PLOS ONE

RESEARCH ARTICLE

HCLC-FC: A novel statistical method for phenome-wide association studies

Xiaoyu Liang[1], Xuewei Cao[2], Qiuying Sha[2], Shuanglin Zhang[2]*

1 Department of Preventive Medicine, Division of Biostatistics, University of Tennessee Health Science Center, Memphis, Tennessee, United States of America, **2** Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan, United States of America

* shuzhang@mtu.edu

Abstract

The emergence of genetic data coupled to longitudinal electronic medical records (EMRs) offers the possibility of phenome-wide association studies (PheWAS). In PheWAS, the whole phenome can be divided into numerous phenotypic categories according to the genetic architecture across phenotypes. Currently, statistical analyses for PheWAS are mainly univariate analyses, which test the association between one genetic variant and one phenotype at a time. In this article, we derived a novel and powerful multivariate method for PheWAS. The proposed method involves three steps. In the first step, we apply the bottom-up hierarchical clustering method to partition a large number of phenotypes into disjoint clusters within each phenotypic category. In the second step, the clustering linear combination method is used to combine test statistics within each category based on the phenotypic clusters and obtain p-values from each phenotypic category. In the third step, we propose a new false discovery rate (FDR) control approach. We perform extensive simulation studies to compare the performance of our method with that of other existing methods. The results show that our proposed method controls FDR very well and outperforms other methods we compared with. We also apply the proposed approach to a set of EMR-based phenotypes across more than 300,000 samples from the UK Biobank. We find that the proposed approach not only can well-control FDR at a nominal level but also successfully identify 1,244 significant SNPs that are reported to be associated with some phenotypes in the GWAS catalog. Our open-access tools and instructions on how to implement HCLC-FC are available at https://github.com/XiaoyuLiang/HCLCFC.

OPEN ACCESS

Citation: Liang X, Cao X, Sha Q, Zhang S (2022) HCLC-FC: A novel statistical method for phenome-wide association studies. PLoS ONE 17(11): e0276646. https://doi.org/10.1371/journal.pone.0276646

Editor: Chunyu Liu, State University of New York Upstate Medical University, UNITED STATES

Received: March 30, 2022

Accepted: October 11, 2022

Published: November 9, 2022

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

Funding: The author(s) received no specific funding for this work.

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

Introduction

Genome-wide association studies (GWAS) have emerged as a common and powerful tool for investigating the genetic architecture of human disease over the last ten years [1, 2]. Over the last decade, numerous disease- and trait-associated common SNPs have been successfully identified by using statistical methods of GWAS.

To date, many software packages, such as PLINK, Gen/ProbABEL, MaCH, SNPTEST, and FaST-LMM, have been developed to support GWAS [3–9]. However, GWAS suffer from important shortcomings. First of all, GWAS usually focus on a pre-defined and limited phenotypic domain and ignore the potential power gained through the use of intermediate phenotypes that may more closely reflect a gene's mechanism, as well as the association between genetic variation and multiple phenotypes [10, 11]. Moreover, it is difficult to reach the threshold of statistical significance by GWAS due to the burden of multiple comparisons conducted; only those associations with a p-value less than 5×10^{-8} are considered statistically significant. GWAS have difficulty in explaining a significant portion of the predicted phenotypic heritability even though a significant number of SNPs are identified [12]. Lastly, the genotype-phenotype association is assessed for millions of SNPs one by one, false-positive results may easily arise due to large-scale multiple testing. Therefore, large sample size is needed to achieve the optimal statistical power and minimize spurious associations. Furthermore, the replication of the significant loci in independent populations is necessary according to the GWAS criteria [13].

Recently, large-scale DNA databanks linked to longitudinal electronic medical records (EMRs) offer the possibility of phenome-wide association studies (PheWAS) and have been proposed as an approach for rapidly generating large, diverse cohorts for the discovery and replication of genotype-phenotype associations [14–16]. In most EMR systems, the whole phenome can be classified into numerous phenotypic categories according to genotypic and phenotypic information such as phenotype similarity [15], genetic architecture [17], and disease network [18]. As a complementary approach to GWAS, PheWAS investigate the association between SNPs and a diverse range of phenotypes. By utilizing all available phenotypic information and all genetic variants in the estimation of associations between genotype and phenotype, a broader picture of the relationship between genetic variation and networks of phenotypes is possible [17]. In summary, GWAS use a phenotype-to-genotype strategy, beginning with a specific phenotype or disease; PheWAS reverse this paradigm by using a genotype-to-phenotype approach, starting with a genotype to test for associations over a wide spectrum of human phenotypes [12].

We are motivated primarily by PheWAS, which aim to assess associations between SNPs and a diverse range of phenotypes. Many of the issues that arise in this setting also occur elsewhere, for example, in clinical trials, the outcomes of cardiovascular risk may include hospitalization, stroke, heart failure, myocardial infarction, cardiac arrest, disability, and death [19]. Therefore, the statistical framework and results given here have a potential for wider application.

Several statistical methods for genetic association studies based on multiple phenotypes have been developed. The traditional Multivariate Analysis of Variance (MANOVA) [20] can take into account multiple continuous phenotypes to essentially test whether or not the independent genetic variant simultaneously explains a statistically significant amount of variance in multiple phenotypes. By performing ordinal regression analysis (proportional odds logistic regression), the joint model of Multiple Phenotypes (MultiPhen) [21] was developed using a reversed analysis by considering a genetic variant of interest as an ordinal response variable and the correlated phenotypes as predictors. A limitation of these multivariate approaches is that their performance depends on the specific configuration of phenotypic correlation structure. To address the limitation of some of the multivariate approaches, the Trait-based Association Test that uses Extended Simes procedure (TATES) [22] was developed to combine p-values obtained in standard univariate GWAS while correcting for the observed correlational structure between phenotypes. However, TATES essentially only depends on the phenotype that has the strongest association with the variant. Thus, MANOVA and MultiPhen are more powerful than TATES when genotypes impact on all phenotypes or on a large proportion of phenotypes because TATES may lose information in this scenario, while TATES is more powerful than MANOVA and MultiPhen when genotypes impact on one or very few phenotypes [23].

In 2019, Sha et al. [24] developed the Clustering Linear Combination (CLC) method that combines univariate test statistics for jointly analyzing multiple phenotypes in association analysis. CLC has been theoretically proved to be the most powerful test among all tests with certain quadratic forms if the phenotypes are clustered correctly. It is not only robust to different signs of means of individual statistics but also reduces the degrees of freedom of the test statistics. Therefore, the CLC method can be applied to PheWAS. However, due to the unknown number of clusters for a given data, the final test statistic of the CLC method is the minimum p-value among all p-values of the test statistics obtained from each possible number of clusters [25], and a simulation procedure is used to estimate the p-value of the final test statistic which would be time-consuming, especially in the PheWAS setting.

In this article, we derive a novel and powerful multivariate method, which we referred to as HCLC-FC (Hierarchical Clustering Linear Combination with False discovery rate Control) to test the association between a genetic variant with a large number of phenotypes. The HCLC-FC method is applicable to PheWAS. In PheWAS, the whole phenome can be classified into numerous phenotypic categories according to genotypic and phenotypic information, and each category contains a certain number of phenotypes. The proposed method (HCLC-FC) involves three steps. In the first step, we use the bottom-up Hierarchical Clustering Method (HCM) [26] to partition a large number of phenotypes into disjoint clusters within each category. In the second step, we apply the CLC method to combine test statistics within each phenotypic category based on the phenotypic clusters and obtain p-values from each phenotypic category. In the third step, we develop a false discovery rate (FDR) control approach based on a large-scale association testing procedure with theoretical guarantees for FDR control under flexible correlation structures [10]. Using extensive simulation studies, we evaluate the performance of the proposed method and compare the power of the proposed method with the powers of three commonly used methods in association studies using multiple phenotypes. These three methods include MANOVA [20], MultiPhen [21], and TATES [22]. Our simulation studies show that the proposed method outperforms the other three methods for different within-group and between-group phenotypic correlation structures we consider. Furthermore, the existing methods using our proposed FDR control procedure can control FDR efficiently. We also evaluate the performance of HCLC-FC through a set of 1,869 EMR-based phenotypes based on the International Classification of Diseases, 10th Revision (ICD-10 code, Data-Field 41202), across more than 300,000 samples from the UK Biobank, where these phenotypes can be classified into 260 ICD-10 level 1 blocks. The real data analysis results show that HCLC-FC can well control the type I error rate and can identify 1,244 SNPs that have previously been reported in the GWAS catalog.

Materials and methods

Statistical methods

Consider a sample with n unrelated individuals for a PheWAS, indexed by $i = 1,2,\ldots,n$. Each individual has the phenome with K phenotypes. The K phenotypes can be divided into M phenotypic categories, indexed by $m = 1, \ldots, M$. Suppose that there are K_m phenotypes in the m^{th} category, where $m = 1,2, \ldots, M$ and $K_1 + \cdots + K_M = K.y_{im} = (y_{im1}, \ldots, y_{imk}, \ldots, y_{imK_m})^T$ is a length of K_m phenotype vector in the m^{th} phenotypic category of the i^{th} individual, where y_{imk} is the k^{th} phenotype in the m^{th} category of the i^{th} individual. Denote $x_i \in \{0,1,2\}$ as the number of minor alleles that the i^{th} individual carries at a genetic variant of interest. We are interested in simultaneously testing the collection of M hypotheses H_{0m}: the m^{th} phenotypic category is not associated with the genetic variant of interest.

We assume that there are no covariates. If there are covariates, such as, gender, age, BMI, and top principal components to adjust for population stratification, we adjust both phenotype and genotype values for the covariates using the method applied by Price et al. [27] and Sha et al. [28]. That is, if there are p covariates, $z_{i1}, \ldots, z_{ip}$, for the i^{th} individual, we adjust both phenotype and genotype values for the covariates through linear models

$$y_{imk} = \alpha_{0mk} + \alpha_{1mk} z_{i1} + \cdots + \alpha_{pmk} z_{ip} + \varepsilon_{imk},$$

$$x_i = \gamma_0 + \gamma_1 z_{i1} + \cdots + \gamma_p z_{ip} + \tau_i.$$

In this article, we derived a novel and powerful multivariate method for PheWAS, which is referred to as HCLC-FC. The proposed method (HCLC-FC) involves three steps. In the first step, we use the bottom-up HCM [26] to partition K_m phenotypes into L_m disjoint clusters within each category, where $m = 1, \ldots, M$. In the second step, we apply the CLC [24] to combine test statistics within each category. The CLC test statistic with L_m clusters follows a chi-square distribution with L_m degrees of freedom. We then obtain the p-value of the CLC test statistic for each phenotypic category. In the third step, we propose an FDR control approach based on the method proposed by Cai et al. [10]. FDR is widely used to claim significance for high-dimensional correlated data. However, most of the existing methods of FDR cannot accurately estimate FDR due to different directions of genetic effects on different phenotypes. Recently, Cai et al. (2019) developed a method to evaluate FDR that works well for PheWAS if only a single phenotype is considered at a time. However, Cai's method is based on test statistics which are difficult to extend to test statistics for multiple phenotypes. Instead of using test statistics, we propose a new approach to evaluate FDR which is based on p-values and does not depend on test statistics. In the following sections, we give a detailed approach for each step.

Step 1: HCM to partition phenotypes in each phenotype category. For the m^{th} phenotypic category, we partition K_m phenotypes into L_m disjoint clusters. Denote $\boldsymbol{D_m} = \mathbf{1} - \boldsymbol{\Sigma_m}$ with entries $d^m_{ll^*}$ as the dissimilarity matrix, where $\boldsymbol{\Sigma_m}$ is $K_m \times K_m$ similarity matrix of $\boldsymbol{Y_m}$ for the m^{th} phenotypic category and $d^m_{ll^*}$ is the dissimilarity value between l^{th} and l^{*th} phenotypes. The HCM is based on the agglomerative clustering algorithm. In agglomerative clustering, all the phenotypes are a cluster of their own, and we merged pairs of clusters until they form a single cluster. In each iteration, we merge two clusters that have the smallest value of the average dissimilarity $d^m_{ll^*}$ between all phenotypes in two clusters and define the smallest average dissimilarity h_b as the height of the b^{th} iteration. The established principle in Bühlmann et al. [29] is used to determine the number of clusters for each phenotypic category. That is, the number of clusters L_m is identified at the $\widehat{b}^{th}$ iteration, where $\widehat{b} = \arg\max_{b \geq 1} (h_{b+1} - h_b)$.

Step 2: CLC to test the association between phenotypes in each category and a genetic variant. For each phenotypic category, we apply the CLC method [24] to combine test statistics among the L_m clusters. We use T_{mk} to denote the score test statistic to test the null hypothesis H_{0mk}: $\beta_{1mk} = 0$ (the k^{th} phenotype in the m^{th} phenotypic category is not associated with the genetic variant) under the generalized linear model $y_{imk} = \beta_{0mk} + \beta_{1mk} x_i + \varepsilon_{imk}$, where $k = 1, \ldots, K_m$. So T_{mk} is given by $T_{mk} = U_{mk}/\sqrt{V_{mk}}$, where $U_{mk} = \Sigma_{i=1}^{n} y_{imk}(x_i - \bar{x})$, $V_{mk} = \frac{1}{n}\Sigma_{i=1}^{n}(y_{imk} - \bar{y}_{mk})^2 \Sigma_{i=1}^{n}(x_i - \bar{x})^2$, $\bar{x} = \frac{1}{n}\Sigma_{i=1}^{n} x_i$, and $\bar{y}_{mk} = \frac{1}{n}\Sigma_{i=1}^{n} y_{imk}$. If we let $T_m = (T_{m1}, \ldots, T_{mK_m})^T$ be the test statistic vector that contains score test statistics for each phenotype in the m^{th} phenotypic category and let $\boldsymbol{B_m}$ be a $K_m \times L_m$ matrix with the indicator entry $b_{kl} = 1$ if the k^{th} phenotype belongs to the l^{th} cluster and $b_{kl} = 0$ otherwise. Then the CLC test statistic for the L_m clusters in the m^{th} phenotypic category is given by $T_{CLC}^{L_m} = (\boldsymbol{W_m} T_m)^T (\boldsymbol{W_m}\boldsymbol{\Sigma_m}\boldsymbol{W_m^T})^{-1}(\boldsymbol{W_m} T_m)$, where $\boldsymbol{W_m} = \boldsymbol{B_m^T}\boldsymbol{\Sigma_m^{-1}}$.$T_{CLC}^{L_m}$ follows a chi-square distribution with L_m degrees of freedom. We denote p_m as the p-value of $T_{CLC}^{L_m}$.

Step 3: Threshold for FDR-controlling. The method proposed by Cai et al. [10] is based on test statistics which are hard to extend to other test statistics. Therefore, in this step, we develop a new approach to evaluate FDR which is based on p-values. In the second step, the p-value for the test statistic in the m^{th} category for $m = 1, \ldots, M$ can be obtained. In this step, we propose a new multiple testing FDR controlling procedure by thresholding the p-values $\{p_m: m = 1, \ldots, M\}$. Under the null hypothesis, each p_m follows a uniform distribution $U(0,1)$. Let t, $0 \leq t \leq 1$, be a rejection threshold so that H_{0m} is rejected if and only if $p_m \leq t$. For any given threshold t, $0 \leq t \leq 1$, the false discovery proportion (FDP) based on a random sample is given by

$$FDP(t) = \frac{\sum_{m \in H_0} I(p_m \leq t)}{\max\left\{\sum_{m=1}^{M} I(p_m \leq t), 1\right\}}.$$

To maximize the power of the test or equivalently the rejection rate among $\mathcal{H}_1$ while maintaining an FDP level of α, the optimal threshold t is $\hat{t}_0 = \sup\{t : \mathrm{FDP}(t) \leq \alpha\}$. The key to empirically controlling the FDP is to find a good estimate of the numerator $\sum_{m \in H_0} I(p_m \leq t)$. Using the idea in Cai et al. [10], we estimate the numerator by $\sum_{m \in H_0} I(p_m \leq t) \approx m_0 G(t)$, where m_0 is the number of categories under the null hypothesis and we can use M to estimate m_0 due to the sparsity in the number of alternative hypotheses in many real data applications, and $G(t) = P(U(0,1) \leq t) = t$.

For a given nominal FDR level $\alpha \in (0,1)$, we reject H_{0i} whenever $p_m \leq \hat{t}$, where

$$\hat{t} = \sup\{t : FDP(t) \leq \alpha\} = \sup\left\{ t : \frac{\sum_{m \in H_0} I(p_m \leq t)}{\max\left\{\sum_{m=1}^{M} I(p_m \leq t), 1\right\}} \leq \alpha \right\}$$

$$= \sup\left\{ t : \frac{m_0 t}{\max\left\{\sum_{m=1}^{M} I(p_m \leq t), 1\right\}} \leq \alpha \right\} = \sup\left\{ 0 \leq t \leq 1 : t \leq \frac{\alpha \max\left\{\sum_{m=1}^{M} I(p_m \leq t), 1\right\}}{m_0} \right\}$$

Comparison of methods

We compare the performance of the proposed method HCLC-FC with those of MultiPhen [21], MANOVA [20], and TATES [22]. To evaluate the FDR-controlling performance, MANOVA, MultiPhen, and TATES are first applied to each category. Then, we apply the third step of HCLC-FC to the three methods to control FDR, which are referred to as MANOVA-FC, MultiPhen-FC, and TATES-FC. That is, we not only compare the performance of different methods for joint analysis of multiple phenotypes but also compare the performance of different methods with the newly developed FDR-controlling process.

In the following sections, we will estimate the FDR and power of each method. FDR is estimated by FDP and the estimated FDR is $\widehat{FDR} = FDP = \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{m=1}^{M} I(p_m \le \hat{t}) - \sum_{m\in H_a} I(p_m \le \hat{t})}{\max\{\sum_{m=1}^{M} I(p_m \le \hat{t}), 1\}}$, where B is the number of replications, H_a is the alternative hypothesis, p_m is the p-value of the test statistic for the m^{th} phenotypic category, $m = 1, \ldots, M$, and $\hat{t}$ is the threshold estimated by HCLC-FC in step 3. The power of each method is the probability of correctly rejecting H_0; it is estimated by $\widehat{Power} = \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{m\in H_a} I(p_m \le \hat{t})}{\#\{m : m \in H_a\}}$.

Simulation study

To evaluate the FDRs and powers of the proposed method, we generate genotypes according to the minor allele frequency (MAF) of a genetic variant and assume Hardy Weinberg equilibrium. Then, we generate K phenotypes by the following models similar to the models used by Sha et al. [2019] and Liang et al. [2018] [24, 26]. We use the same notations in the method section. Suppose there are M categories and $K^* = \frac{K}{M}$ phenotypes in each category, that is, $K_m = K^*$. For the i^{th} individual, let $y_{im} = (y_{im1}, \ldots, y_{imk^*})^T$ denote a length of K^* phenotype vector in the m^{th} phenotypic category. We assume

$$y_{im} = x_i\lambda_m + cf_{im}1_{K^*} + \sqrt{1-c^2}E_{im} \text{ for } i = 1, \ldots, n, m = 1, \ldots, M \quad (1)$$

where x_i is the genotype score at the variant of interest; $\lambda_m = (\lambda_{m1}, \ldots, \lambda_{mK^*})^T$ is the vector of effect sizes of the genetic variant on phenotypes in the m^{th} category; $f_i = (f_{i1}, \ldots, f_{iM})^T \sim MVN_M(0, \boldsymbol{\Sigma}_f)$, $\boldsymbol{\Sigma}_f = (1-\rho_f)\boldsymbol{I} + \rho_f\boldsymbol{A}$, ρ_f is a constant to define the phenotypic correlation between phenotypic categories, $\boldsymbol{A}$ is an $M \times M$ matrix with elements of 1, and $\boldsymbol{I}$ is an $M \times M$ identity matrix; c is a constant; $E_{i1}, \ldots, E_{iM}$ are independent and $E_{im} \sim MVN_{K^*}(0, \Sigma_{\boldsymbol{e}})$ with $\Sigma_{\boldsymbol{e}} = (\sigma_{hh^*})$, where $\sigma_{hh*} = \rho_e^{|h-h^*|}$ and ρ_e is constant to define the phenotypic correlation within each phenotypic category.

Based on Eq (1), we consider the following six models. In these six models, the correlation between the h^{th} and h^{*th} phenotypes within each category is $c^2 + (1-c^2)\rho_e^{|h-h^*|}$, and between categories is $c^2\rho_f$. We set $M = 100$ for Model 1–3 and $M = 50$ for Model 4–6.

Model 1: There are $M = 100$ categories and genotypes impact on only one category. Let $\lambda_1 = \ldots = \lambda_{M-1} = 0$ and $\lambda_M = \beta(1, \ldots, K^*)^T$, β is a constant that is used to define the effect size.

Model 2: There are $M = 100$ categories and genotypes impact on two categories. Let

$$\lambda_1 = \cdots = \lambda_{M-2} = 0, \lambda_{M-1} = \tfrac{2\beta}{K*+1}(1, \ldots, K^*)^T, \text{ and } \lambda_M = 2\beta\left(\underbrace{1, \ldots, 1}_{K*/2}, 0, \ldots, 0\right)^T.$$

Model 3: There are $M = 100$ categories and genotypes impact on three categories. Let $\lambda_1 = \ldots = \lambda_{M-3} = 0, \lambda_{M-2} = \frac{\beta}{K*/2+1}(1, 2, 3, \ldots, K^*/2, K^*/2, \ldots, 3, 2, 1)^T, \lambda_{M-1} = \frac{2\beta}{K*+1}(1, \ldots, K^*)^T$ and $\lambda_M = 2\beta\left(\underbrace{1, \ldots, 1}_{K*/2}, 0, \ldots, 0\right)^T$.

Model 4: Same as Model 1, but there are $M = 50$ categories.

Model 5: Same as Model 2, but there are $M = 50$ categories.

Model 6: Same as Model 3, but there are $M = 50$ categories.

Results

Simulation results

In our simulation studies, we estimate the p-values of all test statistics using their asymptotic distributions. We first set $\rho_f = 0.2$, $\rho_e = 0.3$, $c^2 = 0.5$, and K = 1,000, 2,000 for comparing the performance of different methods for joint analysis of multiple phenotypes, in other words, we consider the proposed FDR-controlling method and compare the performance of HCLC-FC, MANOVA-FC, MultiPhen-FC, and TATES-FC. For FDR evaluation, we consider different numbers of phenotypes, different sample sizes, different values of effect size, and different models.

The estimated FDRs of the four methods are summarized in **Tables 1** and **2**. From these tables, we can see that all methods using our FDR control procedure control their respective targeted error rates very well, which indicates applying our new FDR-controlling procedure to the entire collection of hypotheses can control the rate of FD of associated genetic variants as well as the expected value of the average proportion of FD of phenotypic categories influenced by such variants.

To compare the power of HCLC-FC with that of MANOVA-FC, MultiPhen-FC, and TATES-FC, we consider different numbers of phenotypes, different sample sizes, different models, and different genetic effect sizes. The power of the four tests at an FDR level of 5% for 1,000 phenotypes and 2,000 phenotypes are shown in **Figs 1** and **2**, respectively. According to the power comparison results, we summarize the following conclusions. (1) HCLC-FC

Table 1. The estimated FDR of the four tests under the six models for 1,000 phenotypes ($K = 1{,}000$). MAF is 0.3. The sample size (n) is 2,000. $\rho_f = 0.2$, $\rho_e = 0.3$, and $c^2 = 0.5$. β is the effect size. FDR is evaluated using 200 replicated samples at a nominal FDR level of 5%. All estimated FDR are within the 95% confidence interval (0.0198, 0.0802).

Model	β	Method			
		HCLC-FC	MANOVA-FC	MultiPhen-FC	TATES-FC
1	0.012	0.038	0.039	0.048	0.041
	0.014	0.045	0.029	0.033	0.037
	0.016	0.039	0.049	0.047	0.048
2	0.050	0.034	0.048	0.047	0.044
	0.060	0.041	0.042	0.041	0.037
	0.070	0.049	0.053	0.045	0.073
3	0.050	0.049	0.037	0.048	0.063
	0.090	0.047	0.043	0.046	0.048
	0.130	0.048	0.057	0.057	0.063
4	0.005	0.043	0.063	0.063	0.035
	0.006	0.047	0.061	0.049	0.045
	0.007	0.041	0.050	0.049	0.065
5	0.050	0.048	0.052	0.056	0.030
	0.060	0.048	0.047	0.044	0.034
	0.070	0.042	0.038	0.048	0.050
6	0.050	0.035	0.064	0.065	0.040
	0.090	0.055	0.039	0.049	0.034
	0.130	0.047	0.044	0.043	0.046

https://doi.org/10.1371/journal.pone.0276646.t001

Table 2. The estimated FDR of the four tests under the six models for 2,000 phenotypes ($K = 2{,}000$). MAF is 0.3. The sample size (n) is 4,000. $\rho_f = 0.2$, $\rho_e = 0.3$, and $c^2 = 0.5$. β is the effect size. FDR is evaluated using 200 replicated samples at a nominal FDR level of 5%. All estimated FDR are within the 95% confidence interval (0.0198, 0.0802).

Model	β	Method			
		HCLC-FC	MANOVA-FC	MultiPhen-FC	TATES-FC
1	0.004	0.038	0.055	0.038	0.028
	0.005	0.053	0.054	0.046	0.063
	0.005	0.026	0.047	0.043	0.041
2	0.030	0.043	0.046	0.049	0.050
	0.040	0.034	0.038	0.048	0.046
	0.050	0.054	0.051	0.065	0.042
3	0.050	0.051	0.052	0.052	0.041
	0.070	0.064	0.065	0.062	0.041
	0.090	0.051	0.052	0.054	0.056
4	0.002	0.042	0.057	0.054	0.059
	0.002	0.059	0.063	0.078	0.073
	0.002	0.054	0.077	0.060	0.049
5	0.030	0.024	0.045	0.051	0.043
	0.040	0.039	0.050	0.050	0.052
	0.050	0.034	0.039	0.044	0.042
6	0.050	0.050	0.052	0.042	0.047
	0.070	0.046	0.052	0.046	0.056
	0.090	0.055	0.048	0.053	0.048

https://doi.org/10.1371/journal.pone.0276646.t002

outperforms MANOVA-FC, MultiPhen-FC, and TATES-FC consistently for all models we consider; HCLC-FC is the most powerful test no matter whether the effect sizes show no groups (Model 1 and 4) or show some groups (Model 2, 3, 5, and 6) within the categories impacted by the SNP; (2) MANOVA-FC and MultiPhen-FC have similar power and are more powerful than TATES-FC for all models we consider.

In addition to considering power as a function of genetic effect size, we further evaluate power with varying the correlation between phenotypic categories ρ_f (**S1 Fig in** S1 File), the correlation within each phenotypic category ρ_e (**S2 Fig in** S1 File), the constant c^2 in the model (**S3 Fig in** S1 File), and the MAF (**S4 Fig in** S1 File). These figures show that 1) The powers of HCLC-FC, MANOVA-FC, and MultiPhen-FC slightly decrease with the increasing correlation between phenotypic categories. HCLC-FC outperforms MANOVA-FC, MultiPhen-FC, and TATES-FC consistently for different correlations between phenotypic categories no matter the effect sizes show no groups (Model 1 and 4) or show some groups (Model 2, 3, 5, and 6) (**S1 Fig in** S1 File). 2) The powers of HCLC-FC, MANOVA-FC, and MultiPhen-FC considerably decrease as the within-category correlation increases, while the power of TATE-FC does not change too much as the correlation increases (**S2 Fig in** S1 File). HCLC-FC is the most powerful test for correlation less than or equal to 0.4. For strong correlation within category structures (within-category correlation ≥ 0.6), TATES-FC outperforms other methods when the effect sizes show no groups (Models 1 and 4) or show some groups and genotype impact on multiple categories (Models 3 and 6). The power gap is much larger when the phenotypes are highly correlated and show no groups (Models 1 and 4). The reason is that the p-value of TATES equals the smallest weighted p-value, so TATES is expected to outperform multivariate approaches as the phenotype correlations increase; The power of MANOVA-FC and MultiPhen-FC are nearly identical. 3) For power as a function of c^2 (**S3 Fig in** S1 File), HCLC-FC is either the most powerful test (Model 1, 2, 4, 5, and 6) or comparable with the most powerful

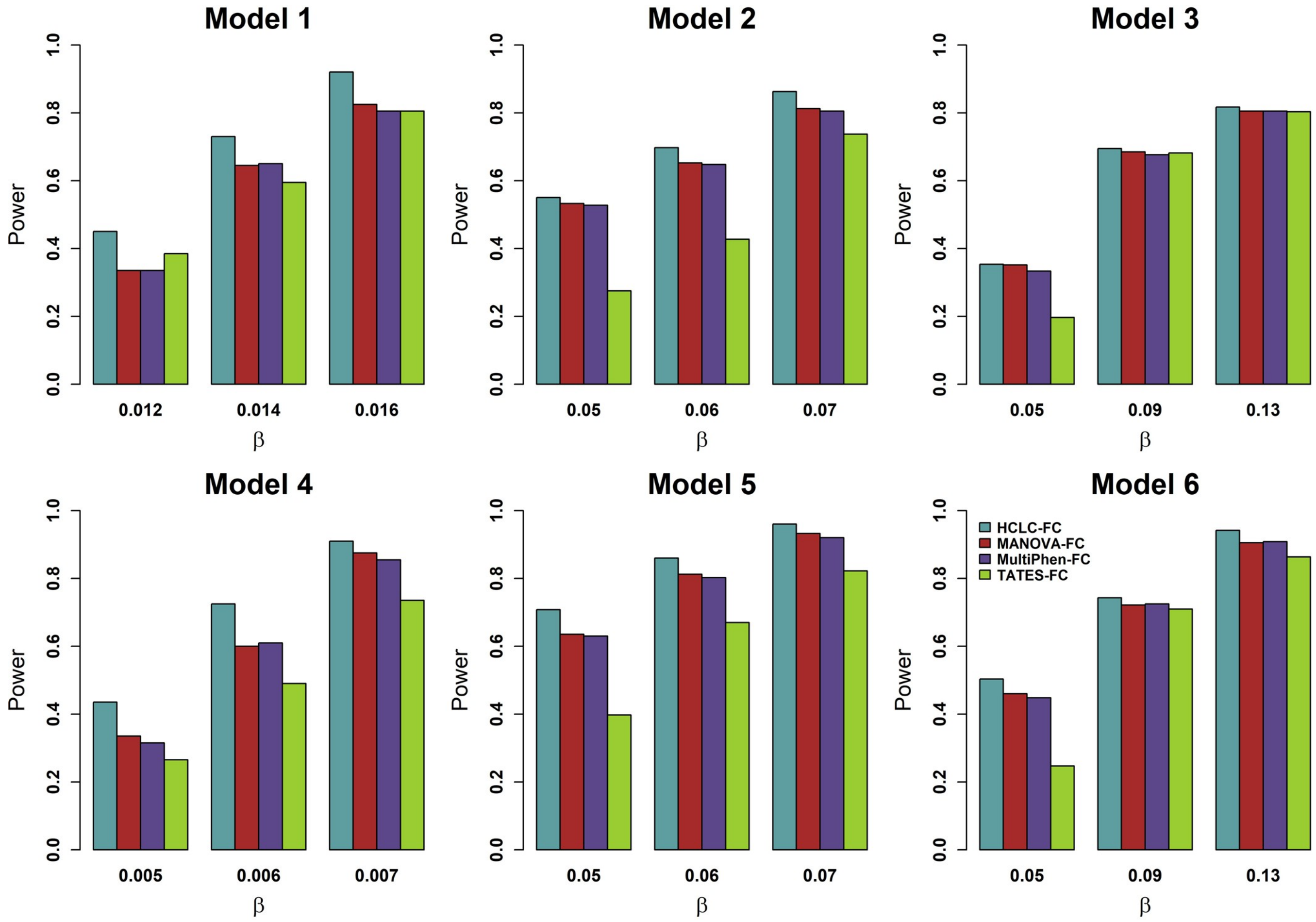

Fig 1. Power comparisons of the four tests for the power as a function of effect size (β) under the six models for 1,000 phenotypes (K = 1,000). MAF is 0.3. The sample size (n) is 2,000. $\rho_f = 0.2$, $\rho_e = 0.3$, and $c^2 = 0.5$. The power of all of the four tests is evaluated using 200 replicated samples at a nominal FDR level of 5%.

https://doi.org/10.1371/journal.pone.0276646.g001

test (In Model 3, $c^2 = 0.3$). The powers of HCLC-FC, MANOVA-FC, and MultiPhen-FC increase with the increase of the constant c^2, but the power of TATES-FC decreases as the increase of the constant c^2. 4) For all the methods we considered, lower MAF decreases the power, but our method has the highest power no matter the effect sizes show no groups or show some groups (**S4 Fig in S1 File**).

One of the important steps of our method, HCLC-FC, is the third step, the FDR controlling procedure. To date, many methods have been developed to address multiple test correction. Here, we compare the performance of using our proposed FDR controlling procedure in step 3 of HCLC-FC with some existing FDR controlling approaches, namely the spectral decomposition-based redundant filtering methods mentioned in Asif et al., 2021 [30]. Nyholt's spectral decomposition method [31] and Li and Ji's method [32] are used to estimate the effective number of independent phenotypes, then, Bonferroni and Sidak [33] corrections are applied to address multiple test corrections. We refer to the combinations of those methods as Nyholt-Sidak (NySi), Nyholt-Bonferroni (NyBo), Ji-Sidak (JiSi), and Ji-Bonferroni (JiBo). **S5 Fig in S1 File** shows FDR comparisons of the methods using our proposed FDR controlling procedure with those of using NySi, NyBo, JiSi, and JiBo for multiple test correction. We can see from **S5 Fig in S1 File**, the methods using our proposed FDR control procedure can control the FDRs

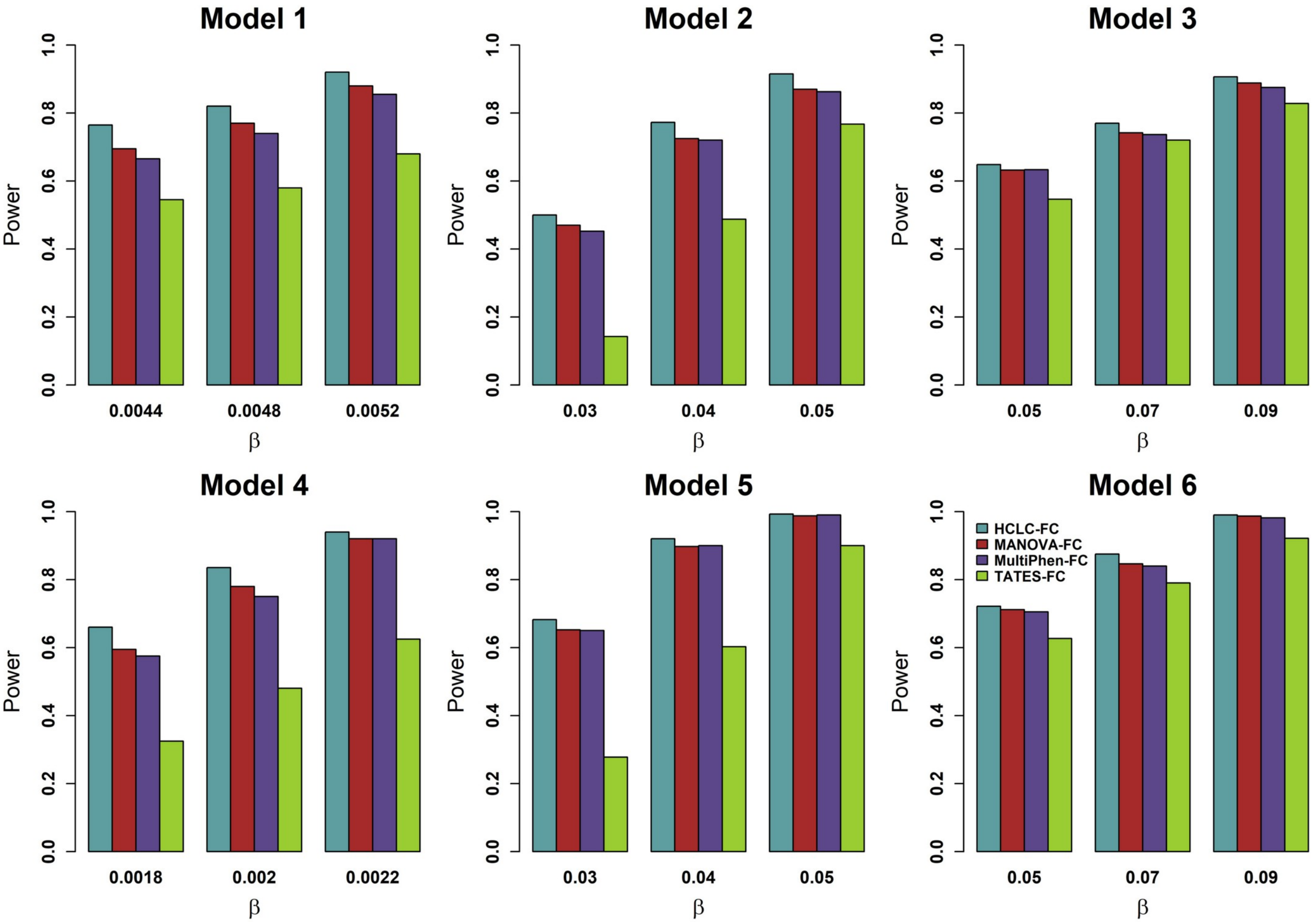

Fig 2. Power comparisons of the four tests for the power as a function of effect size (β) under the six models for 2,000 phenotypes (K = 2,000). MAF is 0.3. The sample size (n) is 4,000. $\rho_f = 0.2$, $\rho_e = 0.3$, and $c^2 = 0.5$. The power of all of the four tests is evaluated using 200 replicated samples at a nominal FDR level of 5%.

https://doi.org/10.1371/journal.pone.0276646.g002

across all six models. In contrast, the tests using NySi, NyBo, JiSi, and JiBo suffer FDR inflation, and the inflation is especially severe when the number of categories is large (Model 1, 2, and 3).

Real data applications

The UK Biobank is a population-based cohort study with a wide variety of genetic and phenotypic information [34]. It includes ~ 500K people from all around the United Kingdom who were aged between 40 and 69 when recruited in 2006–2010 [35]. Genotype and phenotype data from the UK Biobank have 488,377 participants with 784,256 variants on chromosomes 1–22 [36]. The preprocess of genotype is achieved by quality control (QC) which is performed on both genotypic variants and samples using PLINK 1.9 [37] (https://www.cog-genomics.org/plink/1.9/). We summarize the QC procedures in **S6 Fig in S1 File**. In QC, we filter out genetic variants with variant-based missing rates larger than 5%, p-values of Hardy-Weinberg equilibrium exact test less than 10^{-6}, and MAF less than 5%. We also filter out individuals with sample-based genotype missing rates larger than 5% and individuals without sex. After QC, there are 250,850 SNPs and 466,501 individuals remaining in the following analysis.

In this study, we define phenotypes using ICD-10 codes, a standardized coding system for defining disease status as well as for billing purposes [38]. After truncating each full ICD-10 code to the UK Biobank ICD-10 level 2 code (https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=41202), we generate a total of 1,869 unique phenotypes with the names of these phenotypes being the unique truncated ICD codes. For each individual, we denote the EMR-based phenotype for that individual as "1" if a corresponding truncated ICD code ever appears, otherwise, we denote the EMR-based phenotype as "0". To ensure the individuals in our analysis are from the same ancestry, we first restrict individuals to the individuals who self-report themselves from a white British ancestry and have very similar ancestry based on a principal component (PC) analysis of genotypes [34]. To avoid the low quality of phenotype data, we exclude individuals who are marked as outliers for heterozygosity or missing rates and have been identified to have ten or more third-degree relatives or closer. Finally, we also exclude individuals that are recommended for removal by the UK Biobank. After preprocessing the phenotype data, there are 337,285 individuals left (details described in **S6 Fig in S1 File**). It is worth noting that some individuals violate multiple criteria, therefore, the total number of individuals we start with minus the number of individuals that need to be removed does not necessarily equal the number of individuals we keep.

There are 260 blocks based on the UK Biobank ICD-10 level 1 code, therefore, 1,869 phenotypes from the UK Biobank ICD-10 level 2 can be classified into 260 blocks ($M = 260$). We further limit SNPs of interest to those SNPs reaching the genome-wide significance threshold 5×10^{-8}. On Oct. 21st, 2019, the GWAS catalog (https://www.ebi.ac.uk/gwas/) contains a total of 90,428 data entries covering 3,153 publications of 61,613 SNPs which contains 29,297 significant SNPs. Among 250,850 SNPs obtained from the UK Biobank after QC, there are 3,267 SNPs matched with those significant SNPs in GWAS Catalog. After preprocessing procedures, individuals with both genotype and phenotype information are used in our study. There is a total of 322,607 individuals across 3,267 common SNPs and 1,869 case-control phenotypes which are classified into 260 blocks. Furthermore, we adjust each phenotype by thirteen covariates, including age, sex, genotyping array, and the first 10 PCs [28].

Based on the results shown in **Tables 1** and **2**, we know that HCLC-FC, MultiPhen-FC, MANOVA-FC, and TATES-FC can control targeted FDR under all of the simulation models. However, in the UK Biobank data, most of the phenotypes have extremely unbalanced case-control ratios, where the case-control ratios of 1,869 phenotypes are ranged from 3.10×10^{-6} to 1.87×10^{-1}. Meanwhile, many widely used approaches for joint analysis of multiple phenotypes produce inflated type I error rates for such extremely unbalanced case-control phenotypes [39]. Notably, our proposed FDR control method assumes that the p-value of the test statistic in the m^{th} category, p_m, for $m = 1, \ldots, M$, follows a uniform distribution $U(0,1)$. Therefore, we first evaluate the distributions of the p-values under the null hypothesis for each of the four methods based on the UK Biobank data by permutation procedures. For each of the four tests, we randomly permute genotypes for each of the 3,267 SNPs. After permutation, 3,267 SNPs have no association with each of the 260 phenotypic blocks. Therefore, we consider 260 blocks and 3,267 SNPs as $260 \times 3{,}267 = 849{,}420$ replicated samples. For each replicated sample, we apply four tests for testing the association between each permuted SNP and each phenotypic block.

S7 Fig in S1 File shows the histogram of p-values and QQ plot for uniform distribution for each method based on 849,420 replicated samples. The red dashed line in the histogram represents the theoretical frequency ($849{,}420/25 \approx 33{,}977$) for the standard uniform distribution. The frequencies of the p-values of the HCLC method are the only ones that approach the theoretical frequency. We also calculate the genomic inflation factor (λ) and show the observed and expected p-values from the standard uniform distribution in quantile-quantile (QQ) plots for each method. In general, the genomic inflation factor λ should be close to 1 if the p-values fall within the standard uniform distribution [40]. In the QQ plots in **S7 Fig in S1 File**, our proposed HCLC method forms a line that's roughly straight and $\lambda = 0.99$, indicating that the p-values based on 849,420 replicated samples come from the standard uniform distribution. In contrast, $\lambda = 0.58$ for MultiPhen and λ for MANOVA, where the sample quantiles of these methods deviate from the theoretical quantile. Even though the genomic inflation factor of TATES is equal to 0.97 which is pretty satisfactory, the sample quantiles fluctuate around the theoretical quantiles slightly which is not as good as our proposed HCLC method. Here are the possible reasons why the other three methods do not satisfy the uniform distribution assumption of p-values, and only HCLC works. The main assumption of MANOVA is that phenotypes should be continuous. However, all of the phenotypes in the analysis are binary phenotypes that violate the main assumption of MANOVA. MultiPhen uses the likelihood ratio test statistic based on the proportional odds logistic regression and TATES uses the extended Simes procedure to integrate the p-values from the score test statistics for the univariate association tests. It has been shown that the commonly used methods, such as the likelihood ratio test and score test, can inflate type I error rates for unbalanced case-control studies [34] that may result in the non-uniform distribution of the p-values of these two methods under the null hypothesis. Even though our proposed method, HCLC, uses the score test statistic to test the association between each phenotype and a SNP, it then uses the CLC test statistics to combine the individual statistics linearly within each cluster and combine the between-cluster terms in a quadratic form [28]. Our real data analysis shows that CLC is robust to unbalanced case-control studies.

Since MultiPhen is very time-consuming for real data analysis, we apply the other three methods, HCLC, MANOVA, and TATES, to test the association between each of the 3,267 SNPs and each of the 260 phenotypic blocks. **S8 Fig in S1 File** shows the number of SNPs identified by the three methods. Although MANOVA-FC and TATES-FC identified more SNPs than HCLC-FC, they violate the uniform distribution assumption of p-values. Therefore, in the following, we focus on the SNPs identified by HCLC-FC.

There is a total of 3,267 significant SNPs related to different phenotypes in the GWAS Catalog. If a SNP is associated with at least one phenotype in a block, we define this block as a SNP-related phenotypic block in the GWAS Catalog. By controlling the FDR at the 5% level, HCLC-FC identifies 1,244 out of 3,267 SNPs that are significantly associated with at least one phenotypic block. **Table 3** lists the top nine SNPs identified by the HCLC-FC method. We use SNP rs3129716 as an example. rs3129716 is mapped to genes *HLA-DQB1* and *MTCO3P1*. By controlling the FDR at 5%, the FDR threshold is 5.96×10^{-3}. Using this threshold, HCLC-FC identifies 28 phenotypic blocks significantly associated with this SNP. Based on the GWAS catalog, 17 out of 28 phenotypic blocks (bold-faced) are reported to be significantly associated with this SNP.

To visualize the associations between SNPs and phenotypic blocks identified by our proposed HCLC-FC method, we use two sets of phenotypic blocks, the diseases of the circulatory system (I00-I99) and the malignant neoplasms (C00-C97) as examples. **Fig 3** and **S9 Fig in S1 File** are used to showcase interconnections among phenotypes due to shared genetic associations. **Fig 3** shows the associations between SNPs (red circle) and the diseases of the circulatory system phenotypic blocks (I00-I99; blue square) identified by the HCLC-FC method. There are a total of nine phenotypic blocks. **S9 Fig in S1 File** shows the associations between SNPs (red circle) and the set of malignant neoplasms phenotypic blocks (C00-C97; blue square). There are a total of 15 phenotypic blocks. From these two figures, we can see that many SNPs are associated with one phenotypic block while some SNPs are associated with multiple phenotypic blocks, which supports our hypothesis that some SNPs are associated with at least one

Table 3. The top nine SNPs that are associated with multiple phenotypic blocks identified by the HCLC-FC method based on the UK Biobank data. The information of the phenotypic blocks can be found at https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=41202. The bold-faced blocks indicate the associations with the corresponding SNP reported in the GWAS catalog. The number under the rs-number of SNP represents the total number of phenotypic blocks identified. FDR threshold is calculated at a nominal FDR level of 5%.

SNPs	Mapped Gene(s)	FDR Threshold	Phenotypic Blocks
rs3117582 (29)	*APOM*	5.96E-03	**C15-C26,** C81-C96, D50-D53, D60-D64, E10-E14, E15-E16, E20-E35, E70-E90, G50-G59, H00-H06, H30-H36, I70-I79, J40-J47, K20-K31, K40-K46, K50-K52, K70-K77, M15-M19, M20-M25, N00-N08, N20-N23, N30-N39, N40-N51, Q00-Q07, R00-R09, R30-R39, R50-R69, Y90-Y98, Z40-Z54
rs3129716 (28)	*HLA-DQB1*, *MTCO3P1*	5.96E-03	**C15-C26**, **C81-C96**, **D50-D53**, D60-D64, **D65-D69**, E15-E16, **E20-E35**, **E70-E90**, **G50-G59**, H00-H06, **H25-H28**, H30-H36, **H55-H59**, I70-I79, **J40-J47**, K20-K31, K40-K46, **K50-K52**, **L10-L14**, **L40-L45**, M15-M19, **M30-M36**, **N00-N08**, N40-N51, **R00-R09**, R10-R19, R30-R39, **R50-R69**
rs389884 (27)	*STK19*	5.58E-03	**C15-C26**, C81-C96, D50-D53, D60-D64, E10-E14, E15-E16, E20-E35, E70-E90, G50-G59, H00-H06, **H30-H36**, I70-I79, J40-J47, K40-K46, K50-K52, K70-K77, M15-M19, M20-M25, **N00-N08**, N20-N23, N30-N39, N40-N51, R00-R09, R30-R39, R50-R69, Y90-Y98, **Z40-Z54**
rs3134942 (25)	*NOTCH4*	5.19E-03	C81-C96, **D50-D53**, D60-D64, **E10-E14**, E15-E16, **E20-E35**, **E70-E90**, G50-G59, H00-H06, **H30-H36**, I70-I79, **J40-J47**, K40-K46, **K50-K52**, L20-L30, M15-M19, **M20-M25**, M45-M49, M50-M54, M80-M85, N00-N08, N30-N39, R00-R09, R30-R39, R50-R69
rs3130288 (23)	*ATF6B*	4.81E-03	C81-C96, D50-D53, D60-D64, E10-E14, E15-E16, E20-E35, G50-G59, H00-H06, H30-H36, I70-I79, **J40-J47**, K40-K46, K50-K52, M15-M19, **M20-M25**, N00-N08, N20-N23, N30-N39, N40-N51, R00-R09, R30-R39, R50-R69, Y90-Y98
rs3094005 (22)	*MICB*	4.62E-03	**C15-C26**, **C81-C96**, D50-D53, D60-D64, D80-D89, E10-E14, E15-E16, **E20-E35**, G50-G59, H00-H06, H30-H36, **I70-I79**, **J40-J47**, K40-K46, **K50-K52**, K70-K77, **M20-M25**, N00-N08, N20-N23, R00-R09, R30-R39, **Z40-Z54**
rs9270493 (22)	*HLA-DRB1*, *HLA-DQA1*	4.62E-03	**C69-C72**, D50-D53, D60-D64, **D65-D69**, D80-D89, **E10-E14**, E15-E16, **E20-E35**, G50-G59, **H00-H06**, H30-H36, **I70-I79**, K20-K31, **K50-K52**, L00-L08, **L40-L45**, **M05-M14**, **M20-M25**, **N00-N08**, N80-N98, R10-R19, R30-R39
rs35242582 (22)	*HLA-DQA1*	4.62E-03	A75-A79, **C15-C26**, **C43-C44**, **C60-C63**, **C73-C75**, **C76-C80,** D00-D09, **D50-D53**, E00-E07, E10-E14, **E20-E35**, **G35-G37**, **K50-K52**, L55-L59, **L80-L99**, **M05-M14**, M45-M49, M50-M54, N40-N51, **N80-N98**, O20-O29, U00-U49
rs1480380 (22)	*HLA-DMB*, *HLA-DMA*	4.42E-03	D10-D36, D50-D53, D60-D64, E00-E07, E10-E14, E15-E16, **E70-E90**, G10-G14, G50-G59, H00-H06, H30-H36, K20-K31, K40-K46, K50-K52, L10-L14, M15-M19, **M20-M25**, **N00-N08**, N80-N98, R00-R09, V30-V39, X60-X84

https://doi.org/10.1371/journal.pone.0276646.t003

phenotypic block. For example, Fig 3 shows that 108 SNPs are associated with the phenotypic block hypertensive diseases (I10-I15) and 17 out of 108 SNPs are associated with both hypertensive diseases (I10-I15) and Ischaemic heart diseases (I20-I25).

Discussion

GWAS have become a very effective research tool to investigate associations between genetic variation and a disease/phenotype. In spite of the success of GWAS in identifying thousands of reproducible associations between genetic variants and complex diseases, in general, the association between genetic variants and a single phenotype is usually weak. It is increasingly recognized that joint analysis of multiple phenotypes can be potentially more powerful than the univariate analysis and can shed new light on underlying biological mechanisms of complex diseases. As a complementary approach to GWAS, PheWAS analyze many phenotypes with a genetic variant and combine both the exploration of phenotypic structure and genotypic variation [11].

Similar to the widely used GWAS approaches, existing methods for PheWAS largely focus on the association between a single genetic variant with a large number of candidate phenotypes and test the association between one genetic variant and one phenotype at a time. In this paper, we develop a novel and powerful multivariate method, HCLC-FC, to test the association between a genetic variant with multiple phenotypes in each phenotypic category. HCLC-FC

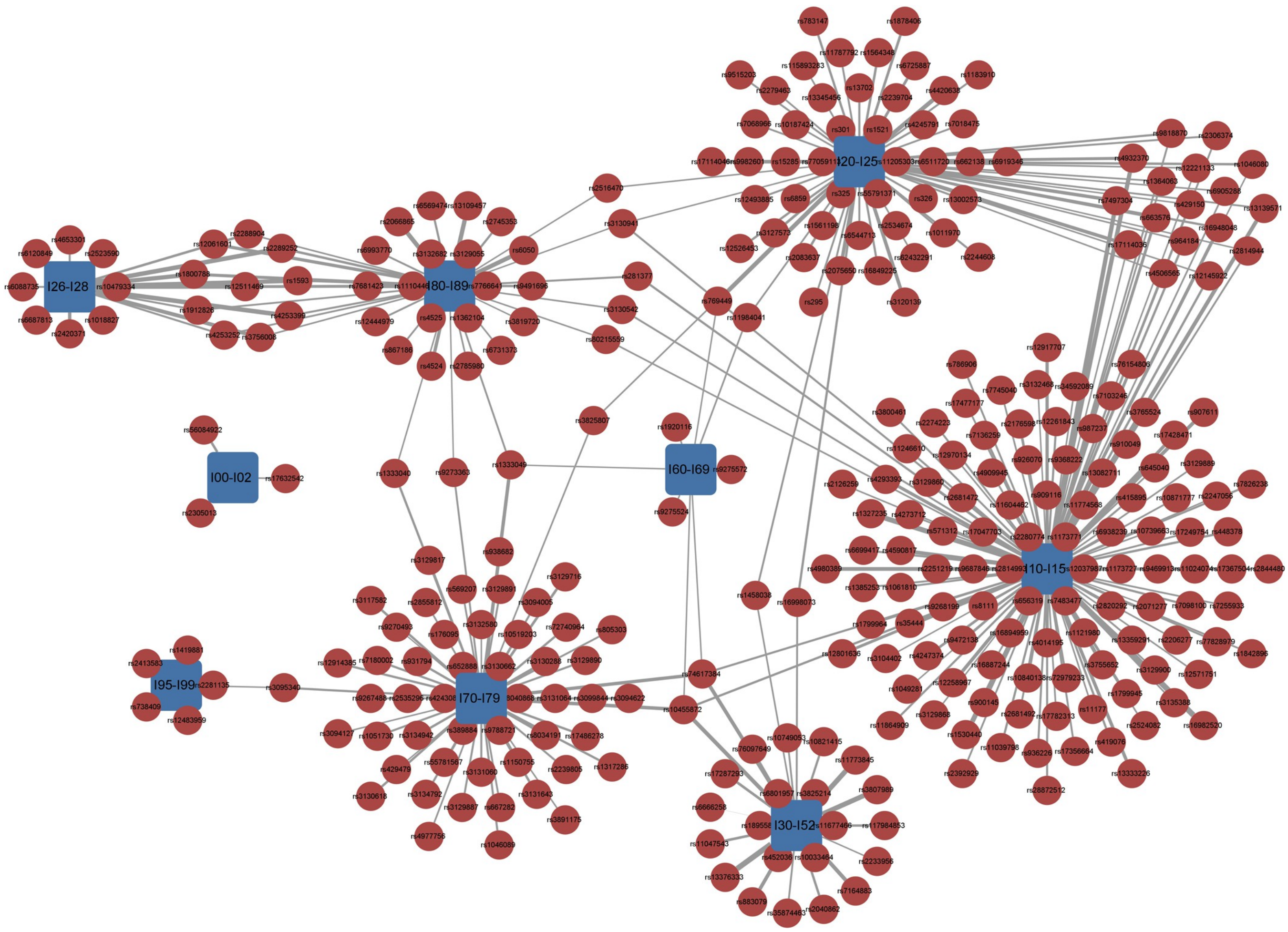

Fig 3. The associations between SNPs and the circulatory system phenotypic blocks identified by the HCLC-FC method. The red circles represent SNPs, and the blue squares represent nine diseases of the circulatory system phenotypic blocks I00-I99 (I00-I02: Acute rheumatic fever; I10-I15: Hypertensive diseases; I20-I25: Ischaemic heart diseases; I26-I28: Pulmonary heart disease and diseases of pulmonary circulation; I30-I52: Other forms of heart disease; I60-I69: Cerebrovascular diseases; I70-I79: Diseases of arteries, arterioles and capillaries; I80-I89: Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified; I95-I99: Other and unspecified disorders of the circulatory system). The width of the connection line represents the strength of association (-log10 scale p-value).

https://doi.org/10.1371/journal.pone.0276646.g003

involves three steps. In the first step, we use the bottom-up hierarchical clustering method [26] to partition a large number of phenotypes into disjoint clusters within each category. In the second step, we apply the clustering linear combination method [24] to combine test statistics within each category based on the phenotypic clusters and obtain a p-value from each phenotypic category. In the third step, we propose a large-scale association testing procedure with theoretical guarantees for FDR control under flexible correlation structures. We perform extensive simulation studies to compare the performance of HCLC-FC with that of other existing methods. The results show that the existing methods using our proposed FDR control procedure can control FDR at a nominal level, and our proposed HCLC-FC method outperforms the other three methods we compare under the six models for different within-group and between-group phenotypic correlation structures. Finally, we also evaluate the performance of HCLC-FC through a set of 1869 case-control phenotypes based on ICD-10 code across more than 300,000 samples from the UK Biobank, where these phenotypes can be classified into 260 ICD-10 level 1 blocks. The real data analysis results show that HCLC-FC not only can well-control type I error rates but also can identify 1,244 SNPs that have previously been reported to be associated with some phenotypes in the GWAS catalog.

As we all know, over the last decades, biobanks have been extremely prevalent in medical research [41] and enable access to a large collection of high-quality biological or medical data and tissue samples, which contain thousands of diseases/traits and a large sample size [42]. However, in biobanks, case-control ratios of most phenotypes are extremely unbalanced. Dey et al. [39] pointed out that a normal approximation of the score test statistic has inflated type I error rates for phenotypes with unbalanced case-control ratios. They proposed a score-test-based single-variant test that estimates the distribution of the test statistic by using the saddle-point approximation (SPA) [39] to control type I error rates and to adjust for covariates even in an extremely unbalanced case-control setting. Based on SPA, the Scalable and Accurate Implementation of GEneralized mixed model (SAIGE) was proposed to analyze large biobank data, controlling for both unbalanced case-control ratio and sample relatedness [43]. It uses SPA [39] to calibrate unbalanced case-control ratios in score tests based on logistic mixed models.

To extend our method to phenotypes with extreme unbalanced case-control ratios, we can apply the SPA method to adjust the score test statistics T_{mk} in Step 2. The adjusted test statistic of T_{mk} is given by $sign(T_{mk})\sqrt{F_{Chi}^{-1}(1-p_{mk}^{SPA})}$, where $F_{Chi}(.)$ denotes the cumulative distribution function of chi-squared distribution with one degree of freedom and p_{mk}^{SPA} is the p-value of T_{mk} calculated using SPA. Then, we can use the adjusted test statistic to calculate $T_{CLC}^{L_m}$. To extend our method to adjust for both case-control imbalance and family relatedness, we can apply the SAIGE method in Step 2 [43]. Instead of fitting the linear regression model, we can fit the null logistic mixed model to estimate the variance component and other model parameters, then test the association between each genetic variant and phenotype by applying SPA to the score test statistics. Finally, the adjusted test statistic can be used to calculate $T_{CLC}^{L_m}$. However, the performance of these approaches for phenotypes with extreme unbalanced case-control ratios needs further evaluations.

TreeWAS is another approach that was developed for identifying cross-disease components of genetic risk across hospital classification codes within a hierarchical ontology in the UK Biobank [44]. It is based on a Bayesian approach that can estimate a Bayes factor statistic for the evidence that genetic coefficients are nonzero for at least one node and also estimate the marginal posterior probability of each node with a nonzero genetic coefficient. The Bayes factor supports the evaluation of evidence in favor of a null hypothesis, rather than only allowing the null to be rejected or not rejected. However, calculating the Bayes factor based on the estimation of the marginal likelihoods of each model requires complicated and extensive time-consuming operations. Moreover, TreeWAS is based on the tree-structured disease and diagnostic ontologies that are built into the systematized coding of medical conditions from the biobank [45]. Therefore, it is developed within two sources of tree-structured phenotypic data sets from the UK Biobank, one is the hospitalization episode statistics data that are coded by ICD-10 codes, and the other one is the self-reported diagnoses that are coded using UK Biobank classification tree [44]. However, our proposed method, HCLC-FC, is not limited to biobank data; it is suitable to be applied to data sets with multiple phenotypes from electronic health records, epidemiological studies, and clinical trial data [46].

Given the extensibility of our method, there are some natural avenues for future work. 1) Our study has been mainly focused on testing the association between one genetic variant with a large number of phenotypes to identify cross-disease components of genetic risk. Future work is needed to extend the current single variant test to gene- or region-based multiple variant tests to improve the power to identify disease susceptibility genes. For example, some existing methods that are developed to test an optimally weighted combination of common and/or rare variants with multiple phenotypes can be used to each phenotypic category [47] in the second step of HCLC-FC. Then, our FDR-control procedure can be used to calculate the rejection threshold. In addition to utilizing existing methods, developing new methods that make use of both gene- or region-based SNPs and a large number of phenotypes is also a direction of our future research. 2) HCLC-FC needs individual-level data for the analysis. We can extend this methodology to use GWAS summary statistics by estimating the dissimilarity matrix using the cross-trait linkage disequilibrium (LD) score regression that requires only GWAS summary statistics [48, 49], then use this dissimilarity matrix to perform the HCM in the first step. However, the performance of these aforementioned extensions need to be evaluated carefully. We would like to pursue these important extensions in our future studies.

Despite the limitations of HCLC-FC, HCLC-FC has several important advantages over other existing methods for association studies using multiple phenotypes. First, it clusters phenotypes within each phenotypic category, which reduces the degrees of freedom of the association tests and has the potential to increase statistical power. Second, it is computationally fast and easy to implement. The CLC approach [24] uses a simulation procedure to estimate the p-value of the final test statistic. HCLC-FC has an asymptotic distribution which avoids the computational burden of permutations. Third, the newly developed FDR controlling process is based on p-values and does not depend on test statistics. Therefore, it is more general and can be applied to other multiple testing procedures to control FDR. Fourth, HCLC-FC can be used for both continuous and discontinuous phenotypes. It can be applied to data sets with multiple phenotypes from electronic health records, epidemiological studies, and clinical trial data.

Supporting information

S1 File.
(DOCX)

Acknowledgments

Part of this research has been conducted using the UK Biobank Resource under application number 41722 and the NHGRI-EBI GWAS Catalog. Superior, a high-performance computing infrastructure at Michigan Technological University, was used in obtaining results presented in this publication.

Author Contributions

Formal analysis: Xiaoyu Liang, Xuewei Cao.

Methodology: Xiaoyu Liang, Xuewei Cao, Qiuying Sha, Shuanglin Zhang.

Resources: Shuanglin Zhang.

Writing – original draft: Xiaoyu Liang, Xuewei Cao, Qiuying Sha, Shuanglin Zhang.

Writing – review & editing: Xiaoyu Liang, Xuewei Cao, Qiuying Sha, Shuanglin Zhang.

References

1. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLoS Comput Biol. 2012; 8(12): e1002822. https://doi.org/10.1371/journal.pcbi.1002822 PMID: 23300413
2. Guo X, Li Y, Ding X, He M, Wang X, Zhang H. Association Tests of Multiple Phenotypes: ATeMP. PLoS One. 2015; 10(10):e0140348. https://doi.org/10.1371/journal.pone.0140348 PMID: 26479245

3. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics. 2007; 23(10):1294–6. https://doi.org/10.1093/bioinformatics/btm108 PMID: 17384015

4. Aulchenko YS, Struchalin MV, van Duijn CM. ProbABEL package for genome-wide association analysis of imputed data. BMC Bioinformatics. 2010; 11:134. https://doi.org/10.1186/1471-2105-11-134 PMID: 20233392

5. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009; 10:387–406. https://doi.org/10.1146/annurev.genom.9.081307.164242 PMID: 19715440

6. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010; 34(8):816–34. https://doi.org/10.1002/gepi.20533 PMID: 21058334

7. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011; 8(10):833–5. https://doi.org/10.1038/nmeth.1681 PMID: 21892150

8. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007; 39(7):906–13. https://doi.org/10.1038/ng2088 PMID: 17572673

9. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81(3):559–75. https://doi.org/10.1086/519795 PMID: 17701901

10. Cai T, Cai TT, Liao K, Liu W. Large-Scale Simultaneous Testing of Cross-Covariance Matrices with Applications to PheWAS. Statistica Sinica. 2019; 29(2):983. https://doi.org/10.5705/ss.202017.0189 PMID: 31889766

11. Pendergrass SA, Brown-Gentry K, Dudek SM, Torstenson ES, Ambite JL, Avery CL, et al. The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. Genet Epidemiol. 2011; 35(5):410–22. https://doi.org/10.1002/gepi.20589 PMID: 21594894

12. Hebbring SJ. The challenges, advantages and future of phenome-wide association studies. Immunology. 2014; 141(2):157–65. https://doi.org/10.1111/imm.12195 PMID: 24147732

13. Du Y, Xie J, Chang W, Han Y, Cao G. Genome-wide association studies: inherent limitations and future challenges. Front Med. 2012; 6(4):444–50. https://doi.org/10.1007/s11684-012-0225-3 PMID: 23124883

14. Denny JC, Crawford DC, Ritchie MD, Bielinski SJ, Basford MA, Bradford Y, et al. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. Am J Hum Genet. 2011; 89(4):529–42. https://doi.org/10.1016/j.ajhg.2011.09.008 PMID: 21981779

15. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010; 26(9):1205–10. https://doi.org/10.1093/bioinformatics/btq126 PMID: 20335276

16. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet. 2010; 86(4):560–72. https://doi.org/10.1016/j.ajhg.2010.03.003 PMID: 20362271

17. Pendergrass SA, Brown-Gentry K, Dudek S, Frase A, Torstenson ES, Goodloe R, et al. Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. PLoS Genet. 2013; 9(1):e1003087. https://doi.org/10.1371/journal.pgen.1003087 PMID: 23382687

18. Verma A, Bang L, Miller JE, Zhang Y, Lee MTM, Zhang Y, et al. Human-Disease Phenotype Map Derived from PheWAS across 38,682 Individuals. Am J Hum Genet. 2019; 104(1):55–64. https://doi.org/10.1016/j.ajhg.2018.11.006 PMID: 30598166

19. Li G, Taljaard M, Van den Heuvel ER, Levine MA, Cook DJ, Wells GA, et al. An introduction to multiplicity issues in clinical trials: the what, why, when and how. Int J Epidemiol. 2017; 46(2):746–55. https://doi.org/10.1093/ije/dyw320 PMID: 28025257

20. Cole DA, Maxwell SE, Arvey R, Salas E. How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. Psychological bulletin. 1994; 115(3):465.

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

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

23. Liang X, Wang Z, Sha Q, Zhang S. An Adaptive Fisher's Combination Method for Joint Analysis of Multiple Phenotypes in Association Studies. Sci Rep. 2016; 6:34323. https://doi.org/10.1038/srep34323 PMID: 27694844

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

25. Li X, Zhang S, Sha Q. Joint analysis of multiple phenotypes using a clustering linear combination method based on hierarchical clustering. Genet Epidemiol. 2020; 44(1):67–78. https://doi.org/10.1002/gepi.22263 PMID: 31541490

26. Liang X, Sha Q, Rho Y, Zhang S. A hierarchical clustering method for dimension reduction in joint analysis of multiple phenotypes. Genet Epidemiol. 2018; 42(4):344–53. https://doi.org/10.1002/gepi.22124 PMID: 29682782

27. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38(8):904–9. https://doi.org/10.1038/ng1847 PMID: 16862161

28. Sha Q, Wang X, Wang X, Zhang S. Detecting association of rare and common variants by testing an optimally weighted combination of variants. Genet Epidemiol. 2012; 36(6):561–71. https://doi.org/10.1002/gepi.21649 PMID: 22714994

29. Bühlmann P, Rütimann P, van de Geer S, Zhang CH. Correlated variables in regression: clustering and sparse estimation. Journal of Statistical Planning and Inference. 2013; 143(no. 11):1835–58.

30. Asif H, Alliey-Rodriguez N, Keedy S, Tamminga CA, Sweeney JA, Pearlson G, et al. GWAS significance thresholds for deep phenotyping studies can depend upon minor allele frequencies and sample size. Mol Psychiatry. 2021; 26(6):2048–55. https://doi.org/10.1038/s41380-020-0670-3 PMID: 32066829

31. Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet. 2004; 74(4):765–9. https://doi.org/10.1086/383251 PMID: 14997420

32. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity (Edinb). 2005; 95(3):221–7. https://doi.org/10.1038/sj.hdy.6800717 PMID: 16077740

33. Šidák Z. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association. 1967; 62(318):626–33.

34. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018; 562(7726):203–9. https://doi.org/10.1038/s41586-018-0579-z PMID: 30305743

35. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015; 12(3):e1001779. https://doi.org/10.1371/journal.pmed.1001779 PMID: 25826379

36. McGuirl MR, Smith SP, Sandstede B, Ramachandran S. Hierarchical clustering of gene-level association statistics reveals shared and differential genetic architecture among traits in the UK Biobank. bioRxiv. 2019:565903.

37. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015; 4(1):s13742-015-0047-8.

38. Denny JC, Bastarache L, Roden DM. Phenome-wide association studies as a tool to advance precision medicine. Annual review of genomics and human genetics. 2016; 17:353–73. https://doi.org/10.1146/annurev-genom-090314-024956 PMID: 27147087

39. Dey R, Schmidt EM, Abecasis GR, Lee S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. Am J Hum Genet. 2017; 101(1):37–49. https://doi.org/10.1016/j.ajhg.2017.05.014 PMID: 28602423

40. Laird NM, Lange C. The fundamentals of modern statistical genetics: Springer; 2011.

41. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. Nature Reviews Genetics. 2019; 20(8):467–84. https://doi.org/10.1038/s41576-019-0127-1 PMID: 31068683

42. Greely HT. The uneasy ethical and legal underpinnings of large-scale genomic biobanks. Annu Rev Genomics Hum Genet. 2007; 8:343–64. https://doi.org/10.1146/annurev.genom.7.080505.115721 PMID: 17550341

43. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet. 2018; 50(9):1335–41. https://doi.org/10.1038/s41588-018-0184-y PMID: 30104761

44. Cortes A, Dendrou CA, Motyer A, Jostins L, Vukcevic D, Dilthey A, et al. Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. Nat Genet. 2017; 49 (9):1311–8. https://doi.org/10.1038/ng.3926 PMID: 28759005

45. Cox NJ. Reaching for the next branch on the biobank tree of knowledge. Nat Genet. 2017; 49(9):1295–6. https://doi.org/10.1038/ng.3946 PMID: 28854181

46. Verma A, Ritchie MD. Current Scope and Challenges in Phenome-Wide Association Studies. Curr Epidemiol Rep. 2017; 4(4):321–9. https://doi.org/10.1007/s40471-017-0127-7 PMID: 29545989

47. Wang Z, Sha Q, Fang S, Zhang K, Zhang S. Testing an optimally weighted combination of common and/or rare variants with multiple traits. PLoS One. 2018; 13(7):e0201186. https://doi.org/10.1371/journal.pone.0201186 PMID: 30048520

48. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015; 47(3):291–5. https://doi.org/10.1038/ng.3211 PMID: 25642630

49. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015; 47(11):1236–41. https://doi.org/10.1038/ng.3406 PMID: 26414676