PhenoSpD: an atlas of phenotypic correlations and a multiple testing correction for the human phenome

Jie Zheng[1,*], Tom G. Richardson[1], Louise A. C. Millard[1,2], Gibran Hemani[1], Christopher Raistrick, Bjarni Vilhjalmsson[3], Philip Haycock[1], Tom R Gaunt[1,*]

[1]MRC Integrative Epidemiology Unit, University of Bristol, Oakfield House, Bristol, UK; [2]Intelligent Systems Laboratory, University of Bristol, Bristol, UK; [3]Århus Center for Bioinformatics BIRC, Aarhus University

*To whom correspondence should be addressed.
Contact: jie.zheng@bristol.ac.uk, tom.gaunt@bristol.ac.uk

Abstract

Summary: Identifying phenotypic correlations between complex traits and diseases can provide useful etiological insights. Restricted access to individual-level phenotype data makes it difficult to estimate large-scale phenotypic correlation across the human phenome. Here, we present a novel method, PhenoSpD, to estimate phenotypic correlations using genome-wide association study (GWAS) summary statistics from the same sample and utilizes the correlations to inform correction of multiple testing for human GWAS studies. In a case study using GWAS summary results, PhenoSpD suggested 324.4 independent tests among 452 metabolites, which is close to the 296 independent tests estimated using true phenotypic correlation. We then estimated 120,713 pair-wise phenotypic correlations among 24 categories of human traits and diseases (total 862 traits) and further corrected multiple testing for these traits using PhenoSpD. The atlas of phenotypic correlations provides novel insights into the relationships between traits, while the PhenoSpD multiple testing correction function provides a simple and conservative way to reduce dimensionality for GWAS of complex molecular traits.

Availability: R codes and Documentation for PhenoSpD V1.0.0 is available online (https://github.com/MRCIEU/PhenoSpD).

1 Introduction

Phenotypic correlations between complex human traits and diseases provide useful etiological insights. For GWAS meta-analysis, a lack of individual-level phenotype data makes it difficult to estimate the phenotypic correlation across human traits and diseases. Here we consider two methods that estimate phenotypic correlations using GWAS summary statistics: metaCCA (Cichonska et al., 2016) and bivariate LD score regression (Bulik-Sullivan et al., 2015). The metaCCA framework estimates phenotypic correlation between two traits based on a Pearson correlation between two univariate regression coefficients (betas) across a set of genetic variants; The bivariate LD score regression approach esti-mates the phenotypic correlation amongst the overlapping samples of two GWAS.

The recently developed MR-Base (Hemani et al.2016) and LD Hub (Zheng et al., 2017) tools include harmonized GWAS summary-level results. This provides an opportunity to estimate the phenotypic correlation structure across a wide array of studies, which is of particular value for high-dimensional, complex molecular traits, such as metabolites, that are potentially highly correlated. Bonferroni correction would markedly overcorrect for the inflated false-positive rate in such correlated datasets, resulting in a reduction in power. An appropriate method to correct for multiple testing among human traits and diseases based on the spectral decomposition of matrices (SpD) (Nyholt, 2004; Li and Ji, 2005) is considered in this study.

2 Methods

2.1 Overview of PhenoSpD

Figure 1 demonstrates the key steps of the proposed method, PhenoSpD: step (1) harmonise GWAS summary results from the same sample; step (2) apply the harmonized GWAS results to metaCCA or LD score regression to estimate the phenotypic correlation matrix of the traits; step (3) apply the phenotypic correlation matrix to the SpD approach and estimate the number of independent variables among the traits.

2.2 Validation of Phenotypic correlation estimation

Firstly, we compared the difference between metaCCA and LD score regression for estimation of phenotypic correlation. We then simulated how sample overlap between two GWASs influence the phenotypic correlation estimation using both methods. In general, only GWASs run on the same set of individuals (or with majority of the individuals overlapped) are suitable for estimating phenotypic correlation using GWAS summary results (Text S1).

2.3 Estimating the phenotypic correlations

Within our database containing 1094 human traits, we selected 862 unique traits (Table S2, Text S2) using the criteria listed in Table S3. We then separated the 862 GWASs into 24 categories (Table S5), e.g. all traits in GIANT consortium were put into the same group to fit the assumption of one sample setting. We then applied metaCCA to each group to estimate the phenotypic correlation matrix of each group of traits. From these 862 traits, we further selected all available traits in LD Hub (219 traits) to conduct the bivariate LD score regression analyses within each of the 24 categories (Table S4) to estimate the phenotypic correlation of these traits.

2.4 Multiple testing correction for human traits

We applied the SpD approach to correct for multiple testing among human traits and diseases. We implemented the R code of the well-known method, SNPSpD (Nyholt, 2004; Li and Ji, 2005), to estimate the number of independent traits using the phenotypic correlation matrix as input (Fig. 1). For each of the 24 categories, we estimate the number of independent tests using the SpD function.

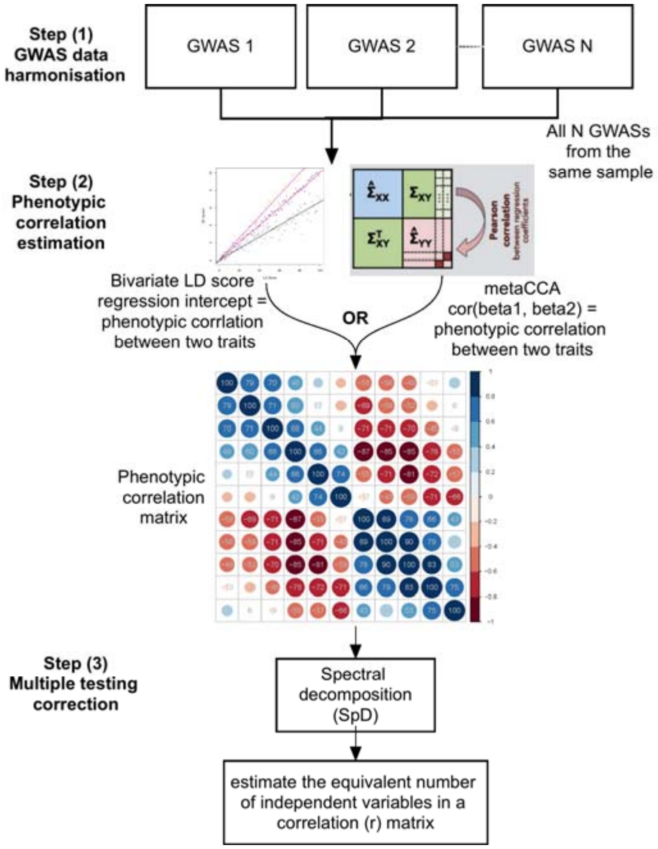


Fig. 1. Flowchart of PhenoSpD

3 Results

3.1 The phenotypic correlations of human phenome

Based on the validation described in Text S1, metaCCA can be applied to almost all GWASs (sample size>300 and number of SNPs>1000). However, 1) the genetic effect of SNPs may bias the phenotypic correlation estimation; 2) it only provides the central estimate of the phenotypic correlation; 3) it is difficult to quantify the effect of sample overlap. LD score regression provides the pair-wise phenotypic correlation with standard errors and it deals with sample overlap automatically (when there is no sample overlap between two GWASs, the phenotypic correlation will be zero). However, it can only be applied to GWAS with large sample size (e.g. N > 5,000) and good SNP coverage (e.g. number of SNPs > 200,000) to fit its assumptions.

We then estimated 120,713 pair-wise phenotypic correlations among 24 categories of human traits and diseases (862 traits) using metaCCA (Table S6 to Table S28). In addition, we also estimate 47,961 pair-wise corre-lations (219 traits) estimated by LD score regression (Table S29).

3.2 Multiple testing correction of human phenome

In a case study, PhenoSpD using GWAs results suggested 324.4 independent tests among 452 metabolites from Shin et al., which is close to 296 independent tests estimated using real phenotypic correlation. Table 1 shows the number of independent traits for three high-dimensional, complex molecular datasets. Table S5 lists the number of tests for the 862 unique GWAS traits within MR-Base.

Table 1. Summary of number of independent traits for three complex molecular networks.

First author	Category	$N_{\_traits}$	$N_{\_SNPs}$	$N_{\_indep}$
Kettunen *et al.*	Metabolites	123	9826292	44.9
Shin *et al.*	Metabolites	452	2482345	324.4
Roederer *et al.*	Immuno-phenotypes	151	1585187	94.2

Note: $N_{\_traits}$ refers to number of traits in each molecular network; $N_{\_SNPs}$ refers to number of SNPs in each network; $N_{\_indep}$ refers to number of independent tests in each network.

4 Discussion

In this study, we present a novel method which allows phenotypic correlation estimation and multiple testing correction for human phenome using GWAS summary statistics. We illustrate the application of PhenoSpD by estimating the phenotypic correlation structure of 24 datasets comprising a total of 862 traits, including the correlation structure of 123 metabolites from Kettunen's study for the very first time (Kettunen et al, 2016). These results showcase the ability of PhenoSpD to estimate an appropriate multiple testing correction for large-scale post-GWAS data mining tools such as MR-Base and LD Hub.

4.1 Advantages and limitations of PhenoSpD

There are two key advantages to the PhenoSpD approach compared with existing methods. Firstly, PhenoSpD builds on the phenotypic correlation function of metaCCA and LD score regression using only GWAS summary results. Secondly, PhenoSpD provides a simple way of correcting multiple testing for the human phenome. Such correction is stringent, but more appropriate than Bonferroni correction, which is particularly valuable for GWAS of complex molecular traits. It can also be used as an indicative threshold for post-GWAS data mining tools such as MR-Base and LD Hub. In addition, based on our simulations, PhenoSpD can only be used for GWAS results from one population sample, which is a general limitation of estimating phenotypic correlation using GWAS results.

Funding

This work was supported by the Medical Research Council (MC_UU_12013/4 and MC_UU_12013/8). This work was in part supported by Cancer Research UK programme grant number C18281/A19169 (the Integrative Cancer Epidemiology Programme). P.H. is a Cancer Research UK Population Research Fellow (grant number C52724/A20138).

Conflict of Interest: none declared.

References

Bulik-Sullivan. et al. (2015b) An atlas of genetic correlations across human diseases and traits. Nat. Genet., 47, 1236–1241.

Cichonska. et al. (2016) metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. Bioinformatics 32 (13): 1981-1989.

Gibran Hemani, et al. (2016) MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. bioRxiv. doi: https://doi.org/10.1101/078972

Kettunen et al. (2016) Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. Nat Commun. 7:11122.

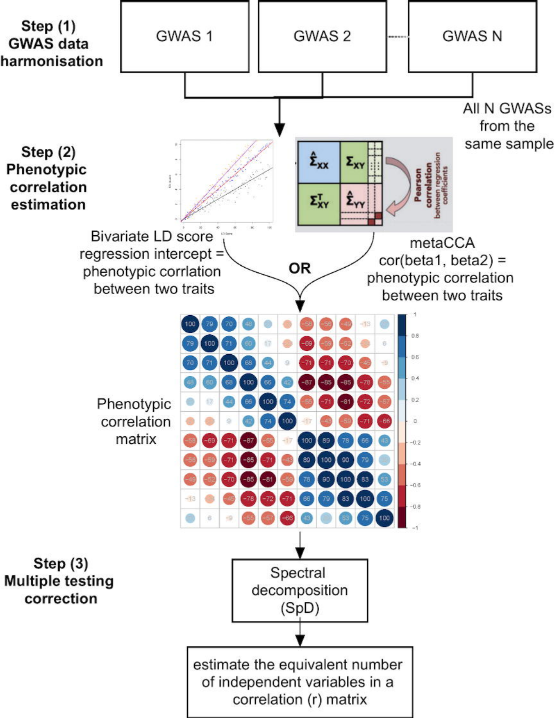
Nyholt DR. (2004) A simple correction for multiple testing for SNPs in linkage disequilibrium with each other. Am J Hum Genet 74(4):765-769.

Li J, Ji L. (2005) Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity 95:221-227

Roederer M et al. (2015) The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis. Cell. 161(2):387-403.

Shin SY et al. (2014) An atlas of genetic influences on human blood metabolites. Nat Genet. 46(6):543-50.

Zheng et al. (2017) LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics. 33 (2): 272-279.


Step (1)
GWAS data harmonisation
GWAS 1
GWAS 2
GWAS N
All N GWASs from the same sample
Step (2)
Phenotypic correlation estimation
$\hat{\Sigma}_{XX}$
Σ_{XY}
Σ_{XY}^T
$\hat{\Sigma}_{YY}$
Pearson correlation between regression coefficients
Bivariate LD score regression intercept = phenotypic corrlation between two traits
OR
metaCCA cor(beta1, beta2) = phenotypic correlation between two traits
Phenotypic correlation matrix
100 79 70 48 -58 -56 -49 -13
79 100 71 60 17 -69 -59 -52 6
70 71 100 68 44 9 -71 -71 -70 -45 -9
48 60 68 100 66 42 -87 -85 -85 -78 -55
17 44 66 100 74 -55 -71 -81 -72 -57
9 42 74 100 -17 -43 -59 -71 -66
-58 -69 -71 -87 -55 -17 100 89 78 66 43
-56 -59 -71 -85 -71 -43 89 100 90 79
-49 -52 -70 -85 -81 -59 78 90 100 83 53
-13 -45 -78 -72 -71 66 79 83 100 75
6 -9 -55 -57 -66 43 53 75 100
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
Step (3)
Multiple testing correction
Spectral decomposition (SpD)
estimate the equivalent number of independent variables in a correlation (r) matrix