# A powerful approach to estimating annotation-stratified genetic covariance using GWAS summary statistics

Qiongshi Lu[1], 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], Hongyu Zhao[1,3,7,*]

[1] Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA
[2] Yale School of Medicine, New Haven, CT, USA
[3] Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA
[4] Yale College, New Haven, CT, USA
[5] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[6] Division of General Internal Medicine, Department of Medicine, University of Washington, Seattle, WA, USA
[7] VA Cooperative Studies Program Coordinating Center, West Haven, CT, USA

* To whom correspondence should be addressed:
Dr. Hongyu Zhao
Department of Biostatistics
Yale School of Public Health
60 College Street,
New Haven, CT, 06520, USA
hongyu.zhao@yale.edu

## **Abstract** (~150 words)

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 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, thus 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's disease (LOAD) and amyotrophic lateral sclerosis (ALS), two major neurodegenerative diseases, in SNPs with high minor allele frequencies and SNPs 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 study (GWAS) has been a success in the past 12 years. Despite a simple study design, GWAS has 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 LD score regression (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's disease (LOAD) and amyotrophic lateral sclerosis (ALS), 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 novel insights into the shared genetic basis of complex traits.

# Results

## Model overview

We briefly outline the statistical framework in this section. Theoretical results and model details are discussed in the **Methods** section and **Supplementary Notes**. We define K possibly overlapping functional annotations $S_1$, $S_2$, ..., $S_K$ (e.g. predicted functional and non-functional genome), and assume two 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 genotype matrices of SNPs in annotation $S_i$; $\beta_i$ and $\gamma_i$ denote the corresponding genetic effects on two traits, which follow an annotation-dependent covariance structure:

$$\mathbb{E}(\beta_i) = \mathbb{E}(\gamma_i) = 0 \ \ and \ \ \mathbb{E}(\gamma_i\beta_i^T) = \frac{\rho_i}{m_i}I, \quad i = 1, \dots, K$$

where $m_i$ denotes the number of SNPs, $\rho_i$ denotes the genetic covariance in annotation category $S_i$, and $\epsilon$ and $\delta$ denote the non-genetic effects for the two traits respectively. To allow for sample overlap between two GWASs with $N_1$ and $N_2$ subjects, assume the first $N_s$ samples in each study are shared. We allow non-genetic effects $\epsilon$ and $\delta$ to be correlated:

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

Next, we study the expectation of $y_1^T A y_2$ where $A$ is an arbitrary matrix. It can be shown that

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

Here, $A_{tt}$ denotes the t[th] diagonal element of matrix $A$. We plug in the following K+1 matrices $\tilde{A}_1$,..., $\tilde{A}_{K+1}$ into the equation above:

$$\tilde{A}_j = \frac{X_jZ_j^T}{m_j}, \quad j = 1, \dots, K$$

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

In addition, we apply method of moments and approximate $\mathbb{E}(y_1^T\tilde{A}_j y_2)$ with $y_1^T\tilde{A}_j y_2$. When the shared sample size is moderate compared to the total sample size (**Supplementary Notes**), we could remove non-genetic covariance $\rho_e$ from the formula and obtain the following K equations.

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

Here, $r^2_{l^{(i)}l'^{(j)}}$ denotes the LD between the $l^{th}$ SNP from category $S_i$ and the $(l')^{th}$ 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 z-scores in annotation $S_j$. We define

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

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

Then, the genetic covariance estimate can be denoted as

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

Of note, $M$ only depends on LD and can be estimated using a reference panel; $v$ can be calculated using GWAS summary statistics. Individual-level genotype or phenotype information is not required in this framework. We apply the following correction if sample overlap is substantial

$$\hat{\rho} = M^{-1}(v - \frac{\hat{\rho}_{pheno}}{N}\mathbf{1})$$

where $\hat{\rho}_{pheno}$ denotes the phenotypic covariance between two traits and can also be estimated using GWAS summary statistics if not provided (**Supplementary Notes**). Standard error is estimated using block-wide jackknife. When performing non-stratified analysis, we can estimate genetic correlation in addition to genetic covariance using the following formula:

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

where heritability is estimated using the estimator proposed in [15] as follows

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

**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 **Methods** section. Since LDSC cannot estimate annotation-stratified genetic covariance, we compared GNOVA and LDSC using data simulated from a non-stratified, infinitesimal genetic covariance structure (**Figures 1A-D**). 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 (**Supplementary Figure 1**). 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 (**Supplementary Figure 2**).

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 (**Methods**). GNOVA provided unbiased estimates for the genetic covariance in each category across all settings (**Figures 1E-F**). 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-H**).

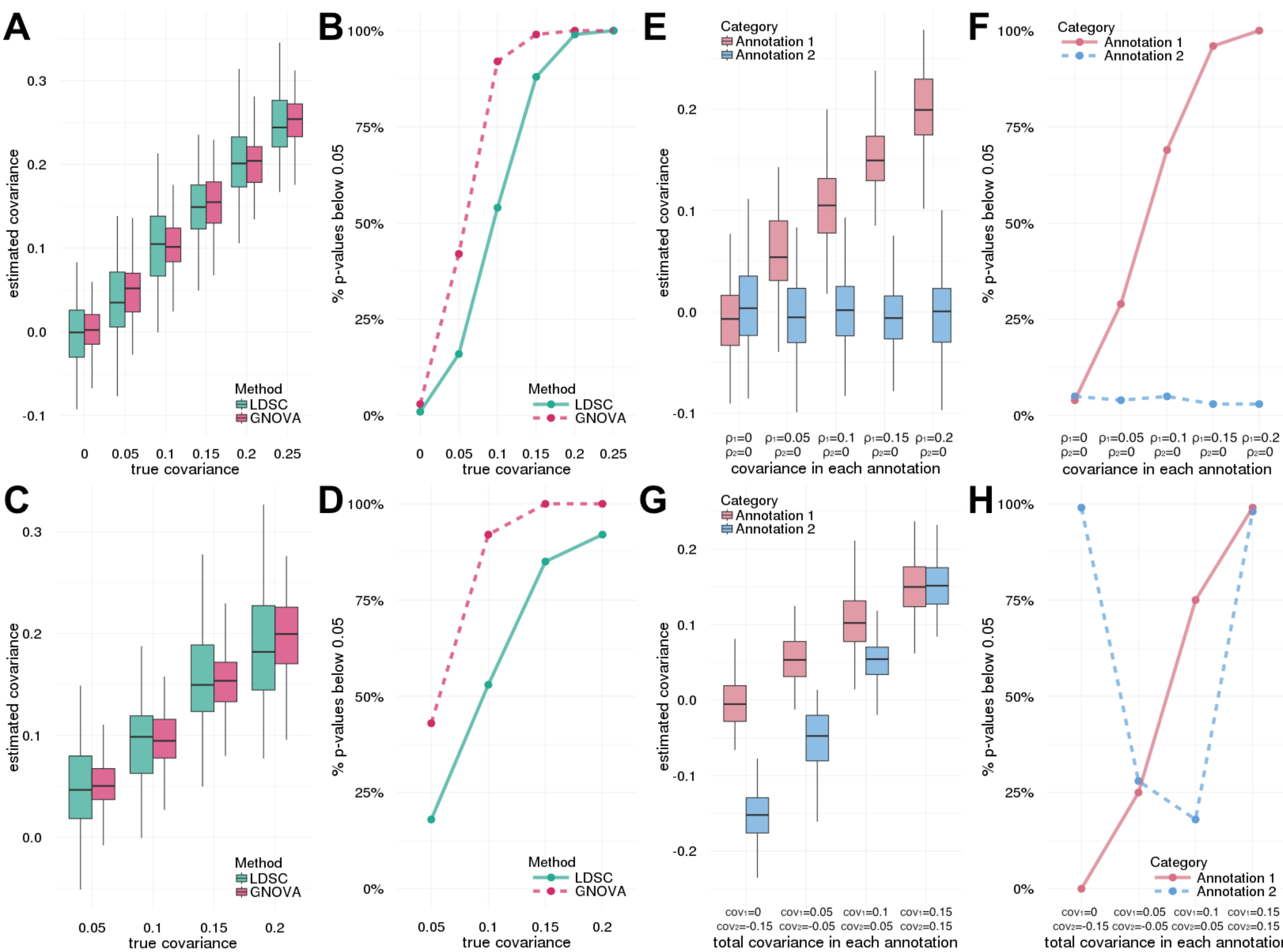


**Figure 1. Evaluation of covariance estimation and statistical power through simulations.** Detailed simulation settings are described in the **Methods** section. **(A-D)** Compare GNOVA and LDSC using traits simulated from a non-stratified covariance structure. We first fixed heritability for both traits but set genetic correlation to different values. The covariance estimates are shown in panel A. Panel 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 panels C and D, respectively. **(E-H)** Estimate annotation-stratified genetic covariance. In panels E and F, we simulated data using two non-overlapping functional annotations. Results in panels 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.

**Table 1.** Acronyms for 50 complex diseases and traits. They will be used throughout this paper.

| Trait | Acronym |
|---|---|
| Age at First Birth | AFB |
| Age at Menarche | AM |
| Age-related Macular Degeneration | AMD |
| Anorexia Nervosa | AN |
| Age at Natural Menopause | ANM |
| Anxiety Disorder | ANX |
| Autism Spectrum Disorder | ASD |
| Asthma | AST |
| Bipolar Disorder | BIP |
| Body Mass Index | BMI |
| Birth Weight | BW |
| Coronary Artery Disease | CAD |
| Crohn's Disease | CD |
| Celiac Disease | CEL |
| Chronotype | CHT |
| Chronic Kidney Disease | CKD |
| Cognitive Performance | COG |
| Diastolic Blood Pressure | DBP |
| Depressive Symptoms | DEP |
| Eczema | ECZ |
| Education Years | EDU |
| Epilepsy | EPL |
| Femoral Neck Bone Mineral Density | FNBMD |
| Fasting Glucose | GLU |
| Gout | GOUT |
| HDL Cholesterol | HDL |
| Height | HGT |
| Inflammatory Bowel Disease | IBD |
| Fasting Insulin | INS |
| LDL Cholesterol | LDL |
| Lumbar Spine Bone Mineral Density | LSBMD |
| Major Depressive Disorder | MDD |
| Multiple Sclerosis | MS |
| Number of Children Ever Born | NCEB |
| Neuroticism | NEU |
| Primary Angle Closure Glaucoma | PACG |
| Primary Billary Cirrhosis | PBC |
| Rheumatoid Arthritis | RA |
| Resting Heart Rate | RHR |
| Systolic Blood Pressure | SBP |
| Schizophrenia | SCZ |
| Systemic Lupus Erythematosus | SLE |
| Smoking Behavior | SMK |
| Serum Urate | SU |
| Subjective Well-being | SWB |
| Type-II Diabetes | T2D |
| Total Cholesterol | TC |
| Triglycerides | TG |
| Ulcerative Colitis | UC |
| Waist Hip Ratio | WHR |

## Estimation of pair-wise genetic correlation for 50 human complex traits

We applied GNOVA to estimate genetic correlations for 50 complex traits using publicly available GWAS summary statistics ($N_{total} \approx 4.7$ million). Trait acronyms are listed in **Table 1** and details of all GWASs are summarized in **Supplementary Table 1**. Out of 1,225 pairs of traits in total, we identified 175 pairs with statistically significant genetic correlation after Bonferroni correction (**Supplementary Table 2** and **Supplementary Figure 3**). We also applied LDSC to the same datasets and only identified 127 significant pairs (**Supplementary Table 3** and **Supplementary Figure 4**). 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.

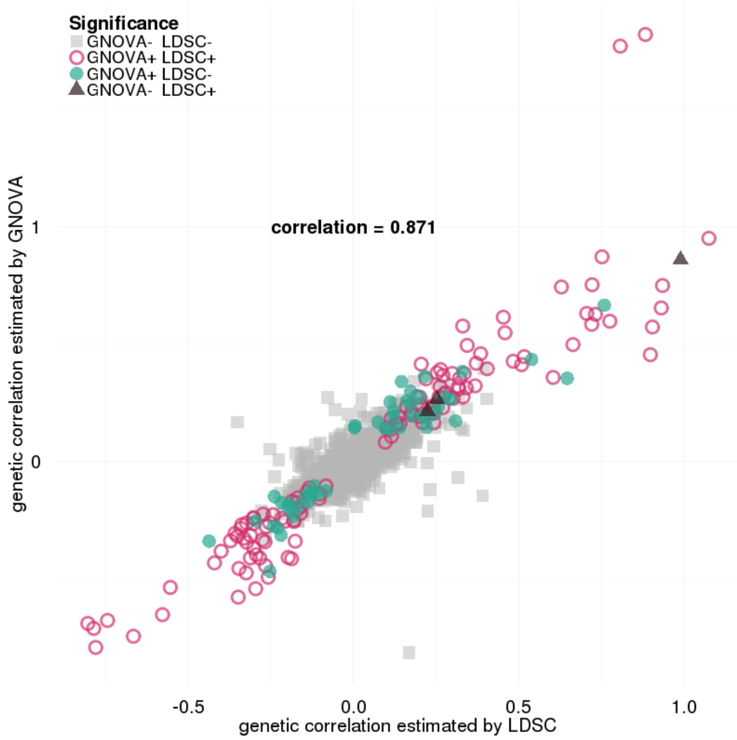


**Figure 2. Comparison of genetic correlations estimated using GNOVA and LDSC.** Each point represents a pair of traits. Overall, genetic correlation estimates are concordant between GNOVA and LDSC, but GNOVA is more powerful when genetic correlation is moderate. Color and shape of each data point represent 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.

To evaluate model validity, we examined correlations between several traits that are closely related either physiologically or epidemiologically (**Supplementary Table 4**). As expected, systolic and diastolic blood pressure (SBP and DBP), femoral and lumbar bone mineral density (FNBMD and LSBMD), and depressive symptoms (DEP) and major depressive disorder (MDD) showed strong positive genetic correlations. We also observed negative correlations between subjective well-being (SWB) and neuropsychiatric disorders such as schizophrenia, anxiety, two depression traits (DEP and MDD) and neuroticism.

We further examined pairwise correlations between 50 traits (**Figure 3**; **Supplementary Figure 3**). Following hierarchical clustering, broad patterns suggesting disease relatedness emerged. These results are well documented in the literature; neuropsychiatric, 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) 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 only passed multiple correction testing at 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) and EDU.

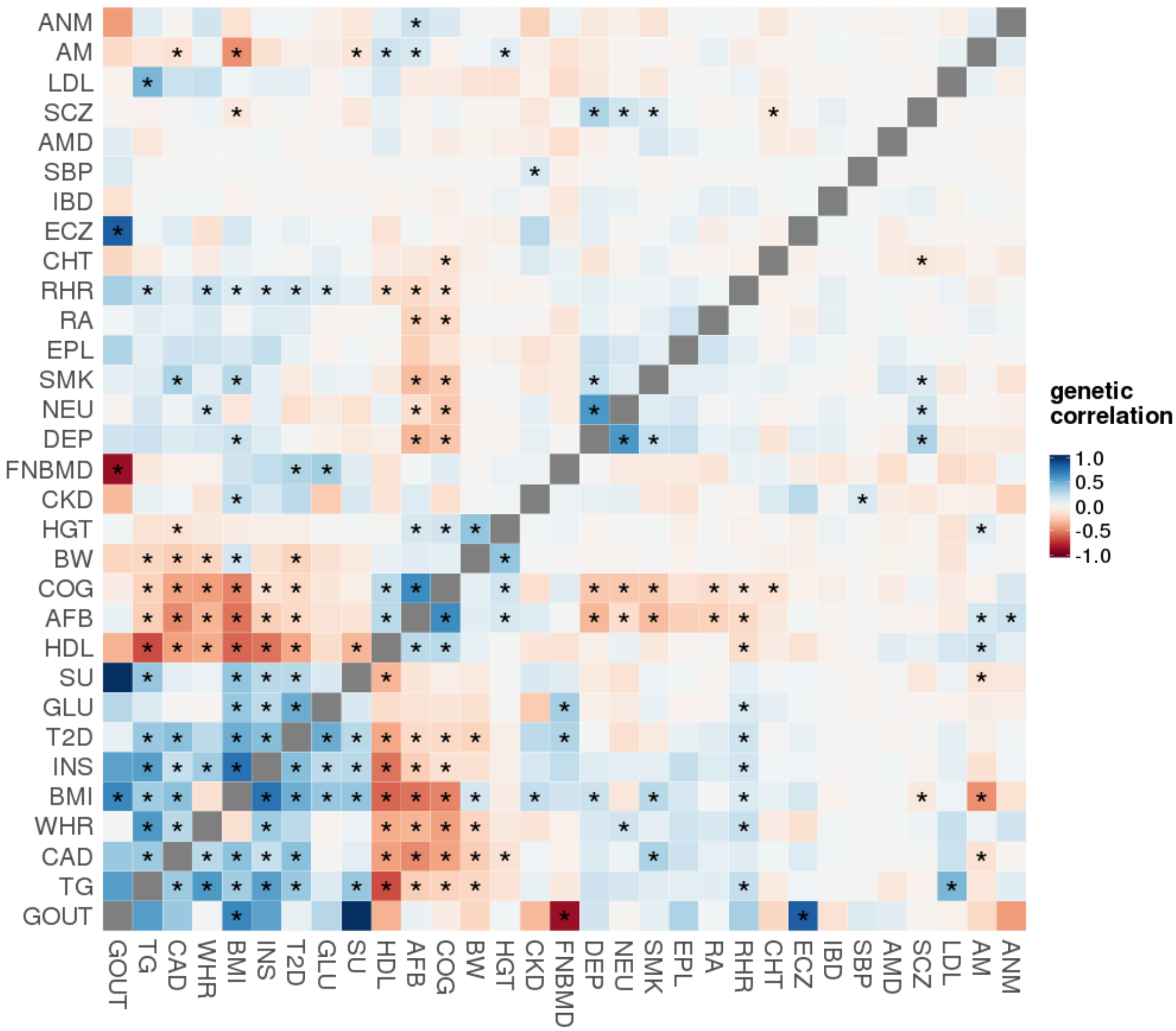


**Figure 3. Estimated genetic correlations of 465 pairs of traits from 31 GWASs.** To visualize a large number of pair-wise correlations more efficiently, we excluded closely related traits and studies with smaller sample sizes (N < 30,000) in this figure. Asterisks highlight significant genetic correlations after Bonferroni correction for all 1,225 pairs ($p < 4.1\times10^{-5}$). The complete heat map matrix is presented in **Supplementary Figure 3**. The order of traits was determined by hierarchical clustering.

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) 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, 16, 17]. We see an unsurprising positive correlation between glucose and insulin levels, which is consistent with our understanding of diabetes [18]. Positive correlations between multiple sclerosis (MS) and Crohn's disease (CD) and more generally, inflammatory bowel disease (IBD), agree with existing reports of shared susceptibility for these diseases [19-21]. We demonstrate a positive correlation between asthma and eczema, which share numerous loci identified in previous GWAS [22]. We found chronic kidney disease (CKD) to be positively correlated with systolic and diastolic blood pressure, consistent with existing evidence for shared genetics via genes such as *UMOD*, which independently predisposes to CKD and hypertension [23]. 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) [24]. 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 [25, 26], and our findings are consistent with previously observed genetic pleiotropy among neuroticism, MDD, BIP, and schizophrenia [27, 28].

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 [29]. Alleles in IRF8, a regulatory factor of type-I interferons, are associated with MS and systemic lupus erythematosus (SLE), 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 [30]. 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-alpha), in serum in ASD and in cerebrospinal fluid in MS [31, 32]. However, previous treatment of MS with anti-TNF-alpha led to an increase in the number of demyelinating lesions and a significantly higher relapse rate [33]. Furthermore, we observed a positive genetic correlation between ulcerative colitis (UC) and primary billary cirrhosis (PBC). CD, also an IBD and thus closely related, has been reported to share susceptibility genes with PBC including *TNFSF15*, *ICOSLG*, and *CXCR5* [34]. 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 50 complex traits. We have previously developed GenoCanyon, a statistical framework to predict functional DNA elements in the human genome through integration of annotation data

[35]. We partitioned the genome into two non-overlapping categories (i.e. functional and non-functional) based on GenoCanyon scores (**Methods**), and estimated genetic covariance within the functional and the non-functional genome for each pair of traits (**Supplementary Table 5**). 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 = 1.0×10$^{-6}$) while the overall covariance did not ($\rho_{overall}$ = 0.062; p = 7.7×10$^{-5}$).

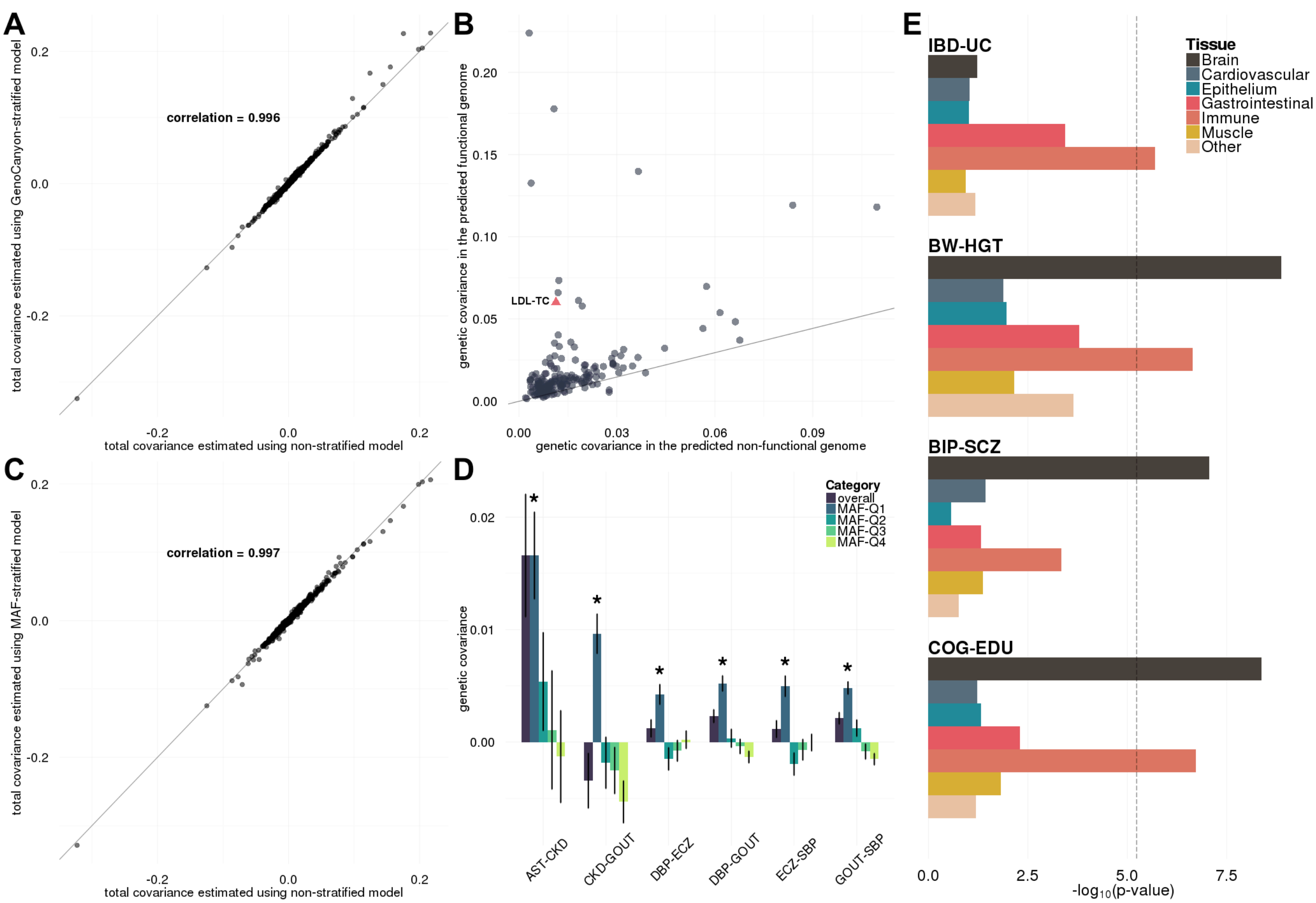


**Figure 4. Annotation-stratified covariance analysis. (A)** Stratify genetic covariance by genome functionality predicted by GenoCanyon. Total genetic covariance estimates were highly concordant between stratified and non-stratified models. **(B)** For significantly correlated pairs of traits based 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 4.1×10$^{-5}$. **(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/(7×1225) = 5.8×10$^{-6}$.

Next, we partitioned genetic covariance based on quartiles of SNPs' minor allele frequencies (MAFs) in subjects with European ancestry from the 1000 Genomes Project (**Methods**; **Supplementary Table 6**). Similar to the previous analysis, we identified high concordance 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 (**Supplementary Figure 5**). However, we identified six pairs of traits that are uniquely correlated in the lowest MAF quartile (**Figure 4D**), namely asthma with CKD ($p = 1.8\times10^{-5}$), gout with CKD ($p = 4.2\times10^{-8}$), DBP ($p = 3.0\times10^{-14}$), and SBP ($p = 4.3\times10^{-18}$), and eczema with DBP ($p = 1.9\times10^{-6}$) and SBP ($p = 6.3\times10^{-8}$). 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, systolic and diastolic BP, as well as immune diseases including asthma and eczema.

Finally, we studied tissue-specificity of genetic covariance through integration of GenoSkyline-Plus annotations (**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 [13]. We utilized seven broadly defined tissue and cell types (i.e. brain, cardiovascular, epithelium, gastrointestinal, immune, muscle, and other) to stratify genetic covariance for 1,225 pairs of traits (**Supplementary Table 7**). Six tests from 4 pairs of traits passed Bonferroni correction, i.e. $p < 0.05/(1225\times7) = 5.8\times10^{-6}$ (**Figure 4E** and **Supplementary Figure 6**). As expected, UC, as an IBD, was significantly and positively correlated with IBD in immune-related functional genome ($p = 2.0\times10^{-6}$); two psychiatric diseases, BIP and schizophrenia, were specifically correlated in the genome predicted to be functional in brain ($p = 8.7\times10^{-8}$). 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 ($p = 3.7\times10^{-4}$). 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 [36, 37], 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 ($N_{LOAD} = 54,162$; $N_{ALS} = 36,052$; **Supplementary Table 8**).

We identified positive and significant genetic correlation between LOAD and ALS (correlation = 0.175, $p = 2.0\times10^{-4}$). LDSC provided similar estimates but failed to achieve significance (**Table 2**). 82.6% of the total genetic covariance between LOAD and ALS is concentrated in 33% of the genome predicted to be functional by GenoCanyon ($p = 8.2\times10^{-5}$). 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. We also performed tissue-stratified analysis using GenoSkyline-Plus annotations (**Supplementary Table 9**). 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.

**Table 2. Dissection of genetic covariance between LOAD and ALS.** Numbers in parentheses indicate standard errors. Significant p-values after adjusting for multiple testing within each section are highlighted in boldface.

| Annotation | Category | Covariance | P-value |
|---|---|---|---|
| Non-stratified | GNOVA | 0.016 (0.004) | **$2.0\times10^{-4}$** |
| | LDSC | 0.012 (0.007) | 0.075[a] |
| GenoCanyon | functional | 0.016 (0.004) | **$8.2\times10^{-5}$** |
| | 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** |

[a] p-value in LDSC was calculated from genetic correlation instead of genetic covariance.

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 (**Supplementary Figure 7**; $p = 0.044$; 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.

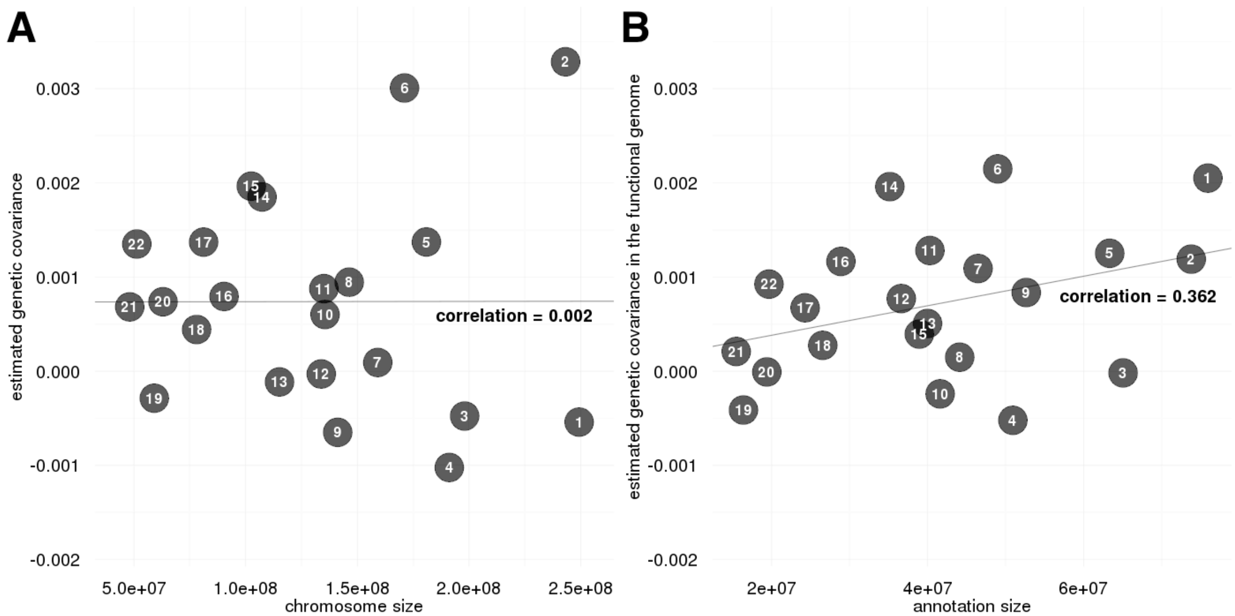


**Figure 5. Stratification of genetic covariance between LOAD and ALS by chromosome. (A)** Comparisons of the estimated per-chromosome genetic covariance with chromosome size. **(B)** Comparisons of the estimated genetic covariance in the predicted functional genome on each chromosome with size of the functional genome.

Finally, we jointly analyzed LOAD, ALS, and 50 other complex traits (**Table 1** and **Supplementary Table 10**). 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 was significant. Instead, ALS was significantly and positively correlated with MS, a neurological disease with a well-established immune component [38]. ALS was also positively correlated with several other immune-related diseases including celiac disease (CEL), asthma, PBC, and IBD (including CD and UC), though none of these was statistically significant. The nominal correlations between ALS and neurological and psychiatric diseases including epilepsy, schizophrenia, BIP, AN, and MDD also remain to be validated in the future using studies with larger sample sizes.

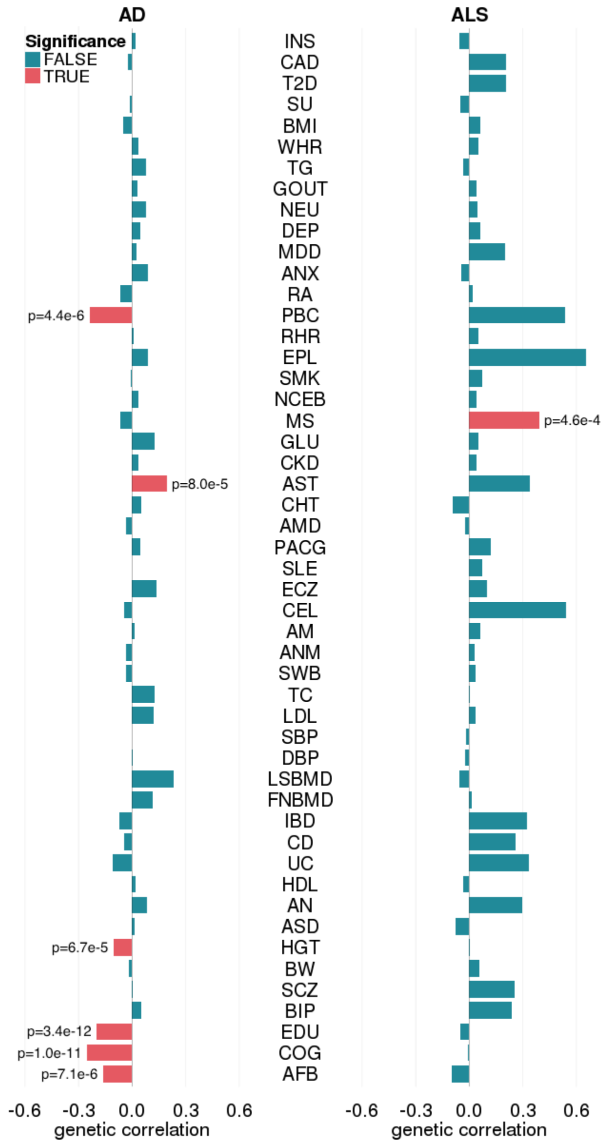


**Figure 6. Genetic correlations between LOAD, ALS, and 50 complex traits.** Significant pairs with $p < 0.05/(50\times2) = 5.0\times10^{-4}$ are highlighted in red.

## Discussion

Although our understanding of complex disease etiology is still far from complete, we have gained valuable knowledge on 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 only focusing 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 [39] and Roadmap Epigenomics Project [40] 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 only requires genome-wide summary statistics and a reference panel as input, and allows stratification of genetic covariance by functional genome annotation.

Numerous studies have hinted at a shared genetic basis among neurodegenerative diseases [41, 42]. 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, 43]. 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 suggested [44]. While associations between hyperuricemia and cardiovascular disease are well described [45], multiple hypotheses exist regarding details of its involvement [46]. For example, hyperuricemia may lead to inflammation in the kidney through vascular smooth muscle proliferation, inducing hypertension via pre-glomerular vascular changes [47]. 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 [48-51]. Despite this evidence, genetic investigations have not identified a strong relationship between hyperuricemia and metabolic syndrome. Polymorphism in gene *SLC22A12* was associated with hyperuricemia but not with metabolic syndrome [52]. 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 [53, 54]. 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 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 [55, 56]. However, due to the large number of SNPs in the genome, statistical power is a critical issue and large-scale inference remains challenging. Second, 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 relationship among traits from pleiotropic effects via independent biological pathways [57]. Han et al. developed a method to distinguish pleiotropy from phenotypic heterogeneity [58]. Although many questions remain unanswered, these recent studies have broadened our view on interpreting complex genetic relationships between human traits. Finally, 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 [59, 60]. This approach provides an alternative methodological option for analyzing genetic effects at specific loci. Our method, in conjunction with these tools, provides the most complete picture to date about shared genetics between complex phenotypes.

In summary, we developed a novel statistical framework to perform powerful, annotation-stratified genetic covariance analysis using GWAS summary statistics. 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.

# Methods

### Model details

Here we outline the genetic covariance estimation framework. The complete derivation, detailed justification for all approximations, and theoretical proofs are presented in the **Supplementary Notes**. We define K functional annotations $S_1$, $S_2$, ..., $S_K$, whose union covers the entire genome; assume two studies share the same list of $m$ SNPs; and assume two 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 genotype matrices defined through annotation $S_i$, $\beta_i$ and $\gamma_i$ denote the corresponding genetic effects for each annotation category. SNPs' genetic effects on two traits follow an annotation-dependent covariance structure:

$$\mathbb{E}(\beta_i) = \mathbb{E}(\gamma_i) = 0 \;\; and \;\; \mathbb{E}\left(\gamma_i\beta_i^T\right) = \frac{\rho_i}{m_i}I, \quad i = 1, \dots, K$$

where $m_i$ and $\rho_i$ denote the total number of SNPs and the total genetic covariance in annotation category $S_i$, respectively. Variables $\epsilon$ and $\delta$ 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 $\epsilon$ and $\delta$ to be correlated:

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

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

### Estimation of covariance parameters

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

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

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 $A_1$,..., $A_{K+1}$ into the equation above. Then, we apply the method of moments to approximate $\mathbb{E}(y_1^T A_j y_2)$ using the observed value $y_1^T A_j y_2$, and get the following equations.

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

Solving this linear system of K+1 equations would get us the covariance estimates.

### Choices of matrix *A*

The estimation procedure described above works for arbitrary $A$ matrices, and it is critical to choose A in practice. Since individual-level genotype and phenotype data from consortium-based GWASs are in many cases not easily accessible, 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:

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

Plugging in these matrices, the first K equations become:

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

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 approximated by (**Supplementary Notes**):

$$\frac{1}{m_j\sqrt{N_1 N_2}}(z_1)_j^T (z_2)_j = \sum_{i=1}^{K} \frac{\rho_i}{m_i m_j} \sum_{l=1}^{m_i} \sum_{l'=1}^{m_j} r^2_{l^{(i)} l'^{(j)}} + \frac{N_s \rho_e}{N_1 N_2}, \quad j = 1, \dots, K$$

Here, $r^2_{l^{(i)} l'^{(j)}}$ denotes the LD between the $l^{\text{th}}$ 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; $(z_1)_j$ and $(z_2)_j$ represent z-scores corresponding to the SNPs in annotation category $S_j$.

Next, we study the $(\text{K}+1)^{\text{th}}$ 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 $N_1 N_2$ on both sides of the $(\text{K}+1)^{\text{th}}$ equation, and we get:

$$\frac{1}{N_1 N_2} \sum_{t=1}^{N_s} (y_1)_t (y_2)_t = \frac{N_s}{N_1 N_2} \sum_{i=1}^{K} \rho_i + \frac{N_s}{N_1 N_2} \rho_e$$

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

$$\begin{pmatrix} \frac{1}{m_1\sqrt{N_1 N_2}}(z_1)_1^T(z_2)_1 - \frac{1}{N_1 N_2}\sum_{t=1}^{N_s}(y_1)_t(y_2)_t \\ \vdots \\ \frac{1}{m_K\sqrt{N_1 N_2}}(z_1)_K^T(z_2)_K - \frac{1}{N_1 N_2}\sum_{t=1}^{N_s}(y_1)_t(y_2)_t \end{pmatrix} = \begin{pmatrix} \frac{1}{m_1 m_1}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_1} r^2_{l^{(1)}l'^{(1)}} - \frac{N_s}{N_1 N_2} & \cdots & \frac{1}{m_K m_1}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_1} r^2_{l^{(K)}l'^{(1)}} - \frac{N_s}{N_1 N_2} \\ \vdots & \ddots & \vdots \\ \frac{1}{m_1 m_K}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_K} r^2_{l^{(1)}l'^{(K)}} - \frac{N_s}{N_1 N_2} & \cdots & \frac{1}{m_K m_K}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_K} r^2_{l^{(K)}l'^{(K)}} - \frac{N_s}{N_1 N_2} \end{pmatrix} \begin{pmatrix} \rho_1 \\ \vdots \\ \rho_K \end{pmatrix}$$

When the sample sizes of both GWASs are large and the sample overlap between two studies is moderate, 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

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

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

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 [61]) and $v$ is only based on GWAS summary statistics. Of note, the same estimation framework can be directly applied to ascertained case-control studies as well (**Supplementary Notes**).

**Special cases**

*1) Two independent GWASs*

If samples from two GWASs do not overlap, then the non-genetic effects $\epsilon$ and $\delta$ are independent and 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 the same covariance estimator:

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

*2) 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 [15].

*3) Two GWASs with substantial sample overlap*

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

$$\begin{pmatrix} \frac{1}{m_1 N}(z_1)_1^T(z_2)_1 - \frac{1}{N^2}\sum_{t=1}^{N}(y_1)_t(y_2)_t \\ \vdots \\ \frac{1}{m_K N}(z_1)_K^T(z_2)_K - \frac{1}{N^2}\sum_{t=1}^{N}(y_1)_t(y_2)_t \end{pmatrix} = \begin{pmatrix} \frac{1}{m_1 m_1}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_1} r_{l^{(1)}l'^{(1)}}^2 & \cdots & \frac{1}{m_K m_1}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_1} r_{l^{(K)}l'^{(1)}}^2 \\ \vdots & \ddots & \vdots \\ \frac{1}{m_1 m_K}\sum_{l=1}^{m_1}\sum_{l'=1}^{m_K} r_{l^{(1)}l'^{(K)}}^2 & \cdots & \frac{1}{m_K m_K}\sum_{l=1}^{m_K}\sum_{l'=1}^{m_K} r_{l^{(K)}l'^{(K)}}^2 \end{pmatrix} \begin{pmatrix} \rho_1 \\ \vdots \\ \rho_K \end{pmatrix}$$

Therefore,

$$\hat{\rho} = M^{-1} \begin{pmatrix} \frac{1}{m_1 N}(z_1)_1^T(z_2)_1 - \frac{1}{N}\hat{\rho}_{pheno} \\ \vdots \\ \frac{1}{m_K N}(z_1)_K^T(z_2)_K - \frac{1}{N}\hat{\rho}_{pheno} \end{pmatrix} = M^{-1}(v - \frac{\hat{\rho}_{pheno}}{N}\mathbf{1})$$

where the phenotypic correlation $\hat{\rho}_{pheno}$ can be either acquired from the literature or estimated using LDSC (**Supplementary Notes**).

**Remarks on overlapping functional annotations**

When functional annotations overlap, the covariance parameter $\rho$ 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

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

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

**Theoretical properties**

Matrices $\tilde{A}_j$ defined in previous sections have two key properties.

1) Vector $v = (y_1^T \tilde{A}_1 y_2, \ldots, y_1^T \tilde{A}_K y_2)^T / N_1 N_2$ can be calculated using GWAS summary statistics.

2) Terms $tr(\tilde{A}_j Z_i X_i^T)$ only depend on LD and can be estimated using a reference panel.

In this section, we show that under reasonable conditions, estimators based on arbitrary $A$ are unbiased but $\hat{\rho} = M^{-1}v$ based on matrices $\tilde{A}_j$ “almost” has the minimum variance. Here we state all the propositions. See **Supplementary Notes** for detailed proofs. To prove the theoretical properties, we need an additional assumption on the distribution of $y_1$ and $y_2$. We assume $y_1$ and $y_2$ follow a multivariate normal distribution:

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

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 (**Supplementary Notes**).

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

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 $p = (p_1, \dots p_K)^T$ 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 $\lambda = (\lambda_1, \dots \lambda_K)^T$ 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 $p = (p_1, \dots p_K)^T$, if $\exists\ \lambda = (\lambda_1, \dots \lambda_K)^T$ 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 $\lambda$, 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 (**Supplementary Notes**). 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 using block-wise jackknife**

Estimating $H_1$ and $H_2$ would require additional assumptions on each trait's 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 GWAS summary statistics. Therefore, following [7], we apply a block-wise jackknife approach to estimate the variance. We divide the genome into $b$ (e.g. $b = 200$) blocks $B_1, \dots, B_b$. Let

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

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:

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

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

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

When annotations overlap,

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

**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 [15] to estimate heritability for each trait.

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

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. Therefore, we only focus on genetic covariance 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 [61]. 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.

$$\begin{pmatrix} \beta \\ \gamma \end{pmatrix} \sim MVN(0, \frac{1}{7959} \begin{pmatrix} h_1^2 I & \rho I \\ \rho I & h_2^2 I \end{pmatrix})$$

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 only assumed the heritability for both traits to be 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. $\rho_1$) is set to be 0.1. We set $\rho_2$ 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 50 GWASs and the URLs for summary statistics files are summarized in **Supplementary Table 1**. For each summary statistics dataset, we applied the same quality control steps described in [7] using the munge_sumstats.py script available at https://github.com/bulik/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 (**Supplementary Figure 8**). 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 [37].

### Annotation data

GenoCanyon and GenoSkyline functional annotations, as previously reported [13, 17, 35], integrate various types of transcriptomic and epigenomic data from ENCODE [39] and Roadmap Epigenomics Project [40] 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 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 available at https://github.com/bulik/ldsc/. 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 [7]. LD scores were estimated using phase-I samples with European ancestry in the 1000 Genomes Project. LD score regression was applied to HAPMAP3 SNPs.

### Software availability

Implemented GNOVA software is available at https://github.com/xtonyjiang/GNOVA

## Acknowledgements

This study was supported in part by the National Institutes of Health grants R01 GM59507, the VA Cooperative Studies Program of the Department of Veterans Affairs, Office of Research and Development, and the Yale World Scholars Program sponsored by the China Scholarship Council. Dr. Crane's and Dr. Mukherjee's efforts were supported by grant R01 AG042437 and U01 AG006781.

This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113, 085475, and 090355. We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. The investigators within IGAP contributed to the design and implementation of IGAP and/or provided data but did not participate in analysis or writing of this report. IGAP was made possible by the generous participation of the control subjects, the patients, and their families. The i-Select chips were funded by the French National Foundation on Alzheimer's disease and related disorders. EADI was supported by the LABEX (laboratory of excellence program investment for the future) DISTALZ grant, Inserm, Institut Pasteur de Lille, Université de Lille 2, and the Lille University Hospital. GERAD was supported by the Medical Research Council (Grant n° 503480), Alzheimer's Research UK (Grant n° 503176), the Wellcome Trust (Grant n° 082604/2/07/Z), and German Federal Ministry of Education and Research (BMBF): Competence Network Dementia (CND) grant n° 01GI0102, 01GI0711, 01GI0420. CHARGE was partly supported by the NIH/NIA grant R01 AG033193 and the NIA AG081220 and AGES contract N01–AG–12100, the NHLBI grant R01 HL105756, the Icelandic Heart Association, and the Erasmus Medical Center and Erasmus University. ADGC was supported by the NIH/NIA grants: U01 AG032984, U24 AG021886, U01 AG016976, and the Alzheimer's Association grant ADGC–10–196728. We are also grateful for all the consortia and investigators that provided publicly accessible GWAS summary statistics.

## Competing financial interests

The authors declare no competing financial interests.

## Author contribution

Q.L. and H.Z. conceived and designed the study. Q.L., B.L., D.C., and W.D. performed the statistical analyses. Q.L., B.L., D.O., and Y.H. developed the statistical model and studied its theoretical properties. M.E. assisted in interpreting genetic correlations. R.L.P. and T.J. developed the software package. Q.L., B.L., Y.H., D.C., W.D., Q.H., and Z.L. implemented the algorithm. B.L. and C.J. implemented jackknife estimation. Q.L., M.E., and H.Z. wrote the manuscript. S.M. and P.K.C. advised on the analysis of neurodegenerative diseases. H.Z. advised on statistical and genetic issues. All authors read and approved the manuscript.

## References

1. Yang, J., et al., *Common SNPs explain a large proportion of the heritability for human height.* Nature genetics, 2010. **42**(7): p. 565-569.
2. Yang, J., et al., *GCTA: a tool for genome-wide complex trait analysis.* The American Journal of Human Genetics, 2011. **88**(1): p. 76-82.
3. Yang, J., et al., *Genome partitioning of genetic variation for complex traits using common SNPs.* Nature genetics, 2011. **43**(6): p. 519-525.
4. Lee, S.H., et al., *Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood.* Bioinformatics, 2012. **28**(19): p. 2540-2542.
5. Vattikuti, S., J. Guo, and C.C. Chow, *Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits.* PLoS Genet, 2012. **8**(3): p. e1002637.
6. Lee, H., et al., *Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs.* Nature Genetics, 2013. **45**(9): p. 984-94.
7. Bulik-Sullivan, B., et al., *An atlas of genetic correlations across human diseases and traits.* Nature genetics, 2015.
8. Pasaniuc, B. and A.L. Price, *Dissecting the genetics of complex traits using summary association statistics.* Nature Reviews Genetics, 2016.
9. Anttila, V., et al., *Analysis of shared heritability in common disorders of the brain.* bioRxiv, 2016: p. 048991.
10. Zheng, J., et al., *LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis.* Bioinformatics, 2016: p. btw613.
11. Zhou, X., *A Unified Framework for Variance Component Estimation with Summary Statistics in Genome-wide Association Studies.* bioRxiv, 2016: p. 042846.
12. Finucane, H.K., et al., *Partitioning heritability by functional annotation using genome-wide association summary statistics.* Nature Genetics, 2015.
13. Lu, Q., et al., *Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease.* bioRxiv, 2016: p. 078865.
14. Gusev, A., et al., *Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases.* The American Journal of Human Genetics, 2014. **95**(5): p. 535-552.
15. Bulik-Sullivan, B., *Relationship between LD Score and Haseman-Elston Regression.* bioRxiv, 2015.
16. Farooqi, I.S., *Defining the neural basis of appetite and obesity: from genes to behaviour.* Clinical Medicine, 2014. **14**(3): p. 286-289.
17. Lu, Q., et al., *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, 2016. **12**(4): p. e1005947.

18. Manning, A.K., et al., *A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance.* Nature genetics, 2012. **44**(6): p. 659-669.
19. Matesanz, F., et al., *A functional variant that affects exon-skipping and protein expression of SP140 as genetic mechanism predisposing to multiple sclerosis.* Human molecular genetics, 2015. **24**(19): p. 5619-5627.
20. Consortium, I.M.S.G. and W.T.C.C.C. 2, *Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis.* Nature, 2011. **476**(7359): p. 214-219.
21. Franke, A., et al., *Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci.* Nat Genet, 2010. **42**(12): p. 1118-25.
22. Marenholz, I., J. Esparza-Gordillo, and Y.-A. Lee, *Shared genetic determinants between eczema and other immune-related diseases.* Current opinion in allergy and clinical immunology, 2013. **13**(5): p. 478-486.
23. Trudu, M., et al., *Common noncoding UMOD gene variants induce salt-sensitive hypertension and kidney damage by increasing uromodulin expression.* Nature medicine, 2013. **19**(12): p. 1655-1660.
24. Reppe, S., et al., *Genetic sharing with cardiovascular disease risk factors and diabetes reveals novel bone mineral density loci.* PloS one, 2015. **10**(12): p. e0144531.
25. Alnaes, R. and S. Torgersen, *Personality and personality disorders predict development and relapses of major depression.* Acta Psychiatrica Scandinavica, 1997. **95**(4): p. 336-342.
26. Kendler, K.S., et al., *A longitudinal twin study of personality and major depression in women.* Archives of general psychiatry, 1993. **50**(11): p. 853-862.
27. De Moor, M.H., et al., *Meta-analysis of genome-wide association studies for neuroticism, and the polygenic association with major depressive disorder.* JAMA psychiatry, 2015. **72**(7): p. 642-650.
28. Gale, C., et al., *Pleiotropy between neuroticism and physical and mental health: findings from 108 038 men and women in UK Biobank.* Translational psychiatry, 2016. **6**(4): p. e791.
29. Sun, H.-L., et al., *Uric acid levels can predict metabolic syndrome and hypertension in adolescents: a 10-year longitudinal study.* PloS one, 2015. **10**(11): p. e0143786.
30. Chrabot, B.S., et al., *Genetic variation near IRF8 is associated with serologic and cytokine profiles in systemic lupus erythematosus and multiple sclerosis.* Genes and immunity, 2013. **14**(8): p. 471-478.
31. Jung, J., I.S. Kohane, and D. Wall, *Identification of autoimmune gene signatures in autism.* Translational psychiatry, 2011. **1**(12): p. e63.
32. Guloksuz, S.A., et al., *Elevated plasma concentrations of S100 calcium-binding protein B and tumor necrosis factor alpha in children with autism spectrum disorders.* Revista Brasileira de Psiquiatria, 2017(AHEAD): p. 0-0.
33. Van Oosten, B., et al., *Increased MRI activity and immune activation in two multiple sclerosis patients treated with the monoclonal anti-tumor necrosis factor antibody cA2.* Neurology, 1996. **47**(6): p. 1531-1534.
34. Aiba, Y., et al., *Disease susceptibility genes shared by primary biliary cirrhosis and Crohn’s disease in the Japanese population.* Journal of human genetics, 2015. **60**(9): p. 525-531.
35. Lu, Q., et al., *A Statistical Framework to Predict Functional Non-Coding Regions in the Human Genome Through Integrated Analysis of Annotation Data.* Sci. Rep., 2015. **5**.
36. Lambert, J.-C., et al., *Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease.* Nature genetics, 2013. **45**(12): p. 1452-1458.

37. van Rheenen, W., et al., *Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis.* Nature genetics, 2016. **48**(9): p. 1043-1048.
38. Haines, J., et al., *A complete genomic screen for multiple sclerosis underscores a role for the major histocompatability complex.* Nature genetics, 1996. **13**(4): p. 469-471.
39. Bernstein, B.E., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
40. Kundaje, A., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-330.
41. Bertram, L. and R.E. Tanzi, *The genetic epidemiology of neurodegenerative disease.* The Journal of clinical investigation, 2005. **115**(6): p. 1449-1457.
42. Nuytemans, K., et al., *Overlap between Parkinson disease and Alzheimer disease in ABCA7 functional variants.* Neurology Genetics, 2016. **2**(1): p. e44.
43. Gagliano, S.A., et al., *Genomics implicates adaptive and innate immunity in Alzheimer's and Parkinson's diseases.* Annals of Clinical and Translational Neurology, 2016. **3**(12): p. 924-933.
44. Reginato, A.M., et al., *The genetics of hyperuricaemia and gout.* Nature reviews Rheumatology, 2012. **8**(10): p. 610-621.
45. Culleton, B.F., et al., *Serum uric acid and risk for cardiovascular disease and death: the Framingham Heart Study.* Annals of internal medicine, 1999. **131**(1): p. 7-13.
46. Kanbay, M., et al., *Uric acid in metabolic syndrome: from an innocent bystander to a central player.* European journal of internal medicine, 2016. **29**: p. 3-8.
47. Watanabe, S., et al., *Uric acid, hominoid evolution, and the pathogenesis of salt-sensitivity.* Hypertension, 2002. **40**(3): p. 355-360.
48. Rao, G.N., M.A. Corson, and B.C. Berk, *Uric acid stimulates vascular smooth muscle cell proliferation by increasing platelet-derived growth factor A-chain expression.* Journal of Biological Chemistry, 1991. **266**(13): p. 8604-8608.
49. Yu, M.-A., et al., *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.* Journal of hypertension, 2010. **28**(6): p. 1234-1242.
50. Sautin, Y.Y., et al., *Adverse effects of the classic antioxidant uric acid in adipocytes: NADPH oxidase-mediated oxidative/nitrosative stress.* American Journal of Physiology-Cell Physiology, 2007. **293**(2): p. C584-C596.
51. Krishnan, E., et al., *Hyperuricemia in young adults and risk of insulin resistance, prediabetes, and diabetes: a 15-year follow-up study.* American journal of epidemiology, 2012. **176**(2): p. 108-116.
52. Jang, W.C., et al., *G109T polymorphism of SLC22A12 gene is associated with serum uric acid level, but not with metabolic syndrome.* Rheumatology international, 2012. **32**(8): p. 2257-2263.
53. Pfister, R., et al., *No evidence for a causal link between uric acid and type 2 diabetes: a Mendelian randomisation approach.* Diabetologia, 2011. **54**(10): p. 2561-2569.
54. Yang, Q., et al., *Multiple Genetic Loci Influence Serum Urate Levels and Their Relationship With Gout and Cardiovascular Disease Risk FactorsClinical Perspective.* Circulation: Cardiovascular Genetics, 2010. **3**(6): p. 523-530.
55. Bhattacharjee, S., et al., *A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits.* The American Journal of Human Genetics, 2012. **90**(5): p. 821-835.
56. Majumdar, A., et al., *An efficient Bayesian meta-analysis approach for studying cross-phenotype genetic associations.* bioRxiv, 2017: p. 101543.
57. Pickrell, J.K., et al., *Detection and interpretation of shared genetic influences on 42 human traits.* Nature genetics, 2016.

58. Han, B., et al., *A method to decipher pleiotropy by detecting underlying heterogeneity driven by hidden subgroups applied to autoimmune and neuropsychiatric diseases.* Nature genetics, 2016. **48**(7): p. 803-810.
59. Shi, H., G. Kichaev, and B. Pasaniuc, *Contrasting the genetic architecture of 30 complex traits from summary association data.* The American Journal of Human Genetics, 2016. **99**(1): p. 139-153.
60. Shi, H., et al., *Local genetic correlation gives insights into the shared genetic architecture of complex traits.* bioRxiv, 2016: p. 092668.
61. Abecasis, G.R., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.