Estimating Trans-Ancestry Genetic Correlation with Unbalanced Data Resources

Bingxin Zhao\textsuperscript{a}, Xiaochen Yang\textsuperscript{b}, and Hongtu Zhu\textsuperscript{c}

\textsuperscript{a}Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA; \textsuperscript{b}Department of Statistics, Purdue University, West Lafayette, IN; \textsuperscript{c}Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC

\textbf{ABSTRACT}

The aim of this article is to propose a novel method for estimating trans-ancestry genetic correlations in genome-wide association studies (GWAS) using genetically predicted observations. These correlations describe how genetic architecture of complex traits varies among populations. Our new estimator corrects for biases arising from prediction errors in high-dimensional weak GWAS signals, while addressing the ethnic diversity inherent in GWAS data, such as linkage disequilibrium (LD) differences. A distinguishing feature of our approach is its flexibility regarding sample sizes: it necessitates a large GWAS sample only from one population, while the secondary population may have a much smaller cohort, even in the hundreds. This design directly addresses the existing imbalance in GWAS data resources, where datasets for European populations typically outnumber those of non-European ancestries. Through extensive simulations and real data analysis from the UK Biobank study encompassing 26 complex traits, we validate the reliability of our method. Our results illuminate the broader implications of transferring genetic findings across diverse populations. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

\textbf{1. Introduction}

The distribution of heritable complex traits and diseases often varies across populations. For example, Hispanics and Blacks may be more susceptible to diseases that affect the white matter of the brain, such as Alzheimer’s disease (Chen and Zissimopoulos 2018). Such disparities in phenotypic manifestations across populations can arise from a combination of genetic, environmental, and demographic factors. These include differences in allele frequency, linkage disequilibrium (LD), genetic influences, and lifestyle patterns. Understanding the trans-ancestry genetic effects on phenotypic variation is paramount for downstream analyses of disease mechanisms and facilitating drug discovery.

In genome-wide association studies (GWAS), the trans-ancestry genetic correlation serves as a measure of genetic similarity between populations (Van Rheenen et al. 2019). Briefly, genetic correlation quantifies the association between two sets of genetic effect sizes spanning the genome. A high trans-ancestry genetic correlation suggests greater applicability and transferability of genetic findings specific to one population to another.

Despite the proliferation of methods to estimate genetic correlations within the same population using either individual-level data or GWAS summary statistics (Loh et al. 2015; Bulik-Sullivan et al. 2015a; Lu et al. 2017; Guo et al. 2019; Speed and Balding 2019; Wang and Li 2022), estimating trans-ancestry genetic correlations in GWAS remains fraught with challenges. The first challenge arises from the inherent heterogeneity in GWAS across different populations. Notably, differences in LD patterns can cause biased findings if analysis neglects these distinctions (Zhang et al. 2021). However, most of the existing methods for estimating genetic correlation have the assumption that the LD patterns of genetic variants remain consistent across GWAS or that the genetic variants are uncorrelated. Furthermore, because of the heterogeneity caused by LD, the definition of trans-ancestry genetic correlation is even unclear (Wang and Li 2022). The second challenge pertains to the unbalanced distribution of GWAS data across global populations. A significant portion of GWAS data originates from European populations, leaving many non-European ancestries underrepresented. Specifically, about 79% of all GWAS participants are of European descent, and the fraction of non-Europeans in GWAS has stagnated or declined since late 2014 (Martin et al. 2019). When non-European GWAS have limited sample sizes, summary statistics-based methods, such as the Popcorn (Brown et al. 2016) and XPASS (Cai et al. 2021), may have poor performance for trans-ancestry genetic correlation estimation (Ni et al. 2018; Zhang et al. 2021).

In this study, we present a novel approach to tackle these two major challenges associated with trans-ancestry genetic correlation estimation. Our method is LD difference-aware and is applicable to GWAS with a small number of subjects. The proposed method is based on constructing genetically predicted traits in non-European GWAS, in which the genetic effects are learned from large-scale European GWAS. As a result, this estimator only needs the European population to have a large GWAS sample size, and the non-European population can only have a much smaller number of individuals (e.g., hundreds).
We estimate and correct the prediction-induced bias in high-dimensional genetic variants data and alleviate the adverse effects of inconsistent LD patterns. Additionally, we examine the popular reference panel-based approaches (Pasaniuc and Price 2017) in trans-ancestry analysis. We develop a pipeline to implement our estimator on real genotype data from the UK Biobank (UKB) (Bycroft et al. 2018). We use extensive simulations and real data analyses to show that our method can provide reliable estimators for complex traits from different domains. The software and pipeline are freely available at https://github.com/FSSKM/TAGC_review.

This article proceeds as follows. In Section 2, we introduce the model setups and definition for trans-ancestry genetic correlation. In Section 3, we develop our estimator and study the influence of LD heterogeneity. Section 4 analyzes reference panel-based approaches in trans-ancestry analysis. Section 5 provides numerical details, including the simulation results, implementation of the estimator in real GWAS data, and real data analysis. We discuss a few future topics in Section 6. Most of the technical details are provided in the supplementary file.

2. Modeling Framework

2.1. Model Setups and Assumptions

Consider two independent GWAS that are conducted on individuals from two different ancestry groups (e.g., European and Asian) with the same p genetic variants, most of which are single nucleotide polymorphisms (SNPs):

- Population-I GWAS: \( (X, y) \) with \( X = (x_1, \ldots, x_p) \in \mathbb{R}^{n \times p} \) and \( y \in \mathbb{R}^{n \times 1} \),
- Population-II GWAS: \( (Z, y_z) \) with \( Z = (z_1, \ldots, z_p) \in \mathbb{R}^{n_z \times p} \) and \( y_z \in \mathbb{R}^{n_z \times 1} \),

where \( y \) and \( y_z \) are continuous complex traits measured in the two GWAS with sample sizes \( n \) and \( n_z \), respectively. In practice, they may represent either the same trait in two different populations, such as height, or two different but genetically related traits, such as regional brain volume and intelligence. The linear additive polygenic models are assumed between complex traits and genetic variants (Jiang et al. 2016) as follows:

\[
y = X \beta + \epsilon \quad \text{and} \quad y_z = Z \alpha + \epsilon_z,
\]

where \( \beta^T = (\beta_1, \ldots, \beta_p)^T \) and \( \alpha^T = (\alpha_1, \ldots, \alpha_p)^T \) are population-specific genetic effects and \( \epsilon \) and \( \epsilon_z \) represent population-specific random error vectors. Then, the genetic heritability of \( y \) and that of \( y_z \) are, respectively, given by

\[
h^2_y = \frac{\beta^T X^T X \beta}{\beta^T X^T X \beta + \epsilon^T \epsilon} \quad \text{and} \quad h^2_z = \frac{\alpha^T Z^T Z \alpha}{\alpha^T Z^T Z \alpha + \epsilon^T_z \epsilon_z}.
\]

The \( h^2_y \) (or \( h^2_z \)) measures the proportion of variation in \( y \) (or \( y_z \)) that can be explained by additive genetic effects across the genome.

We introduce some assumptions on complex traits and genetic variants in order to quantify the effect of LD heterogeneity on trans-ancestry genetic correlations.

SNP data. We summarize the assumptions on SNP data \( X \) and \( Z \) in Condition 1.

**Condition 1.** (a) We assume \( X = X_0 \Sigma_{X}^{1/2} \) and \( Z = Z_0 \Sigma_{Z}^{1/2} \). Entries of \( X_0 \) and \( Z_0 \) are real-value iid random variables with mean zero, variance one, and a finite fourth order moment. The \( \Sigma_X \) and \( \Sigma_Z \) are \( p \times p \) population level deterministic positive definite matrices with uniformly bounded eigenvalues. Specifically, we have \( 0 < c \leq \lambda_{\min}(\Sigma_X) \leq \lambda_{\max}(\Sigma_X) \leq C \) for all \( p \) and some constants \( c, C \), where \( \lambda_{\min}(\cdot) \) and \( \lambda_{\max}(\cdot) \) are the smallest and largest eigenvalues of a matrix, respectively. \( \Sigma_Z \) satisfies similar conditions.

(b) Let \( F_p(\Sigma)(x) = p^{-1} \sum_{i=1}^{p} I(\lambda_i(\Sigma_X) \leq x) \) denote the empirical spectral distributions (ESD) of \( \Sigma_X \), where \( I(\cdot) \) is the indicator function, \( \lambda_i(\cdot) \) is the \( i \)th eigenvalue of a matrix and \( x \in \mathbb{R} \). As \( p \to \infty \), the sequence of ESDs \( \{F_p(\Sigma)(x)\}_{p=1} \) converges weakly to the limiting spectral distribution (LSD) of \( \Sigma_X \), denoted as \( H_X(x) \). Similarly, the LSDs of \( \Sigma_Z \), \( \Sigma_X^{1/2}, \Sigma_Z^{1/2} \), \( \Sigma_Z \) exist and are denoted as \( H_Z(x) \), \( H_{X/Z}(x) \), \( H_{Z/X}(x) \), and \( H_{Z/Z}(x) \), respectively.

(c) As \( \min(n, n_z) \to \infty \), we assume \( p/n \rightarrow \omega \) and \( p/n_z \rightarrow \omega_z \) for \( \omega \) and \( \omega_z \) in \((0, \infty)\).

Conditions 1 (a) and (b) are frequently used in the application of random matrix theory for high-dimensional data (Dobriban and Wager 2018). Moreover, \( \Sigma_X \) and \( \Sigma_Z \) can be different, representing different patterns of LD in diverse populations. In Condition 1 (a), we have assumed that the eigenvalues are bounded away from 0, which allows us to apply existing technical lemmas in the random matrix theory field. In practice, some eigenvalues of \( \Sigma_X \) and \( \Sigma_Z \) might be close to zero due to the high relatedness among certain genetic variants. Nevertheless, the proposed estimators proposed in later sections will remain robust to these small boundary eigenvalues. This is due to the fact that our estimators only require the estimation of a few moments of the entire eigenvalue distribution, and the impact of a few boundary eigenvalues is small. In Condition 1 (c), we assume that the GWAS sample sizes \( n \) and \( n_z \) and the number of genetic variants \( p \) are proportional to each other (Jiang et al. 2016). Moreover, we allow a flexible range for \( \omega \) and \( \omega_z \), where \( \omega \) can be close to one and \( \omega_z \) can be much larger. For GWAS, unimputed genotype data typically have about half a million genetic variants and genotype imputation can increase the number to several millions. In contrast, European GWAS may have large sample sizes (e.g., over 1 million for certain traits), whereas non-European GWAS typically have much smaller sample sizes (e.g., several thousand). The framework and methods developed in this article can also be used to perform within-population genetic correlation analyses between two different traits, while controlling for LD heterogeneity among different datasets.

**Genetic effects and random errors.** Let \( F(0, V) \) denote a generic distribution with mean zero, (co)variance \( V \), and finite 4th order moments. We introduce the following conditions on genetic effects and random errors.
Condition 2. Let $\Phi_{\beta\beta}, \Phi_{\alpha\alpha}$, and $\Phi_{\beta\alpha}$ be diagonal matrices, in which $\Phi_{\beta\beta} = \text{diag}(\phi_{\beta1}^2, \ldots, \phi_{\beta p}^2)$, $\Phi_{\alpha\alpha} = \text{diag}(\phi_{\alpha1}^2, \ldots, \phi_{\alpha p}^2)$, and $\Phi_{\beta\alpha} = \text{diag}(\phi_{\beta\alpha1}, \ldots, \phi_{\beta\alpha p})$ with all diagonal elements in $\in [0, \infty)$. The joint distribution of $\beta$ and $\alpha$ is given by

$$
(\beta, \alpha) \sim F \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, p^{-1} \begin{pmatrix} \Phi_{\beta\beta} & \Phi_{\beta\alpha} \\ \Phi_{\beta\alpha}^t & \Phi_{\alpha\alpha} \end{pmatrix} \right).
$$

In addition, we have $\phi_{\beta\alpha j} = 0$ if either $\phi_{\beta j}^2 = 0$ or $\phi_{\alpha j}^2 = 0$. Let $m_\beta, m_\alpha$, and $m_{\beta\alpha}$ denote the number of positive entries in the $\Phi_{\beta\beta}, \Phi_{\alpha\alpha}$, and $\Phi_{\beta\alpha}$, respectively. As $p \to \infty$, we assume $m_\beta/p \to \kappa_\beta \in (0, 1)$, $m_\alpha/p \to \kappa_\alpha \in (0, 1)$, $m_{\beta\alpha}/p \to \delta_{\beta\alpha} \in (0, 1)$, and $m_{\beta\alpha}/\sqrt{m_\beta m_\alpha} \to \kappa_{\beta\alpha} \in (0, 1)$. For random errors, $\epsilon_j$ in $\epsilon$ and $\epsilon_j$ in $\epsilon_z$ are independent random variables and have distributions $\epsilon_j \sim F(0, \sigma_j^2), j = 1, \ldots, n$ and $\epsilon_{zj} \sim F(0, \sigma_{zj}^2), j = 1, \ldots, n_z$.

(b) We assume $\text{tr}(\Sigma_X \Sigma_Z \Phi_{\beta\alpha}) = \phi_{\beta\alpha} \cdot \text{tr}(\Sigma_X \Sigma_Z) \cdot (1 + o_p(1))$, $\text{tr}(\Sigma_X \Sigma_Z \Phi_{\beta\beta} \Sigma_Z \Phi_{\beta\alpha}) = \phi_{\beta\beta} \cdot \text{tr}(\Sigma_X \Sigma_Z) \cdot (1 + o_p(1))$, and $\text{tr}(\Sigma_X \Sigma_Z \Sigma_X \Phi_{\beta\alpha}) = \phi_{\beta\alpha}^2 \cdot \text{tr}(\Sigma_X^2 \Sigma_Z) \cdot (1 + o_p(1))$, where $\Sigma_X = n^{-1}X^T X$, $\phi_{\beta\alpha} = \sum_{j=1}^p \phi_{\beta\alpha j} / p$, and $\phi_{\beta\alpha}^2 = \sum_{j=1}^p \phi_{\beta\alpha j}^2 / p$. Here, $\phi_{\beta\alpha}^2$ denotes the average genetic effect per variant, while $\phi_{\beta\alpha}^2$ represents the average contribution of each variant to the genetic correlation between $y$ and $y_z$.

Condition 2 (a) details a random effect model, in which genetic effects are independent and may vary in scale and an arbitrary proportion of them is allowed to be zero. Moreover, without further restrictions on their sparsity, $m_\beta, m_\alpha$, and $m_{\beta\alpha}$ are proportional to the number of all genetic variants $p$. Our random effect model weakens the classical iid random effect assumption in GWAS, which typically assumes $\Phi_{\beta\beta} = \text{diag}(\phi_{\beta j}^2)$. $I_{m_\beta}, 0_{p-m_\beta}$ for some constant genetic effect $\phi_{\beta j}^2$ (Yang et al. 2011; Bulik-Sullivan et al. 2015a; Jiang et al. 2016). Condition 2 (b) outlines the relationships between the LD structures and genetic effects that are essential for the non-iid random effect model. The inspiration behind this condition is from the observation that in the iid model, which is a specific case of our framework, the behavior of traces involving both LD structures and random genetic effects are mainly driven by the LD structures. For example, setting $\Phi_{\beta\alpha} = \phi_{\beta\alpha} \cdot I_p$ results in $\text{tr}(\Sigma_X \Sigma_Z \Phi_{\beta\alpha}) = \phi_{\beta\alpha} \cdot \text{tr}(\Sigma_X \Sigma_Z) \cdot (1 + o_p(1))$. Consequently, the behavior of $\text{tr}(\Sigma_X \Sigma_Z \Phi_{\beta\alpha})$ can be approximated by $\text{tr}(\Sigma_X \Sigma_Z)$.

Condition 2(b) ensures that similar approximations are valid for non-iid models. The central insight lies in the necessity to balance the entries of $\Sigma_X \Sigma_Z$ and $\Sigma_Z \Sigma_Z$ throughout the genome. Real LD patterns, which often display a block-diagonal structure, could potentially reinforce this assumption. This condition also provides insights into the robustness of iid random effect models in GWAS.

### 2.2. Heritability and Trans-ancestry Genetic Correlation

According to Conditions 1 and 2, we have the following results for heritability and trans-ancestry genetic correlation.

**Heritability.** The heritability $h_\beta^2$ defined in (2) can be approximated as $h_\beta^2 = \frac{\|\beta\|_2^2}{\text{tr}(\Sigma_X \Phi_{\beta\beta}) / p} + o_p(1)$. In addition, we have $h_\beta^2 = \|\beta\|_2^2 / \|\alpha\|_2^2$.

**Trans-ancestry genetic correlation.** We consider two popular definitions and highlight their differences and connections (Brown et al. 2016). The first one is the Pearson correlation of population-specific genetic effect vectors given by $\psi_{\beta\alpha} = \langle \beta/\|\beta\|_2, \alpha/\|\alpha\|_2 \rangle = \text{tr}(\Phi_{\beta\alpha}) / \text{tr}(\Phi_{\beta\beta})$. However, a major issue is that $\psi_{\beta\alpha}$ does not account for the indirect correlation of genetic effects due to the LD among causal variants (Wang and Li 2022). Incorporating the LDs of both GWAS leads to the second one as follows:

$$
\psi_{\beta\alpha}^* = \frac{\text{tr}(\Sigma_X^{1/2} \Sigma_Z^{1/2} \Phi_{\beta\alpha})}{\text{tr}(\Sigma_X^{1/2} \Sigma_Z^{1/2})} + o_p(1).
$$

We discuss the connections between $\psi_{\beta\alpha}^*$ and $\psi_{\beta\alpha}$. First, we consider the balanced case satisfying $\Phi_{\beta\alpha} = \phi_{\beta\alpha} \cdot I_p$, and $\Phi_{\beta\alpha} = \phi_{\beta\alpha}^2 \cdot I_p$. In this case, we have

$$
\psi_{\beta\alpha}^* = \psi_{\beta\alpha} \cdot b_1(\Sigma_X^{1/2} \Sigma_Z^{1/2}) + o_p(1),
$$

where $b_1(\Sigma_X^{1/2} \Sigma_Z^{1/2}) = \int_{\mathbb{R}} tdH_{X1/2Z1/2}(t)$ is the first moment of the LSD of $\Sigma_X^{1/2} \Sigma_Z^{1/2}$. Thus, the ratio of $\psi_{\beta\alpha}^*$ over $\psi_{\beta\alpha}$ can be approximated by $b_1(\Sigma_X^{1/2} \Sigma_Z^{1/2})$. In Section 5, we will introduce a consistent estimator of $b_1(\Sigma_X^{1/2} \Sigma_Z^{1/2})$ in real GWAS data. Then $\psi_{\beta\alpha}^*$ can be obtained from $\psi_{\beta\alpha}$ by applying this consistent estimator of $b_1(\Sigma_X^{1/2} \Sigma_Z^{1/2})$. Second, we can directly estimate $\psi_{\beta\alpha}^*$ through decorrelating $X$ and $Z$ into $X_0 = L_X X$ and $Z_0 = Z \Sigma_Z^{-1/2}$. Therefore, models in (1) reduce to

$$
\begin{align*}
    y &= X_0 \tilde{\beta} + \epsilon = X_0 \Sigma_X^{-1/2} \beta + \epsilon, \\
    y_z &= Z_0 \tilde{\alpha} + \epsilon_z = Z_0 \Sigma_Z^{-1/2} \alpha + \epsilon_z,
\end{align*}
$$

where $\tilde{\beta}$ and $\tilde{\alpha}$ are the corresponding genetic effects. Therefore, $\psi_{\beta\alpha}$ is equal to the Pearson correlation of genetic effects of decorrelated SNP data $X_0$ and $Z_0$. In practice, SNP data decorrelation can be performed either within each predetermined independent LD block as in Berisa and Pickrell (2016) or with a given window size as in Bulik-Sullivan et al. (2015a). As $\psi_{\beta\alpha}$ and $\psi_{\beta\alpha}^*$ are closely connected, we focus on estimating $\psi_{\beta\alpha}$ from now on.
3. Estimation Using Genetically Predicted Traits

In this section, we propose a consistent estimator of $\varphi_{\beta \alpha}$ and investigate the effects of LD heterogeneity on trans-ancestry analysis.

3.1. Consistent Estimators of Trans-Ancestry Genetic Correlation

Our estimator leverages the widely used GWAS marginal summary association statistics (Pasaniuc and Price 2017) derived from Population-I GWAS and genetically predicted traits for all subjects in the Population-II GWAS. We estimate $\varphi_{\beta \alpha}$ using either predicted or observed values for the same set of individuals in Population-II GWAS, after adjustments for prediction error and LD disparities. For $y$, the Population-I GWAS summary statistics are denoted by $\beta = n^{-1}X^Ty$. The genetically-predicted values for the Population-II GWAS are then determined using $\tilde{Y}_\beta = Z\beta$. This $\tilde{Y}_\beta$ is commonly referred to as cross-population polygenic risk scores (Duncan et al. 2019), the genetic endowments linked to this trait (Barth, Papageorge, and Thom 2020), or the genetically determined trait (Codd et al. 2021). Drawing from $\tilde{Y}_\beta$, a popular estimator for trans-ancestry genetic correlation is $G_{\beta \alpha} = y^T\tilde{Y}_\beta/\|y\|_2 \cdot \|\tilde{Y}_\beta\|$. Although $G_{\beta \alpha}$ has been widely reported in the literature, its asymptotic behavior remains largely unknown. We delve into the asymptotic limit of $G_{\beta \alpha}$ in the following theorem.

**Theorem 1.** Under polygenic model (1) and Conditions 1 and 2, as $n, n_z, m_{\beta \alpha}, p \to \infty$, for any $\omega, \omega_z \in (0, \infty)$, $h_\beta^2, h_\alpha^2 \in (0, 1]$, and $\varphi_{\beta \alpha} \in [-1, 1]$, we have

$$G_{\beta \alpha} = \varphi_{\beta \alpha} \cdot h_\alpha \cdot \left[ \frac{b_1(\Sigma_X^2 \Sigma_Z)}{b_1((\Sigma_X \Sigma_Z))} + \frac{\omega}{h_\beta^2 \cdot b_1(\Sigma_X \Sigma_Z)} \right]^{-1/2} + o_p(1),$$

where $b_1(\Sigma_X^2 \Sigma_Z) = \int_{R} t dH_{XZ}(t)$ and $b_1(\Sigma_X \Sigma_Z) = \int_{R} t dH_{XZ}(t)$.

Theorem 1 reveals that $G_{\beta \alpha}$ acts as a shrinkage estimator of $\varphi_{\beta \alpha}$, impacted significantly by prediction errors and LD differences. Intuitively, the shrinkage is largely caused by using genetically predicted values rather than actual observed observations in estimating the correlation. Even in within-population analyses, a notable mismatch between prediction accuracy and heritability across various complex traits has been observed in GWAS (Daetwyler, Villanueva, and Woolliams 2008). Similarly, using the predicted values to access genetic correlations between two traits may lead to seriously underestimated results. Zhao and Zhu (2022) quantifies the potential bias for within-population analysis, which can be viewed as a special case of our results under independent genetic variants (i.e., $\Sigma_X = \Sigma_Z = I_p$) and iid random effect models. Under more general settings, we use techniques from random matrix theory (Bai and Silverstein 2010) to demonstrate that the shrinkage of $\varphi_{\beta \alpha}$ in trans-ancestry studies is collectively influenced by the heritability metrics across both populations, the sample size in Population-I GWAS, and the first moments of the LSDs of $\Sigma_X \Sigma_Z$ and $\Sigma_X^2 \Sigma_Z$. These insights guide our proposal for a consistent estimator of $\varphi_{\beta \alpha}$. **Consistent estimator of $\varphi_{\beta \alpha}$**. It follows from Theorem 1 that we have

$$G_{\beta \alpha}^M = G_{\beta \alpha} \cdot \left[ \frac{b_1(\Sigma_X^2 \Sigma_Z)}{h_\beta^2 \cdot b_1((\Sigma_X \Sigma_Z))} + \frac{\omega}{h_\alpha^2 \cdot b_1(\Sigma_X \Sigma_Z)} \right]^{1/2},$$

which is a consistent estimator of $\varphi_{\beta \alpha}$. For most complex traits, reliable estimates of $h_\beta^2$ and $h_\alpha^2$ exist (Yang et al. 2011; Hou et al. 2019; Speed and Balding 2019). The major challenge to approximate $G_{\beta \alpha}^M$ is to estimate $b_1(\Sigma_X \Sigma_Z)$ and $b_1(\Sigma_X^2 \Sigma_Z)$, in which the dimensions of $\Sigma_X$ and $\Sigma_Z$ are very large. Details about our implementation will be provided in Section 5.2.

In addition, we assume that $\text{tr}((\Sigma_X \Sigma_Z)) = \varphi_{\beta \alpha} \cdot \text{tr}(\Sigma_Z) \cdot (1 + o_p(1))$, $\text{tr}(\Sigma_X \Sigma_Z) = \varphi_{\beta \alpha} \cdot \text{tr}(\Sigma_X) \cdot (1 + o_p(1))$, and $\text{tr}(\Sigma_X \Sigma_Z) = \varphi_{\beta \alpha} \cdot \text{tr}(\Sigma_X) \cdot (1 + o_p(1))$. Similar to Condition 2(b), these assumptions ensure that the behavior of traces involving both LD structures and genetic effects can be effectively represented by the LD structures. Then, we have

$$\text{var}(G_{\beta \alpha}^M) = O_p\left[ \max\left\{ \frac{1}{n}, \frac{1}{n_z}, \frac{Q_1 + Q_2 + \sum_{i=1}^{p} C_{\beta \alpha}(\Sigma_X \Sigma_Z)^2}{\text{tr}(\Phi_{\alpha \alpha}) \cdot \text{tr}(\Phi_{\beta \beta})} \right\} \right],$$

where $Q_1 = \text{tr}(\Sigma_Z \Sigma_X \Phi_{\beta \alpha} \Sigma_X \Sigma_Z \Phi_{\beta \alpha})$, $Q_2 = \text{tr}(\Sigma_Z \Phi_{\alpha \alpha} \Sigma_X \Sigma_X \Phi_{\beta \alpha})$, and $C_{\beta \alpha} = E[\alpha^2 \beta^2_1] - 2E[\alpha \beta_1]^2 - E[\alpha^2][\beta^2_1]$. The variance of $G_{\beta \alpha}^M$ depends on the sample sizes for Population-I and Population-II GWAS, as well as the degree of signal sparsity related to $O(\text{tr}(\Phi_{\beta \alpha})^{-1}), O(\text{tr}(\Phi_{\beta \alpha})^{-1})$, and $O(\text{tr}(\Phi_{\alpha \alpha})^{-1})$. The exact form of the asymptotic limit of $\text{var}(G_{\beta \alpha}^M)$ is provided in the supplementary file. The variance of $G_{\beta \alpha}^M$ increases as signals become sparser. Because $n$ is much larger than $n_z$ in most cases, $\text{var}(G_{\beta \alpha}^M)$ has a scale of $O(1/n_z)$ when the genetic signals are not very sparse, quantifying by $|Q_1 + Q_2 + \sum_{i=1}^{p} C_{\beta \alpha}(\Sigma_X \Sigma_Z)^2| / [\text{tr}(\Phi_{\alpha \alpha}) \cdot \text{tr}(\Phi_{\beta \beta})] > n_z$. Thus, our estimator is reliable for polygenic or omnigenic traits (Timpson et al. 2018) with a large number of causal variants.

3.2. Effects of LD Heterogeneity on Trans-Ancestry Analysis

In this section, we delve into the influence of LD heterogeneity on trans-ancestry analysis. LD heterogeneity holds significant practical implications, especially when assessing the transferability of GWAS findings between diverse populations. Notably, a clear decline in the effectiveness of genetic predictions of complex traits is observed when European GWAS results are applied to non-European cohorts (Weissbrod et al. 2022). This decrease in performance can be attributed in part to variations in allele-normalized genetic effects between populations, as reflected by values of $|\varphi_{\beta \alpha}| < 1$. However, even in scenarios where the genetic variants demonstrate almost identical effects across the two populations, denoted by $\varphi_{\beta \alpha} \approx 1$, reduced prediction performance can still be observed. For example, while the trans-ancestry genetic correlation of schizophrenia is reported to be 0.98 between East Asian and European groups, predictions can be 50% more accurate in within-European analysis compared to European-Asian analysis (Lam et al. 2019). These remaining discrepancies may be caused by the LD heterogeneity.

First, we introduce an LD-related shrinkage factor as $S_{\beta \alpha} = [b_1(\Sigma_X^2 \Sigma_Z)/b_1((\Sigma_X \Sigma_Z)) + \omega/(h_\beta^2 \cdot b_1(\Sigma_X \Sigma_Z))]^{-1/2}$. Directly
applying Theorem 1 shows that smaller $S_{\beta\alpha}$ indicates more serious shrinkage and smaller $G_{\beta\alpha}$. To study the effects of LD heterogeneity on trans-ancestry analysis, we further consider a generalized version of Theorem 1 by defining $\Sigma(t) = (1 - t)\Sigma_Z + t\Sigma_X$, $t \in [0, 1]$. With $\beta = n^{-1}X^T y$ and SNP data $Z(t) = Z_0Z(t)^{1/2}$, the genetically predicted values generated on the secondary GWAS is $Z(t)\beta$, which results in the estimator $G_{\beta\alpha}(t)$. Then we have a generalized version of $S_{\beta\alpha}$ as follows:

$$S_{\beta\alpha}(t) = \left[ \frac{t \cdot (b_3(\Sigma_X) + \omega/h_{\beta}^2 \cdot b_2(\Sigma_X)) + (1 - t) \cdot (b_1(\Sigma_X^2 \Sigma_Z) + \omega/h_{\beta}^2 \cdot b_1(\Sigma_X \Sigma_Z))}{(t \cdot b_2(\Sigma_X) + (1 - t) \cdot b_1(\Sigma_X \Sigma_Z))^2} \right]^{1/2},$$

where $b_3(\Sigma_X) = \int_{\mathbb{R}} r^2 dh_X(t)$ and $b_1(\Sigma_X) = \int_{\mathbb{R}} r^2 dh_Y(t)$. Here $t = 0$ and $t = 1$ are two special cases. When $t = 0$, we have $S_{\beta\alpha}(t) = S_{\beta\alpha}$, which characterizes the shrinkage when predicting complex traits in Population-II by using the results from the Population-I GWAS. On the other hand, when $t = 1$, $S_{\beta\alpha}(t)$ represents the LD-related shrinkage factor when performing prediction between two Population-I GWAS.

We study the effect of LD heterogeneity on $G_{\beta\alpha}(t)$ by taking the first-order derivative of $S_{\beta\alpha}(t)$ with respect to $t$, which is given by $S_{\beta\alpha}'(t) = \{a(ct - d) + 2bc/2(at + b)^{3/2}\}$, where $a = b_3(\Sigma_X) - b_1(\Sigma_X^2 \Sigma_Z) + \omega/h_{\beta}^2 \cdot [b_2(\Sigma_X) - b_1(\Sigma_X \Sigma_Z)]$, $b = b_1(\Sigma_X^2 \Sigma_Z) + \omega/h_{\beta}^2 \cdot b_2(\Sigma_X \Sigma_Z)$, and $c = b_2(\Sigma_X) - b_1(\Sigma_X \Sigma_Z)$, and $d = b_1(\Sigma_X \Sigma_Z)$. For many real GWAS studies with large $\omega$ (i.e., $nh_{\beta}^2$ is typically much smaller than $p$), we may have $a \approx \omega/h_{\beta}^2 \cdot c$ and $b \approx \omega/h_{\beta}^2 \cdot d$. It follows that $S_{\beta\alpha}'(t) \approx c/[2(\omega h_{\beta}^2)^{1/2} \cdot (ct + d)^{1/2}]$. Thus, if $b_2(\Sigma_X) > b_1(\Sigma_X \Sigma_Z)$, then $\hat{S}_{\beta\alpha}(t) > 0$ for $t \in [0, 1]$ and $S_{\beta\alpha}(t)$ has the largest value at $t = 1$. Otherwise, if $b_2(\Sigma_X) < b_1(\Sigma_X \Sigma_Z)$, then $S_{\beta\alpha}(t)$ has the largest value at $t = 0$. These results suggest that whether GWAS trans-ancestry prediction between two different populations has a lower accuracy than the within-population prediction depends on the eigenvalues of $\Sigma_X$ and $\Sigma_Z$. Specifically, this is largely quantified by the difference between the first moment of the LSD of $\Sigma_X \Sigma_Z$ and the second moment of the LSD of $\Sigma_X$. Moreover, since $b_1(\Sigma_X^2 \Sigma_Z) < \max[b_2(\Sigma_X), b_2(\Sigma_Z)]$, GWAS trans-ancestry prediction between two different populations has a lower accuracy than the best within-population predictions.

In the supplementary file, we also discuss the effect of LD heterogeneity in the classical low-dimensional setting with $\omega = 0$ (that is, $nh_{\beta}^2$ is much larger than $p$). Briefly, when the sample size is much larger than the number of features, we find that the LD mismatch impacts the performance of the cross-population estimates in a more complicated way. We conduct a simulation study to better illustrate these findings in the supplementary file, and the numerical results are presented in Supplementary Figures 1 and 2.

To assess the practical implications of our findings, we examine $S_{\beta\alpha}(t)$ using UKB genetic data. A detailed description of our UKB data analysis is provided in Section 5.2. Briefly, European subjects in the UKB study are used to estimate $\Sigma_X$, and Asian subjects are used to estimate $\Sigma_Z$. The estimates for $b_1(\Sigma_X)$ and $b_1(\Sigma_X \Sigma_Z)$ are 4.41 and 2.86, respectively. Therefore, we may have $b_2(\Sigma_X) > b_1(\Sigma_X \Sigma_Z)$ if we use European GWAS results to generate genetically predicted values for Asian subjects. Figure 1 displays the pattern of $S_{\beta\alpha}(t)$. As expected, $S_{\beta\alpha}(t)$ has the largest value at $t = 1$ for relatively large $\omega$. Due to the LD disparities between European and Asian subjects, European GWAS results may be less accurate in predicting Asian cohorts than they are in predicting European cohorts. In summary, our findings emphasize that LD heterogeneity can significantly influence downstream analyses and predictions.

4. Reference Panels in Trans-Ancestry Analysis

When using GWAS marginal summary statistics, population-specific genotype reference panels are frequently used to account for LD patterns in within-population analyses (Pasaniuc and Price 2017). These panels are typically derived from an independent external database, such as the 1000 Genomes reference panel (1000-Genomes-Consortium 2015), which closely
aligns with the targeted population. In this section, we study the reference panel-based approaches in a unified framework. Specifically, we will show that the shrinkage exists in the naive genetic correlation estimator even after adjusting for LD with a reference panel. Moreover, we will explore the selection criteria for LD reference panels in the context of trans-ancestry analysis, especially considering the ambiguities that arise when LD patterns in the training and testing GWAS diverge.

Let \( W = (w_1, \ldots, w_p) \in \mathbb{R}^{n_w \times p} \) be a reference panel database, which is independent of both \((X, y)\) and \((Z, y_z)\). For trans-ancestry analysis, we examine three different reference panels \( W \) as follows:

- **Reference panel-I:** \( W = W_0 \Sigma X^{-1/2} \), where the entries of \( W_0 \in \mathbb{R}^{n_w \times p} \) are iid random variables with mean zero, variance one and a finite fourth order moment.
- **Reference panel-II:** \( W = W_0 \Sigma X^{1/2} \).
- **Reference panel-III:** \( W^T = (\Sigma X^{1/2} W_{01}, \Sigma X^{1/2} W_{02}) \), where \( W_{01} \) and \( W_{02} \) are sub-matrices of \( W_0^T \) such that \( W_{01}^T = [W_{01} W_{02}], W_{01} \in \mathbb{R}^{n_{w1} \times p}, W_{02} \in \mathbb{R}^{n_{w2} \times p}, \) and \( n_w = n_{w1} + n_{w2} \).

Reference panel-I and Reference panel-II are the panels matched to the LD of Population-I GWAS and that of Population-II GWAS, respectively. Reference panel-III is a mixed reference panel corresponding to both populations.

Let \( \Sigma W = n_w W^T W \) be the estimated LD matrix from the reference panel \( W \). The ridge-type reference panel-adjusted GWAS summary statistics is \( \beta W = (\Sigma W + \lambda I_p)^{-1} \beta \), where \( \lambda \in (0, \infty) \) is a ridge-type tuning parameter. Then the predicted trait in Population-II GWAS is \( \hat{y}_{\beta W} = Z \hat{\beta} W \) and the corresponding trans-ancestry genetic correlation estimator is denoted by \( G_{\beta W}^W = y^T T \hat{y}_{\beta W} /\| y_z \| \cdot \| \hat{y}_{\beta W} \| \). To investigate the asymptotic limit of \( G_{\beta W}^W \), we need to impose an additional condition on the reference panel data as follows.

**Condition 3.** As \( n_w \to \infty \), we assume \( p/n_w \to \omega_w \in (0, \infty) \). We assume \( \text{tr}(\Sigma X (\Sigma W + \lambda I_p)^{-1} \Sigma Z (\Sigma W + \lambda I_p)^{-1} \Sigma Z \Phi_{\beta W}) = \phi_2^2 \cdot \text{tr}(\Sigma X (\Sigma W + \lambda I_p)^{-1} \Sigma Z (\Sigma W + \lambda I_p)^{-1}) \cdot (1 + o_p(1)) \) and \( \text{tr}(\Sigma X (\Sigma W + \lambda I_p)^{-1} \Sigma Z \Phi_{\beta W}) = \phi_1 \cdot \text{tr}(\Sigma X (\Sigma W + \lambda I_p)^{-1} \Sigma Z) \cdot (1 + o_p(1)) \).

Similar to Condition 2(b), Condition 3 establishes relationships between the reference panel LD structures and genetic effects that are required for our non-iid random effect model. Therefore, the behaviors of traces involving both LD structures and genetic effects can be predominantly determined by \( \Sigma X (\Sigma W + \lambda I_p)^{-1} \Sigma Z \) and \( \hat{\Sigma}_X (\hat{\Sigma}_W + \lambda I_p)^{-1} \hat{\Sigma}_Z \). Then, the asymptotic limit of \( G_{\beta W}^W \) is provided in the following theorem.

**Theorem 2.** Under polygenic model (1) and Conditions 1, 2, and 3, as \( \min(n, n_s, n_w, p) \to \infty \), for any \( \omega, \omega_w, \omega_z, \lambda \in (0, \infty), h_p^2, h^2_\alpha \in (0, 1), \) and \( \varphi_{\beta W} \in [-1, 1] \), we have

\[
G_{\beta W}^W = \varphi_{\beta W} \cdot h_\alpha \cdot \left[ \frac{V_1(\lambda) \cdot h_p^2}{V_2(\lambda) \cdot \omega + V_3(\lambda) \cdot h^2_\alpha} \right]^{1/2} + o_p(1),
\]

where \( V_1(\lambda) = p^{-1} \text{tr}(\Sigma Z (\hat{\Sigma}_W + \lambda I_p)^{-1} \Sigma X), V_2(\lambda) = p^{-1} \text{tr}(\hat{\Sigma}_W + \lambda I_p)^{-1} \Sigma Z (\hat{\Sigma}_W + \lambda I_p)^{-1} \Sigma X), \) and \( V_3(\lambda) = p^{-1} \text{tr}(\hat{\Sigma}_W + \lambda I_p)^{-1} \Sigma Z (\hat{\Sigma}_W + \lambda I_p)^{-1} \Sigma Z \).

**Theorem 2** shows that the reference panel-adjusted estimator \( G_{\beta W}^W \) is still a shrinkage estimator of \( \varphi_{\beta W} \). In addition to the sample size of the Population-I GWAS and the heritability measures of both populations, the shrinkage is jointly determined by \( V_1(\lambda), V_2(\lambda), \) and \( V_3(\lambda) \), which are functions of the LD structures in \( X, Z, \) and \( W \). Similar to \( G_{\beta W} \), we can construct a consistent estimator of \( \varphi_{\beta W} \) based on \( G_{\beta W}^W \) as follows:

\[
G_{M\beta W}^W = G_{\beta W}^W \cdot \left[ \frac{V_2(\lambda) \cdot \omega + V_3(\lambda) \cdot h^2_\alpha}{V_1(\lambda) \cdot h_p^2} \right]^{1/2} = \varphi_{\beta W} + o_p(1).
\]

We use a numerical example to compare the three reference panel approaches as well as the marginal estimator. We simulate the data by setting \( \varphi_{\beta W} = 0.3, h_p^2 = 0.4, n = n_w, \) and \( \omega \) ranging from 0.05 to 20. Moreover, we estimate \( \Sigma X \) and \( \Sigma Z \) using real genotype data from the 1000 Genome reference panel. Specifically, we randomly select one genomic region (bp 40–50m on chromosome one) and estimate \( \Sigma X \) and \( \Sigma Z \) separately from the 2000 genetic variants in two different populations. We consider two cases. In Case I, \( \Sigma X \) is estimated from European subjects and \( \Sigma Z \) is estimated from (East) Asian subjects. Case II represents the opposite situation, in which \( \Sigma X \) is estimated from Asian subjects and \( \Sigma Z \) is estimated from European subjects. In each of the two cases, we consider three reference panel options: (i) a reference panel matching the \( \Sigma X \), the Population-I GWAS population; (ii) a reference panel matching the \( \Sigma Z \), the Population-II GWAS population; and (iii) a mixed reference panel with equally-mixed Asian and European samples.

As illustrated in Figure 2, all the presented estimators are smaller than the true genetic correlation \( \varphi_{\beta W} \). Notably, reference panels that align with the Population-I GWAS generally yield better performance. For example, Case I depicts a situation where the genetic variant effects are estimated from a European cohort, and the genetically predicted values are constructed for an Asian population. Here, adopting a reference panel consistent with the European population greatly improve the estimation accuracy over \( G_{\beta W} \). The performance of the mixed reference panel mirrors that of the European one, while a panel matching the Asian LD pattern has a worse performance. There are two main takeaways from our exploration into the role of reference panels in trans-ancestry analysis. First, it may be beneficial to opt for a reference panel that reflects the LD structure of the Population-I GWAS. Second, although the reference panel-aided estimator \( G_{\beta W}^W \) can outperform the straightforward estimator \( G_{\beta W} \), the shrinkage may still exist in \( G_{\beta W}^W \). For a comprehensive overview of more numerical examples and in-depth discussions on various scenarios, please refer to the supplementary file and Supplementary Figure 3.

5. Simulation and Real Data Analysis

5.1. Simulated Genotype Data

We numerically evaluate our theoretical results in Theorems 1 and 2 by using simulated genotype datasets. We consider
two populations with \( n = p = 10,000, n_w = 5000, \) and \( n_z = 500. \) To generate the minor allele frequency (MAF) \( f \) in different populations, we first independently sample the overall MAF from Uniform \([0.05, 0.45]\) and generate \( F_{st} \) values from Uniform \([0.01, 0.04].\) Subsequently, the MAF of each variant for both populations is generated from the Balding-Nichols model (Balding and Nichols 1995). Then each entry of \( X_0, W_0, \) and \( Z_0 \) is independently generated from \([0, 1, 2]\) with probabilities \([(1 - f)^2, 2f(1 - f), f^2], \) respectively. We construct \( \Sigma_X \) and \( \Sigma_Z \) based on the first 10,000 genetic variants on chromosome 22, and they are estimated from 503 European and 504 East Asian samples in the 1000 Genome reference panel, respectively. To align with our model setups, we have standardized the genetic samples in the 1000 Genome reference panel, respectively. To conduct each scenario, we include \( \text{Refer}_{-} \text{Panel}_\text{Asian}, \text{GW}_{\beta\alpha} \) with European reference panel; \( \text{Refer}_{-} \text{Panel}_\text{European}, \text{GW}_{\beta\alpha} \) with Asian reference panel; \( \text{Refer}_{-} \text{Panel}_\text{Mixed}, \text{GW}_{\beta\alpha} \) with a reference panel having equally-mixed Asian and European samples; and Marginal, \( \text{GW}_{\beta\alpha}. \) The vertical line represents \( \omega = 1. \)

We simulate complex traits using model (1) with \( h^2_\beta = h^2_\alpha = 0.3 \) or 0.6. The proportion of variants with nonzero causal genetic effects is set to 0.05 or 0.3. The causal genetic effects in \( \beta \) and \( \alpha \) are sampled from normal distribution \( N(0, 1/p) \) with the true genetic correlation \( \varphi_{\beta\alpha} \) being 0, 0.45 or 0.9. We consider both uncorrected and corrected estimators of \( \varphi_{\beta\alpha}. \) The four naive (uncorrected) estimators of \( \varphi_{\beta\alpha} \) include (i) \( G_{\beta\alpha} \) (Marginal); (ii) \( G_{\beta\alpha}^W \) estimated by Ref-X (Ref-X); (iii) \( G_{\beta\alpha}^W \) estimated by Ref-Z (Ref-Z); and (iv) \( G_{\beta\alpha}^W \) estimated by Ref-Mixed (Ref-Mixed). The four corrected estimators \( \varphi_{\beta\alpha} \) include \( G_{\beta\alpha}^M \) and the three versions of \( G_{\beta\alpha}^M \). A total of 200 replications are conducted for each scenario.

For nonzero \( \varphi_{\beta\alpha}, \) the naive estimators of \( \varphi_{\beta\alpha} \) are all much smaller than \( \varphi_{\beta\alpha}, \) indicating substantial bias in the estimated genetic correlations (Supplementary Figure 4). As expected, the corrected estimators are very close to \( \varphi_{\beta\alpha} \) in all settings. We have also evaluated the performance of the corrected estimators without standardizing the genetic variant data \( (h^2_\beta = h^2_\alpha = 0.6, \varphi_{\beta\alpha} = 0.45, \text{and signal proportion} = 0.3). \) We find that the estimators remain unbiased when applied to genetic variant data in its original scale, providing flexibility in practical applications. These results strongly support our theoretical results, highlighting the importance of correcting for the downstream estimation bias induced by high-dimensional prediction. Since \( G_{\beta\alpha}^M \) performs very similarly to \( G_{\beta\alpha}^W \) and is easier to implement, we focus on \( G_{\beta\alpha}^M \) in later sections when analyzing large-scale real GWAS data.

5.2. UK Biobank Data Analysis

5.2.1. Implementation on Real Genotype Data

In this section, we calculate the corrected genetic correlation estimator \( G_{\beta\alpha}^M \) based on real genotype data. We download the UKB genotype data and apply the following standard quality control (QC) procedures: excluding subjects with more than 10% missing genotypes, only including SNPs with MAF > 0.01, genotyping rate > 90%, and passing Hardy-Weinberg test \((p\text{-value} > 1 \times 10^{-7}).\) After these QC steps, there are 461,488 genetic variants on 488,371 subjects. We focus on the unrelated White (European) and Asian subjects in our analysis (Bycroft et al. 2018), which are the top two largest ancestry groups in the UKB. The sample sizes of unrelated subjects are 366,335 and 8411 for White and Asian groups, respectively. Therefore, we treat the White individuals as the Population-I GWAS with a much smaller sample size, and the Asian individuals as the Population-II GWAS with a larger sample size.

The major difficulty of calculating \( G_{\beta\alpha}^M \) is to estimate \( b_1(\Sigma^2_X \Sigma_Z) \) and \( b_1(\Sigma_X \Sigma_Z). \) The high dimensionality of \( \Sigma_X \) and \( \Sigma_Z \) poses major challenges to estimating their functions, such as \( b_1(\Sigma_X \Sigma_Z) \) (Bickel and Levina 2008). The empirical patterns of LD in GWAS data have been shown to have a block diagonal structure: physically close genetic variants can be highly correlated, while genetic variants far from each other are typically independent. Thus, \( \Sigma_X \) and \( \Sigma_Z \) can be assumed to be banded.
covariance matrices. Based on this assumption, we perform a simultaneous block-diagonal approximation for the two LD structures from both populations. Specifically, we define trans-ancestry independent LD blocks between European and Asian populations. We start from the previous results in Berisa and Pickrell (2016), in which 1701 and 1445 independent LD blocks are defined in European and Asian populations, respectively. We then manually examine these LD blocks and merge them into $L = 253$ trans-ancestry independent LD blocks, which tend to have larger block sizes than population-specific LD blocks. The principle is that genetic variants in two different trans-ancestry blocks are independent in both populations, and the variants within the same block are correlated in at least one population.

We randomly select 8000 unrelated White British and Asian individuals to estimate $b_1(\Sigma_X^1 \Sigma_X^2)$ and $b_2(\Sigma_X^1 \Sigma_X^2)$. The detailed steps can be found in the supplementary file. We also estimate $b_3(\Sigma_X^1)$ and $b_4(\Sigma_X^1)$ in order to quantify the asymptotic shrinkage factor of $G_{\beta\alpha}$ in within-White analysis. Figure 3 presents $G_{\beta\alpha}$ values in White-Asian and within-White analyses based on our LD block approximation. In all settings, $G_{\beta\alpha}$ is much smaller than the underlying true genetic correlation $\varphi_{\beta\alpha} = 0.3$. When $\omega$ is large (say $> 10$), $G_{\beta\alpha}$ in within-White analysis is larger (therefore, has a smaller bias) than that in White-Asian analysis, indicating that the LD heterogeneity between UKB White and Asian populations may lead to smaller genetic correlation estimates in trans-ancestry analysis.

5.2.2. Simulation on Real Genotype Data

We next perform simulations to examine the corrected estimator $G_{\beta\alpha}^{GM}$ based on the trans-ancestry LD block approximation. Among the 366,335 unrelated White British subjects, we randomly select 350,000 or 50,000 as training GWAS samples. Then 1000 or 500 unrelated Asian subjects are randomly selected to construct genetically predicted values. The proportion of causal genetic variants is set to 0.001, 0.01, and 0.1, respectively. The causal variants are randomly selected and the nonzero genetic effects are independently derived from $N(0, 1/p)$ using the GCTA (Yang et al. 2011). We set $h^2_\beta = h^2_\alpha = 0.3$ and $\varphi_{\beta\alpha} = 0.25$; $h^2_\beta = h^2_\alpha = 0.6$ and $\varphi_{\beta\alpha} = 0.5$; or $h^2_\beta = h^2_\alpha = 0.6$ and $\varphi_{\beta\alpha} = 0$. By using the summary statistics from the training GWAS via fastGWA (Jiang et al. 2019), we generate genetically-predicted traits in Asian individuals. We estimate $G_{\beta\alpha}$ and $G_{\beta\alpha}^{GM}$ for each simulated dataset. Each simulation setting is replicated 500 times.

Supplementary Figure 5 shows that the naive estimator $G_{\beta\alpha}$ is much smaller than $\varphi_{\beta\alpha}$ and their gap depends on the training GWAS sample size. For example, when $n_x = 350,000$ and $\varphi_{\beta\alpha} = 0.25$, the mean of $G_{\beta\alpha}$ is 0.041 across all sparsity levels and testing data sample size $n_z$ (standard error = 0.046). These results show that the estimated genetic correlation in this setting is about 5 times smaller than the true genetic correlation. Similarly, for $n_x = 350,000$ and $\varphi_{\beta\alpha} = 0.5$, the mean of $G_{\beta\alpha}$ is 0.128 (standard error = 0.055), which is about 3 times smaller than the true value. The corrected estimator $G_{\beta\alpha}^{GM}$ is much closer to the $\varphi_{\beta\alpha}$ in all settings, with the mean being 0.232 (standard error = 0.259) for $\varphi_{\beta\alpha} = 0.25$ and 0.495 (standard error = 0.211) for $\varphi_{\beta\alpha} = 0.5$. Similar results are observed for the $n = 50,000$ cases. Furthermore, we observe that the performance of the estimator remains consistent when the sample size of genetic variant data used to calculate the moments reduces from 8000 to 5000, 2500, and 500. Our estimator also performs well when estimating the moments using the data from 503 European and 504 Asian subjects in the 1000 Genomes reference panel. Overall, these simulation results show that the trans-ancestry LD block approximation approach performs well with $G_{\beta\alpha}^{GM}$ substantially outperforming $G_{\beta\alpha}$ in real genotype data.

5.2.3. Comparing with Existing Methods

Furthermore, we compare our estimator with existing methods for trans-ancestry genetic correlation, including Popcorn (Brown et al. 2016) and XPASS (Cai et al. 2021). To ensure a substantial overlap with the reference panel and pretrained data used by these methods, we focus on the commonly used HapMap3 variants ($p = 990,761$) available in the UKB imputed genetic data. We apply standard QC steps similar to those in Section 5.2.1. Details can be found in the supplementary file. In our major analysis, we randomly select 350,000 unrelated White British subjects as training GWAS samples, while the sample size of the Asian dataset is set to be 5000, 1000, 500, or 200. We consider $\varphi_{\beta\alpha} = 0$, 0.25, 0.5, and 0.9, with heritability ranging from 0.2 to 0.6. Other simulation setups are similar to those in Section 5.2.2 and are provided in the supplementary information. In total, there are 48 major simulation setups and each setup is repeated 500 times.

Simulation results are presented in Supplementary Figures 6–7. When $\varphi_{\beta\alpha} = 0$, all three estimators provide unbiased estimates, and their empirical standard errors are comparable and increase as the Asian sample size $n_z$ decreases. For nonzero $\varphi_{\beta\alpha}$, we observe that XPASS tends to overestimate the genetic correlation, while Popcorn underestimates it. As $\varphi_{\beta\alpha}$ increases, the bias in both methods also increases. In contrast, the proposed estimator $G_{\beta\alpha}^{GM}$ remains unbiased across all cases. For example, when $\varphi_{\beta\alpha} = 0.5$ and $n_x = 5000$, the mean of estimates for Popcorn, XPASS, and $G_{\beta\alpha}^{GM}$ is 0.359 (standard error = 0.097), 0.546 (standard error = 0.093), and 0.498 (standard error = 0.075).
respectively. The performance patterns observed in our major analysis remain consistent when we reduce the training GWAS sample size from 350,000 to 50,000 and reduce the sample size used to estimate the moments required by $G^{\beta\alpha}_M$. In summary, the proposed estimator demonstrates better performance than the existing methods in scenarios with unbalanced data resources, where one of the two genetic datasets has a small sample size.

### 5.3. Real Data Applications

To evaluate the finite sample performance of $G^{\beta\alpha}_M$, we consider 26 complex traits from different trait domains in the UKB study, similar to those used in Kichaev et al. (2019). The training GWAS is performed on genotype data of the unrelated White British subjects. We use linear models implemented in fastGWA (Jiang et al. 2019) for this analysis. The adjusted covariates include the top 20 genetic principal components, age, sex, age-squared, age-sex interaction, and age-squared-sex interaction. After sub-setting to subjects with complete data of genetic variants, covariates, and phenotypes, the average sample size per trait is 293,953. We construct the genetically predicted values based on two independent UKB datasets. The first is a set of White but non-British subjects (average sample size 15,605) and the second is a group of Asian subjects (average sample size 6491). White non-British and White British groups are known to have similar LD patterns. Accordingly, the White non-British analysis can be viewed as a positive control example, where the underlying genetic correlation is expected to be close to one for every pair of traits. The $G^{\beta\alpha}_M$ is estimated according to model (1), while adjusting for the same set of covariates as in the training GWAS data. Then, $G^{\beta\alpha}_M$ is estimated by plugging the per-trait training GWAS sample size, the heritability estimated using the GREML method from GCTA (Yang et al. 2011), and the LD-related functions estimated in within-White and White-Asian analyses detailed in Section 5.2.

The data analysis results are summarized in Figure 4, Supplementary Figure 8, and Supplementary Tables 1 and 2. In the White non-British analysis, $G^{\beta\alpha}_M$ ranges from 0.065 to 0.211 with mean = 0.142 across the 26 complex traits, all of which have significant $T$-test $p$-values after controlling the false discovery rate (FDR) at 5% level ($P < 2.89 \times 10^{-08}$). The results clearly demonstrate the significant genetic influences on these complex traits. The estimated genetic correlations, however, are much smaller than one. We then correct the genetic correlations and calculate $G^{\beta\alpha}_M$ for each trait pair. The average $G^{\beta\alpha}_M$ of the 26 complex traits is 0.966 (range = [0.650, 1.320]). A genetic correlation close to one is expected in this positive control analysis, indicating the high genetic similarity between White British and White non-British populations. Our results confirm our theoretical analysis and provide strong evidence that our proposed estimator can accurately reflect the underlying genetic similarity. Furthermore, it indicates that the widely reported naive estimator $G^{\beta\alpha}$ in the literature, although suggesting significant genetic controls, might underestimate the shared genetic co-influences between two traits. Next, in the White Asian analysis, the average $G^{\beta\alpha}_M$ is 0.10 (range = [0.014, 0.219]), 24 of which pass the FDR control at
In contrast to the majority of previous literature (e.g., Bulik-Sullivan et al. 2015b), XPASS (Cai et al. 2021), and Popcorn (Brown et al. 2016), alongside our proposed estimator, have occasionally produced point estimates for genetic correlation that breach the theoretical \([-1,1]\) boundary. This is particularly evident when the underlying true genetic correlation approaches the limits of this range. For example, when two populations exhibit high genetic similarity for a specific trait, the true genetic correlation might be just below 1. However, due to the inevitable uncertainties and sample variations inherent to the estimation process, some point estimates can, in practice, exceed this limit. While simplistic solutions, such as truncation to enforce estimate boundaries, might seem tempting, they risk introducing bias and can potentially underestimate the real genetic correlation. In response to this challenge, we have embedded uncertainty quantification within our estimators, harnessing resampling-based confidence intervals. We recommend a heightened awareness of the uncertainties intrinsic to point estimates arising from sample variance and using confidence intervals when interpreting the results.

A few interesting problems can be further explored in trans-ancestry analyses. First, when two populations share a low level of genetic similarity for a trait inherited in both of the populations, it would be of great interest to identify the specific loci contributing to the genetic differences between the two populations. For example, it is helpful to identify the genomic regions where the genetic effects are not zero in two populations, but with heterogeneous effect sizes. Second, trans-ancestry genetic correlation may serve as a window into the genetic similarity of traits across diverse populations. Such information can be incorporated in transfer learning techniques, facilitating the amalgamation of multiple datasets (Li et al. 2022b) or the establishment of auxiliary information (Li et al. 2022a). Moreover, the methods presented in this article are centered around linear models and should only be applied to continuous traits under the assumption of homoscedasticity. Future research could expand the random matrix-based analysis to encompass binary outcomes, ordinal variables, and time-to-event data. Lastly, our rigorous examination of LD heterogeneity serves as an example of the profound influence of the covariance matrix’s heterogeneity on high-dimensional data estimation and prediction. The influence of data distribution shift due to covariance structure differences can be further examined in various data types in future studies (Koh et al. 2021).
Disclosure Statement

The authors report there are no competing interests to declare.

Funding

Research reported in this publication was supported by the National Institute On Aging of the National Institutes of Health under Award Number RF1AG082938. This research has been conducted using the UK Biobank resource (application number 76139), subject to a data transfer agreement.

References

1000-Genomes-Consortium. (2015), "A Global Reference for Human Genetic Variation," Nature, 526, 68–74. [843]
Bai, Z., and Silverstein, J. W. (2010), Spectral Analysis of Large Dimensional Random Matrices (Vol. 20), New York: Springer. [842]
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. [845]
Barth, D., Papageorge, N. W., and Thom, K. (2020), "Genetic Endowments and Wealth Inequality," Journal of Political Economy, 128, 1474–1522. [842]
Berisa, T., and Pickrell, J. K. (2016), "Approximately Independent Linkage Disequilibrium Blocks in Human Populations," Bioinformatics, 32, 283–285. [841,846]
Bickel, P. J., and Levina, E. (2008), "Regularized Estimation of Large Covariance Matrices," The Annals of Statistics, 36, 199–227. [845]
Brown, B. C., Ye, C. J., Price, A. L., Zaitlen, N., Consortium, A. G. E. N. T. D., et al. (2016), "Transethnic Genetic-Correlation Estimates from Summary Statistics," The American Journal of Human Genetics, 99, 76–88. [839,841,846,848]
Bulik-Sullivan, B. K., 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. (2015a), "An Atlas of Genetic Correlations Across Human Diseases and Traits," Nature Genetics, 47, 1236–1241. [839,841,848]
Bulik-Sullivan, B. K., Loh, P. R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., Neale, B. M., of the Psychiatric Genomics Consortium, S. W. G., et al. (2015b), "LD Score Regression Distinguishes Confounding From Polygenicity in Genome-Wide Association Studies," Nature Genetics, 47, 291–295. [848]
Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, I., Sharp, K., Motyer, A., Vuckevic, D., Delaneau, O., O’Connell, J., et al. (2018), "The UK Biobank Resource with Deep Phenotyping and Genomic Data," Nature, 562, 203–209. [840,845]
Cai, M., Xiao, J., Zhang, S., Wan, X., Zhao, H., Chen, G., and Yang, C. (2021), "A Unified Framework for Cross-Population Trait Prediction by Leveraging the Genetic Correlation of Polygenic Traits," The American Journal of Human Genetics, 108, 632–655. [839,846,848]
Chen, C., and Zissimopoulos, J. M. (2018), "Racial and Ethnic Differences in Trends in Dementia Prevalence and Risk Factors in the United States," Alzheimer’s & Dementia: Translational Research & Clinical Interventions, 4, 510–520. [839]
Codd, V., Wang, Q., Allara, E., Musicha, C., Kaptego, S., Stoma, S., Jiang, T., Hamby, S. E., Braund, P. S., Bountziouka, V., et al. (2021), "Polygenic Basis and Biomedical Consequences of Telomere Length Variation," Nature Genetics, 53, 1425–1433. [842]
Van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H., and Wray, N. R. (2019), "Genetic Correlations of Polygenic Disease Traits: From Theory to Practice," *Nature Reviews Genetics*, 20, 567–581. [839]

Wang, J., and Li, H. (2022), "Estimation of Genetic Correlation with Summary Association Statistics," *Biometrika*, 109, 421–438. [839,841]

Weissbrod, O., Kanai, M., Shi, H., Gazal, S., Peyrot, W. J., Khera, A. V., Okada, Y., Martin, A. R., Finucane, H. K., et al. (2022), "Leveraging Fine-Mapping and Multipopulation Training Data to Improve Cross-Population Polygenic Risk Scores," *Nature Genetics*, 54, 450–458. [842]

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011), "Gcta: A Tool for Genome-Wide Complex Trait Analysis," *The American Journal of Human Genetics*, 88, 76–82. [841,842,846,847]

Zhang, Y., Cheng, Y., Ye, Y., Jiang, W., Lu, Q., and Zhao, H. (2021), "Estimating Genetic Correlation Jointly Using Individual-Level and Summary-Level gwas Data," bioRxiv, pp. 456908. [839]

Zhao, B., and Zhu, H. (2022), "On Genetic Correlation Estimation with Summary Statistics from Genome-Wide Association Studies," *Journal of the American Statistical Association*, 117, 1–11. [842]