2-28-2023

A clustering linear combination method for multiple phenotype association studies based on GWAS summary statistics

Meida Wang
Michigan Technological University, meidaw@mtu.edu

Xuewei Cao
Michigan Technological University, xueweic@mtu.edu

Shuanglin Zhang
Michigan Technological University, shuzhang@mtu.edu

Qiuying Sha
Michigan Technological University, qsha@mtu.edu

scientific reports

Check for updates

OPEN

A clustering linear combination method for multiple phenotype association studies based on GWAS summary statistics

Meida Wang, Xuewei Cao, Shuanglin Zhang & Qiuying Sha ✉

There is strong evidence showing that joint analysis of multiple phenotypes in genome-wide association studies (GWAS) can increase statistical power when detecting the association between genetic variants and human complex diseases. We previously developed the Clustering Linear Combination (CLC) method and a computationally efficient CLC (ceCLC) method to test the association between multiple phenotypes and a genetic variant, which perform very well. However, both of these methods require individual-level genotypes and phenotypes that are often not easily accessible. In this research, we develop a novel method called sCLC for association studies of multiple phenotypes and a genetic variant based on GWAS summary statistics. We use the LD score regression to estimate the correlation matrix among phenotypes. The test statistic of sCLC is constructed by GWAS summary statistics and has an approximate Cauchy distribution. We perform a variety of simulation studies and compare sCLC with other commonly used methods for multiple phenotype association studies using GWAS summary statistics. Simulation results show that sCLC can control Type I error rates well and has the highest power in most scenarios. Moreover, we apply the newly developed method to the UK Biobank GWAS summary statistics from the XIII category with 70 related musculoskeletal system and connective tissue phenotypes. The results demonstrate that sCLC detects the most number of significant SNPs, and most of these identified SNPs can be matched to genes that have been reported in the GWAS catalog to be associated with those phenotypes. Furthermore, sCLC also identifies some novel signals that were missed by standard GWAS, which provide new insight into the potential genetic factors of the musculoskeletal system and connective tissue phenotypes.

Over the last decades, genome-wide association studies (GWAS) have been very successful in detecting genetic variants associated with human complex traits or diseases[1–3]. At the same time, a vast majority of GWAS summary statistics obtained from single-trait tests are publicly available, which contain the estimated marginal effect sizes, the corresponding standard deviations, Z scores or p-values. Normally, raw genotypes and phenotypes are not easy to be accessed as a result of privacy concerns and some logistical considerations, thus motivating an extensive interest in developing statistical methods based on GWAS summary statistics[4–6]. On the other hand, because multiple related phenotypes are often measured as indicators for one specific trait, considering the correlated structure between multiple phenotypes and jointly analyzing these phenotypes may increase statistical power in association studies[7–12].

Recently, many multiple phenotype association tests based on GWAS summary statistics have been proposed. CPASSOC[13] contains two separate tests (Hom and Het), where Hom is more powerful when the genetic variant has homogeneous effects on the phenotypes; Het is more powerful when heterogeneous effects are present, whereas Monte-Carlo simulations are needed to calculate the p-value of Het when the number of traits is large, which is computationally intensive. SSU[14,15] is a test statistic based on the sum of squared Z scores, which follows a mixture of chi-squared distributions under the null hypothesis. PCFisher[16] has the test statistic that combines all p-values of independent principal components using Fisher's method, where allocates larger weights to PCs with smaller eigenvalues. The classical Wald test[16] uses the Z score vector and the inverse matrix of the correlation matrix among phenotypes to construct a quadratic test statistic. The adaptive multi-trait association test (aMAT)[17] builds a group of multi-phenotype association tests (MATs) that may have good performance in a specific scenario and then integrates the testing results adaptively.

Mathematical Sciences, Michigan Technological University, Houghton, MI, USA. ✉email: qsha@mtu.edu

In our previous studies, we developed the Clustering Linear Combination (CLC) method[18] and a computationally efficient CLC (ceCLC) method[19] to test the association between multiple phenotypes and a genetic variant based on individual level genotypes and phenotypes. Both of these methods perform very well compared with other multiple phenotypes association tests especially for phenotypes that have natural grouping. In this research, we develop a novel approach called CLC based on GWAS summary statistic (sCLC). In sCLC, we use the LD score regression[20,21] to estimate the correlation matrix among phenotypes. It has been shown that the LD score regression which has been commonly used in recent years can control the potential confounders such as population stratification, unknown sample overlap, cryptic relatedness, and so forth[20–22]. In our simulation studies, we consider a range of simulation settings and compare sCLC with other five commonly used methods for multiple phenotype association studies using GWAS summary statistics to evaluate the performance of sCLC. The simulation results show that sCLC can control the Type I error rate well and has the highest power in most scenarios. We also apply the sCLC method to UK Biobank GWAS summary statistics for 70 related musculoskeletal system and connective tissue phenotypes in the XIII category of UK Biobank. The results show that sCLC identifies the most number of significant SNPs, and most of these SNPs can be matched to the genes that have been reported in the GWAS catalog to be associated with the phenotypes in the XIII category. Furthermore, sCLC also identifies some novel signals that were missed by standard GWAS. The new identified signals may provide new insight into the potential genetic factors of the musculoskeletal system and connective tissue phenotypes.

Materials and methods

We consider a GWAS with M SNPs and K correlated phenotypes of interest. Each time, a single SNP j is considered, then we repeat the same procedure for all SNPs, $j = 1, \cdots, M$. For SNP j, we assume that we have Z score vector $\boldsymbol{Z_j} = (Z_{1j}, Z_{2j}, \cdots, Z_{Kj})^T$ across K phenotypes from GWAS summary statistics. If Z score is not provided, we can compute the Z score as $Z_{kj} = \frac{\widehat{\beta}_{kj}}{\widehat{se}(\widehat{\beta}_{kj})}$, $k = 1, \cdots, K$, where $\widehat{\beta}_{kj}$ is the estimated effect size of SNP j on phenotype k, and $\widehat{se}(\widehat{\beta}_{kj})$ is the standard deviation of $\widehat{\beta}_{kj}$. Based on the GWAS summary statistics, we propose the following sCLC method.

Firstly, sCLC uses the LD score regression (LDSC)[20,21] to estimate the correlation matrix among phenotypes, denoted by $\boldsymbol{R}$. Specifically, consider the pair of phenotypes s and k, the bivariate LDSC[20] regresses the pairwise product of Z scores on the LD scores, the expected value of $Z_{sj}Z_{kj}$ is:

$$E\left(Z_{sj}Z_{kj}\right) = G_g l_j + \rho_{sk},$$

where G_g is related to the genetic covariance between phenotypes s and k; l_j is the LD score of SNP j which can be obtained from the reference panel[20,21]; and ρ_{sk} is the correlation between phenotypes s and k. Therefore, the bivariate LDSC[20] can be applied to each pair of phenotypes, and the estimated intercepts ρ_{sk} are used to estimate the off-diagonal elements of $\boldsymbol{R}$. When $s = k$, it reduces to the univariate LDSC[21] for each phenotype and the estimated intercepts are used to estimate the diagonal elements of $\boldsymbol{R}$. In this procedure, all M SNPs are used to estimate $\boldsymbol{R}$, and the LD scores for SNPs can be obtained from the reference panel, such as the 1000 Genome Project[23]. Moreover, LDSC can control potential confounders such as population stratification, unknown sample overlap, cryptic relatedness, and so forth[20–22].

Secondly, similar to CLC[18], we use the hierarchical clustering approach with similarity matrix $\boldsymbol{R}$ and dissimilarity matrix $1 - \boldsymbol{R}$ to partition the original K phenotypes into L disjoint clusters ($L = 1, 2, \ldots, K$). The agglomerative hierarchical clustering starts with each phenotype as a singleton cluster ($L = K$) and then successively merges pairs of clusters that have the smallest distance (highest similarity) until all clusters have been merged into a single cluster that contains all phenotypes ($L = 1$)[24]. Because we consider a single SNP j and multiple phenotypes at a time, the notation $\boldsymbol{Z_j}$ can be simplified by $\boldsymbol{Z}$. After applying the hierarchical clustering method to partition the original K phenotypes into L disjoint clusters ($L = 1, 2, \ldots, K$), we define a $K \times L$ matrix $\boldsymbol{B}$ with the $(k, l)^{th}$ element equals 1 if the kth phenotype belongs to the lth cluster, otherwise it equals 0. Then the CLC test statistic to test the association between the K phenotypes and a SNP with L clusters is given by:

$$T^L_{CLC} = (\boldsymbol{WZ})^T \left(\boldsymbol{WRW}^T\right)^{-1} (\boldsymbol{WZ}),$$

where $\boldsymbol{W} = \boldsymbol{B}^T \boldsymbol{R}^{-1}$. T^L_{CLC} follows a χ^2 distribution with degrees of freedom L under the null hypothesis. We denote the p-value of T^L_{CLC} by p_L for $1 \le L \le K$.

Finally, we use Cauchy combination[25,26] to integrate the p-values obtained from the second step for all possible number of clusters, p_L for $1 \le L \le K$. The test statistic of sCLC for a SNP is defined as the linear combination of the transformed p-values divided by K (all possible number of clusters), which is given by

$$T_{sCLC} = \frac{1}{K} \sum\nolimits_{L=1}^{K} \tan\left((0.5 - p_L)\pi\right).$$

Under the null hypothesis, p_L follows a standard uniform distribution, so $\tan\left((0.5 - p_L)\pi\right)$ has a standard Cauchy distribution. Because $p_1, \cdots, p_K$ correspond to each possible number of clusters for K phenotypes, there exists a correlated structure between them. Liu et al.[25,26] showed that a weighted sum of "correlated" standard Cauchy variables still has an approximately Cauchy tail, and the influence of the correlated structure on the tail is quite limited because of the heaviness of the Cauchy tail. Therefore, T_{sCLC} is approximately standard Cauchy distributed. Based on the cumulative density distribution of the standard Cauchy distribution, the p-value of T_{sCLC} can be approximated by $0.5 - (\arctan(T_{sCLC})/\pi)$.

Comparison of methods

To better demonstrate the performance of the sCLC approach, we compare sCLC with other five methods for multiple phenotype association studies using GWAS summary statistics: SSU[14,15], Hom[13], PCFisher[16], Wald[16], and aMAT[17]. Below, we briefly summarize these five methods, where $\boldsymbol{Z}$ score vector and the phenotypic correlation matrix $\boldsymbol{R}$ are the same as we define previously.

SSU. The test statistic of SSU is $T_{SSU} = \boldsymbol{Z}^T\boldsymbol{Z}$ and the distribution of T_{SSU} can be well approximated by $a\chi^2_d + b$ with $a = \frac{\sum_{i=1}^K c_i^3}{\sum_{i=1}^K c_i^2}$, $b = \sum_{i=1}^K c_i - \frac{(\sum_{i=1}^K c_i^2)^2}{\sum_{i=1}^K c_i^3}$, and $d = \frac{(\sum_{i=1}^K c_i^2)^3}{(\sum_{i=1}^K c_i^3)^2}$, where c_i s are the eigenvalues of $\boldsymbol{R}$. The p value of T_{SSU} can be obtained by $p(\chi^2_d > (T_{SSU} - b)/a)$. Note that the degrees of freedom of T_{SSU} may be less than K with highly correlated phenotypes.

Hom. Assume that there are summary statistics of GWASs from J cohorts with K traits. Let T_{ijk} be a summary statistic for the ith SNP, jth cohort, and kth trait. Let $\boldsymbol{T}_i = (T_{i11}, \ldots, T_{iJ1}, \ldots, T_{i1K}, \ldots, T_{iJK})^T$. For simplification, we omit the SNP index, then $\boldsymbol{T} = (T_{11}, \ldots, T_{J1}, \ldots, T_{1K}, \ldots, T_{JK})^T$ represents a vector of test statistics for single SNP-trait association tests. The test statistic of Hom is $S_{Hom} = \frac{\boldsymbol{e}^T(\boldsymbol{RV})^{-1}\boldsymbol{T}(\boldsymbol{e}^T(\boldsymbol{RV})^{-1}\boldsymbol{T})^T}{\boldsymbol{e}^T(\boldsymbol{VRV})^{-1}\boldsymbol{e}}$, which follows a χ^2 distribution with one degree of freedom, where $\boldsymbol{e}^T = (1, \ldots, 1)$ is a vector of length $J \times K$ with all elements being 1, $\boldsymbol{V}$ is a diagonal matrix of weights $w_{jk} = \sqrt{n_j}$, and n_j is the sample size in the jth cohort. In this study, we consider $J = 1$ cohort to compare Hom with other methods.

PCFisher. Assume that the spectral decomposition of $\boldsymbol{R}$ is $\boldsymbol{R} = \sum_{m=1}^K \lambda_m \mathbf{u}_m \mathbf{u}_m^T$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K > 0$ are the eigenvalues of $\boldsymbol{R}$, and $\mathbf{u}_m$ is the eigenvector corresponding to the m th largest eigenvalue λ_m. We assume that the K-dimensional vector of the summary statistics $\boldsymbol{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{R})$. It can be shown that[16] $PC_m = \mathbf{u}_m^T\boldsymbol{Z} \sim N\left(\mathbf{u}_m^T\boldsymbol{\mu}, \lambda_m\right), 1 \leq m \leq K$. The non-centrality parameter (ncp) of PC_m under the alternative hypothesis is $ncp_m = (\mathbf{u}_m^T\boldsymbol{\mu})^2/\lambda_m$. PCFisher[16] combines p-values of all K independent principal components using Fisher's method with its null distribution and the test statistic is given by $\mathrm{PCFihser} = -2\sum_{m=1}^K \log(p_m) \sim \chi^2_{2K}$.

Wald. The test statistic of Wald test is defined as $T_{Wald} = \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}$. Assume that the spectral decomposition of $\boldsymbol{R}$ is $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T = \sum_{m=1}^K \lambda_m \mathbf{u}_m \mathbf{u}_m^T$, then the test statistic can be written as $T_{Wald} = \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} = \left(\boldsymbol{U}^T\boldsymbol{Z}\right)^T\boldsymbol{\Lambda}^{-1}\left(\boldsymbol{U}^T\boldsymbol{Z}\right) = \sum_{m=1}^K \frac{PC_m^2}{\lambda_m} \sim \chi^2_K$. So, the Wald test is a special quadratic PC-based test[16].

aMAT. The method was developed to deal with potential (near) singularity problem of $\boldsymbol{R}$. The singular value decomposition (SVD) of $\boldsymbol{R}$ is $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^T$. A modified pseudoinverse $\boldsymbol{R}^+_\gamma$ is calculated by $\boldsymbol{R}^+_\gamma = \boldsymbol{U}\boldsymbol{\Sigma}^+_\gamma\boldsymbol{U}^T$, where Σ^+_γ is formed from Σ by taking the reciprocal of the largest m singular values $\sigma_1, \ldots, \sigma_m$, and setting all other elements to zero, where m is the largest integer that satisfies $\sigma_1/\sigma_m < \gamma$. The test statistic of $\mathrm{MAT}_{(\gamma)}$ is defined as $T_{\mathrm{MAT}_{(\gamma)}} = \boldsymbol{Z}^T\boldsymbol{R}^+_\gamma\boldsymbol{Z}$. Because the optimal value of γ is unknown, aMAT combines the results from a class of MAT tests, $T_{\mathrm{aMAT}} = \min_{\gamma\in\Gamma} p_{\mathrm{MAT}(\gamma)}$, where $p_{\mathrm{MAT}(\gamma)}$ is the p value of $\mathrm{MAT}_{(\gamma)}$, and $\Gamma = (1, 10, 30, 50)$. Finally, a Gaussian copula approximation is applied to calculate the p-value of aMAT. Therefore, aMAT is analogous to a PC-based method which restricts the analysis to the top m axes of the largest variation[17].

Results

Simulation design. Based on a widely used simulation procedure[17,27], we generate Z scores from a multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{R})$. We consider two different correlation matrix structures: (1) $\boldsymbol{R}$ is the sample correlation matrix of 70 related musculoskeletal system and connective tissue phenotypes in the UK Biobank (details of the 70 phenotypes are described in the Application to UK Biobank summary statistics); and (2) $\boldsymbol{R}$ is generated based on the Autoregressive model (AR(1) model)[28] for 40 phenotypes, where $\boldsymbol{R} = Bdiag(\boldsymbol{R}_1, \boldsymbol{R}_2, \boldsymbol{R}_3, \boldsymbol{R}_4)$, a block diagonal matrix, with $\boldsymbol{R}_1 = \boldsymbol{R}_3 = (r_{sk}) = \rho^{|s-k|}$ and $\boldsymbol{R}_2 = \boldsymbol{R}_4 = -\rho^{|s-k|}$. We use $\rho = 0.1$ in the simulation studies.

To investigate how the estimation error of $\boldsymbol{R}$ may affect on the testing results, similar to Wu[17], we consider two cases in the 70 phenotypic correlation matrix structure. In the first case, we suppose that $\boldsymbol{R}$ is known and perform our proposed method, sCLC, and all competing methods based on $\boldsymbol{R}$. In the second case, we suppose that $\boldsymbol{R}$ is unknown and the estimated phenotypic correlation matrix is approximated by $\boldsymbol{R}$ with a small white noise $N(0, \delta)$, denoted by $\boldsymbol{R}(\delta)$. We choose $\delta = 10^{-5}$ and $\delta = 10^{-4}$ in the simulation studies, and use $\boldsymbol{R}(\delta)$ in the association tests for all the methods.

To evaluate Type I error rate of sCLC, we generate 10^8 $\boldsymbol{Z}$ score vectors under the null hypothesis ($\boldsymbol{\mu} = 0$) and choose different significant levels. In order to evaluate power, we generate 10^4 $\boldsymbol{Z}$ score vectors under an alternative with different effect size vector $\boldsymbol{\mu}$ in four scenarios. In the first two scenarios, we assume that the SNP impacts on phenotypes with the same direction. Scenario 3 considers different directions of effects on phenotypes. Scenario 4 is a sparse simulation model, where a SNP impacts on a small proportion of phenotypes. The significant level of 5×10^{-8} is chosen for the power evaluation.

Scenario 1: Generate $\boldsymbol{\mu} = \beta(1/K, 2/K, \ldots, 1)^T$.

Scenario 2: Generate $\boldsymbol{\mu} = (\underbrace{0, 0, ..., 0}_{K/2}, \underbrace{\beta, \beta, ..., \beta}_{K/2})^T$.

Scenario 3: Generate $\boldsymbol{\mu} = (\beta_{11}, \ldots, \beta_{1k}, \beta_{21}, \ldots, \beta_{2k}, \beta_{31}, \ldots, \beta_{3k}, \beta_{41}, \ldots, \beta_{4k}, \beta_{51}, \ldots, \beta_{5k})^T$, where $\beta_{11} = \cdots = \beta_{1k} = \beta_{21} = \cdots = \beta_{2k} = 0, \beta_{31} = \cdots = \beta_{3k} = \beta_{41} = \cdots = \beta_{4k} = \beta, (\beta_{51}, \cdots, \beta_{5k}) = -\frac{2\beta}{k+1}(1, \cdots, k)$, and $k = K/5$.

Scenario 4: Generate $\boldsymbol{\mu} = (\beta_{11}, \ldots, \beta_{1k}, \beta_{21}, \ldots, \beta_{2k}, \beta_{31}, \ldots, \beta_{3k}, \ldots, \beta_{14,1}, \ldots, \beta_{14,k})^T$. $\beta_{11} = \cdots = \beta_{1k} = \beta_{21} = \cdots = \beta_{2k} = \cdots = \beta_{13,1} = \cdots = \beta_{13,k} = 0, (\beta_{14,1}, \cdots, \beta_{14,k}) = \frac{2\beta}{k+1}(1, \cdots, k)$, and $k = K/14$.

Simulation results. *Type I error rates.* Table 1 shows the estimated Type I error rates at different significance levels for all six methods with the phenotypic correlation matrix $\boldsymbol{R}$ of 70 phenotypes. The Type I error rates with the correlation matrix $\boldsymbol{R}(10^{-5})$ and $\boldsymbol{R}(10^{-4})$ of 70 phenotypes are recorded in Tables S1 and S2. From these tables, we can see that the sCLC approach can control the Type I error rates very well at different significant levels α, which indicates that it is a valid test. Among the five competing methods, SSU yields inflated Type I error rates when α is smaller and the other four methods can control Type I error rates very well. Table S3 shows the estimated Type I error rates at different significance levels for all six methods with the phenotypic correlation structure for the 40 phenotypes. We observe that all methods can well-control Type I error rates.

Power comparisons. Power comparison results of the six methods under four scenarios with the phenotypic correlation matrix $\boldsymbol{R}$ of 70 phenotypes are presented in Fig. 1. Figures S1 and S2 show the power comparisons of the six methods with the correlation matrix $\boldsymbol{R}(10^{-5})$ and $\boldsymbol{R}(10^{-4})$ of 70 phenotypes, respectively. From these figures, we can observe that (1) when SNPs have homogeneous effects on the phenotypes (scenarios 1 and 2), our proposed method sCLC, as well as Hom and SSU have higher power than the other three PC-based methods (Wald, aMAT, and PCFisher); whereas all the methods have comparable powers except for Hom when the SNP affects on phenotypes in different directions. (2) The power of Hom dramatically reduces and almost is zero in scenarios 3, while sCLC and SSU are robust to the direction of the genetic effect on the phenotypes. (3) sCLC and SSU are more powerful than other methods when a SNP affects on a small proportion of phenotypes (scenario 4), and Hom is less powerful in this case. (4) In all of the four scenarios, the power patterns observed in Figs. S1 and S2 are very close to that of Fig. 1, indicating that the estimation errors (noise δ) of $\boldsymbol{R}$ have little influence on the powers for all the methods. Figure S3 shows the power comparisons of the six methods with the phenotypic correlation structure for the 40 phenotypes. sCLC is still more powerful than the other five methods under all four scenarios.

Application to UK biobank summary statistics

Connective tissue dysplasia (CTD) and musculoskeletal disorders[29–31], such as Systemic Lupus Erythematosus (SLE), Sjögren Syndrome (SS), and Rheumatoid Arthritis (RA), may influence the physical activity or movement of patients. These kinds of diseases seriously affect the quality of life of people and have been reported to be potentially affected by genetic factors[32]. In this paper, we consider the GWAS summary statistics in the XIII category of UK Biobank with 70 musculoskeletal system and connective tissue phenotypes to detect potential genetic factors.

The UK Biobank is a large long-term biobank study which has recruited almost half a million participants in the UK, enrolled at ages from 40 to 69[33]. Sequenced genotypes for 488,377 participants with 784,256 variants in autosomal chromosomes were extracted by UK Biobank dataset[34]. Similar to Liang et al.[28], we first perform quality controls (QCs) on genotypes and individuals by using PLINK 1.9[35]. We remove SNPs with missing rates larger than 5%, p-values from Hardy–Weinberg equilibrium exact test less than 10^{-6}, and minor allele frequency (MAF) less than 5%. In addition, we screen out individuals with missing genotype rate larger than 5% and without sex information. After these pre-processing, there are 466,580 individuals with 288,647 genetic variants left.

On the other hand, the phenotypes that coded by International Classification of Diseases, the 10th Revision (ICD-10) codes are considered in our study. We truncate the full ICD-10 code to the UK Biobank ICD-10 level 3 code (http://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=41202) to define Electronic Health Record (EHR)-derived phenotypes. When the individual has the truncated ICD-10 code recorded for a specific phenotype, the

α	$\mathbf{1 \times 10^{-3}}$	$\mathbf{1 \times 10^{-4}}$	$\mathbf{1 \times 10^{-5}}$	$\mathbf{1 \times 10^{-6}}$	$\mathbf{1 \times 10^{-7}}$
SSU	1.05×10^{-3}	$\mathbf{1.13 \times 10^{-4}}$	$\mathbf{1.25 \times 10^{-5}}$	$\mathbf{1.61 \times 10^{-6}}$	$\mathbf{2.29 \times 10^{-7}}$
sCLC	1.07×10^{-3}	1.05×10^{-4}	1.06×10^{-5}	1.17×10^{-6}	7.98×10^{-8}
Hom	1.00×10^{-3}	9.82×10^{-5}	1.01×10^{-5}	9.47×10^{-7}	9.97×10^{-8}
Wald	1.01×10^{-3}	1.00×10^{-4}	9.98×10^{-6}	1.17×10^{-6}	1.7×10^{-7}
aMAT	9.97×10^{-4}	1.00×10^{-4}	1.02×10^{-5}	1.17×10^{-6}	1.3×10^{-7}
PCFisher	1.00×10^{-3}	9.90×10^{-5}	1.01×10^{-5}	1.09×10^{-6}	1.5×10^{-7}

Table 1. The estimated Type I error rates at different significance levels for the six methods with the phenotypic correlation structure for the 70 phenotypes. The bold-faced values indicate that the type I error rates cannot be controlled.

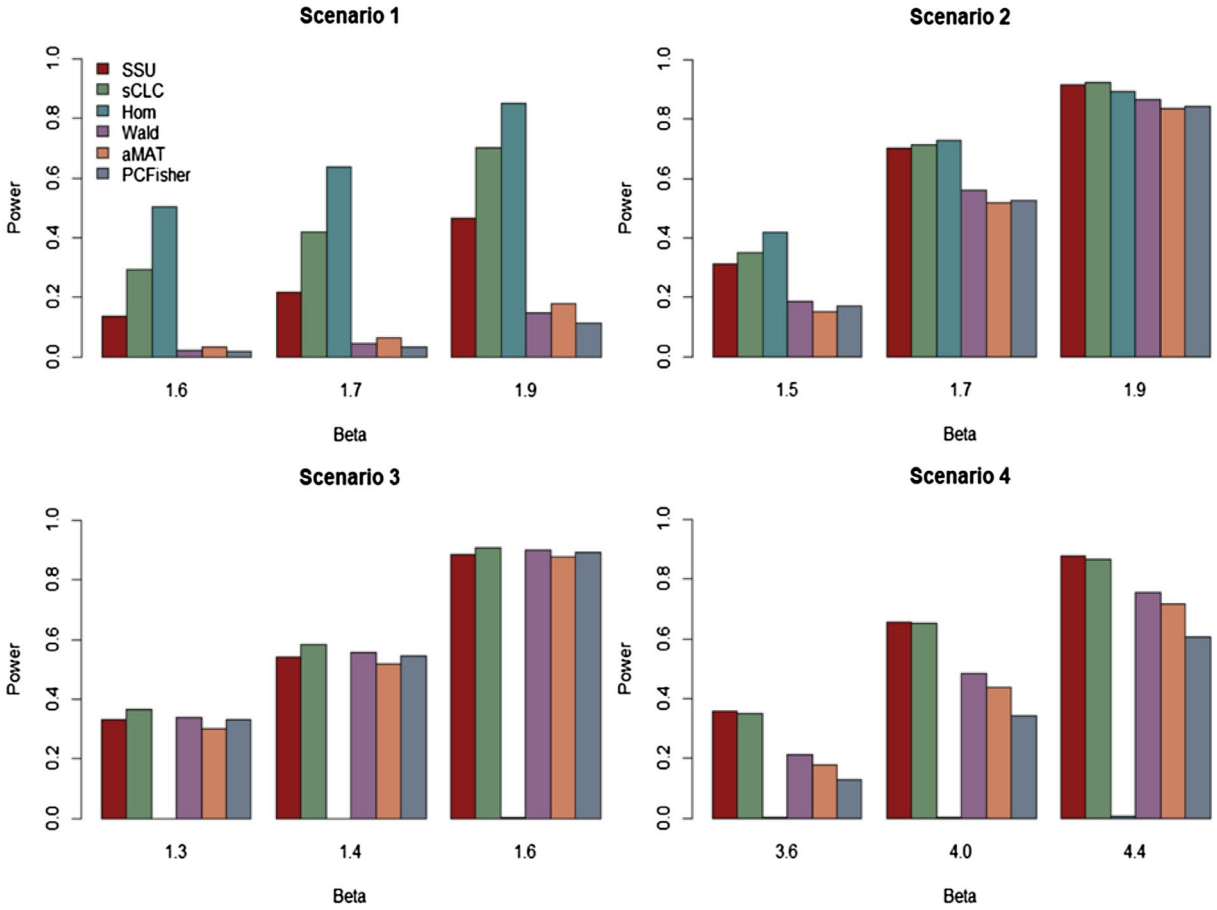

Figure 1. Power comparisons of the six methods, SSU, sCLC, Hom, Wald, aMAT, and PCFisher for the phenotypic correlation structure of the 70 phenotypes at a significant level of 5×10^{-8}.

corresponding EHR-derived phenotype for that individual will be coded as 1, otherwise it will be 0 (1 for cases and 0 for controls). In the XIII category, we only consider phenotypes with more than 200 cases and there are a total of 72 unique phenotypes, such as rheumatoid arthritis (M06.9) and Systemic Lupus Erythematosus (M32.9). Table S4 lists the ICD-10 code, the name of the disease, heritability, and case–control ratio for each of the 72 phenotypes. Since our proposed method is a population-based method and cannot be applied to a mixed population due to population stratification, we analyze 409,672 individuals with the white British ancestry. Similar to Liang et al.[28], we also exclude individuals who are marked as outliers for heterozygosity, and have been identified to have more than ten third-degree relatives or closer, etc. The final dataset includes $N = 322,607$ individuals with $M = 288,647$ common variants across $K = 72$ phenotypes for analyses. All the phenotypes are adjusted by 13 covariates, including age, sex, genotyping array, and the first 10 genetic principal components (PCs).

To apply our method, we first calculate the GWAS summary statistics for the 72 phenotypes based on 288, 647 SNPs. We observed that all of the 72 phenotypes have extremely unbalanced case–control ratios, where the largest case–control ratio is 0.03937 for Gonarthrosis (M17.9) and the smallest case–control ratio is 0.000658 for Lumbar and other intervertebral disk disorders with myelopathy (M51.0). Therefore, we use the saddlepoint approximation (SPA)[36] to calculate the adjusted Z scores. For the jth SNP and kth phenotype $(j = 1, \ldots, M, k = 1, \ldots, K)$, we calculate the score test statistic[37] $S_{kj} = \sum_{i=1}^{N}(Y_{ik} - \overline{Y}_k)G_{ij}$, where $\overline{Y}_k = \sum_{i=1}^{n} Y_{ik}/N$. Y_{ik} denotes the kth phenotype for the ith individual, G_{ij} denotes the jth SNP for the ith individual $(i = 1, \ldots, N)$. The adjusted Z-score is defined as $Z_{kj} = sign(S_{kj})\sqrt{F_{Chi}^{-1}(1 - p_{kj})}$, where $F_{Chi}()$ denotes the cumulative density function of χ_1^2 and p_{kj} is the p-value of S_{kj} obtained using SPA[36]. Based on the adjusted Z-scores, we then apply LDSC to estimate the correlation matrix among phenotypes. We run the single-trait LDSC[21] to estimate the diagonal elements for each phenotype, and the off-diagonal elements are estimated by the cross-trait LDSC[20]. Two phenotypes M79.6 (Enthesopathy of lower limb) and M67.8 (Other specified disorders of synovium and tendon) are excluded in this procedure because the estimators of their heritability are out of bounds. Therefore, there are a total of 70 phenotypes in the simulation studies and real data analysis. The phenotypic correlation matrix only needs to be estimated once for all SNPs. Finally, we apply our proposed sCLC method and the other five methods to test the association between each of 288,647 SNPs and 70 phenotypes, and the commonly used genome-wide significant level $\alpha = 5 \times 10^{-8}$ is considered.

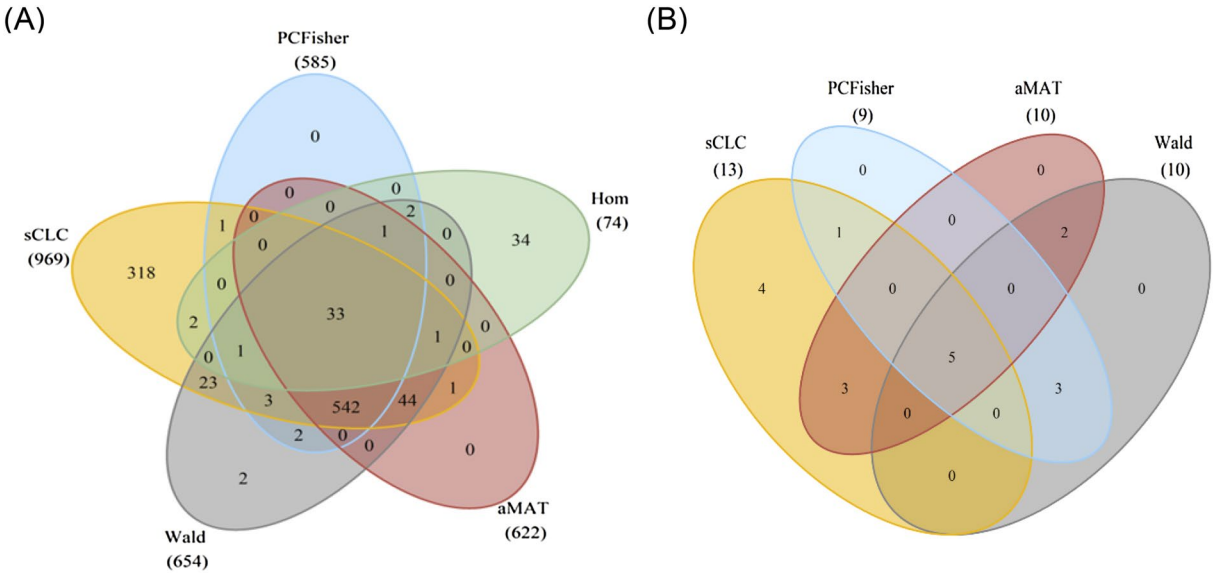

Figure 2. Venn diagram. (**A**) The number of significant SNPs identified by the five methods. (**B**) The number of lead SNPs identified by sCLC, Wald, aMAT, and PCFisher.

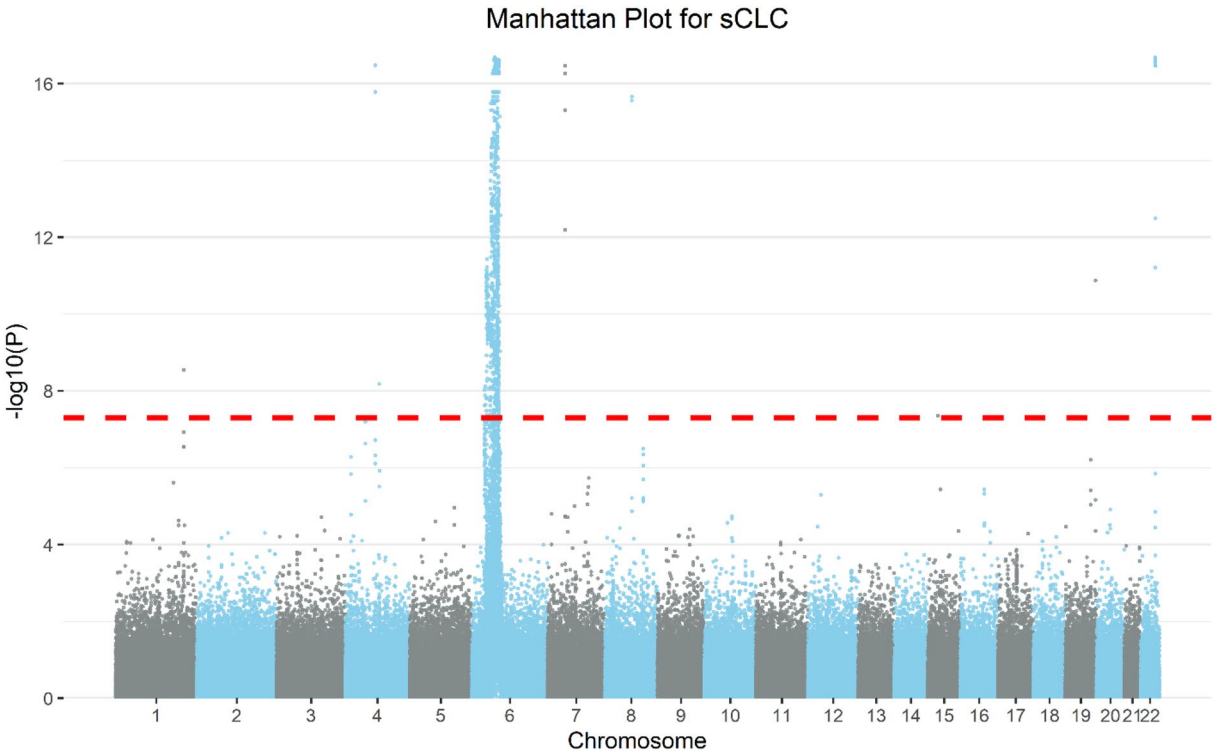

Figure 3. Manhattan Plot from the results of sCLC using multiple phenotypes based on the phenotypes on the UK Biobank XIII category. Each SNP ordered by the genomic position is represented in the x-axis and the association strength with the transformed p-values $-\log_{10}(p)$ is represented in the y-axis.

Among all the six methods, sCLC identifies the largest number of SNPs (969), where Hom identifies 74 SNPs, SSU identifies 872 SNPs, Wald test identifies 654 SNPs, aMAT identifies 622 SNPs, and PCFisher identifies 585 SNPs. Figure 2A shows the Venn Diagram for five methods except for SSU, since SSU cannot control Type I error rates in our simulation studies. There are 33 SNPs identified by all five methods, and 318 SNPs only identified by sCLC. Figure 3 shows the Manhattan plot from the sCLC test results, in which 947 out of 969 SNPs are located in chromosome 6. To evaluate the 969 SNPs identified by sCLC, we first map those SNPs to genes, and we use the commonly used UCSC reference gene file (https://hgdownload-test.gi.ucsc.edu/goldenPath/hg19/bigZips/genes/). Each gene has a position interval. A SNP can be mapped to a gene if its position is within the interval or

20 kb downstream or 20 kb upstream from the interval. These 969 SNPs can be mapped to 235 genes. From the results, we find that 746 out of 969 SNPs can be matched to the genes that have been reported to be associated with the Chapter XIII phenotypes in GWAS catalog. Moreover, among 318 SNPs only identified by sCLC, 229 SNPs can be mapped to the genes that have been reported to be associated with those phenotypes.

However, SNPs within the same LD block are highly correlated and are more likely to be mapped to the same gene. For example, 205 out of 969 identified SNPs are mapped to gene TSBP1-AS1, which is associated with 10 phenotypes in the XIII category; other genes such as NOTCH4, HLA-DRA, and HLA-DRB1 also have many identified SNPs mapped on them. Hence, we are also interested in the independent lead SNPs associated with those phenotypes. We use the Functional Mapping and Annotation (FUMA)[38] platform to obtain independent lead SNPs and distinct risk loci. Here, the independent lead SNPs are defined as $r^2 < 0.1$ and distinct loci are > 250 kb apart. The 969 SNPs identified by sCLC are represented by 13 lead SNPs located in 8 distinct risk loci; the 654 SNPs identified by Wald are represented by 10 lead SNPs located in 6 distinct risk loci; the 622 SNPs identified by aMAT are represented by 10 lead SNPs located in 7 distinct risk loci; and the 585 SNPs identified by PCFisher are represented by 10 lead SNPs located in 6 distinct risk loci. Since the MHC region is excluded by FUMA[38], Hom has no lead SNPs. Figure 2B shows the Venn Diagram of the lead SNPs for sCLC, Wald, aMAT and PCFisher. There are 5 lead SNPs identified by all four methods, and 4 lead SNPs only identified by sCLC. Table 2 shows the details of the summary statistics for all of the 18 independent lead SNPs identified by those four methods. The graying out rows indicate that the SNPs/matched genes have been reported in the GWAS catalog. There are 5 out 13 lead SNPs for sCLC that have not been reported in the GWAS catalog, which may provide us a new insight into the potential genetic factors of the musculoskeletal system and connective tissue phenotypes. Among those 5 SNPs, SNP rs13107325 has the Annotation-Dependent Depletion (CADD) score[39] greater than 20, which means having a high observed probability of a deleterious variant effect. In addition, we compare the p-values of the 13 independent lead SNPs obtained by sCLC with the minimum p-value (MinP) among 70 p-values for testing the association between a SNP and each of the 70 phenotypes. Table S5 shows the comparison results. There are 6 out of 13 SNPs (graying out) with MinP $> 5 \times 10^{-8}$, indicating that these six SNPs have no association with any of the 70 phenotypes by univariate association tests. However, by jointly analyzing the 70 phenotypes, sCLC identified these six SNPs indicating that these 6 SNPs have pleiotropic effects on the phenotypes.

In order to better understand the biological meaning behind 235 mapped genes identified by sCLC, similar to Cao et al.[40], we use DAVID functional annotation software for the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis[41,42]. There are 29 significantly enriched pathways identified by sCLC with FDR < 0.05 and enriched gene count > 2 (Fig. 4). From Fig. 4, we can observe that two related pathways significantly enriched, systemic lupus erythematosus (hsa05322; $FDR = 2.9 \times 10^{-32}$) and rheumatoid arthritis (hsa05323; $FDR = 3.7 \times 10^{-7}$). Especially, there are 32 genes enriched in the systemic lupus erythematosus pathway, including eight genes in HLA-family (*HLA-DMA, HLA-DMB, HLA-DOB, HLA-DQA2, HLA-DQA1, HLA-DRA, HLA-DRB1, HLA-DQB1*), 20 genes in the four core histones (H2A(6): *H2AC6, H2AC13, H2AC14, H2AC15, H2AC16, H2AC17*; H2B(6): *H2BC3, H2BC4, H2BC13, H2BC14, H2BC15, H2BC17*; H3(4): *H3C3, H3C10, H3C11, H3C12*; H4(4): *H4C3, H4C11, H4C12, H4C13*), as well as four genes (*C2, C4B, C4A, TNF*).

Chr	SNP	BP	A1	A2	sCLC P	Wald P	aMAT P	PCFisher P	Mapped gene	Reported trait
1	**rs4846567**	**219,750,717**	**G**	**T**	**2.88E−09**	–	–	–	**ZC3H11B**	**M19.9; M85.8**
4	**rs4148157**	**89,020,934**	**A**	**G**	**1.67E−16**	–	**6.54E−14**	–	**ABCG2**	**M10.9**
4	**rs2231142**	**89,052,323**	**G**	**T**	–	**5.16E−17**	–	**3.96E−16**	**ABCG2**	**M10.9**
4	**rs13107325**	**103,188,709**	**C**	**T**	**6.70E−09**	–	**7.46E−09**	–	**SLC39A8**	**M19.9**
6	rs13212534	25,983,010	A	G	9.47E−09	–	–	–	TRIM38	
6	rs13195040	27,413,924	A	G	–	9.00E−09	1.80E−08	–	ZNF184	
6	**rs13207082**	**27,251,379**	**A**	**G**	**1.08E−10**	–	–	**2.31E−08**	**POM121L2**	**M85.8**
6	rs67340775	28,304,384	A	G	3.78E−12	–	–	–	ZKSCAN3	
6	**rs3117425**	**29,260,431**	**C**	**T**	–	**1.46E−08**	**2.92E−08**	–	**OR14J1**	**M72.9**
6	**rs404240**	**29,523,957**	**A**	**G**	**1.91E−11**	–	–	–	**GABBR1**	**M32.9; M85.8**
7	**rs2598104**	**37,977,249**	**C**	**T**	**5.00E−16**	**1.07E−13**	**2.14E−13**	**5.81E−14**	**EPDR1**	**M72.0; M85.8**
7	**rs2290221**	**37,987,632**	**A**	**G**	–	**5.32E−20**	–	**4.69E−19**	**EPDR1**	**M72.0; M85.8**
7	rs118028828	38,026,155	C	T	5.55E−17	–	2.22E−16	–		
8	rs655028	70,049,047	A	G	2.22E−16	7.08E−16	1.44E−15	4.31E−15		
19	**rs34945782**	**57,678,336**	**C**	**T**	**1.34E−11**	**2.16E−08**	**4.32E−08**	**2.42E−08**	**DUXA**	**M72.0; M85.9**
22	**rs62228062**	**46,381,234**	**A**	**G**	–	**1.74E−35**	–	**2.88E−32**	**WNT7B**	**M85.9**
22	rs28698504	46,403,715	A	G	6.23E−12	1.24E−09	2.48E−09	2.06E−08		
22	**rs9627391**	**46,447,097**	**C**	**T**	**3.27E−13**	**2.50E−12**	**4.99E−12**	**1.50E−11**	**LINC00899**	**M72.0**

Table 2. Summary statistics of the independent lead SNPs identified by sCLC, Wald, aMAT, PCFisher. The bold out rows indicate that the SNPs/mapped genes have been reported in the GWAS Catalog. "–" represents that the SNP is not an independent lead SNP for the corresponding method.

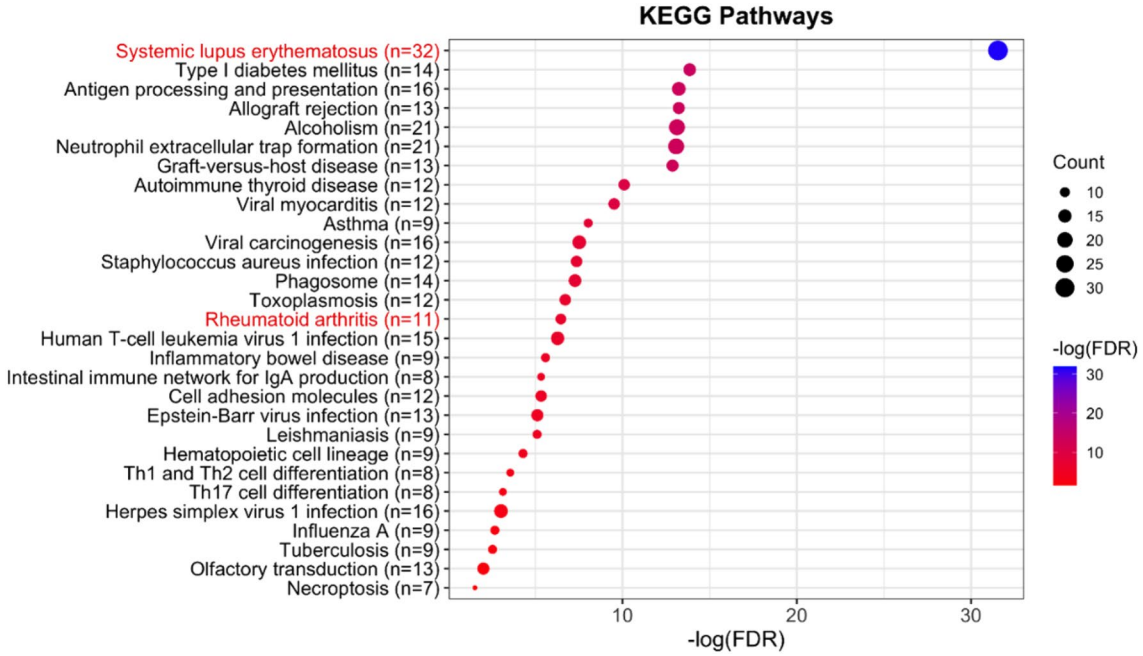

Figure 4. The KEGG pathway enrichment analysis is based on the genes identified by sCLC and the KEGG database. The pathways in red denote the pathways that are related to the diseases of the musculoskeletal system and connective tissue.

For the rheumatoid arthritis pathway, sCLC identifies 104 SNPs mapped to 11 genes that are enriched in this pathway, including *HLA-DMA, HLA-DMB, ATP6V1G2, HLA-DRA, LTB, TNF, HLA-DOB, HLA-DQA2, HLA-DRB1, HLA-DQA1*, and *HLA-DQB1*.

Discussion

In this paper, we propose a multiple-phenotype association test strategy called sCLC which is based on GWAS summary statistics. Through a variety of simulation studies and an application to the UK Biobank XIII category summary statistics, we observed that sCLC is a valid and powerful approach. Specially, sCLC detected some novel signals associated with the musculoskeletal system and connective tissue phenotypes, which provides more evidence to show that those diseases are potentially affected by genetic factors. The sCLC method is also computationally efficient. Since the estimation of the phenotypic correlation matrix $\boldsymbol{R}$ is independent of the association test for each SNP, we only need to estimate $\boldsymbol{R}$ once by using LDSC for all SNPs. In real data analysis with 288,647 SNPs and 70 phenotypes, after estimation of $\boldsymbol{R}$, the running time of sCLC on a computer with 4 Intel Cores @ 3.60 GHz and 16 GB memory is about 4 min 40 s. sCLC as well as many other multiple phenotype association methods, such as the compared methods in this article, test the null hypothesis that a given variant does not contribute to any of the analyzed phenotypes. Therefore, a genetic variant will be identified by these methods even if it is associated with only one phenotype. Hence the identified genetic variants by these methods may not be pleiotropic variants and further analyses are required to interpret the possibility of pleiotropy[43]. This is a limitation of the proposed method in identifying pleiotropic effects. Recently, some methods[43–45] are proposed to evaluate pleiotropic effects. For example, Schaid et al.[43] proposed a new statistical method to evaluate pleiotropy using a sequential testing framework. This approach can determine the number of phenotypes associated with a genetic variant and which phenotypes are associated, while accounting for correlations among the phenotypes. SHAHER[44], a novel framework for analysis of the shared genetic background of correlated phenotypes, can identify genetic factors common for all analyzed phenotypes and specific genetic factors for each phenotype using genetic correlations between phenotypes. PolarMorphism[46] is a summary-statistic-based framework to map and interpret pleiotropic loci in a joint analysis of multiple phenotypes. It identifies horizontally pleiotropic SNPs by converting the trait-specific SNP effect sizes to polar coordinates.

On the other hand, the hierarchical clustering approach in sCLC is applied to cluster multiple phenotypes based on the phenotypic correlation matrix $\boldsymbol{R}$. Therefore, the phenotypes in the same cluster may be affected by non-genetic factors, which may influent the power for disease variant discovery. Instead of using the phenotypic correlation matrix, the genetic correlation matrix among multiple phenotypes[20,21] can also be used in the hierarchical clustering. Furthermore, considering only the phenotypes with a significant non-zero heritability in the estimation of the genetic correlation matrix may also improve the statistical power in the multiple phenotype association studies. Therefore, we would like to consider using the genetic correlation matrix estimated by the LDSC regression[20] or using network-based approaches to cluster phenotypes based on shared genetic architectures in our further work[47].

Data availability

UK Biobank data can be accessed by application through http://www.ukbiobank.ac.uk. UK Biobank has approval by the Research Ethics Committee (REC) under approval number 16/NW/0274. UK Biobank obtained participant's consent for the data to be used for health-related research, and all methods were performed in accordance with the relevant guidelines and regulations.

Received: 17 November 2022; Accepted: 22 February 2023
Published online: 28 February 2023

References

1. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
2. Lutz, S. M., Fingerlin, T. E., Hokanson, J. E. & Lange, C. A general approach to testing for pleiotropy with rare and common variants. *Genet. Epidemiol.* **41**, 163–170 (2017).
3. Pei, G. *et al.* Investigation of multi-trait associations using pathway-based analysis of GWAS summary statistics. *BMC Genomics* **20**, 43–54 (2019).
4. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18**, 117–127 (2017).
5. Kwak, I.-Y. & Pan, W. Gene-and pathway-based association tests for multiple traits with GWAS summary statistics. *Bioinformatics* **33**, 64–71 (2017).
6. Guo, B. & Wu, B. Statistical methods to detect novel genetic variants using publicly available GWAS summary data. *Comput. Biol. Chem.* **74**, 76–79 (2018).
7. 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.* **6**, 1–10 (2016).
8. Deng, Y. & Pan, W. Conditional analysis of multiple quantitative traits based on marginal GWAS summary statistics. *Genet. Epidemiol.* **41**, 427–436 (2017).
9. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* **11**, 407–409 (2014).
10. Liang, X., Sha, Q., Rho, Y. & Zhang, S. A hierarchical clustering method for dimension reduction in joint analysis of multiple phenotypes. *Genet. Epidemiol.* **42**, 344–353 (2018).
11. Jiang, C. & Zeng, Z.-B. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* **140**, 1111–1127 (1995).
12. Stephens, M. A unified framework for association analysis with multiple related phenotypes. *PLoS ONE* **8**, e65245 (2013).
13. Zhu, X. *et al.* Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am. J. Hum. Genet.* **96**, 21–36 (2015).
14. Pan, W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet. Epidemiol.* **33**, 497–507 (2009).
15. Yang, Q. & Wang, Y. Methods for analyzing multivariate phenotypes in genetic association studies. *J. Probab. Stat.* **2012** (2012).
16. Liu, Z. & Lin, X. A geometric perspective on the power of principal component association tests in multiple phenotype studies. *J. Am. Stat. Assoc.* (2019).
17. Wu, C. Multi-trait genome-wide analyses of the brain imaging phenotypes in UK Biobank. *Genetics* **215**, 947–958 (2020).
18. Sha, Q., Wang, Z., Zhang, X. & Zhang, S. A clustering linear combination approach to jointly analyze multiple phenotypes for GWAS. *Bioinformatics* **35**, 1373–1379 (2019).
19. Wang, M., Zhang, S. & Sha, Q. A computationally efficient clustering linear combination approach to jointly analyze multiple phenotypes for GWAS. *PLoS ONE* **17**, e0260911 (2022).
20. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
21. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
22. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).
23. Consortium & G. P,. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56 (2012).
24. Li, X., Zhang, S. & Sha, Q. Joint analysis of multiple phenotypes using a clustering linear combination method based on hierarchical clustering. *Genet. Epidemiol.* **44**, 67–78 (2020).
25. Liu, Y. & Xie, J. Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *J. Am. Stat. Assoc.* **115**, 393–402 (2020).
26. Liu, Y. *et al.* ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am. J. Hum. Genet.* **104**, 410–421 (2019).
27. Guo, B. & Wu, B. Integrate multiple traits to detect novel trait–gene association using GWAS summary data with an adaptive test approach. *Bioinformatics* **35**, 2251–2257 (2019).
28. Liang, X., Cao, X., Sha, Q. & Zhang, S. HCLC-FC: A novel statistical method for phenome-wide association studies. *PLoS ONE* **17**(11), e0276646 (2022).
29. Mosca, M., Tani, C., Vagnani, S., Carli, L. & Bombardieri, S. The diagnosis and classification of undifferentiated connective tissue diseases. *J. Autoimmun.* **48**, 50–52 (2014).
30. Nikolenko, V. *et al.* Morphological signs of connective tissue dysplasia as predictors of frequent post-exercise musculoskeletal disorders. *BMC Musculoskelet. Disord.* **21**, 1–7 (2020).
31. Mosca, M., Neri, R. & Bombardieri, S. Undifferentiated connective tissue diseases (UCTD): A review of the literature and a proposal for preliminary classification criteria. *Clin. Exp. Rheumatol.* **17**, 615–620 (1999).
32. Iudici, M., Cuomo, G., Vettori, S., Avellino, M. & Valentini, G. Quality of life as measured by the short-form 36 (SF-36) questionnaire in patients with early systemic sclerosis and undifferentiated connective tissue disease. *Health Qual. Life Outcomes* **11**, 1–6 (2013).
33. Sudlow, C. *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.* **12**, e1001779 (2015).
34. McGuirl, M. R., Smith, S. P., Sandstede, B. & Ramachandran, S. Detecting shared genetic architecture among multiple phenotypes by hierarchical clustering of gene-level association statistics. *Genetics* **215**, 511–529 (2020).
35. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742-13015-10047-13748 (2015).
36. Daniels, H. E. Saddlepoint approximations in statistics. *Ann. Math. Stat.* 631–650 (1954).
37. Sha, Q., Zhang, Z. & Zhang, S. Joint analysis for genome-wide association studies in family-based designs. *PLoS ONE* **6**, e21957 (2011).
38. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1–11 (2017).

39. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
40. Cao, X., Liang, X., Zhang, S. & Sha, Q. Gene selection by incorporating genetic networks into case-control association studies. *Eur. J. Hum. Genet.* (2022).
41. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
42. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
43. Schaid, D. J. *et al.* Multivariate generalized linear model for genetic pleiotropy. *Biostatistics* **20**, 111–128 (2019).
44. Svishcheva, G. R. *et al.* A novel framework for analysis of the shared genetic background of correlated traits. *Genes* **13**, 1694 (2022).
45. Lee, C. H., Shi, H., Pasaniuc, B., Eskin, E. & Han, B. PLEIO: A method to map and interpret pleiotropic loci with GWAS summary statistics. *Am. J. Hum. Genet.* **108**, 36–48 (2021).
46. von Berg, J., ten Dam, M., van der Laan, S. W. & de Ridder, J. PolarMorphism enables discovery of shared genetic variants across multiple traits from GWAS summary statistics. *Bioinformatics* **38**, i212–i219 (2022).
47. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).

Acknowledgements

Part of this research has been conducted using the UK Biobank Resource under application number 41722 and the NHGRI-EBI GWAS Catalog. X.C. was partially funded by the Michigan Technological University Health Research Institute Fellowship program and the Portage Health Foundation Graduate Assistantship. High-Performance Computing Shared Facility (Superior) at Michigan Technological University was used in obtaining results presented in this publication.

Author contributions

Formal analysis: M.W.; research design: M.W., S.Z., and Q.S.; real data processing: X.C.; visualization: M.W and X.C.; writing original draft: M.W., X.C., and Q.S.; writing review and editing: M.W., S.Z., X.C., and Q.S.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-30415-3.

Correspondence and requests for materials should be addressed to Q.S.

Reprints and permissions information is available at www.nature.com/reprints.

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