Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

On Genetic Correlation Estimation With Summary Statistics From Genome-Wide Association Studies

Bingxin Zhao & Hongtu Zhu

To cite this article: Bingxin Zhao & Hongtu Zhu (2022) On Genetic Correlation Estimation With Summary Statistics From Genome-Wide Association Studies, Journal of the American Statistical Association, 117:537, 1-11, DOI: 10.1080/01621459.2021.1906684

To link to this article: https://doi.org/10.1080/01621459.2021.1906684

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 117, NO. 537, 1–11: Applications and Case Studies
https://doi.org/10.1080/01621459.2021.1906684

On Genetic Correlation Estimation With Summary Statistics From Genome-Wide Association Studies

Bingxin Zhao and Hongtu Zhu

Department of Biostatistics, University of North Carolina at Chapel Hill, NC

ABSTRACT

Cross-trait polygenic risk score (PRS) method has gained popularity for assessing genetic correlation of complex traits using summary statistics from biobank-scale genome-wide association studies (GWAS). However, empirical evidence has shown a common bias phenomenon that highly significant cross-trait PRS can only account for a very small amount of genetic variance (R^2 can be $< 1\%$) in independent testing GWAS. The aim of this paper is to investigate and address the bias phenomenon of cross-trait PRS in numerous GWAS applications. We show that the estimated genetic correlation can be asymptotically biased toward zero. A consistent cross-trait PRS estimator is then proposed to correct such asymptotic bias. In addition, we investigate whether or not SNP screening by GWAS *p*-values can lead to improved estimation and show the effect of overlapping samples among GWAS. We analyze GWAS summary statistics of reaction time and brain structural magnetic resonance imaging-based features measured in the Pediatric Imaging, Neurocognition, and Genetics study. We find that the raw cross-trait PRS estimators heavily underestimate the genetic similarity between cognitive function and human brain structures (mean $R^2 = 1.32\%$), whereas the bias-corrected estimators uncover the moderate degree of genetic overlap between these closely related heritable traits (mean $R^2 = 22.42\%$). Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.

ARTICLE HISTORY
Received August 2019
Accepted March 2021

KEYWORDS
Bias correction; Genome-wide association studies; Marginal screening; Polygenic risk score

1. Introduction

The major aim of many genome-wide association studies (GWAS) is to examine the genetic influences of common genetic variants on complex human traits given that many traits have a polygenic architecture (Boyle, Li, and Pritchard 2017). That is, a large number of genetic variants, typically single nucleotide polymorphisms (SNPs), have small but nonzero contributions to the phenotypic variation. In GWAS, many statistical methods have been developed on the use of individual-level common SNP (minor allele frequency [MAF] ≥ 0.05) data to infer the heritability and cross-trait genetic correlation in general populations. For instance, heritability can be estimated by aggregating the small contributions of a large number of common SNP markers, resulting in the SNP heritability estimator (Yang et al. 2010). Moreover, genetic correlation quantifies the shared genetic influences between two heritable phenotypes and is traditionally estimated in family studies. GWAS data offer an alternative to family studies for genetic correlation estimation using independent individuals. Specifically, GWAS data are able to measure the genetic similarity attributable to common SNPs, which can be calculated as the inner product of genetic effects of SNPs on the two traits (see Definition 1).

Accessing individual-level SNP data is often inconvenient due to policy restrictions, and a recent standard practice in the genetic community is to share the summary association statistics, including the estimated effect size, standard error, *p*-value, and sample size *n*, of all SNPs after GWAS are published. Therefore, it has become an active research area to examine the heritability and cross-trait genetic correlation based on GWAS summary statistics. Among them, the cross-trait polygenic risk score (PRS) (Purcell et al. 2009) has become a popular routine to measure genetic similarity of polygenic traits with widespread applications. Compared with other popular methods such as the cross-trait linkage disequilibrium (LD) score regression (Bulik-Sullivan et al. 2015) (cross-trait LDSC) and BOLT-REML (Loh et al. 2015), cross-trait PRS offers at least two unique strengths as follows. First, cross-trait PRS only requires the summary statistics of one trait obtained from a large discovery GWAS, while it allows the individual-level data of the other trait to be collected in a much smaller testing GWAS. In contrast, most other methods require large GWAS data for both traits at either summary or individual-level. For example, the input summary statistics of cross-trait LDSC should be generated from large-scale GWAS (Ni et al. 2018), whose sample size is typically larger than 5000. The main reason is that summary statistics estimated from small GWAS tend to be noisy and thus the cross-trait LDSC estimates may have large standard errors. However, cross-trait PRS can avoid this issue as it directly utilizes the individual-level data of small testing GWAS. Second, SNP selection can be easily implemented in cross-trait PRS, enabling flexible PRS construction for traits with different genetic architectures. However, given these strengths of cross-trait PRS, empirical evidence has

CONTACT Hongtu Zhu *htzhu@email.unc.edu* Department of Biostatistics, University of North Carolina at Chapel Hill.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

shown a common bias phenomenon that even highly significant cross-trait PRS can only account for a very small amount of variance (R^2 can be $< 1\%$) when dissecting the shared genetic basis among highly related heritable traits (Bogdan, Baranger, and Agrawal 2018). Except for some introductory studies, such as Dudbridge (2013), few attempts have been made to study cross-trait PRS and to explain such a counterintuitive phenomenon in real GWAS applications.

This paper fills this significant gap with the following contributions. By comprehensively investigating the properties of cross-trait PRS for polygenic/omnigenic traits, our first contribution in Section 2 is to show that the estimated genetic correlation may asymptotically biased toward zero, uncovering that the underlying genetic overlap can be seriously underestimated. Furthermore, when all SNPs are used in cross-trait PRS, we show that the asymptotic bias is largely determined by the triple (n, p, h^2) and is independent of the unknown number of causal SNPs of the two traits, where n is GWAS sample size, p is the number of SNPs, and h^2 is heritability. Thus, our second contribution in Section 2 is to propose a consistent estimator by correcting such asymptotic bias in cross-trait PRS. Next, in Section 3, we show that when cross-trait PRS is constructed using q top-ranked SNPs whose GWAS p-values pass a given threshold, in addition to (n, p, h^2), the asymptotic bias will also be determined by the number of causal SNPs m, since the ratio m/n determines the quality of the q selected SNPs. Based on these results, we provide practical guidelines for assessing the m/n ratio and minimizing the potential bias in GWAS applications. In Section 4, we generalize our results to quantify the influence of overlapping samples among GWAS. We show that our bias-corrected estimator for independent GWAS can be smoothly extended to GWAS with partially or even fully overlapping samples.

We apply cross-trait PRS to examine the genetic similarity between cognitive function and human brain structures. Specifically, we use the summary statistics of reaction time generated by a recent large-scale meta-analysis GWAS (Davies et al. 2018) and evaluate the performance of cross-trait PRS in the Pediatric Imaging, Neurocognition, and Genetics (PING) study (Jernigan et al. 2016). In a preliminary positive control analysis, we illustrate that the raw cross-trait PRS estimator is biased toward zero and the proposed bias-corrected estimator provides the expected genetic correlation in the PING study. We then show that raw estimators underestimate the genetic similarity between cognitive function and human brain structures (mean $R^2 = 1.32\%$), whereas the bias-corrected estimators suggest that there are moderate genetic correlations between cognitive function and human brain structures (mean $R^2 = 22.42\%$).

The remainder of this article is structured as follows. Sections 2 and 3 study the cross-trait PRS with all SNPs and selected SNPs, respectively. Section 4 considers the effect of overlapping samples among different GWAS. Section 5 summarizes the results from numerical experiments using simulated and real SNP data. In Section 6, we provide the real data analysis in the PING study. The article concludes with some discussions in Section 7.

2. Cross-Trait PRS With All SNPs

2.1. General Setup

We first introduce the modeling framework to investigate the cross-trait PRS, including the genetic architecture of polygenic traits, distribution of genetic effects, and genetic correlation estimators.

2.1.1. Polygenic Traits

Consider three independent GWAS that are conducted for three different traits as follows: (i) Discovery GWAS-I: $(\boldsymbol{X}, \boldsymbol{y}_\alpha)$, with $\boldsymbol{X} = [\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}] \in \mathbb{R}^{n_1 \times p}$, $\boldsymbol{X}_{(1)} \in \mathbb{R}^{n_1 \times m_\alpha}$, and $\boldsymbol{y}_\alpha \in \mathbb{R}^{n_1 \times 1}$; (ii) Discovery GWAS-II: $(\boldsymbol{Z}, \boldsymbol{y}_\beta)$, with $\boldsymbol{Z} = [\boldsymbol{Z}_{(1)}, \boldsymbol{Z}_{(2)}] \in \mathbb{R}^{n_2 \times p}$, $\boldsymbol{Z}_{(1)} \in \mathbb{R}^{n_2 \times m_\beta}$, and $\boldsymbol{y}_\beta \in \mathbb{R}^{n_2 \times 1}$; and (iii) Target testing GWAS: $(\boldsymbol{W}, \boldsymbol{y}_\eta)$, with $\boldsymbol{W} = [\boldsymbol{W}_{(1)}, \boldsymbol{W}_{(2)}] \in \mathbb{R}^{n_3 \times p}$, $\boldsymbol{W}_{(1)} \in \mathbb{R}^{n_3 \times m_\eta}$, and $\boldsymbol{y}_\eta \in \mathbb{R}^{n_3 \times 1}$. Here $\boldsymbol{y}_\alpha$, $\boldsymbol{y}_\beta$, and $\boldsymbol{y}_\eta$ are three different continuous phenotypes studied in three GWAS with sample sizes n_1, n_2, and n_3, respectively. We will use $(\boldsymbol{y}_\alpha, \boldsymbol{y}_\beta)$ and $(\boldsymbol{y}_\alpha, \boldsymbol{y}_\eta)$ to investigate two different bias phenomena in later sections, respectively. The m_α, m_β, and m_η are different numbers of causal SNPs in these GWAS. The $\boldsymbol{X}_{(1)}$, $\boldsymbol{Z}_{(1)}$, and $\boldsymbol{W}_{(1)}$ denote the causal SNPs of $\boldsymbol{y}_\alpha$, $\boldsymbol{y}_\beta$, and $\boldsymbol{y}_\eta$, respectively, and $\boldsymbol{X}_{(2)}$, $\boldsymbol{Z}_{(2)}$, and $\boldsymbol{W}_{(2)}$ donate the corresponding null SNPs. Thus, $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{W}$ are three matrices of p SNPs. Furthermore, we assume that column-wise standardization on $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{W}$ is performed such that each variable has sample mean zero and sample variance one. Based on these notations, we may introduce the following condition on SNP data:

Condition 1. Entries of $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{W}$ are real-value independent random variables with mean zero, variance one and a finite eighth order moment. For $(\boldsymbol{X}, \boldsymbol{y}_\alpha)$, as $n_1, p \to \infty$, we assume $m_\alpha/n_1 = \gamma_\alpha \to \gamma_{\alpha 0}$ and $m_\alpha/p = \omega_\alpha \to \omega_{\alpha 0}$ for $0 < \gamma_{\alpha 0} \leq \infty$ and $0 \leq \omega_{\alpha 0} \leq 1$. In addition, let $m_\beta/n_2 = \gamma_\beta$, $m_\beta/p = \omega_\beta$, $m_\eta/n_3 = \gamma_\eta$, and $m_\eta/p = \omega_\eta$, similar condition holds for $(\boldsymbol{Z}, \boldsymbol{y}_\beta)$ and $(\boldsymbol{W}, \boldsymbol{y}_\eta)$ as $n_2, n_3, p \to \infty$.

In Condition 1, the number of causal SNPs m is allowed to be proportional to the number of total SNPs p, reflecting the widespread dense signals of polygenic/omnigenic traits across the whole genome (Boyle, Li, and Pritchard 2017). To generate reliable summary statistics for polygenic traits, sample size n in discovery GWAS typically needs to be large enough compared with m. On the other hand, n is allowed to be much smaller than m in testing GWAS. Thus, we assume GWAS sample size to be on the same scale as m or smaller than m. In addition, the standardization assumption on SNP data is for notational convenience and our main conclusions remain unchanged for unstandardized data.

The linear polygenic model assumes

$$\boldsymbol{y}_\alpha = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon}_\alpha, \quad \boldsymbol{y}_\beta = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}_\beta, \quad \text{and} \quad \boldsymbol{y}_\eta = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon}_\eta, \tag{1}$$

where $\boldsymbol{\alpha}^T = (\boldsymbol{\alpha}_{(1)}^T, \boldsymbol{\alpha}_{(2)}^T)$, $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_{(1)}^T, \boldsymbol{\beta}_{(2)}^T)$, and $\boldsymbol{\eta}^T = (\boldsymbol{\eta}_{(1)}^T, \boldsymbol{\eta}_{(2)}^T)$ are $p \times 1$ vectors of SNP effects, in which $\boldsymbol{\alpha}_{(2)}$, $\boldsymbol{\beta}_{(2)}$, and $\boldsymbol{\eta}_{(2)}$ are zeros, and $\boldsymbol{\epsilon}_\alpha$, $\boldsymbol{\epsilon}_\beta$, and $\boldsymbol{\epsilon}_\eta$ represent independent random error vectors. The $\boldsymbol{\alpha}_{(1)}$, $\boldsymbol{\beta}_{(1)}$, and $\boldsymbol{\eta}_{(1)}$ are random

vectors (Dobriban and Wager 2018; Jiang et al. 2016), and the distribution assumption will be detailed below.

The overall genetic heritability of $\boldsymbol{y}_\alpha$ is, therefore, given by $h_\alpha^2 = \mathrm{var}(\boldsymbol{X\alpha})/v(\boldsymbol{y}_\alpha) = \mathrm{var}(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)})/\{\mathrm{var}(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)}) + \mathrm{var}(\boldsymbol{\epsilon_\alpha})\}$, which measures the proportion of variation in $\boldsymbol{y}_\alpha$ that can be explained by the aggregated genetic variation $\mathrm{var}(\boldsymbol{X\alpha})$. The $\boldsymbol{y}_\alpha$ is fully heritable when $h_\alpha^2 = 1$. Similarly, we can define the heritability h_β^2 of $\boldsymbol{y}_\beta$ and h_η^2 of $\boldsymbol{y}_\eta$, respectively. We assume h_α^2, h_β^2, and $h_\eta^2 \in (0, 1]$. The cross-trait genetic correlation in this article is defined as the inner product of SNP effects on pairs of phenotypes (Bulik-Sullivan et al. 2015; Lu et al. 2017; Pasaniuc and Price 2017; Shi et al. 2017; Guo et al. 2019).

Definition 1 (Cross-trait Genetic Correlation). The (cross-trait) genetic correlation between $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\eta$ and that between $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\beta$ are respectively given by inner products $\varphi_{\alpha\eta} = \boldsymbol{\alpha}^T\boldsymbol{\eta}/(\|\boldsymbol{\alpha}\| \cdot \|\boldsymbol{\eta}\|) \cdot \mathrm{I}(\|\boldsymbol{\alpha}\| \cdot \|\boldsymbol{\eta}\| > 0)$ and $\varphi_{\alpha\beta} = \boldsymbol{\alpha}^T\boldsymbol{\beta}/(\|\boldsymbol{\alpha}\| \cdot \|\boldsymbol{\beta}\|) \cdot \mathrm{I}(\|\boldsymbol{\alpha}\| \cdot \|\boldsymbol{\beta}\| > 0)$. where $\mathrm{I}(\cdot)$ is the indicator function, $\|\cdot\|$ is the l_2 norm of a vector, and $\varphi_{\alpha\eta}$ and $\varphi_{\alpha\beta} \in [-1, 1]$.

2.1.2. Genetic Effects

In this section, we introduce the distribution assumption on nonzero genetic effects $\boldsymbol{\alpha}_{(1)}$, $\boldsymbol{\beta}_{(1)}$ and $\boldsymbol{\eta}_{(1)}$. Since m_α, m_β and m_η can be different and the causal SNPs of different phenotypes may partially overlap, we let $m_{\alpha\eta}$ be the number of overlapping causal SNPs of $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\eta$, and $m_{\alpha\beta}$ be the number of overlapping causal SNPs of $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\beta$. Let $F(0, V)$ represent a generic distribution with mean zero, (co)variance V, and finite fourth-order moments. We introduce the following condition on genetic effects and random errors.

Condition 2. As $\min(n_1, n_3, p) \to \infty$, $\min(m_{\alpha\eta}, m_\alpha, m_\eta) \to \infty$, we assume $m_{\alpha\eta}/\sqrt{m_\alpha m_\eta} = \kappa_{\alpha\eta} \to \kappa_{0\alpha\eta} \in (0, 1]$. Similarly, as $\min(n_1, n_2, n_3, p) \to \infty$, $\min(m_{\alpha\beta}, m_\alpha, m_\beta) \to \infty$, we assume $m_{\alpha\beta}/\sqrt{m_\alpha m_\beta} = \kappa_{\alpha\beta} \to \kappa_{0\alpha\beta} \in (0, 1]$. α_i, β_j, and η_k are independent random variables satisfying $\alpha_i \sim F(0, \sigma_\alpha^2/p)$, $i = 1, ..., m_\alpha$; $\beta_j \sim F(0, \sigma_\beta^2/p)$, $j = 1, ..., m_\beta$; $\eta_k \sim F(0, \sigma_\eta^2/p)$, $k = 1, ..., m_\eta$, where σ_α^2, σ_β^2, and σ_η^2 are positive scalars. The $m_{\alpha\eta}$ overlapping nonzero effects (α_i, η_i)s of $(\boldsymbol{y}_\alpha, \boldsymbol{y}_\eta)$ and $m_{\alpha\beta}$ overlapping nonzero effects (α_j, β_j)s of $(\boldsymbol{y}_\alpha, \boldsymbol{y}_\beta)$ satisfy

$$\begin{pmatrix}\alpha_i \\ \eta_i\end{pmatrix} \sim F\left[\begin{pmatrix}0 \\ 0\end{pmatrix}, p^{-1} \cdot \begin{pmatrix}\sigma_\alpha^2 & \sigma_{\alpha\eta} \\ \sigma_{\alpha\eta} & \sigma_\eta^2\end{pmatrix}\right] \quad \text{and}$$
$$\begin{pmatrix}\alpha_j \\ \beta_j\end{pmatrix} \sim F\left[\begin{pmatrix}0 \\ 0\end{pmatrix}, p^{-1} \cdot \begin{pmatrix}\sigma_\alpha^2 & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & \sigma_\beta^2\end{pmatrix}\right],$$

respectively, where $\sigma_{\alpha\eta} = \rho_{\alpha\eta} \cdot \sigma_\alpha\sigma_\eta$ and $\sigma_{\alpha\beta} = \rho_{\alpha\beta} \cdot \sigma_\alpha\sigma_\beta$. And $\epsilon_{\alpha_i}, \epsilon_{\beta_j}$ and ϵ_{η_k} are independent random variables satisfying $\epsilon_{\alpha i} \sim F(0, \sigma_{\epsilon_\alpha}^2)$, $i = 1, ..., n_1$; $\epsilon_{\beta j} \sim F(0, \sigma_{\epsilon_\beta}^2)$, $j = 1, ..., n_2$; $\epsilon_{\eta k} \sim F(0, \sigma_{\epsilon_\eta}^2)$, $k = 1, ..., n_3$, where $\sigma_{\epsilon_\alpha}^2, \sigma_{\epsilon_\beta}^2$, and $\sigma_{\epsilon_\eta}^2$ are positive scalars.

Since the three GWAS have independent samples, their random errors are assumed to be independent. Overlapping samples and the induced nongenetic correlation will be studied in Section 4. The genetic correlation between $\boldsymbol{y}_\beta$ and $\boldsymbol{y}_\eta$ and that between $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\eta$ have similar asymptotic properties. To save space, we do not explicitly study the genetic correlation between $\boldsymbol{y}_\beta$ and $\boldsymbol{y}_\eta$, and the related definitions (e.g., the joint distribution between nonzero effects (β_i, η_i)s) are omitted in Condition 2. The cross-trait genetic correlation between $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\eta$ is asymptotically given by $\varphi_{\alpha\eta} = \boldsymbol{\alpha}^T\boldsymbol{\eta}/(\|\boldsymbol{\alpha}\| \cdot \|\boldsymbol{\eta}\|) = m_{\alpha\eta}/(m_\alpha m_\eta)^{1/2} \cdot \rho_{\alpha\eta} + o_p(1) = \kappa_{0\alpha\eta} \cdot \rho_{\alpha\eta} + o_p(1)$ and the genetic correlation between $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\beta$ is asymptotically given by $\varphi_{\alpha\beta} = \boldsymbol{\alpha}^T\boldsymbol{\beta}/(\|\boldsymbol{\alpha}\| \cdot \|\boldsymbol{\beta}\|) = m_{\alpha\beta}/(m_\alpha m_\beta)^{1/2} \cdot \rho_{\alpha\beta} + o_p(1) = \kappa_{0\alpha\beta} \cdot \rho_{\alpha\beta} + o_p(1)$. As in Jiang et al. (2016), heritability h_α^2, h_β^2, and h_η^2 can be asymptotically represented as follows: $h_\alpha^2 = \{(m_\alpha/p)\sigma_\alpha^2\}/\{(m_\alpha/p)\sigma_\alpha^2 + \sigma_{\epsilon_\alpha}^2\}$, $h_\beta^2 = \{(m_\beta/p)\sigma_\beta^2\}/\{(m_\beta/p)\sigma_\beta^2 + \sigma_{\epsilon_\beta}^2\}$, and $h_\eta^2 = \{(m_\eta/p)\sigma_\eta^2\}/\{(m_\eta/p)\sigma_\eta^2 + \sigma_{\epsilon_\eta}^2\}$.

The aim of introducing the normalizer p^{-1} for nonzero genetic effects is to let the per-SNP contribution vanish and thus the aggregated genetic variation $\mathrm{var}(\boldsymbol{X\beta})$ remains finite (Bulik-Sullivan et al. 2015; Dobriban and Wager 2018). It is also possible to introduce the normalization via SNP data as in Jiang et al. (2016). The following analysis of cross-trait PRS remains the same in both situations, because the normalization will cancel out from the numerator and denominator of genetic correlation estimators. The iid assumption on nonzero genetic effects $\boldsymbol{\alpha}_{(1)}$, $\boldsymbol{\beta}_{(1)}$, and $\boldsymbol{\eta}_{(1)}$ in Condition 2 is introduced for simplicity and can be relaxed. In GWAS context, iid random effect formulation is a popular technique to summarize key global characteristics of widespread small per-SNP genetic contributions (e.g., Yang et al. 2010; Jiang et al. 2016). In practice, however, SNPs in coding, regulatory, and low LD genomic regions may have enriched contributions to heritability. To acknowledge these heterogeneities, it is possible to allow SNP effect sizes to have different magnitudes. For example, by letting $\alpha_i \sim F(0, \sigma_{\alpha_i}^2/p)$ and $\eta_i \sim F(0, \sigma_{\eta_i}^2/p)$ independently for $i = 1, ..., p$, $\varphi_{\alpha\eta}$, h_α^2, and h_η^2 can be redefined as $\varphi_{\alpha\eta} = (\sum_{i=1}^p \alpha_i\eta_i)/\{(\sum_{i=1}^p \alpha_i^2)^{1/2}(\sum_{i=1}^p \eta_i^2)^{1/2}\}$, $h_\alpha^2 = (\sum_{i=1}^p \sigma_\alpha^2/p)/(\sum_{i=1}^p \sigma_\alpha^2/p + \sigma_{\epsilon_\alpha}^2)$, and $h_\eta^2 = (\sum_{i=1}^p \sigma_\eta^2/p)/(\sum_{i=1}^p \sigma_\eta^2/p + \sigma_{\epsilon_\eta}^2)$. Intuitively, the iid assumption takes an average of different effect sizes across a large number of SNPs. Our main conclusions about the asymptotic bias in Section 2.2 remain the same under these more general settings if the parameters such as genetic correlation and heritability are redefined accordingly. For SNP screening studied in Section 3, the performance of cross-trait PRS could be better than expected if the top-ranked SNPs selected by GWAS p-values tend to have larger than average effect sizes, but the pattern across different levels of sparsity will remain unchanged.

2.1.3. PRS-based Genetic Correlation Estimators

For common SNPs, the standard approach in GWAS is marginal screening. That is, the marginal association between the phenotype and single SNP is assessed each at a time, while adjusting for the same set of covariates. Now we introduce the cross-trait PRS and genetic correlation estimators based on GWAS marginal screening. We need the following individual level or summary-level data. As n_1, n_2, and $p \to \infty$, the summary association statistics for $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\beta$ from Discovery GWAS-I & II are given by $\widehat{\boldsymbol{\alpha}} = n_1^{-1}\boldsymbol{X}^T\boldsymbol{y}_\alpha = n_1^{-1}\boldsymbol{X}^T(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha)$ and $\widehat{\boldsymbol{\beta}} = n_2^{-1}\boldsymbol{Z}^T\boldsymbol{y}_\beta = n_2^{-1}\boldsymbol{Z}^T(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta)$. We assume

that the individual-level SNP $\boldsymbol{W}$ and phenotype $\boldsymbol{y}_\eta$ in the Target testing GWAS can be accessed. In addition, h^2_α, h^2_β, and h^2_η are assumed to be estimable by their corresponding either individual-level data (Yang et al. 2010) or summary-level data (Bulik-Sullivan et al. 2015). These SNP heritability estimators generally perform well in practice, see Evans et al. (2018) for a detailed numerical comparison. Theoretically, Jiang et al. (2016) showed that the REML heritability estimator (Yang et al. 2010) is consistent in high-dimensional linear mixed effect model. In summary, besides (n_1, n_2, n_3, p), it is assumed that summary-level data $(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$ and individual-level data $(\boldsymbol{y}_\eta, \boldsymbol{W})$ are available, and consistent estimators of h^2_α, h^2_β, and h^2_η can be obtained in corresponding either GWAS or previous studies of the same traits. We construct cross-trait PRSs $\widehat{\boldsymbol{S}}_\alpha = \sum_{i=1}^p \boldsymbol{w}_i\widehat{a}_i = \boldsymbol{W}\widehat{\boldsymbol{a}} = \boldsymbol{W}_{(1,\alpha)}\widehat{\boldsymbol{a}}_{(1)} + \boldsymbol{W}_{(2,\alpha)}\widehat{\boldsymbol{a}}_{(2)}$ for $\boldsymbol{y}_\alpha$ and $\widehat{\boldsymbol{S}}_{\boldsymbol{\beta}} = \sum_{i=1}^p \boldsymbol{w}_i\widehat{b}_i = \boldsymbol{W}\widehat{\boldsymbol{b}} = \boldsymbol{W}_{(1,\beta)}\widehat{\boldsymbol{b}}_{(1)} + \boldsymbol{W}_{(2,\beta)}\widehat{\boldsymbol{b}}_{(2)}$ for $\boldsymbol{y}_\beta$, where $\widehat{\boldsymbol{a}}^T = (\widehat{a}_1, \cdots, \widehat{a}_{m_\alpha}, \widehat{a}_{m_\alpha+1}, \cdots, \widehat{a}_p) = (\widehat{\boldsymbol{a}}^T_{(1)}, \widehat{\boldsymbol{a}}^T_{(2)})$, in which $\widehat{a}_i = \widehat{\alpha}_i \cdot \mathrm{I}(\mathrm{pval}_{\alpha_i} < c_\alpha)$, $\widehat{\boldsymbol{b}}^T = (\widehat{b}_1, \cdots, \widehat{b}_{m_\beta}, \widehat{b}_{m_\beta+1}, \cdots, \widehat{b}_p) = (\widehat{\boldsymbol{b}}^T_{(1)}, \widehat{\boldsymbol{b}}^T_{(2)})$, in which $\widehat{b}_i = \widehat{\beta}_i \cdot \mathrm{I}(\mathrm{pval}_{\beta_i} < c_\beta)$, pval_{α_i} and pval_{β_i} are p-values correspond to $\widehat{\alpha}_i$ and $\widehat{\beta}_i$, respectively, and c_α and c_β are given thresholds used for SNP screening in order to calculate $\widehat{\boldsymbol{S}}_\alpha$ and $\widehat{\boldsymbol{S}}_\beta$. Moreover, we define $\boldsymbol{W}_{(1,\alpha)} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{m_\alpha}]$, $\boldsymbol{W}_{(2,\alpha)} = [\boldsymbol{w}_{m_\alpha+1}, \cdots, \boldsymbol{w}_p]$, $\boldsymbol{W}_{(1,\beta)} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{m_\beta}]$, $\boldsymbol{W}_{(2,\beta)} = [\boldsymbol{w}_{m_\beta+1}, \cdots, \boldsymbol{w}_p]$, and $\boldsymbol{W} = [\boldsymbol{W}_{(1,\alpha)}, \boldsymbol{W}_{(2,\alpha)}] = [\boldsymbol{W}_{(1,\beta)}, \boldsymbol{W}_{(2,\beta)}]$.

We estimate the genetic correlation between $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\eta$ based on $(\widehat{\boldsymbol{S}}_\alpha, \boldsymbol{y}_\eta)$ and that between $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\beta$ based on $(\widehat{\boldsymbol{S}}_\alpha, \widehat{\boldsymbol{S}}_\beta)$. They represent two common cases in GWAS applications. For $(\widehat{\boldsymbol{S}}_\alpha, \boldsymbol{y}_\eta)$, individual-level data are available for one trait, but not for the other one. It often occurs when the traits are studied in two different GWAS. For $(\widehat{\boldsymbol{S}}_\alpha, \widehat{\boldsymbol{S}}_\beta)$, neither of the two traits has individual-level data. This happens when we have GWAS summary statistics of two traits and estimate their genetic correction on an independent target dataset. We define the PRS-based empirical genetic correlation estimators as follows.

Definition 2 (PRS-based Empirical Genetic Correlation Estimators). Suppose the cross-trait genetic correlation $\varphi_{\alpha\eta}$ between $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\eta$ is estimated by using $(\widehat{\boldsymbol{S}}_\alpha, \boldsymbol{y}_\eta)$, then the PRS-based empirical genetic correlation estimator is $G_{\alpha\eta} = \boldsymbol{y}_\eta^T\widehat{\boldsymbol{S}}_\alpha/(\|\boldsymbol{y}_\eta\| \cdot \|\widehat{\boldsymbol{S}}_\alpha\|) = \{(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta)^T(\boldsymbol{W}_{(1,\alpha)}\widehat{\boldsymbol{a}}_{(1)} + \boldsymbol{W}_{(2,\alpha)}\widehat{\boldsymbol{a}}_{(2)})\}/\{\|\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\| \cdot \|\boldsymbol{W}_{(1,\alpha)}\widehat{\boldsymbol{a}}_{(1)} + \boldsymbol{W}_{(2,\alpha)}\widehat{\boldsymbol{a}}_{(2)}\|\}$. Similarly, suppose the cross-trait genetic correlation $\varphi_{\alpha\beta}$ between $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\beta$ is estimated by $(\widehat{\boldsymbol{S}}_\alpha, \widehat{\boldsymbol{S}}_\beta)$, then the PRS-based empirical genetic correlation estimator is $G_{\alpha\beta} = \widehat{\boldsymbol{S}}_\beta^T\widehat{\boldsymbol{S}}_\alpha/(\|\widehat{\boldsymbol{S}}_\beta\| \cdot \|\widehat{\boldsymbol{S}}_\alpha\|) = \{(\boldsymbol{W}_{(1,\beta)}\widehat{\boldsymbol{b}}_{(1)} + \boldsymbol{W}_{(2,\beta)}\widehat{\boldsymbol{b}}_{(2)})^T(\boldsymbol{W}_{(1,\alpha)}\widehat{\boldsymbol{a}}_{(1)} + \boldsymbol{W}_{(2,\alpha)}\widehat{\boldsymbol{a}}_{(2)})\}/\{\|\boldsymbol{W}_{(1,\beta)}\widehat{\boldsymbol{b}}_{(1)} + \boldsymbol{W}_{(2,\beta)}\widehat{\boldsymbol{b}}_{(2)}\| \cdot \|\boldsymbol{W}_{(1,\alpha)}\widehat{\boldsymbol{a}}_{(1)} + \boldsymbol{W}_{(2,\alpha)}\widehat{\boldsymbol{a}}_{(2)}\|\}$.

Here, $G_{\alpha\eta}$ is defined by the PRS of $\boldsymbol{y}_\alpha$ and measured values of $\boldsymbol{y}_\eta$, whereas $G_{\alpha\beta}$ is defined by the two PRS of $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\beta$. Both of the two estimators can be viewed as empirical values subjected to the set of selected SNPs in PRS construction. Therefore, they have different interpretations from the genetic correlation estimators from other models, such as the genome-wide estimator in cross-trait LDSC (Bulik-Sullivan et al. 2015). Furthermore, $G_{\alpha\eta}$ and $G_{\alpha\beta}$ may not be consistent estimators for the underlying population-level parameters $\varphi_{\alpha\eta}$ and $\varphi_{\alpha\beta}$. We quantify their relationships in the following sections.

2.2. Asymptotic Bias and Correction

We first investigate $G_{\alpha\beta}$ and $G_{\alpha\eta}$ when all of the p candidate SNPs are used, that is, we set $c_\alpha = c_\beta = 0$. Thus, $\widehat{\boldsymbol{a}}_{(1)} = \widehat{\boldsymbol{\alpha}}_{(1)}$, $\widehat{\boldsymbol{a}}_{(2)} = \widehat{\boldsymbol{\alpha}}_{(2)}$, $\widehat{\boldsymbol{b}}_{(1)} = \widehat{\boldsymbol{\beta}}_{(1)}$, and $\widehat{\boldsymbol{b}}_{(2)} = \widehat{\boldsymbol{\beta}}_{(2)}$. We have the following results on the asymptotic properties of $G_{\alpha\eta}$, whose proof can be found in the supplementary file.

Theorem 1. Under polygenic model (1) and Conditions 1 and 2, suppose $\min(m_{\alpha\eta}, m_\alpha, m_\eta) \to \infty$ as $\min(n_1, n_3, p) \to \infty$, and let $p = c \cdot (n_1 n_3)^a$ for some constants $c > 0$ and $a \in (0, \infty]$. If $a \in (0, 1)$, then we have

$$G_{\alpha\eta} = \varphi_{\alpha\eta} + \left(\sqrt{\frac{n_1}{n_1 + p/h^2_\alpha}} \cdot h_\eta - 1\right) \cdot \varphi_{\alpha\eta} + o_p(1).$$

If $a \in [1, \infty]$, then we have $G_{\alpha\eta} \cdot n_3 = O_p(1)$.

An important implication of Theorem 1 is that the asymptotic limit of $G_{\alpha\eta}$ is independent of the unknown numbers m_α, m_η, and $m_{\alpha\eta}$, as well as parameters of genetic effects in Condition 2. In real GWAS, the number of independent common SNPs p can be hundreds of thousands or even more than one million. We allow a to vary widely and depend on GWAS sample sizes n_1 and n_3 such that p and $(n_1 n_3)^a$ are proportional, with a small bounded constant c. For example, suppose $p = 500{,}000$, if $n_1 = 500{,}000$ and $n_3 = 1000$ (large GWAS), then a can be 0.65 with $c = 1.1$; if $n_1 = 1000$ and $n_3 = 200$ (small GWAS), then a can be 1.1 and $c = 0.74$.

If $a \in [1, \infty]$, that is, $n_1 n_3$ is too small compared to p, then $G_{\alpha\eta}$ will have a zero asymptotic limit. In practice, this occurs when the sample size of discovery GWAS is too small to obtain reliable GWAS summary statistics. When these summary statistics are applied on an independent target dataset, the mean of genetic covariance $\boldsymbol{y}_\eta^T\widehat{\boldsymbol{S}}_\alpha$ cannot dominate its standard error. The genetic variance $\widehat{\boldsymbol{S}}_\alpha^T\widehat{\boldsymbol{S}}_\alpha$ is so overwhelming such that $G_{\alpha\eta}$ goes to zero. Details can be found in the supplementary file. If $a \in (0, 1)$, $G_{\alpha\eta}$ is a biased estimator of $\varphi_{\alpha\eta}$ when $\sqrt{n_1/(n_1 + p/h^2_\alpha)} \cdot h_\eta$ is smaller than 1. Formally, let $p = c \cdot n_1^b$ for some constants $c > 0$ and $b \in (0, \infty]$, we have

$$\sqrt{\frac{n_1}{n_1 + p/h^2_\alpha}} \cdot h_\eta = \begin{cases} o_p(1), & \text{if} \quad b > 1; \\ \{h^2_\eta/(1 + c/h^2_\alpha)\}^{1/2}, & \text{if} \quad b = 1; \\ h_\eta, & \text{if} \quad b < 1. \end{cases}$$

It follows that $G_{\alpha\eta}$ is an unbiased estimator of $\varphi_{\alpha\eta}$ only if $h^2_\eta = 1$ and $p = o(n_1)$. For $p = O(n_1)$, $G_{\alpha\eta}$ is a shrinkage estimate of $\varphi_{\alpha\eta}$; and when $n_1 = o(p)$, $G_{\alpha\eta}$ is asymptotically zero. Therefore, $G_{\alpha\eta}$ has nonzero asymptotic limit only when training GWAS sample size n_1 is at least proportional to p (i.e., $0 < b \le 1$). In such situation, a consistent estimator of $\varphi_{\alpha\eta}$ can be given as follows.

Consistent estimator of $\varphi_{\alpha\eta}$. Under the same conditions as in Theorem 1, if $a \in (0, 1)$ and $b \in (0, 1]$, then $G^A_{\alpha\eta} = G_{\alpha\eta} \cdot \sqrt{(n_1 + p/h^2_\alpha)/(n_1 \cdot h^2_\eta)} = \varphi_{\alpha\eta} + o_p(1)$ is a consistent estimator

of $\varphi_{\alpha\eta}$. The variance of $G_{\alpha\eta}$ and $G^A_{\alpha\eta}$ is provided in the following corollary.

Corollary 1. Under polygenic model (1) and Conditions 1 and 2, suppose $\min(m_{\alpha\eta}, m_\alpha, m_\eta) \to \infty$ as $\min(n_1, n_3, p) \to \infty$, and let $p = c \cdot n_1^b$ for some constants $c > 0$ and $b \in (0, 1]$, we have

$$(G_{\alpha\eta}) = \left\{ \frac{(p + 2n_1 + 2n_3)h_\eta^2}{n_3(p/h_\alpha^2 + n_1)} \cdot \varphi_{\alpha\eta}^2 + \frac{n_1 h_\eta^2}{p/h_\alpha^2 + n_1} \cdot \frac{m_{\alpha\eta}(\sigma_{\alpha^2\eta^2} - \sigma_{\alpha\eta}^2)}{m_\alpha m_\eta \sigma_\alpha^2 \sigma_\eta^2} \right\} \cdot \{1 + o_p(1)\},$$

where $\mathrm{E}(\alpha_1^2\eta_1^2) = \sigma_{\alpha^2\eta^2}/p^2$. It follows that $\mathrm{var}(G_{\alpha\eta}) = O_p\{(n_1 + n_3)/(n_3 n_1) + m_{\alpha\eta}/(m_\alpha m_\eta)\} = O_p\{\max(n_1^{-1}, n_3^{-1}, m_{\alpha\eta}^{-1})\}$.

As the discovery GWAS sample size n_1 is often large, we usually have $n_1 > n_3$ in practice. Thus, Corollary 1 shows that the scale of $\mathrm{var}(G_{\alpha\eta})$ is jointly determined by the testing GWAS sample size n_3 and the polygenicity of genetics co-architecture of the two traits, characterized by $m_{\alpha\eta}$. When $m_{\alpha\eta} \geq n_3$, $(G_{\alpha\eta})$ has a scale $O_p(1/n_3)$ and thus the inference of $G_{\alpha\eta}$ can be valid in the testing GWAS even if $G_{\alpha\eta}$ is heavily biased toward zero. For example, if $m_{\alpha\eta} \geq n_3$, the T score for testing $H_0 : \varphi_{\alpha\eta} = 0$ versus $H_1 : \varphi_{\alpha\eta} \neq 0$ is given by $T^2_{\alpha\eta} = G^2_{\alpha\eta}/(G_{\alpha\eta}) = \{(p + 2n_1 + 2n_3)/(n_1 n_3) + (\sigma_{\alpha^2\eta^2} - \sigma^2_{\alpha\eta})/(m_{\alpha\eta}\sigma^2_{\alpha\eta})\}^{-1} \cdot \{1 + o_p(1)\} = O_p(n_3^{-1})$, under H_0. On the other hand, if $m_{\alpha\eta} < n_3$, cross-trait PRS may have large variance with scale $O_p(1/m_{\alpha\eta})$. Notably, the testing power of $G^A_{\alpha\eta}$ and $G_{\alpha\eta}$ is the same under the conditions of Corollary 1, because $G^A_{\alpha\eta}$ can be viewed as $G_{\alpha\eta}$ multiplies some constant.

In summary, estimating genetic correlation with cross-trait PRS requires the training GWAS sample size n_1 is at least proportional to p. The testing sample size n_3 vanishes in the limit of $G_{\alpha\eta}$, which verifies that we can apply the discovery summary statistics onto a much smaller set of target samples. In addition, the variance of cross-trait PRS has scale $O_p(1/n_3)$ for a pair of traits with high polygenicity (i.e., $m_{\alpha\eta} \geq n_3$). Therefore, cross-trait PRS may have good testing power even the estimation is biased. This result matches widespread empirical observations that cross-trait PRS may have small p-value, but the R^2 is small. The asymptotic properties of $G_{\alpha\beta}$ are given as follows.

Theorem 2. Under polygenic model (1) and Conditions 1 and 2, suppose $\min(m_{\alpha\beta}, m_\alpha, m_\beta) \to \infty$ as $\min(n_1, n_2, n_3, p) \to \infty$, and let $p^2 = c \cdot (n_1 n_2 n_3)^a$ for some constants $c > 0$ and $a \in (0, \infty]$. If $a \in (0, 1)$, then we have

$$G_{\alpha\beta} = \varphi_{\alpha\beta} + \left(\sqrt{\frac{n_1}{n_1 + p/h_\alpha^2} \cdot \frac{n_2}{n_2 + p/h_\beta^2}} - 1 \right) \cdot \varphi_{\alpha\beta} + o_p(1).$$

If $a \in [1, \infty]$, then we have $G_{\alpha\beta} \cdot \{n_3(n_1 + p)(n_2 + p)\}/p^2 = O_p(1)$.

If $a \in (0, 1)$, $G_{\alpha\beta}$ is an unbiased estimator of $\varphi_{\alpha\beta}$ for $p = o\{\min(n_1, n_2)\}$. When $p = O(n_1) = O(n_2)$, $\sqrt{n_1/(n_1 + p/h_\alpha^2) \cdot n_2/(n_2 + p/h_\beta^2)}$ is smaller than 1, and thus $G_{\alpha\beta}$ is biased toward zero. Further if $\min(n_1, n_2) = o(p)$, then $G_{\alpha\beta}$ is asymptotically zero. Therefore, to have nonzero asymptotic limit, both of the two sets of summary statistics need to be trained from large-scale GWAS. Giving that n_1, n_2, and p are proportional, the scale of $(G_{\alpha\beta})$ is $(G_{\alpha\beta}) = O_p\{n_3^{-1} + m_{\alpha\beta}/(m_\alpha m_\beta)\} = O_p\{\max(n_3^{-1}, m_{\alpha\beta}^{-1})\}$. A consistent estimator of $\varphi_{\alpha\beta}$ is given as follows.

Consistent estimator of $\varphi_{\alpha\beta}$. Under the same conditions as in Theorem 2, if $a \in (0, 1)$ and n_1, n_2, and p are proportional, then $G^A_{\alpha\beta} = G_{\alpha\beta} \cdot \sqrt{(n_1 + p/h_\alpha^2) \cdot (n_2 + p/h_\beta^2)/(n_1 n_2)} = \varphi_{\alpha\beta} + o_p(1)$ is a consistent estimator of $\varphi_{\alpha\beta}$.

Now we propose and study a novel estimator of $\varphi_{\alpha\beta}$ that can be directly constructed by using two sets of summary statistics $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\beta}}$. Let $\widehat{\varphi}_{\alpha\beta} = \widehat{\boldsymbol{\alpha}}^T\widehat{\boldsymbol{\beta}}/(\|\widehat{\boldsymbol{\alpha}}\| \cdot \|\widehat{\boldsymbol{\beta}}\|) = \{(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha)^T\boldsymbol{X}\boldsymbol{Z}^T(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta)\}/\{\|(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha)^T\boldsymbol{X}\| \cdot \|(\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}_\beta)^T\boldsymbol{Z}\|\}$, we have the following asymptotic properties.

Theorem 3. Under polygenic model (1) and Conditions 1 and 2, suppose $\min(m_{\alpha\beta}, m_\alpha, m_\beta) \to \infty$ as $\min(n_1, n_2, p) \to \infty$, and let $p = c \cdot (n_1 n_2)^a$ for some constants $c > 0$ and $a \in (0, \infty]$. If $a \in (0, 1)$, then we have

$$\widehat{\varphi}_{\alpha\beta} = \varphi_{\alpha\beta} + \left(\sqrt{\frac{n_1}{n_1 + p/h_\alpha^2} \cdot \frac{n_2}{n_2 + p/h_\beta^2}} - 1 \right) \cdot \varphi_{\alpha\beta} + o_p(1).$$

If $a \in [1, \infty]$, then we have $\widehat{\varphi}_{\alpha\beta} \cdot \{(n_1 + p)(n_2 + p)\}/p = O_p(1)$.

The $\widehat{\varphi}_{\alpha\beta}$ is interesting in its own right because it quantifies the potential bias of the inner product of marginal screening estimates in high-dimensions. When n_1, n_2, and p are proportional, $(\widehat{\varphi}_{\alpha\beta}) = O_p\{\max(n_1^{-1}, m_{\alpha\beta}^{-1})\}$ and a consistent estimator of $\varphi_{\alpha\beta}$ can be obtained after bias correction.

Consistent estimator of $\varphi_{\alpha\beta}$. Under the same conditions as in Theorem 3, if $a \in (0, 1)$ and n_1, n_2, and p are proportional, then $\widehat{\varphi}^A_{\alpha\beta} = \widehat{\varphi}_{\alpha\beta} \cdot \sqrt{\{(n_1 + p/h_\alpha^2) \cdot (n_2 + p/h_\beta^2)\}/(n_1 n_2)} = \varphi_{\alpha\beta} + o_p(1)$ is a consistent estimator of $\varphi_{\alpha\beta}$. Since $\widehat{\varphi}_{\alpha\beta}$ and $G_{\alpha\beta}$ have similar asymptotic properties, in what follows we will focus on $G_{\alpha\beta}$ and the general conclusions of $G_{\alpha\beta}$ remain the same for $\widehat{\varphi}_{\alpha\beta}$.

3. SNP Screening

As shown in Theorems 1 and 2, in addition to heritability, the asymptotic limit of $G_{\alpha\eta}$ or $G_{\alpha\beta}$ is largely affected by n/p. These results intuitively suggest to select a subset of p SNPs to construct cross-trait PRS. The common approach in practice is to screen the SNPs according to their GWAS p-values. We investigate this strategy in this section.

For a given threshold $c_\alpha > 0$, let $q_\alpha = p \cdot \pi_\alpha = q_{\alpha 1} + q_{\alpha 2}$ ($\pi_\alpha \in (0, 1]$) be the number of top-ranked SNPs selected for $\boldsymbol{y}_\alpha$, among which there are $q_{\alpha 1}$ true causal SNPs and the remaining $q_{\alpha 2}$ are null SNPs, and we let $q_{\alpha\eta}$ be the number of overlapping causal SNPs of $\boldsymbol{y}_\alpha$ and $\boldsymbol{y}_\eta$, and thus $q_{\alpha 1} \geq q_{\alpha\eta}$. The SNP data are defined accordingly. We write $\boldsymbol{X}_{(1)} = [\boldsymbol{X}_{(11)}, \boldsymbol{X}_{(12)}]$, $\boldsymbol{X}_{(2)} = [\boldsymbol{X}_{(21)}, \boldsymbol{X}_{(22)}]$, $\boldsymbol{W}_{(1,\alpha)} = [\boldsymbol{W}_{(11,\alpha)}, \boldsymbol{W}_{(12,\alpha)}]$, and $\boldsymbol{W}_{(2,\alpha)} = [\boldsymbol{W}_{(21,\alpha)}, \boldsymbol{W}_{(22,\alpha)}]$. Here $\boldsymbol{X}_{(11)}$ and $\boldsymbol{W}_{(11,\alpha)}$ are the selected $q_{\alpha 1}$ causal SNPs of $\boldsymbol{y}_\alpha$, and similarly, $\boldsymbol{X}_{(21)}$ and $\boldsymbol{W}_{(21,\alpha)}$ are the selected $q_{\alpha 2}$ null SNPs of $\boldsymbol{y}_\alpha$. In addition, we let $\widehat{\boldsymbol{\alpha}}^T_{(1)} = [\widehat{\boldsymbol{\alpha}}^T_{(11)}, \widehat{\boldsymbol{\alpha}}^T_{(12)}]$ and $\widehat{\boldsymbol{\alpha}}^T_{(2)} = [\widehat{\boldsymbol{\alpha}}^T_{(21)}, \widehat{\boldsymbol{\alpha}}^T_{(22)}]$, where

$\widehat{\boldsymbol{\alpha}}_{(11)}$ corresponds to the selected causal SNPs of $\boldsymbol{y}_\alpha$ and $\widehat{\boldsymbol{\alpha}}_{(21)}$ corresponds to the selected null ones. Then we have $G_{T\alpha\eta} = \{(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta)^T(\boldsymbol{W}_{(11,\alpha)}\widehat{\boldsymbol{\alpha}}_{(11)} + \boldsymbol{W}_{(21,\alpha)}\widehat{\boldsymbol{\alpha}}_{(21)})\}/\{\|\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\| \cdot \|\boldsymbol{W}_{(11,\alpha)}\widehat{\boldsymbol{\alpha}}_{(11)} + \boldsymbol{W}_{(21,\alpha)}\widehat{\boldsymbol{\alpha}}_{(21)}\|\} = C_{T\alpha\eta}/(V_\eta \cdot V_{T\alpha})$, where $V_\eta = \|\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta\|$, $V_{T\alpha} = \|\boldsymbol{W}_{(11,\alpha)}\boldsymbol{X}_{(11)}^T(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha) + \boldsymbol{W}_{(21,\alpha)}\boldsymbol{X}_{(21)}^T(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha)\|$, and $C_{T\alpha\eta} = (\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)} + \boldsymbol{\epsilon}_\eta)^T\boldsymbol{W}_{(11,\alpha)}\boldsymbol{X}_{(11)}^T(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)}+\boldsymbol{\epsilon}_\alpha)+(\boldsymbol{W}_{(1)}\boldsymbol{\eta}_{(1)}+\boldsymbol{\epsilon}_\eta)^T\boldsymbol{W}_{(21,\alpha)}\boldsymbol{X}_{(21)}^T(\boldsymbol{X}_{(1)}\boldsymbol{\alpha}_{(1)} + \boldsymbol{\epsilon}_\alpha)$.

Corollary 2. Under polygenic model (1) and Conditions 1 and 2, suppose that $\min(m_{\alpha\eta}, m_\alpha, m_\eta) \to \infty$ and $\min(q_{\alpha\eta}, q_{\alpha 1}, q_{\alpha 2}) \to \infty$ as $\min(n_1, n_3, p) \to \infty$, further if $\left\{m_{\alpha\eta}^2(q_{\alpha 1}+q_{\alpha 2})\right\}/(q_{\alpha\eta}^2 n_1 n_3) \to 0$, then we have

$$G_{T\alpha\eta} = \varphi_{\alpha\eta} + \left(\sqrt{\frac{n_1 m_\alpha}{n_1 q_{\alpha 1} + m_\alpha q_\alpha / h_\alpha^2}} \cdot \frac{q_{\alpha\eta}}{m_{\alpha\eta}} \cdot h_\eta - 1\right) \cdot \varphi_{\alpha\eta} + o_p(1).$$

Corollary 2 shows the tradeoff of SNP screening. Given n_1, m_α, $m_{\alpha\eta}$, h_α, and h_η, the potential bias of $G_{T\alpha\eta}$ is also affected by q_α, $q_{\alpha 1}$ and $q_{\alpha\eta}$. As more SNPs are selected, the numerator of $\sqrt{(n_1 m_\alpha)/(n_1 q_{\alpha 1} + m_\alpha q_\alpha/h_\alpha^2)} \cdot (q_{\alpha\eta}/m_{\alpha\eta})$ increases with $q_{\alpha\eta}$, while the denominator increases with $\sqrt{q_\alpha}$ (and $\sqrt{q_{\alpha 1}}$). Therefore, whether or not SNP screening can improve the estimation is largely affected by the quality of the selected SNPs, which is highly related to the m_α/n_1 ratio. In the optimistic case where $q_{\alpha\eta} = m_{\alpha\eta}$ and $q_\alpha = q_{\alpha 1} = m_\alpha$, $G_{T\alpha\eta}$ becomes $\sqrt{n_1/(n_1 + m_\alpha/h_\alpha^2)} \cdot h_\eta \cdot \varphi_{\alpha\eta}$, which is the theoretical upper limit. We note that this optimistic upper limit can still be biased toward zero. An opposite case is that the GWAS summary statistics of causal and null SNPs are totally mixed up, which may occur when m_α/n_1 is large (i.e., sample size is relatively small or trait is highly polygenic). Therefore, we have $q_{\alpha 1}/q_\alpha \approx m_\alpha/p$. Suppose also $q_{\alpha\eta}/q_{\alpha 1} \approx m_{\alpha\eta}/m_\alpha$, we have $G_{T\alpha\eta} \approx \sqrt{n_1/(n_1 p + p^2/h_\alpha^2) \cdot q_\alpha} \cdot h_\eta \cdot \varphi_{\alpha\eta}$, which increases with q_α. As $q_\alpha = p$, $G_{T\alpha\eta}$ reaches its upper bound $\sqrt{n_1/(n_1 + p/h_\alpha^2)} \cdot h_\eta \cdot \varphi_{\alpha\eta}$. That is, $G_{T\alpha\eta}$ achieves the best performance when the cross-trait PRS is constructed without SNP screening. For example, in the left panel of Figure 2, we set $m_\alpha/n_1 = 0.01$ to reflect the sparse signal case, in which causal and null SNPs can be easily separated by SNP screening. Thus, SNP screening can reduce the bias of $G_{\alpha\eta}$ when signals are sparse. However, as the number of causal SNPs increase (from left to right in Figure 2), it becomes much hard to separate causal and null SNPs by their GWAS p-values. Therefore, SNP screening will enlarge the bias.

In conclusion, when causal and null SNPs can be easily separated by GWAS, the top-ranked SNPs are more likely to be causal ones, that is, SNP screening helps. However, for highly polygenic complex traits whose m_α/n_1 is large, SNP screening may result in larger bias. Moreover, since different underlying m_α/n_1 ratio will result in different patterns as shown in Figure 2, the observed pattern can be used to infer the m_α/n_1 ratio (i.e., the degree of polygenicity) and minimize the potential bias in estimation. We display this strategy in Section 6. The $G_{\alpha\beta}$ has similar properties when performing SNP screening, whose results can be found in the supplementary file.

4. Overlapping Samples

In practice, different GWAS may share a subset of participants. It is often inconvenient to recalculate the GWAS summary statistics after removing the overlapping samples. In this section, we examine the effect of overlapping samples on the bias of cross-trait PRS, which provides more insights into the bias phenomenon of cross-trait PRS. Particularly, we focus on one case which is common in practice: n_s overlapping samples between discovery GWAS and Target testing data for $\varphi_{\alpha\eta}$ estimation. We add n_s overlapping samples into Discovery GWAS-I and Target testing GWAS, resulting in the following two new datasets: (i) Dataset IV: $(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{y}_\alpha)$, with $\boldsymbol{X} \in \mathbb{R}^{n_1 \times p}$, $\boldsymbol{S} \in \mathbb{R}^{n_s \times p}$, and $\boldsymbol{y}_\alpha^T = (\boldsymbol{y}_{\alpha X}^T, \boldsymbol{y}_{\alpha S}^T) \in \mathbb{R}^{(n_1+n_s)\times 1}$; and ii) Dataset V: $(\boldsymbol{W}, \boldsymbol{S}, \boldsymbol{y}_\eta)$, with $\boldsymbol{W} \in \mathbb{R}^{n_3 \times p}$, $\boldsymbol{S} \in \mathbb{R}^{n_s \times p}$, and $\boldsymbol{y}_\eta^T = (\boldsymbol{y}_{\eta W}^T, \boldsymbol{y}_{\eta S}^T) \in \mathbb{R}^{(n_3+n_s)\times 1}$. Mimicking h^2, we define $h_{\alpha\eta} \in (0, 1]$ as the proportion of phenotypic correlation that can be explained by the correlation of their genetic components as $h_{\alpha\eta} = (m_{\alpha\eta}/p)\sigma_{\alpha\eta}/\{(m_{\alpha\eta}/p)\sigma_{\alpha\eta} + \sigma_{\epsilon_\alpha\epsilon_\eta}\}$. On the overlapping samples, we allow nonzero correlation between random errors to capture the non genetic contribution to phenotypic correlation. We introduce an additional condition on random errors.

Condition 3. On n_s overlapping samples, ϵ_{α_j} and ϵ_{η_j} are independent random variables satisfying

$$\begin{pmatrix}\epsilon_{\alpha_j}\\ \epsilon_{\eta_j}\end{pmatrix} \sim F\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_{\epsilon_\alpha}^2 & \sigma_{\epsilon_\alpha\epsilon_\eta}\\ \sigma_{\epsilon_\alpha\epsilon_\eta} & \sigma_{\epsilon_\eta}^2\end{pmatrix}\right]$$

for $j = 1, ..., n_s$, where $\sigma_{\epsilon_\alpha\epsilon_\eta} = \rho_{\epsilon_\alpha\epsilon_\eta} \cdot \sigma_{\epsilon_\alpha}\sigma_{\epsilon_\eta}$.

Theorem 4. Under polygenic model (1) and Conditions 1–3, suppose $\min(m_{\alpha\eta}, m_\alpha, m_\eta) \to \infty$ as $\min\{(n_1 + n_s), (n_3 + n_s), p\} \to \infty$, and let $p = c \cdot \{(n_1 + n_s)(n_3 + n_s)\}^a$ for some constants $c > 0$ and $a \in (0, \infty]$. If $a \in (0, 1)$, then $G_{S\alpha\eta}$ can be written as

$$\frac{\left[1 + n_s p/\{(n_1 + n_s)(n_3 + n_s) \cdot h_{\alpha\eta}\}\right] \cdot \left[h_\eta \cdot \varphi_{\alpha\eta} \cdot \{1 + o_p(1)\}\right]}{\left[1 + p/\{(n_1 + n_s) \cdot h_\alpha^2\} + 2n_s p/\{(n_1 + n_s)(n_3 + n_s)\} + n_s p^2/\{(n_1 + n_s)^2(n_3 + n_s) \cdot h_\alpha^2\}\right]^{1/2}}.$$

If $a \in [1, \infty]$, then we have $G_{S\alpha\eta} = o_p(1)$.

Theorem 4 shows the effect of n_s overlapping samples on the estimation of $\varphi_{\alpha\eta}$. Both sample sizes $(n_1 + n_s)$ and $(n_3 + n_s)$ are involved in the limit. An interesting special case is when the two GWAS are fully overlapped, then we have $G_{S\alpha\eta} = (n_s + p/h_{\alpha\eta})/\{n_s^2 + 2n_s p + p(p + n_s)/h_\alpha^2\}^{1/2} \cdot h_\eta \cdot \varphi_{\alpha\eta} + o_p(1)$. In the optimal situation where $h_\alpha^2 = h_\eta^2 = h_{\alpha\eta} = 1$, we have $G_{S\alpha\eta} = \{1 + (p/n_s + n_s/p + 2)^{-1}\}^{-1/2} \cdot \varphi_{\alpha\eta} + o_p(1)$. Therefore, $G_{S\alpha\eta}$ is asymptotically biased unless either $p = o(n_s)$ or $n_s = o(p)$ holds, neither of which is the case in modern GWAS. As n_s and p are more comparable, the asymptotic bias in $G_{S\alpha\eta}$ increases and the largest bias occurs as $p = n_s \to \infty$.

Note that it is not recommended to estimate the genetic correlation between two traits with (fully) overlapping samples due to concerns such as confounding and overfitting (Dudbridge 2013). In our analysis, such concern is quantified by the value of $h_{\alpha\eta}$. That is, when non-genetic correlation exists in error terms, we have $h_{\alpha\eta} < 1$, and the estimation of genetic correlation is inflated. However, on the other hand, our results show that even

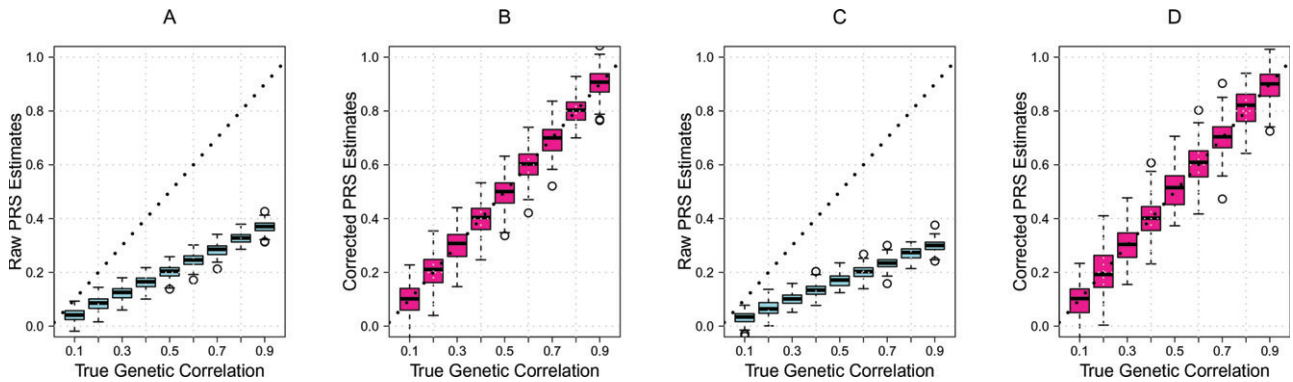

Figure 1. Raw genetic correlations estimated by cross-trait PRS with all SNPs (A: $G_{\alpha\eta}$, C: $G_{\alpha\beta}$) and the bias-corrected genetic correlation estimates (B: $G^A_{\alpha\eta}$, D: $G^A_{\alpha\beta}$). We set $h^2_\alpha = h^2_\beta = h^2_\eta = 0.5$, $n_1 = n_2 = p = 10,000$, and $n_3 = m = 2000$.

in an optimal overlapping setting with $h^2_\alpha = h^2_\eta = h_{\alpha\eta} = 1$, the cross-trait PRS estimator based on GWAS summary statistics can be biased toward zero.

In the supplementary file, we provide a consistent estimator of $\varphi_{\alpha\eta}$ given overlapping samples and further investigate several other specific overlapping cases, which can be useful for quantifying potential bias and perform correction. In summary, these analyses reveal that the bias in cross-trait PRS estimator may result from the following facts: (i) summary statistics are generated from independent GWAS, where the induced bias is largely determined by the n/p ratio; (ii) phenotypes are not fully heritable, that is, heritability is less than one; and (iii) nongenetic correlation exists in the random errors of overlapping samples. This may happen, for example, when confounding effects are not fully adjusted. The first two facts may bias the genetic correlation estimator toward zero, while the last fact may inflate the estimated genetic correlation.

5. Numerical Experiments

5.1. Cross-Trait PRS With All SNPs

To illustrate the finite sample performance of our theoretical results, we simulate 10,000 uncorrelated SNPs. The MAF of each SNP, f, is independently generated from Uniform $[0.05, 0.45]$ based on which the SNP genotypes are independently sampled from $\{0, 1, 2\}$ with probabilities $\{(1-f)^2, 2f(1-f), f^2\}$, respectively. The SNPs are then standardized to satisfy Condition 1. We set the same 2000 causal SNPs on each trait and the nonzero genetic effects are generated from Normal distribution according to Condition 2 with $\sigma_\alpha = \sigma_\eta = \sigma_\beta = 1$. We set all heritability to 0.5 and vary $\sigma_{\alpha\eta}$ and $\sigma_{\alpha\beta}$ (and thus asymptotically $\varphi_{\alpha\eta}$ and $\varphi_{\alpha\beta}$) from 0.1 to 0.9. Model (1) is used to generate continuous phenotypes. We generate 10,000 samples in training dataset and 2000 samples in testing dataset. A total of 200 replicates is conducted. Cross-trait PRS is built with all SNPs. We calculate the raw estimators $G_{\alpha\eta}$ and $G_{\alpha\beta}$ studied in Theorems 1 and 2, and the corresponding bias-corrected estimators $G^A_{\alpha\eta}$ and $G^A_{\alpha\beta}$. The performance of $G_{\alpha\eta}$ and $G_{\alpha\beta}$ is displayed in panels A and C of Figure 1. It is clear that these raw estimates are biased toward zero. For example, when $\sigma_{\alpha\eta} = \sigma_{\alpha\beta} = 0.9$, $G_{\alpha\eta}$ is around 0.37 while $G_{\alpha\beta}$ is about 0.30. The performance of $G^A_{\alpha\eta}$ and $G^A_{\alpha\beta}$ is displayed in panels B and D of Figure 1, which indicates that the two bias-corrected estimators perform well and are close to the true value of $\sigma_{\alpha\eta}$ and $\sigma_{\alpha\beta}$, respectively.

Next, we generate mixed samples from five sub populations. Specifically, the overall MAF of each SNP is simulated from Uniform $[0.05, 0.45]$ and the F_{st} values are independently generated from Uniform $[0.01, 0.04]$. Then the MAF in each sub-population is obtained according to the Balding-Nichols model (Balding and Nichols 1995). The sample size is 2000 in training data and 200 in testing data for each sub population. We perform principal component analysis on all SNPs and use the top four principal components (PCs) to adjust for population substructures. Supplementary material (Figure 1) displays the performance of the raw and bias-corrected estimators, which suggests that our theoretical results remain unchanged after adjusting population substructures by genetic PCs. We also perform power analysis on $G_{\alpha\eta}$ and $G_{\alpha\beta}$. We set $\sigma_{\alpha\eta}$ and $\sigma_{\alpha\beta}$ to be 0, 0.3, or 0.4, simulate 500 testing samples, and vary the number of causal SNPs from 500 to 8000. Other settings remain the same as in Figure 1. A total of 500 replicates is conducted. Although $G_{\alpha\eta}$ and $G_{\alpha\beta}$ are biased toward zero, we find that they can still have good power to detect nonzero genetic correlations (supplementary material, Table 1). When the number of causal SNPs is relatively small (500), the Type I error is slightly inflated, which may due to the larger variance of $G_{\alpha\eta}$ and $G_{\alpha\beta}$ in this situation.

To verify that our bias-corrected estimators are independent of the signal sparsity, we set $m_\alpha = m_\beta = m_\eta = p \cdot a_\alpha$ and vary the sparsity $a_\alpha = 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 0.6, 0.7$, and 0.8 to generate sparse and dense signals. Next, we fix $a_\alpha = 0.2$ and set $m_\beta = m_\eta = k \cdot m_\alpha$ to allow phenotypes to have different number of causal SNPs, where $k = 0.3, 0.4, 0.5, 0.8, 1, 1.25, 2, 2.5$, and 3.3. We set all heritability to 0.5 and let $\sigma_{\alpha\eta} = \sigma_{\alpha\beta} = 0.5$. Sample size of training and testing datasets is set to 10,000 and 2000, respectively. The performance of $G_{\alpha\eta}$ is displayed in the left two panels of Figure 2 (supplementary material). The bias of $G_{\alpha\eta}$ is independent of either the sparsity a_α of a trait or the ratio of sparsity k between two traits, which verifies our results of Theorem 1. The right two panels of Figure 2 (supplementary material) display the performance of $G^A_{\alpha\eta}$. It is clear that $G^A_{\alpha\eta}$ is unbiased regardless of a_α and k. The performance of $G_{\alpha\beta}$ and $G^A_{\alpha\beta}$ is displayed in the supplementary material (Figure 3) and supports our results in Theorem 2. Finally, we illustrate the performance of $\widehat{\varphi}_{\alpha\beta}$ and $\widehat{\varphi}^A_{\alpha\beta}$ in the supplementary material (Fig-

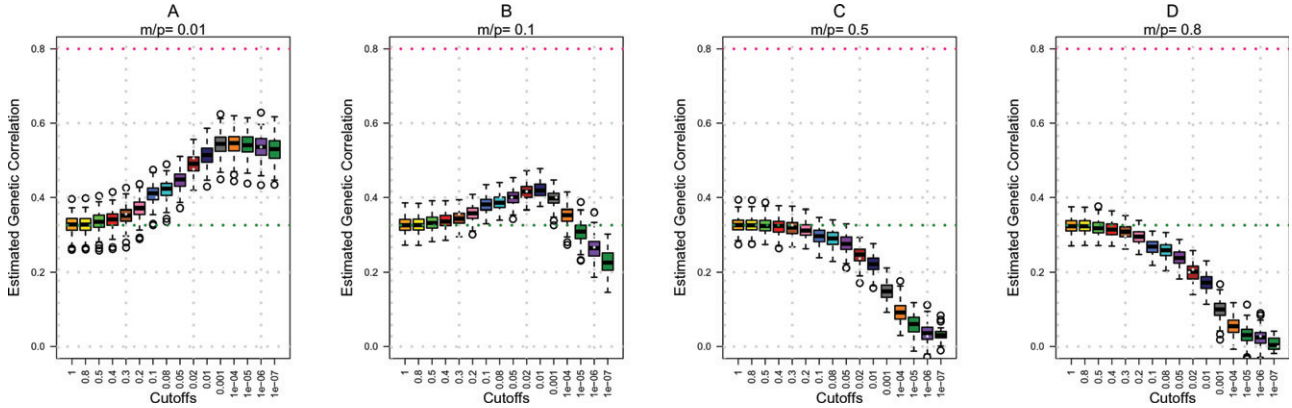

Figure 2. Raw genetic correlation $G_{T\alpha\eta}$ estimated by cross-trait PRS with selected SNPs under different sparsity m/p. We set $h_\alpha^2 = h_\eta^2 = 0.5, \varphi_{\alpha\eta} = 0.8, n_1 = p = 10{,}000$, $n_3 = 2000$, and $m_\alpha/p = m_\eta/p = 0.01, 0.1, 0.5$, and 0.8 (left to right). The red and green dashed lines represent $\varphi_{\alpha\eta}$ and the asymptotic limit of $G_{\alpha\eta}$ according to Theorem 1, respectively.

ure 4), verifying our results in Theorem 3 and the unbiasedness of $\widehat{\varphi}^A_{\alpha\beta}$.

5.2. SNP Screening and Overlapping Samples

Instead of using all the 10,000 SNPs, we construct cross-trait PRS with the top-ranked SNPs whose GWAS p-values pass a pre-specified threshold. We consider a series of thresholds $\{1, 0.8, 0.5, 0.4, 0.3, 0.2, 0.1, 0.08, 0.05, 0.02, 0.01, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$ and generate a series of $G_{T\alpha\eta}$ accordingly. We set heritability to 0.5 and $\varphi_{\alpha\eta} = 0.8$. Four levels of sparsity $m_\alpha/p = m_\eta/p = 0.01, 0.1, 0.5$ and 0.8 are examined. Figure 2 displays the performance of $G_{T\alpha\eta}$ across a series of thresholds. As expected, the pattern of $G_{T\alpha\eta}$ varies dramatically with the sparsity. When signals are sparse, SNP screening helps and $G_{T\alpha\eta}$ outperforms $G_{\alpha\eta}$. However, when signals are dense, the performance of $G_{T\alpha\eta}$ drops as the threshold decreases. The $G_{T\alpha\eta}$ has the best performance when all SNPs are selected, that is, the same as $G_{\alpha\eta}$, which confirms our results of $G_{T\alpha\eta}$ in Corollary 2. In addition, we examine our analyses of overlapping samples. For $G_{S\alpha\eta}$ and $G_{S\alpha\beta}$, half of the 10,000 samples are set to be overlapping. Other settings remain the same as those of Figure 1. The performance of $G_{S\alpha\eta}$, $G_{S\alpha\beta}$, and the corresponding bias-corrected estimators $G^A_{S\alpha\eta}$ and $G^A_{S\alpha\beta}$ (see supplementary file for definitions) is displayed in the supplementary material (Figure 5), which fully supports the results in Theorem 4 and Proposition S4.

5.3. UK Biobank Data Simulation

We perform additional simulation using real SNP data from the UK Biobank (UKB) resources (Bycroft et al. 2018). We download the imputed genotype data and apply the following quality controls (QCs): excluding subjects with more than 10% missing genotypes, only including SNPs with MAF > 0.01, genotyping rate $> 90\%$, and passing Hardy-Weinberg test (p-value $> 1 \times 10^{-7}$). To balance the accuracy and computational burden, we constrain our analysis to 653,122 QC'ed variants that are overlapped with the ones in the HapMap3 reference panel. We randomly select 60,000 unrelated individuals of European (British, Irish and others) ancestry with fine-scale population structures in simulation. Among the 60,000 samples, 50,000 are randomly picked as training data, and the cross-trait PRS is evaluated with the remaining 10,000 testing individuals. We randomly set half of all SNPs to be causal SNPs. The nonzero SNP effects are independently generated from Normal distribution according to Condition 2 with $\sigma_\alpha = \sigma_\eta = 1$, and we set $h_\alpha^2 = h_\eta^2 = \sigma_{\alpha\eta} = 0.6$. We generate cross-trait PRS by summarizing across all relatively independent SNPs after LD-based pruning (window size 50, $R^2 = 0.05$) using the Plink tool set. We use unstandardized SNPs in UKB data simulation and 200 replicates are conducted.

The performance of $G_{\alpha\eta}$ and $G^A_{\alpha\eta}$ is displayed in the left panel of supplementary material (Figure 6). The $G_{\alpha\eta}$ is biased toward zero and the bias-corrected estimator $G^A_{\alpha\eta}$ is close to $\varphi_{\alpha\eta}$. These results illustrate the existence of bias in real SNP data and show the good performance of our bias-corrected estimator. In the right panel of Figure 6 (supplementary material), we further evaluate the sensitivity of $G^A_{\alpha\eta}$ to biased heritability estimates. When the heritability is overestimated (or underestimated), the genetic correlation may be underestimated (or overestimated). The bias of $G^A_{\alpha\eta}$ generally has a similar scale to the bias of heritability estimates. For example, when h_α^2 and h_η^2 are both overestimated by 20% (i.e., $\widehat{h}_\alpha^2 = \widehat{h}_\eta^2 = 0.72$), $\varphi_{\alpha\eta}$ is underestimated by 13% (i.e., $G^A_{\alpha\eta} = 0.52$). To evaluate the performance of cross-trait PRS in the presence of related samples, we rerun our simulation after including related individuals in the training data. Specifically, we replace our training samples with 50,000 European individuals that have relatives in UKB (Bycroft et al. 2018) but are unrelated with the 10,000 testing samples. We find that the performance of $G_{\alpha\eta}$ and $G^A_{\alpha\eta}$ is slightly reduced, indicating the negative influence of sample relatedness on PRS performance.

In PRS construction, we can also directly adjust for the SNP data LD structures with reference panels (e.g., the 1000 Genomes Project LD reference panel), which has shown better performance than LD-based pruning (Mak et al. 2017). We examine one recently the proposed method that incorporates LD using continuous shrinkage priors (PRScs) (Ge et al. 2019). Figure 6 (supplementary material) shows that PRScs estimator has smaller bias than the raw LD-pruned estimator, but still heavily underestimates the underlying genetic correlation.

These results suggest that the bias of cross-trait PRS still exists after taking local LD structures into account. Quantifying and correcting the bias in PRScs and other similar methods can be an interesting future problem.

6. Real Data Analysis

Human brain structural changes are known to be associated with cognitive and mental health traits. It is of great interest to understand the shared genetic influences among these brain-related complex traits. Gray matter volumes of brain region of interest (ROI) (refer to as ROI volumes) are heritable measures of brain structural variation and can be obtained from brain magnetic resonance imaging (MRI). In this section, we apply cross-trait PRS to quantify the genetic similarity between ROI volumes and reaction time, which is a heritable measure of general cognitive functions (Davies et al. 2018). We focus on the volume measures from seven important brain ROIs, including the thalamus proper, caudate, putamen, pallidum, hippocampus, accumbens area, and the total brain volume (TBV). These ROIs are frequently studied in neuroimaging genetics, and common SNPs are able to account for about 50% phenotypic variation in these traits (Biton et al. 2020) (supplementary material, Table 2).

As a positive control, we first estimate the genetic correlation between the TBV phenotype measured in the PING study and the same trait measured in the UKB study. The TBV phenotype in the two studies is generated using consistent standard procedures, and thus the underlying genetic correlation is expected to be close to one. A full description of the PING study, imaging processing, and genotyping data quality controls is documented in the supplementary file. We download the UKB GWAS summary statistics of TBV (Zhao et al. (2019), $n = 19,629$) and construct the PRS on the PING samples ($n = 924$). Specifically, we generate the PRS by summarizing across all the LD-pruned unstandardized SNPs ($R^2 = 0.2$, window size 50), weighed by the UKB GWAS effect sizes. As chromosome strands in UKB and PING data can be different, ambiguous SNPs (i.e., SNP with complementary alleles, either C/G or A/T) are removed in our analysis, after which there are 381,182 overlapped SNPs between UKB and PING. The association between TBV and the constructed PRS is estimated and tested in linear regression, adjusting for the effects of age and sex. The additional phenotypic variation that can be explained by the PRS (i.e., the partial R^2) is interpreted as an estimator of the squared genetic correlation. The estimated genetic correlation is 0.13 (partial $R^2 = 1.77\%$, p-value = 2.6×10^{-6}) in this positive control analysis, which is comparable with reported results for neuroimaging traits and other brain-related complex traits (Bogdan, Baranger, and Agrawal 2018). On the other hand, the bias-corrected genetic correlation is 1.03 according to Theorem 1 ($p = 381,182$, $n = 19,629$, and $h^2 = 0.58$). These results suggest that the raw PRS-based genetic correlation estimator of TBV is heavily biased toward zero, supporting our theoretical results and matches many empirical observations. More importantly, we find that our bias-corrected estimator performs well and can reflect the expected genetic similarity.

Next, cross-trait PRS of reaction time is constructed on these PING samples using the published GWAS summary statistics of reaction time from the largest study so far (Davies et al. (2018), $n = 282,014$). The original GWAS has no overlapping samples with the PING study. After removing ambiguous SNPs, 428, 146 overlapped SNPs are used for PRS construction. We examine the partial R^2 using the same procedure as in the above positive control analysis. The results are summarized in Table 2 (supplementary material). The mean proportion of variation that can be additionally explained by the cross-trait PRS is 1.32% across the seven ROIs. The largest partial R^2 2.80% is found in the thalamus (p-value $=8.74 \times 10^{-9}$), which is known to play integrative roles in cognitive functions (Wolff and Vann 2019). Evidence from imaging studies indicates that the thalamus is associated with reaction time and has predictive power to this cognitive trait (Nikulin et al. 2008). However, though the p-value indicates a significant genetic relationship between thalamus and reaction time, the partial R^2 is small and may underinterpret the genetic similarity of the two traits. Thus, we correct the observed partial R^2 with our formula in Theorem 1. We use the heritability estimate of reaction time reported in Davies et al. (2018) ($h^2 = 0.25$), and the heritability estimates of ROI volumes reported in Biton et al. (2020). The mean partial R^2 across the seven ROIs becomes 22.42% after correction. These findings indicate that the raw PRS-based genetic correlation estimators may substantially underestimate the genetic similarity between reaction time and ROI volumes, whereas our bias-corrected estimator reveals a moderate level of genetic correlation between these brain-related complex traits.

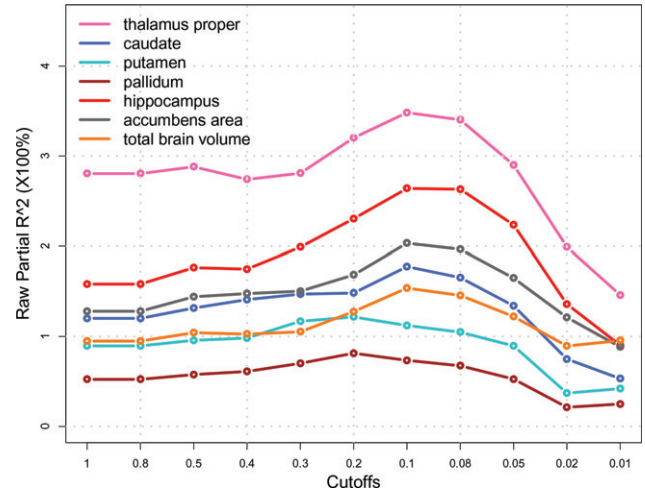

Figure 3. Raw partial R^2 of fitting reaction time PRS on seven regional brain volumes (listed in the figure) in the PING study given different GWAS p-value cutoffs.

To uncover the underlying m/n ratio and minimize the potential bias in the raw partial R^2, we apply SNP screening with multiple GWAS p-value cutoffs and present the trajectory of partial R^2 in Figure 3. The pattern of partial R^2 is similar across the seven ROI volumes, suggesting that these ROI volumes have similar genetic co-architecture with reaction time. The optimal GWAS p-value cutoff for genetic correlation estimation is around 0.1 in this analysis. The partial R^2 of thalamus moves up to 3.48% given this cutoff, which is still much smaller than the bias-corrected estimator. Overall, the trajectory analysis reveals the polygenic genetic co-architecture of reaction time and ROI volumes, indicating that a large number of common genetic variants have small contributions to these traits.

We also run cross-trait LDSC (Bulik-Sullivan et al. 2015) to estimate the genetic correlation between reaction time and the seven ROI volumes. The estimates of LDSC are noisy (supplementary material, Table 3), which may due to the small sample size of the PING study ($n = 924$). These observations make sense, because it is known that cross-trait LDSC may have poor performance when GWAS sample size is small (Ni et al. 2018), for example, smaller than 5000 as mentioned in *https://github.com/bulik/ldsc/wiki/FAQ*.

In summary, we examine the genetic similarity between cognitive function and brain structures using GWAS summary-level data of reaction time and individual-level genotyping and MRI data in the PING study. The raw PRS-based genetic correlation estimators are all small, which may heavily underestimate the genetic overlaps between the two sets of closely related heritable traits. We apply Theorem 1 to generate bias-corrected estimators, which uncover the moderate degree of genetic similarity among these traits. These findings suggest that brain volumetrical variations can serve as important endophenotypes in studying the genetic pathways of human cognitive functions. Finally, we note that methods for genetic correlation estimation based on two sets of GWAS summary statistics, such as cross-trait LDSC, require all the input summary statistics from large-scale GWAS. Thus, one of the main advantages of our bias-corrected cross-trait PRS estimator over cross-trait LDSC is that cross-trait PRS estimator can be applied in small testing GWAS. However, it is also important to note that the genetic correlation estimator in cross-trait LDSC is a genome-wide quantity, while the cross-trait PRS estimator is an empirical value characterizing the genetic correlation attributable to the set of selected SNPs. Although related to each other, these two estimators have completely different definitions and interpretations.

7. Discussion

Understanding the genetic similarity among human complex traits is essential to model biological mechanisms, improve genetic risk prediction, and design personalized prevention/treatment. Cross-trait PRS (Purcell et al. 2009) is one of the most popular methods for genetic correlation estimation with thousands of publications. This paper empirically and theoretically studies the properties of cross-trait PRS in GWAS applications for complex traits. Our analyses demystify the commonly observed small R^2 in GWAS applications, and help avoid over- or under-interpreting of research findings. We demonstrate the importance of our results in a case study using GWAS summary statistics of reaction time and neuroimaging traits measured in the PING study. Our analysis uncovers the moderate degree of genetic correlation between reaction time and brain structures and illustrate the polygenicity of their genetic co-architecture. As more discovery GWAS summary statistics from biobanks become publicly available, our bias-corrected estimators can be used to assess the underlying genetic correlation in many cross-trait PRS applications, especially when the in-house testing GWAS has relatively small sample size.

Our analyses face a few limitations. First, we focus on continuous traits in this article. It is also of great interest to examine binary clinical outcomes/diseases in generalized linear polygenic models. Second, we assume SNPs are independent in our analyses. However, nearby SNPs within the same genomic region (local LD block) may be correlated with each other in real GWAS. Our independence assumption is inspired by the LD-based pruning/clumping step utilized in PRS construction, after which only relatively independent SNPs are kept. We use real SNP data simulation to show that our bias-corrected estimator has good performance after LD-based pruning, but it might be more interesting to develop asymptotic results under a general variance-covariance structure of SNP data in future studies. In addition, the influence of overlapping samples quantified in Section 4 depends on the overlapped sample size n_s, which may be unknown in PRS applications. Thus, the best strategy in practice is to use independent training and testing GWAS samples in cross-trait PRS applications.

Finally, besides the sample size, heritability, and SNP screening analyzed in this study, some other factors may also influence the performance of PRS. For example, the mismatch of the genotyping arrays and imputation platforms in training and testing GWAS may decrease the accuracy of PRS. Moreover, modeling the gene–environment interaction could result in better PRS (Arnau-Soler et al. 2019), and it is well observed that population disparities can introduce additional challenges in PRS applications (Martin et al. 2019). More efforts are needed to explore these problems and improve the performance of PRS.

Acknowledgments

We would like to thank Haiyan Deng for helpful discussion. We thank Tengfei Li and other members of the UNC BIG-S2 lab for processing the raw brain imaging data. We thank the individuals represented in the PING study for their participation and the research team for their work in collecting, processing and disseminating the dataset for analysis. More information of this study can be found in the supplementary material. We thank the Centre for Cognitive Ageing and Cognitive Epidemiology (CCACE, *https://www.ccace.ed.ac.uk/*) for sharing GWAS summary-level data used in the reaction time data analysis. The authors acknowledge the Texas Advanced Computing Center (TACC, *http://www.tacc.utexas.edu/*) at The University of Texas at Austin for providing HPC and storage resources that have contributed to the research results reported within this paper.

Funding

The research is partially supported by NIH grants MH086633 and MH116527.

Supplementary Material

Supplementary Materials: technical proofs, additional simulation results, and details of data analysis using the UKB and PING datasets.

Software codes: Software codes to reproduce Figures 1–3 and Tables 1–3 (supplementary material) have been uploaded with submission, and are also available at *https://github.com/xm1701/paper_figures*. The bias-corrected estimators are implemented in R package bcPRS, which has been uploaded with submission and can be accessed at *https://github.com/xm1701/bcPRS*.

References

Arnau-Soler, A., Macdonald-Dunlop, E., Adams, M. J., Clarke, T.-K., MacIntyre, D. J., Milburn, K., Navrady, L., Hayward, C., McIntosh, A. M., and Thomson, P. A. (2019), "Genome-wide by Environment Interaction Studies of Depressive Symptoms and Psychosocial Stress in UK Biobank and Generation Scotland," *Translational Psychiatry*, 9, 1–13. [10]

Balding, D. J., and Nichols, R. A. (1995), "A Method for Quantifying Differentiation Between Populations at Multi-allelic Loci and Its Implications for Investigating Identity and Paternity," *Genetica*, 96, 3–12. [7]

Biton, A., Traut, N., Poline, J.-B., Aribisala, B. S., Bastin, M. E., Bülow, R., Cox, S. R., Deary, I. J., Fukunaga, M., Grabe, H. J., Hagenaars, S., Hashimoto, S., Muñoz Maniega, R., Nauck, M., Royle, N. A., Teumer, A., Valdes Hernandez, M., Völker, U. (2020), "Polygenic Architecture of Human Neuroanatomical Diversity," *Cerebral Cortex*, 30, 2307–2320. [9]

Bogdan, R., Baranger, D. A., and Agrawal, A. (2018), "Polygenic Risk Scores in Clinical Psychology: Bridging Genomic Risk to Individual Differences," *Annual Review of Clinical Psychology*, 14, 119–157. [2,9]

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017), "An Expanded View of Complex Traits: From Polygenic to Omnigenic," *Cell*, 169, 1177–1186. [1,2]

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. (2015), "An Atlas of Genetic Correlations Across Human Diseases and Traits," *Nature Genetics*, 47, 1236–1241. [1,3,4,10]

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018), "The UK Biobank Resource With Deep Phenotyping and Genomic Data," *Nature*, 562, 203–209. [8]

Davies, G., Lam, M., Harris, S. E., Trampush, J. W., Luciano, M., Hill, W. D., Hagenaars, S. P., Ritchie, S. J., Marioni, R. E., Fawns-Ritchie, C., et al. (2018), "Study of 300,486 Individuals Identifies 148 Independent Genetic Loci Influencing General Cognitive Function," *Nature Communications*, 9, 2098. [2,9]

Dobriban, E., and Wager, S. (2018), "High-dimensional Asymptotics of Prediction: Ridge Regression and Classification," *The Annals of Statistics*, 46, 247–279. [3]

Dudbridge, F. (2013), "Power and Predictive Accuracy of Polygenic Risk Scores," *PLoS Genetics*, 9, e1003348. [2,6]

Evans, L. M., Tahmasbi, R., Vrieze, S. I., Abecasis, G. R., Das, S., Gazal, S., Bjelland, D. W., de Candia, T. R., Goddard, M. E., Neale, B. M., Yang, J., Visscher, P. M., Keller, M. C. (2018), "Comparison of Methods That Use Whole Genome Data to Estimate the Heritability and Genetic Architecture of Complex Traits," *Nature Genetics*, 50, 737–745. [4]

Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A., and Smoller, J. W. (2019), "Polygenic Prediction Via Bayesian Regression and Continuous Shrinkage Priors," *Nature Communications*, 10, 1–10. [8]

Guo, Z., Wang, W., Cai, T. T., and Li, H. (2019), "Optimal Estimation of Genetic Relatedness in High-dimensional Linear Models," *Journal of the American Statistical Association*, 114, 358–369. [3]

Jernigan, T. L., Brown, T. T., Hagler Jr, D. J., Akshoomoff, N., Bartsch, H., Newman, E., Thompson, W. K., Bloss, C. S., Murray, S. S., Schork, N., Kennedy, D. N., Kuperman, J. M., McCabe, C., Chung, Y., Libiger, O., Maddox, M., Casey, B. J., Chang, L., Ernst, T. M., Frazier, J. A., Gruen, J. R., Sowell E. R., Kenet, T., Kaufmann, W. E., Mostofsky, S., Amaral, D. G., and Dale, A. M. (2016), "The Pediatric Imaging, Neurocognition, and Genetics (PING) Data Repository," *Neuroimage*, 124, 1149–1154. [2]

Jiang, J., Li, C., Paul, D., Yang, C., and Zhao, H. (2016), "On High-dimensional Misspecified Mixed Model Analysis in Genome-wide Association Study," *The Annals of Statistics*, 44, 2127–2160. [3,4]

Loh, P.-R., Bhatia, G., Gusev, A., Finucane, H. K., Bulik-Sullivan, B. K., Pollack, S. J., de Candia, T. R., Lee, S. H., Wray, N. R., Kendler, K. S., O'Donovan, M. C., Neale, Benjamin, M., Patterson, N., and Price, A. L. (2015), "Contrasting Genetic Architectures of Schizophrenia and Other Complex Diseases Using Fast Variance-Components Analysis," *Nature Genetics*, 47, 1385–1392. [1]

Lu, Q., Li, B., Ou, D., Erlendsdottir, M., Powles, R. L., Jiang, T., Hu, Y., Chang, D., Jin, C., Dai, W., He, Q., Liu, Z., Mukherjee, S., Crane, P. K., Zhao, H. (2017), "A Powerful Approach to Estimating Annotation-stratified Genetic Covariance via GWAS Summary Statistics," *The American Journal of Human Genetics*, 101, 939–964. [3]

Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X., and Sham, P. C. (2017), "Polygenic Scores Via Penalized Regression on Summary Statistics," *Genetic Epidemiology*, 41, 469–480. [8]

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019), "Clinical Use of Current Polygenic Risk Scores May Exacerbate Health Disparities," *Nature Genetics*, 51, 584–591. [10]

Ni, G., Moser, G., Ripke, S., Neale, B. M., Corvin, A., Walters, J. T., Farh, K.-H., Holmans, P. A., Lee, P., Bulik-Sullivan, B., et al. (2018), "Estimation of Genetic Correlation Via Linkage Disequilibrium Score Regression and Genomic Restricted Maximum Likelihood," *The American Journal of Human Genetics*, 102, 1185–1194. [1,10]

Nikulin, V. V., Marzinzik, F., Wahl, M., Schneider, G.-H., Kupsch, A., Curio, G., and Klostermann, F. (2008), "Anticipatory Activity in the Human Thalamus is Predictive of Reaction Times," *Neuroscience*, 155, 1275–1283. [9]

Pasaniuc, B., and Price, A. L. (2017), "Dissecting the Genetics of Complex Traits Using Summary Association Statistics," *Nature Reviews Genetics*, 18, 117–127. [3]

Purcell, S. M., Wray, R., Stone, L., Visscher, M., O'Donovan, C., Sullivan, F., Sklar, P., Ruderfer, M., McQuillin, A., Morris, W., et al. (2009), "Common Polygenic Variation Contributes to Risk of Schizophrenia and Bipolar Disorder," *Nature*, 460, 748–752. [1,10]

Shi, H., Mancuso, N., Spendlove, S., and Pasaniuc, B. (2017), "Local Genetic Correlation Gives Insights Into the Shared Genetic Architecture of Complex Traits," *The American Journal of Human Genetics*, 101, 737–751. [3]

Wolff, M. and Vann, S. D. (2019), "The Cognitive Thalamus as a Gateway to Mental Representations," *Journal of Neuroscience*, 39, 3–14. [9]

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., M. Goddard, E., and Visscher, P. M. (2010), "Common SNPs Explain a Large Proportion of the Heritability for Human Height," *Nature Genetics*, 42, 565–569. [1,3,4]

Zhao, B., Luo, T., Li, T., Li, Y., Zhang, J., Shan, Y., Wang, X., Yang, L., Zhou, F., Zhu, Z., on behalf of Alzheimer's Disease Neuroimaging Initiative, Pediatric Imaging, Neurocognition and Genetics, & Zhu, H. (2019), "Genome-wide Association Analysis of 19,629 Individuals Identifies Variants Influencing Regional Brain Volumes and Refines Their Genetic co-architecture With Cognitive and Mental Health Traits," *Nature Genetics*, 51, 1637–1644. [9]