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]

[1]Department of Biostatistics, Yale School of Public Health; [2]Yale School of Medicine; [3]Program of Computational Biology and Bioinformatics, Yale University; [4]Yale College; [5]Department of Computer Science and Engineering, Shanghai Jiao Tong University; [6]Division of General Internal Medicine, Department of Medicine, University of Washington

Model overview

What is genetic covariance? Assume two traits y_1 and y_2 follow linear models:

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

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

where the genetic effects follow (genetic covariance)

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

To estimate genetic covariance, we study the expectation of the following quantity

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

By plugging in K different A matrices and applying the method of moments, we get a linear system of K equations

$$y_1^T A_j y_2 = \sum_{i=1}^{K} \frac{\rho_i}{m_i} tr\left(A_jZ_iX_i^T\right), \quad j = 1, \dots, K$$

Solving these equations gives us a set of covariance estimates. If we choose

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

Then the linear system can be represented by **GWAS summary statistics** and **linkage disequilibrium (LD)**.

$$\frac{1}{m_j}\left(X_j^T y_1\right)^T Z_j^T y_2 = \sum_{i=1}^{K} \frac{\rho_i}{m_i m_j} tr\left(Z_j^T Z_i X_i^T X_j\right), \quad j = 1, \dots, K$$

Or in matrix form:

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

Theoretical properties and simulations

We proved that the GNOVA estimator is an unbiased estimator with minimum variance.

Proposition 1. 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 $S_{ll'} = tr(H_1^{-1}X_{l'}Z_{l'}^TH_2^{-1}Z_lX_l^T)/m_lm_{l'}$ 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 $A_* = \sum_{j=1}^{K}\frac{\lambda_j}{m_j}H_1^{-1}X_jZ_j^TH_2^{-1}$.

Then, we have:
1) $\mathbb{E}(y_1^TA_*y_2) = \sum_{t=1}^{K} p_t\rho_t$;
2) Let A be a matrix such that $\mathbb{E}(y_1^TAy_2) = \sum_{t=1}^{K} p_t\rho_t$. Then, $tr(A^TH_1AH_2) \ge tr(A_*^TH_1A_*H_2)$.

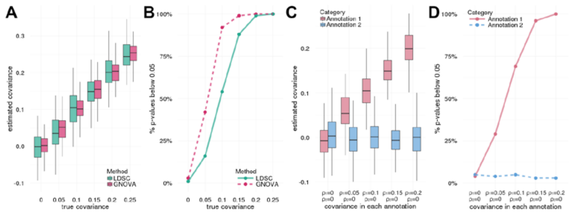

Figure 1. Evaluation of covariance estimation and statistical power through simulations. (A-B) Compare GNOVA and LDSC using traits simulated from a non-stratified covariance structure. The covariance estimates are shown in panel A. Panel B shows the statistical power. **(C-D)** Estimate annotation-stratified genetic covariance. The true covariance values are labeled under each setting. Type-I error was not inflated when the true covariance was zero.

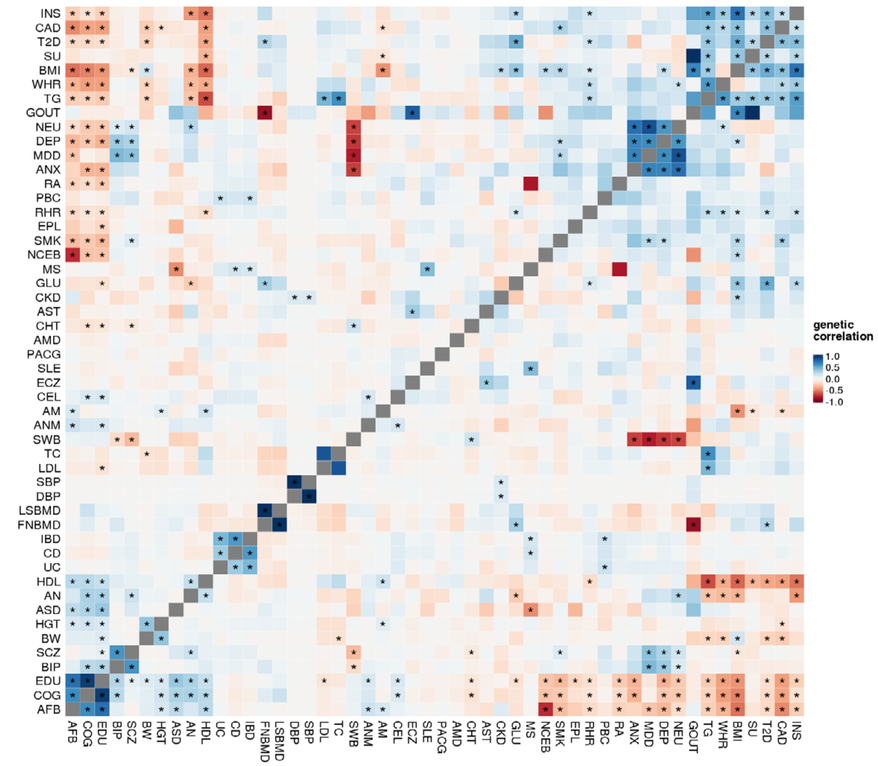

Figure 2. Genetic correlations of 50 complex traits estimated GNOVA. Asterisks highlight significant genetic correlations after Bonferroni correction ($p < 4.1\times10\text{-}5$). The order of traits is determined by hierarchical clustering.

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

Table 1. Acronyms for 50 complex diseases and traits.

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. Out of 1,225 pairs of traits in total, we identified 175 pairs with statistically significant genetic correlation after Bonferroni correction. Consistent with our simulation results, GNOVA is more powerful when the true genetic correlation is moderate.

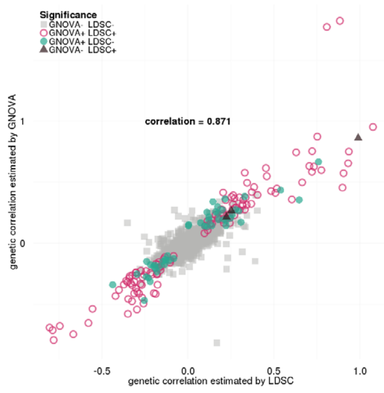

Figure 3. Comparison of genetic correlations estimated using GNOVA and LDSC. Each point represents a pair of traits. Overall, genetic correlation estimates are concordant between GNOVA and LDSC, but GNOVA is more powerful when genetic correlation is moderate. Color and shape of each data point represent the significance status given by GNOVA and LDSC.

Annotation-stratified analyses

Next, we stratified all 1,225 pairs of genetic covariance by predicted genome functionality, tissue-specific functionality, and minor allele frequencies (MAFs).

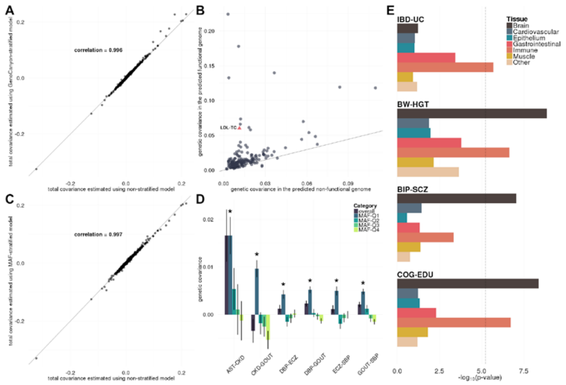

Figure 4. Annotation-stratified covariance analysis. (A) Stratify genetic covariance by genome functionality. Functionality-stratified genetic covariance is concordant with non-stratified estimates. **(B)** Comparison between genetic covariance in the functional and the non-functional genome. Solid line marks the expected value based on annotation size. **(C)** MAF-stratified genetic covariance is concordant with non-stratified estimates. **(D)** Traits that are uniquely correlated in the lowest MAF quartile. **(E)** Stratify genetic covariance by tissue type. Each bar denotes the log-transformed p-value. Dashed line highlights the Bonferroni-corrected significance level $0.05/(7\times1225) = 5.8\times10^{-6}$.

An in-depth case study on LOAD and ALS

Finally, we applied GNOVA to dissect the genetic covariance between late-onset Alzheimer's disease (LOAD) and amyotrophic lateral sclerosis (ALS) using publicly available GWAS summary statistics ($N_{LOAD} = 54{,}162$; $N_{ALS} = 36{,}052$).

Annotation	Category	Covariance	P-value
Non-stratified	GNOVA	0.016 (0.004)	**2.0×10^{-4}**
	LDSC	0.012 (0.007)	0.075[a]
GenoCanyon	functional	0.016 (0.004)	**8.2×10^{-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**

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.

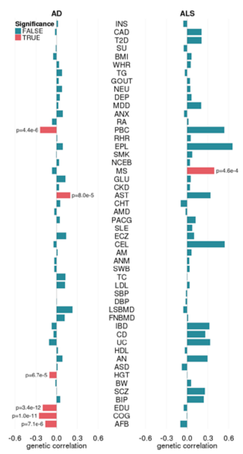

Figure 5. Genetic correlations between LOAD, ALS, and 50 complex traits. Significant pairs with $p < 0.05/(50\times2) = 5.0\times10^{-4}$ are highlighted in red.

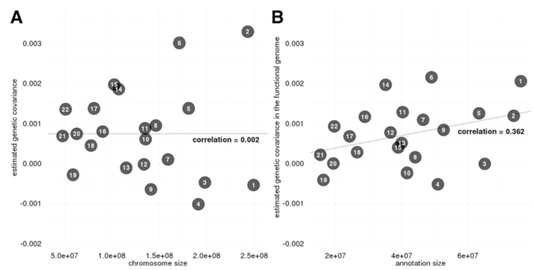

Figure 6. Stratification of genetic covariance between LOAD and ALS by chromosome. (A) Comparisons of the estimated per-chromosome genetic covariance with chromosome size. **(B)** Comparisons of the estimated genetic covariance in the predicted functional genome on each chromosome with size of the functional genome.

Software availability

The GNOVA software is publicly accessible on Github: https://github.com/xtonyjiang/GNOVA

Reference

Lu, Qiongshi, et al. "A powerful approach to estimating annotation-stratified genetic covariance using GWAS summary statistics." bioRxiv (2017): 114561.

Developer

Qiongshi Lu will join University of Wisconsin-Madison as an assistant professor in the Department of Biostatistics & Medical Informatics this summer. His research focuses on genome functional annotations and their applications in human genetics.

Positions Available! (qiongshi.lu@yale.edu)