A Powerful Approach to Estimating Annotation-Stratified Genetic Covariance via GWAS Summary Statistics

Qiongshi Lu,[1,8] Boyang Li,[1] Derek Ou,[2] Margret Erlendsdottir,[2] Ryan L. Powles,[3] Tony Jiang,[4] Yiming Hu,[1] David Chang,[3] Chentian Jin,[4] Wei Dai,[1] Qidu He,[5] Zefeng Liu,[5] Shubhabrata Mukherjee,[6] Paul K. Crane,[6] and Hongyu Zhao[1,3,7,*]

Despite the success of large-scale genome-wide association studies (GWASs) on complex traits, our understanding of their genetic architecture is far from complete. Jointly modeling multiple traits' genetic profiles has provided insights into the shared genetic basis of many complex traits. However, large-scale inference sets a high bar for both statistical power and biological interpretability. Here we introduce a principled framework to estimate annotation-stratified genetic covariance between traits using GWAS summary statistics. Through theoretical and numerical analyses, we demonstrate that our method provides accurate covariance estimates, thereby enabling researchers to dissect both the shared and distinct genetic architecture across traits to better understand their etiologies. Among 50 complex traits with publicly accessible GWAS summary statistics ($N_{total} \approx 4.5$ million), we identified more than 170 pairs with statistically significant genetic covariance. In particular, we found strong genetic covariance between late-onset Alzheimer disease (LOAD) and amyotrophic lateral sclerosis (ALS), two major neurodegenerative diseases, in single-nucleotide polymorphisms (SNPs) with high minor allele frequencies and in SNPs located in the predicted functional genome. Joint analysis of LOAD, ALS, and other traits highlights LOAD's correlation with cognitive traits and hints at an autoimmune component for ALS.

Introduction

Genome-wide association studies (GWASs) have been a success in the past 12 years. Despite a simple study design, GWASs have identified tens of thousands of robust associations for a variety of human complex diseases and traits. Based on the GWAS paradigm, linear mixed models, in conjunction with the restricted maximum likelihood (REML) algorithm, have provided great insights into the polygenic genetic architecture of complex traits.[1–3] The cross-trait extension of linear mixed model has further revealed the shared etiology of many different traits.[4] Compared to traditional, family-based approaches, these methods do not require all the traits to be measured on the same cohort and therefore make it possible to study a spectrum of human complex traits using independent samples from existing GWASs.[5,6] Recently, Bulik-Sullivan et al. developed cross-trait LDSC, a computationally efficient method that utilizes GWAS summary statistics to estimate genetic correlation between complex traits.[7] LDSC is a major advance. As summary statistics from consortium-based GWASs become increasingly accessible,[8] it provides great opportunities for systematically documenting the shared genetic basis of a large number of diseases and traits.[9,10] However, large-scale inference sets a high bar for both estimation accuracy and statistical power. Furthermore, existing methods do not allow explicit modeling of functional genome annotations. As shown in later sections, the estimated genetic correlations in many cases are neither statistically significant nor easy to interpret.

To address these challenges, there is a pressing need for a statistical framework that provides more accurate covariance and correlation estimates and allows integration of biologically meaningful functional genome annotations. The method of moments has recently been shown to outperform LDSC in single-trait heritability estimation.[11] Integrative analysis of GWAS summary statistics and context-specific functional annotations has provided novel insights into complex disease etiology through a variety of applications.[12–14] In this paper, we introduce GNOVA (genetic covariance analyzer), a principled framework to estimate annotation-stratified genetic covariance using GWAS summary statistics. Through extensive numerical simulations, integrative analysis of 50 complex traits, and an in-depth case study on late-onset Alzheimer disease (LOAD [MIM: 104300]) and amyotrophic lateral sclerosis (ALS [MIM: 105400]), we demonstrate that GNOVA provides accurate covariance estimates and powerful statistical inference that are robust to linkage disequilibrium (LD) and sample overlap. Furthermore, we show that annotation-stratified analysis enhances the interpretability of genetic covariance and provides

[1]Department of Biostatistics, Yale School of Public Health, New Haven, CT 06510, USA; [2]Yale School of Medicine, New Haven, CT 06510, USA; [3]Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT 06510, USA; [4]Yale College, New Haven, CT 06520, USA; [5]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; [6]Division of General Internal Medicine, Department of Medicine, University of Washington, Seattle, WA 98195, USA; [7]VA Cooperative Studies Program Coordinating Center, West Haven, CT 06516, USA
[8]Present address: Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53792, USA
*Correspondence: hongyu.zhao@yale.edu
https://doi.org/10.1016/j.ajhg.2017.11.001.
© 2017 American Society of Human Genetics.

novel insights into the shared genetic basis of complex traits.

Material and Methods

Statistical Model

Here we outline the genetic covariance estimation framework. The complete derivation, detailed justification for all approximations, and theoretical proofs are presented in Appendix A. In short, the genetic covariance that we aim to estimate is the covariance between the genetic effects of a group of single nucleotide polymorphisms (SNPs) on two complex traits. When functional genome annotations are present, we allow such covariance to vary in different annotation categories. Specifically, we define K functional annotations S_1, S_2, ..., S_K (e.g., protein-coding genes and non-coding regions), whose union covers the entire genome; assume two studies share the same list of m SNPs; and assume two standardized traits y_1 and y_2 follow the linear models below:

$$y_1 = \sum_{i=1}^{K} X_i\beta_i + \epsilon$$

$$y_2 = \sum_{i=1}^{K} Z_i\gamma_i + \delta,$$

where X_i and Z_i denote the standardized genotype matrices defined through annotation S_i. Random effects terms β_i and γ_i denote the corresponding genetic effects for each annotation category. SNPs' genetic effects on two traits follow an annotation-dependent covariance structure:

$$\mathbb{E}(\beta_i) = \mathbb{E}(\gamma_i) = 0, i = 1, \ldots, K$$

$$Cov(\beta_i, \gamma_i) = \mathbb{E}(\gamma_i\beta_i^T) = \frac{\rho_i}{m_i} I, i = 1, \ldots, K$$

where m_i and ρ_i denote the total number of SNPs and the total genetic covariance in annotation category S_i, respectively. Random variables ϵ and δ denote the non-genetic effects. Of note, this notation implicitly assumes the genetic covariance to follow an additive structure in regions where functional annotations overlap.

In practice, two different GWASs often share a subset of samples. Without loss of generality, we assume N_1 and N_2 to be the sample sizes of two studies and the first N_s samples in each study are shared. To account for the non-genetic correlation introduced by sample overlapping, we allow random error terms ϵ and δ to be correlated:

$$Cov(\epsilon_i, \delta_j) = \mathbb{E}(\epsilon_i\delta_j) = \begin{cases} \rho_e, 1 \le i = j \le N_s \\ 0, otherwise \end{cases}.$$

We note that our model does not require any additional assumption on the heritability structure of either trait.

Estimation of Covariance Parameters via the Method of Moments

To estimate genetic covariance parameters (i.e., $\rho_i, i = 1, \ldots, K$), we developed an analysis framework based on the method of moments. First, we derive equations that relate the population moments to the parameters of interest. For an arbitrary $N_1 \times N_2$ matrix A, we study the expectation of $y_1^T A y_2$. It can be shown that

$$\mathbb{E}(y_1^T A y_2) = \sum_{i=1}^{K} \frac{\rho_i}{m_i} tr(AZ_iX_i^T) + \rho_e \left(\sum_{t=1}^{N_s} A_{tt} \right).$$

Here, quantity A_{tt} denotes the t^{th} diagonal element of matrix A. Since there are K+1 parameters in total in the model (K genetic covariance parameters and ρ_e), we build a linear system of K+1 equations by plugging in K+1 different matrices A_1,...,A_{K+1} into the equation above. Further, we approximate $\mathbb{E}(y_1^T A_j y_2)$ using the sample moments, i.e., the observed value $y_1^T A_j y_2$, and get the following equation:

$$y_1^T A_j y_2 = \sum_{i=1}^{K} \frac{\rho_i}{m_i} tr(A_j Z_i X_i^T) + \rho_e \sum_{t=1}^{N_s} (A_j)_{tt}, j = 1, \ldots, K+1.$$

Solving this linear system of K+1 equations would get us the method of moments estimators for genetic covariance.

Choices of Matrix *A*

The method of moments estimation procedure described above works for arbitrary A matrices. However, it is critical and non-trivial to choose A in practice. Since individual-level genotype and phenotype data from consortium-based GWASs are in many cases difficult to access, it is of practical interest to estimate genetic covariance based on summary statistics only. To achieve this goal, we define the first K matrices as:

$$\tilde{A}_j = \frac{X_j Z_j^T}{m_j}, j = 1, \ldots, K.$$

Plugging in these matrices, the first K equations become:

$$\frac{1}{m_j}\left(X_j^T y_1\right)^T Z_j^T y_2 = \sum_{i=1}^{K} \frac{\rho_i}{m_i m_j} tr\left(Z_j^T Z_i X_i^T X_j\right) + \frac{\rho_e}{m_j} \sum_{t=1}^{N_s} \left(X_j X_j^T\right)_{tt}, j = 1, \ldots, K.$$

The equality is based on the property of trace and the fact that first N_s samples are shared between two studies. These equations can be further approximated by (Appendix A):

$$\frac{1}{m_j\sqrt{N_1N_2}} (z_1)_j^T (z_2)_j = \sum_{i=1}^{K} \frac{\rho_i}{m_i m_j} \sum_{l=1}^{m_i} \sum_{l'=1}^{m_j} r^2_{l^{(i)}l'^{(j)}} + \frac{N_s\rho_e}{N_1N_2}, j = 1, \ldots, K.$$

Here, $r^2_{l^{(i)}l'^{(j)}}$ denotes the LD between the l^{th} SNP from category S_i and the $(l')^{\text{th}}$ SNP from category S_j; z_1 and z_2 denote the z-scores of SNP-level associations from two GWASs; and $(z_1)_j$ and $(z_2)_j$ represent subsets of z-scores corresponding to the SNPs in annotation category S_j. LD can be estimated using an external reference panel. However, if samples in two studies have different ancestries, $X_i^T X_j$ and $Z_j^T Z_i$ need to be estimated separately using two reference panels. When such reference panels do not exist, individual-level genotype data for a subset of study samples may be needed.

Next, we study the $(K+1)^{\text{th}}$ equation. We define:

$$\tilde{A}_{K+1} = \begin{pmatrix} I_{N_s \times N_s} & 0 \\ 0 & 0 \end{pmatrix}_{N_1 \times N_2}.$$

Divide N_1N_2 on both sides of the $(K+1)^{\text{th}}$ equation, and we get:

$$\frac{1}{N_1N_2} \sum_{t=1}^{N_s} (y_1)_t (y_2)_t = \frac{N_s}{N_1N_2} \sum_{i=1}^{K} \rho_i + \frac{N_s}{N_1N_2}\rho_e.$$

Since $\rho_1, \ldots, \rho_K$ are the parameters of interest, we subtract the $(K+1)^{th}$ equation from the first K equations and remove ρ_{K+1} from the linear system. We denote the remaining K equations in matrix form:

$$\begin{pmatrix} \frac{1}{m_1\sqrt{N_1N_2}}(z_1)_1^T(z_2)_1 - \frac{1}{N_1N_2}\sum_{t=1}^{N_s}(y_1)_t(y_2)_t \\ \vdots \\ \frac{1}{m_K\sqrt{N_1N_2}}(z_1)_K^T(z_2)_K - \frac{1}{N_1N_2}\sum_{t=1}^{N_s}(y_1)_t(y_2)_t \end{pmatrix} =$$

$$\begin{pmatrix} \frac{1}{m_1m_1}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_1} r^2_{l^{(1)}l'^{(1)}} - \frac{N_s}{N_1N_2} & \cdots & \frac{1}{m_Km_1}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_1} r^2_{l^{(K)}l'^{(1)}} - \frac{N_s}{N_1N_2} \\ \vdots & \ddots & \vdots \\ \frac{1}{m_1m_K}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_K} r^2_{l^{(1)}l'^{(K)}} - \frac{N_s}{N_1N_2} & \cdots & \frac{1}{m_Km_K}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_K} r^2_{l^{(K)}l'^{(K)}} - \frac{N_s}{N_1N_2} \end{pmatrix}$$

$$\times \begin{pmatrix} \rho_1 \\ \vdots \\ \rho_K \end{pmatrix}.$$

When the sample sizes of both GWASs are large and the sample overlap between two studies is moderate, the K equations can be approximated by:

$$\begin{pmatrix} \frac{1}{m_1\sqrt{N_1N_2}}(z_1)_1^T(z_2)_1 \\ \vdots \\ \frac{1}{m_K\sqrt{N_1N_2}}(z_1)_K^T(z_2)_K \end{pmatrix} = \begin{pmatrix} \frac{1}{m_1m_1}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_1} r^2_{l^{(1)}l'^{(1)}} & \cdots & \frac{1}{m_Km_1}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_1} r^2_{l^{(K)}l'^{(1)}} \\ \vdots & \ddots & \vdots \\ \frac{1}{m_1m_K}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_K} r^2_{l^{(1)}l'^{(K)}} & \cdots & \frac{1}{m_Km_K}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_K} r^2_{l^{(K)}l'^{(K)}} \end{pmatrix} \begin{pmatrix} \rho_1 \\ \vdots \\ \rho_K \end{pmatrix}.$$

We define

$$v = \left(\frac{1}{m_1\sqrt{N_1N_2}}(z_1)_1^T(z_2)_1, \ldots, \frac{1}{m_K\sqrt{N_1N_2}}(z_1)_K^T(z_2)_K\right)^T$$

$$M = \begin{pmatrix} \frac{1}{m_1m_1}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_1} r^2_{l^{(1)}l'^{(1)}} & \cdots & \frac{1}{m_Km_1}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_1} r^2_{l^{(K)}l'^{(1)}} \\ \vdots & \ddots & \vdots \\ \frac{1}{m_1m_K}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_K} r^2_{l^{(1)}l'^{(K)}} & \cdots & \frac{1}{m_Km_K}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_K} r^2_{l^{(K)}l'^{(K)}} \end{pmatrix}.$$

Then, the point estimate of covariance parameters can be denoted as

$$\hat{\rho} = M^{-1}v.$$

Importantly, M can be estimated using a reference panel (e.g., 1000 Genomes Project[15]) and v is based only on GWAS summary statistics. Of note, the same estimation framework can be directly applied to ascertained case-control studies as well (Appendix A).

Special Cases

Two Independent GWASs

If samples from two GWASs do not overlap, then the non-genetic effects ϵ and δ are independent and only K equations are needed for estimating covariance parameters. We still define $\tilde{A}_j = (X_jZ_j^T)/m_j$ for j = 1,...,K. That gives us the same covariance estimator:

$$\hat{\rho} = M^{-1}v.$$

No Annotation Stratification

If no functional annotation is present, it can be shown that

$$\hat{\rho} = \frac{\overline{z_1z_2}}{\overline{r^2}\sqrt{N_1N_2}}.$$

Here, $\overline{z_1z_2}$ is the average product of z-scores from two GWASs; $\overline{r^2}$ is the average LD across all SNP pairs in the study. Under the non-stratified scenario, this estimator can be seen as a two-trait extension of the heritability estimator proposed in Bulik-Sullivan.[16]

Two GWASs with Substantial Sample Overlap

If the two GWASs have substantial sample overlap, some approximations we have applied in previous sections would fail (Appendix A). The problem gets down to solving the following equations:

$$\begin{pmatrix} \frac{1}{m_1N}(z_1)_1^T(z_2)_1 - \frac{1}{N^2}\sum_{t=1}^{N}(y_1)_t(y_2)_t \\ \vdots \\ \frac{1}{m_KN}(z_1)_K^T(z_2)_K - \frac{1}{N^2}\sum_{t=1}^{N}(y_1)_t(y_2)_t \end{pmatrix} = \begin{pmatrix} \frac{1}{m_1m_1}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_1} r^2_{l^{(1)}l'^{(1)}} & \cdots & \frac{1}{m_Km_1}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_1} r^2_{l^{(K)}l'^{(1)}} \\ \vdots & \ddots & \vdots \\ \frac{1}{m_1m_K}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_K} r^2_{l^{(1)}l'^{(K)}} & \cdots & \frac{1}{m_Km_K}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_K} r^2_{l^{(K)}l'^{(K)}} \end{pmatrix} \begin{pmatrix} \rho_1 \\ \vdots \\ \rho_K \end{pmatrix}.$$

Therefore,

$$\hat{\rho} = M^{-1}\begin{pmatrix} \frac{1}{m_1N}(z_1)_1^T(z_2)_1 - \frac{1}{N}\hat{\rho}_{pheno} \\ \vdots \\ \frac{1}{m_KN}(z_1)_K^T(z_2)_K - \frac{1}{N}\hat{\rho}_{pheno} \end{pmatrix} = M^{-1}\left(v - \frac{\hat{\rho}_{pheno}}{N}\mathbf{1}\right)$$

where the phenotypic correlation $\hat{\rho}_{pheno}$ can be either acquired from the literature or estimated using computational methods[7,17,18] (Appendix A).

Remarks on Overlapping Functional Annotations

When functional annotations overlap, the covariance parameter ρ is not the real quantity of interest. Instead, the total covariance in each annotation category is more biologically meaningful and can be estimated using the weighted estimator

$$\hat{\rho}^W = W\hat{\rho}$$

where W is a $K \times K$ matrix with element

$$W_{ij} = \frac{m_{j\cap i}}{m_j}, 1 \le i, j \le K.$$

Here, $m_{j\cap i}$ denotes the number of SNPs in region $S_i \cap S_j$.

Theoretical Properties

In this section, we establish the statistical optimality of our estimator by showing that it is "almost" the unbiased estimator with minimum variance. Here we state all the propositions (see Appendix A for detailed proofs). Assume y_1 and y_2 follow a multivariate normal distribution:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim MVN\left(0, \begin{pmatrix} H_1 & \Theta \\ \Theta^T & H_2 \end{pmatrix}\right).$$

We begin with calculating the variance of the quadratic form-like quantity $y_1^T A y_2$.

Proposition 1. Let A be an $N_1 \times N_2$ matrix. Then $Var(y_1^T A y_2) = tr(A^T H_1 A H_2) + tr(A^T \Theta A^T \Theta)$.

It can be shown that the second part, i.e., $tr(A^T \Theta A^T \Theta)$, is very small compared to the first term $tr(A^T H_1 A H_2)$ in real GWAS data (Appendix A):

$$tr(A^T H_1 A H_2) \gg tr(A^T \Theta A^T \Theta).$$

With this in mind, the following claim is approximately true:

$$Var(y_1^T A y_2) \approx tr(A^T H_1 A H_2).$$

Next, we define a matrix A_* and show that A_* minimizes $tr(A^T H_1 A H_2)$ under some conditions. Based on the argument above, A_* "almost" minimizes $Var(y_1^T A y_2)$ too.

Proposition 2. Assume two GWASs do not share samples. We define the following quantities.

(i) Let $p = (p_1, \ldots p_K)^T$ be an arbitrarily given K-dimensional vector;
(ii) Let S be a $K \times K$ symmetric matrix with element $S_{ll'} = tr(H_1^{-1} X_{l'} Z_{l'}^T H_2^{-1} Z_l X_l^T)/m_l m_{l'}$ for $1 \le l, l' \le K$;
(iii) Let $\lambda = (\lambda_1, \ldots \lambda_K)^T$ be a vector such that $S\lambda = p$;
(iv) Define $A_* = \sum_{j=1}^{K} \frac{\lambda_j}{m_j} H_1^{-1} X_j Z_j^T H_2^{-1}$.

Then, we have:

(1) $\mathbb{E}(y_1^T A_* y_2) = \sum_{t=1}^{K} p_t \rho_t$;
(2) Let A be a matrix such that $\mathbb{E}(y_1^T A y_2) = \sum_{t=1}^{K} p_t \rho_t$. Then, $tr(A^T H_1 A H_2) \ge tr(A_*^T H_1 A_* H_2)$.

Proposition 2 tells us that given arbitrary p = $(p_1, \ldots p_K)^T$, if $\exists\ \lambda = (\lambda_1, \ldots \lambda_K)^T$ such that $S\lambda = p$, then $y_1^T A_* y_2$ is an unbiased estimator for $\sum_{t=1}^{K} p_t \rho_t$. Furthermore, among all unbiased estimators with the form $y_1^T A y_2$, $y_1^T A_* y_2$ has the minimum value of $tr(A_*^T H_1 A_* H_2)$, hence "almost" the minimum variance $Var(y_1^T A_* y_2)$. Interestingly, by carefully choosing p and λ, we can let A_* equal the $\tilde{A}$ matrix we have been using throughout the paper. Therefore, we have the following corollary.

Corollary 1. We assume:

(i) Two GWASs do not overlap;
(ii) The samples in each study are completely independent;
(iii) True LD in both studies (i.e., $Z^T Z$ and $X^T X$) is known.

Consider all matrices A that suffice

$$tr(AZX^T) = \frac{tr(Z^T Z X^T X)}{m}.$$

We define

$$\hat{\rho}_A = m(y_1^T A y_2)/tr(AZX^T).$$

Then, $\hat{\rho}_{\tilde{A}}$ with $\tilde{A} = (XZ^T)/m$ has the lowest variance.

Similarly, we could extend these results to annotation-stratified scenarios (Appendix A). These results show that although we initially defined $\tilde{A}_j$ for the purpose of simplifying calculation, the derived covariance estimator actually enjoys some good theoretical properties.

Variance Estimation via Block-wise Jackknife

Following previous work,[7] we apply a block-wise jackknife approach to estimate the variance. We divide the genome into b (e.g., $b = 200$) blocks $B_1, \ldots, B_b$. Let

$$v_i^{(t)} = \frac{(z_1)_i^T (z_2)_i - (z_1)_{S_i \cap B_t}^T (z_2)_{S_i \cap B_t}}{(m_i - m_{S_i \cap B_t})\sqrt{N_1 N_2}},\ 1 \le i \le K \text{ and } 1 \le t \le b.$$

Here, subscript $S_i \cap B_t$ indicates the subset of SNPs in both functional annotation S_i and block B_t. Then, $Cov(v)$ is estimated as:

$$\left(\widehat{Cov}(v)\right)_{ij} = \frac{b-1}{b} \sum_{t=1}^{b} \left(v_i^{(t)} - \frac{1}{b}\sum_{s=1}^{b} v_i^{(s)}\right)\left(v_j^{(t)} - \frac{1}{b}\sum_{s=1}^{b} v_j^{(s)}\right).$$

Therefore, we get

$$\widehat{Cov}(\hat{\rho}) = M^{-1}\widehat{Cov}(v)M^{-1}.$$

If annotations overlap,

$$\widehat{Cov}(\hat{\rho}^W) = WM^{-1}\widehat{Cov}(v)M^{-1}W^T.$$

Finally, the test statistic for each covariance parameter is

$$z - score_i = \frac{\hat{\rho}_i}{\sqrt{\left(\widehat{Cov}(\hat{\rho})\right)_{ii}}},\ 1 \le i \le K.$$

When annotations overlap,

$$z - score_i^W = \frac{\hat{\rho}_i^W}{\sqrt{\left(\widehat{Cov}(\hat{\rho}^W)\right)_{ii}}},\ 1 \le i \le K.$$

Genetic Correlation

We provide genetic correlation estimates for non-stratified analysis:

$$cor = \frac{\hat{\rho}}{\sqrt{\hat{h}_1^2 \hat{h}_2^2}}.$$

We use the estimator proposed in Bulik-Sullivan[16] to estimate heritability for each trait:

$$\hat{h}_t^2 = \frac{\frac{1}{m}(z_t)^T(z_t) - 1}{\frac{N_t}{m^2}\sum_{l=1}^{m}\sum_{l'=1}^{m} r_{ll'}^2},\ t = 1, 2.$$

When functional annotations are present, the true heritability in each annotation category may be small. Although methods for estimating annotation-stratified heritability have been proposed,[11,12] they may provide unstable, sometimes even negative, heritability estimates, especially when a number of annotation categories are related to the repressed genome. When true heritability is low, variability in the denominator will have great impact

on genetic correlation estimates. Therefore, we use genetic covariance as a more robust metric when performing annotation-stratified analysis.

Simulation Settings

We simulated quantitative traits using real genotype data from the WTCCC1 cohort. We removed individuals with genetic relatedness coefficient greater than 0.05 and filtered SNPs with missing rate above 1% and/or MAF lower than 5% in samples with European ancestry from the 1000 Genomes Project.[15] In addition, we removed all the strand-ambiguous SNPs. After quality control, 15,918 samples and 254,221 SNPs remained in the dataset. Each simulation setting was repeated 100 times.

Setting 1

We equally divided 15,918 samples into two sub-cohorts. We simulated two traits using genetic effects sampled from an infinitesimal model.

$$\begin{pmatrix}\beta\\ \gamma\end{pmatrix} \sim MVN\left(0, \frac{1}{254221}\begin{pmatrix}h_1^2 I & \rho I\\ \rho I & h_2^2 I\end{pmatrix}\right)$$

Heritability for both traits was set as 0.5. We set the genetic covariance to be 0, 0.05, 0.1, 0.15, 0.2, and 0.25.

Setting 2

Instead of fixing the heritability, we assumed only that the heritability for both traits was equal. Genetic correlation was fixed as 0.2. We set the genetic covariance to be 0.05, 0.1, 0.15, and 0.2 and chose heritability value accordingly.

Setting 3

We simulated two traits on the same sub-cohort of 7,959 samples. Heritability was fixed as 0.5 for both traits. We set the genetic covariance to be 0, 0.05, 0.1, 0.15, 0.2, and 0.25. Sample overlap correction was applied to estimate genetic covariance.

Setting 4

We randomly partitioned the genome into two annotation categories of the same size. We set the heritability for both traits to be 0.5, and the heritability structure does not depend on functional annotations. Genetic covariance in the first annotation was set to be 0, 0.05, 0.1, 0.15, and 0.2. Genetic effects for two traits are not correlated in the second annotation category.

Setting 5

We randomly partitioned the genome into three categories of the same size. Define annotation-1 to be the union of the first and the second categories, and let annotation-2 be the union of the second and the third categories. We set the heritability for both traits to be 0.5, and the heritability structure does not depend on functional annotations. Genetic covariance parameter for annotation-1 (i.e., ρ_1) is set to be 0.1. We set ρ_2 to be -0.2, -0.1, 0, and 0.1. The genetic covariance in regions where two annotations overlap follows an additive structure. For example, when $\rho_1 = 0.1$ and $\rho_2 = 0.2$, the total covariance in annotation-1 is

$$\rho_1 + \frac{\rho_2}{2} = 0.$$

Similarly, the total covariance in annotation-2 is

$$\frac{\rho_1}{2} + \rho_2 = -0.15.$$

GWAS Data Analysis

Details of 48 GWASs and the URLs for summary statistics files are summarized in Table S1. For each summary statistics dataset, we applied the same quality-control steps described in Bulik-Sullivan et al.[7] using the munge_sumstats.py script in LDSC. In addition, we removed all the strand-ambiguous SNPs from each dataset. For each pair of complex traits, we took the overlapped SNPs between two summary statistics files, matched the effect alleles, and removed SNPs with MAF below 5% in the 1000 Genomes Project phase III samples with European ancestry. SNPs on sex chromosomes were also removed from the analysis. We then applied the GNOVA framework to the remaining SNPs to estimate genetic covariance. Sample overlap correction was applied when two GWASs have a large sample overlap. When calculating genetic correlation between ALS and other traits, we used previously reported 0.085 as the heritability of ALS due to negative heritability estimates.[19]

Annotation Data

GenoCanyon and GenoSkyline functional annotations, as previously reported,[14,20,21] integrate various types of transcriptomic and epigenomic data from ENCODE[22] and Roadmap Epigenomics Project[23] to predict functional DNA regions in the human genome. GenoCanyon utilizes an unsupervised learning framework to identify non-tissue-specific functional regions. GenoSkyline and GenoSkyline-Plus further extended this framework to identify tissue- and cell type-specific functionality in the human genome. We applied GenoSkyline-Plus annotations for seven broadly defined tissue categories (i.e., brain, cardiovascular, epithelium, gastrointestinal, immune, muscle, and other) to stratify genetic covariance by tissue type. When integrating these annotations in GNOVA, we also included the whole genome as an annotation category to guarantee that the union of all annotations covers the genome. The whole genome was not added as an additional annotation track in analyses or simulations when the functional annotations covered all SNPs in the dataset. The MAF quartiles were calculated using the genotype data of phase III samples with European ancestry from the 1000 Genomes Project after filtering SNPs with MAF below 5%.

LD Score Regression Implementation

We implemented cross-trait LD score regression using the LDSC software package. For the purpose of fair comparison, we ran LD score regression on all SNPs in the dataset in the simulation studies. When analyzing real GWAS data, we followed the protocol suggested in Bulik-Sullivan et al.[7] and used HAPMAP3 SNPs. LD scores were estimated using phase I samples with European ancestry in the 1000 Genomes Project.

Ethical Statement

Procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation. Proper informed consent was obtained when needed.

Results

Simulations

We simulated two traits using genotype data from the Wellcome Trust Case Control Consortium (WTCCC) while assuming a correlated genetic covariance structure. Detailed simulation settings are described in the Material and Methods. Since LDSC cannot estimate annotation-stratified genetic covariance, we compared GNOVA and LDSC using data simulated from a non-stratified,

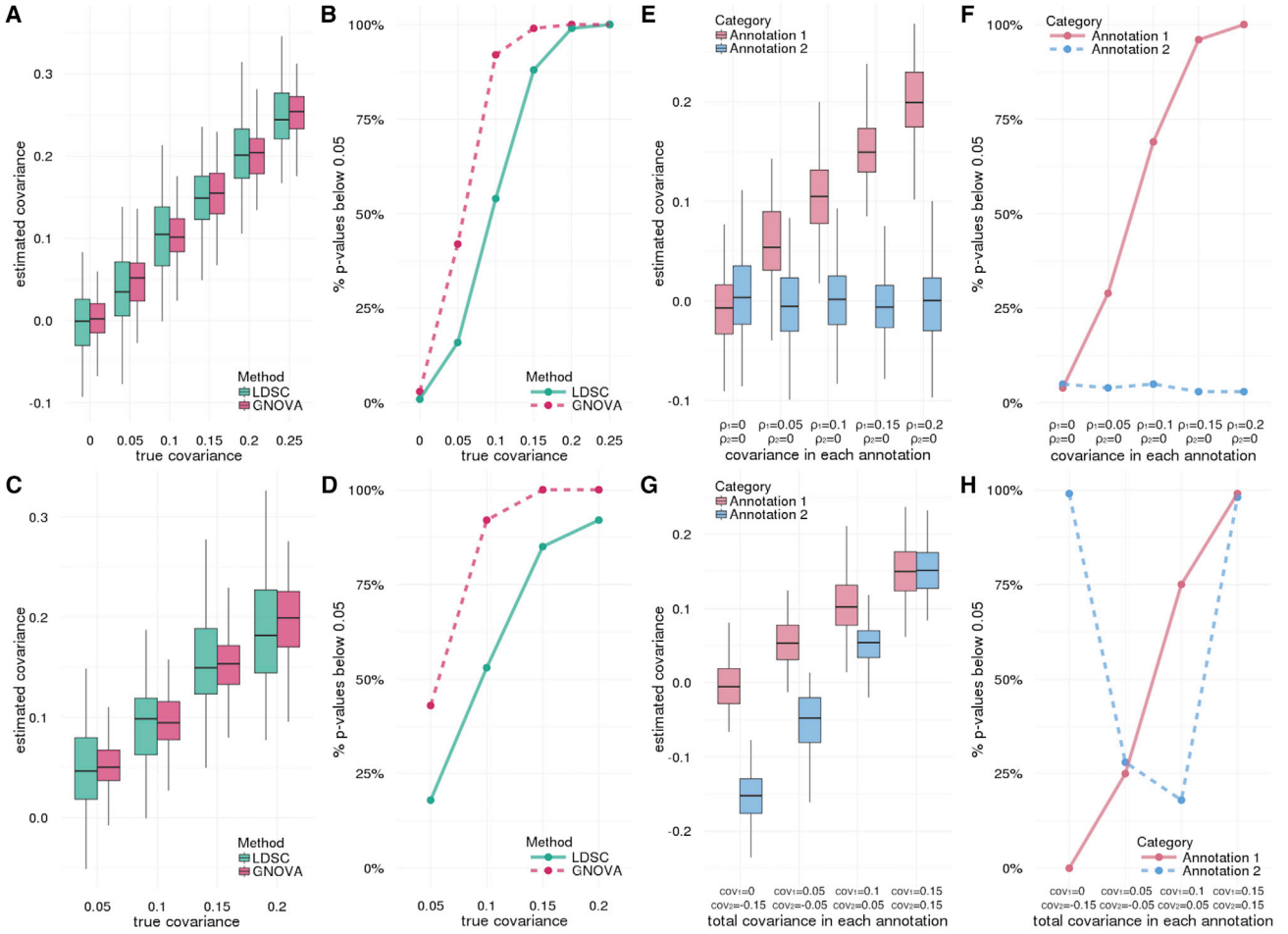

Figure 1. Evaluation of Covariance Estimation and Statistical Power through Simulations
Detailed simulation settings are described in the Material and Methods.
(A–D) Compare GNOVA and LDSC using traits simulated from a non-stratified covariance structure. We first fixed heritability for both traits but set genetic correlation to different values. The covariance estimates are shown in (A). (B) shows the statistical power. Next, we fixed genetic correlation but chose different values for heritability and covariance. Covariance estimates and statistical power are shown in (C) and (D), respectively.
(E–H) Estimate annotation-stratified genetic covariance. In (E) and (F), we simulated data using two non-overlapping functional annotations. Results in (G) and (H) are based on two overlapping annotations. The true covariance values are labeled under each setting. Type I error was not inflated when the true covariance was zero.

infinitesimal genetic covariance structure (Figures 1A–1D). Both methods provided unbiased covariance estimates, but GNOVA estimator had consistently lower variance across all simulation settings. The same pattern could be observed for genetic correlation estimates (Figure S1). Neither method showed inflated type I error when the true covariance is 0. When comparing the frequencies of rejecting the null hypothesis, GNOVA is nearly twice as powerful as LDSC when the true genetic covariance is below 0.1. To evaluate GNOVA's robustness against sample overlap, we simulated two traits using genotype data of the same cohort. After applying sample overlap correction, GNOVA still outperformed LDSC, showing higher estimation accuracy and statistical power (Figure S2).

Next, we investigated GNOVA's capability to estimate annotation-stratified genetic covariance. We randomly partitioned the genome into two non-overlapping annotation categories and simulated two traits using annotation-dependent genetic covariance (Material and Methods). GNOVA provided unbiased estimates for the genetic covariance in each category across all settings (Figures 1E and 1F). Of note, type I error was well controlled in the annotation category without genetic covariance even when the true covariance in the other annotation category was non-zero, suggesting GNOVA's robustness under the influence of LD. Furthermore, when functional annotations overlapped, our method still provided accurate covariance estimates and powerful inference (Figures 1G and 1H).

Estimation of Pairwise Genetic Correlation for 48 Human Complex Traits

We applied GNOVA to estimate genetic correlations for 48 complex traits using publicly available GWAS summary

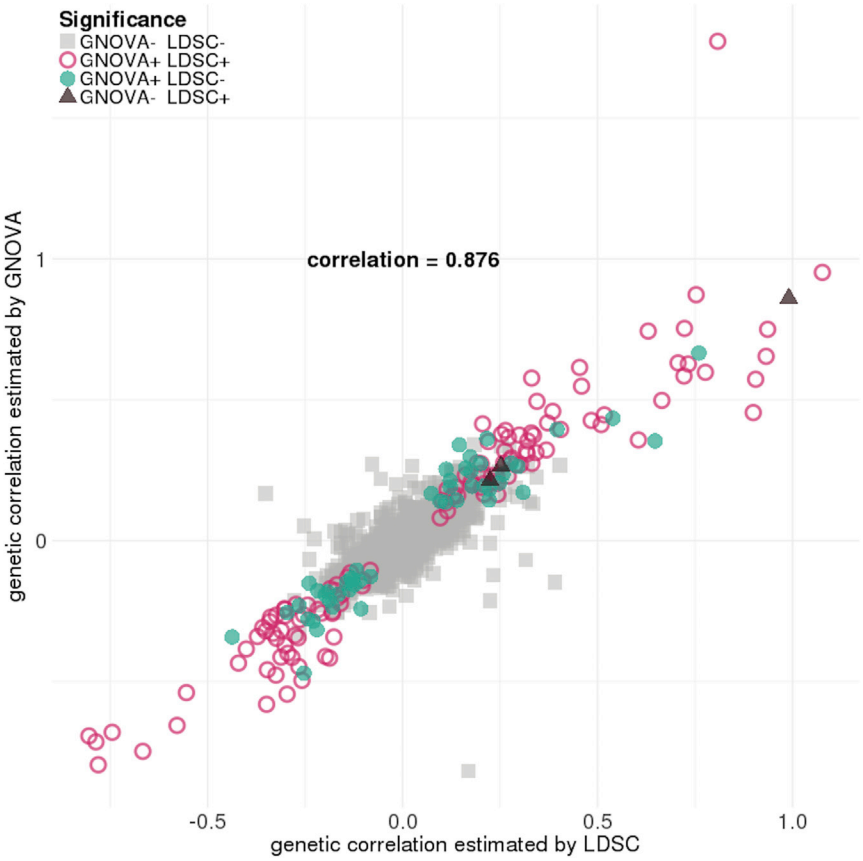

Figure 2. Comparison of Genetic Correlations Estimated via GNOVA and LDSC
Each point represents a pair of traits. Overall, genetic correlation estimates are concordant between GNOVA and LDSC, but GNOVA is more powerful when genetic correlation is moderate. Color and shape of each data point represent the significance status given by GNOVA and LDSC. Trait pairs that involve gout were removed from this figure because LDSC estimated its heritability to be negative and could not properly output p values.

statistics ($N_{total} \approx 4.5$ million). Trait acronyms and other details of all GWASs are summarized in Table S1. Out of 1,128 pairs of traits in total, we identified 176 pairs with statistically significant genetic correlation after Bonferroni correction (Table S2 and Figure S3). We also applied LDSC to the same datasets and identified only 127 significant pairs (Table S3 and Figure S4). A total of 52 significantly correlated trait pairs were uniquely identified by GNOVA while only 3 trait pairs were uniquely identified using LDSC. Overall, the genetic correlations estimated using GNOVA and LDSC are concordant (Figure 2). Consistent with our simulation results, GNOVA is more powerful when genetic correlation is moderate.

To evaluate model validity, we examined correlations between several traits that are closely related either physiologically or epidemiologically (Table S4). As expected, femoral and lumbar bone mineral density (FNBMD and LSBMD) and depressive symptoms (DEP) and major depressive disorder (MDD [MIM: 608516]) showed strong positive genetic correlations. We also observed negative correlations between subjective well-being (SWB) and neuropsychiatric disorders such as schizophrenia (MIM: 181500), anxiety (MIM: 607834), two depression traits (DEP and MDD), and neuroticism.

We further examined pairwise correlations between 48 traits (Figures 3 and S3). Following hierarchical clustering, broad patterns suggesting disease relatedness emerged. These results are well documented in the literature: neuropsychiatric conditions, metabolic diseases, and gastrointestinal inflammatory disorders clustered together with positive correlations within each individual cluster. We replicated several previous genetic correlation findings,[7] including significant correlations of adult height (HGT) with coronary artery disease (CAD [MIM: 608320]) and age at menarche (AM), and of years of education (EDU) with CAD, bipolar disorder (BIP), body-mass index (BMI), triglycerides, and smoking status (SMK). Furthermore, two previous results that passed multiple correction testing at only 1% FDR passed Bonferroni correction in our analysis; namely, we observed a statistically significant negative correlation between AM and CAD and a positive correlation between autism (ASD [MIM: 209850]) and EDU.

We also identified a number of genetic correlations that are consistent with the genetic relationships reported in the previous literature. For example, previous genetic correlation analyses identified a negative correlation between anorexia nervosa (AN [MIM: 606788]) and obesity, a result we also observed.[7] In addition, we found negative correlations of AN with glucose and triglyceride levels, as well as a positive correlation with high-density lipoprotein (HDL). These results provide further support for existing hypotheses proposing an underlying neural, rather than metabolic, etiology for metabolic syndrome.[12,21,24] We see an unsurprising positive correlation between glucose and insulin levels, which is consistent with our understanding of diabetes.[25] Positive correlations between multiple sclerosis (MS [MIM: 126200]) and Crohn disease (CD) and more generally, inflammatory bowel disease (IBD [MIM: 266600]), agree with existing reports of shared susceptibility for these diseases.[26–28] We demonstrate a positive correlation between asthma (MIM: 600807) and eczema (MIM: 603165), which share numerous loci identified in previous GWASs.[29] We also reproduced recent findings linking bone mineral density with metabolic dysfunction with positive correlations between FNBMD and both glucose and type II diabetes (T2D [MIM: 125853]).[30]

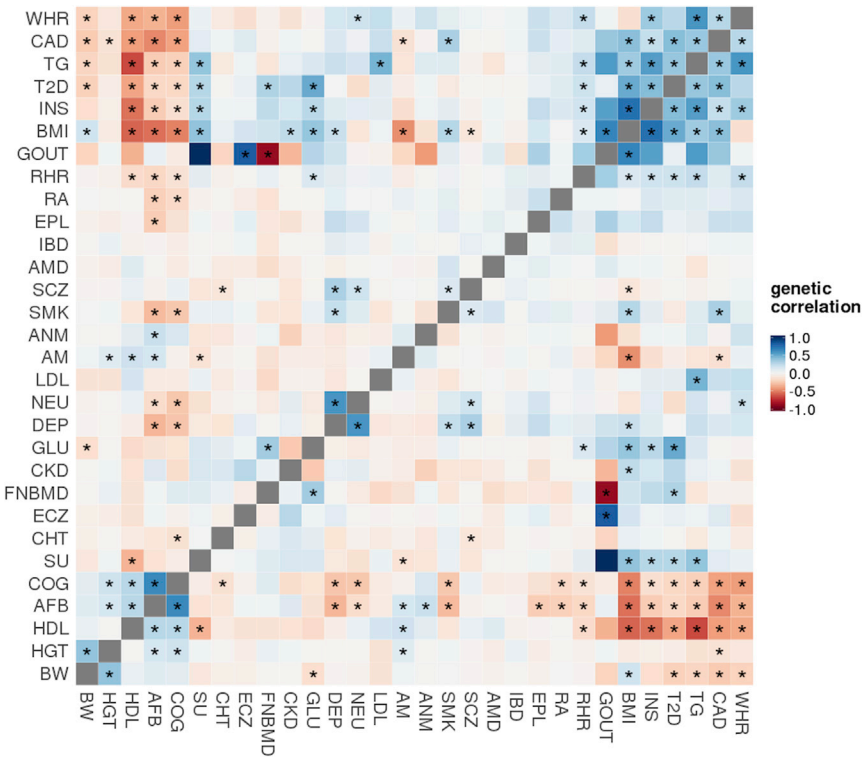

Figure 3. Estimated Genetic Correlations of 435 Pairs of Traits from 30 GWASs
To visualize a large number of pairwise correlations more efficiently, we excluded closely related traits and studies with smaller sample sizes (N < 30,000) in this figure. Asterisks highlight significant genetic correlations after Bonferroni correction for all 1,128 pairs ($p < 4.4 \times 10^{-5}$). The complete heatmap matrix is presented in Figure S3. The order of traits was determined by hierarchical clustering.

Interestingly, however, we did not see significant correlations of bone mineral density with cardiovascular diseases. Among neuropsychiatric disorders, we identified positive correlations between BIP and both depression and neuroticism. Associations between neuroticism and depression are well documented. Neuroticism is highly comorbid with MDD,[31,32] and our findings are consistent with previously observed genetic pleiotropy among neuroticism, MDD, BIP, and schizophrenia.[33,34]

Especially notable are findings that suggest a genetic basis for associations between traits regarding which the literature is either equivocal or absent, and which provide useful information to guide further study. For example, we observed correlations of serum urate (SU) with AM (−0.12), T2D (0.275), and triglycerides (0.38), and we consistently observed associations of SU and markers of metabolic syndrome. In the literature, the genetic architecture of this association has not been extensively studied.[35] Alleles in *IRF8* (MIM: 601565), a regulatory factor of type I interferons, are associated with MS and systemic lupus erythematosus (SLE [MIM: 152700]), but with opposite effect; high type I IFN titers are thought to be causal in SLE but are lower in MS relative to healthy controls.[36] In this analysis, however, we found a positive correlation between MS and SLE. We also draw attention to the significant negative correlation between MS and ASD. This replicates a previous genetic association between MS and ASD, with more recent evidence suggesting shared biomedical markers, such as increase in concentrations of tumor necrosis factor-alpha (TNF-α) in serum in ASD and in cerebrospinal fluid in MS.[37,38] However, previous treatment of MS with anti-TNF-α led to an increase in the number of demyelinating lesions and a significantly higher relapse rate.[39] Furthermore, we observed a positive genetic correlation between ulcerative colitis (UC) and primary billary cirrhosis (PBC [MIM: 109720]). CD, also an IBD and thus closely related, has been reported to share susceptibility genes with PBC including *TNFSF15* (MIM: 604052), *ICOSLG* (MIM: 605717), and *CXCR5* (MIM: 601613).[40] Here we show that ulcerative colitis may also be genetically related to PBC.

Stratification of Genetic Covariance by Functional Annotation

In this section, we apply functional annotations to further dissect the shared genetic architecture of 48 complex traits. We have previously developed GenoCanyon, a statistical framework to predict functional DNA elements in the human genome through integration of annotation data.[20] We partitioned the genome into two non-overlapping categories (i.e., functional and non-functional) based on GenoCanyon scores (Material and Methods) and estimated genetic covariance within the functional and the non-functional genome for each pair of traits (Table S5). The total genetic covariance estimated using the stratified model is highly concordant with covariance estimated using the non-stratified model (Figure 4A). However, genetic covariance is enriched in the predicted functional genome for most traits (Figure 4B). Based on this approach, we identified one more pair of correlated traits, i.e., low-density lipoprotein (LDL) and total cholesterol (TC), whose genetic covariance largely concentrated in the predicted functional genome and achieved significance ($\rho_{func} = 0.060$; $p = 1.0 \times 10^{-6}$) while the overall covariance did not ($\rho_{overall} = 0.062$; $p = 7.7 \times 10^{-5}$).

Next, we partitioned genetic covariance based on quartiles of SNPs' minor allele frequencies (MAFs) in subjects with European ancestry from the 1000 Genomes Project (Material and Methods; Table S6). Similar to the previous analysis, we identified high concordance

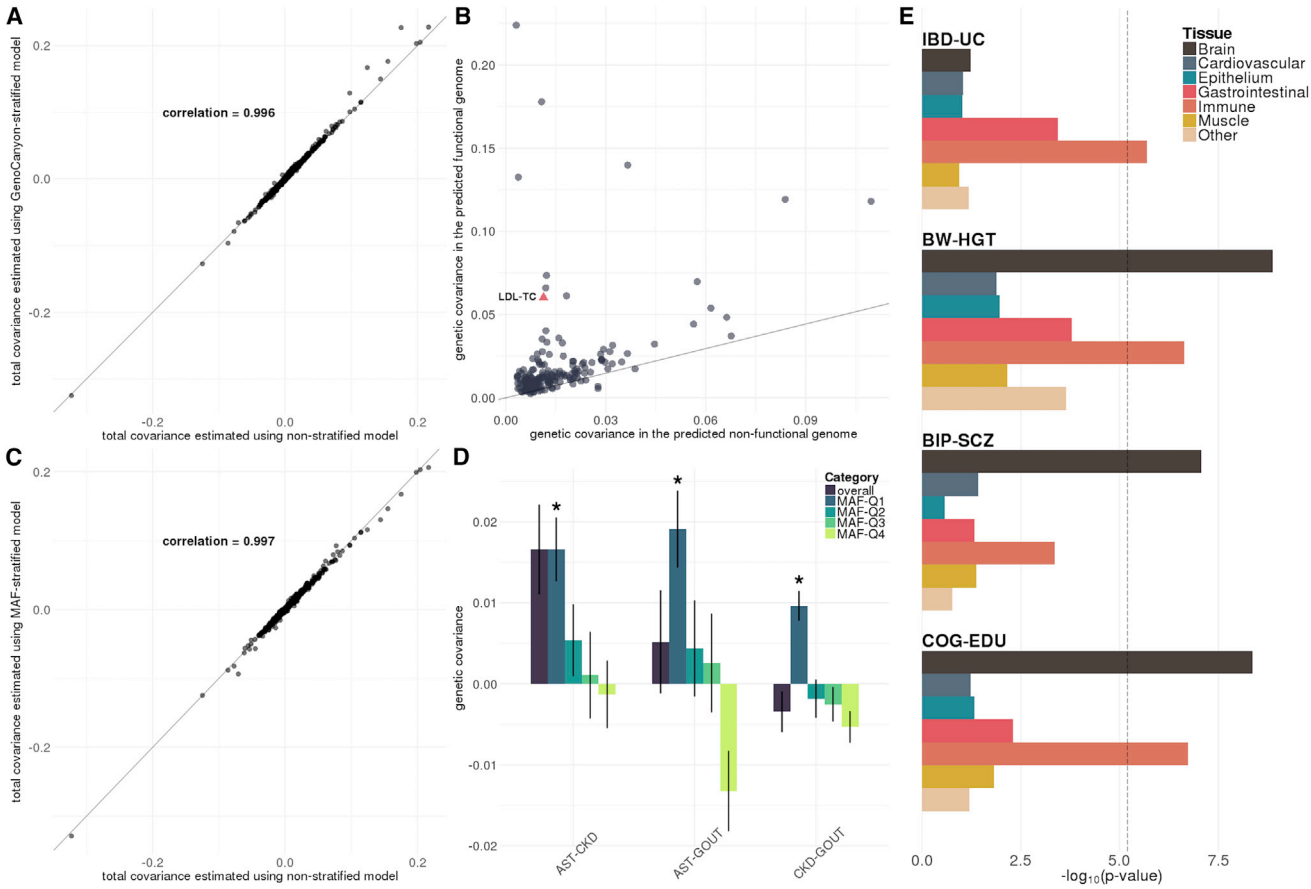

Figure 4. Annotation-Stratified Covariance Analysis

(A) Stratify genetic covariance by genome functionality predicted by GenoCanyon. Total genetic covariance estimates were highly concordant between stratified and non-stratified models.

(B) For significantly correlated pairs of traits based on the non-stratified model, we compared genetic covariance in the functional and the non-functional genome. Solid line marks the expected value based on annotation's size. Trait pair LDL-TC is also plotted.

(C) Stratify genetic covariance by MAF quartile. We compared the genetic covariance estimated by MAF-stratified and non-stratified models.

(D) Six pairs of traits that are uniquely correlated in the lowest MAF quartile. Intervals show the standard error of covariance estimates. Asterisks indicate p values below 4.4×10^{-5}.

(E) Stratify genetic covariance by tissue type. Each bar denotes the log-transformed p value. Dashed line highlights the Bonferroni-corrected significance level $0.05/(7 \times 1128) = 6.3 \times 10^{-6}$.

between the total covariance estimated using MAF-stratified model and the covariance estimates based on non-stratified model (Figure 4C). Overall, the estimated genetic covariance in four MAF quartiles was comparable (Figure S5). However, we identified three pairs of traits that are uniquely correlated in the lowest MAF quartile (Figure 4D), namely asthma with chronic kidney disease (CKD; $p = 1.8 \times 10^{-5}$), gout (MIM: 138900) with CKD ($p = 4.2 \times 10^{-8}$), and asthma with gout ($p = 4.4 \times 10^{-5}$). For several trait pairs, covariance in the lowest MAF quartile showed reversed direction compared to other quartiles. Covariance between CKD and gout even showed reversed direction compared to the estimated total covariance, highlighting the distinction in how common and less common variants are involved in the shared genetic architecture between these traits. Our findings also hint at the possible selection pressure on DNA variations contributing to metabolic traits including CKD and gout, as well as immune diseases including asthma.

Finally, we studied tissue specificity of genetic covariance through integration of GenoSkyline-Plus annotations (Material and Methods). GenoSkyline-Plus integrates multiple epigenomic and transcriptomic annotations from the Roadmap Epigenomics Project to identify tissue- and cell type-specific functional regions in the human genome.[14] We utilized seven broadly defined tissue and cell types (i.e., brain, cardiovascular, epithelium, gastrointestinal, immune, muscle, and other) to stratify genetic covariance for 1,128 pairs of traits (Table S7). Six tests from four pairs of traits passed Bonferroni correction, i.e., $p < 0.05/(1{,}128 \times 7) = 6.3 \times 10^{-6}$ (Figures 4E and S6). As expected, UC, as an IBD, was significantly and positively correlated with IBD in immune-related functional genome ($p = 2.0 \times 10^{-6}$), and two psychiatric diseases, BIP and schizophrenia, were specifically correlated in the genome

Table 1. Dissection of Genetic Covariance between LOAD and ALS

Annotation	Category	Covariance	p Value
Non-stratified	GNOVA	0.016 (0.004)	$*2.0 \times 10^{-4}$
	LDSC	0.012 (0.007)	0.075[a]
GenoCanyon	functional	0.016 (0.004)	$*8.2 \times 10^{-5}$
	non-functional	0.003 (0.004)	0.377
MAF	Q1	−0.001 (0.003)	0.842
	Q2	0.003 (0.004)	0.361
	Q3	0.004 (0.004)	0.327
	Q4	0.008 (0.003)	*0.005

Numbers in parentheses indicate standard errors. Significant p values after adjusting for multiple testing within each section are indicated by an asterisk (*). [a]p value in LDSC was calculated from genetic correlation instead of genetic covariance.

predicted to be functional in brain ($p = 8.7 \times 10^{-8}$). In addition, we identified cognitive function (COG) and EDU, and birth weight (BW) and HGT to be significantly correlated in both brain- and immune-related functional genome. Of note, since the sizes of functional annotations are linked to statistical power, p values here should not be interpreted as reflecting the importance of each tissue. Some tissues may be critically involved in the etiology of analyzed traits even if they may have p values that are not statistically significant. For example, IBD and UC were substantially correlated in the gastrointestinal tract ($p = 3.7 \times 10^{-4}$). Many of these tests may become significant in the near future as GWASs with larger sample sizes are published.

Dissection of Shared and Distinct Genetic Architecture between LOAD and ALS

LOAD and ALS are neurodegenerative diseases. Despite success of large-scale GWASs,[19,41] our understanding of their genetic architecture is still far from complete. We applied GNOVA to dissect the genetic covariance between LOAD and ALS using publicly available GWAS summary statistics ($N_{LOAD} = 54{,}162$; $N_{ALS} = 36{,}052$; Table S8).

We identified positive and significant genetic correlation between LOAD and ALS (correlation = 0.175, $p = 2.0 \times 10^{-4}$). LDSC provided similar estimates but failed to achieve significance (Table 1). 82.6% of the total genetic covariance between LOAD and ALS is concentrated in 33% of the genome predicted to be functional by GenoCanyon ($p = 8.2 \times 10^{-5}$). Furthermore, MAF-stratified analysis showed that 54.6% of the covariance could be explained by the SNPs in the highest MAF quartile ($p = 0.005$). In fact, genetic covariance is lower with lower MAF, and covariance in the lowest MAF quartile is nearly negligible. This is surprising considering that the heritability of ALS is enriched in variants with lower MAF.[19] We also performed tissue-stratified analysis using GenoSkyline-Plus annotations (Table S9). No tissue passed the significance threshold after multiple testing correction, but covariance is more concentrated in immune, brain, and cardiovascular functional genome, and showed nominal significance in the immune annotation track ($p = 0.014$). Whether this will lead to a potential neuroinflammation pathway shared between LOAD and ALS remains to be studied in the future using larger datasets.

Next, we stratified genetic covariance between LOAD and ALS by chromosome. Somewhat surprisingly, we did not observe a linear relationship between per-chromosome genetic covariance and chromosome size (Figure 5A) given that the overall genetic covariance is positive and significant. Since we have observed the concentration of genetic covariance in the functional genome, we further partitioned each chromosome by genome functionality. We identified a clear and positive linear relationship between genetic covariance in the functional genome and the size of predicted functional DNA on each chromosome (Figure 5B). The correlation between per-chromosome genetic covariance in the non-functional genome and the size of non-functional chromosome is negative and significantly smaller than the corresponding quantity in the functional genome (Figure S7; $p = 0.044$; tested using Fisher transformation). Our findings suggest a polygenic covariance architecture between LOAD and ALS and highlight the importance of stratifying genetic covariance by functional annotation.

Finally, we jointly analyzed LOAD, ALS, and 48 other complex traits (Table S10). Interestingly, LOAD and ALS showed distinct patterns of genetic correlations with other complex traits (Figure 6). We identified negative and significant correlations between LOAD and cognitive traits including COG and EDU. HGT and age at first birth (AFB), two traits related to hormonal regulation as well as socio-economic status, were also significantly and negatively correlated with LOAD. Consistent with previous reports, we did not identify substantial correlation between LOAD and other neurological and/or psychiatric diseases.[7,9] We identified negative correlations between LOAD and gastrointestinal inflammatory diseases including a significant correlation with PBC. Asthma and eczema were both positively correlated with LOAD, suggesting a complex genetic relationship between LOAD and different immune-related diseases. Although some of these traits had the same correlation direction with ALS, none of them were significant. Instead, ALS was significantly and positively correlated with MS, a neurological disease with a well-established immune component.[42] ALS was also positively correlated with several other immune-related diseases including celiac disease (CEL [MIM: 212750]), asthma, PBC, and IBD (including CD and UC), though none of these were statistically significant. The nominal correlations between ALS and neurological and psychiatric diseases including

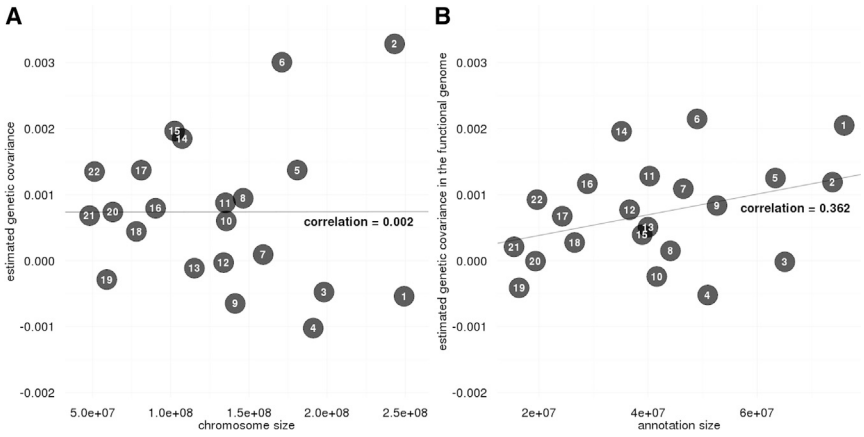

Figure 5. Stratification of Genetic Covariance between LOAD and ALS by Chromosome

(A) Comparisons of the estimated per-chromosome genetic covariance with chromosome size.

(B) Comparisons of the estimated genetic covariance in the predicted functional genome on each chromosome with size of the functional genome.

epilepsy, schizophrenia, BIP, AN, and MDD also remain to be validated in the future using studies with larger sample sizes.

Discussion

Although our understanding of complex disease etiology is still far from complete, we have gained valuable knowledge about the genetic architecture of numerous complex traits from large-scale association studies, partly due to advances in statistical genetics. First, a large proportion of trait heritability can be explained by SNPs that do not pass the Bonferroni-corrected significance threshold.[1] Therefore, it is often helpful to utilize genome-wide data instead of focusing only on significant SNPs in post-GWAS analyses. Second, sample size is critical for many statistical genetics applications. However, individual-level genotype and phenotype data from consortium-based GWASs are not always easily accessible due to policy and privacy concerns. Thanks to the great efforts from large international collaborations such as the Psychiatric Genomics Consortium in promoting open science and data sharing, it has become a tradition for GWAS consortia to share summary statistics to the broader scientific community. Therefore, it is of practical interest to use GWAS summary statistics as the input of downstream analytical methods.[8] Finally, integration of high-throughput transcriptomic and epigenomic annotation data has been shown to improve statistical power as well as interpretability in many recent complex trait studies.[12–14] As large consortia such as ENCODE[22] and Roadmap Epigenomics Project[23] continue to expand, integrative approaches based on functional genome annotations will become an even greater success. In this paper, we developed a novel method to estimate and partition genetic covariance between complex traits. Our method enjoys all the aforementioned advantages. It requires only genome-wide summary statistics and a reference panel as input and allows stratification of genetic covariance by functional genome annotation, which provides novel insights into the shared genetic basis between complex traits and, in some cases, improves the statistical power.

Numerous studies have hinted at a shared genetic basis among neurodegenerative diseases.[43,44] Due to the convenience and efficiency of LDSC and the wide accessibility of GWAS summary statistics, several attempts have been made to estimate genetic correlation between neurodegenerative diseases.[9,45] To date, these efforts have not been as successful as similar studies on psychiatric diseases and immune-related traits. One reason is that existing methods may not be statistically powerful enough to identify moderate genetic correlation using GWASs with limited sample sizes. In addition, the shared genetics among neurodegenerative diseases may not fit the global, infinitesimal covariance structure that most existing tools are based on. In this study, we applied GNOVA to dissect the genetic covariance between LOAD and ALS, two major neurodegenerative diseases, using summary statistics from the largest available GWASs. Our findings suggest that covariance between LOAD and ALS is concentrated in the predicted functional genome and in very common SNPs. Moreover, after applying functional annotations to stratify the genome, estimated per-chromosome genetic covariance is proportional to chromosome size, suggesting a shared polygenetic architecture between LOAD and ALS and also demonstrating the importance of incorporating predicted genetic activity with GenoCanyon. In addition, joint analysis with 50 complex traits also revealed distinctive genetic covariance profiles for LOAD and ALS. LOAD is negatively correlated with multiple traits related to cognitive function and hormonal regulation, while ALS is positively correlated with MS and a few other immune-related traits. Our findings provided novel insights into the shared and distinct genetic architecture between LOAD and ALS and also further demonstrated the benefits of incorporating functional genome annotations into genetic covariance analysis.

Also of note are findings involving serum urate. SU was positively correlated with gout but also with a few metabolic traits. Gout is an arthritic inflammatory process caused by deposition of uric acid crystals in joints, and the role of hyperuricemia in gout is well established. More recently, a role for hyperuricemia in the pathophysiology of metabolic syndrome and CKD has been

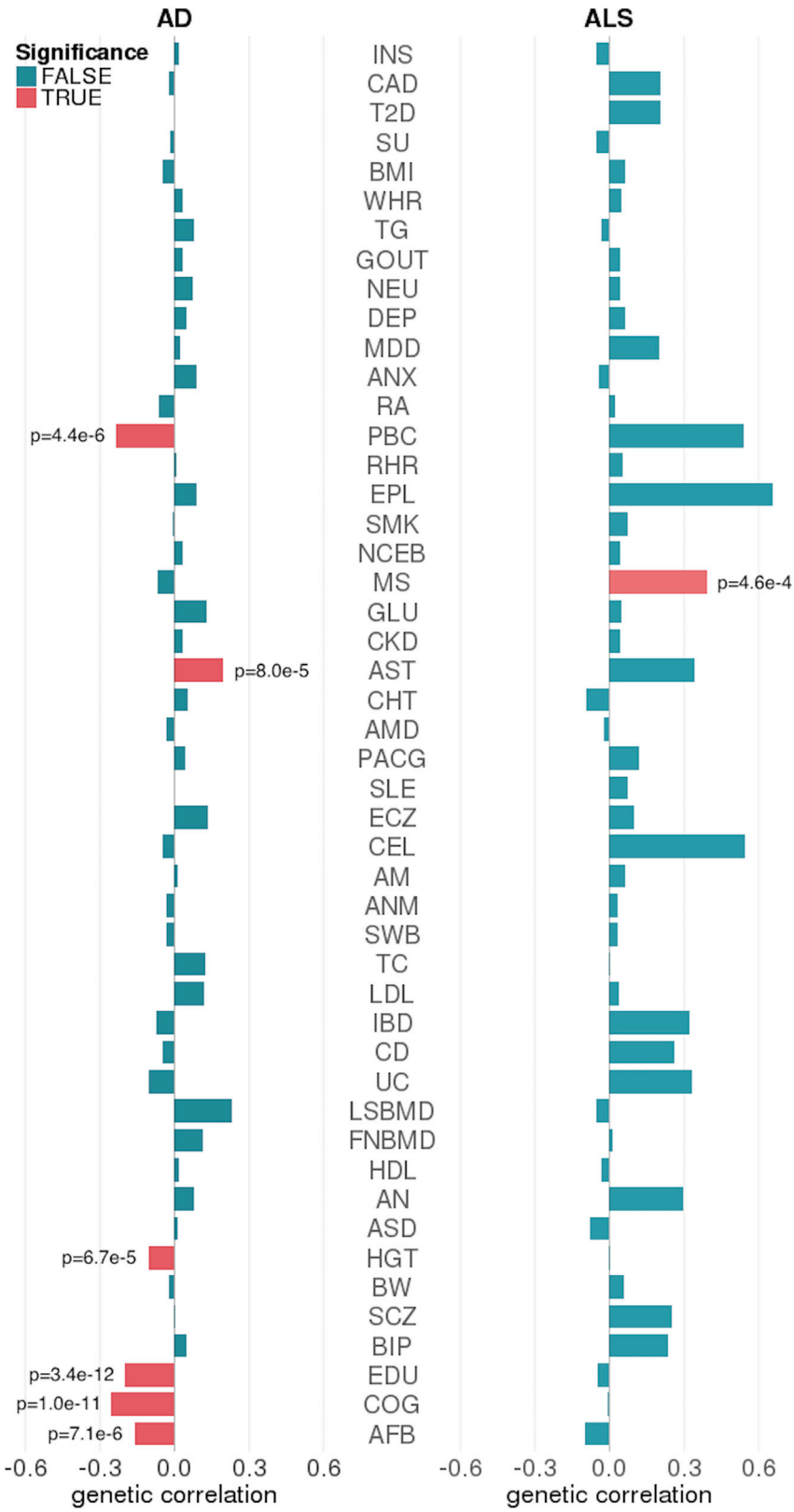

Figure 6. Genetic Correlations between LOAD, ALS, and 48 Complex Traits
Significant pairs with $p < 0.05/(48 \times 2) = 5.2 \times 10^{-4}$ are highlighted in red.

suggested.[46] While associations between hyperuricemia and cardiovascular disease are well described,[47] multiple hypotheses exist regarding details of its involvement.[48] For example, hyperuricemia may lead to inflammation in the kidney through vascular smooth muscle proliferation, inducing hypertension via pre-glomerular vascular changes.[49] It has also been shown to induce oxidative stress in various settings; in adipocytes and islet cells, this may be involved in development of diabetes, and it may also result in impaired endothelin function and activation of the renin-angiotensin-aldosterone system, leading to hypertension.[50–53] Despite this evidence, genetic investigations have not identified a strong relationship between hyperuricemia and metabolic syndrome. Polymorphism in gene *SLC22A12* (MIM: 607096) was associated with hyperuricemia but not with metabolic syndrome.[54] Mendelian randomization studies showed an association between uric acid and gout but did not find an association with T2D or cardiovascular risk factors such as hypertension, glucose, or CAD.[55,56] Our results suggest that GNOVA successfully isolated a signal of biological and clinical significance that provides important impetus for further inquiry in the etiology of metabolic syndrome.

Dissecting relationships among complex traits is a major goal in human genetics research. Genetic covariance is a useful metric to quantify such relationships, but it has its limitations. First, genetic covariance implicitly imposes a strong assumption on the shared genetic basis between complex traits. Not only may the same set of genetic components affect multiple traits, but their effect sizes on both traits are also assumed to be proportional. In the future, it is of interest to extend our method to estimate more generalized metrics, e.g., consistency in effect directions. Second, genetic covariance analysis does not highlight specific DNA segments with pleiotropic effects. Several SNP-based methods have been developed to identify pleiotropic associations using GWAS summary statistics.[57,58] However, due to the large number of SNPs in the genome, statistical

power is a critical issue and large-scale inference remains challenging. In addition, we have demonstrated that integrating functional annotations into genetic covariance analysis could reveal subtle structures in shared genetics between complex traits, but interpretation of genetic covariance remains a challenge. Pickrell et al. recently proposed an approach to distinguishing causal relationships among traits from pleiotropic effects via independent biological pathways.[59] Han et al. developed a method to distinguish pleiotropy from phenotypic heterogeneity.[60] Although many questions remain unanswered, these recent studies have broadened our view on interpreting complex genetic relationships between human traits. Further, statistical power in genetic covariance analysis will be reduced if the shared genetic components have discordant effect directions on different traits. This problem can be partly addressed by the aforementioned SNP-based methods. Recently, Shi et al. developed a method to estimate local heritability and genetic correlation.[61,62] This approach provides an alternative methodological option for analyzing genetic effects at specific loci. Finally, we note that common SNPs in GWASs do not fully explain phenotypic similarity. For example, the estimated genetic covariance among lipid traits explains only 10%–15% of their phenotypic covariance available on LD Hub.[10] Other factors such as rare variants, copy-number variations, and environmental factors may have substantial contributions to the phenotypic covariance among complex traits. Dissection of these complex relationships will be an interesting topic to pursue in the future. Our method, in conjunction with many other tools, provides the most complete picture to date about shared genetics between complex phenotypes.

In summary, we developed GNOVA, a novel statistical framework to perform powerful, annotation-stratified genetic covariance analysis using GWAS summary statistics. Through theoretical proof, we have established GNOVA's statistical optimality within the framework of method of moments. Compared to LD score regression, GNOVA provides more accurate genetic covariance estimates and powerful statistical inference. Its unique feature of performing annotation-stratified analysis also adds depth to existing analysis strategies. Using GNOVA, we were able to expand the discovery of genetic covariance among a spectrum of common diseases and complex traits. Our findings shed light onto the shared and distinct genetic architecture of complex traits. As the sample sizes in genetic association studies continue to grow, our method has the potential to continue identifying shared genetic components and providing novel insights into the etiology of complex diseases.

Appendix A

Model Details

We begin with introducing a general scenario. Assume two standardized traits y_1 and y_2 follow a linear model:

$$y_1 = X\beta + \epsilon$$

$$y_2 = Z\gamma + \delta.$$

Matrices X and Z denote the standardized genotype information for two GWASs. To simplify the algebra, we assume both the genotypes (X and Z) and phenotypes (y_1 and y_2) are standardized. We define K possibly overlapping functional annotations S_1, S_2, …, S_K. All together, these annotations cover the entire genome. We assume two studies share the same list of m SNPs. Vectors β and γ are random effect terms that quantify the genetic effects on traits y_1 and y_2, respectively. Variables ϵ and δ denote the non-genetic effects. Genetic and non-genetic effects on the same trait are assumed to be independent. A SNP's genetic effects on two different traits can be correlated. The genetic covariance depends on functional annotations and follows an additive structure in regions where functional annotations overlap. Specifically, we have

$$\mathbb{E}\left(\beta_{SNP_j}\right) = \mathbb{E}\left(\gamma_{SNP_j}\right) = 0 \; and \; \mathbb{E}\left(\gamma_{SNP_j}\beta_{SNP_j}\right) = \sum_{c:j\in S_c} \frac{\rho_c}{m_c},$$
$$j = 1, \ldots, m$$

$$\mathbb{E}\left(\gamma_{SNP_i}\beta_{SNP_j}\right) = 0, i \neq j,$$

where m_c denotes the total number of SNPs in annotation S_c. Notation $j \in S_c$ indicates that the j^{th} SNP is located in functional annotation S_c. If we use X_i and Z_i to denote the genotype matrices within annotation S_i (some SNPs may be counted multiple times if the functional annotations overlap) and use β_i and γ_i to denote the corresponding genetic effects, the model can be equivalently re-written as follows:

$$y_1 = \sum_{i=1}^{K} X_i\beta_i + \epsilon$$

$$y_2 = \sum_{i=1}^{K} Z_i\gamma_i + \delta$$

$$\mathbb{E}\left(\gamma_i\beta_i^T\right) = \frac{\rho_i}{m_i} I, i = 1, \ldots, K.$$

In practice, two different GWASs often share a subset of samples. Without loss of generality, we assume N_1 and N_2 to be the sample sizes of two studies and the first N_S samples in each study are shared. Therefore, the first N_S rows of matrices X_i and Z_i ($i = 1, \ldots, K$) are identical. To account for the non-genetic correlation introduced by sample overlapping, we allow random error terms ϵ and δ to be correlated.

$$\mathbb{E}\left(\epsilon_i\delta_j\right) = \begin{cases} \rho_e, 1 \leq i = j \leq N_s \\ 0, otherwise \end{cases}$$

To summarize, this framework explicitly models the annotation-stratified genetic covariance in the genome. It also allows functional annotations to overlap, which is important when applied to real-world annotation data. Furthermore, we take the sample overlap between different GWASs into account. Finally, our model does not require any additional assumption on the heritability structure. In following sections, we discuss how to estimate covariance parameters $\rho_1, \ldots, \rho_K$.

Estimate Covariance Parameters

First, for an arbitrary $N_1 \times N_2$ matrix A, we study the expectation of $y_1^T A y_2$.

$$\begin{aligned}
\mathbb{E}(y_1^T A y_2) &= \mathbb{E}\left(\left(\sum_{i=1}^{K} \beta_i^T X_i^T + \epsilon^T\right) A \left(\sum_{i=1}^{K} Z_i \gamma_i + \delta\right)\right) \\
&= tr\left(\mathbb{E}\left(\left(\sum_{i=1}^{K} \beta_i^T X_i^T + \epsilon^T\right) A \left(\sum_{i=1}^{K} Z_i \gamma_i + \delta\right)\right)\right) \\
&= tr\left(\mathbb{E}\left(\sum_{i=1}^{K} \beta_i^T X_i^T A Z_i \gamma_i\right) + \mathbb{E}(\epsilon^T A \delta)\right) \\
&= \mathbb{E}\left(tr\left(\sum_{i=1}^{K} \beta_i^T X_i^T A Z_i \gamma_i\right) + tr(\epsilon^T A \delta)\right) \\
&= \mathbb{E}\left(\sum_{i=1}^{K} tr(A Z_i \gamma_i \beta_i^T X_i^T)\right) + \mathbb{E}(tr(A \delta \epsilon^T)) \\
&= \sum_{i=1}^{K} tr(A Z_i \mathbb{E}(\gamma_i \beta_i^T) X_i^T) + tr(A \mathbb{E}(\delta \epsilon^T)) \\
&= \sum_{i=1}^{K} \frac{\rho_i}{m_i} tr(A Z_i X_i^T) + \rho_e \left(\sum_{t=1}^{N_s} A_{tt}\right)
\end{aligned}$$

Here, quantity A_{tt} denotes the t^{th} diagonal element of matrix A. To estimate the covariance parameters, we plug in K+1 different matrices $A_1,\ldots,A_{K+1}$ into the equation above. Next, we apply method of moments to approximate $\mathbb{E}(y_1^T \tilde{A}_j y_2)$ using the observed value $y_1^T \tilde{A}_j y_2$. After these steps, we get the following equations:

$$y_1^T A_j y_2 = \sum_{i=1}^{K} \frac{\rho_i}{m_i} tr(A_j Z_i X_i^T) + \rho_e \sum_{t=1}^{N_s} (A_j)_{tt}, j = 1, \ldots, K+1.$$

Solving this linear system of K+1 equations would get us a set of point estimates $\hat{\rho}_1, \ldots, \hat{\rho}_K$ for covariance parameters. We discuss the details in the following section.

Choose Matrix *A*

The estimation approach described the previous section works for an arbitrary set of A matrices. So how do we properly choose them in practice? We begin with solving a practical issue. Processing large-scale GWASs requires a substantial amount of resource for both computation and data storage. Moreover, individual-level genotype and phenotype data from consortium-based GWASs are often non-accessible due to policy concerns. However, sharing the summary statistics has become a common practice in the field of complex disease genetics. Summary data for many GWASs are openly accessible online. Therefore, it is of practical interest to estimate genetic covariance based on summary statistics only. To achieve this goal, we define the first K matrices as:

$$\tilde{A}_j = \frac{X_j Z_j^T}{m_j}, j = 1, \ldots, K.$$

Plugging in these matrices, the first K equations become:

$$\begin{aligned}
\frac{1}{m_j}\left(X_j^T y_1\right)^T Z_j^T y_2 &= \sum_{i=1}^{K} \frac{\rho_i}{m_i m_j} tr\left(X_j Z_j^T Z_i X_i^T\right) + \frac{\rho_e}{m_j} \sum_{t=1}^{N_s} \left(X_j Z_j^T\right)_{tt} \\
&= \sum_{i=1}^{K} \frac{\rho_i}{m_i m_j} tr\left(Z_j^T Z_i X_i^T X_j\right) \\
&\quad + \frac{\rho_e}{m_j} \sum_{t=1}^{N_s} \left(X_j X_j^T\right)_{tt}, j = 1, \ldots, K.
\end{aligned}$$

The second equality is based on the property of trace and the fact that first N_S samples are shared between two studies. To calculate all the terms in these equations, we note that

$$\begin{aligned}
\sum_{t=1}^{N_s} \left(X_j X_j^T\right)_{tt} &= \sum_{t=1}^{N_s} \sum_{l=1}^{m_j} (X_j)_{tl} = \sum_{l=1}^{m_j} \sum_{t=1}^{N_s} (X_j)_{tl} \approx \sum_{l=1}^{m_j} N_s \\
&= m_j N_s.
\end{aligned}$$

Here, we used approximation $(\sum_{t=1}^{N_s} (X_j)_{tl})/N_s \approx 1$. This is because genotype data are standardized and the shared sub-cohort is a subset of all individuals. In practice, if two GWASs share samples, the shared sample size is usually greater than several hundred, which is sufficient to make this approximation reasonable.

$$\begin{aligned}
\frac{1}{m_i m_j} tr\left(Z_j^T Z_i X_i^T X_j\right) &= \frac{N_1 N_2}{m_i m_j} tr\left(\left(\frac{Z_j^T Z_i}{N_2}\right)\left(\frac{X_i^T X_j}{N_1}\right)\right) \\
&\approx \frac{N_1 N_2}{m_i m_j} tr\left(D_{ij}^T D_{ij}\right) = \frac{N_1 N_2}{m_i m_j} \sum_{l=1}^{m_i} \sum_{l'=1}^{m_j} r^2_{l^{(i)} l'^{(j)}}.
\end{aligned}$$

We approximate the sample linkage disequilibrium (LD) matrices from both studies, i.e., $(Z_i^T Z_j)/N_2$ and $(X_i^T X_j)/N_1$, using the population LD matrix D_{ij}. In practice, we estimate LD using a reference panel, e.g., samples from the 1000 Genomes Project with European ancestry. In the formula, $r^2_{l^{(i)} l'^{(j)}}$ denotes the LD between the l^{th} SNP from category S_i and the $(l')^{\text{th}}$ SNP from category S_j.

$$\begin{aligned}
\left(X_j^T y_1\right)^T Z_j^T y_2 &= \sqrt{N_1 N_2}\left(\frac{1}{\sqrt{N_1}} X_j^T y_1\right)^T \left(\frac{1}{\sqrt{N_2}} Z_j^T y_2\right) \\
&= \sqrt{N_1 N_2} (z_1)_j^T (z_2)_j
\end{aligned}$$

Here, z_1 and z_2 denote the z-scores of SNP-level associations from two GWASs; $(z_1)_j$ and $(z_2)_j$ represent z-scores corresponding to the SNPs in annotation category S_j.

We plug in these quantities and divide N_1N_2 on both sides of the K equations, then we get:

$$\frac{1}{m_j\sqrt{N_1N_2}}(z_1)_j^T(z_2)_j = \sum_{i=1}^{K}\frac{\rho_i}{m_im_j}\sum_{l=1}^{m_i}\sum_{l'=1}^{m_j} r^2_{l^{(i)}l'^{(j)}} + \frac{N_s\rho_e}{N_1N_2}, j = 1,\ldots,K.$$

Next, we study the (K+1)$^{\text{th}}$ equation. We define:

$$\tilde{A}_{K+1} = \begin{pmatrix} I_{N_s\times N_s} & 0 \\ 0 & 0 \end{pmatrix}_{N_1\times N_2}.$$

We make the following observations.

$$y_1^T\tilde{A}_{K+1}y_2 = \sum_{t=1}^{N_s}(y_1)_t(y_2)_t$$

$$\frac{1}{m_i}tr\left(\tilde{A}_{K+1}Z_iX_i^T\right) = \frac{1}{m_i}\sum_{t=1}^{N_s}\left(Z_iX_i^T\right)_{tt} = \frac{1}{m_i}\sum_{t=1}^{N_s}\left(X_iX_i^T\right)_{tt} \approx \frac{1}{m_i}m_iN_s = N_s$$

Again, the approximation is based on the facts that the genotype data are standardized, the first N_S rows of matrices X_i and Z_i are identical, and the shared subcohort is a subset of the complete study with sufficient sample size.

$$\sum_{t=1}^{N_s}\left(\tilde{A}_{K+1}\right)_{tt} = \sum_{t=1}^{N_s}1 = N_s$$

Plugging in these quantities and dividing N_1N_2 on both sides of the (K+1)$^{\text{th}}$ equation, we get:

$$\frac{1}{N_1N_2}\sum_{t=1}^{N_s}(y_1)_t(y_2)_t = \frac{N_s}{N_1N_2}\sum_{i=1}^{K}\rho_i + \frac{N_s}{N_1N_2}\rho_e.$$

We denote all K+1 equations in matrix form:

$$\begin{pmatrix} \frac{1}{m_1\sqrt{N_1N_2}}(z_1)_1^T(z_2)_1 \\ \vdots \\ \frac{1}{m_K\sqrt{N_1N_2}}(z_1)_K^T(z_2)_K \\ \frac{1}{N_1N_2}\sum_{t=1}^{N_s}(y_1)_t(y_2)_t \end{pmatrix} = \begin{pmatrix} \frac{1}{m_1m_1}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_1}r^2_{l^{(1)}l'^{(1)}} & \cdots & \frac{1}{m_Km_1}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_1}r^2_{l^{(K)}l'^{(1)}} & \frac{N_s}{N_1N_2} \\ \vdots & \ddots & \vdots & \vdots \\ \frac{1}{m_1m_K}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_K}r^2_{l^{(1)}l'^{(K)}} & \cdots & \frac{1}{m_Km_K}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_K}r^2_{l^{(K)}l'^{(K)}} & \frac{N_s}{N_1N_2} \\ \frac{N_s}{N_1N_2} & \cdots & \frac{N_s}{N_1N_2} & \frac{N_s}{N_1N_2} \end{pmatrix} \times \begin{pmatrix} \rho_1 \\ \vdots \\ \rho_K \\ \rho_{K+1} \end{pmatrix}.$$

Since $\rho_1, \ldots, \rho_K$ are the parameters of interest, we subtract the (K+1)$^{\text{th}}$ equation from the first K equations and remove ρ_{K+1} from the linear system:

$$\begin{pmatrix} \frac{1}{m_1\sqrt{N_1N_2}}(z_1)_1^T(z_2)_1 - \frac{1}{N_1N_2}\sum_{t=1}^{N_s}(y_1)_t(y_2)_t \\ \vdots \\ \frac{1}{m_K\sqrt{N_1N_2}}(z_1)_K^T(z_2)_K - \frac{1}{N_1N_2}\sum_{t=1}^{N_s}(y_1)_t(y_2)_t \end{pmatrix} = \begin{pmatrix} \frac{1}{m_1m_1}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_1}r^2_{l^{(1)}l'^{(1)}} - \frac{N_s}{N_1N_2} & \cdots & \frac{1}{m_Km_1}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_1}r^2_{l^{(K)}l'^{(1)}} - \frac{N_s}{N_1N_2} \\ \vdots & \ddots & \vdots \\ \frac{1}{m_1m_K}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_K}r^2_{l^{(1)}l'^{(K)}} - \frac{N_s}{N_1N_2} & \cdots & \frac{1}{m_Km_K}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_K}r^2_{l^{(K)}l'^{(K)}} - \frac{N_s}{N_1N_2} \end{pmatrix} \begin{pmatrix} \rho_1 \\ \vdots \\ \rho_K \end{pmatrix}$$

When the sample sizes of both GWASs are large and the sample overlap between two studies is moderate, then $N_S / (N_1N_2)$ is a small quantity. We use the approximation $N_s/(N_1N_2) \approx 0$. Similarly, we have:

$$\frac{1}{N_1N_2}\sum_{t=1}^{N_s}(y_1)_t(y_2)_t = \frac{N_s}{N_1N_2}\left(\frac{1}{N_s}\sum_{t=1}^{N_s}(y_1)_t(y_2)_t\right) \approx 0.$$

Here, $(\sum_{t=1}^{N_s}(y_1)_t(y_2)_t)/N_s$ is the phenotypic correlation between two traits among the shared N_S samples and is bounded by 1. Therefore the approximation is reasonable. Additional justification on these approximations will be given in the next section.

In summary, the K equations can be approximated by:

$$\begin{pmatrix} \frac{1}{m_1\sqrt{N_1N_2}}(z_1)_1^T(z_2)_1 \\ \vdots \\ \frac{1}{m_K\sqrt{N_1N_2}}(z_1)_K^T(z_2)_K \end{pmatrix} =$$

$$\begin{pmatrix} \frac{1}{m_1m_1}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_1} r^2_{l^{(1)}l'^{(1)}} & \cdots & \frac{1}{m_Km_1}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_1} r^2_{l^{(K)}l'^{(1)}} \\ \vdots & \ddots & \vdots \\ \frac{1}{m_1m_K}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_K} r^2_{l^{(1)}l'^{(K)}} & \cdots & \frac{1}{m_Km_K}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_K} r^2_{l^{(K)}l'^{(K)}} \end{pmatrix} \begin{pmatrix} \rho_1 \\ \vdots \\ \rho_K \end{pmatrix}.$$

We denote

$$v = \left(\frac{1}{m_1\sqrt{N_1N_2}}(z_1)_1^T(z_2)_1, \ldots, \frac{1}{m_K\sqrt{N_1N_2}}(z_1)_K^T(z_2)_K\right)^T$$

$$M = \begin{pmatrix} \frac{1}{m_1m_1}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_1} r^2_{l^{(1)}l'^{(1)}} & \cdots & \frac{1}{m_Km_1}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_1} r^2_{l^{(K)}l'^{(1)}} \\ \vdots & \ddots & \vdots \\ \frac{1}{m_1m_K}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_K} r^2_{l^{(1)}l'^{(K)}} & \cdots & \frac{1}{m_Km_K}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_K} r^2_{l^{(K)}l'^{(K)}} \end{pmatrix}.$$

Then, the point estimate of covariance parameters can be denoted as

$$\widehat{\rho} = M^{-1}v.$$

Importantly, we emphasize that M can be estimated using a reference panel and v is based only on GWAS summary data. No individual-level genotype or phenotype information from the original GWASs is needed in this framework. Finally, we note that in some rare cases (e.g., very similar annotations are used simultaneously in the analysis), matrix M may not be invertible. In that case, we can acquire the genetic covariance estimator through the following minimization problem.

$$\widehat{\rho} = \min_{\rho} \|M\rho - v\|_2^2$$

Remarks on Approximation

Several approximations are critical in the estimation framework described above. In this section, we discuss why these approximations are reasonable.

Approximation 1.

$$\frac{1}{N_s}\sum_{t=1}^{N_s}(X_j)_{tl} \approx 1$$

This approximation is based on law of large numbers and two assumptions. (1) The genotype matrix is standardized. (2) If two GWASs share samples, the shared sample size N_S needs to be sufficiently large. The first assumption is commonly seen in complex trait genetic models. It is actually not a required condition, but it simplifies the algebra. The second assumption is also most likely going to hold in practice. If two GWASs have a sample overlap, it is often because one or more cohorts were used in both studies. A cohort like this usually has a sample size that ranges from several hundred to a few thousand, which is sufficiently large for the law of large numbers to hold.

Approximation 2.

$$\frac{1}{m_im_j}\sum_{l=1}^{m_i}\sum_{l'=1}^{m_j} r^2_{l^{(i)}l'^{(j)}} - \frac{N_s}{N_1N_2} \approx \frac{1}{m_im_j}\sum_{l=1}^{m_i}\sum_{l'=1}^{m_j} r^2_{l^{(i)}l'^{(j)}}$$

Since the first term $(\sum_{l=1}^{m_i}\sum_{l'=1}^{m_j} r^2_{l^{(i)}l'^{(j)}})/m_im_j$ does not depend on GWAS sample size, this approximation holds when N_1 and N_2 are large and the shared sample size N_S is moderate. Notably, this condition does not contradict with the condition in approximation 1. In approximation 1, we require the value of N_S to exceed several hundred so that the law of large numbers could hold. Here, we require the ratio between N_S and the actual GWAS sample size to be small. Since large-scale GWAS meta-analyses published in recent years often have sample sizes on the scale of 10^4 or 10^5, this approximation is reasonable. Of note, the term N_S / N_1N_2 is introduced when we remove parameter ρ_{K+1} from the linear system by subtracting the $(K+1)^{th}$ equation from the first K equations. If the true value of ρ_{K+1}, i.e., the non-genetic covariance introduced by sample overlap, is in fact very small compared with the genetic covariance, then this approximation can be omitted. Finally, we note that even if the two GWASs are performed on the identical cohort (i.e., complete sample overlap), then N_S / $N_1N_2 = 1/N$ is still a small quantity as long as the sample size N is big.

Approximation 3.

$$\frac{1}{\sqrt{N_1N_2}}\left(\frac{1}{m_j}(z_1)_j^T(z_2)_j\right) - \frac{N_s}{N_1N_2}\left(\frac{1}{N_s}\sum_{t=1}^{N_s}(y_1)_t(y_2)_t\right) \approx \frac{1}{\sqrt{N_1N_2}}\left(\frac{1}{m_j}(z_1)_1^T(z_2)_1\right)$$

The z-scores in GWASs usually do not deviate much from the standard normal distribution. Therefore $(z_1)_j^T(z_2)_j/m_j$ is close to the true correlation between z_1 and z_2. Similarly, since we assume the phenotypes are standardized,

$(\sum_{t=1}^{N_s}(y_1)_t(y_2)_t)/N_s$ is the phenotypic correlation between two traits among the shared N_S samples. Therefore, as long as

$$\frac{N_s}{N_1N_2} \ll \frac{1}{\sqrt{N_1N_2}}$$

or equivalently,

$$\frac{N_s}{\sqrt{N_1N_2}} \ll 1,$$

then it is reasonable to omit the term $(\sum_{t=1}^{N_s}(y_1)_t(y_2)_t)/N_1N_2$. Therefore, similar to the condition in approximation 2, if the GWAS sample sizes N_1 and N_2 are large and the shared sample size N_S is moderate, then approximation 3 holds. However, we note that if there is a substantial overlap between two studies (e.g., when analyzing two traits measured on the same cohort), then $N_s/\sqrt{N_1N_2} \approx 1$ and we can no longer omit the term $(\sum_{t=1}^{N_s}(y_1)_t(y_2)_t)/N_1N_2$ from the equation.

Special Cases

(1) Two Independent GWASs

If samples from two GWASs do not overlap, then the non-genetic effects ϵ and δ are independent and $\rho_e = 0$. So only K equations are needed for estimating covariance estimators. We still define $\tilde{A}_j = (X_jZ_j^T)/m_j$ for j = 1,...,K. That gives us:

$$\begin{pmatrix} \frac{1}{m_1\sqrt{N_1N_2}}(z_1)_1^T(z_2)_1 \\ \vdots \\ \frac{1}{m_K\sqrt{N_1N_2}}(z_1)_K^T(z_2)_K \end{pmatrix} = \begin{pmatrix} \frac{1}{m_1m_1}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_1} r^2_{l^{(1)}l'^{(1)}} & \cdots & \frac{1}{m_Km_1}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_1} r^2_{l^{(K)}l'^{(1)}} \\ \vdots & \ddots & \vdots \\ \frac{1}{m_1m_K}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_K} r^2_{l^{(1)}l'^{(K)}} & \cdots & \frac{1}{m_Km_K}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_K} r^2_{l^{(K)}l'^{(K)}} \end{pmatrix} \begin{pmatrix} \rho_1 \\ \vdots \\ \rho_K \end{pmatrix}.$$

Therefore, none of the approximations discussed in the previous section is needed in this simple scenario. The covariance estimator remains the same:

$$\hat{\rho} = M^{-1}\nu.$$

(2) No Annotation Stratification

If we do not stratify covariance by functional annotation, then $\hat{\rho}$ is just a one-dimensional estimator for the overall genetic covariance.

$$\hat{\rho} = \nu/M = \left(\frac{1}{m\sqrt{N_1N_2}}(z_1)^T(z_2)\right) \Big/ \left(\frac{1}{m^2}\sum_{l=1}^{m}\sum_{l'=1}^{m} r^2_{ll'}\right) = \frac{\overline{z_1z_2}}{\overline{r^2}\sqrt{N_1N_2}}$$

Here, $\overline{z_1z_2}$ is the average product of z-scores from two GWASs; $\overline{r^2}$ is the average LD across all SNP pairs in the study, or equivalently, the average LD score across all SNPs in the study. Interestingly, this estimator can be seen as a two-trait extension of the heritability estimator proposed by Bulik-Sullivan.[16]

(3) Two Different Traits Measured on the Same Cohort

If the samples completely overlap between two GWASs (i.e., $N_1 = N_2 = N_S = N$), as we discussed in the previous section, approximations 1 and 2 still hold as long as the sample size is large but approximation 3 would fail. Therefore, after subtracting the $(K+1)^{th}$ equation from the first K equations and removing parameter ρ_{K+1}, we get:

$$\begin{pmatrix} \frac{1}{m_1N}(z_1)_1^T(z_2)_1 - \frac{1}{N^2}\sum_{t=1}^{N}(y_1)_t(y_2)_t \\ \vdots \\ \frac{1}{m_KN}(z_1)_K^T(z_2)_K - \frac{1}{N^2}\sum_{t=1}^{N}(y_1)_t(y_2)_t \end{pmatrix} = \begin{pmatrix} \frac{1}{m_1m_1}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_1} r^2_{l^{(1)}l'^{(1)}} & \cdots & \frac{1}{m_Km_1}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_1} r^2_{l^{(K)}l'^{(1)}} \\ \vdots & \ddots & \vdots \\ \frac{1}{m_1m_K}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_K} r^2_{l^{(1)}l'^{(K)}} & \cdots & \frac{1}{m_Km_K}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_K} r^2_{l^{(K)}l'^{(K)}} \end{pmatrix} \begin{pmatrix} \rho_1 \\ \vdots \\ \rho_K \end{pmatrix}.$$

In practice, since we do not assume access to the individual-level phenotype data, an estimate of phenotypic correlation $\hat{\rho}_{pheno}$ needs to be acquired elsewhere (since we assumed phenotypes to be standardized, this is equivalent to phenotypic covariance). Then, we could get the covariance estimate under sample overlap correction:

$$\hat{\rho} = M^{-1}\begin{pmatrix} \frac{1}{m_1N}(z_1)_1^T(z_2)_1 - \frac{1}{N}\hat{\rho}_{pheno} \\ \vdots \\ \frac{1}{m_KN}(z_1)_K^T(z_2)_K - \frac{1}{N}\hat{\rho}_{pheno} \end{pmatrix} = M^{-1}\left(\nu - \frac{\hat{\rho}_{pheno}}{N}\mathbf{1}\right).$$

For some traits, $\hat{\rho}_{pheno}$ may have been reported in the literature. Otherwise, we need to estimate $\hat{\rho}_{pheno}$ using GWAS summary statistics. Bulik-Sullivan et al. showed the following LD score regression equation without annotation structure in the genome:[7]

$$\mathbb{E}\Big((z_1)_j(z_2)_j\Big) = \frac{\sqrt{N_1N_2}\rho}{m}\sum_{i=1}^{m} r_{ij}^2 + \frac{N_s\rho_{pheno}}{\sqrt{N_1N_2}}.$$

In this special case when two traits are measured on the same cohort, the formula becomes

$$\mathbb{E}\Big((z_1)_j(z_2)_j\Big) = \frac{N\rho}{m}\sum_{i=1}^{m} r_{ij}^2 + \rho_{pheno}.$$

Therefore, we could apply LD score regression and use the estimated intercept as $\widehat{\rho}_{pheno}$.

(4) Binary Traits

In this section, we investigate whether we could analyze ascertained case-control studies using our framework. It has been previously shown that the following formula holds under the liability threshold model:[7]

$$\mathbb{E}\left(\frac{1}{\sqrt{N_1N_2}}(z_1)_{SNP_j}(z_2)_{SNP_j}\right) = \frac{\rho_{obs}}{m}\sum_{i=1}^{m} r_{ij}^2 + \sqrt{P_1(1-P_1)P_2(1-P_2)} \times \left(\frac{N^{cas,cas}}{N_1^{cas}N_2^{cas}} + \frac{N^{con,con}}{N_1^{con}N_2^{con}} - \frac{N^{cas,con}}{N_1^{cas}N_2^{con}} - \frac{N^{con,cas}}{N_1^{con}N_2^{cas}}\right),$$

where ρ_{obs} denotes the covariance on the observed scale, $N^{a,b}$ denotes the number of samples with phenotype a in study 1 and phenotype b in study 2, N_i^a denotes the total number of samples with phenotype a in study i, and P_i denotes the sample prevalence of trait y_i. Using the same approximation we used in method of moments,

$$\mathbb{E}\left(\frac{1}{\sqrt{N_1N_2}}(z_1)_{SNP_j}(z_2)_{SNP_j}\right) \approx \frac{1}{\sqrt{N_1N_2}}(z_1)_{SNP_j}(z_2)_{SNP_j},$$

it is straightforward to extend it to the following matrix form that allows annotation stratification:

$$\begin{pmatrix} \frac{1}{m_1\sqrt{N_1N_2}}(z_1)_1^T(z_2)_1 - \eta \\ \vdots \\ \frac{1}{m_K\sqrt{N_1N_2}}(z_1)_K^T(z_2)_K - \eta \end{pmatrix} = \begin{pmatrix} \frac{1}{m_1m_1}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_1} r^2_{l^{(1)}l'^{(1)}} & \cdots & \frac{1}{m_Km_1}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_1} r^2_{l^{(K)}l'^{(1)}} \\ \vdots & \ddots & \vdots \\ \frac{1}{m_1m_K}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_K} r^2_{l^{(1)}l'^{(K)}} & \cdots & \frac{1}{m_Km_K}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_K} r^2_{l^{(K)}l'^{(K)}} \end{pmatrix} \begin{pmatrix} \rho_{obs,1} \\ \vdots \\ \rho_{obs,K} \end{pmatrix},$$

where

$$\eta = \sqrt{P_1(1-P_1)P_2(1-P_2)}\left(\frac{N^{cas,cas}}{N_1^{cas}N_2^{cas}} + \frac{N^{con,con}}{N_1^{con}N_2^{con}} - \frac{N^{cas,con}}{N_1^{cas}N_2^{con}} - \frac{N^{con,cas}}{N_1^{con}N_2^{cas}}\right).$$

We note that $\eta = 0$ when two GWASs do not share any sample. In that case, covariance estimator remains the same:

$$\widehat{\rho} = M^{-1}v.$$

We just need to interpret it as the covariance on the observed scale.

When two studies have a substantial sample overlap, η cannot be ignored in the equations and therefore needs to be estimated. Notably, $\sqrt{N_1N_2}\eta$ is in fact the intercept term in cross-trait LD score regression. Therefore, similar to the previous scenario in this section, we could estimate $\widehat{\eta}$ by running LD score regression first and then plug it in the equations to calculate the covariance estimate:

$$\widehat{\rho} = M^{-1}\begin{pmatrix} \frac{1}{m_1N}(z_1)_1^T(z_2)_1 - \widehat{\eta} \\ \vdots \\ \frac{1}{m_KN}(z_1)_K^T(z_2)_K - \widehat{\eta} \end{pmatrix} = M^{-1}(v - \widehat{\eta}\,\mathbf{1}).$$

Finally, we note that the argument can be extended to estimate the covariance between a continuous trait and a binary trait. However, the estimated genetic covariance will be on the half-observed scale.[7]

Remarks on Overlapping Functional Annotations

We have discussed parameter estimation in previous sections. Our framework allows functional annotations to overlap, which is an important feature in real data analysis. However, when functional annotations overlap, the covariance parameter ρ is not the real quantity of interest. Instead, the total covariance in each annotation category is more meaningful biologically. For instance, the total covariance in functional annotation S_1 is

$$covariance(S_1) = \sum_{i=1}^{K}\rho_i\frac{m_{i\cap 1}}{m_i},$$

where $m_{i\cap 1}$ denotes the number of SNPs in region $S_1 \cap S_i$. Of note, this quantity equals to ρ_1 when S_1 does not overlap with any other functional annotation. Therefore, we use the weighted estimator $\widehat{\rho}^W$ to estimate the total covariance in each category when functional annotations overlap:

$$\widehat{\rho}^W = W\widehat{\rho}.$$

Here, W is a $K \times K$ matrix with element

$$W_{ij} = \frac{m_{j\cap i}}{m_j}, 1 \le i,j \le K.$$

Theoretical Properties of Covariance Estimator and Some Numerical Justifications

As discussed in previous sections, matrices $\tilde{A}_j$ have two major properties under the ideal case where two GWASs do not share samples.

(1) Vector $v = (y_1^T\tilde{A}_1y_2, \ldots, y_1^T\tilde{A}_Ky_2)^T/N_1N_2$ can be directly calculated using GWAS summary statistics.
(2) $\mathbb{E}(y_1^T\tilde{A}_jy_2) = \sum_{i=1}^{K} tr(\tilde{A}_jZ_iX_i^T)\rho_i/m_i$, where terms $tr(\tilde{A}_jZ_iX_i^T)$ only depends on LD and therefore can be estimated using a reference panel.

In this section, we investigate whether changing A could get us another covariance estimator, $\widehat{\rho}_A$, that is even better than $\widehat{\rho}$, which is based on $\tilde{A}_j$. We show that under reasonable conditions, all estimators in our framework are unbiased but $\widehat{\rho} = M^{-1}\nu$ "almost" has the minimal variance. The proof is an extension of the MINQUE theory developed in Rao et al.[63]

To prove the theoretical properties, we need an additional assumption on the distribution of y_1 and y_2. We assume that y_1 and y_2 are marginally standardized and follow a multivariate normal distribution:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim MVN\left(0, \begin{pmatrix} H_1 & \Theta \\ \Theta^T & H_2 \end{pmatrix}\right).$$

H_1 and H_2 denote the variance-covariance matrices of two traits; Θ denotes the covariance elements between two traits. Based on the model we have described throughout the paper:

$$\Theta = \mathbb{E}(y_1 y_2^T) = \sum_{i=1}^{K} \frac{\rho_i}{m_i} X_i Z_i^T.$$

We begin with calculating the variance of the quadratic form-like quantity $y_1^T A y_2$.

Proposition 1. Let A be a $N_1 \times N_2$ matrix. Then $Var(y_1^T A y_2) = tr(A^T H_1 A H_2) + tr(A^T \Theta A^T \Theta)$.

Proof:

We note that

$$\mathbb{E}(y_1 \mid y_2) = \Theta H_2^{-1} y_2$$

$$Var(y_1 \mid y_2) = H_1 - \Theta H_2^{-1} \Theta^T.$$

Therefore,

$$\begin{aligned} Var(y_1^T A y_2) &= \mathbb{E}_2(Var_{1\mid 2}(y_1^T A y_2)) + Var_2(\mathbb{E}_{1\mid 2}(y_1^T A y_2)) \\ &= \mathbb{E}_2(Var_{1\mid 2}(y_2^T A^T y_1)) + Var_2(\mathbb{E}_{1\mid 2}(y_2^T A^T y_1)) \\ &= \mathbb{E}_2(y_2^T A^T (H_1 - \Theta H_2^{-1}\Theta^T) A y_2) \\ &\quad + Var_2(y_2^T A^T \Theta H_2^{-1} y_2) \\ &= tr(A^T (H_1 - \Theta H_2^{-1}\Theta^T) A H_2) \\ &\quad + Var_2(y_2^T A^T \Theta H_2^{-1} y_2) \\ &= tr(A^T H_1 A H_2) - tr(A^T \Theta H_2^{-1}\Theta^T A H_2) \\ &\quad + Var_2(y_2^T A^T \Theta H_2^{-1} y_2). \end{aligned}$$

Since

$$y_2^T A^T \Theta H_2^{-1} y_2 = y_2^T H_2^{-1} \Theta^T A y_2,$$

we have

$$\begin{aligned} y_2^T A^T \Theta H_2^{-1} y_2 &= \frac{1}{2}(y_2^T A^T \Theta H_2^{-1} y_2 + y_2^T H_2^{-1} \Theta^T A y_2) \\ &= \frac{1}{2} y_2^T (A^T \Theta H_2^{-1} + H_2^{-1} \Theta^T A) y_2. \end{aligned}$$

Matrix $A^T \Theta H_2^{-1} + H_2^{-1}\Theta^T A$ is symmetric; therefore $y_2^T (A^T \Theta H_2^{-1} + H_2^{-1}\Theta^T A) y_2$ is a quadratic form. This gives us

$$\begin{aligned} Var(y_2^T A^T \Theta H_2^{-1} y_2) &= \frac{1}{4} \times 2tr((A^T \Theta H_2^{-1} \\ &\quad + H_2^{-1}\Theta^T A) H_2 (A^T \Theta H_2^{-1} \\ &\quad + H_2^{-1}\Theta^T A) H_2) \\ &= \frac{1}{2} tr(A^T \Theta A^T \Theta + A^T \Theta H_2^{-1}\Theta^T A H_2 \\ &\quad + H_2^{-1}\Theta^T A H_2 A^T \Theta + H_2^{-1}\Theta^T A \Theta^T A H_2) \\ &= \frac{1}{2}(2tr(A^T \Theta A^T \Theta) \\ &\quad + 2tr(A^T \Theta H_2^{-1}\Theta^T A H_2)) \\ &= tr(A^T \Theta A^T \Theta) + tr(A^T \Theta H_2^{-1}\Theta^T A H_2). \end{aligned}$$

Therefore,

$$\begin{aligned} Var(y_1^T A y_2) &= tr(A^T H_1 A H_2) - tr(A^T \Theta H_2^{-1}\Theta^T A H_2) \\ &\quad + tr(A^T \Theta A^T \Theta) + tr(A^T \Theta H_2^{-1}\Theta^T A H_2) \\ &= tr(A^T H_1 A H_2) + tr(A^T \Theta A^T \Theta). \end{aligned}$$

Notably, if $y_1 = y_2$ and A is symmetric, then this result becomes the well-known variance formula for quadratic forms.

$$Var(y_1^T A y_1) = 2tr(A H_1 A H_1)$$

Proposition 1 tells us that the variance of $y_1^T A y_2$ contains two parts. Later we will show that the second part, i.e., $tr(A^T \Theta A^T \Theta)$, is very small compared to the first term, $tr(A^T H_1 A H_2)$, when analyzing real GWAS data. This is because the individuals in GWASs are almost independent samples and the elements of matrix Θ are small. On the contrary, we assume the data to be standardized, so the diagonal elements of matrices H_1 and H_2 are always 1. This leads to

$$tr(A^T H_1 A H_2) \gg tr(A^T \Theta A^T \Theta).$$

With this in mind, the following claim is approximately true.

$$Var(y_1^T A y_2) \approx tr(A^T H_1 A H_2)$$

In the next proposition, we define a $N_1 \times N_2$ matrix A_* and show that A_* minimizes $tr(A^T H_1 A H_2)$ under some conditions. Based on the argument above, A_* "almost" minimizes $Var(y_1^T A y_2)$ too.

Proposition 2. Assume two GWASs do not share samples. We define the following quantities.

(i) Let $p = (p_1, \ldots p_K)^T$ be an arbitrarily given K-dimensional vector;
(ii) Let S be a $K \times K$ symmetric matrix with element $S_{ll'} = tr(H_1^{-1} X_{l'} Z_{l'}^T H_2^{-1} Z_l X_l^T)/m_l m_{l'}$ for $1 \le l, l' \le K$;
(iii) Let $\lambda = (\lambda_1, \ldots \lambda_K)^T$ be a vector such that $S\lambda = p$;
(iv) Define $A_* = \sum_{j=1}^{K} (\lambda_j / m_j) H_1^{-1} X_j Z_j^T H_2^{-1}$.

Then, we have:

(1) $\mathbb{E}(y_1^T A_* y_2) = \sum_{t=1}^{K} p_t \rho_t$;
(2) Let A be a matrix such that $\mathbb{E}(y_1^T A y_2) = \sum_{t=1}^{K} p_t \rho_t$. Then, $tr(A^T H_1 A H_2) \ge tr(A_*^T H_1 A_* H_2)$.

Proof:

(1)

Note that

$$\mathbb{E}\left(y_1^T A_* y_2\right) = \sum_{t=1}^{K} \frac{tr\left(A_* Z_t X_t^T\right)\rho_t}{m_t}.$$

Therefore, it is equivalent to show

$$\frac{tr\left(A_* Z_t X_t^T\right)}{m_t} = p_t, 1 \le t \le K.$$

In fact,

$$\begin{aligned}\frac{1}{m_t} tr\left(A_* Z_t X_t^T\right) &= \sum_{j=1}^{K} \frac{\lambda_j}{m_j m_t} tr\left(H_1^{-1} X_j Z_j^T H_2^{-1} Z_t X_t^T\right) \\ &= \sum_{i=1}^{K} \lambda_j S_{tj} = p_t.\end{aligned}$$

(2)

Let $B = A - A_*$, then

$$\begin{aligned}tr\left(A^T H_1 A H_2\right) &= tr\left((A_* + B)^T H_1 (A_* + B) H_2\right) \\ &= tr\left(A_*^T H_1 A_* H_2\right) + tr\left(A_*^T H_1 B H_2\right) \\ &\quad + tr\left(B^T H_1 A_* H_2\right) + tr\left(B^T H_1 B H_2\right).\end{aligned}$$

First, we show that $tr(A_*^T H_1 B H_2) = tr(B^T H_1 A_* H_2) = 0$.

$$\begin{aligned}tr\left(A_*^T H_1 B H_2\right) &= tr\left(H_2^T B^T H_1^T A_*\right) = tr\left(H_2 B^T H_1 A_*\right) \\ &= tr\left(B^T H_1 A_* H_2\right) \\ &= tr\left(B^T H_1 \sum_{j=1}^{K} \lambda_j H_1^{-1} \tilde{A}_j H_2^{-1} H_2\right) \\ &= \sum_{j=1}^{K} \lambda_j tr\left(B^T H_1 H_1^{-1} \tilde{A}_j H_2^{-1} H_2\right) \\ &= \sum_{j=1}^{K} \lambda_j tr\left(B^T \tilde{A}_j\right)\end{aligned}$$

In the first part of this proof, we have shown that

$$\frac{tr\left(A_* Z_j X_j^T\right)}{m_j} = p_j, 1 \le j \le K.$$

Since

$$\mathbb{E}\left(y_1^T A y_2\right) = \sum_{t=1}^{K} p_t \rho_t$$

or equivalently,

$$p_j = \frac{tr\left(A Z_j X_j^T\right)}{m_j} = tr\left((A_* + B)\tilde{A}_j^T\right), 1 \le j \le K.$$

Therefore,

$$tr\left(B^T \tilde{A}_j\right) = tr\left(B \tilde{A}_j^T\right) = 0, 1 \le j \le K.$$

This gives us

$$tr\left(A_*^T H_1 B H_2\right) = tr\left(B^T H_1 A_* H_2\right) = 0.$$

Thus, all that remains is to show that $tr(B^T H_1 B H_2) \ge 0$. Since H_1 and H_2 are positive definite, $\exists D_1$, D_2 such that $H_1 = D_1^T D_1$ and $H_2 = D_2^T D_2$. Then,

$$\begin{aligned}tr\left(B^T H_1 B H_2\right) &= tr\left(B^T D_1^T D_1 B D_2^T D_2\right) = tr\left(D_1 B D_2^T D_2 B^T D_1^T\right) \\ &= tr\left(D_1 B D_2^T \left(D_1 B D_2^T\right)^T\right) \ge 0.\end{aligned}$$

Hence,

$$tr\left(A^T H_1 A H_2\right) \ge tr\left(A_*^T H_1 A_* H_2\right).$$

Proposition 2 tells us that given arbitrary $p = (p_1, \ldots p_K)^T$, if $\exists \lambda = (\lambda_1, \ldots \lambda_K)^T$ such that $S\lambda = p$, then $y_1^T A_* y_2$ is an unbiased estimator for $\sum_{t=1}^{K} p_t \rho_t$. Furthermore, among all unbiased estimators with the form $y_1^T A y_2$, $y_1^T A_* y_2$ has the minimum value of $tr(A_*^T H_1 A_* H_2)$, hence "almost" the minimum variance $Var(y_1^T A_* y_2)$.

Corollary 1. **(without annotation stratification)** We assume:

(i) Samples from two GWASs do not overlap;
(ii) The samples in each study are completely independent;
(iii) True LD in both studies (i.e., $Z^T Z$ and $X^T X$) is known.

Consider all matrices A that suffice

$$tr\left(A Z X^T\right) = \frac{tr(Z^T Z X^T X)}{m}.$$

We define

$$\widehat{\rho}_A = m\left(y_1^T A y_2\right) / tr\left(A Z X^T\right).$$

Then, $\widehat{\rho}_{\tilde{A}}$ with $\tilde{A} = (XZ^T)/m$ has the lowest variance.

Proof:

Let A be a matrix that suffices $tr(AZX^T) = tr(Z^T Z X^T X)/m$. The goal is to show that

$$Var(\widehat{\rho}_A) \ge Var(\widehat{\rho}_{\tilde{A}})$$

Since the samples in each GWAS are completely independent, we have:

$$H_1 = I_{N_1 \times N_1}$$

$$H_2 = I_{N_2 \times N_2}$$

Therefore,

$$S = \frac{1}{m^2} tr\left(H_1^{-1} X Z^T H_2^{-1} Z X^T\right) = \frac{1}{m^2} tr\left(Z^T Z X^T X\right).$$

Let $p = S$ and $\lambda = 1$. Then, by definition we have

$$A_* = \frac{H_1^{-1} X Z^T H_2^{-1}}{m} = \frac{X Z^T}{m} = \tilde{A}.$$

Since we have

$$\frac{1}{m}tr(AZX^T) = \frac{1}{m^2}tr(Z^TZX^TX) = p,$$

by proposition 2, we know that

$$Var(y_1^TAy_2) \geq Var(y_1^T\tilde{A}y_2).$$

Therefore,

$$\begin{aligned} Var(\widehat{\rho}_A) &= Var\left(\frac{y_1^TAy_2}{\frac{1}{m^2}tr(Z^TZX^TX)}\right) \\ &= \left(\frac{m^2}{tr(Z^TZX^TX)}\right)^2 Var(y_1^TAy_2) \\ &\geq \left(\frac{m^2}{tr(Z^TZX^TX)}\right)^2 Var(y_1^T\tilde{A}y_2) = Var(\widehat{\rho}_{\tilde{A}}). \end{aligned}$$

Of note, $\widehat{\rho}_{\tilde{A}}$ is identical to the non-annotation-stratified covariance estimator we developed in previous sections (see section Special Cases). Although we initially defined $\tilde{A} = (XZ^T)/m$ for the purpose of simplifying calculation, corollary 1 tells us that $\tilde{A}$ actually enjoys some good theoretical properties. As we have emphasized before, matrix $\tilde{A}$ could greatly simplify the estimation procedure because (1) $y_1^T\tilde{A}y_2$ can be calculated from GWAS summary statistics and (2) $tr(\tilde{A}ZX^T) = tr(Z^TZX^TX)/m$ depends only on LD. In corollary 1 we showed that if we want to keep the convenient property $tr(AZX^T) = tr(Z^TZX^TX)/m$, then it is impossible to improve the variance of estimator $\widehat{\rho}$ by choosing another matrix A. We note, however, additional variability may be introduced when we estimate LD using a reference panel in practice.

Similarly, we have a corollary for annotation-stratified covariance estimator.

***Corollary 2.* (with annotation stratification)** We assume:

(i) Samples from two GWASs do not overlap;
(ii) The samples in each study are completely independent;
(iii) The two LD matrices are known and identical (i.e., $X^TX/N_1 = Z^TZ/N_2$);
(iv) SNPs in different functional annotations are not in LD.

Consider all matrix sets A_j $(1 \leq j \leq K)$ that suffice

$$tr\left(A_jZ_jX_j^T\right) = \frac{tr\left(Z_j^TZ_jX_j^TX_j\right)}{m_j}, 1 \leq j \leq K.$$

We define

$$\widehat{\rho}_A = \begin{pmatrix} \frac{1}{m_1}tr(\tilde{A}_1Z_1X_1^T) & \cdots & \frac{1}{m_K}tr(\tilde{A}_1Z_KX_K^T) \\ \vdots & \ddots & \vdots \\ \frac{1}{m_1}tr(\tilde{A}_KZ_1X_1^T) & \cdots & \frac{1}{m_K}tr(\tilde{A}_KZ_KX_K^T) \end{pmatrix}^{-1} \begin{pmatrix} y_1^T\tilde{A}_1y_2 \\ \vdots \\ y_1^T\tilde{A}_Ky_2 \end{pmatrix}.$$

Then, $\widehat{\rho}_{\tilde{A}}$ with $\tilde{A}_j = (X_jZ_j^T)/m_j$ $(1 \leq j \leq K)$ has the lowest variance.

Proof:

Since the samples in each GWAS are completely independent, we have:

$$H_1 = I_{N_1 \times N_1}$$

$$H_2 = I_{N_2 \times N_2}$$

Therefore,

$$S_{ll'} = \frac{1}{m_lm_{l'}}tr(H_1^{-1}X_{l'}Z_{l'}^TH_2^{-1}Z_lX_l^T) = \frac{1}{m_lm_{l'}}tr(X_{l'}Z_{l'}^TZ_lX_l^T).$$

Given integer c such that $1 \leq c \leq K$, let $p^{(c)} = (p_1, \ldots p_K)^T$ where

$$p_i = \frac{1}{m_cm_i}tr(X_cZ_c^TZ_iX_i^T), 1 \leq i \leq K$$

and $\lambda^{(c)} = (\lambda_1, \ldots \lambda_K)^T$ where

$$\lambda_i = \begin{cases} 1, i = c \\ 0, i \neq c \end{cases}.$$

Then, by definition we have

$$A_* = \sum_{j=1}^{K}\frac{\lambda_j}{m_j}H_1^{-1}X_jZ_j^TH_2^{-1} = \frac{X_cZ_c^T}{m_c} = \tilde{A}_c.$$

By condition (iii), it is straightforward to check

$$p^{(c)} = S\lambda^{(c)}.$$

Therefore, by proposition 2 we know that

$$Var(y_1^TA_cy_2) \geq Var(y_1^TA_*y_2) = Var(y_1^T\tilde{A}_cy_2).$$

Since c is arbitrary, we have

$$Var(y_1^TA_cy_2) \geq Var(y_1^T\tilde{A}_cy_2), 1 \leq c \leq K.$$

Finally, since SNPs in different functional annotations are not in LD,

$$\frac{1}{m_j}tr\left(\tilde{A}_iZ_jX_j^T\right) = \frac{1}{m_im_j}tr\left(Z_i^TZ_jX_j^TX_i\right) = 0, \ \forall i \neq j.$$

Therefore, the variance of the c[th] estimated covariance component:

$$\begin{aligned} Var((\widehat{\rho}_A)_c) &= Var\left(\frac{y_1^TA_cy_2}{\frac{1}{m_c^2}tr(Z_c^TZ_cX_c^TX_c)}\right) \\ &= \left(\frac{m_c^2}{tr(Z_c^TZ_cX_c^TX_c)}\right)^2 Var(y_1^TA_cy_2) \\ &\geq \left(\frac{m_c^2}{tr(Z_c^TZ_cX_c^TX_c)}\right)^2 Var(y_1^T\tilde{A}_cy_2) \\ &= Var((\widehat{\rho}_{\tilde{A}})_c), 1 \leq c \leq K. \end{aligned}$$

It is straightforward to check that $\widehat{\rho}_{\tilde{A}}$ is identical to the annotation-stratified covariance estimator we developed in previous sections. In corollary 2, we showed that under some reasonable conditions, the annotation-stratified covariance estimator also has the minimal variance property. However, since we assumed linkage equilibrium for SNPs in different functional annotations, this result does not apply to overlapping annotations.

In this section, we have shown some theoretical properties of our covariance estimator. However, the claim

$$tr(A^T H_1 A H_2) \gg tr(A^T \Theta A^T \Theta)$$

is critical in our argument. Moreover, each proposition and corollary has its assumptions, which may or may not hold in practice. Therefore, we provide numerical justifications to our claims.

Numerical Study 1. Compare $tr(A^T H_1 A H_2)$ and $tr(A^T \Theta A^T \Theta)$

Simulation workflow:

Step 1. We simulate a 500×500 matrix A whose elements are independently sampled from standard normal distribution $N(0,1)$.

Step 2. We simulate the $1,000 \times 1,000$ matrix

$$\mathbb{E}\left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\begin{pmatrix} y_1^T & y_2^T \end{pmatrix}\right) = \begin{pmatrix} H_1 & \Theta \\ \Theta^T & H_2 \end{pmatrix}$$

by fixing the diagonal elements to be 1 and sampling the non-diagonal elements from uniform distribution $Unif(0, 0.05)$. Sample pairs with a genetic relatedness coefficient greater than 0.05 are often removed from GWAS analysis. Therefore, the matrix we simulate here closely mimics the phenotypic covariance matrices we see in real studies.

Step 3. We sample 10,000 independent vectors from the distribution

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim MVN\left(0, \begin{pmatrix} H_1 & \Theta \\ \Theta^T & H_2 \end{pmatrix}\right).$$

Step 4. We calculate and record $tr(A^T H_1 A H_2)$, $tr(A^T \Theta A^T \Theta)$, and the sample variance of $y_1^T A y_2$.

Step 5. Repeat steps 1–4 100 times.

From the simulations, we can see that $tr(A^T H_1 A H_2)$ closely approximates the sample variance of $y_1^T A y_2$, while $tr(A^T \Theta A^T \Theta)$ is a negligible term (Figure S9). The median log fold, i.e., $\log_{10}(|tr(A^T H_1 A H_2)/tr(A^T \Theta A^T \Theta)|)$, is 3.12. Therefore, $tr(A^T H_1 A H_2)$ is on average around 1,300 times greater than $tr(A^T \Theta A^T \Theta)$ in our simulation, which is consistent with our claim.

Numerical Study 2. Compare $Var(y_1^T A y_2)$ and $Var(y_1^T \tilde{A} y_2)$

Simulation workflow:

Step 1. Randomly divide 15,918 samples from the Wellcome Trust Case Control Consortium (WTCCC) dataset into two subgroups (each with 7,959 samples and m = 254,221 SNPs after quality control). We simulate 100 independent sets of continuous traits y_1 and y_2 using real genotype data from WTCCC and the following covariance structure on heritability and genetic covariance:

$$\begin{pmatrix} \beta \\ \gamma \end{pmatrix} \sim MVN\left(0, \frac{1}{m}\begin{pmatrix} \frac{1}{2}I & \frac{1}{10}I \\ \frac{1}{10}I & \frac{1}{2}I \end{pmatrix}\right).$$

Step 2. We simulate a 7,959 × 7,959 matrix A whose elements are independently sampled from standard normal distribution $N(0,1)$. We also simulate a matrix A' of the same size by permuting elements of $\tilde{A}$. Then we rescale matrices A and A' so that

$$\mathbb{E}(y_1^T A y_2) = \mathbb{E}(y_1^T A' y_2) = \mathbb{E}(y_1^T \tilde{A} y_2)$$

or equivalently,

$$tr(AZX^T) = tr(A'ZX^T) = tr(\tilde{A}ZX^T).$$

This makes all three matrices comparable.

Step 3. Calculate $y_1^T A y_2$ and $y_1^T A' y_2$ for all 100 independent sets. Record the sample variance for each quantity.

Step 4. Repeat steps 2–3 100 times and get a distribution for $\widehat{Var}(y_1^T A y_2)$ and $\widehat{Var}(y_1^T A' y_2)$. Compare them with the sample variance of $y_1^T \tilde{A} y_2$.

The results are consistent with our previous conclusions. Both $\widehat{Var}(y_1^T A y_2)$ and $\widehat{Var}(y_1^T A' y_2)$ are consistently and substantially greater than $\widehat{Var}(y_1^T \tilde{A} y_2)$. In fact, the variances are not on the same scale. Median $\widehat{Var}(y_1^T A y_2)$ is 8.4×10^7 times greater than $\widehat{Var}(y_1^T \tilde{A} y_2)$ and median $\widehat{Var}(y_1^T A' y_2)$ is also 2.5×10^7 times greater (Figure S10). These results suggest that matrix $\tilde{A}$ indeed enjoys the minimal variance property when applied to real genetic data.

Estimate Variance via Block-wise Jackknife

In the previous section, we showed that if two traits follow multivariate normal distributions, then $Var(y_1^T \tilde{A}_j y_2) = 2tr(\tilde{A}_j^T H_1 \tilde{A}_j H_2)$. In fact, we could get similar results for covariance, too.

$$Cov(y_1^T \tilde{A}_i y_2, y_1^T \tilde{A}_j y_2) = 2tr\left(\tilde{A}_i^T H_1 \tilde{A}_j H_2\right)$$

Therefore, the variance-covariance matrix of $\widehat{\rho}$ can be calculated accordingly:

$$\begin{aligned} Cov(\widehat{\rho}) &= Cov(M^{-1}v) = Cov\left(\frac{1}{N_1 N_2} M^{-1}\begin{pmatrix} y_1^T \tilde{A}_1 y_2 \\ \vdots \\ y_1^T \tilde{A}_K y_2 \end{pmatrix}\right) \\ &= \frac{2}{(N_1 N_2)^2} M^{-1} \\ &\quad \times \begin{pmatrix} tr\left(\tilde{A}_1^T H_1 \tilde{A}_1 H_2\right) & \cdots & tr\left(\tilde{A}_K^T H_1 \tilde{A}_1 H_2\right) \\ \vdots & \ddots & \vdots \\ tr\left(\tilde{A}_1^T H_1 \tilde{A}_K H_2\right) & \cdots & tr\left(\tilde{A}_K^T H_1 \tilde{A}_K H_2\right) \end{pmatrix} M^{-1}. \end{aligned}$$

However, it is difficult to calculate $tr(\tilde{A}_i^T H_1 \tilde{A}_j H_2)$. Estimating H_1 and H_2 would involve additional assumptions on the heritability structure. Even if we could accurately

estimate H_1 and H_2, $tr(\tilde{A}_i^T H_1 \tilde{A}_j H_2)$ cannot be calculated using standard GWAS summary statistics. Therefore, following Bulik-Sullivan et al.,[7] we apply a block-wise jackknife approach to estimate the variance.

First, we estimate the variance-covariance matrix of

$$v = \frac{1}{\sqrt{N_1 N_2}} \begin{pmatrix} \frac{1}{m_1}(z_1)_1^T (z_2)_1 \\ \vdots \\ \frac{1}{m_K}(z_1)_K^T (z_2)_K \end{pmatrix}.$$

We divide the genome into b (e.g., $b = 200$) blocks B_1, …, B_b. Let

$$v_i^{(t)} = \frac{(z_1)_i^T (z_2)_i - (z_1)_{S_i \cap B_t}^T (z_2)_{S_i \cap B_t}}{(m_i - m_{S_i \cap B_t})\sqrt{N_1 N_2}}, 1 \le i \le K \text{ and } 1 \le t \le b.$$

Here, subscript $S_i \cap B_t$ indicates the subset of SNPs in both functional annotation S_i and block B_t. Therefore, $v_i^{(t)}$ is the re-calculated v_i after removing all SNPs in block B_t from the analysis. Then, $Cov(v)$ is estimated as:

$$\left(\widehat{Cov}(v)\right)_{ij} = \frac{b-1}{b}\sum_{t=1}^{b}\left(v_i^{(t)} - \frac{1}{b}\sum_{s=1}^{b} v_i^{(s)}\right)\left(v_j^{(t)} - \frac{1}{b}\sum_{s=1}^{b} v_j^{(s)}\right).$$

Therefore, we get

$$\widehat{Cov}(\hat{\rho}) = M^{-1}\widehat{Cov}(v)M^{-1}.$$

If annotations overlap,

$$\widehat{Cov}(\hat{\rho}^W) = WM^{-1}\widehat{Cov}(v)M^{-1}W^T.$$

Finally, the test statistic for each covariance parameter is

$$z - score_i = \frac{\hat{\rho}_i}{\sqrt{\left(\widehat{Cov}(\hat{\rho})\right)_{ii}}}, 1 \le i \le K.$$

When annotations overlap,

$$z - score_i^W = \frac{\hat{\rho}_i^W}{\sqrt{\left(\widehat{Cov}(\hat{\rho}^W)\right)_{ii}}}, 1 \le i \le K.$$

Genetic Correlation

In non-stratified analysis, we could provide the genetic correlation estimate as follows:

$$cor = \frac{\hat{\rho}}{\sqrt{\hat{h}_1^2 \hat{h}_2^2}}.$$

We use the estimator proposed in Bulik-Sullivan[16] to estimate heritability for each trait.

$$\hat{h}_t^2 = \frac{\frac{1}{m}(z_t)^T(z_t) - 1}{\frac{N}{m^2}\sum_{l=1}^{m}\sum_{l'=1}^{m} r_{ll'}^2} = \frac{\overline{\chi_t^2} - 1}{N\overline{r^2}}, t = 1, 2$$

Compared to genetic covariance, genetic correlation is a more interpretable metric. It is also robust against certain systematic bias that exists in both genetic covariance and heritability (e.g., genomic control correction). However, statistical inference based on genetic covariance is equivalent to that based on genetic correlation. Estimating heritability requires additional model assumptions on the heritability structure and introduces additional variability into the estimation framework. Therefore, although we report the point estimate for genetic correlation, the statistical inference in our method is completely based on genetic covariance only.

In annotation-stratified analysis, the heritability in each annotation category may be small. This is especially true when applying annotations related to the repressed genome. Although methods for estimating annotation-stratified heritability have been proposed,[11,12] they may provide unstable, sometimes even negative, heritability estimates. Therefore, we focus on genetic covariance only when performing annotation-stratified analysis.

Supplemental Data

Supplemental Data include 10 figures, 10 tables, and Supplemental Acknowledgments and can be found with this article online at https://doi.org/10.1016/j.ajhg.2017.11.001.

Web Resources

GNOVA, https://github.com/xtonyjiang/GNOVA
LD Hub, http://ldsc.broadinstitute.org/ldhub/
LDSC, https://github.com/bulik/ldsc/
OMIM, http://www.omim.org/

Received: June 14, 2017
Accepted: October 25, 2017
Published: December 7, 2017

References

1. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
2. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
3. Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G., et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. Nat. Genet. *43*, 519–525.
4. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M., and Wray, N.R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics *28*, 2540–2542.

5. Vattikuti, S., Guo, J., and Chow, C.C. (2012). Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. PLoS Genet. *8*, e1002637.
6. Lee, S.H., Ripke, S., Neale, B.M., Faraone, S.V., Purcell, S.M., Perlis, R.H., Mowry, B.J., Thapar, A., Goddard, M.E., Witte, J.S., et al.; Cross-Disorder Group of the Psychiatric Genomics Consortium; and International Inflammatory Bowel Disease Genetics Consortium (IIBDGC) (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nat. Genet. *45*, 984–994.
7. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., Duncan, L., Perry, J.R., Patterson, N., Robinson, E.B., et al.; ReproGen Consortium; Psychiatric Genomics Consortium; and Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3 (2015). An atlas of genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.
8. Pasaniuc, B., and Price, A.L. (2017). Dissecting the genetics of complex traits using summary association statistics. Nat. Rev. Genet. *18*, 117–127.
9. Anttila, V., Bulik-Sullivan, B., Finucane, H.K., Bras, J., Duncan, L., Escott-Price, V., Falcone, G., Gormley, P., Malik, R., and Patsopoulos, N. (2016). Analysis of shared heritability in common disorders of the brain. bioRxiv. https://doi.org/10.1101/048991.
10. Zheng, J., Erzurumluoglu, A.M., Elsworth, B.L., Kemp, J.P., Howe, L., Haycock, P.C., Hemani, G., Tansey, K., Laurin, C., and Pourcain, B.S. (2016). LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics *33*, 272–279.
11. Zhou, X. (2016). A unified framework for variance component estimation with summary statistics in genome-wide association studies. bioRxiv. https://doi.org/10.1101/042846.
12. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al.; ReproGen Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; and RACI Consortium (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.
13. Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjálmsson, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium; SWE-SCZ Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; and SWE-SCZ Consortium (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet. *95*, 535–552.
14. Lu, Q., Powles, R.L., Abdallah, S., Ou, D., Wang, Q., Hu, Y., Lu, Y., Liu, W., Li, B., Mukherjee, S., et al. (2017). Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease. PLoS Genet. *13*, e1006933.
15. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A.; and 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
16. Bulik-Sullivan, B. (2015). Relationship between LD Score and Haseman-Elston Regression. bioRxiv. https://doi.org/10.1101/018283.
17. Cichonska, A., Rousu, J., Marttinen, P., Kangas, A.J., Soininen, P., Lehtimäki, T., Raitakari, O.T., Järvelin, M.R., Salomaa, V., Ala-Korpela, M., et al. (2016). metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. Bioinformatics *32*, 1981–1989.
18. Zheng, J., Richardson, T., Millard, L., Hemani, G., Raistrick, C., Vilhjalmsson, B., Haycock, P., and Gaunt, T. (2017). PhenoSpD: an integrated toolkit for phenotypic correlation estimation and multiple testing correction using GWAS summary statistics. bioRxiv. https://doi.org/10.1101/148627.
19. van Rheenen, W., Shatunov, A., Dekker, A.M., McLaughlin, R.L., Diekstra, F.P., Pulit, S.L., van der Spek, R.A., Võsa, U., de Jong, S., Robinson, M.R., et al.; PARALS Registry; SLALOM Group; SLAP Registry; FALS Sequencing Consortium; SLAGEN Consortium; and NNIPPS Study Group (2016). Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. Nat. Genet. *48*, 1043–1048.
20. Lu, Q., Hu, Y., Sun, J., Cheng, Y., Cheung, K.-H., and Zhao, H. (2015). A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Sci. Rep. *5*, 10576.
21. Lu, Q., Powles, R.L., Wang, Q., He, B.J., and Zhao, H. (2016). Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. PLoS Genet. *12*, e1005947.
22. Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., Snyder, M.; and ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.
23. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al.; Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.
24. Farooqi, I.S. (2014). Defining the neural basis of appetite and obesity: from genes to behaviour. Clin. Med. (Lond.) *14*, 286–289.
25. Manning, A.K., Hivert, M.-F., Scott, R.A., Grimsby, J.L., Bouatia-Naji, N., Chen, H., Rybin, D., Liu, C.-T., Bielak, L.F., Prokopenko, I., et al.; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium; and Multiple Tissue Human Expression Resource (MUTHER) Consortium (2012). A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat. Genet. *44*, 659–669.
26. Matesanz, F., Potenciano, V., Fedetz, M., Ramos-Mozo, P., Abad-Grau, Mdel.M., Karaky, M., Barrionuevo, C., Izquierdo, G., Ruiz-Peña, J.L., García-Sánchez, M.I., et al. (2015). A functional variant that affects exon-skipping and protein expression of SP140 as genetic mechanism predisposing to multiple sclerosis. Hum. Mol. Genet. *24*, 5619–5627.
27. Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C.C., Patsopoulos, N.A., Moutsianas, L., Dilthey, A., Su, Z., Freeman, C., Hunt, S.E., et al.; International Multiple Sclerosis Genetics Consortium; and Wellcome Trust Case Control Consortium 2 (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature *476*, 214–219.
28. Franke, A., McGovern, D.P., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J.,

Roberts, R., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat. Genet. *42*, 1118–1125.
29. Marenholz, I., Esparza-Gordillo, J., and Lee, Y.-A. (2013). Shared genetic determinants between eczema and other immune-related diseases. Curr. Opin. Allergy Clin. Immunol. *13*, 478–486.
30. Reppe, S., Wang, Y., Thompson, W.K., McEvoy, L.K., Schork, A.J., Zuber, V., LeBlanc, M., Bettella, F., Mills, I.G., Desikan, R.S., et al.; GEFOS Consortium (2015). Genetic sharing with cardiovascular disease risk factors and diabetes reveals novel bone mineral density loci. PLoS ONE *10*, e0144531.
31. Alnaes, R., and Torgersen, S. (1997). Personality and personality disorders predict development and relapses of major depression. Acta Psychiatr. Scand. *95*, 336–342.
32. Kendler, K.S., Neale, M.C., Kessler, R.C., Heath, A.C., and Eaves, L.J. (1993). A longitudinal twin study of personality and major depression in women. Arch. Gen. Psychiatry *50*, 853–862.
33. de Moor, M.H., van den Berg, S.M., Verweij, K.J., Krueger, R.F., Luciano, M., Arias Vasquez, A., Matteson, L.K., Derringer, J., Esko, T., Amin, N., et al.; Genetics of Personality Consortium (2015). Meta-analysis of genome-wide association studies for neuroticism, and the polygenic association with major depressive disorder. JAMA Psychiatry *72*, 642–650.
34. Gale, C.R., Hagenaars, S.P., Davies, G., Hill, W.D., Liewald, D.C., Cullen, B., Penninx, B.W., Boomsma, D.I., Pell, J., McIntosh, A.M., et al.; International Consortium for Blood Pressure GWAS, CHARGE Consortium Aging and Longevity Group (2016). Pleiotropy between neuroticism and physical and mental health: findings from 108 038 men and women in UK Biobank. Transl. Psychiatry *6*, e791.
35. Sun, H.-L., Pei, D., Lue, K.-H., and Chen, Y.-L. (2015). Uric acid levels can predict metabolic syndrome and hypertension in adolescents: a 10-year longitudinal study. PLoS ONE *10*, e0143786.
36. Chrabot, B.S., Kariuki, S.N., Zervou, M.I., Feng, X., Arrington, J., Jolly, M., Boumpas, D.T., Reder, A.T., Goulielmos, G.N., and Niewold, T.B. (2013). Genetic variation near IRF8 is associated with serologic and cytokine profiles in systemic lupus erythematosus and multiple sclerosis. Genes Immun. *14*, 471–478.
37. Jung, J.Y., Kohane, I.S., and Wall, D.P. (2011). Identification of autoimmune gene signatures in autism. Transl. Psychiatry *1*, e63.
38. Guloksuz, S.A., Abali, O., Aktas Cetin, E., Bilgic Gazioglu, S., Deniz, G., Yildirim, A., Kawikova, I., Guloksuz, S., and Leckman, J.F. (2017). Elevated plasma concentrations of S100 calcium-binding protein B and tumor necrosis factor alpha in children with autism spectrum disorders. Rev. Bras. Psiquiatr. *39*, 195–200.
39. van Oosten, B.W., Barkhof, F., Truyen, L., Boringa, J.B., Bertelsmann, F.W., von Blomberg, B.M., Woody, J.N., Hartung, H.-P., and Polman, C.H. (1996). Increased MRI activity and immune activation in two multiple sclerosis patients treated with the monoclonal anti-tumor necrosis factor antibody cA2. Neurology *47*, 1531–1534.
40. Aiba, Y., Yamazaki, K., Nishida, N., Kawashima, M., Hitomi, Y., Nakamura, H., Komori, A., Fuyuno, Y., Takahashi, A., Kawaguchi, T., et al. (2015). Disease susceptibility genes shared by primary biliary cirrhosis and Crohn's disease in the Japanese population. J. Hum. Genet. *60*, 525–531.
41. Lambert, J.-C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., et al.; European Alzheimer's Disease Initiative (EADI); Genetic and Environmental Risk in Alzheimer's Disease; Alzheimer's Disease Genetic Consortium; and Cohorts for Heart and Aging Research in Genomic Epidemiology (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat. Genet. *45*, 1452–1458.
42. Haines, J.L., Ter-Minassian, M., Bazyk, A., Gusella, J.F., Kim, D.J., Terwedow, H., Pericak-Vance, M.A., Rimmler, J.B., Haynes, C.S., Roses, A.D., et al.; The Multiple Sclerosis Genetics Group (1996). A complete genomic screen for multiple sclerosis underscores a role for the major histocompatability complex. Nat. Genet. *13*, 469–471.
43. Bertram, L., and Tanzi, R.E. (2005). The genetic epidemiology of neurodegenerative disease. J. Clin. Invest. *115*, 1449–1457.
44. Nuytemans, K., Maldonado, L., Ali, A., John-Williams, K., Beecham, G.W., Martin, E., Scott, W.K., and Vance, J.M. (2016). Overlap between Parkinson disease and Alzheimer disease in ABCA7 functional variants. Neurol. Genet. *2*, e44.
45. Gagliano, S.A., Pouget, J.G., Hardy, J., Knight, J., Barnes, M.R., Ryten, M., and Weale, M.E. (2016). Genomics implicates adaptive and innate immunity in Alzheimer's and Parkinson's diseases. Ann. Clin. Transl. Neurol. *3*, 924–933.
46. Reginato, A.M., Mount, D.B., Yang, I., and Choi, H.K. (2012). The genetics of hyperuricaemia and gout. Nat. Rev. Rheumatol. *8*, 610–621.
47. Culleton, B.F., Larson, M.G., Kannel, W.B., and Levy, D. (1999). Serum uric acid and risk for cardiovascular disease and death: the Framingham Heart Study. Ann. Intern. Med. *131*, 7–13.
48. Kanbay, M., Jensen, T., Solak, Y., Le, M., Roncal-Jimenez, C., Rivard, C., Lanaspa, M.A., Nakagawa, T., and Johnson, R.J. (2016). Uric acid in metabolic syndrome: From an innocent bystander to a central player. Eur. J. Intern. Med. *29*, 3–8.
49. Watanabe, S., Kang, D.-H., Feng, L., Nakagawa, T., Kanellis, J., Lan, H., Mazzali, M., and Johnson, R.J. (2002). Uric acid, hominoid evolution, and the pathogenesis of salt-sensitivity. Hypertension *40*, 355–360.
50. Rao, G.N., Corson, M.A., and Berk, B.C. (1991). Uric acid stimulates vascular smooth muscle cell proliferation by increasing platelet-derived growth factor A-chain expression. J. Biol. Chem. *266*, 8604–8608.
51. Yu, M.-A., Sánchez-Lozada, L.G., Johnson, R.J., and Kang, D.-H. (2010). Oxidative stress with an activation of the renin-angiotensin system in human vascular endothelial cells as a novel mechanism of uric acid-induced endothelial dysfunction. J. Hypertens. *28*, 1234–1242.
52. Sautin, Y.Y., Nakagawa, T., Zharikov, S., and Johnson, R.J. (2007). Adverse effects of the classic antioxidant uric acid in adipocytes: NADPH oxidase-mediated oxidative/nitrosative stress. Am. J. Physiol. Cell Physiol. *293*, C584–C596.
53. Krishnan, E., Pandya, B.J., Chung, L., Hariri, A., and Dabbous, O. (2012). Hyperuricemia in young adults and risk of insulin resistance, prediabetes, and diabetes: a 15-year follow-up study. Am. J. Epidemiol. *176*, 108–116.
54. Jang, W.C., Nam, Y.H., Ahn, Y.C., Park, S.M., Yoon, I.K., Choe, J.-Y., Park, S.-H., Her, M., and Kim, S.-K. (2012). G109T polymorphism of SLC22A12 gene is associated with serum uric acid level, but not with metabolic syndrome. Rheumatol. Int. *32*, 2257–2263.

55. Pfister, R., Barnes, D., Luben, R., Forouhi, N.G., Bochud, M., Khaw, K.-T., Wareham, N.J., and Langenberg, C. (2011). No evidence for a causal link between uric acid and type 2 diabetes: a Mendelian randomisation approach. Diabetologia *54*, 2561–2569.
56. Yang, Q., Köttgen, A., Dehghan, A., Smith, A.V., Glazer, N.L., Chen, M.-H., Chasman, D.I., Aspelund, T., Eiriksdottir, G., Harris, T.B., et al. (2010). Multiple genetic loci influence serum urate levels and their relationship with gout and cardiovascular disease risk factors. Circ Cardiovasc Genet *3*, 523–530.
57. Bhattacharjee, S., Rajaraman, P., Jacobs, K.B., Wheeler, W.A., Melin, B.S., Hartge, P., Yeager, M., Chung, C.C., Chanock, S.J., Chatterjee, N.; and GliomaScan Consortium (2012). A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. Am. J. Hum. Genet. *90*, 821–835.
58. Majumdar, A., Haldar, T., Bhattacharya, S., and Witte, J. (2017). An efficient Bayesian meta-analysis approach for studying cross-phenotype genetic associations. bioRxiv. https://doi.org/10.1101/101543.
59. Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y., and Hinds, D.A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. Nat. Genet. *48*, 709–717.
60. Han, B., Pouget, J.G., Slowikowski, K., Stahl, E., Lee, C.H., Diogo, D., Hu, X., Park, Y.R., Kim, E., Gregersen, P.K., et al.; Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium (2016). A method to decipher pleiotropy by detecting underlying heterogeneity driven by hidden subgroups applied to autoimmune and neuropsychiatric diseases. Nat. Genet. *48*, 803–810.
61. Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the genetic architecture of 30 complex traits from summary association data. Am. J. Hum. Genet. *99*, 139–153.
62. Shi, H., Mancuso, N., Spendlove, S., and Pasaniuc, B. (2016). Local genetic correlation gives insights into the shared genetic architecture of complex traits. bioRxiv. https://doi.org/10.1101/092668.
63. Rao, C.R. (1972). Estimation of variance and covariance components in linear models. J. Am. Stat. Assoc. *67*, 112–115.