의학박사 학위논문

Development of a statistical model to detect pleiotropic loci shared by multiple traits

여러 형질이 공유하는 다면발현 유전자 탐색을 위한 통계모델의 개발

2021 년 8 월

서울대학교 대학원

의과학과 의과학전공

이 현 규

Development of a statistical model to detect pleiotropic loci shared by multiple traits

Advisor: Buhm Han

Submitting a Ph.D. Dissertation of Medicine

April 2021

Graduate School of Biomedical Sciences
Seoul National University
Biomedical Science Major

Cue Hyunkyu Lee

Confirming the Ph.D. Dissertation written by
Cue Hyunkyu Lee
June 2021

Chair _____

Vice-Chair _____

Examiner _____

Examiner _____

Examiner _____

Abstract

Development of a statistical model to detect pleiotropic loci shared by multiple traits

Cue Hyunkyu Lee

The Department of Biomedical Science
The Graduate School of Medicine
Seoul National University

Introduction: GWAS have been used widely for mapping disease-associated genetic variants. Some of these variants exhibited pleiotropic effects in which a locus affects multiple traits simultaneously. Detecting and interpreting pleiotropic loci provides important information to understand the genetic structure shared between diseases and complex traits. A common approach to detecting pleiotropic loci is to perform a meta-analysis with multi-trait GWAS summary statistics. However, existing meta-analysis methods do not model complex genetic structures such as genetic correlations and heritability. In addition, these multi-trait analyses are often

difficult to interpret the analysis results due to the differences in units or scales in phenotypes across traits.

Method: In this paper, I propose PLEIO, a summary statistics-based framework that can map and interpret pleiotropic loci by jointly analyzing diseases and complex traits. The method maximizes the performance of the association test by using a novel statistical model that comprehensively describes the genetic correlations and heritabilities of the traits. PLEIO uses standardized metrics to account for differences in phenotypic units and scales; This generalized model can seamlessly combine any sets of traits. To reduce the computation time for the multi-trait analysis, I used an optimization process using novel mathematical techniques such as importance sampling and eigenvalue decomposition. In addition, PLEIO provides an interpretation and visualization tool that supports downstream analysis of the identified loci.

Results: To verify the performance of PLEIO, I carried out extensive simulations and real data analysis. Simulations, assuming various genetic correlations and heritabilities structures, have confirmed that PLEIO has good control over false-positive rates and outperforms other multi-trait analysis methods. In the real data analysis, I applied PLEIO to 18 traits related to cardiovascular disease and detected 13 novel (newly identified) pleiotropic loci showing four different association patterns. In terms of computational efficiency, the real data analysis that combines 18 traits used less than 4 hours per one CPU unit to test 1,777,411 association tests.

Conclusion: PLEIO is a multi-trait analysis framework, which uses genetic structure between traits to detect pleiotropic loci. The statistical model implemented in PLEIO uses a generalized model that includes the assumptions used by the existing models. The software can be downloaded for free from the following Github webpage: https://github.com/cuelee/pleio.

Keywords: multi-trait analysis, pleiotropy, association test, heritability, genetic correlation, meta-analysis, GWAS, variance component

Student number: 2019-33070

* This work has been published in the American Journal of Human Genetics. (Lee, C. H., Shi, H., Pasaniuc, B., Eskin, E., & Han, B. (2021). “PLEIO: a method to map and interpret pleiotropic loci with GWAS summary statistics.” Am J Hum Genet, 108(1), 36-48. doi:10.1016/j.ajhg.2020.11.017)

Contents

List of Tables

List of Figures

List of Abbreviations

< Sorted in A-Z order >

Age_Smo (phenotype), Age at smoking

ASSET (software), association analysis based on subsets

BMI, body mass index

CAD (phenotype), coronary artery disease

CARDIo+C4D consortium, Coronary artery disease genome-wide replication and meta-analysis plus the coronary artery disease consortium

CDF, cumulative density function

CFTR, cystic fibrosis transmembrane conductance regulator

cIQ (phenotype), childhood IQ

EMMA, Efficient Mixed-Model Association

FPR, false-positive rate

GWAS, genome-wide association studies

HbA1C (phenotype), hemoglobin

LD, linkage disequilibrium

LDL (phenotype), low-density lipoprotein

LDSC (software), linkage disequilibrium score regression

LRT, log-likelihood ratio test

MAF, population minor allele frequency

MAGIC consortium, the meta-analyses of glucose and insulin-related traits consortium

MLE, maximum likelihood estimate

MTAG (software), multi-trait analysis of GWAS

PDF, probability density function

PLEIO (software), Pleiotropic Locus Exploration and Interpretation using Optimal test

RE2C (software), random effect 2 complement

SNP, single nucleotide polymorphism

TG (phenotype), triglyceride

WHR (phenotype), waist-hip-ratio

Chapter 1. Introduction

1.1 Study background

The thesis contains contents that require a high level of background in bioinformatics and statistics. This section is intended to provide background information that helps readers, especially non-majors. Throughout this chapter, I avoided the use of math equations and complicated jargon.

1.1.1 Mendelian and complex disorders

Genetic diseases are health problems caused by abnormalities in the genome, classified as Mendelian disorders or complex disorders. Mendelian (monogenic) disease is caused by a single mutated gene. This type of diseases includes Cystic fibrosis, Sickle cell disease, and Hemophilia diseases. For example, Cystic fibrosis is a recessive genetic disorder caused by mutations in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene[1]. Complex disease (or multifactorial disease) is the parlance of genomics, which denotes that a disease is not a simple Mendelian single-gene disorder but is caused by a combination of many genes and significant environmental contributions. One of the good examples of complex diseases is cardiovascular disease. The risk of cardiovascular disease can be increased by genetic factors[2], such as an ancestor's medical history[3], and by environmental factors, such as long-term consumption of salty foods[4]. Similarly, some human traits such as height and body mass index (BMI) are multifactorial, of which phenotypes are determined by many genetic and environmental factors[5, 6].

1.1.2 Genetic liability-threshold model

The liability-threshold model is a threshold model that describes categorical outcomes (usually binary) with liability scores obtained by summing over many variables. The model assumes that the observed outcome is determined by whether the latent score (l) is less than or greater than the threshold value (z). In a genetic context, the liability score can be described as the sum of two components: genetic (g) and environmental (e), and thus $l = g + e$, and the threshold z defines the limit by which the disease is determined by of genetic and environmental factors. For diseases (whose phenotypic outcomes are binary or categorical), the threshold can be estimated from the population prevalence of the disease (which is typically low). The threshold is defined relative to the population and environment, so the liability score is generally considered as a $N(0,1)$ normally distributed random variable.

1.1.3 Genome-wide association studies (GWAS)

GWAS is a type of research design that identifies genetic variants associated with a trait by performing association tests at the whole genome level using the genotypes and phenotypes generated from many individuals in the general population. To date, GWAS has been performed on a large scale for several complex diseases (or traits) using single nucleotide polymorphisms (SNPs) as independent variables. Here, each SNP represents a single genetic variation that has a modest fraction of mutant alleles in the population (e.g., 1 % or more). Since a single nucleotide can be one of adenine (A), thymine (T), guanine (G), and cytosine (C), a SNP can be one of bi-, tri-, or tetra-allelic. Usually, however, SNPs are considered bi-allelic. Alternatively, GWAS

can be performed using genetic variations other than SNPs (e.g., deletions, insertions, copy number variations, CNVs).

The traits used in GWAS can be broadly divided into two categories (binary and quantitative), and each type uses a different method to estimate the magnitude of individual SNP associations. For quantitative traits (e.g., height, BMI, fasting glucose concentration), we usually use a linear regression model that regresses the standardized phenotypes (continuous) on standardized genotypes. The term 'standardized' here means that the phenotypes (or genotypes) are normalized to follow the standard distribution, $N(0,1)$. For binary traits (e.g., cardiovascular diseases or type 2 diabetes), we can use a logistic regression model that regresses the phenotypes (0 or 1) on genotypes. To perform the above regression analysis, we transform each genotype of an individual into a dosage (0/1/2) representing the frequency of the minor allele of the individual.

Since 2005, the first successful GWAS study[7], to the present, GWAS have identified more than 55,000 significant ($p-value < 5 \times 10^{-8}$; genome wide significance threshold, 2021-06-15) genome-wide associations between genetic variations and common diseases or traits collected from 5106 GWAS publications (**Figure 1**)[8]. These associations have led to many important scientific discoveries: understanding disease mechanisms by identifying novel associated loci causing the disease, identifying therapeutic targets of diseases, and developing methods for diagnosing and predicting prognosis[9].

Figure 1 GWAS diagram from the NHGRI-EBI catalog In this diagram, associations with p-values less than 5×10^{-8} are shown and colored according to trait categories.

Despite these great successes, GWAS have several limitations[9]. A representative example is the problem of missing heritability, a phenomenon in which associations identified by GWAS explain only a small fraction of the heritability of complex traits (e.g., height)[10]. This may be because GWAS does not study all types of genetic variations that affect complex traits. For example, most GWAS does not include rare variants (or ultra-rare variants) and the effects due to epistasis[9]. In many cases, GWAS do not directly identify disease- or trait-causal variants. Instead, they detect tag-SNPs correlated with nearby causal variants through the genetic structure called linkage disequilibrium (LD)[11]. Some critics argue that GWAS associations may be spurious associations caused by the cryptic relatedness between individuals[12].

1.2 Purpose of research

GWAS (Genome-wide association studies) have discovered many genetic variants with pleiotropic effects that affect several traits simultaneously[13, 14]. For example, the GWAS catalog, a database that summarizes the associated variants identified by GWAS performed to date, contains several pleiotropic variants for which pleiotropic effects have been firmly established (e.g., hypertension and myocardial infarction)[15]. Recently, Watanabe et al. conducted a study to examine the GWAS catalog data collected up to 2019 to map pleiotropic variants' position and interpret their genetic structures[14]. In this study, Watanabe et al. found that a large part of the human genome is related to pleiotropy (**Figure 2**)[14].

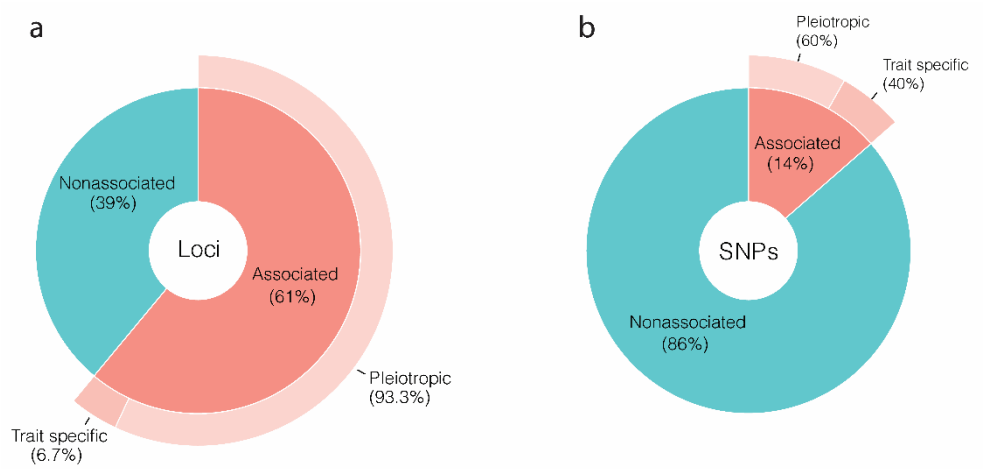


Figure 2 Proportion of pleiotropic trait-associated loci and SNP. a. Each value is based on the summed length of the associated loci, and each associated locus represents a genomic segment correlated by linkage disequilibrium (LD). Here, each locus may contain more than one trait-associated SNP. The definition of an associated locus can be found elsewhere[14] b. Each percentage value uses the number of SNPs in the GWAS included in the analysis as the denominator.

Identifying pleiotropic loci is vital as they can understand the physiological mechanism of complex diseases or develop common therapeutic targets. Most of the resulting summary statistics of the GWAS performed to date are publicly available, which can be used to detect pleiotropic loci. Existing methods for detecting pleiotropic loci are based on a meta-analysis[16-18], trait-specific effect size estimation[19], or Bayesian approaches. Meta-analysis-based methods are suitable for variant mapping because they provide one p-value for each variant. However, the pooled statistics and p-values alone are insufficient to determine how much a gene is associated with each trait. In other words, there are limits to interpreting the results. The trait-specific methods are advantageous for interpretation and genetic risk prediction because they provide updated effect sizes and p-values for each trait and variant. However, variant mapping is difficult for trait-specific methods as additional multiple testing corrections may be required to obtain per variant p-values. In this thesis, I developed a meta-analysis-based method to detect pleiotropic loci.

There are several problems with the strategy of naïve application of the existing meta-analysis method to multi-trait analysis. First, existing methods do not adequately model the genetic structure of diseases and complex traits. The problem can be ameliorated by explicitly modeling the genetic correlation and heritabilities and providing information about the magnitude and direction of effect sizes across traits. Second, conventional methods use the assumption that the phenotypic unit and scale are the same. In a multi-trait analysis, units can differ between quantitative traits, and the definitions of the effect size can vary between binary and quantitative

traits. Currently, most meta-analysis methods ignore these differences and use the observed effect size estimates as input, so multi-trait analysis using the existing meta-analysis methods does not provide optimal results. For the same reason, the interpretation of the analysis using a forest plot or m-value may not provide optimal results. Third, there may be environmental correlations between traits resulting from using the same sample in multiple GWAS. Without systemic estimation and correction for environmental correlations, the use of meta-analysis can inflate false positives.

In this study, I propose a multi-trait method PLEIO (Pleiotropic Locus Exploration and Interpretation using Optimal test), which maps and interprets pleiotropic loci (**Figure 3**). PLEIO uses GWAS summary statistics as input. The multi-trait analysis begins with estimating the genetic correlations, heritabilities, and environmental correlations using whole-genome GWAS summary statistics. Then, it transforms the observed effect size estimates into standardized estimates. For quantitative traits, the standardization makes the phenotypes and genotypes follow a standard normal distribution. For binary traits, standardization means converting per sample genetic contributions into a liability. This process is necessary to analyze diseases and complex traits with different units and compare the magnitude of their effect sizes. PLEIO uses a variance component model and assumes genetic effects as a random variable. The model tests the non-zero genetic variance component where the covariance matrix is proportional to the cross-trait genetic covariance matrix. The statistical model can account for genetic correlations and heritabilities to maximize statistical power and control the false positive rate by taking environmental

correlations into account. To increase the computational efficiency in maximum likelihood estimation, PLEIO uses an optimization technique that applies spectral decomposition to the covariance matrix of the linearly transformed effect sizes. While using the proposed variance component model, I discovered bias in the p-values in multi-trait analysis using a small number of traits due to the small sample problem. I addressed this problem by implementing a novel importance sampling method that accurately estimates the p-value.

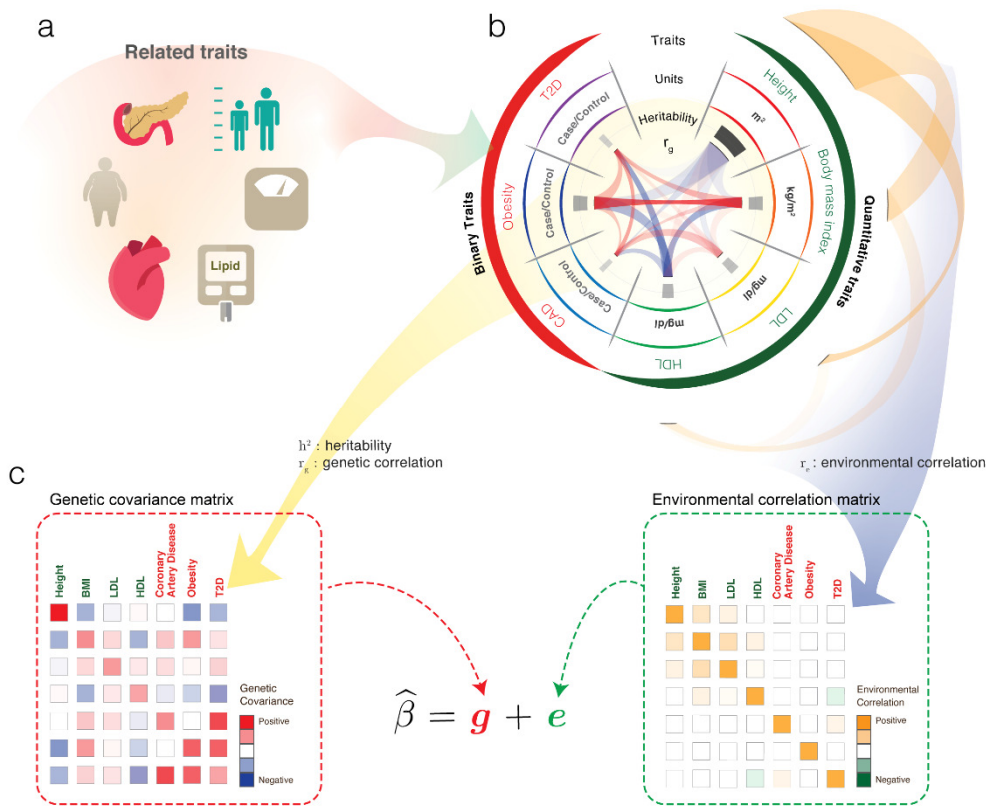


Figure 3. Overview of the PLEIO framework. a. One collects summary statistics of genetically correlated traits. b. One estimates genetic correlations, heritabilities, and environmental correlations across the traits using LDSC. c. PLEIO performs association analysis by modeling the effect sizes as the sum of genetic and environmental effects.

I have validated the power of PLEIO to detect pleiotropic loci through extensive simulations and real data analysis. The simulations assumed different cross-trait genetic correlation structures and compared the performance of several multi-trait methods. PLEIO outperformed almost all competitors in various simulation settings. The results showed that PLEIO, unlike other methods, can adapt flexibly to changes in genetic structures. Next, I collected summary statistics of 18 GWAS related to cardiovascular disease and performed real-data analysis. Through this joint analysis of 18 traits, PLEIO discovered 13 novel pleiotropic loci. I have divided these 13 pleiotropic loci into four groups according to their association patterns, where each group may represent a distinct pathway. To visualize the association patterns of each pleiotropic locus, I used the visualization tool, "pleiotropicPlot" implemented in PLEIO. The software is available to the research community on the GitHub page: https://github.com/cuelee/pleio.

Chapter 2. Material and method

2.1 The whole process of PLEIO analysis

Below, I describe the framework, PLEIO (Pleiotropic Locus Exploration and Interpretation using Optimal test). To jointly analyze multiple traits, the user needs to select the Q traits and collect T GWAS summary statistics. Each of T summary statistics is a SNP association test result of a trait in Q traits, and the traits can be any combination of binary and quantitative phenotypes. One can collect more than one GWAS for one trait so that T can be greater than Q. PLEIO takes M common SNPs shared by T summary statistics as input. The risk and reference alleles of each SNP should be matched across all T summary statistics. Let $\hat{\boldsymbol{\beta}}_{\boldsymbol{it}}$ be the observed effect size of the ith SNP and tth trait, $\mathrm{SE}[\hat{\beta}_{it}]$ be the corresponding standard error, and N_t is the number of the sample size of tth trait. PLEIO uses the T summary statistics as input and performs five analysis steps described below.

2.1.1 Step 1: Estimation of correlation matrices

PLEIO assumes that the correlation of GWAS marginal effect sizes is the summation of the correlation due to causal genetic effects and the correlation due to environmental effects. Here, each marginal effect size measures additive genetic effects. Let $\boldsymbol{C_g}$ is a $T \times T$ matrix of the genetic correlation matrix, $\boldsymbol{C_e}$ is a $T \times T$ matrix of the environmental correlation, and $\boldsymbol{h^2}$ is a $T \times 1$ vector of narrow sense heritabilities. It is straightforward to obtain $\boldsymbol{C_g}$ and $\boldsymbol{h^2}$ by applying the linkage-disequilibrium score regression (LDSC) to a pair of studies [19, 20]. For $\boldsymbol{C_g}$, we use

the "Genetic Covariance" value in LDSC analysis output[21], and for $\boldsymbol{C_e}$, we use "Intercept of Genetic Covariance" as suggested by multi-trait analysis of GWAS (MTAG)[19].

Additionally, I suggest another method for estimating $\mathbf{C_e}$. The proposed method is a two-step procedure. First, a pair of traits are combined using the fixed-effects meta-analysis method based on the inverse variance of the effect size. Then, I apply the single trait LDSC method to this pooled summary statistics, which gives a LDSC intercept. Let this LDSC intercept be α_{meta}, the environmental correlation (ρ_e) becomes

$$\rho_e \approx \frac{N_j + N_k}{2\sqrt{N_j N_k}}(\alpha_{meta} - 1) \qquad \text{Equation 1}$$

where N_j and N_k are the sample sizes of the two studies. I found that the two approaches described above give similar estimates for $\mathbf{C_e}$. For details, see **Estimation of environmental correlations using LDSC**.

2.1.2 Step 2: Standardization of the input statistics

In the collection of T summary statistics (or traits), the scale of observed effect sizes can be heterogeneous. Instead of using the observed (reported) effect sizes, PLEIO uses the effect sizes in a standardized metric derived from each summary statistics. The standardized effect size of SNP i for trait t can be shown as follows:

$$\hat{\eta}_{it} = \frac{\sqrt{\delta_t}\dfrac{\hat{\beta}_{it}}{SE\left[\hat{\beta}_{it}\right]}}{\sqrt{N_t + \delta_t\theta_t\left[\dfrac{\hat{\beta}_{it}}{SE\left[\hat{\beta}_{it}\right]}\right]^2}}, \text{and } \mathrm{SE}[\hat{\eta}_{it}] = \frac{SE\left[\hat{\beta}_{it}\right]}{\hat{\beta}_{it}}\hat{\eta}_{it}.$$

Equation 2

δ_t is a scaling factor that is 1 for quantitative trait and $\frac{K_t^2(1-K_t)^2}{P_t(1-P_t)} \cdot \frac{1}{[\psi(\phi^{-1}(1-K_t))]^2}$ for a binary trait, where K_t refers to the disease prevalence, $P_t = (N_t|y=1)/N_t$ refers to the sample prevalence, $\boldsymbol{\psi}$ refers to the probability density function (PDF) of the standard normal distribution, and ϕ^{-1} refers to the inverse of the cumulative density function (CDF) of the standard normal distribution. θ_t is a scaling factor that is 0 for quantitative trait and $\left(i_t \times \frac{P_t-K_t}{1-K_t}\right)\left(t - i_t \times \frac{P_t-K_t}{1-K_t}\right)$ for a binary trait, where $i_t = \frac{\psi\left(\phi^{-1}(1-K_t)\right)}{K_t}$ refers to the mean liability of cases, and $t = \phi^{-1}(1-K_t)$ refers to the liability threshold for cases. For quantitative traits, $\hat{\eta}_{it} = \frac{\hat{\beta}_{it}}{SE\left[\hat{\beta}_{it}\right]} \cdot \frac{1}{\sqrt{N_t}}$, which corresponds to the regression coefficient of the simple linear regression model using the standardized phenotypes and the standardized genotypes. For binary traits, $\hat{\eta}_{it}$ is the standardized effect size with a liability scale derived from a linear model, assuming a non-randomly ascertained case-control study (by setting the phenotypes 1 and 0). The use of the two scaling factors in **Equation 2** was suggested by S.H Lee et al. For binary traits, the observed statistics are often obtained from logistic regression model rather than linear regression model, but it has been customary to assume that the statistics were obtained from the simple linear regression model[20]. The accuracy of the proposed scaling function was verified by performing extensive simulations with different $\boldsymbol{K_t}$ and $\boldsymbol{P_t}$ (**Figure 4**). The use of $\hat{\eta}_{it}$ is useful for

downstream analysis as it is convenient to interpret the pleiotropic association pattern across traits. Note that the proposed standardized effect size is independent of phenotypic and genotypic units.

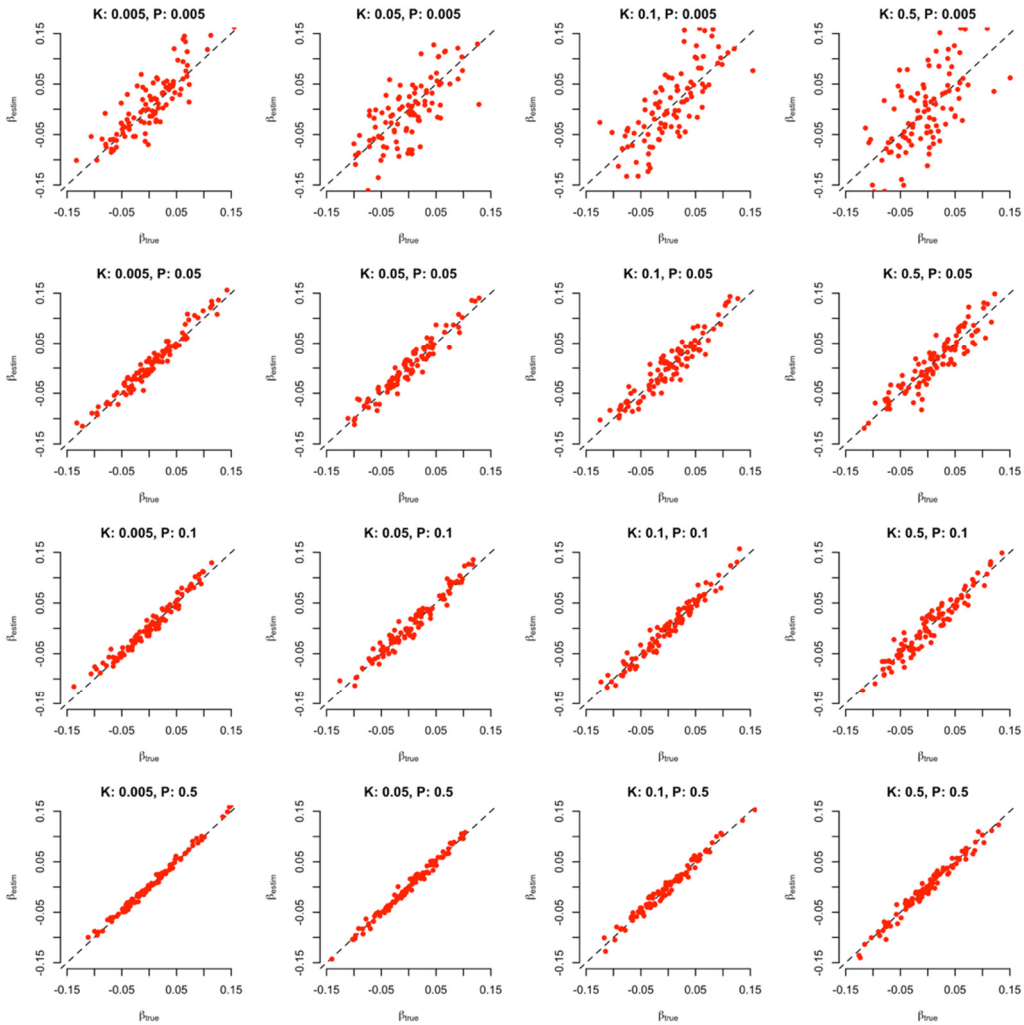


Figure 4. Simulation to verify the scaling of the effect sizes for binary traits. I generated liabilities of $2M$ individuals where $\boldsymbol{l} \sim N(0,1)$ and $\boldsymbol{l} = \boldsymbol{g} + \boldsymbol{\epsilon}$ where $\boldsymbol{g}$ is the random genetic effects and $\boldsymbol{\epsilon}$ is the random errors. In this simulation, I assumed $\boldsymbol{g} = \boldsymbol{X}\boldsymbol{\beta_{true}}$ where $\boldsymbol{\beta_{true}}$ is a 100×1 vector of causal SNPs and $\boldsymbol{\beta_{true}} \sim N(0, \frac{\boldsymbol{h^2}}{100})$, $\boldsymbol{h^2}$ is fixed to 0.3, and X is the genotypes whose minor allele frequency (p) is fixed to 0.3. Then, I generated a case-control study whose total sample number is fixed to $20{,}000$ and estimated the regression coefficients

of 100 variants using a linear regression model. Let β_{estim} be the estimates of the effect sizes after transforming the observed estimate with the proposed scaling scheme. To show the validity of the scaling, we plotted β_{estim} and β_{true} on a 2-D plane (total a hundred red dots). The test was repeated using 16 different combinations of population prevalence and sample prevalence

2.1.3 Step 3: A variance component model to identify pleiotropic loci using GWAS summary statistics

Below, I describe the statistical model of the PLEIO, which is optimized to identify pleiotropic loci. I assume that a phenotype is influenced by K causal SNPs whose individual contribution is very small. For simplicity, I assume that all K SNPs are shared by the T traits. Let $\boldsymbol{\eta_i}$ be a $T \times 1$ vector denoting the true effect sizes of ith causal SNP. Inspired by the LDSC model used in **Step 1: Estimation of correlation matrices**, I assume that all K SNPs have equal contributions such that $\boldsymbol{\eta_i} \sim \mathrm{MVN}(\mathbf{0}, \frac{\boldsymbol{\Omega}}{K})$ where $\boldsymbol{\Omega}$ is the genetic covariance matrix whose diagonal elements are the narrow sense heritabilities. For non-causal SNPs, I assume $\boldsymbol{\eta_i} = \mathbf{0}$. Let $\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}$ is the observed effect sizes of ith SNP and $\boldsymbol{SE}(\hat{\boldsymbol{\eta}}_{\boldsymbol{i}})$ is the corresponding standard errors. I model $\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}$ as the sum of the true genetic effect and the error as follows:

$$\hat{\boldsymbol{\eta}}_{\boldsymbol{i}} = \boldsymbol{\eta_i} + \boldsymbol{\epsilon_i} \qquad \textbf{Equation 3}$$

where $\boldsymbol{\epsilon_i}$ is a random variable of the errors such that $\boldsymbol{\epsilon_i} \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = diag(\boldsymbol{SE}[\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}]) \cdot \mathbf{C}_{\boldsymbol{e}} \cdot diag(\boldsymbol{SE}[\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}])$. Thus, $\hat{\boldsymbol{\eta}}_{\boldsymbol{i}} \sim \mathrm{MVN}(\mathbf{0}, \frac{\boldsymbol{\Omega}}{K} + \boldsymbol{\Sigma})$ for causal SNP and $\hat{\boldsymbol{\eta}}_{\boldsymbol{i}} \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$ for non-causal SNPs. As described earlier, LDSC uses $\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}$ and $\boldsymbol{SE}[\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}]$ for M observed SNPs to estimate the genetic covariance matrix $\widehat{\boldsymbol{\Omega}}$ and

environmental correlation matrix $\widehat{\boldsymbol{\Sigma}}$ (see **Step 1: Estimation of correlation matrices** for details).

In the remaining half of this section, I demonstrate how to test the pleiotropic association using $\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}$, $\boldsymbol{SE}[\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}]$, $\widehat{\boldsymbol{\Omega}}$, and $\widehat{\boldsymbol{\Sigma}}$. First, I relax the assumption that K SNPs have equal contributions and model $\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}$ as the sum of two random variables as follows

$$\hat{\boldsymbol{\eta}}_{\boldsymbol{i}} = \boldsymbol{\gamma}_{\boldsymbol{i}} + \boldsymbol{\epsilon}_{\boldsymbol{i}} \qquad \textbf{Equation 4}$$

where $\boldsymbol{\gamma}_{\boldsymbol{i}}$ is a new random variable of genetic effects that follows $\boldsymbol{\gamma}_{\boldsymbol{i}}$~$\mathrm{MVN}(\mathbf{0}, \tau_i^2\boldsymbol{\Omega})$, and τ_i^2 is a scaling factor of the variance-covariance matrix $\boldsymbol{\Omega}$ so that $\tau_i^2 > 0$ for causal SNPs and $\tau_i^2 = 0$ for non-causal SNPs. Note that, in a special case, $\boldsymbol{\gamma}_{\boldsymbol{i}}$ is equivalent to $\boldsymbol{\eta}_{\boldsymbol{i}}$ when τ_i^2 has the fixed value of $\frac{1}{K}$. In this model, PLEIO tests pleiotropic association by testing $\tau_i^2 = 0$ under the null hypothesis and $\tau_i^2 > 0$ under the alternative hypothesis.

I describe intuitions of our model as follows. The key assumption is that the genetic component $\boldsymbol{\gamma}_{\boldsymbol{i}}$ is a random variable whose variance is proportional to the genetic covariance matrix $\boldsymbol{\gamma}_{\boldsymbol{i}}$. This implies the following two: First, phenotypes with larger heritability show larger genetic effects. Second, phenotypes of multiple traits show genetic effects concordance to their genetic correlations. The statistical model is optimized to have maximized power with $\widehat{\boldsymbol{\Omega}}$ and $\widehat{\boldsymbol{\Sigma}}$. The use of estimates obtained

using whole-genome data, such as $\widehat{\boldsymbol{\Omega}}$ and $\widehat{\boldsymbol{\Sigma}}$, is similar to the approach of the empirical Bayes approaches[22].

I can test the hypothesis by fitting the variance component model in **Equation 4** and obtaining the maximum likelihood estimate (MLE) $\hat{\tau}_i^2$ that maximize $\boldsymbol{\mathcal{L}}\left(\tau_i^2 \middle| \hat{\boldsymbol{\eta}}_{\boldsymbol{i}}; \widehat{\boldsymbol{\Omega}}, \widehat{\boldsymbol{\Sigma}}\right)$. This can be done with numerical solution such as the pseudo-Newton-Raphson method. However, updating the parameter τ_i^2 in the likelihood function above requires a matrix inversion, $\left[\tau_i^2 \widehat{\boldsymbol{\Omega}} + \widehat{\boldsymbol{\Sigma}}\right]^{-\mathbf{1}}$, in every iteration, which has a polynomial time complexity. To solve this challenge, I developed an optimization technique that considerably reduces the computational burden for finding MLE (see **Appendix A**).

In the suggested optimization technique, I apply a linear transformation to $\hat{\boldsymbol{\eta}}_i$ as follows:

$$\widehat{\boldsymbol{\Omega}}^{-\frac{1}{2}} \hat{\boldsymbol{\eta}}_i \sim \text{MVN}\left(\mathbf{0}, \tau_i^2 \boldsymbol{I} + \widehat{\boldsymbol{\Omega}}^{-\frac{1}{2}} \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\Omega}}^{-\frac{1}{2}}\right). \qquad \text{Equation 5}$$

The goal of the optimization is to find the MLE $\hat{\tau}_i^2$ that maximize $\boldsymbol{\mathcal{L}}\left(\tau_i^2 \middle| \widehat{\boldsymbol{\Omega}}^{-\frac{1}{2}} \hat{\boldsymbol{\eta}}_{\boldsymbol{i}}; \widehat{\boldsymbol{\Omega}}^{-\frac{1}{2}} \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\Omega}}^{-\frac{1}{2}}\right)$ with $\hat{\tau}_i^2 > 0$ as a constraint. By applying spectral decomposition, $\boldsymbol{D} = \widehat{\boldsymbol{\Omega}}^{-\frac{1}{2}} \widehat{\boldsymbol{\Sigma}} \widehat{\boldsymbol{\Omega}}^{-\frac{1}{2}} = \boldsymbol{P_D}(\boldsymbol{\Lambda_D}) \boldsymbol{P_D^T}$ where $\boldsymbol{\Lambda_D}$ is a diagonal matrix of the eigenvalues that are arranged in ascending order, and $\boldsymbol{P_D}$ is an eigenvector matrix whose ith column corresponds to the $\boldsymbol{i}$th eigenvalue. Then, $\boldsymbol{D^{-1}}$ can be simplified as $\boldsymbol{P_D}(\boldsymbol{\Lambda_D} + \tau_i^2 \boldsymbol{I})^{-\mathbf{1}} \boldsymbol{P_D^T}$. Note that the computation of $\boldsymbol{P_D}(\boldsymbol{\Lambda_D} +$

$\tau_i^2 \boldsymbol{I})^{-1} \boldsymbol{P_D^T}$ is much easier than the computation of $(\tau_i^2 \widehat{\boldsymbol{\Omega}} + \widehat{\boldsymbol{\Sigma}})^{-1}$. The log-likelihood function obtained through the linear transformation (ℓ_1') can be shown as follows:

$$\begin{aligned}\ell_1' &= -\frac{\mathbf{1}}{\mathbf{2}}\Bigg[T\ln(2\pi) + \sum\nolimits_{t=1}^{R}\ln(\boldsymbol{\xi_t} + \tau_i^2) \\ &\qquad + \left(\boldsymbol{P_D E}\left[\widehat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}\right)^{T}[\boldsymbol{\Lambda}_{\boldsymbol{D}}^{+}]^{-1}\left(\boldsymbol{P_D E}\left[\widehat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}\right)\Bigg] \\ &= -\frac{1}{2}\left[T\ln(2\pi) + \sum\nolimits_{t=1}^{R}\ln(\xi_t + \tau_i^2) + \sum_{t=1}^{R}\frac{\boldsymbol{\delta_t^2}}{\xi_t + \tau_i^2}\right]\end{aligned} \qquad \text{Equation 6}$$

where R is the number of non-zero eigenvalues, $\boldsymbol{\Lambda_D^+}$ is a $R \times R$ matrix that removed columns with zero diagonal elements of $\boldsymbol{\Lambda_D}$, ξ_t is the tth diagonal element of $\boldsymbol{\Lambda_D^+}$, $\boldsymbol{\delta_t^2}$ is the tth element of the vector $\boldsymbol{P_D E}\widehat{\boldsymbol{\Omega}}^{-\frac{1}{2}}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}$, and $\boldsymbol{E}$ is a diagonal matrix of which the first p elements are ones, and the rest are zeros.

The first and second derivatives of ℓ_1' with respect to τ_i^2 are:

$$\begin{aligned}\frac{d\ell_1'}{d\tau_i^2} &= -\frac{1}{2}\left[\sum\nolimits_{t=1}^{R}\frac{1}{\xi_t + \tau_i^2} - \sum_{t=1}^{R}\frac{\boldsymbol{\delta_t^2}}{(\xi_t + \tau_i^2)^2}\right] \\ \frac{d^2\ell_1'}{d(\tau_i^2)^2} &= -\frac{1}{2}\left[\sum\nolimits_{t=1}^{R}\frac{1}{(\xi_t + \tau_i^2)^2} + 2\sum_{t=1}^{R}\frac{\boldsymbol{\delta_t^2}}{(\xi_t + \tau_i^2)^3}\right]\end{aligned} \qquad \text{Equation 7}$$

With **Equation 7**, I can obtain optimal $\hat{\tau}_i^2$ using the Newton Raphson method. the resulting log-likelihood ratio test (LRT) statistic can be shown as follows:

$$S_{PLEIO} = \left[\sum_{t=1}^{R} \ln\left(\frac{\xi_t}{\xi_t + \hat{\tau}_i^2}\right)\right] + \left[\sum_{t=1}^{R} \frac{\boldsymbol{\delta_t^2}}{\xi_t} - \sum_{t=1}^{R} \frac{\boldsymbol{\delta_t^2}}{\xi_t + \hat{\tau}_i^2}\right] \quad \text{Equation 8}$$

This technique can substantially reduce the time to complete the association tests where the amount of the time reduction increases with the increase of the number of traits (**Figure 5** and **Figure 6**). The use of spectral decomposition was inspired by the technique used in the Efficient Mixed-Model Association (EMMA)[23]. Kang et al. applied eigendecomposition on the variance of a linear mixed model to reduce the time complexity of solving REML.

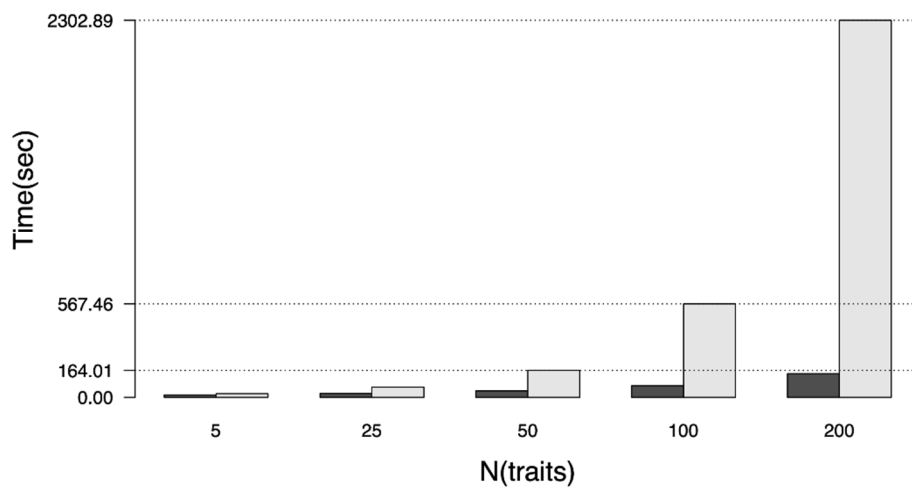


Figure 5. Comparison of the computational efficiency between the proposed Newton Raphson (NR) technique and the pseudo-NR technique implemented in Python's Scipy library. I measured the time for a 10K SNP test with PLEIO, where I changed the number of traits from 5 to 200. The dark gray bar represents the NR method implemented in PLEIO, and the light gray bar represents the pseudo-Newton-Raphson method implemented in Python's Scipy library.

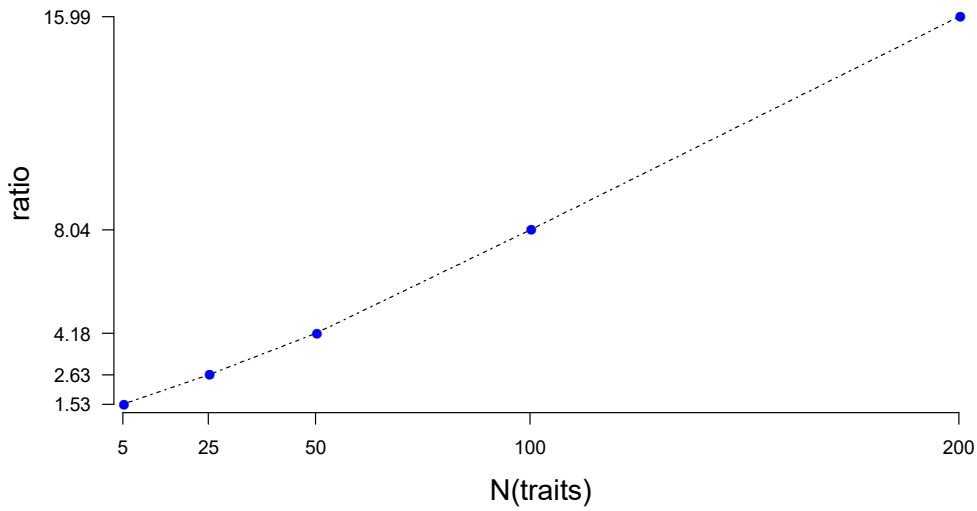


Figure 6. Line plot comparing the computational time of the NR method proposed by PLEIO and the pseudo-NR method implemented in Scipy library. Each point indicates the ratio of computational time for testing 10K simulations using the NR implemented in PLEIO to the pseudo NR implemented in Scipy.

2.1.4 Step 4: P-value estimation using a novel importance sampling method

I describe how the statistical significance (p-value) of the LRT statistic, S_{PLEIO}, is evaluated. PLEIO's LRT statistic uses only one variance estimate as to the parameter. According to Self and Liang, the statistic asymptotically follows a 50:50 mixture of χ_0^2 and χ_1^2 under the null hypothesis[24]. However, the asymptotic approximation is inaccurate when the number of traits (T) is small. I found that the null p-values calculated from the asymptotic distribution deviate from the uniform distribution (the left column of **Figure 7**). S_{PLEIO} has a unique null distribution for every combination of $\widehat{\boldsymbol{\Omega}}$ and $\widehat{\boldsymbol{\Sigma}}$. In this case, a reasonable solution is to estimate null distribution using by applying a simulation-based approach (e.g., Monte Carlo

method) to the analysis that uses those study-specific factors ($\widehat{\boldsymbol{\Omega}}$ and $\widehat{\boldsymbol{\Sigma}}$). The suggested approach has to accurately approximate the p-value at a very small quantity (e.g., 5×10^{-8}). For this reason, the use of the Monte Carlo method is not an optimal solution, as it will significantly increase the total analysis time.

Below, I suggest a novel importance sampling method to assess the p-value of S_{PLEIO}. Let x be a random variable of the standard effect sizes, and $q(x)$ be the probability density function (PDF) of x under the null hypothesis. Thus, $q(\boldsymbol{x}) \sim \mathrm{MVN}\left(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}\right)$. By the definition of the probability distribution, $\int_D q(x)\, d\boldsymbol{x} = 1$ when $\boldsymbol{D} = \mathbb{R}^T$. In this section, Below, I treat S_{PLEIO} as a function of x given $\widehat{\boldsymbol{\Omega}}$ and $\widehat{\boldsymbol{\Sigma}}$, and let θ be the observed statistics of S_{PLEIO}. Using definitions above, I define an indicator function $f(x, \theta)$ as follows:

$$f\left(x, \theta | \widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\Omega}}\right) = \begin{cases} 1 \; if \; S_{PLEIO}(x | \widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\Omega}}) \geq \theta \\ 0 \; if \; S_{PLEIO}\left(x \middle| \widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\Omega}}\right) < \theta \end{cases}. \qquad \text{Equation 9}$$

For simplicity, I replace $f\left(\boldsymbol{x}, \theta | \widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\Omega}}\right)$ with $f(\boldsymbol{x})$. Then, the p-value of θ can be shown as:

$$I = \int_{\boldsymbol{D}} m(x) dx \qquad \text{Equation 10}$$

where $m(x) = f(x)q(x)$. To estimate the value of I, one can use the importance sampling approaches. Let $p(x)$ be a sampling distribution that differs from $q(x)$, and $\boldsymbol{X}^p \sim p(x)$ denote a $M \times T$ matrix of the effect sizes sampled from $p(x)$ where M

is the number of samples. Note that M can be any number but is usually smaller than the number of samples using the Monte Carlo method. Then, the estimate I using $\boldsymbol{X}^p$ can be shown as follows:

$$\hat{I} = \mathrm{E}^p\left[\frac{f(x)q(x)}{p(x)}\right] = \frac{1}{M}\left[\sum_{i=1}^{M}\frac{f(X_i^p)q(X_i^p)}{p(X_i^p)}\right] \qquad \text{Equation 11}$$

where $\mathrm{E}^p[\cdot]$ denotes the expectation over $\boldsymbol{X}^p$, and $\boldsymbol{X}_i^p$ is the ith row vector of $\boldsymbol{X}^p$. The challenge in the above importance sampling is to choose an $p(x)$ that minimize the variance of $\hat{I}$. In GWAS, this can be more challenging because the range of $\hat{I}$ is very wide, from 1.0 to 5×10^{-8}. In other words, each θ may have an optimal $p(x)$ that minimizes the variance of $\hat{I}$. To solve this problem, I applied the importance sampling method suggested by Owen and Zhou[25]. The proposed method generates samples from a mixture distribution. Let $p_j(x)$ be the jth sampling distribution where $j = \{1,2,\dots,F\}$. Unlike the conventional importance sampling method, $p_j(x)$ can include $q(x)$. Let $p_\alpha(x)$ be the mixture distribution of F sampling distribution. In PLEIO, I assume 11 sampling distributions ($F = 11$) including $q(x)$ and assume the equal contribution of jth sampling distribution such that $p_\alpha(x) = \frac{1}{F}\sum_{j=1}^{F} p_j(x)$. Detailed information on how I selected $p_j(x)$ can be found in **Appendix B.**

In the suggested importance sampling, Owen and Zhou used each $p_j(x)$ as a control variate of $m(x) = f(x)q(x)$ to reduce the variance of $\hat{I}$[25]. Let $m^*(x,\beta)$ be an estimator of $m(x)$ which can be shown as follows:

$$m^*(x,\beta) = m(x) - \sum_{j=1}^{K} \beta_j \left(p_j(x) - \int_{\boldsymbol{D}} p_j(x)\, dx \right) \qquad \text{Equation 12}$$

where $\mathrm{E}[m^*] = \mathrm{E}[m] = I$, and $\int_D p_j(x)\, dx = 1$. The control variate method minimizes the variance of $\hat{I}$ with the optimal control variate coefficient $(\boldsymbol{\beta}^*)$ where $\boldsymbol{\beta}^* = \{\beta_1, \beta_2, \ldots, \beta_F\}$. Then, the variance $\mathrm{Var}(m^*)$ is equivalent to or smaller than $\mathrm{Var}(m)$. Following **Equation 12**, the p-value estimate of θ can be shown as follows:

$$\begin{aligned} \hat{I} &= \mathrm{E}^{p_\alpha}[m^*] \\ &= \frac{1}{M} \left(\sum_{i=1}^{M} \frac{f(X_i^p) q(X_i^p) - \sum_{j=1}^{K} \beta_j p_j(X_i^p)}{p_\alpha(X_i^p)} \right) + \sum_{k=1}^{K} \beta_k \end{aligned} \qquad \text{Equation 13}$$

In PLEIO, I implemented the suggested importance sampling above as follows. Instead of estimating p-values of every variant being tested, I approximate the null distribution of S_{PLEIO} using 40 different θ that roughly correspond to the p-values from 1.0 to 5×10^{-8}. Note that the p-value estimation in **Equation 13** requires optimization of $\boldsymbol{\beta}$ to maximize the variance reduction of each p-value estimate. Using these 40 p-values, I interpolate p-values for $\theta < 40$ using B-spline fit and extrapolate p-values for $\theta > 40$ using the linear fit on the logarithmic p-value scale. A detailed description of how to get the optimal control variate coefficients $(\boldsymbol{\beta}^*)$ can be found in **Appendix B**.

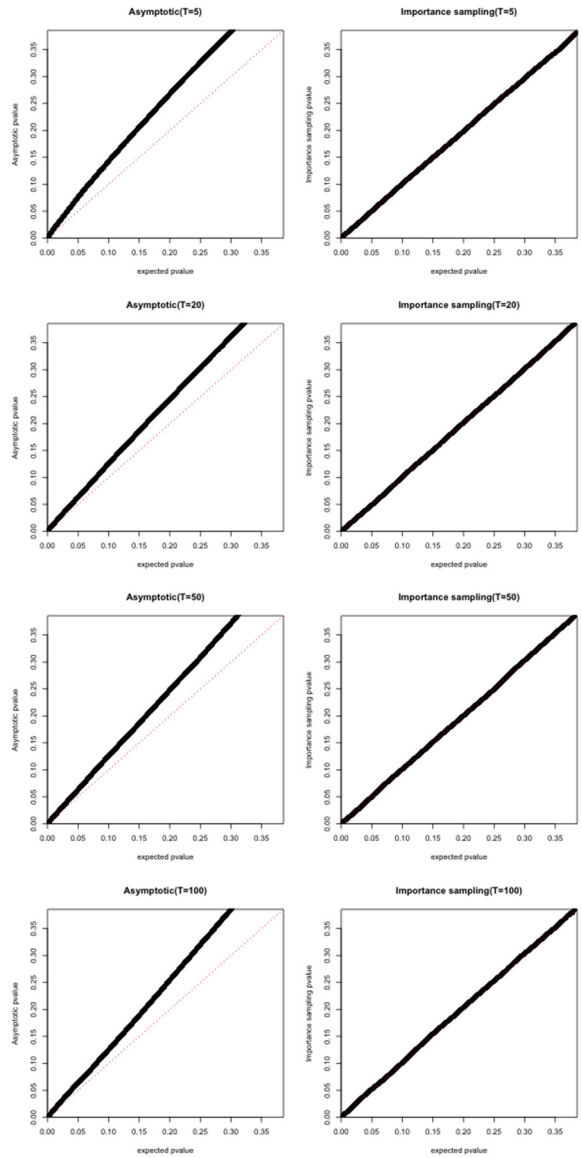


Figure 7 PLEIO's p-value distribution plot. We compared the probability plots of p-values obtained using the importance sampling method and asymptotic distribution under the null hypothesis. The p-values were sorted in ascending order and compared with the expected p-values. I have changed the number of traits from 5 to 100.

2.1.5 Step 5: Visualization of pleiotropic association pattern

PLEIO offers PleiotropyPlot, which visualizes the pleiotropic effects of a SNP in a circular plot[26]. Each plot contains information on the normalized effect sizes, the local heritability, the genetic correlation structure, and the local Manhattan plots of the SNP. The information in the outer part of the plot is as follows: 1. textual information of the effect sizes and p-values obtained from raw summary statistics of the traits. 2. per trait regional Manhattan plots showing the p-values of the SNPs within $1M$ base-pair window. 3. bar plots whose length indicates the magnitude of the standard effect and whose color indicates the direction of the effects. The inner part of this plot is a ribbon plot where each ribbon connects two traits. The color of the ribbon indicates the magnitude of the genetic correlation between the two traits, and the width at the end of each ribbon indicates the locus heritability of the trait on a relative scale (square of the normalized effect size).

2.2 Simulations

2.2.1 Evaluation method of false-positive rate

Below, I describe how to generate SNP effects in the false positive rate (FPR) simulation. In each simulation of the given $\mathbf{\Sigma}_{sim}$ and $\mathbf{C}_{\boldsymbol{e}}$, I generated standard effect sizes ($\hat{\boldsymbol{\eta}}_{sim}$) from $\mathrm{MVN}(\mathbf{0}, \mathbf{\Sigma}_{sim})$ where $\mathbf{\Sigma}_{\mathrm{sim}} = \mathbf{C}_{\boldsymbol{e}}$, and $\boldsymbol{SE}[\hat{\boldsymbol{\eta}}_{sim}] = 1$. I assumed a significance threshold level α and estimated FPR at $\alpha = 0.05$ as the proportion of simulations whose $P \leq \alpha$.

2.2.2 Generation of effect sizes used in power simulation

I provide a detailed description of how I generate SNP effects used in the power simulation. I assumed a SNP whose minor allele frequency p is fixed to 0.3 under the Hardy Weinberg equilibrium. I performed a joint analysis of seven traits ($T = 7$) for each simulation setting. Each setting differed following factors: heritability $\boldsymbol{h^2}$, the genetic correlation $\boldsymbol{C_g}$, phenotypic unit ($\boldsymbol{U}$), and phenotypic type (either B or Q). For simplicity, I treated the environmental correlation matrix, $\boldsymbol{C_e}$, as a diagonal matrix of ones.

First, I describe how to generate effect sizes for quantitative traits (Q). Let $\boldsymbol{\beta_i} = (\beta_{i1}, \beta_{i2}, \dots, \beta_{iT})$ be the $T \times 1$ vector of true effect sizes of the SNP i. I generated $\boldsymbol{\beta}_i \sim \mathrm{MVN}\left(\mathbf{0}, \frac{1}{M_{true}} \boldsymbol{\Omega}_{\mathrm{sim}}\right)$ where $\boldsymbol{\Omega}_{\mathrm{sim}} = \mathrm{diag}\left(\sqrt{\boldsymbol{h^2}}\right) \cdot \boldsymbol{C_g} \cdot \mathrm{diag}\left(\sqrt{\boldsymbol{h^2}}\right)$, and M_{true} is the number of causal variants. By default, $M_{true} = 1{,}000$. Let N be the number of samples, h_t^2 is the tth element of $\boldsymbol{h}^2$, $\boldsymbol{x_t}$ is a $N \times 1$ vector of genotypes. For each trait, I simulated $\boldsymbol{x_t}$ from $Binomial(2, p)$ and standardized them such that $\boldsymbol{x}_{t,std} = \frac{x_t - 2p}{2p(1-p)}$. Then, I generated $\boldsymbol{y_t}$ from the linear model $\boldsymbol{y_t} = \boldsymbol{\beta_{it}} \boldsymbol{x_{t,std}} + \boldsymbol{\epsilon_t}$ where $\boldsymbol{\epsilon_t} = (\epsilon_{t1}, \epsilon_{t2}, \dots, \epsilon_{tN})$ is a vector of errors and $\boldsymbol{\epsilon_{tj}} \sim \mathrm{N}\left(0, 1 - \frac{1}{M_{true}}\ h_t^2\right)$. the observed effect size $\hat{\beta}_{it}$ is the regression coefficient of the linear regression of $\boldsymbol{y_t}$ to $\boldsymbol{x}_{t,std}$.

Second, I describe how to generate effect sizes for binary traits (B). I assumed that each binary study contains half cases ($\frac{N}{2}$) and half controls ($\frac{N}{2}$).

[Step 1: Sampling of true effect size under observed linear scale]

Let $\boldsymbol{\beta_i}$ be the $T \times 1$ vector of true effect sizes of the SNP i and $\boldsymbol{\beta_i} \sim \mathrm{MVN}\left(\mathbf{0}, \frac{1}{M_{true}} \boldsymbol{\Omega_{sim}}\right)$. For binary traits, I treat $\boldsymbol{\beta_i}$ as the true effect sizes under the liability scale. I converted this true effect size into the observed scale $\boldsymbol{\beta}_{i,obs} = \boldsymbol{\beta_i}/\delta_t$. where δ_t is a scaling factor and $\delta_t = \frac{K_t^2(1-K_t)^2}{P_t(1-P_t)} \cdot \frac{1}{[\psi(\phi^{-1}(1-K_t))]^2}$, where K_t refers to the disease prevalence of trait t, P_t refers to the sample prevalence (fixed to 0.5 in this power test), ψ refers to the probability density function of the standard normal distribution, and ϕ^{-1} refers to the inverse of the cumulative density function of the standard normal distribution.

[Step 2: Searching for relative risk corresponding to the true effect size]

Now, I search for the value of the relative risk γ giving $\boldsymbol{\beta}_{i,obs}$ under the observed linear scale. Suppose that γ is any scalar value. Given the disease prevalence K_t and the population minor allele frequency (MAF) p (which is 0.5), the expected case MAF is $p^+ = \frac{\gamma p}{p(\gamma-1)+1}$ and the expected control MAF is $p^- = \frac{p - p^+ K_t}{1-K_t}$. Assuming the Hardy-Weinberg equilibrium, I can construct a set of reference genotypes such that the case MAF is exactly p^+ and control MAF is exactly p^- (after ignoring the integer rounding). Let $\overline{\boldsymbol{x}_t}$ is a $N \times 1$ vector of the reference genotypes and $\overline{\boldsymbol{x}_{t,std}} = \frac{\overline{x_t} - 2p}{2p(1-p)}$. As mentioned earlier, the vector of phenotypes $\boldsymbol{y_t}$ consists of half ones and half zeros. Thus, I can perform a linear regression analysis of $\boldsymbol{y_t}$ on $\boldsymbol{y_t}$ to get the effect size under the observed linear scale $\overline{\boldsymbol{\beta}_{i,obs}}$. Finally, I search the relative risk estimate $\hat{\gamma}$ that satisfies $\boldsymbol{\beta}_{i,obs} = \overline{\boldsymbol{\beta}_{i,obs}}$.

[Step 3: Generating genotypes]

Now I can create the genotypes $\boldsymbol{x_t}$ using the value of $\hat{p}^+$ and $\hat{p}^-$ that corresponds to $\hat{\gamma}$, I generated the case genotypes from $x_{tj} \sim Binomial(2, \hat{p}^+)$ and the control genotypes from $x_{tj} \sim Binomial(2, \hat{p}^-)$.

[Step 4: Logistic regression]

Commonly, the heritability calculations of binary traits are based on the (observed and liability scale) linear model. This was why I had to derive the relative risk and the case and control MAFs through the observed scale linear model. However, in association analyses, the logistic regression model is commonly used. To simulate a realistic situation, I applied logistic regression to $\boldsymbol{y_t}$ and the sampled $\boldsymbol{x_t}$. This way, I obtained the log odds ratios and the standard errors for *T* traits. This information was used as the final input for binary traits in our power analysis.

2.3 Real data analysis

2.3.1 Data collection

I collected 18 public GWAS summary statistics on traits related to cardiovascular disease. These 18 traits include the diseases (binary phenotypes) and the complex traits (quantitative phenotypes).

Table 1 provides a detailed summary of these 18 traits. I selected the most recent study where the consortium has more than one summary statistics for the same trait.

From the Global Lipid consortium[27], I collected summary statistics on four quantitative traits. Each summary statistics is the result of GWAS from 94,595 individuals from 23 studies genotyped with GWAS array and 93,982 individuals from 37 studies genotyped with Metabochip array. From the GWAS results of the UK biobank, I collected summary statistics of twelve binary traits(**Table 2**)[28]. Each summary statistics is the result of GWAS from 361,193 individuals. From the CARDIo+C4D consortium[2], I collected summary statistics of one binary trait (coronary artery disease), which is the result of GWAS from 60,801 cases and 123,504 controls of 48 studies. Finally, from the MAGIC consortium[29], I collected summary statistics of one quantitative trait (fasting glucose), GWAS from 46,186 non-diabetic patients of 21 studies. All study subjects were of European ancestry.

Phenotype	$\mathbf{S_{prev}}$	$\mathbf{P_{prev}}$	Year	Phenotype	Database	$\mathbf{N_{total}}$	$\mathbf{N_{controls}}$	$\mathbf{N_{cases}}$	LDSC_h2
Heart attack: 6150_1	0.024	0.024	2018	Binary	UK Biobank	360419	352132	8287	0.14
Hypertension: I9	0.003	0.003	2018	Binary	UK Biobank	361193	359957	1236	0.11
Essential (primary) hypertension: I10	0.002	0.002	2018	Binary	UK Biobank	361193	360329	864	0.11
Acute myocardial infarction: I21	0.018	0.018	2018	Binary	UK Biobank	361193	354782	6411	0.13
Myocardial infarction: I9	0.020	0.020	2018	Binary	UK Biobank	361193	354175	7018	0.13
Major coronary heart disease: I9	0.029	0.029	2018	Binary	UK Biobank	361193	351037	10156	0.13
Ischemic heart disease: I9	0.061	0.061	2018	Binary	UK Biobank	361193	340337	20856	0.13
Coronary atherosclerosis: I9	0.041	0.041	2018	Binary	UK Biobank	361193	346860	14333	0.15
Heart failure	0.004	0.004	2018	Binary	UK Biobank	361193	359789	1404	0.16
Obesity: E66	0.001	0.001	2018	Binary	UK Biobank	361193	360752	441	0.36
Type 1 diabetes: E4	0.002	0.002	2018	Binary	UK Biobank	361193	360611	582	0.17
Type 2 diabetes: E4	0.002	0.002	2018	Binary	UK Biobank	361193	360305	888	0.17
Coronary artery disease	0.492	0.050	2015	Binary	CARDIo+C4D	184305	123504	60801	0.06
High-density lipoprotein	NA	NA	2013	Quantitative	Global Lipids	188577	NA	NA	0.21
Low-density lipoprotein	NA	NA	2013	Quantitative	Global Lipids	188577	NA	NA	0.20
Total cholesterol	NA	NA	2013	Quantitative	Global Lipids	188577	NA	NA	0.21
Triglycerides	NA	NA	2013	Quantitative	Global Lipids	188577	NA	NA	0.21
Fasting glucose	NA	NA	2011	Quantitative	Magic	46186	NA	NA	0.09

Table 1. The list of phenotypes included in the PLEIO's real data analysis. For binary phenotypes, S_{prev} means sample prevalence, and P_{prev} means the population prevalence.

Phenotype Code	Phenotype Description	Sex	File
6150_1	Vascular/heart problems diagnosed by doctor: Heart attack	both_sexes	6150_1.gwas.imputed_v3.both_sexes.tsv.bgz
I9_HYPTENS	Hypertension	both_sexes	I9_HYPTENS.gwas.imputed_v3.both_sexes.tsv.bgz
I10	Diagnoses - main ICD10: I10 Essential (primary) hypertension	both_sexes	I10.gwas.imputed_v3.both_sexes.tsv.bgz
I21	Diagnoses - main ICD10: I21 Acute myocardial infarction	both_sexes	I21.gwas.imputed_v3.both_sexes.tsv.bgz
I9_MI	Myocardial infarction	both_sexes	I9_MI.gwas.imputed_v3.both_sexes.tsv.bgz
I9_CHD	Major coronary heart disease event	both_sexes	I9_CHD.gwas.imputed_v3.both_sexes.tsv.bgz
I9_IHD	Ischaemic heart disease, wide definition	both_sexes	I9_IHD.gwas.imputed_v3.both_sexes.tsv.bgz
I9_CORATHER	Coronary atherosclerosis	both_sexes	I9_CORATHER.gwas.imputed_v3.both_sexes.tsv.bgz
I9_HEARTFAIL	Heart failure,strict	both_sexes	I9_HEARTFAIL.gwas.imputed_v3.both_sexes.tsv.bgz
E66	Diagnoses - main ICD10: E66 Obesity	both_sexes	E66.gwas.imputed_v3.both_sexes.tsv.bgz
E4_DM1	Type 1 diabetes	both_sexes	E4_DM1.gwas.imputed_v3.both_sexes.tsv.bgz
E4_DM2	Type 2 diabetes	both_sexes	E4_DM2.gwas.imputed_v3.both_sexes.tsv.bgz

Table 2 The detailed description of twelve UKB traits. The data above can be found at Neale lab's UKB summary statistics portal.

2.3.2 Quality control of the data

Each study underwent the following standard quality control protocols. For each summary statistics, I excluded SNPs not in the 1000 Genomes and checked the consistency of allele pair of each SNP with the corresponding allele pair of the SNP in 1000 Genomes. In addition, I removed all strand-ambiguous SNPs that have allele pair GC or AT. A total of 1,777,411 SNPs were included in the joint analysis of the 18 traits. Summary statistics of these remaining SNPs were used to estimate the genetic covariance and error correlation.

Chapter 3. Results

3.1 Overview of the method

PLEIO is a multi-trait framework to identify and interpret pleiotropic loci. PLEIO uses genetic covariance and environmental correlation across these traits to optimize the statistical power. I described the statistical model used in PLEIO using a toy model of three traits (A, B, and C) in **Figure 8**. Let X_1 be a SNP that has observed effect sizes of $(2.2, 2.8, -1.2)$, and X_2 be another SNP that has observed effect sizes of $(-1.5, 0.4, -2.7)$. For simplicity, the variances of all estimates were assumed to be one. If I test the SNP association using the fixed-effects meta-analysis method, these SNPs have the exact p-value $(P = 0.03)$ as the magnitude of the mean effect size is the same. However, suppose we know that A and B have a positive correlation, and C has a negative correlation with the rest. Then, taking into account the genetic correlation between traits, SNP X_1 is more likely to be a true signal compared to X_2. Additionally, suppose we know that B has the most significant heritability and C has the least heritability, which makes the association of X_1 much more likely to be a true signal because the relative strength of the effect sizes is similar to the heritabilities. PLEIO accounts for the genetic covariance and environmental correlation and gives a more significant p-value at SNP X_1 $(P = 0.0006)$ than X_2 $(P = 0.1)$.

The complete analysis of PLEIO consists of five steps. First, PLEIO uses LD score regression (LDSC)[20] to estimate the genetic correlation $\boldsymbol{C_g}$, environmental

correlation $\boldsymbol{C_e}$, and the heritabilities $\boldsymbol{h^2}$. Note that the genetic covariance $\boldsymbol{\Omega}$ is obtained by summarizing $\boldsymbol{C_g}$ and $\boldsymbol{h^2}$. Second, it changes the scale of observed effect sizes $\hat{\boldsymbol{\beta}}$ into the standardized scale $\hat{\boldsymbol{\eta}}$. For quantitative traits, $\hat{\boldsymbol{\eta}}$ corresponds to the regression coefficient of simple linear regression whose dependent and independent variables follow $\mathrm{N}(0,1)$. For binary traits, $\hat{\boldsymbol{\eta}}$ is the standardized effect sizes for liability. Third, PLEIO uses a variance component model $\hat{\boldsymbol{\eta}} = \boldsymbol{g} + \boldsymbol{e}$ to map pleiotropic loci (**Figure 3**). The primary assumption of this statistical model is that the genetic effects $\boldsymbol{g}$ follows the genetic covariance, $\mathrm{Var}(\mathbf{g}) = \tau^2\boldsymbol{\Omega}$. Each SNP association test, PLEIO performs hypothesis test of $H_0: \tau^2 = 0$ versus $H_1: \tau^2 > 0$. To increase the computational efficiency, an optimization technique using spectral decomposition of the variance, $\mathrm{Var}(\boldsymbol{g}) + \mathrm{Var}(\boldsymbol{e})$, was applied. Fourth, PLEIO uses an importance sampling method to assess the one-tailed p-value per SNP. Lastly, PLEIO provides a visualized summary of the analysis results to help interpretation.

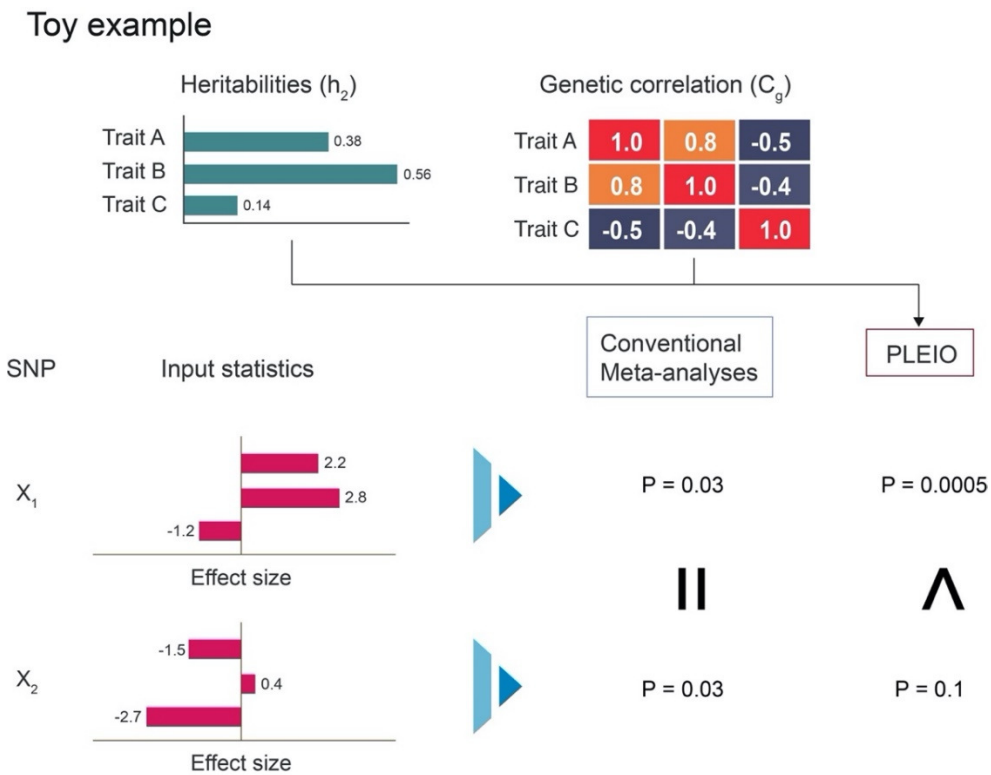


Figure 8. A toy example designed to understand the association analysis carried out by PLEIO.

3.2 Evaluation of false-positive rates in null simulations

I evaluated the false positive rate (FPR) of PLEIO using extensive simulations. I assumed the null hypothesis of no genetic effects at a SNP for all T traits. I tested FPR by differing the following four factors: 1. the number of traits T, 2. the heritabilities $\boldsymbol{h}^2$, 3. the genetic correlation matrix $\boldsymbol{C_g}$, and 4. the environmental correlation matrix $\boldsymbol{C_e}$. Note that, the $\boldsymbol{h}^2$ and $\boldsymbol{\Omega}$ are zeros under the null, but I treated $\boldsymbol{h}^2$ and $\boldsymbol{\Omega}$ as input parameters generated from an external dataset. In other words, $\boldsymbol{h}^2$ and $\boldsymbol{\Omega}$ describe what PLEIO thought to be true. In a real data analysis, $\boldsymbol{h}^2$ and $\boldsymbol{\Omega}$ (and $\boldsymbol{C_e}$) are estimates generated from GWAS summary statistics.

I tested three different *T*= 5, 10, and 20. I set off-diagonal elements of $\boldsymbol{C_e}$ to 0.0 and 0.5 to simulate uncorrelated and correlated errors, respectively. I simulated two different $\boldsymbol{h^2}$. For "equal $\boldsymbol{h^2}$", I set the same heritability for all traits as 0.5. For "different $\boldsymbol{h^2}$", I simulated heritabilities ranged from 0.1 to 0.5. I simulated two different $\boldsymbol{C_g}$. For "uniform $\boldsymbol{C_g}$", I set off-diagonal elements of $\boldsymbol{C_g}$ to 0.3. For "partitioned $\boldsymbol{C_g}$", I set up two sub-groups and set the off-diagonal elements to 0.3 within a group and to 0 between groups. A total of 24 FPR tests were conducted. In each simulation, I generated one million null datasets for each situation and calculated FPR at $a = 0.05$. **Table 3** shows that PLEIO's FPR is well calibrated in all situations.

Next, I examined if PLEIO's FPR is well-calibrated at a very low threshold of 5×10^{-8}, which is used as the statistical significance of the conventional GWAS. I assumed three situations that use $T = 5, 10, 20$ and increased the number of null datasets to a billion (10^9). For each situation, I assumed that the equal $\boldsymbol{h^2}$, partitioned $\boldsymbol{C_g}$, and no sample overlap (uncorrelated errors). **Table 4** shows that PLEIO's FPR is well calibrated for α down to $\alpha = 5 \times 10^{-8}$. See **Evaluation method of false-positive rate** for a detailed explanation for the simulation.

		$\mathbf{T = 5}$		$\mathbf{T = 10}$		$\mathbf{T = 20}$	
		Equal $\mathbf{h_g^2}$	Diff $\mathbf{h_g^2}$	Equal $\mathbf{h_g^2}$	Diff $\mathbf{h_g^2}$	Equal $\mathbf{h_g^2}$	Diff $\mathbf{h_g^2}$
Uniform $\mathbf{C_g}$	No $\mathbf{C_e}$	0.0499	0.0497	0.0500	0.0497	0.0505	0.0499
	Uniform $\mathbf{C_e}$	0.0499	0.0499	0.0496	0.0499	0.0496	0.0500
Partitioned $\mathbf{C_g}$	No $\mathbf{C_e}$	0.0497	0.0498	0.0499	0.0498	0.0509	0.0502
	Uniform $\mathbf{C_e}$	0.0499	0.0500	0.0495	0.0499	0.0505	0.0502

Table 3. PLEIO's FPR in various simulation conditions. In this simulation, 10^7 null study sets were generated for each of the 24 situations, and the FPR was calculated at $\alpha = 0.05$. Each test consisted of a unique combination of four parameters: T, $\mathbf{h}^2$, C_g, and C_e. I changed the number of studies (T) to 5, 10, and 20. "Equal $\mathbf{h^2}$" denotes that the heritability is fixed to the value of 0.5, and "Diff $\mathbf{h^2}$" denotes that the values of the heritabilities increase from 0.1 to 0.5. "Uniform $\mathbf{C_g}$" denotes a genetic correlation matrix whose non-diagonal value is fixed to 0.3, and "partitioned $\mathbf{C_g}$" denotes a genetic correlation matrix consisting of two sub-groups where each non-diagonal value within a group is fixed to 0.3, and each non-diagonal value between groups is fixed to 0. "No $\mathbf{C_e}$" denotes an environmental correlation matrix whose non-diagonal value is fixed to 0, and "Uniform $\mathbf{C_e}$" denotes an environmental correlation matrix whose non-diagonal value is fixed to 0.5.

FPR	$T = 5$	$T = 10$	$T = 20$
5×10^{-2}	5.02×10^{-2}	5.03×10^{-2}	5.09×10^{-2}
5×10^{-3}	5.02×10^{-3}	5.05×10^{-3}	5.07×10^{-3}
5×10^{-4}	5.00×10^{-4}	5.07×10^{-4}	5.06×10^{-4}
5×10^{-5}	5.01×10^{-5}	5.04×10^{-5}	4.99×10^{-5}
5×10^{-6}	5.00×10^{-6}	4.91×10^{-6}	5.06×10^{-6}
5×10^{-7}	4.86×10^{-7}	5.94×10^{-7}	5.57×10^{-7}
5×10^{-8}	5.50×10^{-8}	5.70×10^{-8}	4.00×10^{-8}

Table 4. PLEIO's FPR at genome-wide thresholds. In this simulation, I generated 10^9 null study sets to test FPR in different α ranging from 5×10^{-2} to 5×10^{-8}. I changed the number of studies (*T*) to 5, 10, and 20, and the fixed $\mathbf{h}^2, \mathbf{C_g},$ and $\mathbf{C_e}$ as follows: For $\mathbf{h}^2$, I used a $T \times 1$ vector whose values increase in the range of (0.1, 0.5). For $\mathbf{C_g}$, I used a genetic correlation matrix consisting of two sub-groups where each non-diagonal value within a group is fixed to 0.3. Each non-diagonal value between groups is fixed to 0. For $\mathbf{C_e}$, I used a diagonal matrix.

3.3 Evaluation of power in alternate simulations

I compared the power of PLEIO with conventional meta-analysis methods: the fixed effects meta-analysis and association analyses based on SubSETs (ASSET)[30]. For the fixed-effects method, I used the inverse variance weighted sum method implemented in METAL[31]. In addition, I applied Lin-Sullivan's approach to the inverse variance weighted sum method above to explain the correlation due to sample overlap between traits. Finally, I generated a simple R test code of the fixed effect method above. For ASSET, I downloaded and used the R package "ASSET."

Additionally, I compared the power of PLEIO with a trait-specific approach (MTAG)[19]. MTAG jointly analyzes summary statistics of GWAS as in the meta-analysis methods, but there are differences in identifying pleiotropic loci. A meta-analytic method gives a single p-value per SNP, but MTAG gives multiple p-values per SNP (T p-values per SNP). A straightforward solution is to choose a minimum p-value per SNP, but it leads to multiple testing problems. In the FPR test result, I observed inflated FPR in MTAG that uses the minimum p-value approach. To correct the multiple testing problem, I applied the Bonferroni correction by multiplying the minimum p-value by T. I observed that Bonferroni correction can control FPR but is conservative due to the correlation between T effect size estimates from MTAG. Therefore, I reported the powers of MTAG both before the Bonferroni correction (MTAG-U; uncorrected) and after the Bonferroni correction (MTAG-C; corrected). Since MTAG-U is anti-conservative and MTAG-C is conservative, they can be treated as the upper and lower bounds of the MTAG's

power. I implemented the MTAG method using test code written in Python because The MTAG software thought the input was defective if the median z-score was far from zero, such as the input used in the power simulation.

I evaluated the power of PLEIO, MTAG-U, MTAG-C, ASSET, and METAL using various simulation settings. For each simulation setting, I defined a specific genetic correlation structure ($\boldsymbol{C_g}$), heritability ($\boldsymbol{h^2}$), phenotypic unit ($\boldsymbol{U}$), and trait type (quantitative; Q or binary; B). In the power simulation, $\boldsymbol{C_g}$ and $\boldsymbol{h^2}$ given to PLEIO and MTAG are not estimates but true genetic correlation and heritability. I assumed the seven traits ($T = 7$) and repeated the simulation $10{,}000$ times. For each method, the statistical power was estimated as the proportion of simulations with estimated $P < 5 \times 10^{-8}$. In this power simulation, instead of directly sampling the effect sizes from a multivariate distribution, I generated the actual genotypes (See **Generation of effect sizes used in power simulation**).

First, I assumed a fixed heritability of 0.4 and perfect correlation ($r^2 = 1.0$) across seven traits. This represents the scenario that collects the multiple GWAS of the same traits. In this situation, PLEIO, METAL, MTAG-U performed better than MTAG-C and ASSET (**Figure 9a**). With a sample size of $N = 50{,}000$, the power of PLEIO, METAL, MTAG-U, MTAG-C, and ASSET were 63.79%, 63.81%, 63.81%, 63.81%, and 61.67%. As expected, METAL performed well because it is optimized to aggregate multiple GWAS with the same trait. MTAG-U and METAL are analytically identical[19]; therefore, MTAG-U performed the same as the METAL. PLEIO attained similar (or slightly less) power of METAL and MTAG-U

as it can account for the genetic correlations. In this scenario with one trait, the multiple testing correction using Bonferroni is not necessary for MTAG-U. Because of this, the power of MTAG-C was overly conservative.

Second, I changed the heritability for seven traits from 0.005 to 0.7. I assumed a uniform genetic correlation of $r = 0.5$ of all trait pairs. In this scenario, PLEIO outperformed other methods (**Figure 9b**). With a sample size of $N = 50{,}000$, PLEIO attained a power of 77.6%, while the second-best method (MTAG-U) attained 67.2%, and the third-best method (MTAG-C) attained 62.7%. The result indicates that PLEIO is optimized for a joint analysis of multi-trait with different heritabilities.

Third, I simulated a complex genetic correlation pattern with both positive and negative correlations. I divided seven groups into two groups (three traits and four traits). I set the within-group correlation of the first group to 0.95 and the second group to 0.9, and I set the correlation between groups to -0.9. I assumed a uniform heritability of 0.4 for all traits. PLEIO outperformed other methods (**Figure 9c**). With a sample size of $N = 50{,}000$, PLEIO attained a power of 78.6%, while the second-best method (MTAG-U) attained 66.3%, and the third-best method (MTAG-C) attained 62.6%. The result indicates that PLEIO is optimized for a joint analysis of multi-trait with a complex correlation pattern.

Fourth, I simulated a mixture of quantitative and binary traits. I assumed four quantitative traits and three binary traits. For quantitative traits, I assumed that

phenotypic units could differ between traits. When U is the standard phenotypic unit I assumed, I changed the units of four traits from $0.1U$ to $10U$. I assumed a uniform heritability of 0.4 and a uniform genetic correlation of 0.5. Again, PLEIO outperformed other methods (**Figure 9d**). With a sample size of $N = 50{,}000$, PLEIO attained a power of 80.1%, while the second-best method (MTAG-U) attained 63.4%, and the third-best method (MTAG-C) attained 57.3%. The result indicates that PLEIO systematically combines heterogeneous traits by standardizing the effect sizes.

So far, I tested the power by changing one factor per simulation: different heritabilities, a complex genetic correlation pattern, different phenotypic units. In a real data analysis, all three can occur together. I tested such a combined situation (**Figure 9e**). With a sample size of $N = 50{,}000$, PLEIO attained a power of 49.2%, while the second-best method (MTAG-U) attained 59.3%.

Next, I wanted to test a power simulation using real data-based parameters. In this simulation of seven studies, I assumed one focal trait and six non-focal traits where the focal trait shows strong genetic correlations with the non-focal traits. Here, I assumed that MTAG could selectively take the p-values of the focal trait only, which I call MTAG-F.

Based on the information provided by LD-HUB[32], I chose LDL as the focal trait and selected six traits that are strongly correlated to LDL $(0.35 \geq |r_g| \geq 0.17)$: triglyceride (TG), coronary artery disease (CAD), Age at Smoking (Age_Smo),

childhood IQ (cIQ), Hemoglobin A1c (HbA1C), and Waist-Hip-Ratio (WHR). For simplicity, I assumed that all seven traits share 1000 causal variants. Unlike MTAG-F, PLEIO and MTAG-U can have strong associations driven by one or some non-focal traits with the large $\boldsymbol{h^2}$ if I assume the same sample size. To compensate for this difference in heritability, the samples sizes were adjusted so that $\boldsymbol{Nh^2}$ is constant for all traits. Then, I doubled the sample size of the focal trait.

Figure 10 shows the result of the power simulation above. Again, PLEIO outperformed other methods. With sample sizes that meet $\boldsymbol{Nh^2} = 10{,}000$, PLEIO attained a power of 72.6%, while the second-best method (MTAG-U) attained 52.8%, and the third-best method (ASSET) attained 37.3%. Note that MTAG-F is a trait-specific method, and the interpretation is different for MTAG-F than other methods. Therefore, a careful interpretation is required for other methods before concluding that the focal trait drives the association.

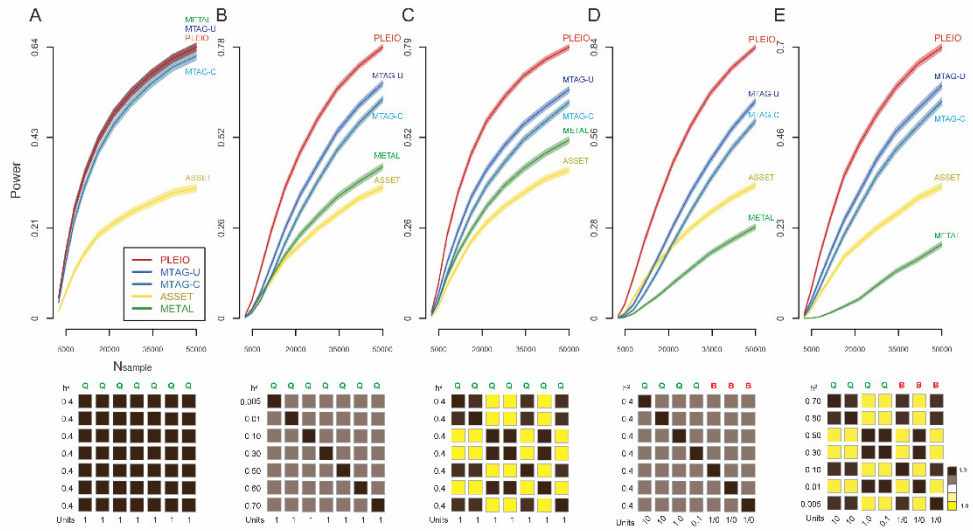


Figure 9 The results of the power test. I performed a total of five power tests. Each line shows the statistical power of a model gained from an association test using seven summary statistics: PLEIO (red), MTAG-U (blue), MTAG-C (light blue), METAL (green), and ASSET (yellow). At the bottom of the figure, I visualized the simulation setting of each test. The box plot shows the genetic correlation. Q and B indicate whether the phenotype is quantitative or binary. The heritability values of the traits are shown on the left side of the boxplot. The trait phenotype units are shown at the bottom of the box plot. The line thickness indicates the 95% confidence interval.

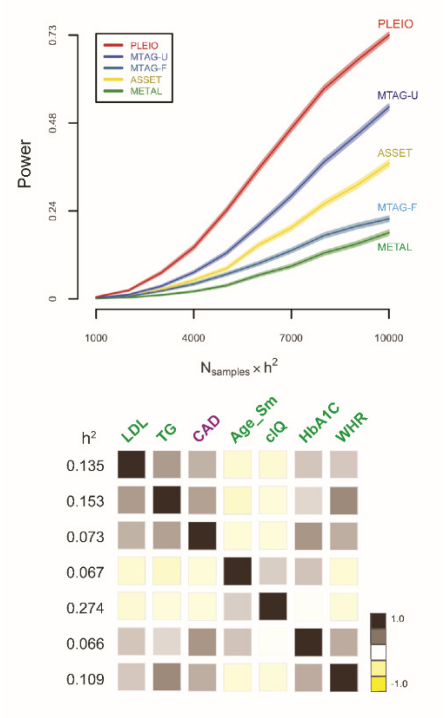


Figure 10. Power test results assuming LDL as the focal trait. Each line shows the statistical power of a model gained from an association test using seven summary statistics: PLEIO (red), MTAG-U (blue), MTAG-F (light blue), METAL (green), and ASSET (yellow). Note that the x-axis is the product of the sample number and heritability. For example, the number of samples of a trait with a heritability value of 0.1 is $100{,}000$ for $N\boldsymbol{h}^2 = 10{,}000$ and $40{,}000$ for $N\boldsymbol{h}^2 = 4{,}000$. At the bottom of the figure, I visualized the simulation setting of each test. The box plot shows the genetic correlation. The color of the trait name indicates whether the phenotype is quantitative (green) or binary (purple). Since the focal trait is the main interest of this analysis, I assumed that the focal trait collected twice as many samples as a non-focal trait. In other words, a point of the focal trait in the x-axis means $\frac{N\boldsymbol{h}^2}{2}$.

3.4 Measuring computation time and memory usage

Here, I compared the computation time and maximum memory usage of PLEIO, MTAG, ASSET, and METAL. I assumed the simulation setting used in the focal-trait power simulation above. The source codes of MTAG and METAL are implemented in test codes and used for the simulation. For importance sampling, I used $N_{sample} = 100K$. I generated two sets of simulation inputs for $10K$ and $1M$ association tests and tested each method with one CPU.

Table 5 shows that all methods except ASSET can perform $1M$ association tests in an hour with less than $4\ Gb$ of free memory usage.

	10K association tests					1M association tests				
	PLEIO		MTAG	ASSET	METAL	PLEIO		MTAG	ASSET	METAL
	Prep. null. dist	assoc. tests				Prep. null. dist	assoc. tests			
Total analysis time using 1 CPU (sec)	1125.04	45.48	8.8	2150.7	3.71	1125.04	1607.1	967.6	N/A	46.6
Maximum memory usage	177.7 Mb		108 Mb	80.4 Mb	73.6 Mb	1.02 Gb		2.8 Gb	N/A	0.7 Gb

Table 5. Comparison of the computational efficiency of PLEIO, MTAG, ASSET, and METAL. I measured the runtime and maximum memory usage of each method required to perform $\mathbf{10}\boldsymbol{K}$ and $\mathbf{1}\boldsymbol{M}$ association tests with one CPU. We used our own implementation of MTAG (python) and METAL (R). For PLEIO, we distinguished the time used for the association test (assoc. tests) and the time used for the importance sampling procedure (Prep. null. dist). By default, the importance sampling procedure performs $\mathbf{100,000}$ association tests to compute the probability distribution of the PLEIO statistics under the null hypothesis.

3.5 Joint analysis of multiple traits related to cardiovascular disease

I used PLEIO to identify pleiotropic loci of cardiovascular disease (CVD) related traits. To this end, I reviewed several GWAS consortia open to the public and collected 18 GWAS summary statistics on disease status and complex traits (**Table 1**). I reviewed the UK Biobank GWAS results of Neale Lab and selected twelve binary traits that included one or more of the following keywords: heart, hypertension, obesity, lipoprotein, cholesterol, and diabetes (**Table 2**)[28]. I selected four lipid traits from the Global Lipid consortium[27], one binary trait (coronary artery disease) from CARDIoGRAM+C4D consortium[2], and one quantitative trait (fasting glucose) from the MAGIC (Meta-Analysis of Glucose and Insulin-related traits Consortium)[29]. As a result, I collected a total of 13 binary traits and five quantitative traits. See **Data collection** for details of the trait selection procedure. For quantitative traits, I found differences in the phenotypic units. For example, Lipid traits had the unit of mg/dl, whereas the fasting glucose uses the unit of mmol/l[27, 29]. Below, I tested SNP associations of 1,777,412 SNPs shared by these 18 summary statistics. In the pre-analysis phase with LDSC, 18 traits showed differing heritabilities and non-zero environmental and genetic correlations (**Figure 11**).

I perform SNP association tests of 1,777,412 SNPs using PLEIO and identified 625 independent GWAS hits that exceeded the threshold $P < 5 \times 10^{-8}$ (**Figure 12**). Among those, I found 13 independent novel variants, each of which locus was not listed in the GWAS catalog and was not identified by one of the 18 GWAS (**Table**

6). **Figure 13** shows local Manhattan plots of these 13 variants. **Figure 12a** shows a circular plot whose radial position denotes the genomic position and heights of points denote the statistical significance of variants. **Figure 12b** shows the genome-wide Manhattan plot of the association test results. Next, I compared the results of PLEIO to input summary statistics using a mirrored Manhattan plot in **Figure 14**. Finally, I applied LDSC to the association results of PLEIO and estimated the LDSC intercept ($\alpha = 1.11$) to see if PLEIO's log-likelihood statistics had systematic inflation.

To investigate the biological role of these identified variants, I conducted a functional analysis using Variant Effect Predictor (VEP v.97.2) in ENSEMBL GRCh37[33]. The 13 novel variants included six intronic variants, three non-coding transcript variants, three intergenic variants, one upstream gene variant (**Table 7**). The 625 top hits included 374 intronic variants, 112 intergenic variants, 41 upstream gene variants, 25 downstream variants, 23 missense variants, 21 3-prime UTR variants, 12 non-coding transcript exon variants, 12 synonymous variants, and five 5-prime UTR variants.

I did additional analysis on 625 top hits using DAVID v.6.8[34]. Here, the gene list obtained from the VEP was used as an input for DAVID to search for the existence of known trait-gene associations in the Genetic Association Database (GAD). I curated the results to get eight categories of traits: coronary artery disease, fasting glucose, high blood pressure, diabetes, high-density lipoprotein, low-density lipoprotein, total cholesterol, and total glycerides. In other words, I obtained eight

sets of genes where each gene set corresponds to a trait above. Finally, I visualized the results in the circular ribbon plot in **Figure 12a**. Each ribbon represents a pair of genes in the same phenotypic category.

I performed an additional real data analysis using the same data for MTAG (**Table 8**). Since MTAG produced as many p-values as the number of studies per SNP, I converted MTAG results to MTAG-U and MTAG-C and compared the results to PLEIO as in the power test. As a result, MTAG-U found 622 independent GWAS top hits variants, slightly fewer than the 625 variants found by PLEIO. As explained earlier, MTAG-U is one method of selecting the minimum p per SNP, which leads to multiple testing problems. Applying LDSC to MTAG-U confirmed the strong inflation in the LDSC intercept (α = 3.89). Next, I compared MTAG-C and PLEIO using Bonferroni correction on MTAG-U. MTAG-C found 493 GWAS top hits. In addition, as an alternative to solve the multiple testing problem in MTAG-U, I corrected the chi-square statistics χ^2 of MTAG-U so that the LDSC-intercept estimate is the same as 1.10, which is the LDSC intercept estimated from PLEIO. I referred to the approach as MTAG-$\boldsymbol{\alpha}$ and compared MTAG-$\boldsymbol{\alpha}$ with other approaches. The number of GWAS top hits obtained from MTAG-α was only 102, confirming that using the LDSC section did not solve the multi-test problem well.

I measured the computation time and maximum memory usage needed for this real data analysis using a single CPU core. The estimated time to run single-trait LDSC analysis took 0.2 hours and to run pairwise LDSC analysis for $\binom{18}{2}$ pairs took 1.5 hours. Estimation of null distribution took 1.89 hours. Lastly, the association test

analysis of 1,777,141 SNPs took 1.83 hours. In total, PLEIO required 3.72 hours, excluding LDSC preprocessing using 2.1 GB memory at peak.

In this real data analysis, I assumed that the samples used in each UKB GWAS represented the European population and used the frequency of the case sample among the total sample as both the sample prevalence and the disease prevalence. However, the UKB cohort consists of the individuals who volunteered for this study, not a random sampling process. Therefore, we cannot say that the cohort represents the European population. For some traits (e.g., coronary atherosclerosis, obesity, etc.), there is a possibility that an individual's phenotype has not yet been explicitly diagnosed or expressed, in which case a control sample may later turn out to be a case sample. In this case, the disease prevalence and sample prevalence are not the same as the frequency of the case samples.

To determine whether the issues described above influenced the results of the real data analysis, I performed an additional analysis. In this analysis, I conducted a literature review and updated the disease prevalence of the 13 UKB binary traits (see **Table 9**). I then completed the real data analysis again and compared the PLEIO's p-values of the 13 novel loci between the two analyses (old and new). **Table 10** shows the resulting PLEIO's p-values for 13 novel pleiotropic variants obtained from the two analyses, and I confirmed that there was no significant difference between the p-values.

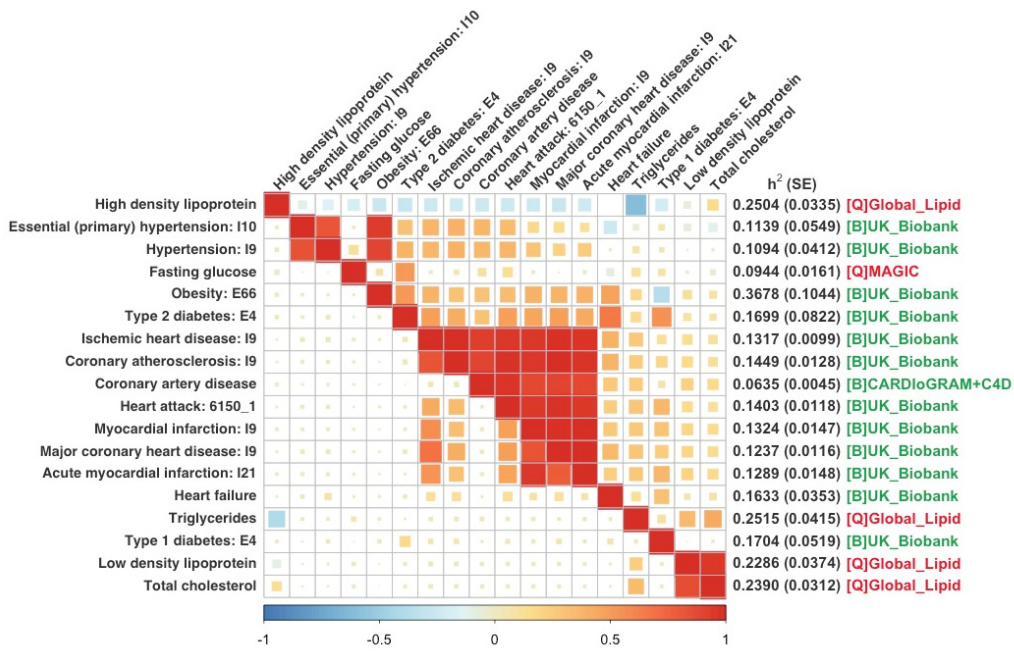


Figure 11. Genetic correlation and environmental correlation among 18 traits. The 18×18 matrix shows the genetic (upper triangular) and environmental (lower triangular) correlations among 18 traits. The labels on the left and the top are the names of the traits, and the labels on the right indicate the heritabilities along with the names of the database and the types of the phenotypes (green: binary, red: quantitative).

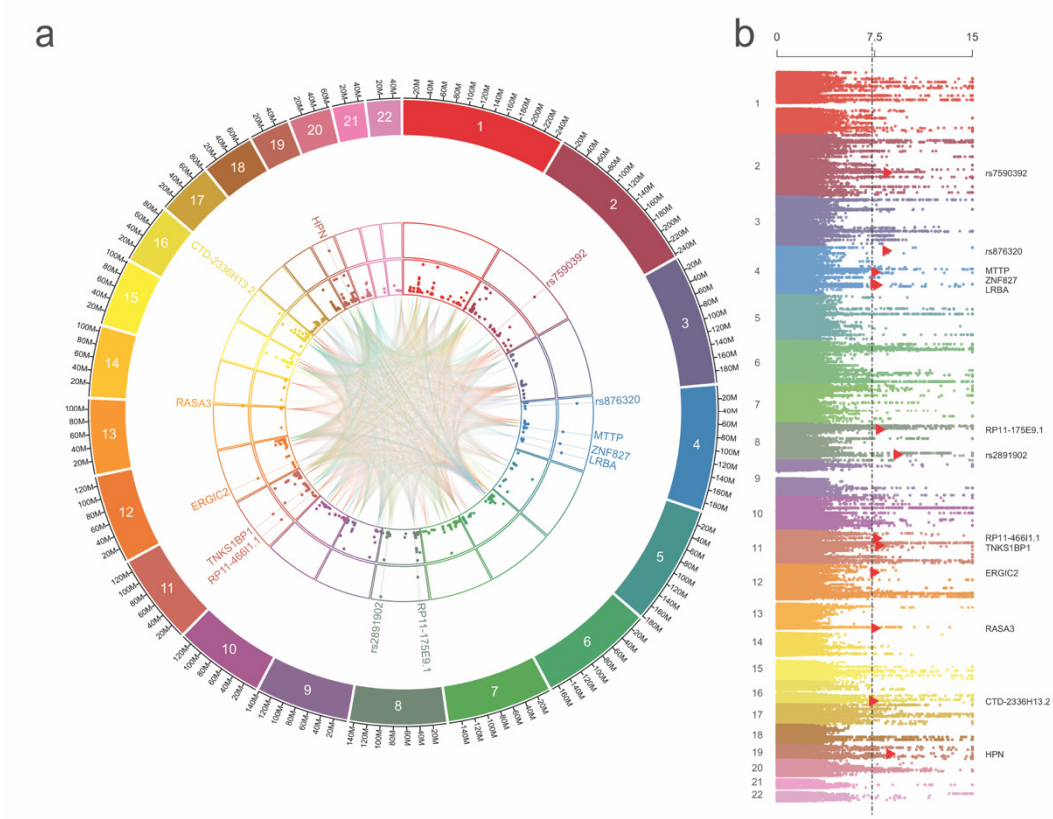


Figure 12. The summary of the real data analysis. a. The circular plot shows the locations and the statistical significances of the 13 novel variants (outer edge) and the 625 GWAS top SNPs (inner edge). The inner ribbons connect the variants in the same functional category found by the DAVID analysis. b. The Manhattan plot of the PLEIO association results. Red triangles indicate the 13 novel loci.

SNP	CHR	BP	A1	A2	S_{PLEIO}	P_{PLEIO}	HGNC Symbols
rs7590392	2	148379602	T	C	32.46	3.69E-09	
rs876320	4	15930961	A	G	32.06	4.38E-09	
rs7693203	4	100500130	T	C	27.77	3.27E-08	MTTP
rs1979974	4	146800815	G	A	27.5	3.76E-08	ZNF827
rs6817572	4	151303318	A	G	28.51	2.24E-08	LRBA
rs1561105	8	23610799	G	T	29.43	1.43E-08	RP11-175E9.1
rs2891902	8	122422130	C	T	36.36	6.18E-10	
rs2055014	11	29195732	A	G	28.38	2.39E-08	RP11-466I1.1
rs12787728	11	57069056	G	A	29.25	1.56E-08	TNKS1BP1
rs2278093	12	29534209	A	C	27.48	3.80E-08	ERGIC2
rs4393438	13	114821075	C	T	27.76	3.28E-08	RASA3
rs1039119	16	76946526	T	C	27.08	4.71E-08	CTD-2336H13.2
rs1688030	19	35556744	C	T	33.53	2.30E-09	HPN

Table 6 The summary of 13 NOVEL GWAS hits identified by PLEIO. S_{PLEIO} column contains the PLEIO's log-likelihood ratio test statistics; P_{PLEIO} column contains PLEIO's p-values. HGNC Symbols column gives the human genome nomenclature (HGNG) gene names of the genetic loci associated with the 13 novel variants. SNP denote rsID of SNPs, CHR denote chromosome number, BP denote a base position, A1 denote risk allele, A2 denote reference allele.

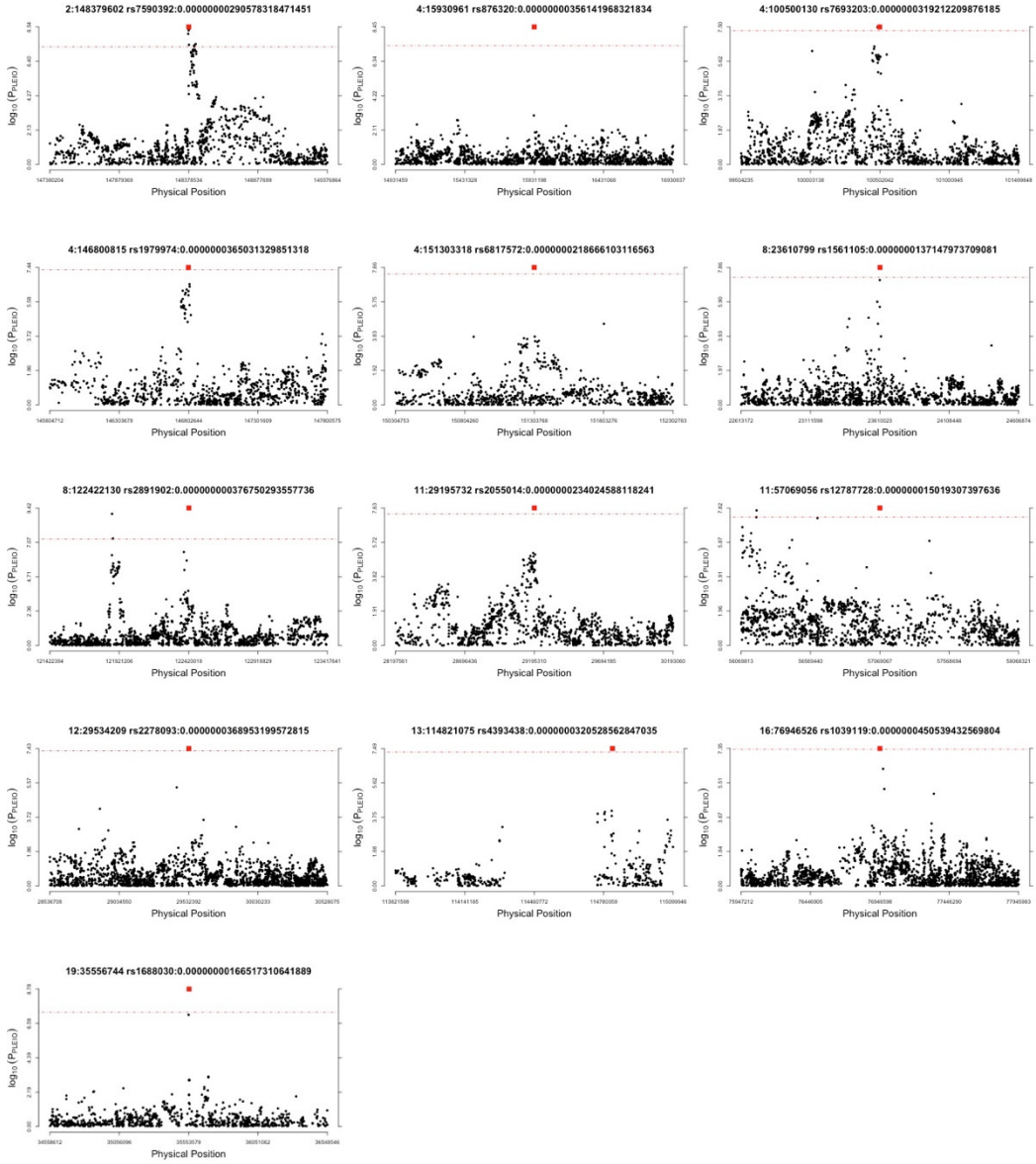


Figure 13. Local Manhattan plots of the 13 novel loci identified by PLEIO.

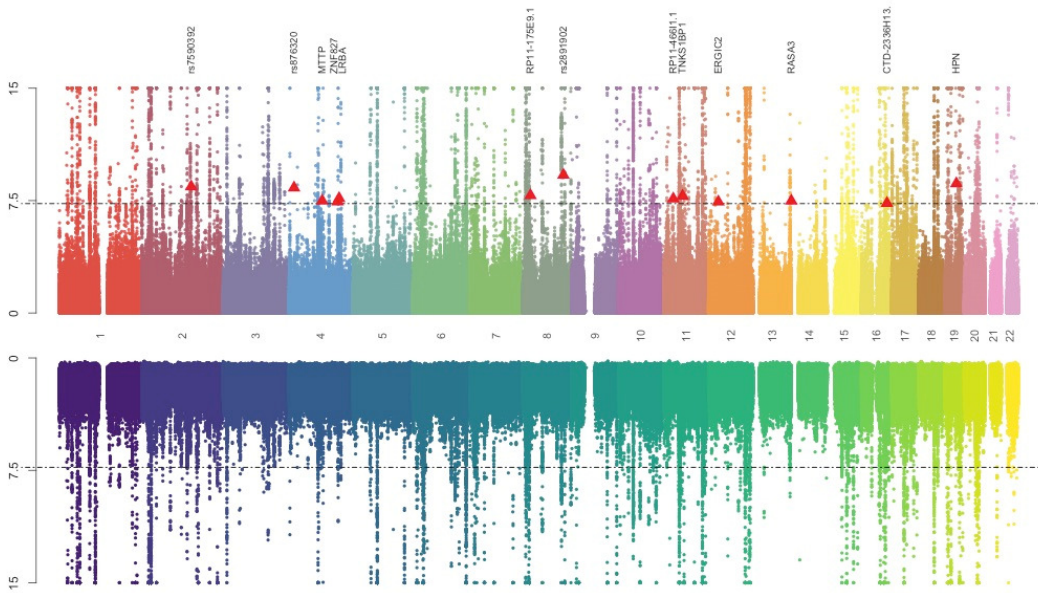


Figure 14. Manhattan plots showing the association analysis results using real data. The top shows the PLEIO's p-values, and the bottom shows the minimum p-values of 18 summary statistics included in the PLEIO analysis. I set the maximum value of the -log(p) to 15.

Location (CHR: BP)	Allele (Ref/Risk)	HGNC gene symbol	VEP Consequence	Alias	Cellular location (3-5; confidence level)	Known function	GWAS catalog (S):$p < 5 \times 10^{-8}$ (W):$p > 5 \times 10^{-8}$
2:148379602	C/T		Intergenic				
4:15930961	G/A		Intergenic				
4:100500130	C/T	MTTP	Intronic	Microsomal Triglyceride Transfer Protein	Endoplasmic reticulum(5) Golgi apparatus(5) Plasma membrane(4) Extracellular (3)	Required for the assembly and secretion of plasma lipoproteins that contain apolipoprotein B	triglyceride measurement(S), high-density lipoprotein cholesterol measurement(S)
4:146800815	A/G	ZNF827	Intronic	Zinc Finger Protein 827	Nucleus(4)	May be involved in transcriptional regulation	sleep duration, low-density lipoprotein cholesterol measurement(S), coronary artery disease(S)
4:151303318	G/A	LRBA	Intronic	Lipopolysaccharide-Responsive And Beige-Like Anchor Protein	Plasma membrane(5) Cytosol(5) Endoplasmic reticulum(4) Lysosome(4) Golgi apparatus(4) Nucleus(3)	May be involved in coupling signal transduction and vesicle trafficking to enable polarized secretion and/or membrane deposition of immune effector molecules.	systolic blood pressure(S), peripheral arterial disease, traffic air pollution measurement(W),
8:23610799	T/G	RP11-175E9.1	Intronic non_coding_transcript	Antisense RNA			
8:122422130	T/C		Intergenic				
11:29195732	G/A	RP11-466I1.1	Intronic non_coding_transcript	LincRNA			
11:57069056	A/G	TNKS1BP1	Intronic	Tankyrase 1 Binding Protein 1	Cytoskeleton(5) Nucleus(5) Cytosol(5) Plasma membrane(3)	Deadenylation-dependent mRNA decay	apolipoprotein A 1 measurement(S)

12:29534209	C/A	ERGIC2	Upstream	ERGIC And Golgi 2	Nucleus(5) Golgi apparatus(5) Endoplasmic reticulum(4) Plasma	Possible role in transport between endoplasmic reticulum and Golgi.	low-density lipoprotein cholesterol measurement(S)
13:114821075	T/C	RASA3	Intronic	RAS P21 Protein Activator 3	plasma membrane(5) cytosol(5)	Inhibitory regulator of the Ras-cyclic AMP pathway.	Lymphocyte percentage of leukocytes(S), monocyte count(S)
16:76946526	C/T	CTD-2336H13.2	Intronic non_coding_transcript	LincRNA			
19:35556744	T/C	HPN	Intronic	Hepsin	plasma membrane(5), extracellular(5)	Plays a role in cell growth and maintenance of cell morphology	triglyceride measurement(S)

Table 7. The functional analysis of the 13 GWAS novel hits using ENSEMBL VEP[33], Gene Cards[35], and GWAS catalog[8].

The 13 variants are in ascending order by chromosomal position and then in ascending order according to by genomic position (BP)

	PLEIO	MTAG-U	MTAG-C	MTAG-λ
#independent GWAS hits	625	622	493	102
intercept (LDSC)	1.109	3.893	0.556	1.113
λ_{GC}	1.5883	8.456	0.0367	2.413

Table 8. Comparison of the number of GWAS-TOP hits of PLEIO and MTAG identified in post GWAS analysis. I applied real data of CVD-related traits to PLEIO, MTAG-U, MTAG-C, and MTAG-λ and identified GWAS top hits. I obtained the intercept of the LDSC heritability estimate and the genomic inflation factor (λ_{GC}) from the output of LDSC software. MTAG-U: Selecting minimum p-value among the multi-trait MTAG p-values; MTAG-C: Bonferroni correction applied to MTAG-U; MTAG-λ: Intercept correction applied to MTAG-U so that intercept can be comparable to PLEIO.

Trait name	Sprev	Pprev	Pprev_literature	phenotypic type	Database
Heart attack: 6150_1	0.023533	0.023533	0.043	Binary	UK Biobank
Hypertension: I9	0.003433	0.003433	0.436	Binary	UK Biobank
Essential (primary) hypertension: I10	0.002397	0.002397	0.4033	Binary	UK Biobank
Acute myocardial infarction: I21	0.018070	0.018070	0.043	Binary	UK Biobank
Myocardial infarction: I9	0.019815	0.019815	0.043	Binary	UK Biobank
Major coronary heart disease: I9	0.028931	0.028931	0.07	Binary	UK Biobank
Ischemic heart disease: I9	0.061280	0.061280	0.035	Binary	UK Biobank
Coronary atherosclerosis: I9	0.041322	0.041322	0.485	Binary	UK Biobank
Heart failure	0.003902	0.003902	0.06	Binary	UK Biobank
Obesity: E66	0.001222	0.001222	0.424	Binary	UK Biobank
Type 1 diabetes: E4	0.001616	0.001616	0.095	Binary	UK Biobank
Type 2 diabetes: E4	0.002464	0.002464	0.0628	Binary	UK Biobank
Coronary artery disease	0.492299	0.05	0.05	Binary	CARDIo+C4D
High density lipoprotein	NaN	NaN	NaN	Quantitative	Global Lipids
Low density lipoprotein	NaN	NaN	NaN	Quantitative	Global Lipids
Total cholesterol	NaN	NaN	NaN	Quantitative	Global Lipids
Triglycerides	NaN	NaN	NaN	Quantitative	Global Lipids
Fasting glucose	NaN	NaN	NaN	Quantitative	Magic

Table 9. Disease prevalence of 13 UKB traits, updated based on a literature review. **Trait name**; the name of the trait, **Sprev**; sample prevalence used in the real data analysis, Pprev; disease prevalence used in the real data analysis, **Pprev_literature**; disease prevalence obtained from a literature search, **phenotypic type**; the type of the phenotype (either binary or quantitative), **Database**; The source of the GWAS summary statistics.

SNP	$P_{PLEIO,l}$	P_{PLEIO}
rs7590392	2.418961e-09	2.91E-09
rs1979974	2.500851e-08	3.65E-08
rs6817572	2.150352e-08	2.19E-08
rs12787728	1.627141e-08	1.50E-08
rs2278093	3.807202e-08	3.69E-08
rs1688030	1.823994e-09	1.67E-09
rs7693203	3.789116e-08	3.19E-08
rs4393438	3.281006e-08	3.21E-08
rs876320	3.071378e-09	3.56E-09
rs1561105	1.372890e-08	1.37E-08
rs2891902	5.758588e-10	3.77E-10
rs2055014	2.479782e-08	2.34E-08
rs1039119	6.511051e-08	4.51E-08

Table 10. Comparison of PLEIO p-value results for 13 new pleiotropic loci before and after adjusting for disease prevalence values. P_{PLEIO}; the p-value obtained from the real data analysis, $P_{PLEIO,l}$; The p-value obtained from the real data analysis to which the updated disease prevalence obtained from the literature search was applied.

3.6 Interpretation of the joint analysis results

I interpreted the 13 novel multi-trait associations with the visualization tool implemented in PLEIO, called the "*pleiotropy plot.*" The R package software can produce a circular plot, which gives a detailed summary of the association pattern of the pleiotropic variant. The outer plot area includes the local Manhattan plots and the bar plots of the standard effect sizes. The inner ribbons show the genetic correlations as colors and the locus heritability as widths. I drew pleiotropic plots of the 13 novel variants identified by PLEIO (**Figure 15**). Based on the association patterns observed in these plots, I divided these 13 novel variants into four groups without overlapping (**Figure 16**).

The first group of variants had associations with seven binary traits that include six traits (acute myocardial infarction, myocardial infarction, heart attack, major coronary heart disease, coronary atherosclerosis, and ischemic heart disease) from the UK Biobank and one trait from CARDIoGRAM+C4D. These seven traits showed high genetic correlations (**Figure 15**). The variants in this group showed the strongest association with one of the seven traits and had associations ($P < 0.001$) with at least three of the seven traits. The variants showing this pattern were rs7590392 near the *ACVR2A* gene (2q22.3) and rs1979974 in the *ZNF827* gene (4q31.22).

The second group of variants had an association with four lipid phenotypes (triglycerides, low-density lipoprotein; LDL, high-density lipoprotein; HDL, and total cholesterol). The variants in this group showed the strongest association with

one of four traits and had associations ($P < 0.001$) with at least two of the four traits. The variants showing this pattern were rs6817572 in the *LRBA* gene (6p22.3), rs12787728 in the *TNKS1BP1* gene (11q12.1), rs2278093 in the *ERGIC2* gene (12p11.22), and rs1688030 in the *HPN* gene (19q13.12). These variants were associated with some (but not all) of the lipid phenotypes. rs6817572 showed the strongest associations to the total cholesterol and LDL. rs12787728 showed the strongest associations to the total cholesterol and HDL. rs2278093 and 1688030 showed the strongest associations to the total cholesterol and triglycerides.

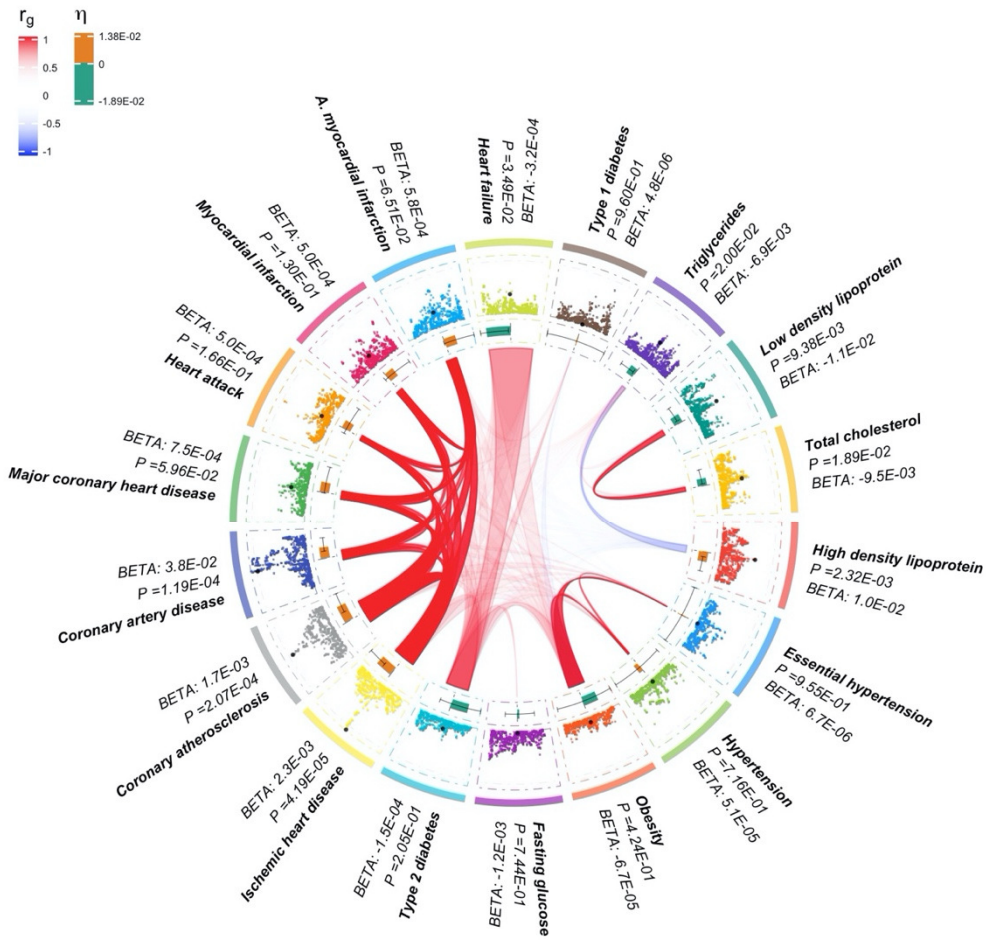
The third group of variants had associations with both coronary artery disease and lipid phenotypes. These variants in this group showed association ($P < 0.001$) with both coronary artery disease and one of the lipid traits at the same time. These variants met both the condition for group 1 and the condition for group 2, but I categorized them separately into the third group. The variants showing this pattern were rs7693203 in the *MTTP* gene (4q23) and rs4393438 in the *RASA3* gene (13q34). The variants in this group showed strong associations $(P < 0.0001)$ to the total cholesterol and LDL.

The fourth group of variants was a set of not categorized variants into the three aforementioned groups. The variants in this group were rs876320 near the *FGFBP1* gene (4p15.32), rs1561105 in the *RP11-175E9.1* gene (8p21.2), rs2891902 near the *RPL35AP19* gene (8q24.12), rs2055014 in the *RP11-466I1.1* gene (8q24.12), and rs1039119 in the *AC106729.1* gene (16q23.1). rs2891902 showed the strongest association to obesity $(P < 0.001)$ and weak associations to type 2 diabetes and

hypertension. Rs876320, rs1561105, and rs1039119 were interesting because their associations to all traits were weak $(P > 0.01)$. The strongest associations of rs1039119 were to coronary atherosclerosis $(P = 0.02)$ and triglycerides $(P = 0.08)$. However, this SNP's effect size directions to the seven binary traits in the first group were all concordant to the genetic correlations of these traits. The strongest associations of rs1561105 were to triglycerides $(P = 0.005)$ and major coronary heart disease $(P = 0.03)$, acute myocardial infarction $(P = 0.04)$, and myocardial infarction $(P = 0.05)$. This SNP's effect size directions to these three traits were all concordant to the genetic correlations. The strongest associations of rs876320 were to acute myocardial infarction $(P = 0.01)$, myocardial infarction $(P = 0.04)$, and heart attack $(P = 0.04)$. This SNP's effect size directions to these three traits were all concordant to the genetic correlations. Thus, PLEIO seems to have captured the aggregate information in multiple weak associations by considering the fact that the effect size directions were concordant to the genetic correlations. Follow-up studies will be needed to determine whether loci with weak associations for all traits are true associations or false positives.

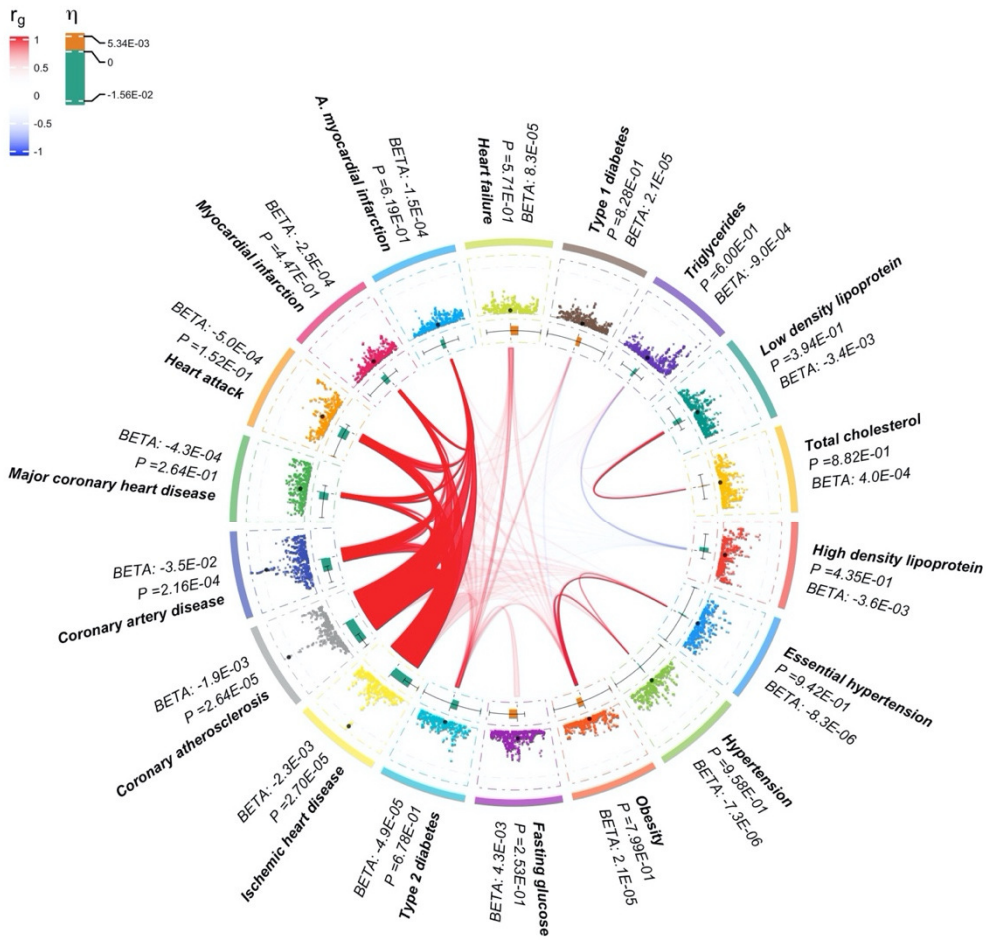
A. rs7590392.

[Group 1: Driven by seven binary traits]



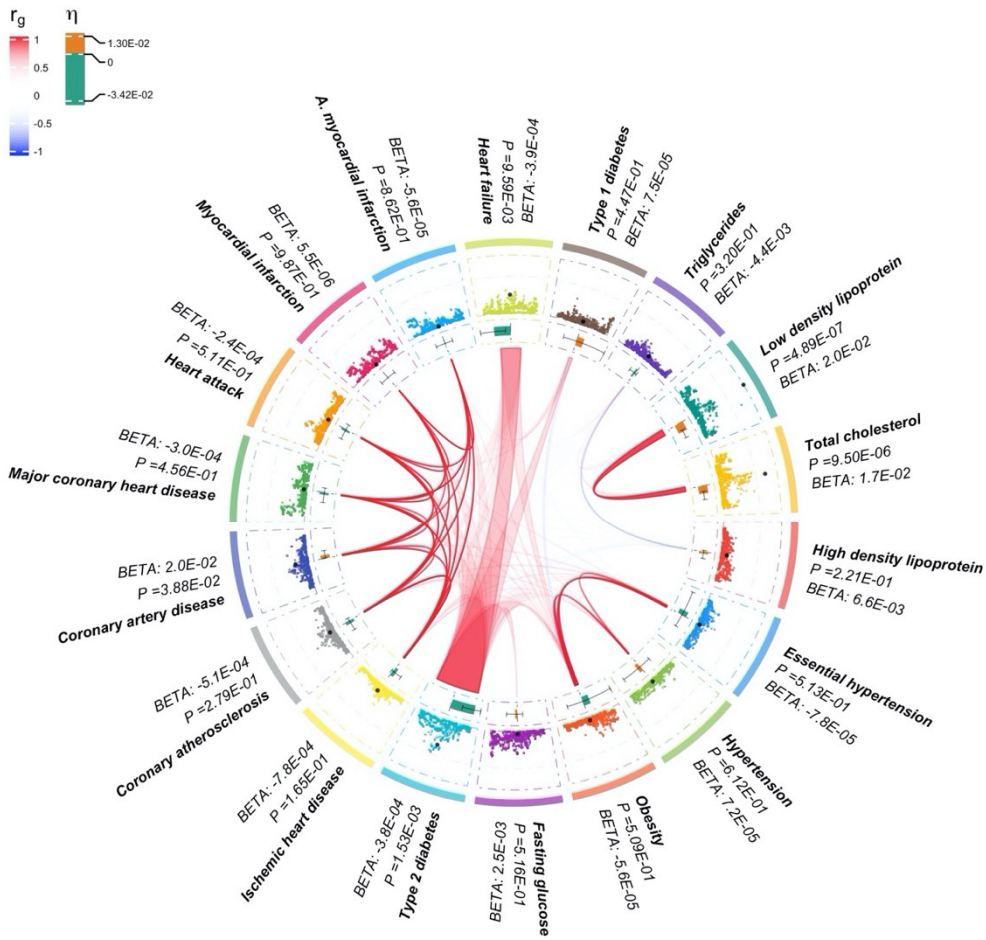
B. rs1979974

[Group 1: Driven by seven binary traits]



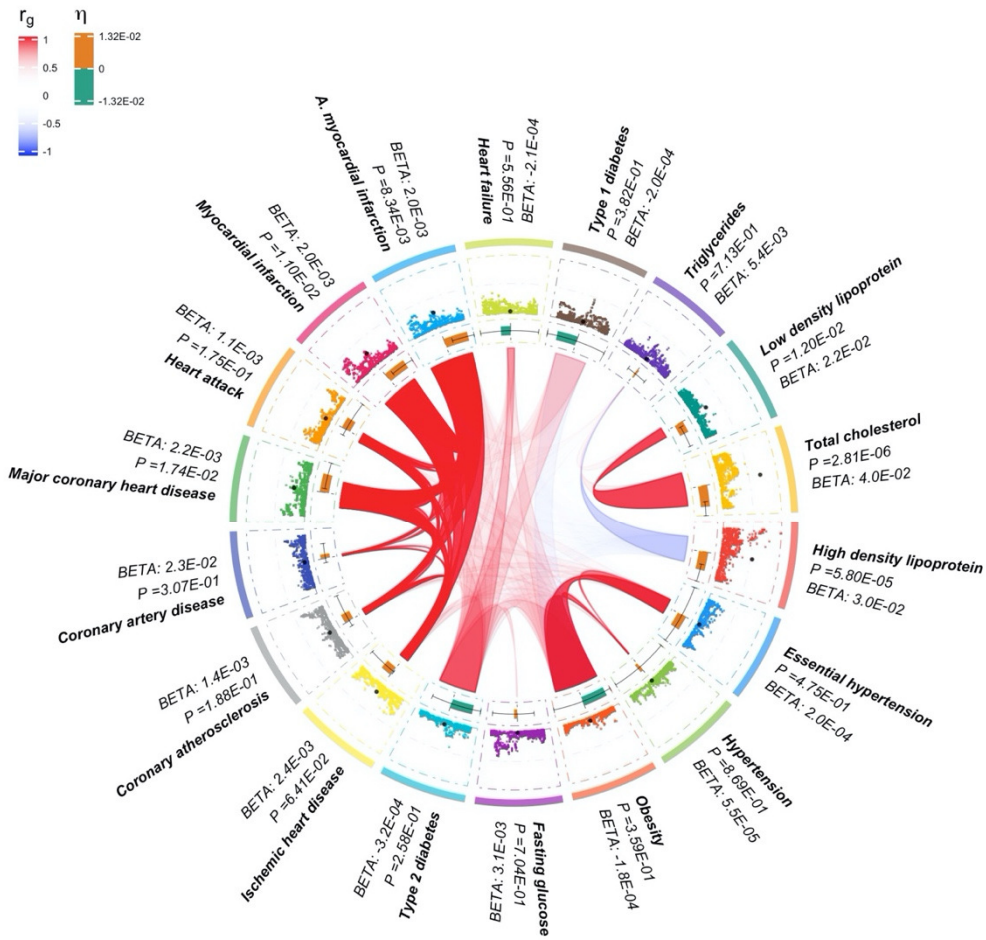
C. rs6817572.

[Group 2: Driven by lipid phenotypes (triglycerides, LDL, HDL, and total cholesterol)]



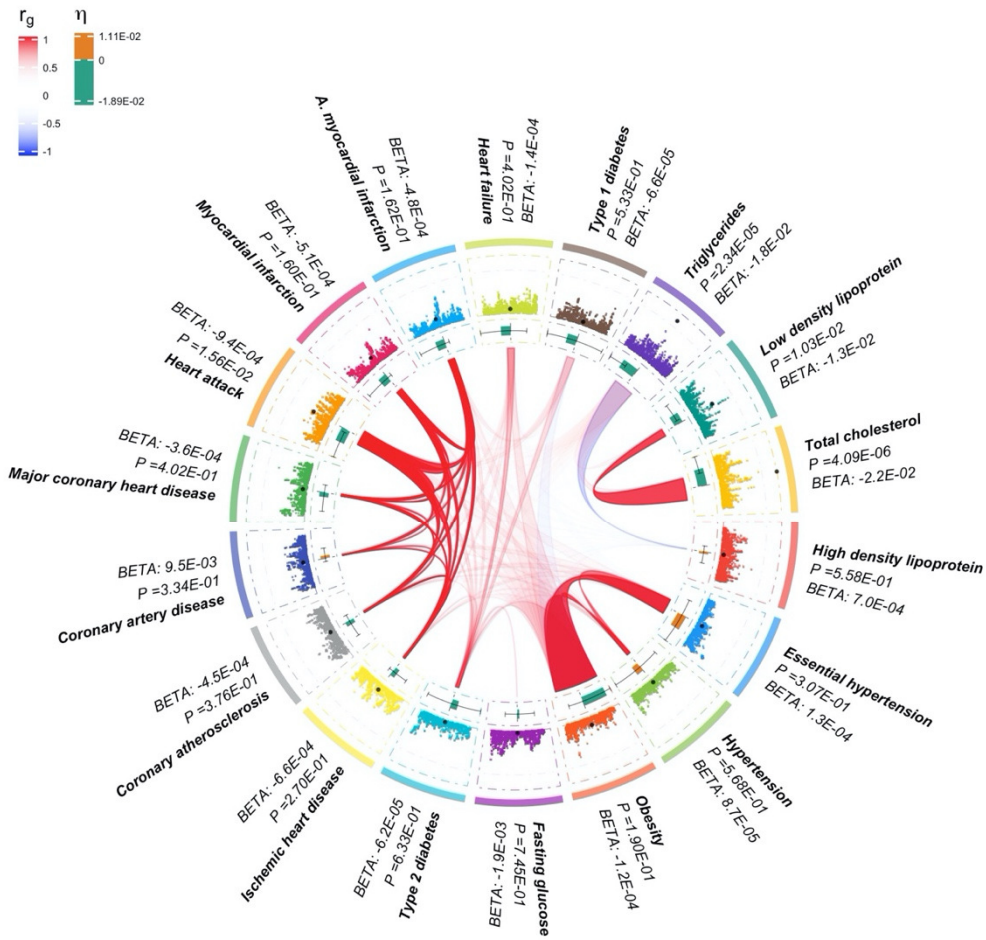
D. rs12787728

[Group 2: Driven by lipid phenotypes (triglycerides, LDL, HDL, and total cholesterol)]



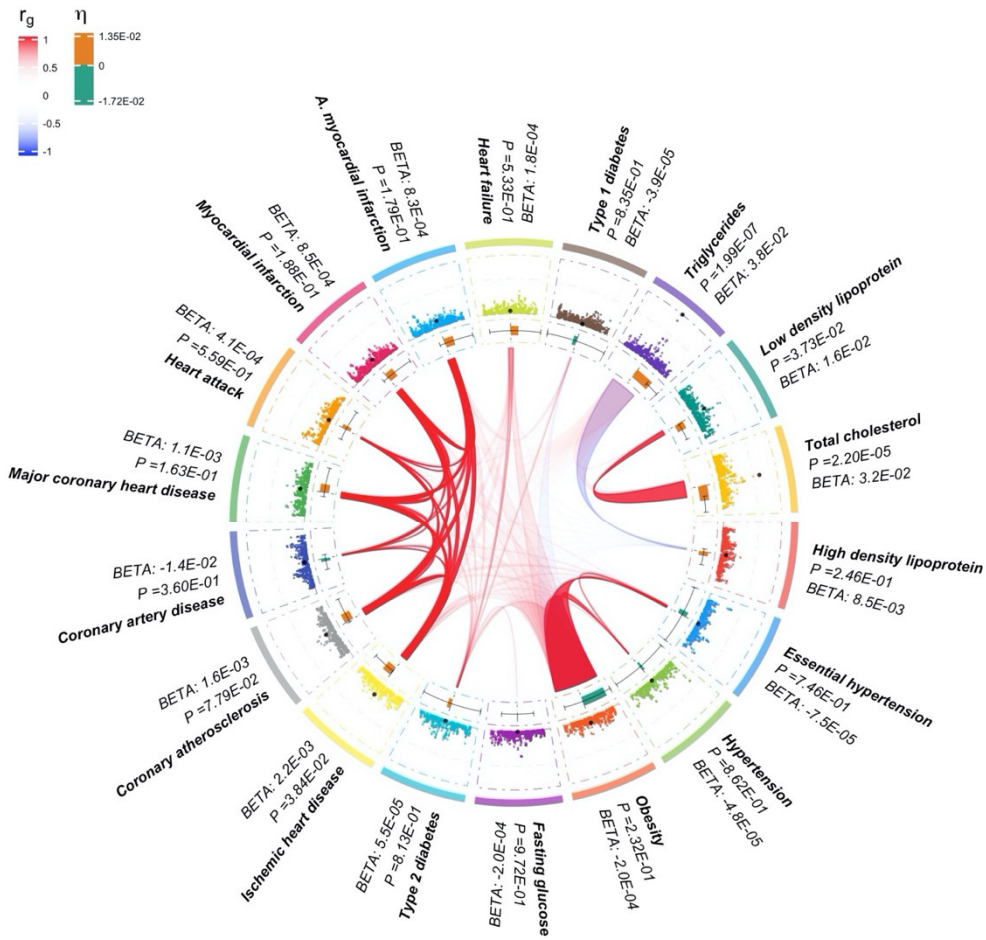
E. rs2278093.

[Group 2: Driven by lipid phenotypes (triglycerides, LDL, HDL, and total cholesterol)]



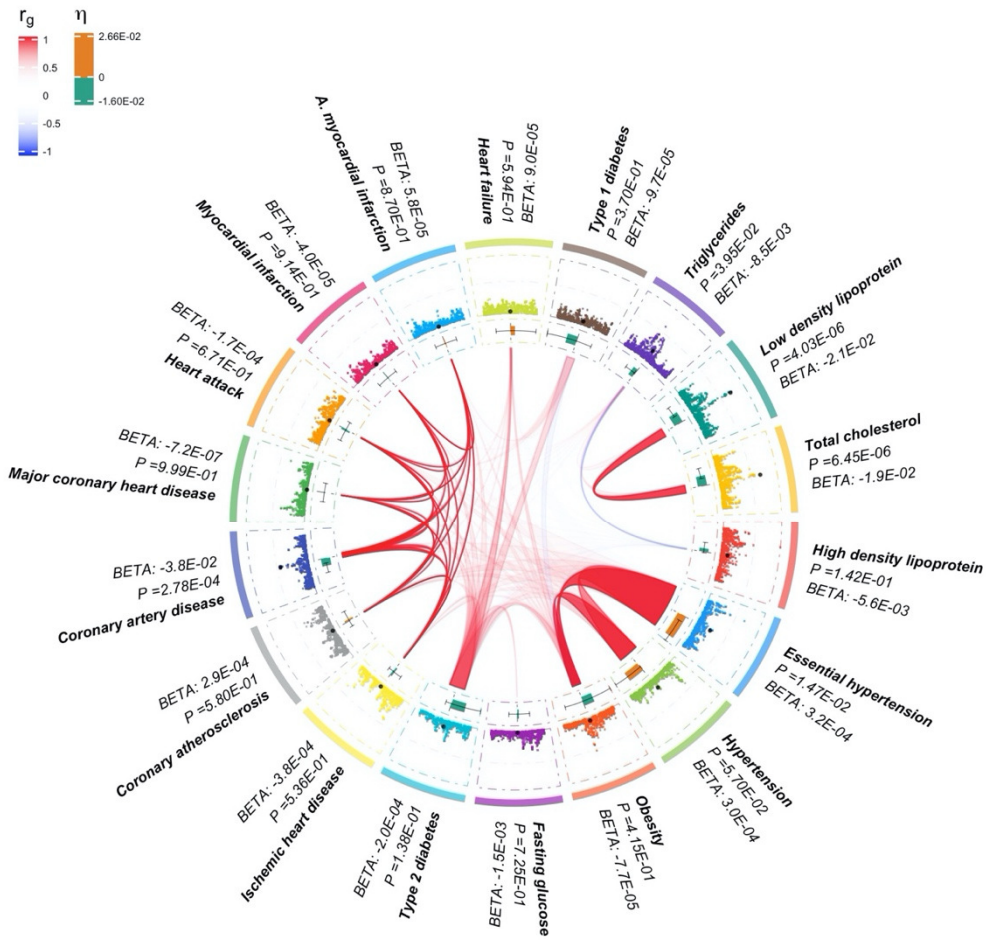
F. rs1688030

[Group 2: Driven by lipid phenotypes (triglycerides, LDL, HDL, and total cholesterol)]



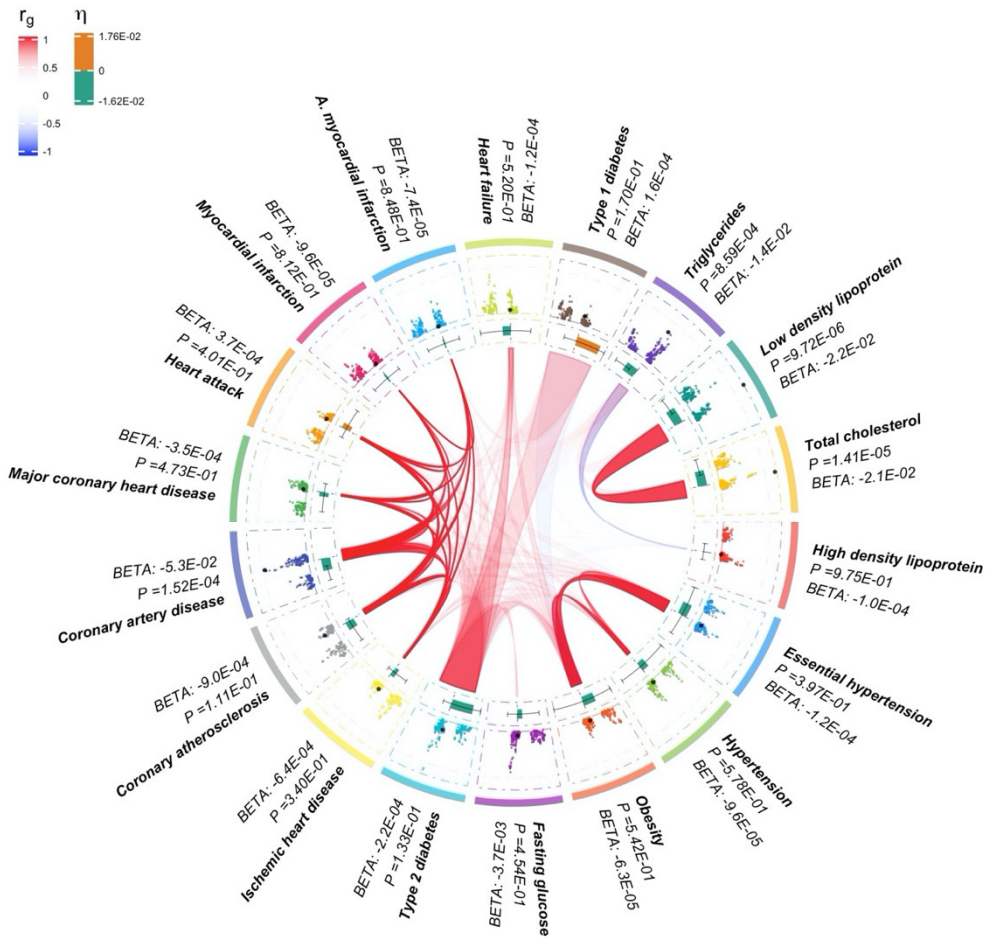
G. rs7693203.

[Group 3: Driven by both coronary artery disease and lipid]



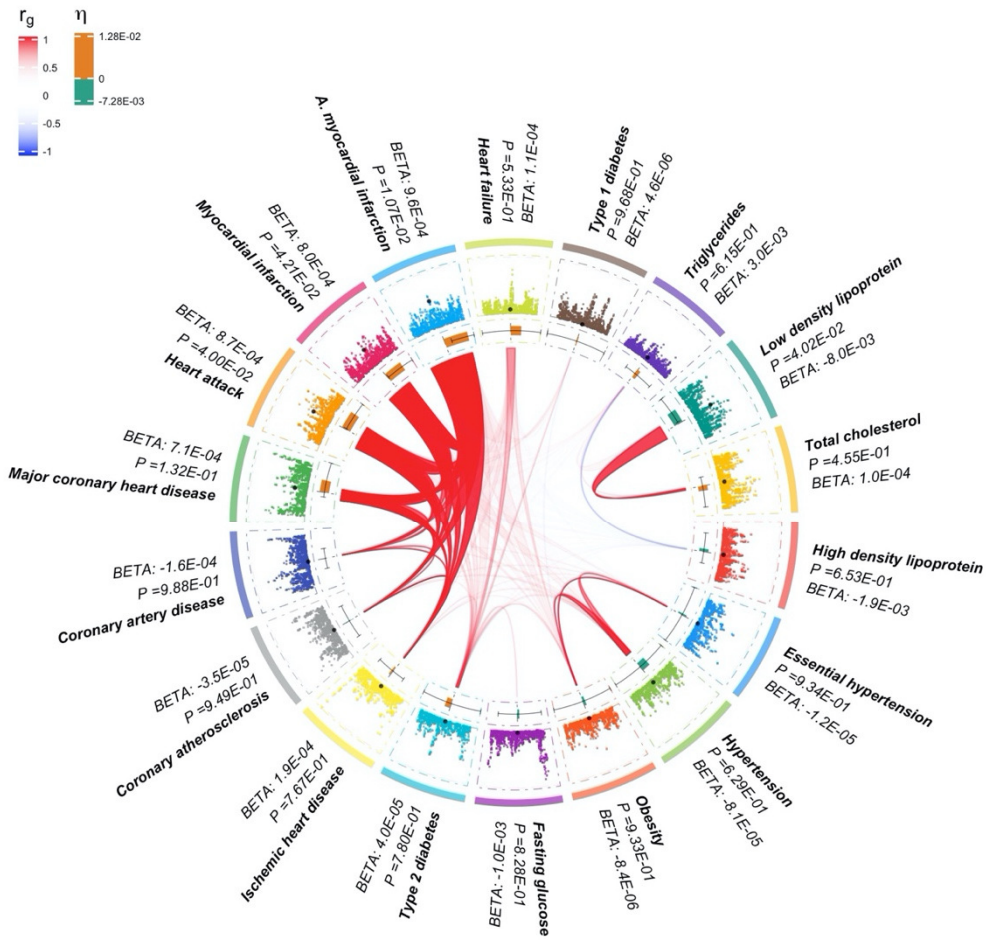
H. rs4393438.

[Group 3: Driven by both coronary artery disease and lipid]



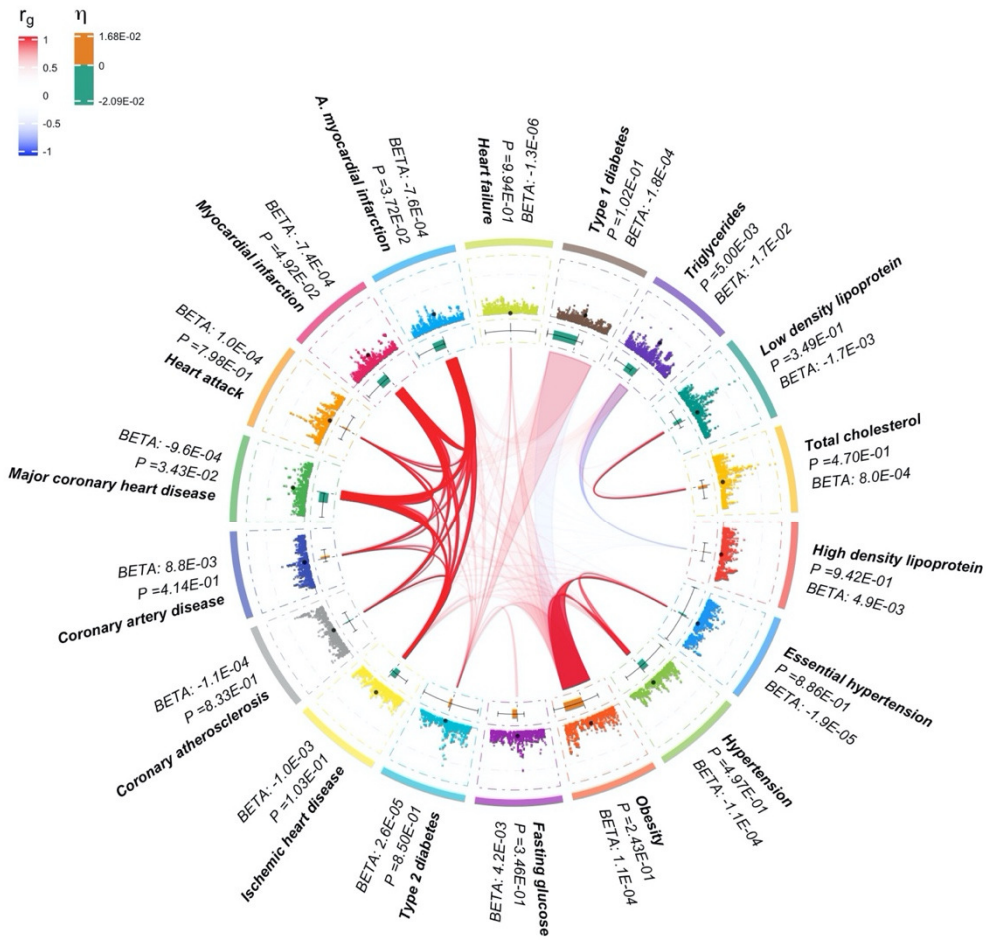
I. rs876320.

[Group 4: Others]



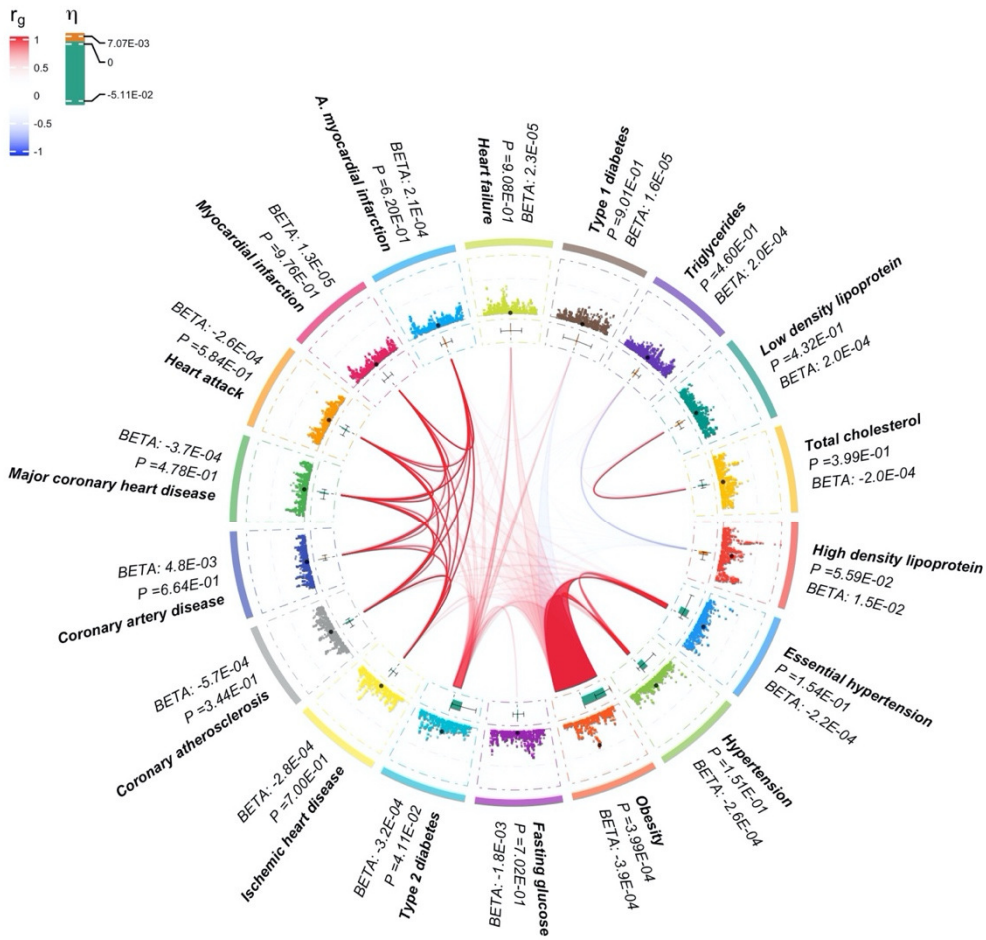
J. rs1561105.

[Group 4: Others]



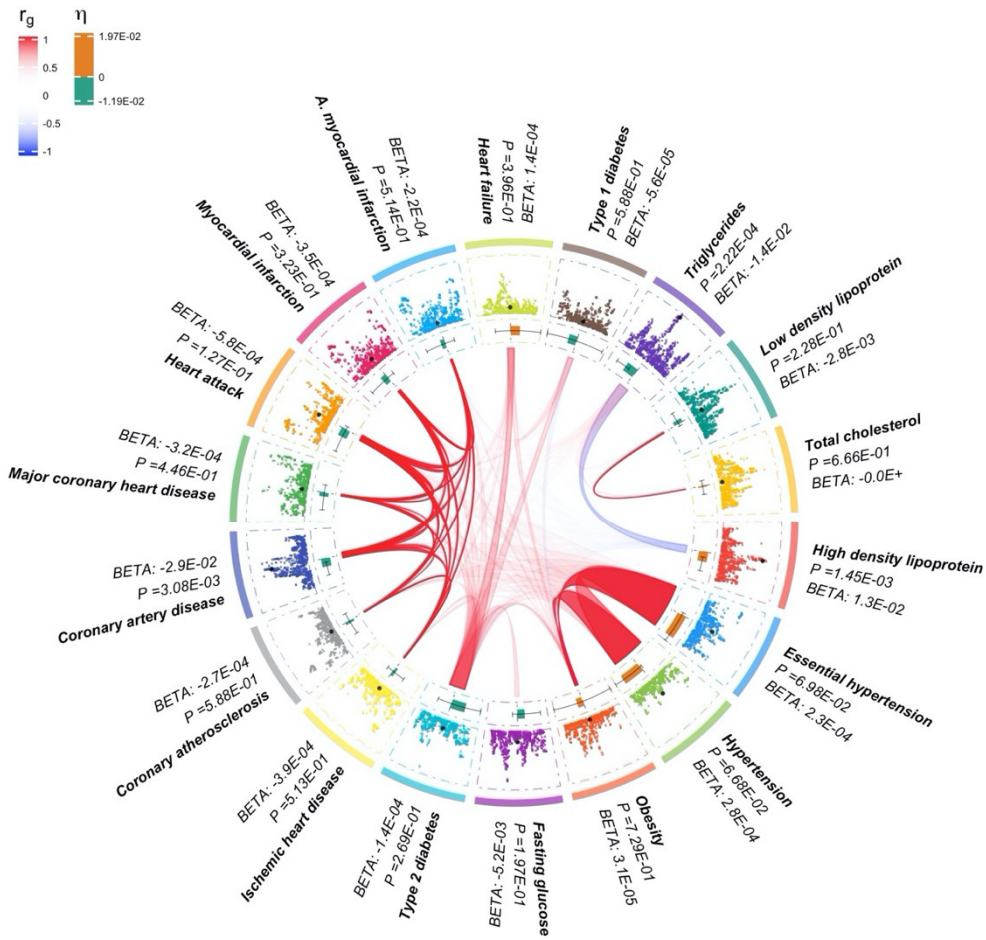
K. rs2891902.

[Group 4: Others]



L. rs2055014.

[Group 4: Others]



M. rs1039119.

[Group 4: Others]

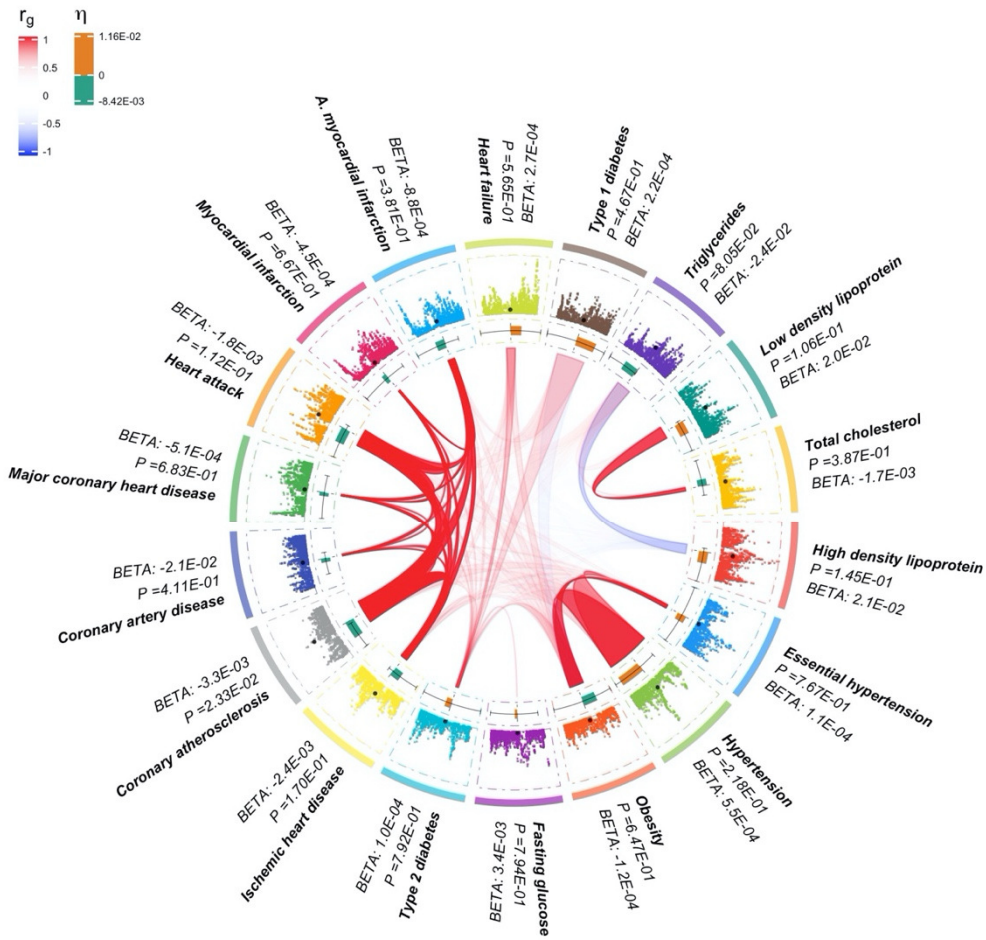


Figure 15. Pleiotropy plots of 13 novel loci identified by PLEIO.

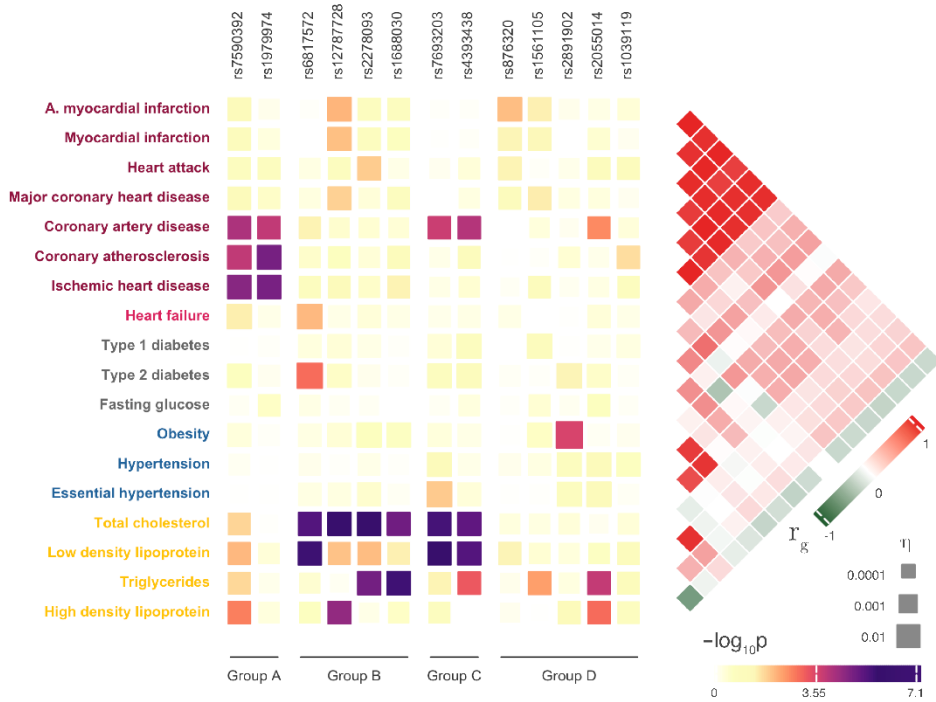


Figure 16. Distinct association patterns of 13 novel variants identified by PLEIO. Each box represents the association of a variant with a trait, where the size of the box indicates the magnitude of the standardized effect size (η) and the color of the box indicates the statistical significance. The right-side heatmap shows the genetic correlations. We divided the variants into four groups based on their association patterns.

3.7 Comparison of the association patterns between known and novel pleiotropic loci.

Of the 625 GWAS hits identified by PLEIO, I performed additional analysis on 612 known pleiotropic loci previously identified and reported in a GWAS study. In this analysis, I visualized the association patterns of the 625 variants. Then I compared the similarities and differences in association patterns between the known 612 pleiotropic loci and the 13 novel pleiotropic loci. For visualization, I collected p-values of 625 variants for 18 GWAS summary statistics used in my real data analysis and generated a heatmap with the $-log$ scaled p-values.

Figure 17 shows distinct association patterns of 625 pleiotropic variants for a total of 18 traits. In this analysis, I first divided these 625 variants into 16 categories using k-mean clustering. The majority of the pleiotropic loci identified by PLEIO showed strong associations $(p \cong 1 \times 10^{-6})$ with either four lipid traits (low-density lipoproteins, high-density lipoproteins, triglycerides, total cholesterol) or three cardiovascular diseases (coronary artery disease, ischemic heart disease, coronary atherosclerosis). This observation appears to be related to the number of samples used in each GWAS. For example, the summary statistics of the four lipid traits include more than $180{,}000$ samples, which is three times the number of GWAS samples for fasting glucose $(N = 46{,}186)$. The three cardiovascular disease traits from the UKB include many case samples over $10{,}000$. In contrast, the number of case samples for other UKB traits, including hypertension, type 1 diabetes, type 2 diabetes, and obesity, was less than $1{,}500$. For some pleiotropic loci, I found no strong associations with any of the 18 traits.

In summary, I found that the association patterns between known and novel pleiotropic loci were similar. As observed in the analysis above, PLEIO have identified some novel pleiotropic loci that have strong associations with either four lipid traits or three cardiovascular diseases (see **Figure 16**). However, I also confirmed that the magnitudes of the trait-specific associations of 13 novel pleiotropic loci were small compared to 612 known pleiotropic loci; therefore, they have not been identified as associated loci to date.

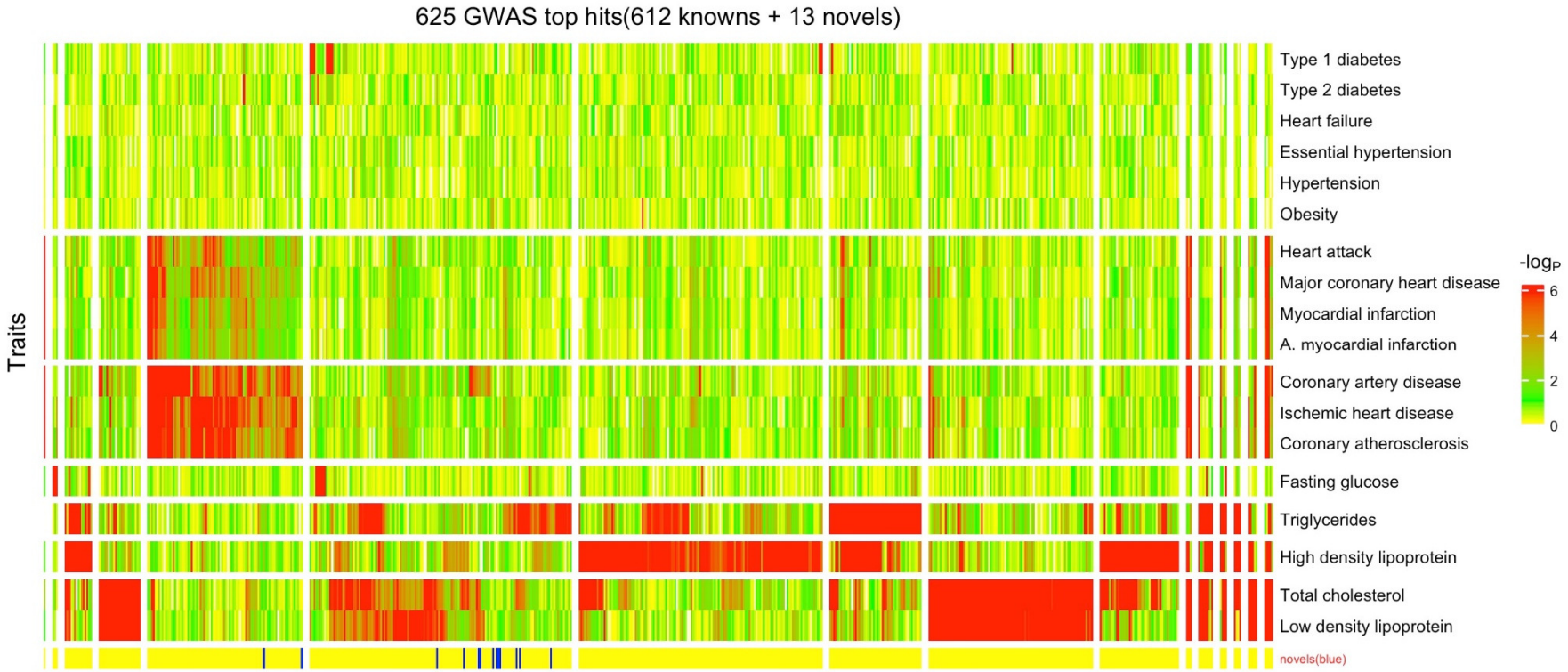


Figure 17. A heatmap created using the p-values of 625 pleiotropic variants for a total of 18 traits. The x-axis represents genetic variants, and the y-axis represents traits. Since the x-axis contains both known and novel pleiotropic variations, I added an annotation at the bottom of the heatmap to distinguish the two. To identify association patterns, I performed k-mean clustering on each axis. To find an optimal k for an axis, I tested several k values and selected one that would allow for a straightforward interpretation of the heatmap.

Chapter 4. Discussion

In this study, I proposed PLEIO, a statistical framework that identifies and interprets pleiotropic loci using GWAS summary statistics of multiple traits as input. PLEIO increases its statistical power by using a variance component model, which can account for genetic correlations and heritabilities across traits. Furthermore, PLEIO can seamlessly combine any set of quantitative and binary traits whose phenotypic units and scales can vary and provides an interpretation of the analysis results. This can be possible through the process of converting the observed effect sizes into standardized metrics. Finally, we provide an extension (R package) named 'pleiotropyPlot' to visualize and interpret the results of PLEIO's analysis.

PLEIO is a generalized method that can replace the traditional meta-analysis in special cases. If I set the genetic covariance matrix to a matrix of ones and the environmental correlation to zeros, the test is almost identical to the fixed effects meta-analysis method. If we assume non-zero environmental correlations, the test is nearly identical to the Lin-Sullivan method[36]. Suppose I set the genetic covariance matrix to the identity matrix and the environmental correlation to zeros. In that case, it is similar to the heterogeneity test in the Han-Eskin random-effects model[37]. If we assume non-zero environmental correlations, it is similar to the heterogeneity test in the RE2C model[38]. In contrast to the conventional meta-analysis methods, PLEIO optimizes model performance by learning genetic covariances and environmental correlations based on data (GWAS summary statistics). For example,

suppose you have a collection of multiple GWAS summary statistics on the same trait. In that case, PLEIO will learn the information and work as if it were a fixed-effects meta-analysis method.

We can do a fine-mapping analysis to test whether an identified pleiotropic locus has a true signal[39]. However, PLEIO does not tell us which traits were attributed to the pleiotropic association of the variant we tested. One way is to do a fine-mapping analysis for each trait, which requires a lot of labor. To reduce the number of traits to analyze, we can use a screening strategy to select strongly associated traits. One option is to use ASSET and select traits having strong signals, or one can manually select traits by interpreting the pleiotropy plot.

PLEIO can be extended to a single trait analysis that spans multiple ethnic groups. Assuming that each GWAS is carried out on one ethnic group, the genetic correlations can be estimated by considering the population-specific LD structures [40, 41]. In this case, the estimated genetic correlation between two GWAS summary statistics of the same trait from an ethnic group is 1. However, the estimated genetic correlation between the two GWAS summary statistics obtained from two different ethnic groups is generally positive but imperfect ($0 < r_g < 1$).

In a Multi-trait analysis, one must make a careful decision when choosing which traits to include in the analysis. This process can be performed based on literature describing comorbidities, shared candidate genes, or observed genetic correlations. Choosing a trait that does not have pleiotropic correlations with other traits reduces

the power to identify pleiotropic loci. In the real data analysis, the trait selection process was based on the literature search, and the observed r_g between selected traits was greater than 0.15. The method of selecting traits based on r_g estimated from the whole genome has a potential risk that can neglect the region-specific pleiotropic effects. This can happen when there are certain regions in which the regional co-heritabilities are greater than in other regions.

During the data collection, we may collect two or more GWAS that share overlapping samples. Failure to adequately account for the sample overlap between these summary statistics can inflate the error of the pooled estimates. For PLEIO, the proposed variance component model can account the sample overlaps across traits with an environmental correlation matrix. For example, our real data analysis showed strong environmental correlations between some traits collected by the UK Biobank and the Global Lipid Consortium that contains many overlapping samples.

There are two types of multi-trait analysis. The first type is a joint meta-analysis, in which statistics of several traits are combined into one. The goal of this type of analysis is to find pleiotropic loci that have associations with several traits. This type shares the advantages and disadvantages of conventional meta-analysis methods. Aggregating more traits can provide additional power, but modeling the heterogeneity between traits and interpreting results can often be challenging. The second type is a trait-specific analysis, in which multiple related traits are used to help with association tests of a specific trait. This type of analysis aims to maximize the statistical power of the analysis of individual traits. PLEIO is an analysis method

using a meta-analytic approach. Utilizing the visualization tool to facilitate interpretations, PLEIO can minimize the weaknesses of the joint meta-analysis.

PLEIO has similarities and differences to MTAG, which is used as the gold standard for multi-trait analyses. For example, both methods model genetic correlations, heritabilities, and environmental correlations. Although each has applied a different strategy, both methods can combine binary and quantitative traits of different units seamlessly. The main difference is that PLEIO is a meta-analysis approach, whereas MTAG is a trait-specific approach. For example, for a set of T traits, PLEIO provides one p-value per SNP, whereas MTAG provides T p-values per SNP. Therefore, if one aims to estimate a single p-value per SNP to map pleiotropic associations throughout the genome, PLEIO will be the optimal choice. On the other hand, one advantage of MTAG is the ability to assess the polygenic risk predictions more accurately using updated trait-specific effect sizes.

For PLEIO, the trait-specific effect sizes can be updated manually with the results of the PLEIO analysis by estimating BLUP (best linear unbiased predictor)[42]. Using $\widehat{\boldsymbol{\Omega}}$, $\widehat{\boldsymbol{\Sigma}}$, and standardized effect sizes($\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}$), the updated trait-specific effect sizes can be estimated as follows:

$$\boldsymbol{u_i} = \left[\left(\hat{\tau}_i^2\widehat{\boldsymbol{\Omega}}\right)^{-1} + \widehat{\boldsymbol{\Sigma}}^{-1}\right]^{-1}\widehat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}$$

and

$$Var(\boldsymbol{u_i}) = \left[\left(\hat{\tau}_i^2\widehat{\boldsymbol{\Omega}}\right)^{-1} + \widehat{\boldsymbol{\Sigma}}^{-1}\right]^{-1}$$

where the $\boldsymbol{u_i}$ is a $T \times 1$ vector representing the BLUP estimators of the observed effect sizes $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$. Note that the estimates of $\hat{\tau}_i^2$ for ith association test can be found in PLEIO's output file.

The pleiotropic loci found by PLEIO can be attributed to biological or mediated pleiotropy[43]. In the former case, the variant has an independent association for each trait tested. However, the variant will have non-independent associations in the latter case due to the causal relationship of two or more traits being tested. In the case of PLEIO, the association test results identify both biological and mediated pleiotropic associations, and the model does not discriminate the type of pleiotropy of the identified pleiotropic loci. Later, examining the extent of the pleiotropic association due to biological pleiotropy using the analysis results of PLEIO will be an exciting research direction.

Although not mentioned in this study, there exist multi-trait analysis methods that apply individual levels of genotyping data to a multivariate regression model[44-46]. These methods can utilize individual-level information to control confounding factors consistently across traits. However, to use this model, sample data for all traits must be collected in one place. Furthermore, the transmission of the genotyping data itself is becoming more and more difficult due to privacy concerns, hindering the use of these methods[47, 48]. In addition, models that use individual genotypes typically require a lot of computing resources. In terms of statistical power, Lin and Zeng[49] have shown that using data at the individual level does not significantly improve statistical power over using summary statistics in the context of traditional

meta-analysis. It would be interesting to compare the power between the two types of methods in future studies.

In this study, I assume that the additive effects can explain a large part of the genetic contribution of a locus. However, an average gene effect can also be modulated by either dominance effects or epistasis (how genes interact with genes at other loci). In PLEIO, I used a variance component model of two random effects (genetic and environmental). Here, the variance-covariance matrices were obtained from a method that only models additive genetic effects. To add a new variance component (either dominance effects or epistasis) into the variance component model of PLEIO, it might be required to estimate each variance component's covariance matrix and collect summary statistics for each genetic effect and variant.

PLEIO can be extended as a web application that helps many researchers around the world. Below, I provide the details of the proposed web application. The suggested web application includes a database capable of storing several GWAS summary statistics for various complex diseases and traits. Here, we consolidate the format of each summary statistics to simplify the analysis in the backend. The front end provides the user with the following functions: navigation and selection of the traits to be analyzed and various options for interpreting the analysis results. Finally, the back end provides core algorithms required for the PLEIO analysis and the interpretation of the analysis results.

In summary, I proposed a general and flexible meta-analysis framework to identify and interpret pleiotropic loci. I expect that our framework can help discover core genes that contribute to multiple phenotypes, leading us to a better understanding of the common etiology of traits and the development of shared drug targets.

Reference

1. Mall, M.A. and D. Hartl, CFTR: cystic fibrosis and beyond. Eur Respir J, 2014. **44**(4): p. 1042-54.
2. Nikpay, M., A. Goel, H.H. Won, L.M. Hall, C. Willenborg, S. Kanoni, D. Saleheen, T. Kyriakou, C.P. Nelson, J.C. Hopewell, et al., A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat Genet, 2015. **47**(10): p. 1121-1130.
3. Lloyd-Jones, D.M., B.H. Nam, R.B. D'Agostino, D. Levy, J.M. Murabito, T.J. Wang, P.W.F. Wilson, and C.J. O'Donnell, Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults - A prospective study of parents and offspring. Jama-Journal of the American Medical Association, 2004. **291**(18): p. 2204-2211.
4. Micha, R., G. Michas, and D. Mozaffarian, Unprocessed red and processed meats and risk of coronary artery disease and type 2 diabetes--an updated review of the evidence. Curr Atheroscler Rep, 2012. **14**(6): p. 515-24.
5. Boyle, E.A., Y.I. Li, and J.K. Pritchard, An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell, 2017. **169**(7): p. 1177-1186.
6. Marouli, E., M. Graff, C. Medina-Gomez, K.S. Lo, A.R. Wood, T.R. Kjaer, R.S. Fine, Y.C. Lu, C. Schurmann, H.M. Highland, et al., Rare and low-frequency coding variants alter human adult height. Nature, 2017. **542**(7640): p. 186-190.
7. Klein, R.J., C. Zeiss, E.Y. Chew, J.Y. Tsai, R.S. Sackler, C. Haynes, A.K. Henning, J.P. SanGiovanni, S.M. Mane, S.T. Mayne, et al., Complement

factor H polymorphism in age-related macular degeneration. Science, 2005. **308**(5720): p. 385-389.

8. MacArthur, J., E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, et al., The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Research, 2017. **45**(D1): p. D896-D901.
9. Tam, V., N. Patel, M. Turcotte, Y. Bosse, G. Pare, and D. Meyre, Benefits and limitations of genome-wide association studies. Nature Reviews Genetics, 2019. **20**(8): p. 467-484.
10. Manolio, T.A., F.S. Collins, N.J. Cox, D.B. Goldstein, L.A. Hindorff, D.J. Hunter, M.I. McCarthy, E.M. Ramos, L.R. Cardon, A. Chakravarti, et al., Finding the missing heritability of complex diseases. Nature, 2009. **461**(7265): p. 747-753.
11. Altshuler, D., M.J. Daly, and E.S. Lander, Genetic Mapping in Human Disease. Science, 2008. **322**(5903): p. 881-888.
12. McClellan, J. and M.C. King, Genetic Heterogeneity in Human Disease. Cell, 2010. **141**(2): p. 210-217.
13. Gratten, J. and P.M. Visscher, Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. Genome Medicine, 2016. **8**.
14. Watanabe, K., S. Stringer, O. Frei, M.U. Mirkov, C. de Leeuw, T.J.C. Polderman, S. van der Sluis, O.A. Andreassen, B.M. Neale, and D. Posthuma, A global overview of pleiotropy and genetic architecture in complex traits. Nature Genetics, 2019. **51**(9): p. 1339-+.

15. Welter, D., J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, et al., The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research, 2014. **42**(D1): p. D1001-D1006.

16. Bhattacharjee, S., P. Rajaraman, K.B. Jacobs, W.A. Wheeler, B.S. Melin, P. Hartge, M. Yeager, C.C. Chung, S.J. Chanock, N. Chatterjee, et al., A Subset-Based Approach Improves Power and Interpretation for the Combined Analysis of Genetic Association Studies of Heterogeneous Traits. American Journal of Human Genetics, 2012. **90**(5): p. 821-835.

17. Han, B. and E. Eskin, Interpreting Meta-Analyses of Genome-Wide Association Studies. Plos Genetics, 2012. **8**(3).

18. Kang, E.Y., Y. Park, X. Li, A.V. Segre, B. Han, and E. Eskin, ForestPMPlot: A Flexible Tool for Visualizing Heterogeneity Between Studies in Meta-analysis. G3-Genes Genomes Genetics, 2016. **6**(7): p. 1793-1798.

19. Turley, P., R.K. Walters, O. Maghzian, A. Okbay, J.J. Lee, M.A. Fontana, T.A. Nguyen-Viet, R. Wedow, M. Zacher, N.A. Furlotte, et al., Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet, 2018. **50**(2): p. 229-237.

20. Bulik-Sullivan, B.K., P.R. Loh, H.K. Finucane, S. Ripke, J. Yang, C. Schizophrenia Working Group of the Psychiatric Genomics, N. Patterson, M.J. Daly, A.L. Price, and B.M. Neale, LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet, 2015. **47**(3): p. 291-5.

21. Bulik-Sullivan, B., H.K. Finucane, V. Anttila, A. Gusev, F.R. Day, P.R. Loh, C. ReproGen, C. Psychiatric Genomics, C. Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control, L. Duncan, et al., An atlas of genetic correlations across human diseases and traits. Nat Genet, 2015. **47**(11): p. 1236-41.

22. Liley, J. and C. Wallace, A Pleiotropy-Informed Bayesian False Discovery Rate Adapted to a Shared Control Design Finds New Disease Associations From GWAS Summary Statistics. Plos Genetics, 2015. **11**(2).

23. Kang, H.M., J.H. Sul, S.K. Service, N.A. Zaitlen, S.Y. Kong, N.B. Freimer, C. Sabatti, and E. Eskin, Variance component model to account for sample structure in genome-wide association studies. Nat Genet, 2010. **42**(4): p. 348-54.

24. Self, S.G. and K.Y. Liang, Asymptotic Properties of Maximum-Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions. Journal of the American Statistical Association, 1987. **82**(398): p. 605-610.

25. Owen, A. and Y. Zhou, Safe and effective importance sampling. Journal of the American Statistical Association, 2000. **95**(449): p. 135-143.

26. Lee, C.H. Rcode: PleiotropyPlot. 2020; Available from: https://github.com/cuelee/pleiotropyPlot.

27. Willer, C.J., E.M. Schmidt, S. Sengupta, G.M. Peloso, S. Gustafsson, S. Kanoni, A. Ganna, J. Chen, M.L. Buchkovich, S. Mora, et al., Discovery and refinement of loci associated with lipid levels. Nat Genet, 2013. **45**(11): p. 1274-1283.

28. Liam Abbott, S.B., Claire Churchhouse, Andrea Ganna, Daniel Howrigan, Duncan Palmer, Ben Neale, Raymond Walters, Caitlin Carey, The Hail team. UK buibank GWAS results. Available from: http://www.nealelab.is/uk-biobank/.

29. Dupuis, J., C. Langenberg, I. Prokopenko, R. Saxena, N. Soranzo, A.U. Jackson, E. Wheeler, N.L. Glazer, N. Bouatia-Naji, A.L. Gloyn, et al., New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet, 2010. **42**(2): p. 105-16.

30. Bhattacharjee, S., P. Rajaraman, K.B. Jacobs, W.A. Wheeler, B.S. Melin, P. Hartge, C. GliomaScan, M. Yeager, C.C. Chung, S.J. Chanock, et al., A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. Am J Hum Genet, 2012. **90**(5): p. 821-35.

31. Willer, C.J., Y. Li, and G.R. Abecasis, METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics, 2010. **26**(17): p. 2190-1.

32. Zheng, J., A.M. Erzurumluoglu, B.L. Elsworth, J.P. Kemp, L. Howe, P.C. Haycock, G. Hemani, K. Tansey, C. Laurin, G. Early, et al., 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, 2017. **33**(2): p. 272-279.

33. McLaren, W., L. Gil, S.E. Hunt, H.S. Riat, G.R. Ritchie, A. Thormann, P. Flicek, and F. Cunningham, The Ensembl Variant Effect Predictor. Genome Biol, 2016. **17**(1): p. 122.

34. Dennis, G., B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane, and R.A. Lempicki, DAVID: Database for annotation, visualization, and integrated discovery. Genome Biology, 2003. **4**(9).

35. Stelzer, G., N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, T.I. Stein, R. Nudel, I. Lieder, Y. Mazor, et al., The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Curr Protoc Bioinformatics, 2016. **54**: p. 1 30 1-1 30 33.

36. Lin, D.Y. and P.F. Sullivan, Meta-analysis of genome-wide association studies with overlapping subjects. Am J Hum Genet, 2009. **85**(6): p. 862-72.

37. Han, B. and E. Eskin, Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. American Journal of Human Genetics, 2011. **88**(5): p. 586-598.

38. Lee, C.H., E. Eskin, and B. Han, Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. Bioinformatics, 2017. **33**(14): p. I379-I388.

39. Schaid, D.J., W.N. Chen, and N.B. Larson, From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics, 2018. **19**(8): p. 491-504.

40. Brown, B.C., C. Asian Genetic Epidemiology Network Type 2 Diabetes, C.J. Ye, A.L. Price, and N. Zaitlen, Transethnic Genetic-Correlation Estimates from Summary Statistics. Am J Hum Genet, 2016. **99**(1): p. 76-88.

41. Galinsky, K.J., Y.A. Reshef, H.K. Finucane, P.R. Loh, N. Zaitlen, N.J. Patterson, B.C. Brown, and A.L. Price, Estimating cross-population genetic correlations of causal effect sizes. Genet Epidemiol, 2019. **43**(2): p. 180-188.

42. Robinson, G.K., That BLUP is a Good Thing: The Estimation of Random Effects. Statistical Science, 1991. **6**(1): p. 15-32, 18.

43. Solovieff, N., C. Cotsapas, P.H. Lee, S.M. Purcell, and J.W. Smoller, Pleiotropy in complex traits: challenges and strategies. Nat Rev Genet, 2013. **14**(7): p. 483-95.

44. Korte, A., B.J. Vilhjalmsson, V. Segura, A. Platt, Q. Long, and M. Nordborg, A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet, 2012. **44**(9): p. 1066-71.

45. Yang, J., S.H. Lee, M.E. Goddard, and P.M. Visscher, GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet, 2011. **88**(1): p. 76-82.

46. Zhou, X. and M. Stephens, Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods, 2014. **11**(4): p. 407-9.

47. Erlich, Y. and A. Narayanan, Routes for breaching and protecting genetic privacy. Nat Rev Genet, 2014. **15**(6): p. 409-21.

48. Kim, K., H. Baik, C.S. Jang, J.K. Roh, E. Eskin, and B. Han, Genomic GPS: using genetic distance from individuals to public data for genomic analysis without disclosing personal genomes. Genome Biology, 2019. **20**(1).

49. Lin, D.Y. and D. Zeng, On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. Biometrika, 2010. **97**(2): p. 321-332.

50. Kang, H.M., N.A. Zaitlen, C.M. Wade, A. Kirby, D. Heckerman, M.J. Daly, and E. Eskin, Efficient control of population structure in model organism association mapping. Genetics, 2008. **178**(3): p. 1709-23.

51. Lee, C.H., S. Cook, J.S. Lee, and B. Han, Comparison of Two Meta-Analysis Methods: Inverse-Variance-Weighted Average and Weighted Sum of Z-Scores. Genomics Inform, 2016. **14**(4): p. 173-180.

Appendix

Appendix A Optimization strategy for variance component test

a) LRT statistic

Below, I describe our optimization strategy that increases the computational efficiency for determining the maximum likelihood estimate (MLE) in the variance component test of PLEIO. Note that I use the letter L to denote the number of traits in the multi-trait joint analysis (not T, because I use the letter T to denote the matrix transpose). Let $\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}$ denote a $L \times 1$ vector representing the observed standardized effect sizes for SNP i and $\boldsymbol{SE}(\hat{\boldsymbol{\eta}}_{\boldsymbol{i}})$ denote the vector of the corresponding standard errors. Let $\widehat{\boldsymbol{\Omega}}$ denote a $L \times L$ matrix representing the genetic covariance matrix of L traits, and $\widehat{\boldsymbol{\Sigma}}$ denote a $L \times L$ matrix representing the environmental covariance matrix. Note that $\widehat{\boldsymbol{\Sigma}} = diag\big(\boldsymbol{SE}(\hat{\boldsymbol{\eta}}_{\boldsymbol{i}})\big) \cdot \hat{\mathbf{C}}_{\mathbf{e}} \cdot diag\big(\boldsymbol{SE}(\hat{\boldsymbol{\eta}}_{\boldsymbol{i}})\big)$, where $\hat{\mathbf{C}}_{\boldsymbol{e}}$ is a $L \times L$ matrix representing the environmental correlation matrix, and $diag\big(\boldsymbol{SE}(\hat{\boldsymbol{\eta}}_{\boldsymbol{i}})\big)$ is a diagonal matrix whose diagonal values are $\boldsymbol{SE}(\hat{\boldsymbol{\eta}}_{\boldsymbol{i}})$. As described, $\widehat{\boldsymbol{\Sigma}}$ is independent of SNP i under the standardized scale. The PLEIO's statistic is a log-likelihood ratio test statistic (LRT). The likelihood functions under the null and alternative hypotheses can be shown as follows:

$$\mathcal{L}_0\big(\cdot \,|\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}; \widehat{\boldsymbol{\Sigma}}\big) = \frac{1}{(2\pi)^{L/2}\big|\widehat{\boldsymbol{\Sigma}}\big|^{1/2}} \exp\left(-\frac{1}{2}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}^T \widehat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\eta}}_{\boldsymbol{i}}\right)$$

$$\mathcal{L}_1\left(\tau^2\middle|\hat{\boldsymbol{\eta}}_{\boldsymbol{i}};\widehat{\boldsymbol{\Omega}},\widehat{\boldsymbol{\Sigma}}\right)=\frac{1}{(2\pi)^{L/2}\left|\widehat{\boldsymbol{\Omega}}\tau_i^2+\widehat{\boldsymbol{\Sigma}}\right|^{1/2}}\exp\left(-\frac{1}{2}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}^T\left(\widehat{\boldsymbol{\Omega}}\tau_i^2+\widehat{\boldsymbol{\Sigma}}\right)^{-1}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}\right),$$

and the corresponding LRT statistic is

$$S_{PLEIO}=-2\ \ln\left[\frac{\mathcal{L}_0\left(\cdot\,|\hat{\boldsymbol{\eta}}_{\boldsymbol{i}};\widehat{\boldsymbol{\Sigma}}\right)}{\sup\left\{\mathcal{L}_1\left(\tau_i^2\middle|\hat{\boldsymbol{\eta}}_{\boldsymbol{i}};\widehat{\boldsymbol{\Omega}},\widehat{\boldsymbol{\Sigma}}\right):\tau^2\geq 0\right\}}\right].$$

b) Efficient optimization

Our goal is to find $\boldsymbol{\tau^2}$ that satisfies $\sup\left\{\mathcal{L}_1\left(\tau_i^2\middle|\hat{\boldsymbol{\eta}}_{\boldsymbol{i}};\widehat{\boldsymbol{\Omega}},\widehat{\boldsymbol{\Sigma}}\right):\tau_i^2\geq 0\right\}$ under the alternative model, $\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}\sim\mathrm{MVN}\left(\mathbf{0},\tau^2\widehat{\boldsymbol{\Omega}}+\widehat{\boldsymbol{\Sigma}}\right)$, to obtain the LRT statistic. To find MLE $\hat{\tau}^2$, one possible way is to use an iterative optimization technique (e.g., quasi Newton's method). This optimization, however, requires a burdensome calculation of the matrix inversion of the variance-covariance matrix in the function $\mathcal{L}_1$ for each iteration. Instead, I propose a novel optimization technique that avoids the repeated inversions.

I first define a spectral (eigen) decomposition of a symmetric and positive semidefinite matrix $\mathbf{A}$ of size $L\times L$ as follows: let $\boldsymbol{\xi_A}$ denote the $L\times 1$ vector representing eigenvalues of A, and $\boldsymbol{P_A}$ denote the $L\times L$ matrix whose $\boldsymbol{i}$th column vector indicates the corresponding eigenvector. By definition, $\boldsymbol{P_A}$ is an orthonormal matrix so that $\boldsymbol{P_A^{-1}}=\boldsymbol{P_A^T}$. Suppose the values of $\boldsymbol{\xi_A}$ are sorted in descending order, and the corresponding column order of $\boldsymbol{P_A}$ is also sorted as well. Let $\xi_{A,i}$ be the $\boldsymbol{i}$th eigenvalue. Then, $\xi_{A,i}\geq 0$ where $i=\{1,2,\dots,L\}$.

Note that any covariance matrix (including $\widehat{\boldsymbol{\Omega}}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{i}}$) is symmetric and positive semidefinite by definition. Let $\widehat{\boldsymbol{\Omega}}^{\boldsymbol{g}}$ be the generalized (pseudo) inverse of $\widehat{\boldsymbol{\Omega}}$. Since $\frac{1}{\xi_{\widehat{\Omega},i}} > 0$ for all non-zero $\xi_{\widehat{\Omega},i}$, $\widehat{\boldsymbol{\Omega}}^{\boldsymbol{g}}$ is also positive semidefinite (PSD) and symmetric, as is $\left[\widehat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}}$. Then, I apply the following linear transformation to $\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}$:

$$\left[\widehat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}} \sim MVN\left(\mathbf{0}, \left[\widehat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}\left[\widehat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}} + \tau_i^2\boldsymbol{I}\right)$$

whose log-likelihood functions under the null and alternative hypotheses can be shown as:

$$\ell_0'\left(\cdot \middle| \left[\widehat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}; \boldsymbol{D}\right)$$

$$= -\frac{1}{2}\left[L\ln(2\pi) + \ln(|\boldsymbol{D}|) + \left(\left[\widehat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}\right)^T \boldsymbol{D}^{-1}\left(\left[\widehat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}\right)\right]$$

and

$$\ell_1'\left(\tau_i^2 \middle| \left[\widehat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}; \boldsymbol{K}\right)$$

$$= -\frac{1}{2}\left[L\ln(2\pi) + \ln(|\boldsymbol{K}|) + \left(\left[\widehat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}\right)^T \boldsymbol{K}^{-1}\left(\left[\widehat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}\right)\right],$$

where $\boldsymbol{D} = \left[\widehat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}\left[\widehat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}}$ and $\boldsymbol{K} = \boldsymbol{D} + \tau_i^2\boldsymbol{I}$. The product of two symmetric real PSD matrices, $\mathbf{A}^{-\frac{1}{2}}\boldsymbol{B}\mathbf{A}^{-\frac{1}{2}}$, is also symmetric real PSD, and the value of τ_i^2 is strictly non-negative. Therefore, $\boldsymbol{D}$ and $\boldsymbol{K}$ are symmetric real PSD, and are covariance matrices. Note that by applying this transformation, I made the second term in $\boldsymbol{K}$ a diagonal matrix. Under this condition, I can apply an optimization technique similar to ones used in EMMA[50] or RE2C[38]. The following equalities hold:

$$\boldsymbol{K}\boldsymbol{P_D} = (\boldsymbol{D} + \tau^2\boldsymbol{I})\boldsymbol{P_D}$$

$$= (\boldsymbol{D}\boldsymbol{P_D} + \boldsymbol{P_D}\tau^2) = \boldsymbol{P_D}\boldsymbol{\Lambda_D} + \boldsymbol{P_D}\tau^2$$

$$= \boldsymbol{P_D}(\boldsymbol{\Lambda_D} + \tau^2\boldsymbol{I}),$$

and

$$\boldsymbol{K} = \boldsymbol{P_D}(\boldsymbol{\Lambda_D} + \tau_i^2\boldsymbol{I})\boldsymbol{P_D}^T$$

where $\boldsymbol{\Lambda_D}$ is a $L \times L$ diagonal matrix whose ith element is $\xi_{\boldsymbol{D},i}$. By the definition of spectral decomposition, the following statements are true:

For a real symmetric PSD matrix $\boldsymbol{A}$, $|\boldsymbol{A}| = \prod_{i=1}^{p} \xi_{A,i}$.

$$\boldsymbol{A^g} = (\boldsymbol{P_A}\boldsymbol{E})^T[\boldsymbol{\Lambda_A^+}]^{-1}(\boldsymbol{P_A}\boldsymbol{E})$$

where p is the number of positive eigenvalues of A, $\boldsymbol{E}$ is a diagonal matrix whose first p diagonal elements are 1 and 0 otherwise, and $\boldsymbol{\Lambda_A^+}$ is a $p \times p$ diagonal matrix whose ith diagonal element is $\xi_{A,i}$. Then, I can rewrite ℓ_1' as follows:

$$\ell_1' = -\frac{1}{2}\Bigg[L\,ln(2\pi) + \sum\nolimits_{t=1}^{p} ln(\xi_{D,t} + \tau_i^2)$$

$$+ \left(\left[\hat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}\right)^T (\boldsymbol{P_D}\boldsymbol{E})^T[\boldsymbol{\Lambda_D^+}]^{-1}(\boldsymbol{P_D}\boldsymbol{E})\left(\left[\hat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}\right)\Bigg]$$

$$= -\frac{1}{2}\Bigg[L\,ln(2\pi) + \sum\nolimits_{t=1}^{p} ln(\xi_{D,t} + \tau_i^2)$$

$$+ \left(\boldsymbol{P_D}\boldsymbol{E}\left[\hat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}\right)^T [\boldsymbol{\Lambda_D^+}]^{-1}\left(\boldsymbol{P_D}\boldsymbol{E}\left[\hat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}\right)\Bigg]$$

$$= -\frac{1}{2}\left[L\,ln(2\pi) + \sum\nolimits_{t=1}^{p} ln(\xi_{D,t} + \tau_i^2) + \sum_{t=1}^{p}\frac{\boldsymbol{\delta_t^2}}{\xi_{D,t} + \tau_i^2}\right]$$

where $\boldsymbol{\delta_t^2}$ is the tth element of the vector $\boldsymbol{P_D}\boldsymbol{E}\left[\widehat{\boldsymbol{\Omega}}^{\boldsymbol{g}}\right]^{\frac{1}{2}}\hat{\boldsymbol{\eta}}_{\boldsymbol{i}}$. Using the equation above, I can derive the first and second derivative of the likelihood function ℓ_1' as:

$$\frac{d\ell_1'}{d\tau_i^2} = -\frac{1}{2}\left[\sum\nolimits_{t=1}^{p} \frac{1}{\xi_{D,t}+\tau_i^2} - \sum_{t=1}^{p} \frac{\boldsymbol{\delta_t^2}}{(\xi_{D,t}+\tau_i^2)^2}\right]$$

$$\frac{d^2\ell_1'}{d(\tau_i^2)^2} = -\frac{1}{2}\left[\sum\nolimits_{t=1}^{p} \frac{1}{(\xi_{D,t}+\tau_i^2)^2} + 2\sum_{t=1}^{p} \frac{\boldsymbol{\delta_t^2}}{(\xi_{D,t}+\tau_i^2)^3}\right].$$

In PLEIO, I use $\frac{d\ell_1'}{d\tau_i^2}$ as the root function of the Newton Raphson algorithm and $\frac{d^2\ell_1'}{d(\tau_i^2)^2}$ as its first derivative function. The initial value of the algorithm can be found using a grid search algorithm. I generate multiple values of τ_i^2 ranged from 10^{-9} to 10^6 and use τ_i^2 that maximizes the likelihood as an initial value. Finally, I estimate the value of PLEIO as follows:

$$S_{PLEIO} = \left[\sum\nolimits_{t=1}^{p} \ln\left(\frac{\xi_{D,t}}{\xi_{D,t}+\widehat{\tau}_i^2}\right)\right] + \left[\sum_{t=1}^{p} \frac{\boldsymbol{\delta_t^2}}{\xi_{D,t}} - \sum_{t=1}^{p} \frac{\boldsymbol{\delta_t^2}}{\xi_{D,t}+\widehat{\tau}_i^2}\right]$$

This simple form of S_{PLEIO} shows that our method can optimize τ_i^2 with a single matrix inversion of $\boldsymbol{D}$. Recall that optimization of τ_i^2 using a naïve quasi-Newton Raphson approach would have required multiple matrix inversions. I tested the computing efficiency of our method by comparing it to that of the standard approach using the optimization function implemented in the python Scipy library (scipy.optimize.minimize), taking into account the two variance-covariance matrices. **Figure 5** and **Figure 6** show that the computational time of the proposed model was faster than the time of the standard optimization. The reduction rate of computational time was approximately linear in relation to the number of studies, where the reduction rate increased approximately 8% per number of studies. In other words, the proposed model can compute 16-fold (1600%) faster than scipy.optimization for cross-disease joint analysis of 200 studies ($8\% \times 200 = 1600\%$) (**Figure 6**).

Appendix B P-value estimation

a) Problem definition

I describe how PLEIO estimates the p-value. Suppose I have an observed statistic $\hat{S}_{PLEIO}$. Then, the p-value is $\mathrm{P}\left(S_{PLEIO} \geq \hat{S}_{PLEIO} \middle| H_0\right)$. Let x denote a random variable representing the standardized effect size $\boldsymbol{\eta}$, $\widehat{\boldsymbol{\Omega}}$ denote the $L \times L$ genetic covariance matrix, and $\widehat{\boldsymbol{\Sigma}}$ denote the $L \times L$ environmental covariance matrix. In the following, I will assume that the estimated $\widehat{\boldsymbol{\Sigma}}$ represents the true environmental variance. Under the null, x will follow $\mathrm{MVN}\left(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}\right)$. Let the statistic S_{PLEIO} denote a function of x given $\widehat{\boldsymbol{\Omega}}$ and $\widehat{\boldsymbol{\Sigma}}$. I can define an indicator function

$$f\left(x, \theta | \widehat{\Sigma}, \widehat{\Omega}\right) = \begin{cases} 1 \; if \; S_{PLEIO}(x | \widehat{\Sigma}, \widehat{\Omega}) \geq \theta \\ 0 \; if \; S_{PLEIO}\left(x \middle| \widehat{\Sigma}, \widehat{\Omega}\right) < \theta \end{cases}.$$

For simplicity, I replace $f\left(x, \theta | \widehat{\Sigma}, \widehat{\Omega}\right)$ with a simpler expression, $f(x)$. Let $q(x)$ be the probability density function of $\mathrm{MVN}\left(0, \widehat{\Sigma}\right)$. For a given observed LRT statistic θ, the p-value can be defined as the expected value of the integrand $f(x)q(x)$ as follows,

$$I = \int_{\boldsymbol{D}} f(x)q(x)dx$$

where $\boldsymbol{D} = \mathbb{R}^{\boldsymbol{L}}$. Our goal is to estimate $\boldsymbol{I}$ accurately and efficiently.

b) Asymptotic approach

The simplest way to approximate the p-value I is to use the asymptotic distribution. In general, an LRT statistic using the value of the likelihood for random components in a linear mixed model asymptotically follows a mixture of Chi-squared

distributions under the null. In our situation, since τ^2 is a non-negative variance estimate, the statistics asymptotically follow a $50{:}\,50$ mixture of zero and one degrees of freedom Chi-square distributions[24]. In a cross-disease joint analysis, however, the validity of this asymptote holds if only if the number of combined studies is large. In practice, it is uncommon to combine more than 100 statistics. Therefore, the asymptotic approach will not be an exact solution for us.

c) Standard Monte Carlo approach

A possible alternative is the Monte Carlo approach. In the Monte Carlo method, I repeatedly draw samples from q. Let $\boldsymbol{X}^q$ denote a set of samples generated from q. Then, $I = \mathrm{E}^{\mathrm{q}}[f(X^q)]$ where $\mathrm{E}^{\mathrm{q}}[\cdot]$ denotes expectation for $\boldsymbol{X}^q \sim q$. Assume that I have sampled N observations and let $\boldsymbol{X}_i^q$ be the $\boldsymbol{i}$th observation. The Monte Carlo estimator of I is $\hat{I} = \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{X}_i^q)$. However, the use of Monte Carlo integration can be computationally intensive if the region of interest in $\boldsymbol{X}$ is located at the tails of q distribution. In genome-wide analyses, the p-value of interest is often as small as 5×10^{-8}. To get reasonable accuracy for such a small value, more than 10 million samples are required. This can be computationally intensive since the maximum likelihood estimation must be carried out for each sample to calculate S_{PLEIO}.

In previous studies, the standard Monte Carlo approach was used in the context of meta-analysis of GWAS. Han and Eskin adapted a strategy to pre-calculate I for every θ using the Monte Carlo sampling[17]. For each possible number of studies (L), they generated $10M$ null samples and tabulated the relationship between I and θ. This was possible because they need not assume any genetic correlation or

environmental correlation. Since their statistical model did not include any correlations, the standard MVN with mean zero and variance I (identity matrix) represented the null distribution of various situations well; therefore, pre-sampling from that distribution was sufficient. In a subsequent study, Cue et al.[38] extended the model to include the environmental correlation caused by sample overlap. Because the environmental correlation can vary from situation to situation, it was not possible to calculate the table in advance for all possible situations. Fortunately, the environmental correlation always becomes positive if it is due to a sample overlap of controls. Inspired by this, Cue et al. *[38]* developed a heuristic to summarize the strength of overall positive correlations between studies in one value (the average correlation $\bar{r}$), and tabulated I for each possible $\bar{r}$. Although these previous studies have used the standard Monte Carlo approach or its variation, I cannot apply these approaches directly in our context. This is because every analysis will have unique $\widehat{\boldsymbol{\Omega}}$ and $\widehat{\boldsymbol{\Sigma}}$, and it is not possible to pre-calculate the p-value table for every possible $\widehat{\boldsymbol{\Omega}}$ and $\widehat{\boldsymbol{\Sigma}}$.

d) Importance sampling approach

I have developed an importance sampling approach to solve this challenge. Since each analysis study with our method will have unique $\widehat{\boldsymbol{\Omega}}$ and $\widehat{\boldsymbol{\Sigma}}$, our strategy is to calculate the p-value table in a study-specific manner. The use of the standardized effect size $\boldsymbol{\eta}$ helps in this situation, because $\boldsymbol{SE}(\boldsymbol{\eta})$ is independent of the SNPs. Therefore, under the null, $\boldsymbol{\eta}$ always follows $MVN\left(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}\right)$, regardless of SNPs. Thus,

once I successfully build the null distribution of S_{PLEIO}, I can use the distribution repeatedly for all SNPs.

In importance sampling, I define the sampling distribution $p(x)$, which is a positive probability density function in $\boldsymbol{D}$. Then,

$$I = \int_D f(x)q(x)dx = \int_D \frac{f(x)q(x)}{p(x)}p(x)dx = E^p\left[\frac{f(\boldsymbol{X}^p)q(\boldsymbol{X}^p)}{p(\boldsymbol{X}^p)}\right]$$

and

$$\hat{I} = \frac{1}{M}\sum_{i=1}^{M}\frac{f(\boldsymbol{X}_i^p)q(\boldsymbol{X}_i^p)}{p(\boldsymbol{X}_i^p)},$$

where $\mathrm{E}^{\mathrm{p}}[\cdot]$ denotes expectation for $\boldsymbol{X}^p \sim p$. The two practical constraints on importance sampling are; it must be feasible to generate $\boldsymbol{X}^p$, and I must be able to compute $\frac{f(x)q(x)}{p(x)}$. In the importance sampling method, the variance of $\hat{I}$ can be shown as $\mathrm{Var}(\hat{I}) = \frac{\sigma_p^2}{M}$ where σ_p^2 is the standard deviation of the random variable $\int \frac{f(x)q(x)}{p(x)}$ where

$$\begin{aligned}\sigma_p^2 &= \int_{\boldsymbol{D}} \left(\frac{f(x)q(x)}{p(x)} - I\right)^2 p(x)dx \\ &= \int_{\boldsymbol{D}} \frac{f(x)^2q(x)^2}{p(x)}dx - 2\int_{\boldsymbol{D}} f(x)q(x)dx + I^2 \\ &= \int_{\boldsymbol{D}} \frac{f(x)^2q(x)^2}{p(x)}dx - I^2.\end{aligned}$$

Suppose $f(x) > 0$ and $I > 0$. In this case, the optimal $p(x)$ can be defined as $p^*(x) = \frac{f(x)q(x)}{I}$ where this density gives $\sigma_p^2 = 0$. Since I is an unknown constant that depends on the value of θ, I cannot use $p^*(x)$. However, it is clear that a good

sampling density of the importance sampling method will be roughly proportional to $f(x)q(x)$.

In PLEIO, it is challenging to choose a good sampling distribution p. In GWAS, the p-values can be as big as 1.0 or as small as 5×10^{-8}, or can be even smaller. Thus, I have a wide range of θ. Depending on θ, $f(x)$ changes. If I choose p that resembles $f(x)q(x)$ for a large θ, it can give a large variance for a small θ. Reversely, if I choose p that resembles $f(x)q(x)$ for a small θ, it can give a large variance for a large θ. To solve this challenge, I decided to use multiple sampling distributions $p_j(x)$ where $j = \{1,2, \dots K\}$. $p_1(x)$ is $q(x)$, the probability density function of the original null distribution that follows $\mathrm{MVN}\left(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}\right)$. Then I increase the variance of each coordinate by a factor of φ^2, where $\varphi \in \{1.1, 1.2, 1.3, 1.4, 1.7, 2, 2.5, 3.0, 4.0, 5.0\}$. Thus, in total, I use 11 sampling distributions such that each $p_j(x)$ follows $\mathrm{MVN}\left(\mathbf{0}, \mathrm{c_j^2}\widehat{\boldsymbol{\Sigma}}\right)$, where c_j is a constant value ranged from 1 to 5. Let $\boldsymbol{X}^{p_j}$ denote a matrix representing samples generated from the jth sampling distribution, and let α_j denote the proportion of samples generated from the jth sampling distribution. Then, the matrix $\boldsymbol{X}^{p_j}$ has a size of $\alpha_j M \times T$. In PLEIO, I generate $\boldsymbol{X}^p$ from each $p_j(x)$ uniformly, so that $\alpha_j \approx \frac{1}{K}$. In the importance sampling method using multiple sampling distributions, the p-value of a given θ can be estimated by follows:

$$\hat{I} = \frac{1}{M}\sum_{i=1}^{M}\frac{f(\boldsymbol{X}_i^p)q(\boldsymbol{X}_i^p)}{p_\alpha(\boldsymbol{X}_i^p)},$$

where $p_\alpha(\boldsymbol{X}_i^p) = \sum_{j=1}^{K}\alpha_j p_j(\boldsymbol{X}_i^p)$, and $\sum_{j=1}^{K}\alpha_j = 1$.

Recently, Owen and Zhou (2000) [25] proposed a novel importance sampling approach to minimize the variance of the estimate in the situation where multiple sampling distributions are used. I employed this approach. The method generates $\boldsymbol{X}^p$ from $p_j(x)$ and uses $p_j(\boldsymbol{X}^p)/p_\alpha(\boldsymbol{X}^p)$ as the control variates of $f(\boldsymbol{X}^p)q(\boldsymbol{X}^p)/p_\alpha(\boldsymbol{X}^p)$. Assuming high correlations between $f(x)q(x)$ and $p_j(x)$, I expect a large reduction in variance when estimating $\mathrm{E}\big(f(x)\big)$ using the control variate method. Note that $p_j(x)$ is a probability density function and therefore has the expected value $\int_{\boldsymbol{D}} p_j(x)dx = 1$. The expression of the importance sampling with the control variate method can be shown as follows:

$$\hat{I} = \frac{1}{M}\left(\sum_{i=1}^{M} \frac{f(\boldsymbol{X}_i^p)q(\boldsymbol{X}_i^p) - \sum_{j=1}^{K}\beta_j p_j(\boldsymbol{X}_i^p)}{p_\alpha(\boldsymbol{X}_i^p)}\right) + \sum_{k=1}^{K}\beta_k\mu_{p_k}$$

where $\mu_{p_j} = \mathrm{E}^p[p_j(\boldsymbol{X}_i^p)/p_\alpha(\boldsymbol{X}_i^p)] = \int p_j(x)dx = 1$ for any j. Define $m^* = \frac{f(x)q(x)-\sum_{j=1}^{K}\beta_j p_j(x)}{p_\alpha(x)} + \sum_{k=1}^{K}\beta_k\mu_{p_k}$ such that $\mathrm{E}[m^*] = I$. The variance of m^* is then,

$$\begin{aligned}
\mathrm{Var}(m^*) = &-\beta_1\mathrm{Cov}\left(\frac{f(x)q(x)}{p_\alpha(x)}, \frac{p_1(x)}{p_\alpha(x)}\right) + \sum_{l=1}^{K}\beta_1\beta_l\mathrm{Cov}\left(\frac{p_1(x)}{p_\alpha(x)}, \frac{p_l(x)}{p_\alpha(x)}\right) \\
&-\beta_2\mathrm{Cov}\left(\frac{f(x)q(x)}{p_\alpha(x)}, \frac{p_2(x)}{p_\alpha(x)}\right) + \sum_{l=1}^{K}\beta_2\beta_l\mathrm{Cov}\left(\frac{p_2(x)}{p_\alpha(x)}, \frac{p_l(x)}{p_\alpha(x)}\right) \\
&-\beta_3\mathrm{Cov}\left(\frac{f(x)q(x)}{p_\alpha(x)}, \frac{p_3(x)}{p_\alpha(x)}\right) + \sum_{l=1}^{K}\beta_3\beta_l\mathrm{Cov}\left(\frac{p_3(x)}{p_\alpha(x)}, \frac{p_l(x)}{p_\alpha(x)}\right) \\
&\vdots
\end{aligned}$$

$$-\beta_K \mathrm{Cov}\left(\frac{f(x)q(x)}{p_\alpha(x)}, \frac{p_K(x)}{p_\alpha(x)}\right) + \sum_{l=1}^{K} \beta_K \beta_l \mathrm{Cov}\left(\frac{p_K(x)}{p_\alpha(x)}, \frac{p_l(x)}{p_\alpha(x)}\right)$$

$$+ \mathrm{Var}\left(\frac{f(x)q(x)}{p_\alpha(x)}\right)$$

because $\mathrm{Var}(\beta_k \mu_{p_k}) = 0$. By definition, the optimal $\boldsymbol{\beta}$ can be found by solving the following partial derivatives.

$$\frac{\partial \mathrm{Var}(m^*)}{\partial \beta_j} = -\mathrm{Cov}\left(\frac{f(x)q(x)}{p_\alpha(x)}, \frac{p_j(x)}{p_\alpha(x)}\right) + \sum_{l=1}^{K} \beta_l \mathrm{Cov}\left(\frac{p_j(x)}{p_\alpha(x)}, \frac{p_l(x)}{p_\alpha(x)}\right)$$

which generates $K(=11)$ equations and K unknown variables (β) as follows:

$$\frac{\partial \mathrm{Var}(m^*)}{\partial \beta_1} = -\mathrm{Cov}\left(\frac{f(x)q(x)}{p_\alpha(x)}, \frac{p_1(x)}{p_\alpha(x)}\right) + \sum_{l=1}^{K} \beta_l \mathrm{Cov}\left(\frac{p_1(x)}{p_\alpha(x)}, \frac{p_l(x)}{p_\alpha(x)}\right)$$

$$\frac{\partial \mathrm{Var}(m^*)}{\partial \beta_2} = -\mathrm{Cov}\left(\frac{f(x)q(x)}{p_\alpha(x)}, \frac{p_2(x)}{p_\alpha(x)}\right) + \sum_{l=1}^{K} \beta_l \mathrm{Cov}\left(\frac{p_2(x)}{p_\alpha(x)}, \frac{p_l(x)}{p_\alpha(x)}\right)$$

$$\vdots$$

$$\frac{\partial \mathrm{Var}(m^*)}{\partial \beta_K} = -\mathrm{Cov}\left(\frac{f(x)q(x)}{p_\alpha(x)}, \frac{p_K(x)}{p_\alpha(x)}\right) + \sum_{l=1}^{K} \beta_l \mathrm{Cov}\left(\frac{p_K(x)}{p_\alpha(x)}, \frac{p_l(x)}{p_\alpha(x)}\right).$$

Here, the function of $\mathrm{Var}(m^*)$ is a quadratic function for β_j. Therefore, the β^* which maximizes the variance of $\mathrm{Var}(m^*)$ satisfies $\frac{\partial \mathrm{Var}(m^*)}{\partial \beta_j^*} = 0$, the root of the partial derivative. Thus, the optimal β^* $\left(\beta \middle| \frac{\partial \mathrm{Var}(m^*)}{\partial \beta_j} = 0 \text{ and } X_i^p\right)$ can be obtained by solving the linear equation:

$$\mathrm{Var}(P_1)\beta_1^* + \mathrm{Cov}(P_1, P_2)\beta_2^* + \cdots + \mathrm{Cov}(P_1, P_K)\beta_K^* = \mathrm{Cov}(FQ\ , P_1)$$

$$\mathrm{Cov}(P_2, P_1)\beta_1^* + \mathrm{Var}(P_2)\beta_2^* + \cdots + \mathrm{Cov}(P_2, P_K)\beta_K^* = \mathrm{Cov}(FQ\ , P_2)$$

$$\vdots$$

$$\mathrm{Cov}(P_K, P_1)\beta_1^* + \mathrm{Cov}(P_K, P_2)\beta_2^* + \cdots + \mathrm{Var}(P_K)\beta_K^* = \mathrm{Cov}(FQ\ , P_K)$$

which can be shown as follows:

$$\begin{bmatrix} \mathrm{Var}(P_1) & \mathrm{Cov}(P_1,P_2) & \dots & \mathrm{Cov}(P_1,P_K) \\ \mathrm{Cov}(P_2,P_1) & \mathrm{Var}(P_2) & \cdots & \mathrm{Cov}(P_2,P_K) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(P_K,P_1) & \mathrm{Cov}(P_K,P_2) & \dots & \mathrm{Var}(P_K) \end{bmatrix} \begin{bmatrix} \beta_1^* \\ \beta_2^* \\ \vdots \\ \beta_K^* \end{bmatrix} = \begin{bmatrix} \mathrm{Cov}(FQ\ ,P_1) \\ \mathrm{Cov}(FQ\ ,P_2) \\ \vdots \\ \mathrm{Cov}(FQ\ ,P_K) \end{bmatrix}$$

and

$$\begin{bmatrix} \beta_1^* \\ \beta_2^* \\ \vdots \\ \beta_K^* \end{bmatrix} = \begin{bmatrix} \mathrm{Var}(P_1) & \mathrm{Cov}(P_1,P_2) & \dots & \mathrm{Cov}(P_1,P_K) \\ \mathrm{Cov}(P_2,P_1) & \mathrm{Var}(P_2) & \cdots & \mathrm{Cov}(P_2,P_K) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(P_K,P_1) & \mathrm{Cov}(P_K,P_2) & \dots & \mathrm{Var}(P_K) \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{Cov}(FQ\ ,P_1) \\ \mathrm{Cov}(FQ\ ,P_2) \\ \vdots \\ \mathrm{Cov}(FQ\ ,P_K) \end{bmatrix},$$

where $P_l = \frac{p_j(\boldsymbol{X}_i^p)}{p_\alpha(\boldsymbol{X}_i^p)}$, and $FQ = \frac{f(\boldsymbol{X}_i^p)q(\boldsymbol{X}_i^p)}{p_\alpha(\boldsymbol{X}_i^p)}$. Owen and Zhou[25] showed that If at least one of $p_j(x) > 0$ whenever $f(x)q(x) > 0$, then $\hat{I}_{\alpha,\beta}$ is unbiased and $\mathrm{Var}\big(\hat{I}_{\alpha,\beta}\big) \le \mathrm{Var}\big(\hat{I}_{\alpha_j p_j}\big)$ for any j where $\hat{I}_{\alpha_j p_j} = \frac{1}{n_j}\sum_{i=1}^{n_j} \frac{f(\boldsymbol{X}_i^p)q(\boldsymbol{X}_i^p)}{p_j(\boldsymbol{X}_i^p)}$.

e) Implementation

The implementation of the p-value estimation in PLEIO is as follows. After calculating $\widehat{\boldsymbol{\Omega}}$ and $\widehat{\boldsymbol{\Sigma}}$ using LDSC, I assume that these values are true values and generate the null samples by using an importance sampling method. The default number of sampling is $100K$, where each of 11 distributions being used equally. For each sample, I use our efficient transformation technique for the Newton Raphson method to determine the maximum likelihood estimate $\hat{\tau}^2$ and calculate S_{PLEIO}. Then, I calculate p-values of 40 different θ that are in the range $(0, 40)$. For each θ, I calculate the optimal $\boldsymbol{\beta}$ for the control variate method and use the method to calculate the p-value from our null samples. Using these 40 points, I interpolate p-values for $\theta < 40$ using B-spline fitting and extrapolate p-values for $\theta > 40$ using linear fitting on the log p-value scale.

Appendix C Estimation of environmental correlations using LDSC

I first describe the LDSC framework of Bulik-Sullivan et al. (2015)[20, 21]. Let A and B denote two different traits. I assume that we genotyped N_A and N_B samples at M SNPs. Let $\boldsymbol{y_A}$ and $\boldsymbol{y_B}$ denote $N_A \times 1$ and $N_B \times 1$ vectors of phenotypes, and let $\boldsymbol{G_A}$ and $\boldsymbol{G_B}$ denote $N_A \times M$ and $N_B \times M$ genotype matrices. $\boldsymbol{G_A}$ and $\boldsymbol{G_B}$ are standardized so that each column follows $N(0,1)$. The z-scores of the SNP j can be obtained as $z_{A,j} := \mathbf{G}_{\boldsymbol{A},j}^T \boldsymbol{y_A}/\sqrt{N_A}$ and $z_{B,j} := \mathbf{G}_{\boldsymbol{B},j}^T \boldsymbol{y_B}/\sqrt{N_B}$. Let $\boldsymbol{z_A}$ and $\boldsymbol{z_B}$ denote $M \times 1$ vectors of z-scores of traits $\boldsymbol{A}$ and $\boldsymbol{B}$.

Bulik-Sullivan et al., (2015) derived the following equations[20, 21]:

$$E[z_{A,j}^2|\ell_j] = \frac{N_A h_A^2}{M}\ell_j + N_A\alpha_A + 1$$

$$E[z_{B,j}^2|\ell_j] = \frac{N_B h_B^2}{M}\ell_j + N_B\alpha_B + 1 \qquad (1)$$

where h_A^2 and h_B^2 are narrow sense heritabilities, and $\boldsymbol{\ell}_j$ is the value of the LD-score of jth SNP, which can be obtained from an external reference. LDSC uses both summary statistics and LD-score information to estimate the trait heritability (h_A^2 and h_B^2) along with the intercepts ($N_A\alpha_A + 1$ and $N_B\alpha_B + 1$). To estimate genetic correlation, LDSC uses the following equality:

$$E[z_{A,j}z_{B,j}|\ell_j] = \frac{\sqrt{N_A N_B}\sigma_g}{M}\ell_j + \frac{\sigma N_{AB}}{\sqrt{N_A N_B}} \qquad (2)$$

where σ_g denotes the genetic covariance between trait A and B, N_{AB} denotes the shared individuals between samples N_A and N_B, and σ denotes phenotypic correlations.

To estimate environmental correlation, I apply genomic control to both traits to make intercept one. After the genomic-control correction, the second term in equation (2) reflects the environmental correlation of z-scores attributable to shared individuals. Another similar approach is to use the weighted sum of z-scores to combine z-scores of A and B,

$$z_{AB,j} = \frac{\sqrt{N_A} z^c_{A,j} + \sqrt{N_B} z^c_{B,j}}{\sqrt{N_A + N_B}} \qquad (3)$$

where the superscript c denotes that the z-score was corrected with genomic control. The weighted sum of z-scores is approximately equivalent to the popular inverse-variance weighted method[51].

I can decompose z-scores to the genetic effect and the environmental error, such that $z^c_{A,j} = g_{A,j} + \epsilon_{A,j}$ *and* $z^c_{B,j} = g_{B,j} + \epsilon_{B,j}$. Let ρ_e denote the environmental correlation. That is, $\rho_e = \mathrm{Cor}(\epsilon_{A,j}, \epsilon_{B,j})$. Note that $\mathrm{Var}[g_{A,j}] = \frac{N_A h^2_A}{M} \ell_j$ and $\mathrm{Var}[g_{B,j}] = \frac{N_B h^2_B}{M} \ell_j$.

If there is a sample overlap,

$$E[(z_{AB,j}|N_{AB} \neq 0)^2|\ell_j]$$

$$= E\left[\left(\frac{\sqrt{N_A}(g_{A,j} + \epsilon_{A,j}) + \sqrt{N_B}(g_{B,j} + \epsilon_{B,j})}{\sqrt{N_A + N_B}}|N_{AB} \neq 0\right)^2 |\ell_j\right]$$

$$= f(\ell_j) + \left[\frac{N_A}{N_A + N_B} Var(\epsilon_{A,j}) + \frac{N_B}{N_A + N_B} \mathrm{Var}(\epsilon_{B,j}) \right.$$

$$\left. + 2\sqrt{N_A N_B}\mathrm{Cov}(\epsilon_{A,j}, \epsilon_{B,j})\right]$$

$$= \ f(\ell_j) + \left[1 + 2\frac{\sqrt{N_A N_B}}{N_A + N_B}\rho_e\right]$$

where $f(\ell_j)$ is a first-order function of ℓ_j without a constant term. Therefore, $\gamma_{AB} = 2\frac{\sqrt{N_A N_B}}{N_A+N_B}\rho_e$ is the inflation of the intercept caused by environmental correlation. Since I can estimate γ_{AB} from LDSC, I can estimate the environmental correlation as follows:

$$\rho_e = \frac{N_A + N_B}{2\sqrt{N_A N_B}}\gamma_{AB}.$$

국문초록

전장 유전체 연관성 분석 연구(GWAS)는 질병과 관련된 유전변이들의 위치를 탐색하는 방법으로 널리 사용되어왔습니다. 수행된 여러 전장 유전체 연관성 분석의 결과를 해석하여 연구자들은 질병과 연관된 유전자의 일부가 여러 특성과 동시에 연관성을 보이는 다발성 효과(pleiotropic effects)를 가짐을 확인했습니다. 이 다발성 유전자(pleiotropic loci)를 탐색하고 해석하는 것은 질병과 복잡한 특성간에 공유되는 유전적 생리기전을 이해하는 데 매우 중요합니다. 일반적인 다발성 유전자의 탐색은 여러 특성에 대한 GWAS 요약통계를 메타 분석수행하여 진행되었습니다. 그러나 기존의 방법들은 유전 적 상관 관계(genetic correlation) 및 유전율(heritability)과 같은 형질의 복잡한 유전 적 구조를 설명하지 않으며, 특성들간 표현형의 단위(unit) 및 척도(scale)가 다를 수 있음을 고려하지 않아 그 통계적 검사의 힘이 떨어집니다.

본 연구에서는 여러 질병과 복합 형질을 공동 분석하여 다발성 유전자좌를 탐색하고 해석 할 수있는 요약 통계 기반 분석 프레임 워크 인 PLEIO 를 제안합니다. 이 방법론은 형질의 유전 적 상관 관계와 유전성을 체계적으로 설명하여 연관성 검사의 성능을 극대화하며 이 과정에서 단위 및 척도가 다른 특성들을 원활하게 공동 분석 할 수 있도록 만들어 졌습니다. 추가적으로, 결과 해석을 지원하는 시각화 도구인 pleiotropyPlot 을 제공합니다.

PLEIO 의 성능을 확인하기 위해 본 연구자는 광범위한 시뮬레이션과 실제 데이터 분석을 수행했습니다. 7 개의 연구에서 서로 다른 유전 적 상관 관계 구조와 유전성을 가정 한 시뮬레이션을 수행하여 PLEIO 의 오탐율(FPR)이 잘 조정됨을 확인했고, 경쟁 방법론보다 뛰어난 통계적 힘을 가짐을 확인했습니다. 실제 데이터 분석에서는 심혈관 질환과 관련된 18 가지 특성을 PLEIO 에 적용했습니다. 그 결과 PLEIO 는 4 개의 서로 다른 연관 패턴을 가진 13 개의 새롭게 발견된 다발성 유전자좌를 식별했습니다. 계산 효율성 측면에서 PLEIO 는 100 개의 특성을 공동분석하여 1M 개의 SNP 를 1 일 내로 검사 할 수 있도록 최적화 되어 있습니다.

요약해서, PLEIO 는 형질간에 공통적으로 연관되는 다발성 유전자좌를 찾기 위해 형질 간의 유전 적 구조를 사용하는 공동 분석 다중 형질의 통계 프레임 워크입니다. PLEIO 에서 구현 된 통계 모델은 기존의 공동분석 모델에서 사용하는 가정들을 포괄하는 일반화 된 모델을 사용하며, Importance sampling 및 eigenvector decomposition 와 같은 수학적 기법을 사용하여 모델 최적화를 수행했습니다. PLEIO 소프트웨어는 다음 Github 웹 페이지에서 무료로 다운로드 할 수 있습니다: https://github.com/cuelee/pleio.