bangladeshi	0.05
Chinese	0.31
Any other Asian background	0.37
Black	1.61
African	0.68
Caribbean	0.90
Any other Black background	0.03
Mixed	0.59
White and Asian	0.17
White and Black African	0.08
White and Black Caribbean	0.12
Any other mixed background	0.22