Genetics and population analysis

XPXP: Improving polygenic prediction by cross-population and cross-phenotype analysis

Jiashun Xiao [1,2,#], Mingxuan Cai [1,2,#], Xianghong Hu [1,2], Xiang Wan [3,4,*], Gang Chen [5,*], and Can Yang [1,2,*]

[1]Guangzhou HKUST Fok Ying Tung Research Institute, Guangzhou 511458, China. [2]Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong SAR, China. [3]Shenzhen Research Institute of Big Data, Shenzhen 518172, China. [4]Pazhou Lab, Guangzhou, 510330, China. [5]Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, 410083, China. [#]These authors contributed equally to this work

[*]To whom correspondence should be addressed.

Abstract

Motivation: As increasing sample sizes from genome-wide association studies (GWASs), polygenic risk scores (PRSs) have shown great potential in personalized medicine with disease risk prediction, prevention and treatment. However, the PRS constructed using European samples becomes less accurate when it is applied to individuals from non-European populations. It is an urgent task to improve the accuracy of PRSs in under-represented populations, such as African populations and East Asian populations.

Results: In this paper, we propose a cross-population and cross-phenotype (XPXP) method for construction of PRSs in under-represented populations. XPXP can construct accurate PRSs by leveraging biobank-scale datasets in European populations and multiple GWASs of genetically correlated phenotypes. XPXP also allows to incorporate population-specific and phenotype-specific effects, and thus further improves the accuracy of PRS. Through comprehensive simulation studies and real data analysis, we demonstrated that our XPXP outperformed existing PRS approaches. We showed that the height PRSs constructed by XPXP achieved 9% and 18% improvement over the runner-up method in terms of predicted R^2 in East Asian and African populations, respectively. We also showed that XPXP substantially improved the stratification ability in identifying individuals at high genetic risk of Type 2 Diabetes.

Availability: The XPXP software and all analysis code are available at github.com/YangLabHKUST/XPXP

Contact: wanxiang@sribd.cn, chengangcs@gmail.com, or macyang@ust.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genome-wide association studies (GWASs) have been widely conducted to understand the genetic basis of complex traits/diseases. As of August, 2021, more than 270,000 genome-wide significant associations (p-value$\leq 5 \times 10^{-8}$) have been identified between the single nucleotide polymorphisms (SNPs) and complex traits (https://www.ebi.ac.uk/gwas/). Investigations into these findings have revealed that most complex traits are affected by many genetic variants with small effect sizes, which is referred to as "polygenicity". Given this fact, polygenic risk scores (PRS) constructed by a collective contribution of SNPs across the genome from GWAS has shown great potential in personal and clinical utility for a number of heritable diseases [Torkamani, 2018], including aid diagnosis by stratifying patients into different risk groups, early and cost-effective interventions, and improved therapeutic strategies [Chatterjee, 2016, Khera, 2018]. As an example, a recent PRS model for type 2 diabetes (T2D) was trained on more than 600,000 European participants from a consumer genetic database. The constructed PRS achieved the area under the receiver operator curve (AUC) as high as 0.65 for individuals from European ancestry [Multhaup, 2019]. In terms of odds ratio, participants with the top 5% PRS have a three-fold increased risk of T2D. However, the AUC of this PRS model reduces to 0.57 when applying to individuals from African ancestries, suggesting limited transferability of PRS across populations as reported in recent studies [Lam, 2019, Martin, 2019].

Owing to the fact that 89% of GWAS participants to date are from European ancestry, earlier constructed PRSs tend to be biased to the over-represented European populations [Mills and Rahal, 2020]. Recently, much efforts have been devoted to improving the prediction power of

PRSs in non-European populations by incorporating information from large-scale European datasets. Several methods have been developed to account for heterogeneous genetic architectures across populations, including differences in linkage disequilibrium (LD) patterns and minor allele frequencies (MAF). To name a few, MultiPRS [Márquez-Luna, 2017] linearly combines PRSs from multiple populations using mixing weights estimated by cross-validation. Linear mixed models (LMM) are widely used in GWAS and the estimated SNP effects are known as the best unbiased linear predictor (BLUP) under the random-effects assumption [Lee, 2011]. XP-BLUP [Coram, 2017] assumes significant SNPs identified in a large-scale auxiliary population (e.g., European population) play an important role for PRS construction in the under-represented populations. It uses an extra component in LMM to characterize pre-selected significant SNPs from well-powered GWAS of the auxiliary population. Bivariate BLUP (bvBLUP) implemented in the GCTA software [Lee, 2011] uses the standard bivariate linear mixed model to characterize SNP effect sizes of two traits and their genetic correlation. However, the current implementation is too memory-consuming to be applicable for biobank-scale GWAS data analysis. By the novel data structure and algorithm design, XPA [Cai, 2021] can leverage biobank-scale data to construct PRSs for the under-represented population. Very recently, Huang [2021] proposed PRS-CSx to jointly models GWAS summary statistics from multiple populations. By introducing a continuous shrinkage prior, PRS-CSx can model the shared genetic effects for causal variants across populations while allowing SNPs effect sizes to vary across populations.

Despite the above advances in constructing cross-population PRS, existing approaches primarily focus on a single phenotype at a time. However, genetic variants have been commonly found to affect multiple phenotypes, which is known as pleiotropy [Solovieff, 2013, Yang, 2015]. In a systematic analysis of 4,155 publicly available GWASs, 90% trait-associated loci showed pleiotropic effects [Watanabe, 2019]. These loci can have correlated effect sizes on multiple phenotypes and induce their correlation [Van Rheenen, 2019, Guo, 2021]. In fact, genetic correlation estimated from GWASs also reveals the ubiquity of pleiotropy in complex human traits/diseases. Well-known examples include strong genetic correlation (0.68, s.e.=0.04) between bipolar disorder and schizophrenia [Lee, 2013], and moderate genetic correlation (0.4, s.e.=0.04) between T2D and body mass index (BMI) [Zheng, 2017]. Therefore, constructing PRS cross multiple correlated phenotypes can further improve risk prediction by leveraging their shared genetic basis [Li *et al.*, 2014]. Actually, the benefits of multi-phenotype over a single phenotype have been well documented in the PRS analysis, including MTGBLUP [Maier, 2015], PleioPred [Hu, 2017], and SMTpred [Maier, 2018]). However, these PRS methods mainly focus on the datasets of European population. It remains unclear how much genetic information can be transferred from large-scale European datasets to under-represented non-European populations when combining multi-phenotypes for constructing PRSs [Martin, 2019, Cai, 2021].

Here we propose a unified method to construct PRS by cross-population and cross-phenotype (XPXP) analysis. The keys to the success of XPXP are threefold. First, it can greatly improve PRS of the target population by making use of biobank-scale datasets of the auxiliary population through trans-ancestry genetic correlation. Second, it can exploit genetic correlation among multiple phenotypes within the same population by leveraging pleiotropy. Third, it allows to incorporate phenotype-specific or population-specific genetic effects to improve the accuracy of PRS. XPXP is widely applicable because it only requires the summary statistics of multiple GWASs as its input. In terms of prediction accuracy, we first demonstrated that XPXP can outperform existing PRS approaches through comprehensive simulation studies. Then we applied XPXP to construct PRS for height of individuals from African ancestry by integrating a relatively small-scale African training dataset (about 7K participants) with the GIANT [Wood, 2014], the UK-Biobank (UKBB) and BioBank Japan (BBJ) datasets. Based on an independent African testing dataset (1K participants), we showed that XPXP achieved a substantial improvement of prediction accuracy in terms of R^2. To demonstrate the generality of our methods, we further applied XPXP to construct PRS of height for East Asian populations. By integrating multiple datasets from the BBJ and UKBB datasets, we showed that the PRS constructed by XPXP achieved 9% accuracy gain in terms of predicted R^2 compared to the runner-up. We also showed that XPXP substantially improved predictive power in identifying high-risk groups for T2D when integrating GWASs of multi-phenotypes from both BBJ and UKBB datasets, highlighting the value of well-powered European GWASs and shared genetic basis among correlated phenotypes in constructing PRSs for individuals of non-Europeans ancestries.

2 Materials and methods

2.1 The XPXP model

We consider observations from two populations: the target population (e.g., East Asian population) often has a limited number of GWAS samples while the auxiliary population (e.g., European population) has collected biobank-scale GWAS data. Let $\{(\mathbf{Z}_k, \mathbf{G}_k, \mathbf{y}_k), k = 1, .., K\}$ be the collected datasets in the target population, where $\mathbf{Z}_k$ is an $n_k \times c_k$ matrix collecting all covariates for the k-th phenotype, $\mathbf{G}_k$ is an $n_k \times p$ matrix containing p genotypes of n_k samples, $\mathbf{y}_k$ is an $n_k \times 1$ vector of corresponding phenotypic values, and K is the number of phenotypes under consideration. Without loss of generality, we assume that $\mathbf{y}_k$ has been standardized to have zero mean and unit variance. Similarly, we consider the collected datasets $\{(\mathbf{Z}'_m, \mathbf{G}'_m, \mathbf{y}'_m), m = 1, ..., M\}$ from the auxiliary population, where $\mathbf{Z}'_m$ is the $n'_m \times c'_m$ covariate matrix, $\mathbf{G}'_m$ is the $n'_m \times p$ genotype matrix, $\mathbf{y}'_m$ is the $n'_m \times 1$ phenotype vector, and M is the number of phenotypes from the auxiliary population.

Now we consider modeling the relationship between genotypes and phenotypes in the cross-population setting. To reconcile the difference of allele frequencies in the two populations, XPXP works with standardized genotype matrices by assuming that the SNP effect sizes in both populations increase as the allele frequencies decrease [Speed, 2012]. Let $\mathbf{g}_{kj} \in \mathbb{R}^{n_k}$ and $\mathbf{g}'_{mj} \in \mathbb{R}^{n'_m}$ denote the j-th column of $\mathbf{G}_k$ and $\mathbf{G}'_m$, respectively. The corresponding column means and standard deviations are given as $\bar{\mathbf{g}}_{kj}$ and $\bar{\mathbf{g}'}_{mj}$, s_{kj} and s'_{mj}, respectively. Then we have standardized genotype matrices as $\mathbf{X}_k = [\mathbf{x}_{k1}, \mathbf{x}_{k2}, \ldots, \mathbf{x}_{kp}] \in \mathbb{R}^{n_k \times p}$ and $\mathbf{X}'_m = [\mathbf{x}'_{m1}, \mathbf{x}'_{m2}, \ldots, \mathbf{x}'_{mp}] \in \mathbb{R}^{n'_m \times p}$, where the j-th column of $\mathbf{X}_k$ and $\mathbf{X}'_m$ is given as $\mathbf{x}_{kj} = (\mathbf{g}_{kj} - \bar{\mathbf{g}}_{kj})/(s_{kj}\sqrt{p})$ and $\mathbf{x}'_{mj} = (\mathbf{g}'_{mj} - \bar{\mathbf{g}'}_{mj})/(s'_{mj}\sqrt{p})$, respectively. Clearly, each column of $\mathbf{X}_k$ and $\mathbf{X}'_m$ has mean 0 and variance $1/p$. We begin with the following linear mixed models to relate genotypes and phenotypes:

$$\begin{aligned} \mathbf{y}_k &= \mathbf{Z}_k\boldsymbol{\omega}_k + \mathbf{X}_k\boldsymbol{\beta}_k + \boldsymbol{\epsilon}_k,\ k = 1, \ldots, K, \\ \mathbf{y}'_m &= \mathbf{Z}'_m\boldsymbol{\omega}'_m + \mathbf{X}'_m\boldsymbol{\beta}'_m + \boldsymbol{\epsilon}'_m,\ m = 1, \ldots, M, \end{aligned} \tag{1}$$

where $\boldsymbol{\omega}_k \in \mathbb{R}^{c_k}$ and $\boldsymbol{\omega}'_m \in \mathbb{R}^{c'_m}$ are the vectors of fixed effects, $\boldsymbol{\beta}_k \sim \mathcal{N}(\mathbf{0}, h^2_k\mathbf{I}_p)$ and $\boldsymbol{\beta}'_m \sim \mathcal{N}(\mathbf{0}, h^2_m\mathbf{I}_p)$ are the vectors of random effects, $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, (1-h^2_k)\mathbf{I}_{n_k})$ and $\boldsymbol{\epsilon}'_m \sim \mathcal{N}(\mathbf{0}, (1-h^2_{m'})\mathbf{I}_{n'_m})$ are the noise terms, and h^2_k and h^2_m are known as heritabilities of the k-th trait and the m-th trait in the target population and the auxiliary population, respectively. Note that we have implicitly imposed that the SNP effect sizes increases as the allele frequencies decrease at the rate of $1/\sqrt{2f(1-f)}$, where f is the minor allele frequency [Speed, 2012, Cai, 2021]. For easier introduction of the key idea but without loss of generality, we present XPXP by using two phenotypes in both target and auxiliary populations,

i.e., $K = 2$ and $M = 2$. Now we can explicitly write down Eq. (1) as

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{Z}_1\boldsymbol{\omega}_1 + \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1, \quad \mathbf{y}_2 = \mathbf{Z}_2\boldsymbol{\omega}_2 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2, \\ \mathbf{y}'_1 &= \mathbf{Z}'_1\boldsymbol{\omega}'_1 + \mathbf{X}'_1\boldsymbol{\beta}'_1 + \boldsymbol{\epsilon}'_1, \quad \mathbf{y}'_2 = \mathbf{Z}'_2\boldsymbol{\omega}'_2 + \mathbf{X}'_2\boldsymbol{\beta}'_2 + \boldsymbol{\epsilon}'_2. \end{aligned} \tag{2}$$

A major difficulty for constructing accurate PRS in the underrepresented population is the limited sample size. XPXP is able to improve the accuracy of PRSs by leveraging genetic correlations from three sources simultaneously. First, trans-ancestry genetic correlation of the same phenotype (corr($\boldsymbol{\beta}_1$,$\boldsymbol{\beta}'_1$)) can be strong. When $n_1 \ll n'_1$, information from biobank-scale data in the auxiliary population can be leveraged through a strong trans-ancestry genetic correlation. Second, genetic correlation of two phenotypes within the target population (corr($\boldsymbol{\beta}_1$,$\boldsymbol{\beta}_2$)) can be also very useful when the environmental correlation has been appropriately accounted for. Third, there may exist a distinct phenotype in the auxiliary population (e.g., $\mathbf{y}'_2$) genetically correlated with $\mathbf{y}_1$. We refer corr($\boldsymbol{\beta}_1$,$\boldsymbol{\beta}'_2$) as trans-ancestry genetic correlation of two distinct phenotypes.

To model the above genetic correlation of effect sizes, we introduce the following probabilistic structures on $\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\}$ and $\{\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2\}$:

$$\begin{aligned} \boldsymbol{\beta} &= [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2] \sim \mathcal{MN}(\mathbf{0}, \mathbf{I}_p, \boldsymbol{\Sigma}_\beta), \\ \boldsymbol{\Sigma}_\beta &= \begin{bmatrix} h_1^2 & h_{12} & h_{11'} & h_{12'} \\ h_{12} & h_2^2 & h_{21'} & h_{22'} \\ h_{11'} & h_{21'} & h_{1'}^2 & h_{1'2'} \\ h_{12'} & h_{22'} & h_{1'2'} & h_{2'}^2 \end{bmatrix}, \end{aligned} \tag{3}$$

where $\mathcal{MN}(\mathbf{A}, \mathbf{B}, \mathbf{C})$ denotes the matrix normal distribution. In Eq. (3), we assume that rows of $\boldsymbol{\beta}$ are uncorrelated but columns of $\boldsymbol{\beta}$ can be correlated with covariance matrix $\boldsymbol{\Sigma}_\beta$. The diagonal elements of $\boldsymbol{\Sigma}_\beta$ are the heritabilities of corresponding phenotypes, and the off-diagonal elements include the cross-trait co-heritability within population (h_{12} and $h_{1'2'}$), the cross-population co-heritability of the same trait ($h_{11'}$ and $h_{22'}$), and cross-population cross-trait co-heritability ($h_{12'}$ and $h_{21'}$). As the phenotype vectors are assumed to be standardized with unit variance, these co-heritabilities can be also called genetic covariances. From the statistical point of view, XPXP can improve the accuracy of PRS because it can borrow information from both within-population and cross-population large-scale GWASs through nonzero genetic correlation and yield more accurate posterior mean of $\boldsymbol{\beta}$ for PRS construction.

To model GWAS data from the same population, it is also very important to account for the residual correlation within the same population, i.e., corr($\boldsymbol{\epsilon}_1$, $\boldsymbol{\epsilon}_2$) and corr($\boldsymbol{\epsilon}'_1$, $\boldsymbol{\epsilon}'_2$). Such a correlation could be attributed to the common environmental factors introduced by sample overlapping [Bulik-Sullivan, 2015a, Gao, 2021], so it is also referred to as the environmental correlation. With these consideration, we assume the following covariance on $\{\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2\}$ and $\{\boldsymbol{\epsilon}'_1, \boldsymbol{\epsilon}'_2\}$ for overlapped samples:

$$cov(\boldsymbol{\epsilon}_{1i}, \boldsymbol{\epsilon}_{2i}) = \rho^{\epsilon}_{12},\ cov(\boldsymbol{\epsilon}'_{1i'}, \boldsymbol{\epsilon}'_{2i'}) = \rho^{\epsilon}_{1'2'}, \tag{4}$$

where $i = 1, ..., n_s$, $i' = 1, ..., n'_s$, n_s and n'_s are the number of overlapped individuals in the target and auxiliary population, respectively. For non-overlapped individuals, we simply assume their environmental correlation is zero.

So far, we have modeled shared genetic information of multiple phenotypes and accounted for the environmental correlation within population. Now we consider incorporating population-specific and phenotype-specific information into XPXP. In the target population, let $\mathbf{X}_{l1} \in \mathbb{R}^{n_1 \times l_1}$ and $\mathbf{X}_{l2} \in \mathbb{R}^{n_2 \times l_2}$ be the standardized genotype matrices of SNPs with large effects corresponding to $\mathbf{y}_1$ and $\mathbf{y}_2$, respectively. In the auxiliary population, we denote the standardized genotype matrices as $\mathbf{X}'_{l1} \in \mathbb{R}^{n'_1 \times l'_1}$ and $\mathbf{X}'_{l2'} \in \mathbb{R}^{n'_2 \times l'_2}$ for $\mathbf{y}'_1$ and $\mathbf{y}'_2$, respectively. SNPs in these genotype matrices are selected based on their p-values from SNP-phenotype association tests, which are available in the GWAS summary statistics. For example, we select SNPs in $\mathbf{X}_{l1}$ if their p-values are smaller than a given p-value threshold, e.g., 1×10^{-6}, and then apply LD pruning to ensure that they are nearly independent. In practice, the number of selected SNPs will be much smaller than the sample size, i.e., $l_1 \ll n_1$. We apply the same procedure to select SNPs in $\mathbf{X}_{l2}$, $\mathbf{X}'_{l1}$ and $\mathbf{X}'_{l2}$. In such a way, SNPs in $\mathbf{X}_{l1}$ $\mathbf{X}_{l2}$, $\mathbf{X}'_{l1}$ and $\mathbf{X}'_{l2}$ carry large population-specific effects corresponding to $\mathbf{y}_1$, $\mathbf{y}_2$, $\mathbf{y}'_1$, and $\mathbf{y}'_2$, respectively. Now we modify model (2) to incorporate population-specific and phenotype-specific effects as

$$\begin{aligned} \mathbf{y}_k &= \mathbf{Z}_k\boldsymbol{\omega}_k + \mathbf{X}_{lk}\boldsymbol{\gamma}_k + \mathbf{X}_k\boldsymbol{\beta}_k + \boldsymbol{\epsilon}_k, k = 1, 2, \\ \mathbf{y}'_m &= \mathbf{Z}'_m\boldsymbol{\omega}'_m + \mathbf{X}'_{lm}\boldsymbol{\gamma}'_m + \mathbf{X}'_m\boldsymbol{\beta}'_m + \boldsymbol{\epsilon}'_m, m = 1, 2, \end{aligned} \tag{5}$$

where $\boldsymbol{\gamma}_k \in \mathbb{R}^{l_k}$ is the vector of fixed effects corresponding to $\mathbf{y}_k$ in the target population, $\boldsymbol{\gamma}'_m \in \mathbb{R}^{l'_m}$ is the vector of fixed effects corresponding to $\mathbf{y}'_m$ in the auxiliary population.

In summary, XPXP uses the random effects $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2]^T$ to capture the shared polygenic effects across phenotypes populations, and uses fixed effects $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}'_1, \boldsymbol{\gamma}'_2]^T$ to characterize population-specific and phenotype-specific large effects. XPXP also accounts for the environmental correlations, corr($\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2$) and corr($\boldsymbol{\epsilon}'_1, \boldsymbol{\epsilon}'_2$) introduced by the sample overlap. Although we present XPXP by setting $K = 2$ and $M = 2$, XPXP is generally applicable for $K \geq 2$ and $M \geq 2$ as we demonstrate in real data analysis.

We have formulated our model using individual-level genotypes and phenotypes. However, XPXP only requires GWAS summary statistics and LD matrices computed by external reference panels to build a PRS model. To illustrate this, we dissect the procedure of PRS construction into two steps: estimation of model parameters and computation of SNPs effect size. In each step, we show how to bypass the computation involving the individual-level data by using the GWAS summary statistics and LD reference panels.

2.2 Parameter estimation using GWAS summary statistics

To obtain the posterior mean of $\boldsymbol{\beta}$ and the analytic solution of $\boldsymbol{\gamma}$, we first need to estimate the unknown parameters in $\boldsymbol{\Sigma}_\beta$ and environmental covariance if sample overlap exists. We divide those parameters into two groups: within-population and cross-population. The former group includes heritabilities $\{h_1^2, h_2^2, h_{1'}^2, h_{2'}^2\}$, genetic covariance of two phenotypes within a population $\{h_{12}, h_{1'2'}\}$, and environmental covariance $\{\rho^{\epsilon}_{12}, \rho^{\epsilon}_{1'2'}\}$. The later group includes trans-ancestry genetic covariance $\{h_{11'}, h_{12'}, h_{21'}, h_{22'}\}$. Instead of handling all $K + M$ traits simultaneously, we apply a pair-wise strategy to estimate one unknown covariance at a time. This pair-wise analysis is guaranteed to give consistent results based on the composite likelihood approach [Varin, 2011, Ming, 2020].

Regarding the within-population group, we consider parameter estimation for a pair of phenotypes in the target population as an example. LD score regression (LDSC) is a widely used method to estimate heritability and co-heritability for GWAS within a single population [Bulik-Sullivan, 2015a,b]. LDSC assumes a random-effect model to characterize polygenic genetic architecture of complex traits. Under the assumptions of LDSC, the polygenic effects can be tagged by LD and the confounding factors (e.g., population stratification or cryptic relatedness) are uncorrelated with LD. Specifically, the LD tagging effect implies that the expected squared z-score of SNP j is proportional to its LD score $l_j = \sum_s r_{js}^2$, where r_{js} is the correlation between SNP j and SNP s.

This relationship is precisely given by the following equations:

$$\mathrm{E}\left[z_{kj}^2\right] = \frac{n_k h_k^2}{p} l_j + 1 + a_k n_k,\ k = 1, 2, \tag{6}$$

where z_{kj} is z-score of j-th SNP for phenotype k, l_j is the LD score for j-th SNP, which can be obtained from publicly available LD reference panel (e.g. The 1000 Genomes Project), a_k is inflation constants due to confounding bias. By regressing the observed squared z-scores to the LD score, we can obtain the estimated heritabilities $\hat{h}_1^2$ and $\hat{h}_2^2$. Similarly, the genetic covariance h_{12} can be estimated using the following relationship [Bulik-Sullivan, 2015a]:

$$\mathbb{E}\left[z_{1j} z_{2j}\right] = \frac{\sqrt{n_1 n_2} h_{12}}{p} l_j + \frac{n_s \rho_{12}^{pheno}}{\sqrt{n_1 n_2}}, \tag{7}$$

where n_s is the number of overlapped individuals in both studies, $\rho_{12}^{pheno} = h_{12} + \rho_{12}^{\epsilon}$ is the phenotypic correlation among the overlapping samples. By regressing the product of z_{1j} and z_{2j} to the LD score l_j, we can obtain the estimated co-heritability $\hat{h}_{12}$ from the fitted slope.

In principle, we can easily estimate environmental covariance ρ_{12}^{ϵ} from the intercept term of Eq. (7). However, the number of overlapped samples n_s may not be known exactly in practice. To handle this problem, we consider two cases: (i) When samples from two GWASs do not overlap ($n_s = 0$) or slightly overlap ($n_s/\sqrt{n_1 n_2} \ll 1$), we simply ignore the environmental correlation and set $\hat{\rho}_{12}^{\epsilon} = 0$. This case corresponds to two GWASs from different cohorts, e.g., BBJ and the China Kadoorie Biobank (CKB). (ii) When samples substantially overlap between two GWASs (i.e., two phenotypes measured on the same biobank or cohort), we follow [Lu, 2017] and approximate $n_s/\sqrt{n_1 n_2}$ by 1. In such a way, we can easily obtain environmental covariance as: $\hat{\rho}_{12}^{\epsilon} = \hat{\rho}_{12}^{pheno} - \hat{h}_{12}$. For parameters within the auxiliary population, we apply the same procedure to obtain their estimates.

To estimate cross-population genetic covariance $\{h_{11'}, h_{12'}, h_{21'}, h_{22'}\}$, we use our recently developed XPA method [Cai, 2021] which has been shown effective for cross-population analysis. The details are provided in Section 3.3.1 of the supplementary note.

2.3 PRS construction using GWAS summary statistics

To obtain the estimate of fixed effects $\boldsymbol{\gamma}$ and posterior mean of random effects $\boldsymbol{\beta}$, we define:

$$\tilde{\mathbf{y}} = \left[\mathbf{y}_1 - \mathbf{Z}_1\boldsymbol{\omega}_1,\ \ \mathbf{y}_2 - \mathbf{Z}_2\boldsymbol{\omega}_2,\ \ \mathbf{y}_1' - \mathbf{Z}_1'\boldsymbol{\omega}_1',\ \ \mathbf{y}_2' - \mathbf{Z}_2'\boldsymbol{\omega}_2'\right]^T,$$
$$\boldsymbol{\Sigma}_e = \begin{bmatrix} (1-h_1^2)\mathbf{I}_{n_1} & \rho_{12}^{\epsilon}\boldsymbol{I}_{n_s} & \mathbf{0} & \mathbf{0} \\ \rho_{12}^{\epsilon}\boldsymbol{I}_{n_s} & (1-h_2^2)\mathbf{I}_{n_2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & (1-h_{1'}^2)\mathbf{I}_{n_1'} & \rho_{1'2'}^{\epsilon}\boldsymbol{I}_{n_s'} \\ \mathbf{0} & \mathbf{0} & \rho_{1'2'}^{\epsilon}\boldsymbol{I}_{n_s'} & (1-h_{2'}^2)\mathbf{I}_{n_2'} \end{bmatrix}, \tag{8}$$

where $\boldsymbol{I}_{n_s} = \begin{bmatrix} \mathbf{I}_{n_s} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ and $\boldsymbol{I}_{n_s'} = \begin{bmatrix} \mathbf{I}_{n_s'} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$. By combining Eqs. (3,4,5) and taking integration over $\boldsymbol{\beta}$, the marginal distribution of $\tilde{\mathbf{y}}$ can be obtained as

$$\tilde{\mathbf{y}} \sim \mathcal{N}(\mathbf{X_1}\boldsymbol{\gamma}, \boldsymbol{\Omega}),\ \ \boldsymbol{\Omega} = \mathbf{X}(\boldsymbol{\Sigma}_\beta \otimes \mathbf{I}_p)\mathbf{X}^T + \boldsymbol{\Sigma}_e, \tag{9}$$

where $\mathbf{X_1} = diag\{\mathbf{X}_{l1}, \mathbf{X}_{l2}, \mathbf{X}_{l1}', \mathbf{X}_{l2}'\}$ and $\mathbf{X} = diag\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_1', \mathbf{X}_2'\}$. Given the estimates $\hat{\boldsymbol{\Sigma}}_\beta$ and $\hat{\boldsymbol{\Sigma}}_e$, the fixed large genetic effects in Eq. (5) can be computed using generalized least squares as:

$$\hat{\boldsymbol{\gamma}} = (\mathbf{X_1}^T\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X_1})^{-1}\mathbf{X_1}^T\hat{\boldsymbol{\Omega}}^{-1}\tilde{\mathbf{y}}. \tag{10}$$

Then the posterior mean of $\boldsymbol{\beta}$ is given as:

$$\hat{\boldsymbol{\mu}} = \left(\mathbf{X}^T\hat{\boldsymbol{\Sigma}}_e^{-1}\mathbf{X} + \hat{\boldsymbol{\Sigma}}_\beta^{-1} \otimes \mathbf{I}_p\right)^{-1}\mathbf{X}^T\hat{\boldsymbol{\Sigma}}_e^{-1}(\tilde{\mathbf{y}} - \mathbf{X_1}\hat{\boldsymbol{\gamma}}). \tag{11}$$

Despite the closed-form solutions given in Eq. (10) and Eq. (11), individual-level GWAS data are required to obtain $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\mu}}$. To address this issue, we show that Eq. (10) and Eq. (11) can be approximated by using GWAS summary statistics and LD references (see supplementary note 3.3.2 and 3.3.3 for more details). For example, the posterior mean given in Eq. (11) can be approximated as:

$$\hat{\boldsymbol{\mu}} = \left[\hat{\boldsymbol{\mu}}_1,\ \hat{\boldsymbol{\mu}}_2,\ \hat{\boldsymbol{\mu}}_1',\ \hat{\boldsymbol{\mu}}_2'\right]^T \approx$$
$$\left(\begin{bmatrix} \left(\begin{bmatrix} n_1 & n_s \\ n_s & n_2 \end{bmatrix} \circ \hat{\Sigma}_e\right) \otimes \hat{\mathbf{R}} & \mathbf{0} \\ \mathbf{0} & \left(\begin{bmatrix} n_1' & n_s' \\ n_s' & n_2' \end{bmatrix} \circ \hat{\Sigma}_e'\right) \otimes \hat{\mathbf{R}}' \end{bmatrix} + \hat{\boldsymbol{\Sigma}}_\beta^{-1} \otimes \mathbf{I}_p\right)^{-1}$$
$$\begin{bmatrix} \Sigma_e \begin{bmatrix} \sqrt{\frac{n_1}{p}}\mathbf{z}_1 - \hat{\mathbf{R}}_{sl}\hat{\boldsymbol{\gamma}}_1 \\ \sqrt{\frac{n_2}{p}}\mathbf{z}_2 - \hat{\mathbf{R}}_{sl}\hat{\boldsymbol{\gamma}}_2 \end{bmatrix} \\ \Sigma_e' \begin{bmatrix} \sqrt{\frac{n_1'}{p}}\mathbf{z}_1' - \hat{\mathbf{R}}_{sl}'\hat{\boldsymbol{\gamma}}_1' \\ \sqrt{\frac{n_2'}{p}}\mathbf{z}_2' - \hat{\mathbf{R}}_{sl}'\hat{\boldsymbol{\gamma}}_2' \end{bmatrix} \end{bmatrix}, \tag{12}$$

where $\hat{\Sigma}_e = \begin{bmatrix} 1-\hat{h}_1^2 & \hat{\rho}_{12}^{\epsilon} \\ \hat{\rho}_{12}^{\epsilon} & 1-\hat{h}_2^2 \end{bmatrix}^{-1}$, $\hat{\Sigma}_e' = \begin{bmatrix} 1-\hat{h}_{1'}^2 & \hat{\rho}_{1'2'}^{\epsilon} \\ \hat{\rho}_{1'2'}^{\epsilon} & 1-\hat{h}_{2'}^2 \end{bmatrix}^{-1}$, notations $\circ$ and $\otimes$ are element-wise product and Kronecker product, $\hat{\mathbf{R}} = \hat{\mathbf{X}}^T\hat{\mathbf{X}}/n_r$ and $\hat{\mathbf{R}}' = \hat{\mathbf{X}}'^T\hat{\mathbf{X}}'/n_r'$ are the LD matrices of all SNPs, and $\hat{\mathbf{R}}_{sl} = \hat{\mathbf{X}}^T\hat{\mathbf{X}}_l/n_r$ and $\hat{\mathbf{R}}_{sl}' = \hat{\mathbf{X}}'^T\hat{\mathbf{X}}_l'/n_r'$ are the LD matrices between all SNPs and large-effect SNPs for the two populations, $n_r \times p$ matrix $\hat{\mathbf{X}}$ and $n_r' \times p$ matrix $\hat{\mathbf{X}}'$ are standardized genotype matrices (with mean 0 and variance $1/p$) from the reference panels of the target and auxiliary populations, respectively. By computing the LD matrices using external reference genotypes from the two populations, XPXP can account for heterogeneity of the LD pattern across populations.

Nevertheless, solving the linear systems in Eq. (12) involves a computationally intensive $4p \times 4p$ matrix inversion. Therefore, we further apply a block-diagonal matrix approximation [Berisa and Pickrell, 2016] to the LD matrices, $\hat{\mathbf{R}}$ and $\hat{\mathbf{R}}'$, and a fast conjugate gradient (CG) algorithm to estimate the posterior mean of $\boldsymbol{\beta}$ in a computationally efficient and parallel fashion. This algorithmic design makes the computational complexity of XPXP nearly linear to the number of SNPs. More specifically, we divide the whole genome into thousands of approximately independent LD blocks (i.e., 1,445 blocks for East Asian ancestry). We compute the LD matrices within each block while ignoring SNP correlations between different LD blocks. Given the block diagonal LD matrices, XPXP computes $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\mu}}$ in one block at a time and uses CG algorithm to efficiently solve the linear systems (see Table 2 for more details).

Finally, to obtain the effect sizes for the dosage genotypes, we re-scale both the estimated posterior mean and fixed large genetic effects by $\tilde{\boldsymbol{\mu}}_{kj} = \hat{\boldsymbol{\mu}}_{kj}/(s_{kj}\sqrt{p})$, for $j = 1, ..., p$, $k = 1, ..., K$ and $\tilde{\boldsymbol{\gamma}}_{kj} = \hat{\boldsymbol{\gamma}}_{kj}/(s_{kj}\sqrt{p})$, for $j = 1, ..., l_k$, $k = 1, ..., K$. When a new observation with genotype $\mathbf{g}_{new} \in \mathbb{R}^p$ from the target population is available, its PRS for k-th trait can be computed as $\mathrm{PRS}_{k,new} = \mathbf{g}_{new}^T\tilde{\boldsymbol{\mu}}_k + \mathbf{g}_{lk,new}^T\tilde{\boldsymbol{\gamma}}_k$, where $\mathbf{g}_{lk,new} \in \mathbb{R}^{l_k}$ is the vector of dosage genotypes corresponding to the SNPs with large effects for trait k in target population. When the true phenotypic value $\mathbf{y}_{k,new}$, covariates $\mathbf{Z}_{k,new}$ and covariates effect estimates $\hat{\boldsymbol{\omega}}_k$ are also available, we can evaluate the prediction accuracy of $\mathrm{PRS}_{k,new}$ using the predicted R^2 defined as:

$$R^2 = corr(\mathbf{y}_{k,new} - \mathbf{Z}_{k,new}\hat{\boldsymbol{\omega}}_k, \mathrm{PRS}_{k,new})^2. \tag{13}$$

3 Results

3.1 Overview of PRS analysis

We compared the performance of XPXP with several summary-level PRS methods, including the P+T procedure [Consortium, 2009], LDpred-inf [Vilhjálmsson, 2015], LDpred2 [Privé, 2020], PRS-CSx [Huang,

Table 1. Summary of multivariate PRS methods

Characteristics	PRS-CSx	MultiPRS	SMTpred	XPXP
Training by summary statistics	yes	yes	yes	yes
Require LD reference panel	yes	yes	yes	yes
Explicitly model multiple populations	yes	yes	no	yes
Explicitly model multiple phenotypes	no	yes	yes	yes
Explicitly model sample overlapping	no	no	no	yes
High computational efficiency	no	no	no	yes
Individual-level validation data required for tuning parameters	no	yes	no	no
Program language	Python3	NA	Python2	Python3

2021], MultiPRS [Márquez-Luna, 2017] and SMTpred [Maier, 2018]. The P+T procedure, LDpred and LDpred2 are widely used PRS methods in the single population setting. We included these three methods as the baseline. LDpred2 assumes a mixture model to characterize SNP effect sizes and its model parameters can be estimated automatically in the Bayesian framework. PRS-CSx is developed for construction of cross-population PRSs by introducing a shared continuous shrinkage prior to borrow information from well-powered auxiliary GWASs. MultiPRS is a risk profile approach that linearly combines PRSs from multiple populations using mixing weights estimated in a validation dataset. However, MultiPRS may perform poorly when the validation sample size is insufficient to accurately estimate the linearly composed weights. SMTpred is designed for analyzing multiple traits in a single population. To evaluate whether SMTpred has satisfactory performance in the cross-population setting, we include SMTpred in the comparison.

Although both XPXP and SMTpred can construct PRS by using multi-trait information, they differ in three aspects: (i) XPXP incorporates population-specific signals by including a subset of SNPs with large effects as fixed effects while SMTpred only considers the polygenic effects as random effects. (ii) XPXP allows the existence of sample overlap by modeling the environmental covariance among phenotypes measured on the overlapping individuals while SMTpred assumes that all GWAS datasets used for constructing PRS are nearly independent. (iii) XPXP approximates the individual-level exact posterior mean given in Eq. (11) by using summary-level data (i.e., z-score) and genotypes from reference panels. SMTpred adopts a simplified approximation which only uses SNP effects estimated from LDpred-inf and their genetic correlation. This method does not fully account for heterogeneity of LD in the cross-population set and thus may lead to a sub-optimal approximation. For a better overview of those multivariate PRS methods, we provide a fundamental comparison in Table 1. We have also been aware of several other methods that construct PRS in a single population, such as lassosum [Mak, 2017], MTAG [Turley, 2018]. We do not include them because they have been compared with XPA [Cai, 2021], which is a special case of XPXP when analyzing single phenotype across two populations only.

3.2 Simulation study

In our simulations study, LDpred-inf was trained using the ldpred software version 1.0.11, and LDpred2 was trained using the R package 'bigsnpr'. For P+T procedure, we followed the XPA and applied R package 'ieugwasr' to compute PRS under the setting of 1000 kb region size, LD threshold $r^2 = 0.1$, and nine candidate p-values thresholds 5×10^{-8}, 1×10^{-6}, 1×10^{-4}, 0.001, 0.01, 0.05, 0.1, 0.2, and 0.5. The optimal threshold was then selected by tuning on testing data. Because LDpred-inf, LDpred2 and P+T methods cannot handle GWAS summary statistics from multiple phenotypes, we only trained these three models with either the target dataset or the auxiliary dataset for the phenotype of interest. PRS-CSx was trained using cross-population GWAS datasets with the parameters a=1, b=0.5, phi=0.01, n_iter=1000*2, n_burnin=500*2, and thin=5. For MultiPRS, we combined the resulting PRSs computed from LDpred2 to construct a final PRS, and the mixing weights were estimated by cross-validation. To assess the prediction utility of SMTpred, we applied SMTpred to construct multi-trait PRSs by combining SNP effects inferred by LDpred-inf for all GWAS training datasets. For XPXP, we applied the P+T procedure with LD threshold $r^2 = 0.1$ and p-value threshold 1×10^{-6} to identify the SNPs with large effects on the target phenotype. In the following analysis, we applied the same procedure and parameters to include large population-specific effects in the target population.

First, we evaluated the performance of XPXP in the cross-population analysis by using one phenotype in both target and auxiliary populations. We constructed the target and auxiliary training samples from the Chinese and UKBB genotypes, respectively. Specifically, 10,000 samples randomly drawn from the Chinese dataset described in [Cai, 2021] were used as the target population. For the auxiliary population, we explored five different sample sizes using random samples from the UKBB dataset: 0, 10,000, 20,000, 40,000, and 60,000. We included about $300,000$ approximately independent SNPs from their overlapped genotypes after LD pruning. Then, we generated a moderately polygenic scenario by setting 1% SNPs with shared non-zero effects and 0.01% SNPs with ancestry-specific large effects for each population and varied the trans-ancestry genetic correlation (denoted as TG_{corr}) among $\{0, 0.3, 0.6\}$. The shared effects were simulated from the bivariate normal distribution $\mathcal{N}(\mathbf{0}, \begin{bmatrix} \frac{h_1^2}{0.01p} & \frac{TG_{corr}h_1h_{1'}}{0.01p} \\ \frac{TG_{corr}h_1h_{1'}}{0.01p} & \frac{h_{1'}^2}{0.01p} \end{bmatrix})$ and ancestry-specific large effects were generated by two independent normal distributions $\mathcal{N}(0, \frac{h_{l1}^2}{0.0001p})$ and $\mathcal{N}(0, \frac{h_{l1'}^2}{0.0001p})$, where $h_1^2 = h_{1'}^2 = 0.48$ and $h_{l1}^2 = h_{l1'}^2 = 0.02$. Here 0.01 means that 1% of SNPs jointly contribute to heritabilities h_1^2 and $h_{1'}^2$ with per-SNP heritabilities $\frac{h_1^2}{0.01p}$ and $\frac{h_{1'}^2}{0.01p}$. Similarly, for population-specific large effects, 0.0001 means that 0.01% of SNPs jointly contribute heritabilities h_{l1}^2 and $h_{l1'}^2$ with per-SNP heritabilities $\frac{h_{l1}^2}{0.0001p}$ and $\frac{h_{l1'}^2}{0.0001p}$. Given the effect sizes and genotype matrices, quantitative phenotypes in both populations were simulated using Eq. (5). Next, we computed the z-scores of the two datasets by marginally regressing the simulated phenotypes on each SNP. Finally, we sampled 2,000 individuals in each population serving as LD reference panels for parameter estimation and PRS construction. To evaluate the prediction performance, we further sampled 2,000 individuals from the Chinese dataset as the independent test set of the target population. For each simulation setting, we computed the averaged predicted R^2 from 10 replications.

As expected, the prediction accuracy of single-trait PRS methods (LDpred2, LDpred-inf and P+T) trained on the auxiliary dataset could not be improved regardless of the auxiliary sample size when the trans-ancestry genetic correlation (TG_{corr}) was zero (Figure 1a). However, as the TG_{corr} becomes moderate (i.e., 0.3) or strong (i.e., 0.6), the performance of single-trait PRS methods steadily improved as the auxiliary sample size increased. Consistent with our previous study [Cai, 2021], single-trait PRS models trained on the auxiliary dataset only were more accurate than those trained on the target population when the correlation was strong and the auxiliary sample size was large. For multivariate PRS models, the improvement of PRS-CSx seemed to be very small even though the TG_{corr} is as high as 0.6 and the auxiliary sample size increased to 60K. MultiPRS performed worst when the TG_{corr} is 0. However, as TG_{corr} becomes strong (i.e., 0.6), MultiPRS achieved comparable performance to SMTpred. XPXP performed satisfactorily in different settings. In the presence of strong TG_{corr} and a large auxiliary sample size (e.g. $n =$

60,000), XPXP ($R^2 = 4.9\%$) was able to construct more accurate PRS than SMTpred ($R^2 = 4.4\%$). The advantage of XPXP over SMTpred can be attributed to the population-specific effects incorporated in XPXP and more accurate approximation accounting for heterogeneity in LD. From this experiment, we realized that TG_{corr} is the key to utilize a large amount of information in the biobank-scale data of EUR ancestry for construction of PRSs in the under-represented ancestry. In other words, the effective sample size of the under-represented target population increases when TG_{corr} is strong. This helps a lot to improve the PRS accuracy of the target population.

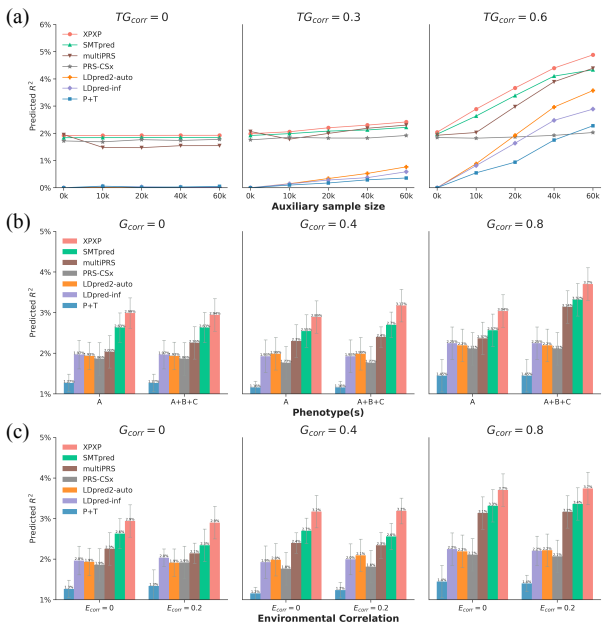

Fig. 1. Comparison of XPXP, SMTpred, MultiPRS, PRS-CSx, LDpred2, LDpred-inf, and P+T procedure in simulation studies. (a) Mean predicted R^2 in three simulation scenarios with different trans-ancestry genetic correlations. For XPXP, SMTpred MultiPRS and PRS-CSx, the solid lines show the R^2 obtained by combining both target and auxiliary datasets. For other univariate PRS methods, the solid lines show the R^2 obtained by training with auxiliary dataset only. (b) Mean predicted R^2 in six simulation scenarios with three different genetic correlations and two different numbers of phenotypes. (c) Mean predicted R^2 in six simulation scenarios with three different genetic correlations and two different environmental correlations. Results are summarized from 10 replications. Error bars represent the standard errors of predicted R^2 evaluated on 10 replications.

Second, we evaluate the performance of XPXP with multiple-phenotype in the cross-population setting. We used three phenotypes in both populations and fixed the auxiliary sample size and trans-ancestry correlation at 10,000 and 0.6, respectively. We considered three settings of the genetic correlation between phenotypes from the same population $G_{corr} = \{0, 0.4, 0.8\}$, corresponding to zero, moderate, and strong genetic correlations, respectively. The visualization of genetic correlation pattern for three phenotypes from two populations was shown in Figure S1. When $G_{corr} = 0$, the prediction accuracy can not be improved as expected. When G_{corr} became moderate or strong, the prediction accuracy of MultiPRS, SMTpred and XPXP were substantially improved as the number of phenotypes increased (Figure 1b). When $G_{corr} = 0.8$, XPXP achieved $R^2 = 3.7\%$ while the runner-up method SMTpred achieved $R^2 = 3.3$, indicating 12% improvement of PRS accuracy.

Third, we evaluated the performance of XPXP in the presence of environmental correlation and sample overlap. We considered three phenotypes in both target and auxiliary populations. We fixed the auxiliary sample size at 10,000 and trans-ancestry correlation at 0.6. As visualized in Figure S2, we considered three settings of genetic correlation $G_{corr} = \{0, 0.4, 0.8\}$ and two settings of environmental correlation $E_{corr} = \{0, 0.2\}$ between phenotypes within a population. As we shall demonstrate in the real data analysis, $E_{corr} = 0.2$ is quite a common scenario. XPXP had satisfactory performance in the presence of moderate environmental correlation (Figure 1c). In contrast, the performance of SMTpred degraded due to the unaccounted environmental correlation.

Table 2. CPU timing of multivariate PRS methods under different scenarios

Methods	0.3M SNPs		1M SNPs	
	1 phenotype	2 phenotypes	1 phenotype	2 phenotypes
XPXP	1min	2min	3min	5min
SMTpred	21min x2	21min x4	29min x2	29min x4
MultiPRS	7min x2	7min x4	25min x2	25min x4
PRS-CSx	5h	NA	19h	NA

Finally, we evaluated the CPU times and memory usage of those multivariate PRS methods when different numbers of SNPs were included in the model: 300,000 and 1,000,000. Under each scenario, we further considered a single phenotype across two populations (a total of two GWAS datasets) and two phenotypes across two populations (a total of four GWAS datasets) as training data. All of the tests were performed on the same computing environment (20 CPU cores of Intel(R) Xeon(R) Gold 6230N CPU @ 2.30GHz processor, 1TB of memory, and a 22 TB solid-state disk). For SMTpred, we first used LDpred-inf to estimate SBLUP (summary statistic approximate BLUP) SNP effects, then we combined SBLUP SNP effects of multiple GWAS datasets using the optimal index weighting proposed in SMTpred. Therefore, the CPU times and memory usage of SMTpred are mainly attributed to LDpred-inf. Similarly, the CPU times and memory usage of MultiPRS should also be attributed to LDpred2. Because LDpred-inf or LDpred2 was performed independently on a single GWAS data set, we reported the CPU times of SMTpred and MultiPRS as the running time of analyzing one data set multiplied by the number of GWAS datasets. As shown in Table 2 and Table S1, our XPXP has computational advantages over other PRS methods.

3.3 Real data applications

3.3.1 Height

To study the performance of XPXP and other PRS methods in real applications, we considered the construction of PRS for human height of African population (AFR) using two height GWASs from UKBB ($n = 458{,}303$) and GIANT ($n = 252{,}682$) in European population (EUR) and one heigh GWAS from BBJ ($n = 159{,}095$) in East Asian population (EAS). To investigate the role of the sample size of the auxiliary population, we applied these PRS methods to construct height PRS in African population by integrating African training dataset and subsampled UKBB datasets with different sample sizes. After that, we showed how to further improve the prediction performance using multiple GWAS datasets from multiple populations. African data was obtained by combining the African samples in Personalized Medicine (IPM) BioMe biobank ($n = 5{,}491$) and UK biobank ($n = 2{,}931$). After quality control and overlapping, 2,690,737 autosomal SNPs were retained for PRS construction. We randomly selected 7,422 samples for training and used the remaining 1K samples for testing. Height GWASs of UKBB, GIANT and BBJ are summarized in Table 3. The details of datasets and pre-processing procedures are given in supplementary note 3.1.

First, we systematically investigated how the sample size of auxiliary datasets influences the predictive performance of XPXP. For the auxiliary dataset, we randomly subsampled 0∼400K individuals from the UKBB

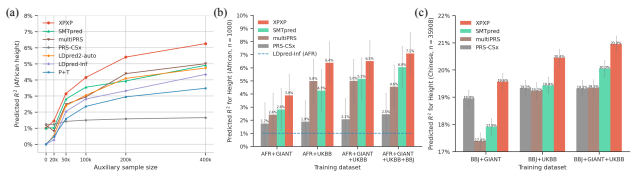

Fig. 2. Prediction performance for height in African and Chinese populations. (a) Influence of the auxiliary sample size on predictive performance of XPXP, SMTpred, MultiPRS, PRS-CSx, LDpred2, LDpred-inf, and P+T procedure. We trained XPXP, SMTpred MultiPRS and PRS-CSx using about 7,000 African training samples with random subsamples drawn from UKBB individuals as the auxiliary dataset and evaluated the predicted R^2 in an independent African testing data. For those multivariate PRS methods, the solid lines show the predicted R^2 obtained by combining both target and auxiliary datasets. For LDpred2, LDpred-inf and P+T procedure, the solid lines show the predicted R^2 obtained by training with the auxiliary dataset only. The predicted R^2 for height in African (b) and Chinese (c) when combining different sources of height GWAS data. Error bars represent the standard errors of predicted R^2 estimated by block-jackknife based on the testing data [Weissbrod, 2021]. As standard error depends on the testing sample size, the estimated standard errors for AFR height PRS are relatively large due to the small sample size of testing data.

European samples. As a comparison, we also trained single-trait PRS methods, LDpred2, LDpred-inf and P+T procedure, using the auxiliary dataset only. As shown in Figure 2a, due to the heterogeneity between African and European populations, LDpred2 trained on 7K African samples achieved a better prediction performance than that trained on 20K UKBB samples. The performance of LDpred-inf and LDpred2 was gradually improved when more UKBB samples ($\geq$ 50K) were included for training. For multivariate PRS methods, the improvement of PRS-CSx seemed to be very minor when the auxiliary sample size increased from 20K to 400K. The performance of MultiPRS and SMTpred steadily improved as the auxiliary sample size increased. Notably, our XPXP always achieved the best performance and effectively improved the prediction accuracy by integrating African training data with the UKBB samples. Notice that XPXP trained on 7K Africans and 100K Europeans achieved comparable performance with LDpred-inf which was trained using 400K European samples, indicating that samples from the target population play an essential role in PRS construction.

Next, we built a height PRS model for the African population by combining the small-scale African training data and three publicly available height GWAS summary statistics from EUR (UKBB and GIANT) and EAS (BBJ). After z-score imputation [Pasaniuc, 2014] and quality control for GIANT data, 2,689,436 SNPs were used to construct PRS. As shown in Figure S6, the estimated genetic correlation between UKBB and GIANT was 0.82, higher than the trans-ancestry genetic correlation between African and GIANT (0.58), UKBB (0.61) or BBJ (0.52). In view of the substantial genetic correlations among the four height GWAS data, PRS constructed by XPXP was able to effectively improve prediction accuracy in the African population by leveraging shared genetic basis from the UKBB, GIANT and BBJ data. As summarized in Figure 2b, XPXP largely outperformed other methods for different combinations of training datasets. When XPXP was trained on AFR+GIANT, the predicted R^2 increased from 1% to 3.9%. We further increased the sample size of training data by including the UKBB and BBJ data. The predicted R^2 obtained by XPXP increased to 7.1%, achieving (7.1%-6.0%)/6.0% $\approx$ 18% improvement compared to the SMTpred. On the contrary, the predicted R^2 of MultiPRS declined from 5.0% (AFR+GIANT+UKBB) to 4.6% (AFR+GIANT+UKBB+BBJ), suggesting no information was borrowed from BBJ (see more discussion in Figure S12). To include PRS-CSx, we first applied MTAG for meta-analysis of the GIANT and UKBB because PRS-CSx only supports at most one GWAS summary for each population as input.

Table 3. Sources of GWAS summary statistics.

Trait name	Full name	Sample size	Reference
T2D-EAS	Type 2 Diabetes	108,479	[Ishigaki, 2020]
T2D-EUR (UKBB)	Type 2 Diabetes	459,324	[Loh, 2018]
BMI-EAS	Body Mass Index	158,284	[Akiyama, 2017]
BMI-EUR (UKBB)	Body Mass Index	457,824	[Loh, 2018]
height-EAS	height	159,095	[Akiyama, 2019]
height-UKB	height	458,303	[Loh, 2018]
height-GIANT	height	252,682	[Wood, 2014]
HDL-EAS	High-density-lipoprotein cholesterol	70,657	[Kanai, 2018]
LDL-EAS	Low-density-lipoprotein cholesterol	72,866	[Kanai, 2018]
TC-EAS	Total cholesterol	128,305	[Kanai, 2018]
TG-EAS	Triglyceride	105,597	[Kanai, 2018]

To evaluate the generalization ability of XPXP for other populations, we applied it to EAS by constructing a height PRS model using publicly available height GWAS summary data from BBJ, GIANT, and UKBB. We evaluated the predicted R^2 in two independent Chinese testing data: UKBB Chinese (n = 1,439, see supplementary note 3.1.2 for more details) and a medium-scale Chinese cohort (n = 35,908) recruited from WeGene platform [Cai, 2021]. The WeGene dataset consists of a diverse cohort that covers 43 out of 56 ethnic groups in China. Participants were genotyped on the Illumina or Affymetrix platforms to enable the GWASs of anthropometric traits (e.g., height and BMI) in Chinese population. After the same procedure of z-score imputation for GIANT data and quality control, 3,756,616 SNPs remained for construction of PRS. In both evaluation datasets, XPXP showed the best performance for construction of PRS among all these methods. In particular, the predicted R^2 obtained by XPXP is as high as 20.9% in WeGene Chinese test data (Figure 2c). For Chinese testing samples from UKBB, XPXP trained on all three GWAS datasets achieved 17% R^2, leading to a 9% improvement compared to the runner-up (Figure S13). Because SMTpred and XPXP have shown stable advantages over other methods, we only considered these two multivariate PRS methods in the analysis of disease trait.

3.3.2 Type 2 diabetes

To demonstrate the utility of XPXP in dichotomous traits, we evaluated the PRS accuracy of XPXP on T2D in EAS. We used T2D GWAS summary statistics derived from BBJ (n = 108,479) and UKBB (n = 459,324) as the training datasets. For evaluation of prediction accuracy, we selected 4,367 East Asian samples from GERA cohort (dbGaP phs000674.v3.p3) serving as testing data, which we had access to individual-level GWAS data (see supplementary note 3.1.4). To leverage cross-phenotype genetic correlation, we additionally included the summary data of 6 phenotypes that are genetically correlated to T2D, including BMI, Total cholesterol (TC), Triglycerides (TG), high-density lipoproteins (HDL) and Low-density lipoprotein (LDL) from BBJ, and BMI from UKBB. These datasets were summarized in Table 3, where we added '-EAS' and '-EUR' behind the abbreviation of a trait name for GWAS summary statistics from BBJ and UKBB, respectively. We first estimated the genetic and environmental correlations among those 8 GWAS summary statistics. Note that phenotypes measured on two different populations have no sample overlap, we thus simply set their environmental correlations as zero. As shown in Figure 3, we observed a very strong trans-ancestry genetic correlation between T2D-EUR and T2D-EAS (0.678) and trans-ancestry correlation (0.555) between BMI-EAS and BMI-EUR, suggesting the shared genetic architecture between Europeans and East Asians. We also observed substantial genetic correlations between T2D-EAS and two metabolic traits: TG-EAS, HDL-EAS, indicating the widespread pleiotropy effects. Regarding the environmental correlation, phenotypes measured on the BBJ cohort had a mean absolute value of 0.11. In addition,

we observed a strong environmental correlation (0.548) between TC-EAS and LDL-EAS, implying the importance of modeling environmental correlation in Eq. (4).

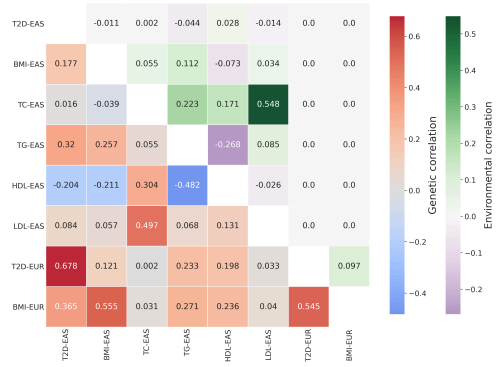

Fig. 3. Estimated genetic and environmental correlation among eight T2D related GWAS data. For genetic correlation, positive correlations are colored in red and negative correlations are colored in blue. For environmental correlation, positive correlations are colored in green and negative correlations are colored in purple. The corresponding estimated genetic and environmental covariances are shown in Figure S10.

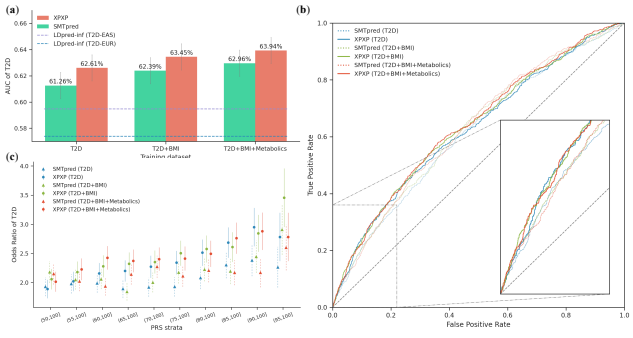

Fig. 4. Prediction performance for T2D in East Asian sample. (b) An ROC curve to compare the sensitivity and specificity of PRS generated by XPXP (solid line) and SMTpred (dash line) when different sets of GWAS data (T2D: T2D-EAS and T2D-EUR, T2D+BMI: T2D-EAS, T2D-EUR, BMI-EAS and BMI-EUR, T2D+BMI+Metabolics: T2D-EAS, T2D-EUR, BMI-EAS, BMI-EUR, TC-EAS, TG-EAS, HDL-EAS and LDL-EAS) were used for training, and the corresponding AUC scores were summarized in (a). Error bars represent the standard errors estimated by block-jackknife based on testing data. (c) The odds ratio of T2D across ten unequal PRS strata generated by XPXP (dot marker) and SMTpred (triangle marker), highlighting the increased risk among individuals in the top percentiles of PRS. Different colors corresponding to different combinations of GWAS training data. ORs and standard error (error bars) were estimated using logistic regression on the continuous scores.

To compare the prediction performance of XPXP and SMTpred, we trained the PRS models with different configurations of training datasets: T2D from two populations only, T2D and BMI from two populations, and all 8 GWAS summary statistics. We first constructed the ROC curve and evaluated the AUC value as an overall metric of the prediction performance. As shown in Figure 4a, XPXP largely outperformed SMTpred for different configurations of training datasets. More specifically, the AUC score obtained by XPXP achieved 62.61% using T2D GWAS summary statistics only. As we included more and more GWAS summary statistics of phenotypes correlated with T2D, the AUC score increased to 63.94%. For disease traits, it is more important to stratify individuals with high genetic predisposition from the general population. We therefore computed the odds ratio by contrasting the true disease status with predicted risk groups defined by a series of PRS strata (Figure 4c). We observed that XPXP trained on T2D GWAS data from two populations achieved 2.78 odds ratio in the top 5% quantile of PRS. As we included BMI summary statistics, the odds ratio was improved as high as 3.46, achieving (3.46-2.91)/2.91 $\approx$ 19% improvement compared to SMTpred trained on the same GWAS datasets. Then we evaluated XPXP and SMTpred trained by using all 8 GWAS datasets. Although the overall prediction accuracy (measured by AUC) increased, we observed a decline of odds ratio in the top 5% quantile of PRS. The stratification ability of both XPXP and SMTpred decreased when including 4 metabolically related GWAS summary statistics from BBJ, as shown in Figure 4b. The above results suggest a different perspective for us to consider the inclusion of multiple GWAS. In terms of overall accuracy, it is often very helpful to include more genetically related phenotypes. In terms of stratification of high-risk individuals in the general population, it may not be necessarily true since the improvement may not be equal across the PRS strata. Therefore, we would like to suggest that multiple phenotypes should be incorporated into the XPXP model in a step-wise manner and the stratification ability should be evaluated accordingly (see more discussion in Table S2).

4 Discussion and conclusion

In this study, we have introduced a novel and computationally efficient approach, XPXP, for constructing cross-population and cross-phenotype PRS. XPXP can substantially improve the genetic prediction of under-represented populations by leveraging the GWAS datasets of the well-powered auxiliary population through trans-ancestry genetic correlation. Given the widespread pleiotropic effects, XPXP also combines multiple phenotypes within the same population into a unified framework while taking the sample overlap into account. By incorporating the population-specific or phenotype-specific large genetic effects, XPXP allows a flexible model structure to accommodate both small polygenic effects and large genetic effects. Through comprehensive simulations and real data applications, we showed that XPXP achieved stable improvement over existing PRS methods. We believe that XPXP can serve as an effective tool of constructing PRSs for personal and clinic utility.

Although it is convenient to construct PRSs based on summary statistics, we should be aware of some potential limitations. First, confounding biases, such as population stratification, may still remain in the released GWAS summary statistics [Bulik-Sullivan, 2015b]. For our XPXP, we adopted the assumptions of LDSC to address the confounding issue. Under these assumptions, polygenic effects and confounding biases are distinguishable, where polygenic effects and confounding biases can be captured by the first-order term and the zero-order term of LD score, respectively. In such a way, we have tried to minimize the influence of confounding when we use XPXP to construct PRS. However, population structure driven by socioeconomic status [Tyrrell, 2016] or geographic structure [Abdellaoui, 2019] may not be fully corrected by routine adjustments. More careful investigations based on individual-level data are needed. Second, summary statistics obtained from different association methods may affect the accuracy of PRS construction. Linear mixed models are widely used for association mapping of quantitative traits and balanced case-control studies [Loh, 2015]. For a disease trait with extremely unbalanced case-control ratio (e.g., ratio $< 1/10$), the logistic mixed model implemented in SAIGE [Zhou, 2018] is preferred. For summary statistics released by genomic consortiums, they are often based on meta-analysis of GWASs in several cohorts, where fixed-effect models or random-effect models are widely used. Therefore, the PRS accuracy depends on the methods for meta-analysis. Third, in the early stage of GWAS, samples from different ancestries are often included for meta-analysis. The produced summary statistics can be inaccurate

because the genetic background of different ancestries has not been taken into account. For our method XPXP, we assumed the homogeneous ancestry in a single population and accounted for different ancestry across populations. However, our model did not consider admixed populations, such as Hispanic/Latinos. How to construct accurate PRSs for admixed populations is an interesting direction for future work. Fourth, summary statistics-based methods require relatively large sample sizes. As a rule of thumb, the LDSC estimator [Bulik-Sullivan, 2015a] of heritability and co-heritability needs at least 5,000 samples and 8,000 samples, respectively. Fifth, it is hard for summary statistics-based methods to examine the sample quality. Due to the presence of cryptic relatedness, the reported GWAS sample sizes could be larger than the effective sample sizes. Given the above concerns, we would like to recommend the construction of PRSs using the individual-level data when they are available. Our previous work XPA [Cai, 2021] offers a computationally efficient way to handle bio-bank scale individual-level data and it can provide more accurate PRS than its summary statistics version.

Funding

This work is supported in part by National Key R&D Program of China (2020YFA0713900), Hong Kong Research Grant Council [16307818, 16301419, 16308120], Hong Kong Innovation and Technology Fund [PRP/029/19FX], Hong Kong University of Science and Technology [startup grant R9405, Z0428 from the Big Data Institute] and the Open Research Fund from Shenzhen Research Institute of Big Data [2019ORF01004]. The computational task for this work was partially performed using the X-GPU cluster supported by the RGC Collaborative Research Fund: C6021-19EF.

References

Abdellaoui, A. *et al*. (2019). Genetic correlates of social stratification in great britain. *Nature human behaviour*, **3**(12), 1332–1342.

Akiyama, M. *et al*. (2017). Genome-wide association study identifies 112 new loci for body mass index in the japanese population. *Nature genetics*, **49**(10), 1458–1467.

Akiyama, M. *et al*. (2019). Characterizing rare and low-frequency height-associated variants in the japanese population. *Nature communications*, **10**(1), 1–11.

Berisa, T. and Pickrell, J. K. (2016). Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, **32**(2), 283.

Bulik-Sullivan, B. *et al*. (2015a). An atlas of genetic correlations across human diseases and traits. *Nature genetics*, **47**(11), 1236.

Bulik-Sullivan, B. K. *et al*. (2015b). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, **47**(3), 291–295.

Cai, M. *et al*. (2021). A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. *The American Journal of Human Genetics*, **108**(4), 632–655.

Chatterjee, N. *et al*. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics*, **17**(7), 392.

Consortium, I. S. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**, 6.

Coram, M. A. *et al*. (2017). Leveraging multi-ethnic evidence for risk assessment of quantitative traits in minority populations. *The American Journal of Human Genetics*, **101**(2), 218–226.

Gao, B. *et al*. (2021). Accurate genetic and environmental covariance estimation with composite likelihood in genome-wide association studies. *PLoS genetics*, **17**(1), e1009293.

Guo, H. *et al*. (2021). Detecting local genetic correlations with scan statistics. *Nature communications*, **12**(1), 1–13.

Hu, Y. *et al*. (2017). Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLoS genetics*, **13**(6), e1006836.

Huang, H. . *et al*. (2021). Improving polygenic prediction in ancestrally diverse populations. *medRxiv*, page doi:10.1101/2020.12.27.20248738.

Ishigaki, K. *et al*. (2020). Large-scale genome-wide association study in a japanese population identifies novel susceptibility loci across different diseases. *Nature genetics*, **52**(7), 669–679.

Kanai, M. *et al*. (2018). Genetic analysis of quantitative traits in the japanese population links cell types to complex human diseases. *Nature genetics*, **50**(3), 390–400.

Khera, A. V. *et al*. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics*, **50**(9), 1219–1224.

Lam, M. *et al*. (2019). Comparative genetic architectures of schizophrenia in east asian and european populations. *Nature genetics*, **51**(12), 1670–1678.

Lee, S. H. *et al*. (2011). Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, **88**(3), 294–305.

Lee, S. H. *et al*. (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature genetics*, **45**(9), 984.

Li, C., Yang, C., Gelernter, J., and Zhao, H. (2014). Improving genetic risk prediction by leveraging pleiotropy. *Human genetics*, **133**(5), 639–650.

Loh, P.-R. . *et al*. (2015). Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, **47**(3), 284.

Loh, P.-R. *et al*. (2018). Mixed-model association for biobank-scale datasets. *Nature genetics*, **50**(7), 906–908.

Lu, Q. *et al*. (2017). A powerful approach to estimating annotation-stratified genetic covariance via GWAS summary statistics. *The American Journal of Human Genetics*, **101**(6), 939–964.

Maier, R. *et al*. (2015). Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *The American Journal of Human Genetics*, **96**(2), 283–294.

Maier, R. M. *et al*. (2018). Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nature communications*, **9**(1), 1–17.

Mak, T. S. H. *et al*. (2017). Polygenic scores via penalized regression on summary statistics. *Genetic epidemiology*, **41**(6), 469–480.

Márquez-Luna, C. *et al*. (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic epidemiology*, **41**(8), 811–823.

Martin, A. R. *et al*. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, **51**(4), 584–591.

Mills, M. C. and Rahal, C. (2020). The GWAS diversity monitor tracks diversity by disease in real time. *Nature genetics*, **52**(3), 242–243.

Ming, J. *et al*. (2020). LPM: a latent probit model to characterize the relationship among complex traits using summary statistics from multiple GWASs and functional annotations. *Bioinformatics*, **36**(8), 2506–2514.

Multhaup, M. L. *et al*. (2019). 304-OR: Polygenic risk score predicts type 2 diabetes susceptibility in a diverse consumer genetic database. *Diabetes*, **68**.

Pasaniuc, B. *et al*. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, **30**(20), 2906–2914.

Privé, F. *et al*. (2020). Ldpred2: better, faster, stronger. *Bioinformatics*, **36**(22-23), 5424–5431.

Solovieff, N. *et al*. (2013). Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, **14**(7), 483–495.

Speed, D. *et al*. (2012). Improved heritability estimation from genome-wide SNPs. *The American Journal of Human Genetics*, **91**(6), 1011–1021.

Torkamani, A. *et al*. (2018). The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, **19**(9), 581–590.

Turley, P. *et al*. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature genetics*, **50**(2), 229–237.

Tyrrell, J. *et al*. (2016). Height, body mass index, and socioeconomic status: mendelian randomisation study in uk biobank. *bmj*, **352**.

Van Rheenen, W. *et al*. (2019). Genetic correlations of polygenic disease traits: from theory to practice. *Nature Reviews Genetics*, **20**(10), 567–581.

Varin, C. *et al*. (2011). An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42.

Vilhjálmsson, B. J. *et al*. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, **97**(4), 576–592.

Watanabe, K. *et al*. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nature genetics*, **51**(9), 1339–1348.

Weissbrod, O. *et al*. (2021). Leveraging fine-mapping and non-european training data to improve trans-ethnic polygenic risk scores. *medRxiv*.

Wood, A. R. *et al*. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, **46**(11), 1173–1186.

Yang, C. *et al*. (2015). Implications of pleiotropy: challenges and opportunities for mining Big Data in biomedicine. *Frontiers in genetics*, **6**, 229.

Zheng, J. *et al*. (2017). LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics*, **33**(2), 272–279.

Zhou, W. *et al*. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature genetics*, **50**(9), 1335–1341.