ARTICLE

A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits

Mingxuan Cai,[1,2,7] Jiashun Xiao,[1,2,7] Shunkang Zhang,[1,2,7] Xiang Wan,[3] Hongyu Zhao,[5,6] Gang Chen,[4,*] and Can Yang[1,2,*]

Summary

The development of polygenic risk scores (PRSs) has proved useful to stratify the general European population into different risk groups. However, PRSs are less accurate in non-European populations due to genetic differences across different populations. To improve the prediction accuracy in non-European populations, we propose a cross-population analysis framework for PRS construction with both individual-level (XPA) and summary-level (XPASS) GWAS data. By leveraging trans-ancestry genetic correlation, our methods can borrow information from the Biobank-scale European population data to improve risk prediction in the non-European populations. Our framework can also incorporate population-specific effects to further improve construction of PRS. With innovations in data structure and algorithm design, our methods provide a substantial saving in computational time and memory usage. Through comprehensive simulation studies, we show that our framework provides accurate, efficient, and robust PRS construction across a range of genetic architectures. In a Chinese cohort, our methods achieved 7.3%–198.0% accuracy gain for height and 19.5%–313.3% accuracy gain for body mass index (BMI) in terms of predictive R^2 compared to existing PRS approaches. We also show that XPA and XPASS can achieve substantial improvement for construction of height PRSs in the African population, suggesting the generality of our framework across global populations.

Introduction

In the past 15 years, genome-wide association studies (GWASs) have been performed on a wide spectrum of complex traits and diseases, providing an unprecedented opportunity to stratify the general population into different risk groups. With the availability of large-scale GWAS data, polygenic risk scores (PRSs) have been constructed to estimate the genetic predisposition of complex phenotypes by collecting contributions of many single-nucleotide polymorphisms (SNPs). The accurate construction of PRSs holds promise in disease risk prediction, personalized healthcare guidance, disease screening, and therapeutic intervention.[1] As one example, it was shown that an appropriately constructed PRS can identify 8% of the population with three-fold increased risk of coronary artery disease (CAD), while monogenic mutations with comparable risk can only cover 0.4% of the population.[2] In terms of the area under the receiver operator curve (AUC), the PRS's accuracy in predicting CAD onset can be as high as 0.81 based on 288,978 independent testing participants from the UK Biobank.[3] More recently, a significant improvement of AUC (from 0.73 to 0.80) was achieved in glaucoma prediction by incorporating PRSs in the traditional risk prediction model without genetic information.[4]

Despite the great promise of PRSs, one major limitation for its broader applications is the fact that most GWASs have been performed on samples with European (EUR) ancestry.[5–11] According to the GWAS diversity monitor,[8] about 89% of GWAS participants to date are from European ancestry, while less than 8% of the participants are from East Asian (EAS) ancestry and less than 0.45% are from African (AFR) ancestry. Despite the wealth of GWAS findings derived from Europeans, such an unbalanced sample makeup across global populations may exacerbate the disparities in genetic studies of non-Europeans.[6] Recent studies have reported that the PRSs derived from European samples are often less accurate when applied to other populations.[12,13] It remains unclear how much the genetic discovery from the European population can be transferred to non-European populations.

The challenges of transferable genetic studies arise from three aspects. First, SNPs with biologically important roles in the non-European populations may be neglected in GWASs if they are absent or have very low allele frequencies in Europeans.[14,15] Second, the same SNP may have different effect sizes on the same phenotype across different populations,[10,14] limiting the extrapolation value of GWAS findings and the GWAS-derived PRS power in non-discovery populations. Third, the linkage disequilibrium (LD) patterns vary across populations,[12,16–20] exacerbating the bias in extrapolating the PRS for risk prediction.

PRSs can be constructed by using either individual-level or summary-level GWAS data. The individual-level methods

[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]Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, China; [5]SJTU-Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University, Shanghai 201111, China; [6]Department of Biostatistics, Yale School of Public Health, New Haven, CT 06510, USA
[7]These authors contributed equally to this work
*Correspondence: chengangcs@gmail.com (G.C.), macyang@ust.hk (C.Y.)
https://doi.org/10.1016/j.ajhg.2021.03.002.

construct PRSs by directly taking the genotype and phenotype data as their input. Well-known individual-level approaches include best linear unbiased predictor (BLUP),[21] LASSO,[22] BayesR,[23,24] and BayesS.[25] These methods have their limitations in construction of PRS in the trans-ancestry setting. First, they are often too time consuming or memory consuming to be applicable for the biobank-scale GWAS data. While the recently proposed snpnet[26] method implements LASSO in a memory-efficient manner, it does not provide much gain in computational efficiency. Second, these methods cannot be easily extended to integrate datasets from heterogeneous ethnic backgrounds, since they do not account for different genetic architectures. While the bivariate BLUP (bvBLUP) is useful to reconcile the different genetic effects, the most popular bvBLUP approach provided by the GCTA software (GCTA-bvBLUP)[27] is not scalable to the biobank-scale dataset, and thus cannot fully utilize the large-scale datasets to improve prediction performance. Different from individual-level methods, summary-level PRS methods use only summary statistics (e.g., *z*-scores or marginal effect size estimates) that are widely available for large-scale GWASs to approximate the prediction of individual-level methods. Therefore, they are more flexible, efficient, and easily scalable to well-powered GWAS data. Notable summary-level methods include P+T procedure,[28] LDpred,[29] and lassosum.[30] By taking the advantage of large sample size, existing PRS models have proved informative in predicting genetic risk of European ancestry. However, in view of the imbalanced population composition and the different genetic structures across populations, existing PRS approaches that do not take the heterogeneous genetic architectures into account may not be easily applied in *trans*-ancestry genetic prediction, leading to sub-optimal prediction accuracy for under-represented populations. The recently proposed MTAG approach[31] provides an effective data integration framework to combine GWAS summary data of multiple phenotypes, which has been proved useful to improve power of association mapping or construction of PRSs in the same population. Despite the success of MTAG, its performance in the construction of PRSs across populations remains largely unknown. To generate more accurate PRSs by leveraging trans-ancestry information from large-scale European GWASs, a pioneer approach, XP-BLUP,[32] was recently proposed, which generalizes BLUP by introducing an extra variance component to model the effect sizes of significant SNPs derived from well-powered GWASs of the auxiliary population. The underlying assumption is that the significant SNPs from the auxiliary population are more useful for improving PRS accuracy in the under-represented populations. However, as XP-BLUP only incorporates information from significant SNPs, its improvement in prediction accuracy is limited. Therefore, there is a great need for a comprehensive investigation on the transferability of genetic study in the PRS construction among global populations.

With our innovations in data structure and algorithm design, XPA is able to handle bio-bank scale individual-level datasets and achieves the accurate prediction with the computational cost nearly linear to the SNP number and sample size. To demonstrate the effectiveness of our methods, we considered the data collected from about 33,000 Chinese participants through a direct-to-customer platform. For two anthropometric traits, height and body mass index (BMI), we applied XPA to integrate the Chinese training dataset (about 20,000 participants) with the UKBB dataset and evaluated its prediction accuracy on an independent Chinese testing dataset (about 13,000 participants). XPA achieved $R^2 = 18.92\%$ for height and $R^2 = 6.06\%$ for BMI. Compared to the runner up, XPA improved the relative prediction accuracy by 7.3% for height and 19.5% for BMI. We also evaluated XPASS and other related methods that only use summary statistics. Given the summary statistics obtained from the same individual-level GWAS data as its input, XPASS is slightly worse than XPA as expected, but it is more broadly applicable when individual-level data are not accessible. Since the ancestry profiles of Japanese and Chinese are close to each other, summary statistics from a larger East Asian GWAS, the BioBank Japan (BBJ) project,[33,34] was used as the training data to construct PRSs for testing dataset. By integrating summary data from BBJ with UKBB, XPASS was able to further improve the prediction accuracy for height with $R^2 = 19.54\%$, offering 12.7% relative improvement compared to training with Chinese and UKBB summary data. We also demonstrated that our proposed methods was able to improve the construction of PRSs in African population by integrating UKBB.

Material and methods

Method overview: XPA and XPASS

Due to the polygenicity of human complex traits, it is challenging to construct accurate PRSs for an under-studied target population. On one hand, the PRSs constructed using samples from EUR ancestry becomes less accurate when it is applied to non-European samples. On the other hand, the accuracy of PRSs constructed only using samples from the under-represented target population is limited by the sample size. By integrating small-scale or medium-scale data from the target population with existing large-scale data resources from EUR ancestry, our methods can robustly improve the PRS prediction performance. The key idea of our methods is based on the observed substantial genetic correlation that largely remains for the same trait between populations due to the shared genetic basis. Therefore, a large amount of information in the biobank-scale data of EUR ancestry can be utilized for risk prediction in the under-represented ancestry. By taking individual-level GWAS data as input, XPA offers analytic estimate for the SNP effect sizes using shared information across populations. With our innovations in data structure and algorithm design, such as Boolean representation[35,36] and stochastic approximation,[37] the analytic SNP effect size estimates can be computed efficiently, allowing us to construct PRSs in the target population accurately. Because of the unavailability of individual-level data for many traits, we have extended XPA to XPASS which requires only the GWAS *z*-scores and SNP correlation matrices from the target and auxiliary populations, where the

SNP correlation matrices can be approximated using block-diagonal correlation matrices from a genotype reference panel of the two populations. The computational complexity of XPA and XPASS is approximately linear to the number of SNPs and sample size, making our framework appealing in cross-population risk prediction with biobank-scale data.

XPA statistical framework

Consider a GWAS dataset $\{\mathbf{G}_1,\mathbf{Z}_1,\mathbf{y}_1\}$ from an under-represented ancestry, where $\mathbf{G}_1$ is an $n_1 \times p$ genotype matrix, $\mathbf{Z}_1$ is an $n_1 \times c_1$ matrix collecting all covariates (e.g., age, sex, and principal components), and $\mathbf{y}_1$ is an $n_1 \times 1$ phenotype vector. Due to the polygenicity of complex traits, the accuracy of risk prediction is limited by the sample size. Now suppose a biobank-scale dataset $\{\mathbf{G}_2,\mathbf{Z}_2,\mathbf{y}_2\}$ from European ancestry is also available, where $\mathbf{G}_2$ is an $n_2 \times p$ genotype matrix, $\mathbf{Z}_2$ is an $n_2 \times c_2$ covariate matrix, $\mathbf{y}_2$ is an $n_2 \times 1$ phenotype vector, and $n_2 \gg n_1$. Since we are mainly interested in improving risk prediction in the under-represented population, we shall regard it as the target population and the biobank-scale data from European ancestry as the auxiliary dataset. To reconcile the difference of allele frequencies in the two populations, XPA assumes that the SNP effect sizes in both the target and auxiliary populations increase as the allele frequencies decrease[38] and works with standardized genotype matrices. Specifically, let $\mathbf{g}_{1j} \in \mathbb{R}^{n_1}$ and $\mathbf{g}_{2j} \in \mathbb{R}^{n_2}$ denote the j-th column of $\mathbf{G}_1$ and $\mathbf{G}_2$, respectively. The corresponding column means and standard deviations are given as $\overline{\mathbf{g}}_{1j}$ and $\overline{\mathbf{g}}_{2j}$, s_{1j} and s_{2j}, respectively. Then we have the corresponding standardized genotype matrices as $\mathbf{X}_1 = [\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1p}] \in \mathbb{R}^{n_1 \times p}$ and $\mathbf{X}_2 = [\mathbf{x}_{21}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{2p}] \in \mathbb{R}^{n_2 \times p}$, where the j-th column of $\mathbf{X}_1$ and $\mathbf{X}_2$ is given as $\mathbf{x}_{1j} = (\mathbf{g}_{1j} - \overline{\mathbf{g}}_{1j})/(s_{1j}\sqrt{p})$ and $\mathbf{x}_{2j} = (\mathbf{g}_{2j} - \overline{\mathbf{g}}_{2j})/(s_{2j}\sqrt{p})$, respectively. In such a way, each column of $\mathbf{X}_1$ and $\mathbf{X}_2$ has mean 0 and variance $1/p$. We relate genotypes and phenotypes using the following linear models:

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{Z}_1\boldsymbol{\omega}_1 + \mathbf{X}_1\boldsymbol{\beta}_1 + \varepsilon, \\ \mathbf{y}_2 &= \mathbf{Z}_2\boldsymbol{\omega}_2 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\xi}, \end{aligned} \qquad \text{(Equation 1)}$$

where $\boldsymbol{\omega}_1 \in \mathbb{R}^{c_1}$ and $\boldsymbol{\omega}_2 \in \mathbb{R}^{c_2}$ are fixed effects of covariates, $\boldsymbol{\beta}_1 = [\beta_{1,1}, \beta_{1,2}, \ldots, \beta_{1,p}]^T \in \mathbb{R}^p$ and $\boldsymbol{\beta}_2 = [\beta_{2,1}, \beta_{2,2}, \ldots, \beta_{2,p}]^T \in \mathbb{R}^p$ are vectors collecting the SNP effect sizes from the two populations, and $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathbf{I}_{n_1})$ and $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma_\xi^2 \mathbf{I}_{n_2})$ are independent errors. Of note, the two datasets do not share samples because they are from different populations. Therefore, we can assume that the residual vectors ε and $\boldsymbol{\xi}$ are independent. To model the polygenic effects and their correlation between two populations, we introduce the following probabilistic structures on $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$:

$$\begin{pmatrix} \beta_{1,j} \\ \beta_{2,j} \end{pmatrix} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\beta) = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right), \quad j = 1, \ldots, p, \qquad \text{(Equation 2)}$$

where σ_1^2 and σ_2^2 are the variance components characterizing the polygenic effects in the two populations, respectively, ρ is the trans-ancestry genetic correlation of the same trait across population, and $\delta := \rho\sigma_1\sigma_2$ is the corresponding covariance. XPA computes the posterior mean of $\boldsymbol{\beta}_1$ by combining the target dataset with the auxiliary dataset through their genetic correlation, and therefore constructs an improved genetic prediction when the genetic correlation ρ is nonzero. In contrast to XP-BLUP,[32] XPA leverages genome-wide information by using all SNPs rather than only the top SNPs from the auxiliary population. While the inclusion of all SNPs from the biobank-scale auxiliary dataset introduces challenges in computation and data storage, we show that they can be properly addressed by the novel data structure and algorithm design in XPA.

XPA$_+$ for capturing population-specific effects

The XPA framework can incorporate population-specific genetic effects to improve prediction performance. To see this, we denote $\mathbf{Z}_{c1} \in \mathbb{R}^{n_1 \times c_1}$ and $\mathbf{Z}_{c2} \in \mathbb{R}^{n_2 \times c_2}$ as matrices collecting all the covariates, and $\mathbf{X}_{l1} \in \mathbb{R}^{n_1 \times l_1}$ and $\mathbf{X}_{l2} \in \mathbb{R}^{n_2 \times l_2}$ as the standardized genotype matrices of SNPs with large effects in populations one and two, respectively. By constructing $\mathbf{Z}_1$ and $\mathbf{Z}_2$ in Equation 1 as $\mathbf{Z}_1 = [\mathbf{Z}_{c1}, \mathbf{X}_{l1}]$ and $[\mathbf{Z}_{c2}, \mathbf{X}_{l2}]$, respectively, we can re-write Equation 1 as:

$$\begin{aligned} \mathbf{y}_1 &= \mathbf{Z}_{c1}\boldsymbol{\omega}_{c1} + \mathbf{X}_{l1}\boldsymbol{\gamma}_1 + \mathbf{X}_1\boldsymbol{\beta}_1 + \varepsilon, \\ \mathbf{y}_2 &= \mathbf{Z}_{c2}\boldsymbol{\omega}_{c2} + \mathbf{X}_{l2}\boldsymbol{\gamma}_2 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\xi}, \end{aligned} \qquad \text{(Equation 3)}$$

where $\boldsymbol{\gamma}_1 \in \mathbb{R}^{l_1}$ and $\boldsymbol{\gamma}_2 \in \mathbb{R}^{l_2}$ are fixed effects of the pre-selected SNPs, $\boldsymbol{\omega}_{c1} \in \mathbb{R}^{c_1}$ and $\boldsymbol{\omega}_{c2} \in \mathbb{R}^{c_2}$ are fixed effects of covariates, and $\boldsymbol{\omega}_1 = [\boldsymbol{\omega}_{c1}^T, \boldsymbol{\gamma}_1^T]^T$, $\boldsymbol{\omega}_2 = [\boldsymbol{\omega}_{c2}^T, \boldsymbol{\gamma}_2^T]^T$. SNPs in $\mathbf{X}_{l1}$ and $\mathbf{X}_{l2}$ are selected by applying the P+T procedure to the GWAS summary statistics of the target and the auxiliary populations, respectively. Because $l_1 \ll n_1$ and $l_2 \ll n_2$ in practice, the vectors of fixed effects $\boldsymbol{\gamma}_1 \in \mathbb{R}^{l_1}$ and $\boldsymbol{\gamma}_2 \in \mathbb{R}^{l_2}$ can be accurately estimated.

This flexible model structure can accommodate polygenic effects across population and large population-specific effects. On one hand, the probabilistic structures of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ capture the polygenic effects that are correlated between populations, allowing the auxiliary samples to be effectively utilized. On the other hand, $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ capture the large population-specific genetic effects, which further boost the prediction performance. We call this extension of XPA as XPA$_+$. When no SNPs with large population-specific effects are selected, XPA$_+$ is equivalent to XPA. Once the large-effect SNPs have been selected, the parameters can be estimated in the same way as in XPA.

Parameter estimation in XPA

To obtain the posterior mean of $\boldsymbol{\beta}_1$, we first need to estimate the unknown parameters $\{\sigma_1^2, \sigma_2^2, \delta, \sigma_\varepsilon^2, \sigma_\xi^2\}$. For convenience, we define $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}$, $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & 0 \\ 0 & \mathbf{Z}_2 \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & 0 \\ 0 & \mathbf{X}_2 \end{bmatrix}$, $\boldsymbol{\omega} = \begin{bmatrix} \boldsymbol{\omega}_1 \\ \boldsymbol{\omega}_2 \end{bmatrix}$, and $\boldsymbol{\Sigma}_e = \begin{bmatrix} \sigma_\varepsilon^2 \mathbf{I}_{n_1} & 0 \\ 0 & \sigma_\xi^2 \mathbf{I}_{n_2} \end{bmatrix}$. The marginal distribution of $\mathbf{y}$ can be obtained by combining Equations 1 and 2 and taking integration over $\boldsymbol{\beta}$:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{Z}\boldsymbol{\omega}, \boldsymbol{\Omega}), \quad \boldsymbol{\Omega} = \mathbf{X}(\boldsymbol{\Sigma}_\beta \otimes \mathbf{I}_p)\mathbf{X}^T + \boldsymbol{\Sigma}_e, \qquad \text{(Equation 4)}$$

where $\mathbf{X}(\boldsymbol{\Sigma}_\beta \otimes \mathbf{I}_p)\mathbf{X}^T = \begin{bmatrix} \sigma_1^2\mathbf{K}_1 & \delta\mathbf{K}_{12} \\ \delta\mathbf{K}_{12}^T & \sigma_2^2\mathbf{K}_2 \end{bmatrix}$, $\mathbf{K}_1 = \mathbf{X}_1\mathbf{X}_1^T$, $\mathbf{K}_2 = \mathbf{X}_2\mathbf{X}_2^T$, and $\mathbf{K}_{12} = \mathbf{X}_1\mathbf{X}_2^T$. To estimate unknown parameters $\{\sigma_1^2, \sigma_2^2, \delta, \sigma_\varepsilon^2, \sigma_\xi^2\}$, we get rid of the covariates by multiplying Equation 4 with the projection matrix $\mathbf{M} \equiv \begin{bmatrix} \mathbf{M}_1 & 0 \\ 0 & \mathbf{M}_2 \end{bmatrix} = \mathbf{I}_{n_1+n_2} - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$, which leads to $\mathbf{M}\mathbf{y} \sim \mathcal{N}(0, \mathbf{M}\mathbf{X}(\boldsymbol{\Sigma}_\beta \otimes \mathbf{I}_p)\mathbf{X}^T\mathbf{M} + \mathbf{M}\boldsymbol{\Sigma}_e)$. For convenience, we use $\tilde{\mathbf{A}}$ to denote $\mathbf{M}\mathbf{A}$ for any matrix $\mathbf{A}$ involved in our notation, hence we have $\tilde{\mathbf{y}} \sim \mathcal{N}(0, \tilde{\Omega})$, where $\tilde{\Omega} = \tilde{\mathbf{X}}(\boldsymbol{\Sigma}_\beta \otimes \mathbf{I}_p)\tilde{\mathbf{X}}^T + \tilde{\boldsymbol{\Sigma}}_e$. The method of moments (MoM) offers a computationally efficient estimator of $\boldsymbol{\theta} = \{\sigma_1^2, \sigma_2^2, \delta, \sigma_\varepsilon^2, \sigma_\xi^2\}$ by matching the second-order moment based

on the criterion of least-squares: $\text{argmin}_{\boldsymbol{\theta}} \left|\left| \tilde{\mathbf{y}}\tilde{\mathbf{y}}^T - \tilde{\Omega} \right|\right|_F^2$. Taking derivatives of the objective function w.r.t $\boldsymbol{\theta}$ and setting them to zero leads to the estimating equations (see Appendix A):

$$\underbrace{\begin{bmatrix} \text{tr}\left(\tilde{\mathbf{K}}_1^2\right) & \text{tr}\left(\tilde{\mathbf{K}}_1\right) & 0 & 0 & 0 \\ \text{tr}\left(\tilde{\mathbf{K}}_1\right) & \text{tr}(\mathbf{M}_1) & 0 & 0 & 0 \\ 0 & 0 & \text{tr}\left(\tilde{\mathbf{K}}_2^2\right) & \text{tr}\left(\tilde{\mathbf{K}}_2\right) & 0 \\ 0 & 0 & \text{tr}\left(\tilde{\mathbf{K}}_2\right) & \text{tr}(\mathbf{M}_2) & 0 \\ 0 & 0 & 0 & 0 & \text{tr}\left(\tilde{\mathbf{K}}_{12}\tilde{\mathbf{K}}_{12}^T\right) \end{bmatrix}}_{\mathbf{S}} \underbrace{\begin{bmatrix} \sigma_1^2 \\ \sigma_\varepsilon^2 \\ \sigma_2^2 \\ \sigma_\xi^2 \\ \delta \end{bmatrix}}_{\boldsymbol{\theta}} = \underbrace{\begin{bmatrix} \\ \tilde{\mathbf{y}}_1^T\tilde{\mathbf{K}}_1\tilde{\mathbf{y}}_1 \\ \tilde{\mathbf{y}}_1^T\tilde{\mathbf{y}}_1 \\ \tilde{\mathbf{y}}_2^T\tilde{\mathbf{K}}_2\tilde{\mathbf{y}}_2 \\ \tilde{\mathbf{y}}_2^T\tilde{\mathbf{y}}_2 \\ \tilde{\mathbf{y}}_1^T\tilde{\mathbf{K}}_{12}\tilde{\mathbf{y}}_2. \end{bmatrix}}_{\mathbf{q}} \quad \text{(Equation 5)}$$

The bottle neck of solving Equation 5 is computing the traces of squared kinship matrices, which has a computational complexity $\mathcal{O}(\max(n_1, n_2)^2 p)$. This computational overhead can become very large when dealing with biobank-scale data. Instead of computing the traces exactly, we apply a stochastic approximation to derive unbiased estimates of the trace terms.[37] Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and a vector of random variables $\mathbf{d} \in \mathbb{R}^n$ with $\mathbb{E}(\mathbf{d}) = 0$ and $\text{Cov}(\mathbf{d}) = \mathbf{I}_n$, we have the identity $\mathbb{E}(\mathbf{d}^T\mathbf{A}\mathbf{d}) = \text{tr}(\mathbf{A})$. Using this identity, we can construct the following estimates:

$$\mathbf{L}_{B_1} \equiv \widehat{\text{tr}\left(\tilde{\mathbf{K}}_1^2\right)} = \widehat{\text{tr}\left(\tilde{\mathbf{K}}_1\tilde{\mathbf{K}}_1^T\right)} = \frac{1}{B}\sum_{b=1}^{B} \mathbf{d}_b^T\tilde{\mathbf{K}}_1\tilde{\mathbf{K}}_1^T\mathbf{d}_b$$

$$= \frac{1}{B}\sum_{b=1}^{B} \| \tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1^T\mathbf{d}_b \|_2^2,$$

$$\mathbf{L}_{B_2} \equiv \widehat{\text{tr}\left(\tilde{\mathbf{K}}_2^2\right)} = \frac{1}{B}\sum_{b=1}^{B} \| \tilde{\mathbf{X}}_2\tilde{\mathbf{X}}_2^T\mathbf{d}_b \|_2^2, \quad \mathbf{L}_{B_{12}} \equiv \widehat{\text{tr}\left(\tilde{\mathbf{K}}_{12}\tilde{\mathbf{K}}_{12}^T\right)}$$

$$= \frac{1}{B}\sum_{b=1}^{B} \| \tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_2^T\mathbf{d}_b \|_2^2, \quad \text{(Equation 6)}$$

where $\mathbf{d}_1, \ldots, \mathbf{d}_B$ are B random vectors drawn independently from a distribution with zero mean and identity covariance matrix $\mathbf{I}_p$. The stochastic approximation in Equation 6 requires a computational complexity of $\mathcal{O}(\max(n_1, n_2)pB)$. In practice, we found that the traces can be effectively approximated with $B \sim 50$. Replacing $\text{tr}(\tilde{\mathbf{K}}_1^2)$, $\text{tr}(\tilde{\mathbf{K}}_2^2)$, and $\text{tr}(\tilde{\mathbf{K}}_{12}\tilde{\mathbf{K}}_{12}^T)$ by $\mathbf{L}_{B_1}$, $\mathbf{L}_{B_2}$, and $\mathbf{L}_{B_{12}}$, respectively, we obtained $\widehat{\boldsymbol{\theta}} = \mathbf{S}^{-1}\mathbf{q}$. Then heritablities of the given phenotype in the target and auxiliary populations are estimated as $\widehat{h}_1^2 = \frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_1^2 + \widehat{\sigma}_\varepsilon^2}$ and $\widehat{h}_2^2 = \frac{\widehat{\sigma}_2^2}{\widehat{\sigma}_2^2 + \widehat{\sigma}_\xi^2}$, respectively. The shared genetic basis between the two populations can be quantified by the estimated trans-ancestry genetic correlation, computed as $\widehat{\rho} = \frac{\widehat{\delta}}{\widehat{\sigma}_1\widehat{\sigma}_2}$, which is the key to utilizing information from the biobank-scale EUR data for risk prediction in the target population.

PRS construction in XPA

Given the estimated parameters $\widehat{\boldsymbol{\theta}} = \{\widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \widehat{\delta}, \widehat{\sigma}_\varepsilon^2, \widehat{\sigma}_\xi^2\}$, the fixed effects in Equation 1 are estimated by:

$$\widehat{\boldsymbol{\omega}} \equiv \begin{bmatrix} \widehat{\boldsymbol{\omega}}_1 \\ \widehat{\boldsymbol{\omega}}_2 \end{bmatrix} = \left(\mathbf{Z}^T\widehat{\Omega}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^T\widehat{\Omega}^{-1}\mathbf{y}, \quad \text{(Equation 7)}$$

where $\widehat{\Omega} = \mathbf{X}(\widehat{\boldsymbol{\Sigma}}_\beta \otimes \mathbf{I}_p)\mathbf{X}^T + \widehat{\boldsymbol{\Sigma}}_e, \widehat{\boldsymbol{\Sigma}}_\beta = \begin{bmatrix} \widehat{\sigma}_1^2 & \widehat{\delta} \\ \widehat{\delta} & \widehat{\sigma}_2^2 \end{bmatrix}$,

$$\widehat{\boldsymbol{\Sigma}}_e = \begin{bmatrix} \widehat{\sigma}_\varepsilon^2\mathbf{I}_{n_1} & 0 \\ 0 & \widehat{\sigma}_\xi^2\mathbf{I}_{n_2} \end{bmatrix}$$

.

The posterior mean of $\boldsymbol{\beta}_1$ can be derived as (see Appendix A):

$$\widehat{\boldsymbol{\mu}}_1^{\mathbf{XPA}} = \begin{bmatrix} \widehat{\sigma}_1^2\mathbf{X}_1 \\ \widehat{\delta}\mathbf{X}_2 \end{bmatrix}^T \underbrace{\begin{bmatrix} \widehat{\sigma}_1^2\mathbf{X}_1\mathbf{X}_1^T + \widehat{\sigma}_\varepsilon^2\mathbf{I}_{n1} & \widehat{\delta}\mathbf{X}_1\mathbf{X}_2^T \\ \widehat{\delta}\mathbf{X}_2\mathbf{X}_1^T & \widehat{\sigma}_2^2\mathbf{X}_2\mathbf{X}_2^T + \widehat{\sigma}_\xi^2\mathbf{I}_{n2} \end{bmatrix}^{-1}}_{\widehat{\Omega}^{-1} \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}} \begin{bmatrix} \mathbf{y}_1 - \mathbf{Z}_1\widehat{\boldsymbol{\omega}}_1 \\ \mathbf{y}_2 - \mathbf{Z}_2\widehat{\boldsymbol{\omega}}_2 \end{bmatrix}. \quad \text{(Equation 8)}$$

Obviously, when the genetic correlation $\rho \neq 0$ (i.e., $\delta \neq 0$), information could be borrowed from the auxiliary population to the target population to improve the posterior mean estimates.

Finally, to obtain the effect size of dosage genotypes $\tilde{\boldsymbol{\mu}}_1^{\mathbf{XPA}} \in \mathbb{R}^p$, we need to re-scale the posterior mean by $\tilde{\boldsymbol{\mu}}_{1j}^{\mathbf{XPA}} = \widehat{\boldsymbol{\mu}}_{1j}^{\mathbf{XPA}}/(s_j\sqrt{p})$, for $j = 1, \ldots, p$. When a new observation of genotype $\mathbf{g}_{1,new} \in \mathbb{R}^p$ is available, the associated PRS can be computed by its inner product with $\tilde{\boldsymbol{\mu}}_1^{\mathbf{XPA}}$, i.e., $\text{PRS}_{new} = \mathbf{g}_{1,new}^T\tilde{\boldsymbol{\mu}}_1^{\mathbf{XPA}}$. If the covariates of the new observation $\mathbf{z}_{1,new} \in \mathbb{R}^{c_1}$ are also available, we can predict the phenotype at its original scale by $\widehat{y}_{new} = \text{PRS}_{new} + \mathbf{z}_{1,new}^T\widehat{\boldsymbol{\omega}}_1 - \sum_j \tilde{\mu}_{1j}^{\mathbf{XPA}}\overline{\mathbf{g}}_{1j}$.

For XPA $_+$, we note that by constructing $\mathbf{Z}_1 = [\mathbf{Z}_{c1}, \mathbf{X}_{I1}]$ and $\mathbf{Z}_2 = [\mathbf{Z}_{c2}, \mathbf{X}_{I2}]$, the effect sizes of the pre-selected SNPs $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ can be estimated jointly with $\boldsymbol{\omega}_{c1}$ and $\boldsymbol{\omega}_{c2}$ in Equation 7, and the posterior mean in the target population $\widehat{\boldsymbol{\mu}}_1^{\mathbf{XPA}_+}$ can be computed in the same way as in Equation 8. Similarly, the posterior mean $\tilde{\boldsymbol{\mu}}_1^{\mathbf{XPA}_+}$ should be re-scaled in the same way as in XPA and the estimated effects of large-effect SNPs should be re-scaled by $\tilde{\boldsymbol{\gamma}}_{1j}^{\mathbf{XPA}_+} = \widehat{\boldsymbol{\gamma}}_{1j}^{\mathbf{XPA}_+}/(s_j\sqrt{p})$, for $j = 1, \ldots, l_1$. Then, the PRSs associated with the new observation

$\mathbf{g}_{1,new}$ is computed as $\text{PRS}_{new} = \mathbf{g}_{l1,new}^T \tilde{\boldsymbol{\gamma}}_1^{\mathbf{XPA}_+} + \mathbf{g}_{1,new}^T \tilde{\boldsymbol{\mu}}_1^{\mathbf{XPA}_+}$, where $\mathbf{g}_{l1,new} \in \mathbb{R}^{l_1}$ is the sub-vector of $\mathbf{g}_{1,new}$ corresponding to the SNPs with large effects.

Despite the simple form of Equation 8, it is highly nontrivial to obtain the posterior mean in biobank-scale data due to the challenges in both data storage in computer memory and computation. To boost the computational speed, we solve the large-scale linear system in Equation 8 (i.e., $(n_1+n_2)\times(n_1+n_2)$) using the efficient conjugate gradient (CG) method. Because the classical CG requires storing the whole $\widehat{\Omega}$ matrix, which is infeasible for bio-bank scale dataset, we developed a memory-efficient strategy (as described in the following section) for storing standardized genotype matrices, and designed a highly efficient CG algorithm (see Appendix A). Our CG algorithm has the time complexity of $\mathcal{O}(p(n_1+n_2)(c_1+c_2+1)\sqrt{\kappa})$, where κ is the condition number of $\widehat{\Omega}$. Because κ is usually small, the CG algorithm offers substantial computational improvement in solving the linear system.

Data storage for large genotype matrices in XPA

In practice, the sample size of the auxiliary dataset can be very large (e.g., $n_2 \approx 400,000$ for UKBB), producing both computation and data storage problems. To address this difficulty, we apply the memory-efficient Boolean representation proposed in our previous work BOOST[35] and the hash table data structure[36] to store the standardized genotypes.

Suppose that we are handling a standardized biobank-scale genotype matrix of 400,000 × 3,000,000, the memory usage of storing such a matrix in double-precision floating-point format is 8 bytes× $400,000\times3,000,000 = 9.6\times10^{12}$ bytes ≈ 8.7 Tb. Therefore, directly storing such a large matrix is usually impractical. However, we note that each genotype only takes 4 possible values even after standardization: $\frac{0-\overline{\mathbf{g}}_j}{s_j}, \frac{1-\overline{\mathbf{g}}_j}{s_j}, \frac{2-\overline{\mathbf{g}}_j}{s_j}$, and $\tilde{\mathbf{g}}_j$, where $\overline{\mathbf{g}}_j$ is the mean genotype value at SNP j, s_j is the standard deviation of the genotype j, and $\tilde{\mathbf{g}}_j$ is the value to be filled in for the missing genotypes. For each SNP, we index the standardized genotype values using a Boolean representation. Specifically, we can use 00 to index $\frac{0-\overline{\mathbf{g}}_j}{s_j}$, 01 to index $\frac{1-\overline{\mathbf{g}}_j}{s_j}$, 10 to index $\frac{2-\overline{\mathbf{g}}_j}{s_j}$, and 11 to index $\tilde{\mathbf{g}}_j$. We then save 4 consecutive individual's genotypes at a time using the 8-digit Boolean representation, which has $2^8 = 256$ combinations. This constructs a hash table of size 256×4, where each cell contains a standardized genotype value in double precision. The total size of this hash table is 256×4×8 bytes = 8,192 bytes. Finally, we need to construct such a hash table for all 3,000,000 SNPs, which amounts to 8,192 bytes ×3,000,000 = 24.576 Gb. In addition to the hash tables, we need to save the Boolean representations for retrieving genotypes, which takes 400,000 × 3,000,000 / 4 = 3×10^{11} bytes $\approx$ 280 Gb. In total, this memory-efficient storing strategy requires only 305 Gb for storing the UKBB genotype matrix.

Constructing PRS using summary statistics by XPASS

We consider the datasets $\{\mathbf{z}_1, \mathbf{G}_1', \mathbf{Z}_1'\}$ and $\{\mathbf{z}_2, \mathbf{G}_2', \mathbf{Z}_1'\}$ from the two populations. The vectors $\mathbf{z}_1 = [z_{11},\ldots,z_{1j},\ldots,z_{1p}]^T \in \mathbb{R}^p$ and $\mathbf{z}_2 = [z_{21},\ldots,z_{2j},\ldots,z_{2p}]^T \in \mathbb{R}^p$ contain the z-scores derived from the two populations, where $z_{1j} = \frac{(\mathbf{x}_{1j}^T\mathbf{x}_{1j})^{-1}\mathbf{x}_{1j}^T\mathbf{y}_1}{\sqrt{\widehat{\sigma}_{1j}^2(\mathbf{x}_{1j}^T\mathbf{x}_{1j})^{-1}}}$ and $z_{2j} = \frac{(\mathbf{x}_{2j}^T\mathbf{x}_{2j})^{-1}\mathbf{x}_{2j}^T\mathbf{y}_2}{\sqrt{\widehat{\sigma}_{2j}^2(\mathbf{x}_{2j}^T\mathbf{x}_{2j})^{-1}}}$, $\widehat{\sigma}_{1j}^2$ and $\widehat{\sigma}_{2j}^2$ are the residual variance of regressing $\mathbf{y}_1$ on $\mathbf{x}_{1j}$ and $\mathbf{y}_2$ on $\mathbf{x}_{2j}$, respectively. Following Vilhjálmsson et al.,[29] we assume the z-scores are derived from GWAS datasets with phenotype vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ standardized to have mean of zero and standard deviation of one. The first few PCs of the reference genotypes from the two populations are given in $\mathbf{Z}_1' \in \mathbb{R}^{m_1\times c_1}$ and $\mathbf{Z}_2' \in \mathbb{R}^{m_2\times c_2}$. Similar to XPA, we first standardize the reference genotype matrices $\mathbf{G}_1'$ and $\mathbf{G}_2'$ to obtain the corresponding $\mathbf{X}_1'$ and $\mathbf{X}_2'$ that have column means zero and variances $1/p$.

In real applications, the individual-level GWAS data may not be easily accessible. To effectively make use of publicly available GWAS summary statistics, we extend the XPA method as XPASS which requires only the z-scores from GWAS results and reference genotypes from the target and auxiliary populations. For XPASS, we consider the vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ of the z-scores for p SNPs derived from the target and auxiliary populations, respectively, and the $m_1 \times p$ matrix $\mathbf{X}_1'$ and $m_2 \times p$ matrix $\mathbf{X}_2'$ being the standardized reference genotype matrices from the corresponding populations. With these summary-level data, we show that XPASS approximates the posterior mean of SNP effect sizes in XPA as (see Appendix A):

$$\begin{bmatrix}\widehat{\boldsymbol{\mu}}_1^{\mathbf{XPASS}} \\ \widehat{\boldsymbol{\mu}}_2^{\mathbf{XPASS}}\end{bmatrix} = \left(\begin{bmatrix} n_1\widehat{\mathbf{R}}_1 & 0 \\ 0 & n_2\widehat{\mathbf{R}}_2\end{bmatrix} + \begin{bmatrix} \frac{\widehat{h}_1^2}{1-\widehat{h}_1^2} & \frac{\widehat{h}_{12}}{1-\widehat{h}_2^2} \\ \frac{\widehat{h}_{12}}{1-\widehat{h}_1^2} & \frac{\widehat{h}_2^2}{1-\widehat{h}_2^2}\end{bmatrix}^{-1} \otimes \mathbf{I}_p\right)^{-1} \begin{bmatrix}\sqrt{\frac{n_1}{p}}\mathbf{z}_1 \\ \sqrt{\frac{n_2}{p}}\mathbf{z}_2\end{bmatrix},$$

where $\otimes$ denotes the Kronecker product, $\widehat{h}_1^2$ and $\widehat{h}_2^2$ are estimates of heritabilities h_1^2 and h_2^2 for the two populations, respectively, $\widehat{h}_{12}$ is the estimate of co-heritability $h_{12} := \rho h_1 h_2$ between two populations, and $\widehat{\mathbf{R}}_1 = \mathbf{X}_1'^T\mathbf{X}_1'/m_1$ and $\widehat{\mathbf{R}}_2 = \mathbf{X}_2'^T\mathbf{X}_2'/m_2$ are the LD matrices of target and auxiliary populations, respectively. The parameter estimates $\widehat{h}_1^2$, $\widehat{h}_2^2$, and $\widehat{h}_{12}$ can be computed using the summary-level datasets (Appendix A).[39,40] Again, the information is shared across populations through the genetic correlation. In practice, XPASS takes the heterogeneous LD patterns into account by computing the LD matrices using $\mathbf{X}\prime_1$ and $\mathbf{X}\prime_2$ from either subsamples of $\mathbf{X}_1$ and $\mathbf{X}_2$ or external reference genotypes from the two populations. Because the LD between SNPs decreases exponentially with their distance, we approximate the LD matrices using block diagonal matrices by assuming the SNPs between LD blocks are approximately independent.[41] The heterogeneous LD patterns result in different LD partitions across populations. Therefore, we build the approximated LD matrices using partitions derived from either the target or the auxiliary population. The application of XPASS to real datasets suggests that our method is insensitive to the partition strategy. To obtain the dosage scale effect size, we compute $\tilde{\boldsymbol{\mu}}_{1j}^{\mathbf{XPASS}} = \widehat{\boldsymbol{\mu}}_{1j}^{\mathbf{XPASS}}/(s_j\sqrt{p}),\ j = 1,\ldots,p$. When the genotypes of a new sample is available, we can obtain the PRS by $\text{PRS}_{new} = \mathbf{g}_{1,new}^T\tilde{\boldsymbol{\mu}}_1^{\mathbf{XPASS}}$.

XPASS$_+$ for capturing large population-specific effects

Similarly to XPA$_+$, we extend XPASS as XPASS$_+$ to incorporate large population-specific effects to improve prediction performance. We denote $\mathbf{X}\prime_{l1} \in \mathbb{R}^{m_1\times l_1}$ and $\mathbf{X}\prime_{l2} \in \mathbb{R}^{m_1\times l_2}$ as the standardized genotype matrices collecting the columns of $\mathbf{X}\prime_1$ and $\mathbf{X}\prime_2$ that correspond to the SNPs with large effects, and $\mathbf{z}_{l1} \in \mathbb{R}^{l_1}$ and $\mathbf{z}_{l2} \in \mathbb{R}^{l_2}$ as sub-vectors of $\mathbf{z}_1$ and $\mathbf{z}_2$ corresponding to the SNPs with large effects, respectively. The large genetic effects can be estimated by XPASS$_+$ as (see Appendix A):

$$\begin{bmatrix}\widehat{\gamma}_1^{\mathbf{XPASS}_+}\\ \widehat{\gamma}_2^{\mathbf{XPASS}_+}\end{bmatrix}=\left(\begin{bmatrix}n_1\widehat{\mathbf{R}}_{l1} & 0\\ 0 & n_2\widehat{\mathbf{R}}_{l2}\end{bmatrix}-\begin{bmatrix}n_1\widehat{\mathbf{R}}_{ls1} & 0\\ 0 & n_2\widehat{\mathbf{R}}_{ls2}\end{bmatrix}\left(\begin{bmatrix}\frac{\widehat{h}_1^2}{1-\widehat{h}_1^2} & \frac{\widehat{h}_{12}}{1-\widehat{h}_2^2}\\ \frac{\widehat{h}_{12}}{1-\widehat{h}_1^2} & \frac{\widehat{h}_2^2}{1-\widehat{h}_2^2}\end{bmatrix}^{-1}\otimes\mathbf{I}_p+\begin{bmatrix}n_1\widehat{\mathbf{R}}_1 & 0\\ 0 & n_2\widehat{\mathbf{R}}_2\end{bmatrix}\right)^{-1}\begin{bmatrix}n_1\widehat{\mathbf{R}}_{sl1} & 0\\ 0 & n_2\widehat{\mathbf{R}}_{sl2}\end{bmatrix}\right)^{-1}$$

$$\left(\begin{bmatrix}\sqrt{\frac{n_1}{p}}\mathbf{z}_{l1}\\ \sqrt{\frac{n_2}{p}}\mathbf{z}_{l2}\end{bmatrix}-\begin{bmatrix}n_1\widehat{\mathbf{R}}_{ls1} & 0\\ 0 & n_2\widehat{\mathbf{R}}_{ls2}\end{bmatrix}\left(\begin{bmatrix}\frac{\widehat{h}_1^2}{1-\widehat{h}_1^2} & \frac{\widehat{h}_{12}}{1-\widehat{h}_2^2}\\ \frac{\widehat{h}_{12}}{1-\widehat{h}_1^2} & \frac{\widehat{h}_2^2}{1-\widehat{h}_2^2}\end{bmatrix}^{-1}\otimes\mathbf{I}_p+\begin{bmatrix}n_1\widehat{\mathbf{R}}_1 & 0\\ 0 & n_2\widehat{\mathbf{R}}_2\end{bmatrix}\right)^{-1}\begin{bmatrix}\sqrt{\frac{n_1}{p}}\mathbf{z}_1\\ \sqrt{\frac{n_2}{p}}\mathbf{z}_2\end{bmatrix}\right),\qquad\text{(Equation 9)}$$

where $\widehat{\mathbf{R}}_{l1}=\mathbf{X}_{l1}'^T\mathbf{X}_{l1}/m_1$ and $\widehat{\mathbf{R}}_{l2}=\mathbf{X}_{l2}'^T\mathbf{X}_{l2}/m_2$ are the LD matrices of large-effect SNPs and $\widehat{\mathbf{R}}_{ls1}=\mathbf{X}_{l1}'^T\mathbf{X}_1/m_1$ and $\widehat{\mathbf{R}}_{ls2}=\mathbf{X}_{l2}'^T\mathbf{X}_2/m_2$ are the SNP correlation matrices between large-effect SNPs and all SNPs from the two populations, respectively. With the estimated fixed effects given in Equation 9, the posterior means can be computed as:

$$\begin{bmatrix}\widehat{\mu}_1^{\mathbf{XPASS}_+}\\ \widehat{\mu}_2^{\mathbf{XPASS}_+}\end{bmatrix}=\left(\begin{bmatrix}n_1\widehat{\mathbf{R}}_1 & 0\\ 0 & n_2\widehat{\mathbf{R}}_2\end{bmatrix}+\begin{bmatrix}\frac{\widehat{h}_1^2}{1-\widehat{h}_1^2} & \frac{\widehat{h}_{12}}{1-\widehat{h}_2^2}\\ \frac{\widehat{h}_{12}}{1-\widehat{h}_1^2} & \frac{\widehat{h}_2^2}{1-\widehat{h}_2^2}\end{bmatrix}^{-1}\otimes\mathbf{I}_p\right)^{-1}\times\begin{bmatrix}\sqrt{\frac{n_1}{p}}\mathbf{z}_1-\widehat{\mathbf{R}}_{sl1}\widehat{\gamma}_1\\ \sqrt{\frac{n_2}{p}}\mathbf{z}_2-\widehat{\mathbf{R}}_{sl2}\widehat{\gamma}_2\end{bmatrix}.\qquad\text{(Equation 10)}$$

Finally, to obtain the effect sizes for the dosage genotypes, we rescale both the posterior mean and the estimated fixed effects by $\tilde{\mu}_{1j}^{\mathbf{XPASS}_+}=\widehat{\mu}_{1j}^{\mathbf{XPASS}_+}/(s_j\sqrt{p})$, for $j=1,\ldots,p$ and $\tilde{\gamma}_{1j}=\widehat{\gamma}_{1j}^{\mathbf{XPASS}_+}/(s_j\sqrt{p})$, for $j=1,\ldots,l_1$, respectively. When a new observation with genotype $\mathbf{g}_{1,new}\in\mathbb{R}^p$ is available, its PRS can be computed as $\mathrm{PRS}_{new}=\mathbf{g}_{1,new}^T\tilde{\boldsymbol{\mu}}_1^{\mathbf{XPASS}_+}+\mathbf{g}_{l1,new}^T\tilde{\boldsymbol{\gamma}}_1^{\mathbf{XPASS}_+}$, where $\mathbf{g}_{l1,new}\in\mathbb{R}^{l_1}$ is the vector of dosage genotypes corresponding to the SNPs with large effects.

Simulation design

We conducted a comprehensive simulation study to compare the performance of XPA and XPASS with other PRS approaches for individual-level data and summary data, respectively. For individual-level approaches, we investigated the prediction accuracy of XPA and XPA_+ in comparison with three scalable PRS models, including BLUP and bvBLUP implemented in the GCTA,[21] LASSO,[22] and XP-BLUP.[32] The BLUP and bvBLUP were fitted using the GCTA software v.1.93 (GCTA-BLUP and GCTA-bvBLUP). The LASSO was fitted using the R package glmnetPlus. For single-population-base approaches, GCTA-BLUP and LASSO, we trained them with either only the target dataset or only the auxiliary dataset. In addition, we trained GCTA-BLUP with the combined dataset obtained by directly merging the target and the auxiliary data (GCTA-BLUP-combine). For GCTA-bvBLUP, we followed the instruction from the GCTA forum for integrating two independent samples. For XP-BLUP, we considered 6 settings of p value threshold for selecting candidate SNPs from the auxiliary GWAS results: 5×10^{-6}, 1×10^{-6}, 5×10^{-7}, 1×10^{-7}, 5×10^{-8}, 1×10^{-8}. We revised the original XP-BLUP script to allow for the support of multi-threading computation. For XPA_+, we selected the large-effect SNPs by applying the P+T procedure to the target dataset with LD threshold $r^2=0.1$ and p value threshold 1×10^{-6}.

For summary-level approaches, we compared XPASS and XPASS_+ with three alternative summary-level PRS models: P+T procedure, LDpred-inf,[29] and lassosum.[30] As the performance of non-infinitesimal LDpred is often similar to lassosum,[30,42] we only considered LDpred-inf, a special case of LDpred with closed form solution. The LDpred-inf was computed using the ldpred software v.1.0.11 with LD radius set at the recommended value 300,000/3,000 = 100. For P+T procedure, we set the region size as 1,000 kb and the LD threshold as 0.1 and considered 10 p value thresholds according to a previous study:[43] 5×10^{-8}, 1×10^{-6}, 1×10^{-4}, 0.001, 0.01, 0.05, 0.1, 0.2, 0.5, and 1. We used R package ieugwasr to compute effect sizes in P+T procedure. The lassosum was fitted using the R package v.0.4.4 with default tuning parameter settings. Because these summary-level methods cannot handle multi-ancestry datasets, we trained these models with either only the target dataset or only the auxiliary dataset. To assess the prediction utility of MTAG, we applied MTAG to combine the datasets from the target and auxiliary populations, and then applied LDpred-inf to construct PRS (MTAG+LDpred-inf). For XPASS_+, we selected the large-effect SNPs by applying the P+T procedure to the target dataset with LD threshold $r^2=0.1$ and p value threshold 1×10^{-6}.

To mimic realistic LD patterns, we used the genotypes from the Chinese and UKBB samples to generate the target and auxiliary individuals, respectively. For the target population, 5,000 samples were randomly drawn from the Chinese dataset. For the auxiliary population, we explored seven different sample sizes using

random samples from the UKBB dataset: 0, 5,000, 10,000, 30,000, 50,000, 70,000, and 90,000. We included 300,000 SNPs in total by selecting the first 30,000 SNPs from each of chromosomes 1 to 10. Given these SNPs, we simulated their effect sizes with heritability $h_1^2 = h_2^2 = 0.5$ and generated the phenotypes in both populations. To investigate a wide spectrum of genetic architectures, we varied the proportion of non-zero genetic effects and the genetic correlation between the two populations. Specifically, we set the proportion of non-zero genetic effects to be 0.9, 0.01, or 0.001, corresponding to the highly polygenic scenario, the moderately sparse scenario, and the sparse scenario, respectively. We considered three settings of the overall genetic correlation ρ: 0, 0.4, and 0.8, corresponding to no, moderate, and high genetic correlation, respectively. As the effect sizes may not be correlated for all SNPs, we generated the nonzero effects by simulating 80% of them from the bi-variate normal distribution $\mathcal{N}\left(0, \begin{bmatrix} \frac{h_1^2}{p} & \frac{\rho h_1 h_2}{0.8p} \\ \frac{\rho h_1 h_2}{0.8p} & \frac{h_2^2}{p} \end{bmatrix}\right)$ and the rest from two independent normal distributions $\mathcal{N}\left(0, \frac{h_1^2}{p}\right)$ and $\mathcal{N}\left(0, \frac{h_2^2}{p}\right)$ for populations one and two, respectively, where p is the number of nonzero effects. By the combinatorial configurations of the proportion of non-zero effects and genetic correlation, nine scenarios were considered in our analysis. To evaluate the prediction performance, we sampled 3,000 additional individuals from the Chinese dataset serving as the test set of the target population. For each simulation setting, we computed the averaged prediction R^2 from 10 replications.

Among the individual-level methods considered, XPA, GCTA-BLUP, GCTA-bvBLUP, and XP-BLUP support multi-threading computation. To examine the computational and memory efficiency of these methods, we further evaluated their CPU time and memory usage when different numbers of SNPs were included in the model: 100,000, 200,000, and 300,000. The analyses were performed with 16 threads on the platform of Intel Xeon Gold 6152 CPU.

Application of XPA and XPASS to predict height and BMI in the Chinese population

Following the simulation studies, we applied XPA and XPASS to construct PRSs for height and BMI in the Chinese population by integrating European samples from UKBB and Asian samples from Chinese cohort and BBJ. In the individual-level PRS analysis, we split the Chinese data into the training and testing sets and only included the SNPs overlapping with UKBB. For height, we used 21,069 samples for training and held out 11,852 samples for testing. After quality control and overlapping, 3,776,575 SNPs were used to fit the model. For BMI, we used 18,575 samples for training and 10,572 for testing. After quality control and overlapping, 3,777,871 SNPs were used to fit the model. For Chinese population, we included age, sex, and first 10 principal components as covariates. For UKBB, we used the top 20 principal components, age, squared age, sex, genotyping arrays, and sequencing platforms as covariates.

To benchmark the performance of XPA and XPA$_+$ with existing approaches, we compared the predictive R^2 of XPA and XPA$_+$ with six other PRS models. Four of the six methods are designed only for single population analysis, including BLUP, which serves as a baseline model; snpnet, a memory-efficient LASSO implementation for large-scale genetic prediction based on R packages glmnetPlus and glmnet;[26] BayesR, a hierarchical Bayesian mixture model;[23,24] and BayesS, which accounts for the impact of natural selection.[25] We trained all these four models on the Chinese dataset, and additionally trained BLUP on the UKBB dataset using our efficient implementation. The fifth method, GCTA-bvBLUP, was trained with all Chinese samples and 150K subsamples randomly drawn from the UKBB dataset because the large memory requirement of GCTA-bvBLUP makes it infeasible to include more UKBB samples (see the results section for details). We also included a recently proposed approach for cross-population prediction,[32] XP-BLUP, in our comparison. We implemented the conventional BLUP method in our XPA software, making it more efficient and scalable to biobank-scale data. Since parameter-tuning is required in snpnet, we randomly took one-third of training samples as validation set and fitted the LASSO model on the rest of the training samples. The GCTB v2.0 was used to fit the BayesR and BayesS models with 25,000 MCMC iterations in total and 5,000 burn-in iterations. We set the initial value of heritability at 0.3 for height and 0.15 for BMI. In BayesS, we set the initial value of the proportion of non-zero effects at 0.05 for both height and BMI. The XP-BLUP requires a list of candidate SNPs selected from the UKBB GWAS results by taking a threshold of p values. We considered eight thresholds for selecting SNPs: 5×10^{-6}, 1×10^{-6}, 5×10^{-7}, 1×10^{-7}, 5×10^{-8}, 1×10^{-8}, 5×10^{-9}, and 1×10^{-9}. After tuning, the optimal thresholds were found at 1×10^{-8} and 5×10^{-7} for height and BMI, respectively. For XPA$_+$, we set the LD threshold at $r^2 = 0.1$ and the p value threshold at 5^{-8}.

In the summary-level PRS analysis, we obtained the summary statistics of the Chinese dataset and UKBB using the BOLT-LMM software[36] and additionally included summary statistics from BBJ (~170,000 Japanese samples)[33,34] as an alternative training data from the target population. We took the intersection of SNPs in the Chinese dataset, UKBB and BBJ, leading to 3,621,504 SNPs to be included in height and 3,562,502 SNPs to be included in BMI, respectively. We first randomly sub-sampled 2,000 individuals form both UKBB and the Chinese dataset as the LD reference panels for the two populations.

Here we mainly compared XPASS and XPASS$_+$ with LDpred and P+T because other related methods with different assumptions on effect sizes, such as SBayesR,[43] are expected to have very minor improvement, as shown in the aforementioned individual-level analysis. Because LDpred and P+T were developed for single-population analysis, we considered two ways for training the PRS models. The first way was to train the models separately using the Chinese cohort, BBJ, or UKBB. The second way was to first combine the target and auxiliary datasets using MTAG and then train the models with the combined datasets. When MTAG was applied, a covariance matrix of the estimation error should be first constructed and provided as an input.[31] Because the two datasets were from different populations, we applied LD score regression to estimate the intercepts of the Chinese cohort and UKBB summary datasets with their corresponding reference genotypes, respectively. After that, the estimated intercepts were used to construct the diagonal elements of the covariance matrix. The off-diagonal elements of the covariance matrix were set to zero because there was no sample overlap between populations. For XPASS and XPASS$_+$, we considered two configurations of the training sets, i.e., Chinese + UKBB and BBJ + UKBB. Assuming only the SNPs within the same LD block are correlated, XPASS approximates the LD matrices by partitioning

the genome into nearly independent blocks. Because the LD block partition is not aligned in EAS and EUR,[41] we used the LD block partition derived from both EAS and EUR to construct PRS and then evaluated the sensitivity of PRS to the two of LD block partition strategies.

Because both LDpred and P+T have tuning parameters, we considered a number of parameter settings and determine the optimal values by evaluating prediction performance on the test set. For LDpred, we considered nine settings of the proportion of non-zero effects: 1×10^{-4}, 5×10^{-4}, 1×10^{-3}, 5×10^{-3}, 1×10^{-2}, 5×10^{-2}, 1×10^{-1}, 5×10^{-1}, and 1. For height, the optimal values turned out to be 5×10^{-2} in Chinese, 1 in BBJ, 1 in UKBB, 10^{-1} in MTAG-Chinese, and 1 in MTAG-UKBB (Figures S25 and S29). For BMI, the optimal values were 1×10^{-3} in Chinese, 5×10^{-1} in BBJ, 1 in UKBB, 5×10^{-1} in MTAG-Chinese, and 1 in MTAG-UKBB (Figures S26 and S30).

For P+T procedure, we set the LD threshold at $r^2 = 0.1$ and considered ten settings of the proportion of p value threshold: 5×10^{-8}, 1×10^{-6}, 1×10^{-4}, 1×10^{-3}, 1×10^{-2}, 5×10^{-2}, 1×10^{-1}, 2×10^{-1}, 5×10^{-1}, and 1. For height, the optimal values turned out to be 1×10^{-4} in Chinese, 1×10^{-3} in BBJ, 1×10^{-4} in UKBB, 5×10^{-2} in MTAG-Chinese, and 1×10^{-2} in MTAG-UKBB (Figures S25 and S29). For BMI, the optimal values were 1×10^{-6} in Chinese, 5×10^{-2} in BBJ, 1×10^{-2} in UKBB, 1×10^{-1} in MTAG-Chinese, and 1×10^{-2} in MTAG-UKBB (Figures S26 and S30). Since the parameter tuning process involved the testing data, the performance of LDpred and P+T reported here could be slightly optimistic.

When XPASS$_+$ was applied, we set the LD threshold $r^2 = 0.1$ and varied the p value threshold at $\{10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}\}$ to include large population-specific effects. The selected SNPs are treated as covariates only in the target population. As shown in Figure S33, the prediction performance of XPASS$_+$ was insensitive to the p value threshold. Therefore, we reported the results obtained with p value threshold at 10^{-6} and used 10^{-6} as the default setting in the XPASS software.

We used the prediction R^2 to examine the prediction performance. Given a test set $\{\mathbf{y}_{new}, \mathbf{G}_{new}, \mathbf{Z}_{new}\}$, we used the predictive R^2 to evaluate the performance of PRS. The PRS for the testing samples is constructed by $\mathrm{PRS}_{new} = \mathbf{G}_{new}\tilde{\boldsymbol{\mu}}_1$, where $\tilde{\boldsymbol{\mu}}_1$ is the posterior mean of the effect sizes on the dosage genotypes. Then, the predictive R^2 of PRS was defined as the squared correlation between PRS_{new} and the residual obtained by regressing $\mathbf{y}_{new}$ on $\mathbf{Z}_{new}$, denoted as $\mathbf{y}_{res}$:

$$R^2_{\mathrm{PRS}} = corr(\mathbf{y}_{res}, \mathrm{PRS}_{new})^2.$$

When the covariates were taken into account, XPA generated prediction on the original scale of phenotype: $\hat{\mathbf{y}}_{new} = \mathrm{PRS}_{new} + \mathbf{Z}_{1,new}\hat{\boldsymbol{\omega}}_1 - \sum_j \tilde{\mu}_{1j}^{\mathbf{XPA}}\overline{\mathbf{g}}_{1j}$. In this case, we evaluated the R^2 by computing the squared correlation between $\mathbf{y}_{new}$ and $\hat{\mathbf{y}}_{new}$:

$$R^2_{\hat{\mathbf{y}}} = corr\left(\mathbf{y}_{new}, \hat{\mathbf{y}}_{new}\right)^2.$$

Collection, genotyping, and imputation of Chinese sample

To evaluate the prediction performance of our framework, we have collected genotypes of Chinese individuals from the WeGene platform and participants have signed the consent form. The study was reviewed and approved by the Committee on Research Practices of HKUST in strict compliance with regulations regarding ethical considerations and personal data protection. To comply with the regulations of the Human Genetic Resources Administration of China (HGRAC), all Chinese genotypic and phenotypic data were processed and analyzed in a server located in Shenzhen, China. Researchers who request access to the summary statistics from the Chinese samples must get permission from Ministry of Science and Technology of the People's Republic of China.

DNA extraction and genotyping were performed on saliva samples. A total of 35,908 Chinese participants were genotyped on the Illumina or Affymetrix platforms. Among all participants, 21,830 were genotyped on the Affymetrix WeGene V1 Arrays covering 596,744 SNPs at the WeGene genotyping center, Shenzhen. The WeGene V1 Array optimized the identification of all known paternal and maternal lineages through adding EAS-relevant 18,963 Y chromosome and 4,448 mtDNA phylogenetic SNPs.[44] The remaining 14,078 individuals were genotyped on Illumina WeGene V2 Arrays with a total of about 700,000 SNPs at the WeGene genotyping center, Shenzhen. The WeGene V2 array was designed based on the Illumina Infinium Global Screening Array.

Imputation of unobserved genetic variants was performed using the 1000 Genomes Project Phase 3 reference panel.[15] All datasets were phased by SHAPEIT2[45] and imputed by IMPUTE2[46] using regular steps and parameters. SNP-level (INFO score>0.5) and genotype-per-participant-level (genotype probability>0.9) filters were used to exclude poorly imputed variants.

Sample description of the Chinese cohort

The 35,908 genotyped Chinese participants in the Chinese cohort cover 33 out of 34 administrative divisions and 43 out of 56 ethnic groups in China, with the majority of samples from the Southeastern area (Figure S2). Among the 28,796 participants with self-reported ethnic information, there are 26,953 (93.6%) Han Chinese, 440 (1.5%) Manchu, 385 (1.3%) Hui, 201 Mongols (0.7%), and 817 (2.8%) from other minority ethnic groups.

To explore the population structure of the Chinese cohort, we first combined genotypes data from the Chinese cohort and the 1000 Genomes Project and performed a principal component analysis (PCA). As shown in Figure S2, the Chinese samples are overlapped with EAS from the 1000 Genomes Project, while its variance is larger because it is comprised of both Han Chinese and multiple minority groups. We then carried out PCA within Chinese to study differentiation across minority ethnic groups in China. Since the Han Chinese dominates the sample makeup, directly applying PCA to all samples fails to capture the variation among minority groups. Therefore, we first obtained the PC loadings from a subset of samples that included 500 randomly selected Han Chinese (roughly matching the number of Manchu people) and then computed the PC scores of all samples using these PC loadings (Figure S2). The first two principal components represent the latitudinal and longitudinal differentiation behind Chinese population structure. The Han Chinese differs substantially along the latitudinal gradient, while less differentiation is found along the longitude direction, which is consistent with previous reports.[47–49] Among the minority ethnic groups, Manchus are genetically closest to the Han Chinese in the northeastern China. In the same area, Koreans in China are more distant to Han Chinese compared to Manchus. The most differentiated groups are Mongols, Hui, and Tibetan from the northwestern area. In the

South China, Zhuang people also differ substantially from Han Chinese.

We considered two anthropometric phenotypes in our analysis, height and BMI. After quality control of genotype and phenotype data (see supplementary note), there were 32,921 samples with self-reported height and 29,147 samples with BMI computed from height and weight. For both phenotypes, there are slightly fewer males than females (15,406 compared to 17,515 for height and 13,721 compared to 15,426 for BMI). The overall distributions of height and BMI are summarized in Figure S1. Regarding the geometric distribution, the Northern Chinese are generally higher in both males and females (Figures S2 and S4). A similar latitudinal differentiation is observed in BMI, where the individuals from the north have higher obesity indices than those from the south (Figures S2 and S6). Besides, the older people are generally shorter and tend to have higher BMI (Figures S5 and S7).

Sample description of UKBB data

The UKBB genotype and phenotype data were obtained from the UK Biobank Access Management System (see web resources). We used the measured height phenotype extracted from Data Field 50 and the BMI value constructed from height and weight from Data Field 21001. To restrict the samples within EUR ancestry, we identified the individuals with self-reported ethnic background as "white" in Data Field 21000 and included only these samples for analysis. After phenotype and genotype quality control (see supplemental note), the UKBB data contain genotype information of 3,776,575 SNPs for 429,312 individuals in the analysis of height and 3,777,818 SNPs for 428,864 individuals in the analysis of BMI. We obtained the summary statistics of the UKBB datasets using the BOLT-LMM software[36] with age, squared age, sex, and the first 20 genomic PCs as covariates. The same set of covariates was included in the construction of PRS when applying XPA.

Sample description of GWAS data from African population

To demonstrate the generality of our framework to other populations, we applied XPA and XPASS to two GWAS datasets comprising thousands of samples from African ancestry: Institute for Personalized Medicine (IPM) BioMe biobank[10] (phs000925.v1.p1) and UKBB. IPM BioMe biobank aimed to study clinical care processes, with 28% samples of which were African Americans. The African participants from the IPM BioMe cohort were genotyped on Illumina Human OmniExpressExome Chip with a total of about 500K SNPs that passed an initial quality control (QC) process. Imputation of unobserved genetic variants was performed using the 1000 Genomes Project Phase 3 reference panel. All datasets were phased by SHAPEIT2 and imputed by Minimac4[46,50] using regular steps and default parameters.

After removing ancestry and phenotype outliers and samples with ambiguous sex (see supplemental note), 5,491 confirmed African participants from IPM were included in our study. For UKBB, 3,323 participants with self-reported African ancestry were also included, after the same procedure of sample QC, 2,931 samples remained for our analysis. By projecting the genotypes of IPM and UKBB samples to the PC coordinates derived from the 1000 Genomes Project, we found that the Africans from both datasets overlapped with the AFR samples from the 1000 Genomes Project (Figure S3). We observed similar phenotypic distributions for height between IPM and UKBB Africans, before or after we regressed the covariates (e.g., age, squared age, sex, first 20 genomic PCs) out.

To evaluate the performance of PRS approaches, we combined the African samples from IPM and UKBB, leading to a total of 8,422 African samples with 2,690,737 overlapping SNPs. Then, we randomly selected 1K samples as testing data and used the remaining 7.4K samples as training data. We obtained the summary statistics of the training set using the BOLT-LMM software[36] with age, squared age, sex, and the first 20 genomic PCs as covariates. The same set of covariates was included in the construction of PRSs when applying XPA.

Results

Simulation study

With data simulated as described above, we investigated the prediction accuracy of XPA in comparison with alternative methods. To gain some intuition, we first considered the single-population-based methods, such as BLUP and LASSO. These predictive models can be trained using samples from either the target population or the auxiliary population. The performance of BLUP and LASSO trained on the target population can serve as the reference results (dashed lines in Figure 1A). When the genetic correlation was zero, the prediction accuracy of BLUP and LASSO trained on the auxiliary dataset could not be improved regardless of the auxiliary sample size. When the genetic correlation became moderate ($\rho = 0.4$) or strong ($\rho = 0.8$), the prediction accuracy of BLUP and LASSO trained on the auxiliary dataset steadily improved as the auxiliary sample size increased. It is worthwhile to note that BLUP and LASSO trained on the auxiliary dataset were more accurate than those trained on the target population when the correlation was strong and the auxiliary sample size was large.

Among the methods that combine both datasets, GCTA-BLUP-combine had the worst performance in most settings. As expected, when there was no genetic correlation, the inclusion of auxiliary dataset led to worse performance than using only the target dataset. When the genetic correlation was nonzero, the predictive R^2 first dropped and then increased with increasing auxiliary sample size. When the auxiliary sample size was large enough (e.g., $n_2 > 30$K), the performance of GCTA-BLUP-combine gradually converged to GCTA-BLUP. For GCTA-bvBLUP, it had similar predictive R^2 with XPA when the auxiliary sample size was comparable with the target sample size (i.e., $n_2 \in \{0, 5K, 10K\}$). However, GCTA-bvBLUP does not account for the allele frequency difference between two populations. Therefore, as we can expect, its prediction accuracy became worse than XPA when the auxiliary sample size was large (i.e., $n_2 > 30$K). Between the cross-population methods XPA and XP-BLUP, XPA was clearly the overall winner, as shown in Figure 1A. When the genetic correlation became moderate ($\rho = 0.4$) or strong ($\rho = 0.8$), the prediction accuracy of XPA steadily improved as the sample size of auxiliary population increased, suggesting that

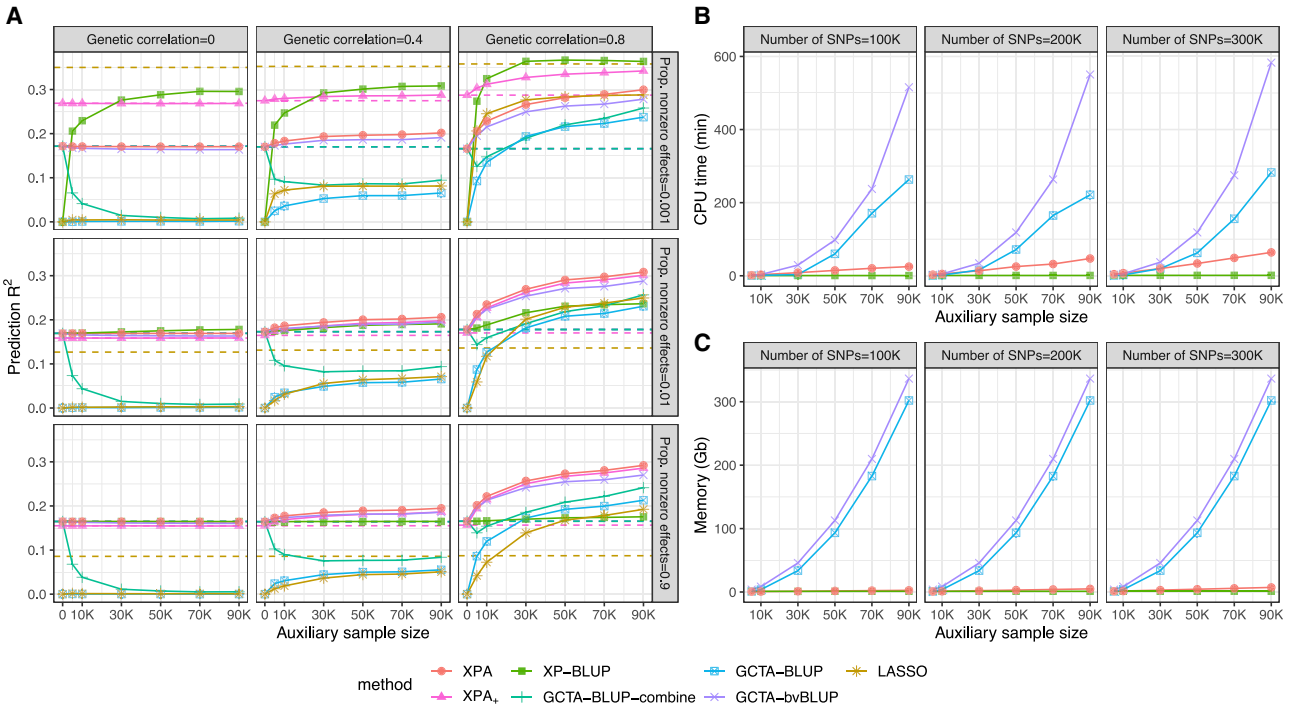

Figure 1. Comparison of individual-level approaches in simulation studies
(A) Mean predictive R^2 of XPA, XPA$_+$, GCTA-BLUP, LASSO, and XP-BLUP in each of nine simulation scenarios. The dashed lines show the R^2 obtained by training with target dataset only. For XPA, XPA$_+$, XP-BLUP, and GCTA-bvBLUP, the solid lines show the R^2 obtained by combining both target and auxiliary datasets. For GCTA-BLUP-combine, the solid line shows the R^2 obtained by merging the target and auxiliary datasets. For GCTA-BLUP and LASSO, the solid lines show the R^2 obtained by training with auxiliary dataset only.
(B) CPU timings for XPA, XP-BLUP, and GCTA-BLUP are shown for increasing auxiliary sample size based on different numbers of SNPs.
(C) Memory usages for XPA, XP-BLUP, and GCTA-BLUP are shown for increasing auxiliary sample sizes based on different numbers of SNPs. Results are summarized from ten replicates.

XPA was able to leverage the trans-ancestry genetic correlation for constructing PRSs. When either the genetic correlation or the auxiliary sample size approached zero, XPA reduced to BLUP which was trained on the target dataset, as no information could be borrowed from the auxiliary population in these cases. For XP-BLUP, it assumes that the top SNPs from the auxiliary dataset are more likely to have non-zero effects in the target population, but it does not model the correlation of effect sizes between populations. Therefore, it was worse than XPA in the polygenic and moderately sparse settings, but had better performance in the setting of highly sparse effects. However, in the highly sparse setting, XPA$_+$ extension achieved comparable performance with XP-BLUP by incorporating large population-specific effects. We also noted that the causal SNPs between the target and auxiliary populations largely overlapped in our simulation setting, which was preferred by XP-BLUP. We expect the performance of XPA$_+$ will be better than XP-BLUP when the causal SNPs with large effects are not largely overlapping between the two populations.

Among methods with the support of multi-threading computation (Figure 1B), XP-BLUP had the lowest computational cost and memory usage, as its computation was mostly based on the target dataset. While GCTA-BLUP, GCTA-bvBLUP, and XPA all had increased computational cost as the scale of the auxiliary dataset became larger, the CPU time of XPA was nearly linear in the sample size, providing higher efficiency than GCTA-BLUP and GCTA-bvBLUP when the auxiliary sample size was larger than 50,000. In addition, because we have adopted a memory-efficient strategy in storing genotypes, the memory cost of XPA was also linear in the auxiliary sample size and the number of SNPs. However, the memory cost of GCTA-BLUP and GCTA-bvBLUP increased quadratically with the sample size, limiting its usage in the biobank-scale data analysis.

Next, we compared XPA with XPASS which takes the summary-level data as its input. When the auxiliary data are not available, XPA and XPASS reduce to their special cases, BLUP and LDpred-inf, respectively. When the auxiliary sample size increased in our simulation, we measured the relative improvement of individual-level and summary-level approaches as the difference of prediction R^2 between XPA and BLUP, and that of XPASS and LDpred-inf, respectively. As shown in Figure 2B, XPA was slightly better than XPASS, suggesting that XPASS can provide comparable prediction improvement using only summary-level datasets. The advantage of XPA became more apparent as the auxiliary sample size increased, suggesting the importance of developing methods to handle biobank-scale individual-level data. These observations highlight

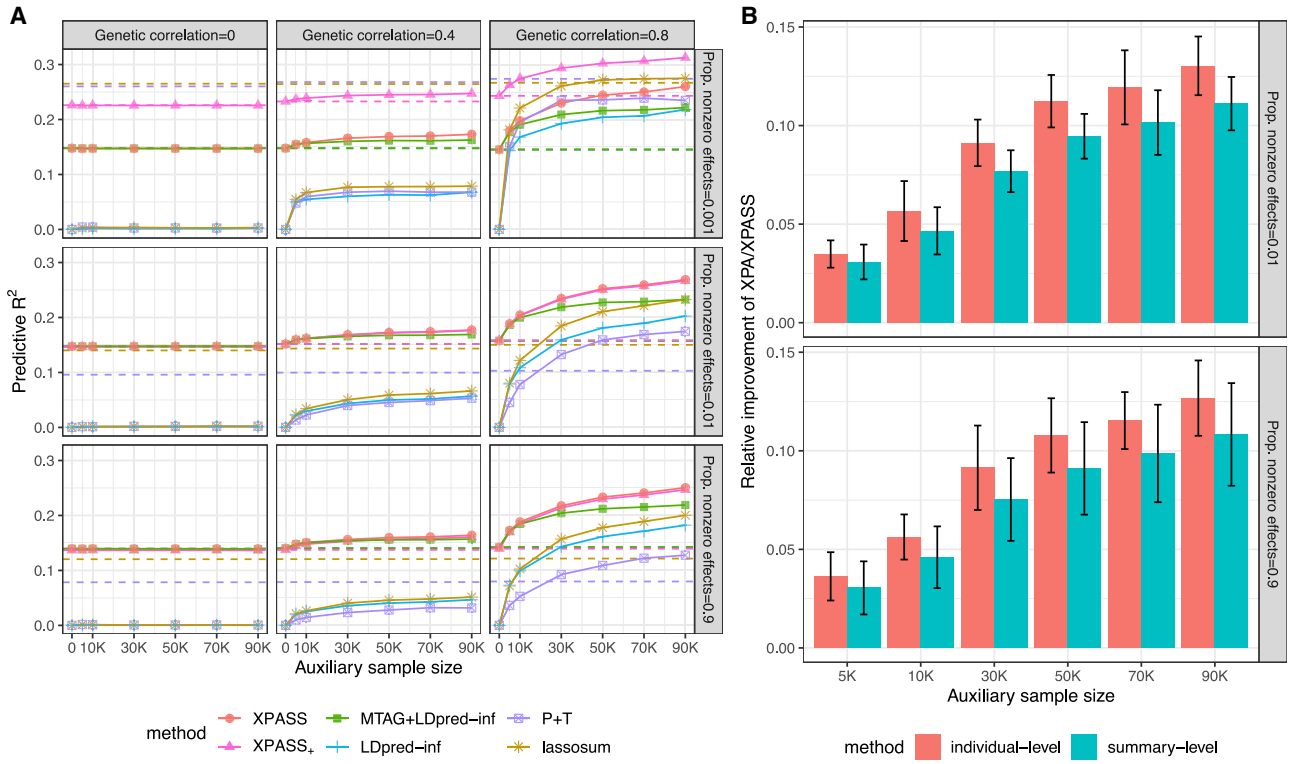

Figure 2. Comparison of summary-level approaches in simulation studies
(A) Mean prediction R^2 in each of nine simulation scenarios. Compared methods include XPASS, XPASS$_+$, LDpred-inf, MTAG+LDpred-inf, P+T procedure, and lassosum. The dashed lines show the R^2 obtained by training with target dataset only. For XPASS, XPASS$_+$, and MTAG+LDpred-inf, the solid lines show the R^2 obtained by combining both target and auxiliary datasets. For other methods, the solid lines show the R^2 obtained by training with auxiliary dataset only.
(B) Relative improvement in predictive R^2 of XPA and XPASS as compared to GCTA-BLUP and LDpred-inf, respectively. Results are summarized from ten replications. Error bars represent ± 1.96 of the standard error.

the value of XPA and XPASS in different practical scenarios: XPASS provides well-powered PRSs using cross-population information based on summary-level datasets while XPA can achieve higher accuracy with the availability of individual-level datasets.

As a promising approximation to XPA, XPASS also outperformed existing summary-level PRS models. As shown in Figure 2, XPASS had nearly the same performance with LDpred-inf when either the genetic correlation was zero or the auxiliary dataset was unavailable, consistent with previous observations for individual-level methods. With the availability of the auxiliary dataset and non-zero genetic correlation, XPASS achieved the highest prediction R^2 among all compared methods in the polygenic and moderately sparse settings. In the highly sparse setting, its extension XPASS$_+$ had the best performance when the genetic correlation was high ($\rho = 0.8$) and was comparable to alternative approaches with smaller genetic correlation ($\rho = 0$ and 0.4). The prediction accuracy of both XPASS and XPASS$_+$ increased with larger auxiliary sample size and stronger genetic correlation. Of note, the improvement of XPASS had a very similar pattern in both the sparse scenario and the polygenic scenario, suggesting the robustness of XPASS to different genetic architectures.[51] For models trained on the auxiliary dataset, P+T procedure had the lowest prediction accuracy, followed by LDpred-inf. Because lassosum adopts an elastic net model, it had comparable R^2 to LDpred-inf in the sparse scenario and outperformed LDpred-inf with larger sample size.

Construction of PRS for the Chinese population by XPA using the individual-level data from the Chinese cohort and UKBB

To study the performance of XPA and XPASS in real applications, we applied our approaches to construct PRSs for height and BMI in the Chinese population by integrating Chinese and UKBB data. We first investigate the performance of XPA using the individual-level data from Chinese and UKBB. To benchmark the performance of XPA with existing approaches, we compared the predictive R^2 of XPA with five other PRS models. XPA estimated the genetic correlations between Chinese and UKBB as 0.71 for height ($\widehat{h}^2_{\text{Chinese}} = 33.6\%$ with SE $= 1.8\%$, $\widehat{h}^2_{\text{UKBB}} = 41.2\%$ with SE $= 0.7\%$) and 0.66 for BMI ($\widehat{h}^2_{\text{Chinese}} = 16.7\%$ with SE $= 1.7\%$, $\widehat{h}^2_{\text{UKBB}} = 18.1\%$ with SE $= 0.1\%$), suggesting substantial genetic sharing between the two populations. As summarized in Figures 3A and 3D, regarding the overall performance, XPA had the highest accuracy for both height and

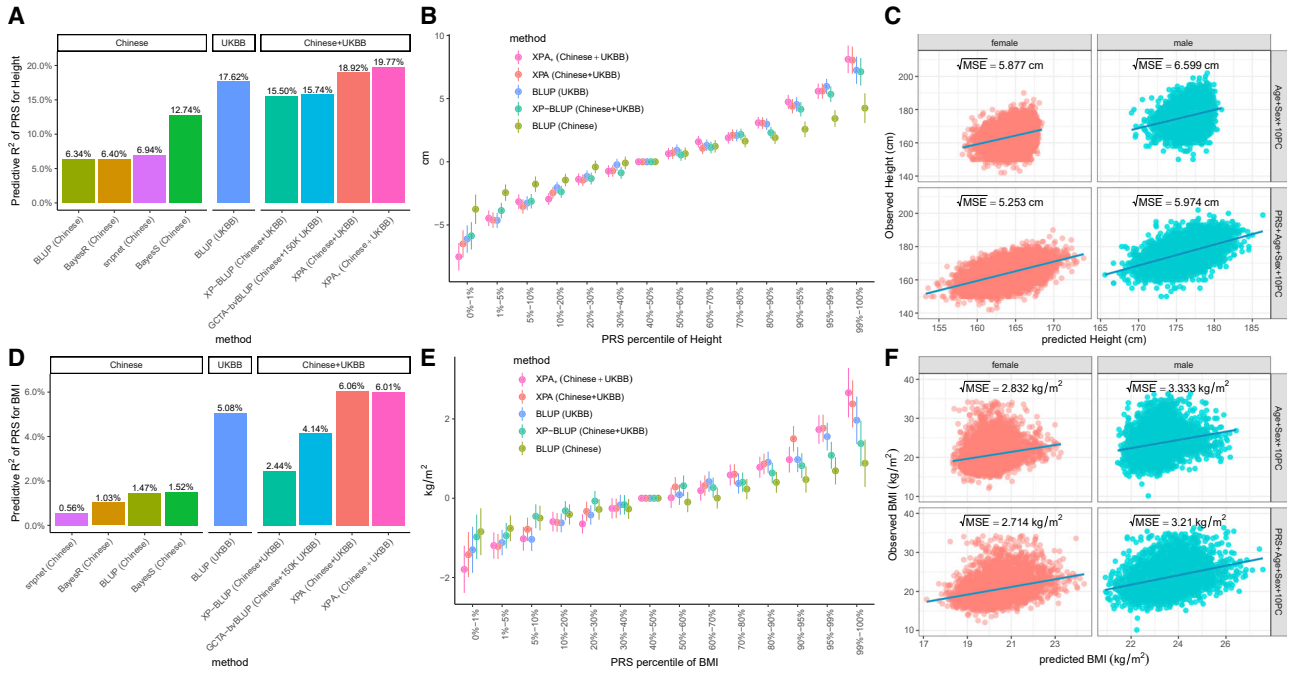

Figure 3. Prediction performance of XPA and related individual-level methods for height and BMI in the Chinese population
Predictive R^2 for height and BMI are shown in (A) and (D). Stratification ability of compared methods for height and BMI are shown in (B) and (E). Error bars represent ± 1.96 of the standard error. (C) and (F) show the comparison of XPA with traditional risk factor models in height and BMI.

BMI, with a substantial improvement compared to the baseline BLUP model trained on the Chinese data only. XPA outperformed the runner up (BLUP trained on UKBB) by 7.3% and 19.5% improvements for height and BMI, respectively. In contrast, both XP-BLUP and GCTA-bvBLUP, the two methods that integrated Chinese and UKBB datasets, had lower predictive R^2 than XPA. XP-BLUP was only slightly better than the methods trained on the Chinese data, but inferior to both XPA and BLUP trained on the UKBB only. This is because XP-BLUP only includes information from the significant SNPs in UKBB while XPA can borrow information from UKBB across the whole genome. Due to the memory bottleneck, GCTA-bvBLUP failed to include all UKBB samples, and so could not further improve the prediction accuracy (see Figure S23 for details of comparing the memory usage and computational time). GCTA-bvBLUP took 75.8 h and required 1.07 Tb to integrate 150K UKBB samples with all Chinese samples and achieved its best performance for constructing PRSs with predictive $R^2 = 15.74\%$. In contrast, XPA used only 54.5 h (including 9 h for loading data, 3 h for estimating variance components, and 42.5 h for computing the posterior means and estimating fixed effects) and 385 Gb to analyze all Chinese and UKBB samples while achieving 20.2% and 46.4% improvement compared to GCTA-bvBLUP for height and BMI, respectively. When the population-specific large-effect SNPs were utilized by XPA_+, we found the predictive R^2 further increased to 19.72% in height and remained roughly the same with XPA in BMI.

In clinical applications, it is critical to stratify individuals into different genetic risk groups. In our discussion, we use 'risk' as a generic term to describe a trait. To measure the ability of risk stratification, we compared the observed phenotypic values of individuals from the top PRS quantiles with those from the reference group (i.e., 40%–50% quantile in our analysis). For both height and BMI, XPA was most effective in screening individuals with high genetic risk (Figures 3B and 3E). Compared to their reference groups, the individuals in the top 1% of PRS were 8.04 cm (SE = 0.54 cm) taller in height and 2.38 kg/m^2 larger in BMI. When XPA_+ was applied, the stratification ability was further improved, with 8.27 cm increased height and 2.67 kg/m^2 larger BMI for the individuals in the top 1% of PRSs. By incorporating covariates, such as age, sex, and first 10 principal components, XPA can construct prediction on the original phenotypic scale. Compared to the risk prediction model using these covariates only, XPA achieved a three-fold improvement of R^2 in height and two-fold improvement of R^2 in BMI for both males and females. In terms of the square root of mean squared error ($\sqrt{\text{MSE}}$), XPA achieved a 10% improvement in height and 4% improvement in BMI for both males and females (Figures 3C and 3F), respectively. We also found XPA can improve the stratification ability of PRSs for different ethnic groups in the Chinese population (see supplemental note and Figure S24). These results suggest that the prediction accuracy in a target population can be greatly improved by XPA which can effectively incorporates trans-ancestry genetic information.

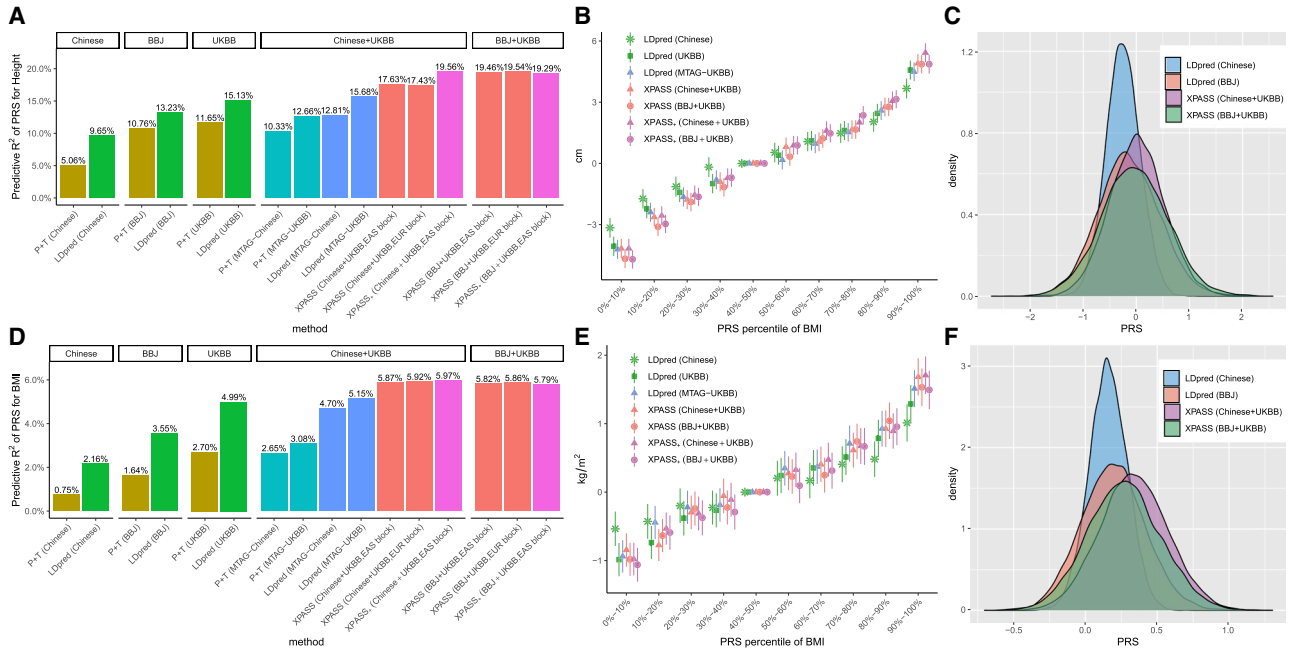

Figure 4. Prediction performance of XPASS and related summary-level methods for height and BMI in the Chinese population
Compared methods include XPASS, XPASS$_+$, LDpred, and P+T. For LDpred and P+T, one of the five sets of GWAS summary statistics were used as training set: Chinese only, BBJ only, UKBB only, improved Chinese and UKBB summary statistics obtained by combining the two datasets using MTAG (MTAG-Chinese and MTAG-UKBB). Predictive R^2 for height and BMI are shown in (A) and (D). Panels in (A) and (D) represent the datasets used for training. Stratification ability of XPASS and LDpred for height (B) and BMI (E). Error bars represent ±1.96 of the standard error. The distributions of PRSs constructed by XPASS and LDpred for height (C) and BMI (F).

Construction of PRSs for the Chinese population by XPASS using the summary-level data from trans-ancestry groups

When the individual-level GWAS data are not accessible, XPASS can take summary statistics as its input to construct PRSs. As estimated by XPASS, the genetic correlation between Chinese and UKBB was 0.78 for height ($\widehat{h}^2_{\text{Chinese}} = 35.0\%$ with SE = 2.4%, $\widehat{h}^2_{\text{UKBB}} = 43.6\%$ with SE = 2.3%) and 0.68 for BMI ($\widehat{h}^2_{\text{Chinese}} = 16.7\%$ with SE = 1.9%, $\widehat{h}^2_{\text{UKBB}} = 19.4\%$ with SE = 0.7%), comparable to those estimated by XPA. For the genetic correlations between BBJ and UKBB, the estimated genetic correlation computed by XPASS was 0.71 for height ($\widehat{h}^2_{\text{BBJ}} = 36.7\%$ with SE = 1.9%) and 0.68 for BMI ($\widehat{h}^2_{\text{BBJ}} = 13.0\%$ with SE = 0.5%). Given the substantial genetic correlations, XPASS could effectively leverage the UKBB summary data to improve prediction accuracy in the Chinese population. As summarized in Figures 4A and 4D, the PRSs derived by XPASS largely outperformed LDpred and P+T trained with the training set from a single population. When XPASS was trained on Chinese+UKBB, the predictive R^2 for height increased from 9.65% (LDpred trained on the Chinese data only) to 17.6%. By either applying XPASS$_+$ or using BBJ+UKBB as training set, the predictive R^2 further increased to 19.5%, achieving $(19.5\% - 15.1\%)/15.1\% \approx 29\%$ improvement compared to the best method using a single population. XPASS also outperformed MTAG with $(19.5\% - 15.7\%)/15.7\% \approx 24\%$ improvement in terms of R^2, when the MTAG output was used as training data for LDpred and P+T. A similar trend of improvement was observed for BMI (R^2 = 5.9% for XPASS compared to R^2 = 2.16% for LDpred trained on Chinese and R^2 = 5.15% for LDpred trained on MTAG-UKBB), although the amount of improvement for BMI was less than that of height. This can be attributed to the lower heritability of BMI. We found that the choice of LD block partitions in XPASS had little effect on its performance, suggesting that XPASS is quite robust to the partitioning strategy.

When the PRS was used to stratify individuals, we found XPASS is more effective in identifying the groups with extreme genetic values (Figures 4B and 4E), consistent with the analyses conducted on the individual-level data. By comparing the PRS distributions (Figures 4C and 4F), we observed that PRSs derived by XPASS from the larger training sets often had broader distribution than those derived by LDpred from the smaller training sets. We note that this observation is consistent with the stratification analysis, since the model with higher stratification ability pulls individuals with extreme PRSs farther away from its mean value.

Influence of the auxiliary sample size on prediction performance

As we can observe in the simulation analyses, a large auxiliary dataset is critical for the performance of cross-

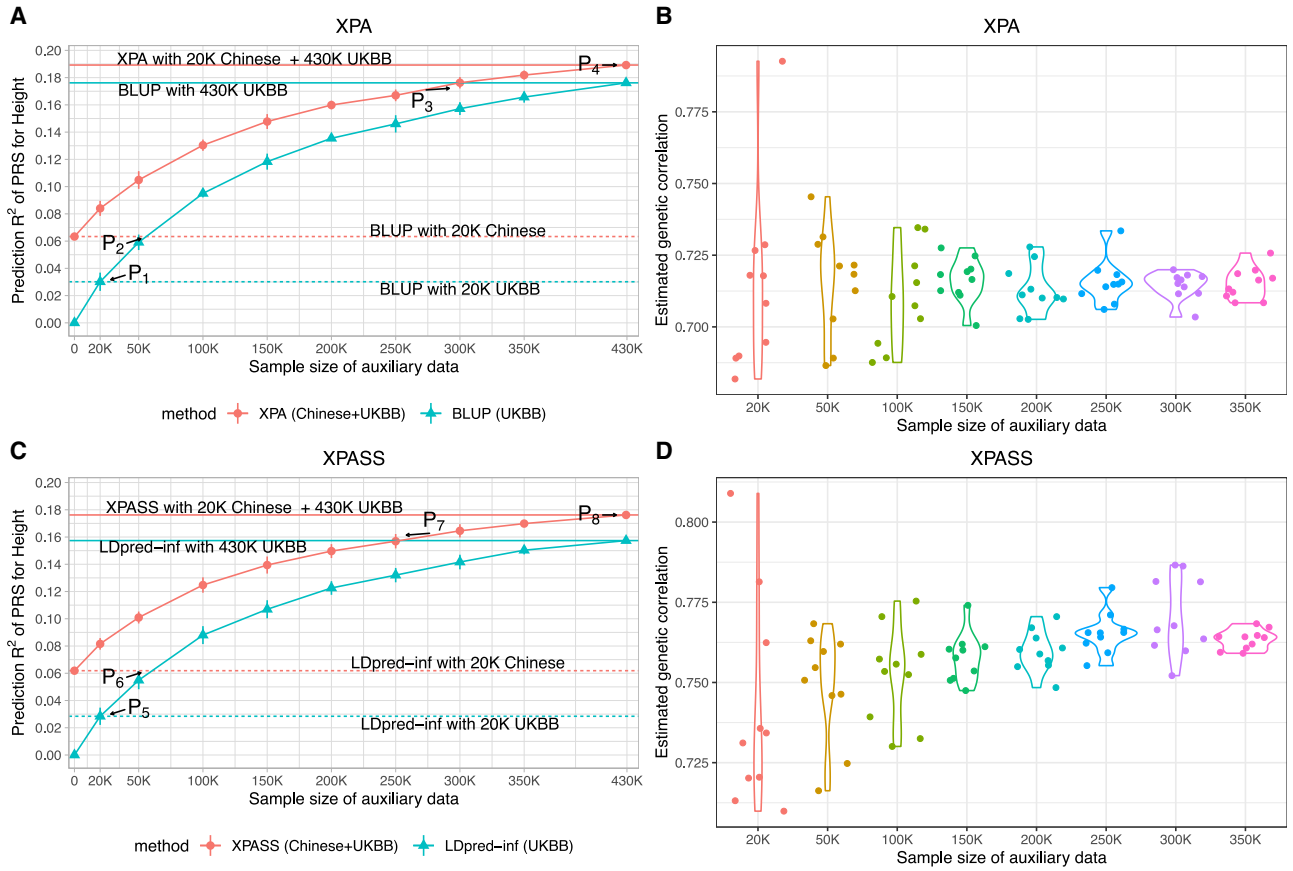

Figure 5. Influence of the auxiliary sample size on the prediction performance of XPA and XPASS for predicting height
Predictive R^2 of XPA and XPASS are shown in (A) and (C). The corresponding trans-ancestry genetic correlations estimated by XPA and XPASS in each replicate are shown in (B) and (D). We trained XPA and XPASS by integrating 21,069 Chinese training samples with 20,000–300,000 random subsamples drawn from UKBB, where samples from UKBB could be viewed as the auxiliary dataset. The results are summarized from ten replications. Dashed horizontal lines in (A) and (C) mark the BLUP/LDpred-inf results obtained by using 20,000 samples from Chinese (red) and UKBB (cyan). Solid horizontal lines in (A) and (C) mark the results obtained by using all UKBB samples with (red) or without (cyan) Chinese. Points P_1–P_4 in (A) represent the situations where the auxiliary sample size achieves 20,000 (P_1), BLUP trained on about 50,000 UKBB samples achieves equivalent performance with that trained on 20,000 Chinese samples (P_2), XPA achieves identical performance with BLUP trained on all UKBB samples (P_3), and XPA is trained with all UKBB samples (P_4). Points P_5–P_8 in (C) represent the similar situations for summary-level approaches XPASS and LDpred-inf. Error bars represent ±1.96 of the standard error.

population prediction. To systematically investigate how the sample size of auxiliary dataset influences the predictive performance of XPA, we randomly subsampled 20,000–300,000 UKBB individuals as an auxiliary dataset to construct a PRS for height in the Chinese population and evaluated the predictive R^2. We also trained BLUP using only auxiliary datasets as benchmark. As expected, due to the population difference between EAS and EUR, BLUP trained on 20,000 sample from UKBB was significantly inferior to that trained on 20,000 samples from Chinese (point P_1 in Figure 5A). When the UKBB sample size became larger than 50,000 (point P_2 in Figure 5A), it achieved better performance than the model trained on 20,000 samples from Chinese. In contrast, XPA provided stable estimate of genetic correlation (Figure 5B) regardless of the auxiliary sample size, so it always outperformed BLUP and effectively improved the prediction accuracy with the inclusion of more UKBB samples. It is also worth noting that XPA used only 20,000 Chinese and 300,000 Europeans to achieve the comparable performance with BLUP that was trained using all 430,000 European samples (point P_3 in Figure 5A), highlighting the importance of including samples from the target population in PRS construction. Comparing XPASS with its special case LDpred-inf[29] led to similar conclusions (Figures 5C and 5D). When 250,000 samples were included in the auxiliary dataset, XPASS achieved comparable performance with LDpred-inf using all 430,000 UKBB samples (point P_7 in Figure 5C). When XPASS$_+$ was applied, it further improved the performance and outperformed LDpred-inf when only 150,000 UKBB samples were included (Figure S34). By contrasting Figures 5A and 5C, we found individual-level approaches were generally better than summary-level approaches, which is consistent with our simulation results.

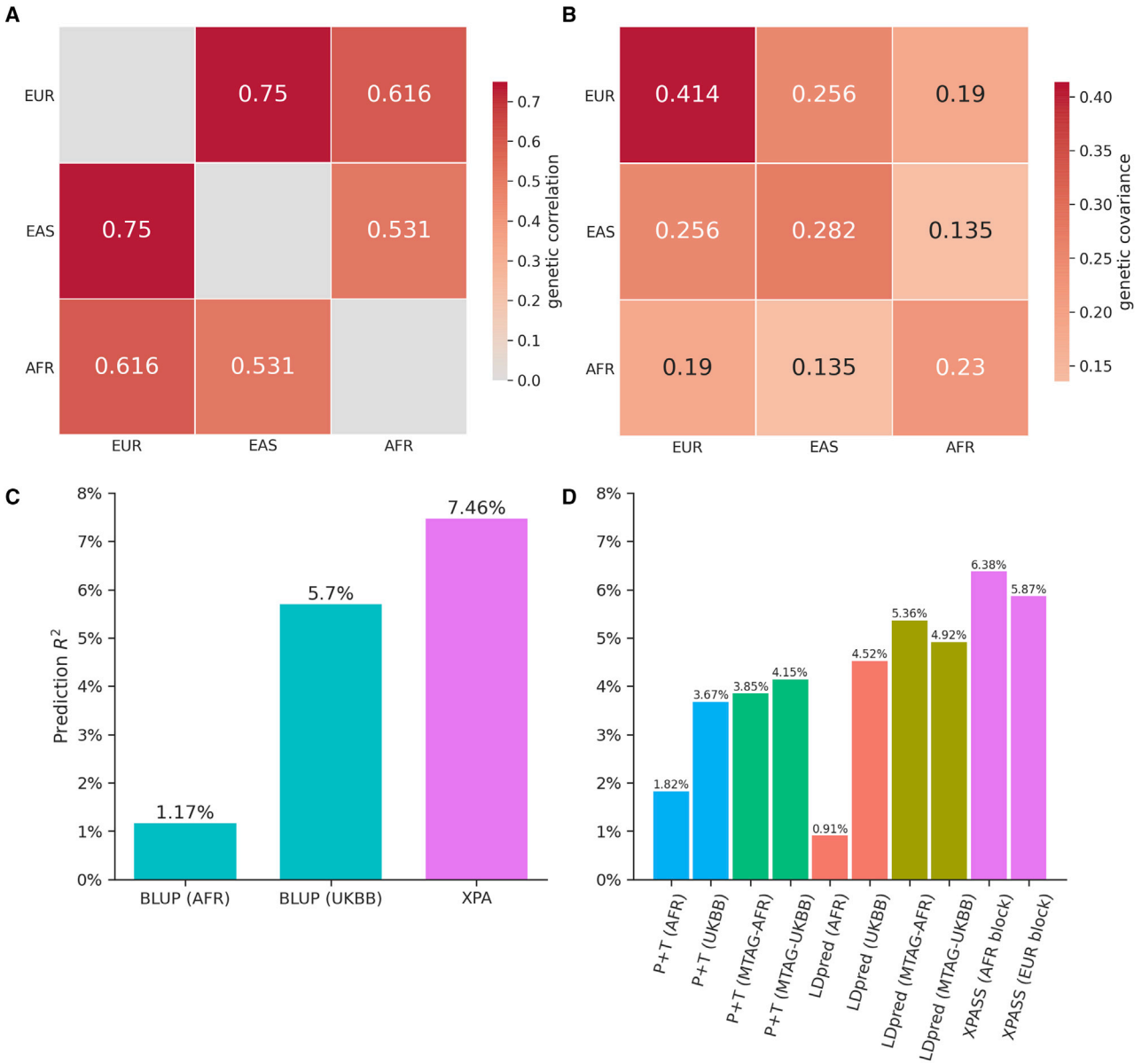

Figure 6. Application of XPA and XPASS for predicting height in the African population
Trans-ancestry genetic correlation (A) and genetic covariance (B) among European, African, and East Asian populations for height. (C) Prediction performance of XPA and BLUP for height measured by predictive R^2. (D) Prediction performance of XPASS, LDpred, and P+T for height measured by predictive R^2. For LDpred and P+T, one of the four sets of GWAS summary statistics were used as training set: African only, UKBB only, improved African and UKBB summary statistics obtained by combining the two datasets using MTAG (MTAG-AFR and MTAG-UKBB). For XPASS, we used the LD reference from either AFR or EUR population to construct independent LD blocks (AFR block and EUR block).

Construction of PRS for the African population by XPA and XPASS

To demonstrate the generality of our framework to other populations, we compared the prediction performance of XPA and XPASS with existing approaches using GWAS data from the African population. First, we estimated the trans-ancestry genetic correlation among European, African, and East Asian populations for height. Our results suggest that trans-ancestry genetic correlations of East Asian and European populations are often stronger than those of African and European populations (Figure 6A), consistent with a previous report.[52] To construct PRSs, we applied XPA and XPASS to integrate 7.4K African training samples with 430K European samples from UKBB. The prediction accuracy was evaluated on the 1K testing African samples. Clearly, both XPA (Figure 6C) and XPASS (Figure 6D) outperformed BLUP and LDpred and effectively improved the prediction accuracy with the inclusion of UKBB samples. The replication of better prediction performance by combining African and well-powered auxiliary European

populations reassures us that, despite the more significant genetic distance, our XPA and XPASS framework still achieved state-of-the-art performance when compared to alternative methods.

Discussion

The genetic architectures of various human traits have been mostly studied in samples with European ancestry, while non-European populations are still under-represented. Whether the genetic discoveries derived from Europeans can be transferred to non-Europeans remains unclear. In this study, we have proposed a unified framework for cross-population analysis (XPA and XPASS) to improve genetic prediction of under-represented populations by leveraging their trans-ancestry genetic correlations with a large and well-powered auxiliary GWAS dataset from another population. By combining the individual-level UKBB samples and Chinese samples, we were able to construct improved PRSs for height and BMI in the Chinese population, demonstrating the utility of trans-ancestry genetic prediction. We also showed that XPASS can achieve comparable prediction performance while only requiring summary-level data. When XPASS was trained using the summary-level BBJ and UKBB data, it produced even better prediction performance than XPA trained with the individual-level UKBB and Chinese data. As we do not have access to the individual-level BBJ data, XPASS offers an effective strategy to make use of existing data resources. We also observed improved prediction accuracy of PRSs in the African population when we applied XPA and XPASS to combine AFR and EUR GWAS data, suggesting the generality of our framework across global populations. When XPA_+ and $XPASS_+$ were applied, the population-specific effects can also be incorporated, leading to higher prediction accuracy in height. Because of the better performance and robustness to the choice of p value threshold, we recommend XPA_+ and $XPASS_+$ in practice.

Existing studies[53,54] have established the connection between the prediction accuracy of PRSs and various parameters from the theoretical perspective, including heritability, sample size, genetic correlation, and the effective number of chromosome segments. These theoretical properties have been supported by practical evidence. For example, Truong et al.[55] showed that the prediction accuracy of PRSs increased when individuals with higher-level relatedness were included in the analysis, which was due to the decreasing number of effective chromosome segments. Our work mainly focuses on the side of practice, aiming to develop a scalable and accurate method for the construction of PRSs in the cross-ancestry setting. We have observed that higher genetic correlation and inclusion of more target samples in the training set can improve prediction performance, which is consistent with the theory.[53]

In real applications, by taking the advantage of widespread pleiotropy among phenotypes, many successful multi-trait models[56–58] have been developed to produce powerful PRSs with increased prediction accuracy. While these approaches have been widely used in risk prediction within a single population, they may not be easily applied to integrate datasets from multi-ancestry backgrounds, since they do not take the heterogeneous genetic architectures into account. Our XPA framework, as compared to existing multi-trait models, provides a scalable solution to effectively combine multi-ancestry datasets by leveraging the stable estimate of trans-ancestry genetic correlation while accounting for the heterogeneous LD structure between populations. The success of XPA sheds light on the transferable genetic basis among global populations and demonstrates the benefits of integrating multi-ancestry datasets in genetic prediction.

While we have mainly focused on the anthropometric traits in this study, it is worth noting that XPA and XPASS can be applied to a wide class of phenotypes, such as complex diseases and molecular phenotypes. Due to the binary nature of diseases, their relationships with genotypes are often better captured by the liability threshold model (LTM). When the individual relatedness is low, the univariate linear mixed model (LMM) can be viewed as an approximation of the LTM.[27] A recent study[59] found that the bi-variate LMM can approximate the bi-variate LTM and produce consistent genetic correlation estimate for binary responses, suggesting the potential of applying XPA and XPASS to predict disease risks.

Our XPA framework needs more investigation in the following directions. First, XPA could be improved by allowing more flexible assumptions on SNP effect sizes. XPA assumes that, for a given population, the variance of the effect sizes of standardized genotypes is a constant, implicitly assuming that the SNP effect sizes increase as the allele frequencies decrease at the rate $1/\sqrt{2f(1-f)}$, where f is the allele frequency (AF). This assumption was also adopted in the previous trans-ancestry analysis.[60] Some recent studies have suggested that the effect sizes may not keep increasing when the allele frequency is small due to the negative selection.[25,61,62] It was shown that the prediction can benefit from introducing a selection parameter in BayesS (Figure 3A).

Second, the trans-ancestry genetic correlation may not be homogeneous across the genome. The differential selection pressure between populations can induce differences in their AF.[34,63] By sorting the SNPs according to the normalized AF difference between EAS and EUR, we estimated the genetic correlation of the effect sizes corresponding to the SNPs with the largest AF differences. As shown in the Figure S35, the trans-ancestry genetic correlation decreases as the AF difference increases. To assess the effect of AF difference on the prediction accuracy, we extended the XPASS model to include an additional genetic component that captures the effects of SNPs with large AF differences across populations (see supplemental

note). From our real data analysis, we did not observe significant enrichment of heritability among these SNPs. As a result, we did not obtain a better PRS by modeling the effect sizes of these SNPs as an additional variance component in the extended model (see Table S1). Our results suggest that modeling the effect sizes of SNPs with large AF difference may not be the key to improve PRSs.

Third, the integration of multiple GWASs across populations may further improve the prediction accuracy of XPA and XPASS. It has been shown that jointly modeling multiple genetically correlated phenotypes can produce more accurate prediction within EUR population.[56–58] By applying XPASS to estimate genetic correlations for a wide spectrum of phenotypes between EUR and EAS, we found that many genetically correlated traits discovered in EUR studies are replicated between EAS and EUR (supplemental note). Therefore, when a number of correlated phenotypes are simultaneously available in both populations, a more flexible model that jointly considers these phenotypes may further increase the prediction power.

Fourth, functional annotations may also be included to inform the prediction. The SNPs with biological functions, such as gene regulatory elements,[64–66] epigenomic regulations,[40,67,68] and tissue-specific functional pathways, are usually enriched for the heritability of complex traits.[69–73] Recent studies suggest that leveraging the functional annotations in prediction models trained on the EUR datasets produced PRSs with higher accuracy in both EUR[74,75] and EAS,[76] indicating the substantial overlap of functionally important variants across populations. Hence, integrating functional information to prioritize biologically relevant SNPs in complex traits can potentially increase the prediction accuracy of the XPA framework.

Appendix A

Derivation of normal Equation 5

Note $\tilde{\mathbf{y}} \sim \mathcal{N}(0, \tilde{\Omega})$ has its first-order moment as zero. Thus, the MoM estimator is then obtained by matching the second-order moment based on the criterion of least-squares:

$$\text{argmin}_{\sigma_1^2,\sigma_2^2,\delta,\sigma_\varepsilon^2,\sigma_\xi^2} \left|\left|\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T - \left(\tilde{\mathbf{X}}(\boldsymbol{\Sigma}_\beta \otimes \mathbf{I}_p)\tilde{\mathbf{X}}^T + \tilde{\boldsymbol{\Sigma}}_e\right)\right|\right|_F^2. \quad \text{(Equation A1)}$$

Knowing the fact that $\left|\left|\mathbf{B}\right|\right|_F = \sqrt{\text{tr}(\mathbf{B}\mathbf{B}^T)}$, the OLS objective function in Equation A1 can be re-written as

$$\begin{aligned}
&\left[\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T - \left(\tilde{\mathbf{X}}(\boldsymbol{\Sigma}_\beta \otimes \mathbf{I}_p)\tilde{\mathbf{X}}^T + \tilde{\boldsymbol{\Sigma}}_e\right)\right]^T \\
&= \text{tr}\Big[\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T + \tilde{\mathbf{X}}(\boldsymbol{\Sigma}_\beta \otimes \mathbf{I}_p)\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}(\boldsymbol{\Sigma}_\beta \otimes \mathbf{I}_p)\tilde{\mathbf{X}}^T + \tilde{\boldsymbol{\Sigma}}_e\tilde{\boldsymbol{\Sigma}}_e \\
&- 2\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T\tilde{\mathbf{X}}(\boldsymbol{\Sigma}_\beta \otimes \mathbf{I}_p)\tilde{\mathbf{X}}^T - 2\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T\tilde{\boldsymbol{\Sigma}}_e + 2\tilde{\mathbf{X}}(\boldsymbol{\Sigma}_\beta \otimes \mathbf{I}_p)\tilde{\mathbf{X}}^T\tilde{\boldsymbol{\Sigma}}_e\Big] \\
&= \text{tr}\left[\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T + \begin{bmatrix} \sigma_1^2\tilde{\mathbf{K}}_1 & \delta\tilde{\mathbf{K}}_{12} \\ \delta\tilde{\mathbf{K}}_{12}^T & \sigma_2^2\tilde{\mathbf{K}}_2 \end{bmatrix}\begin{bmatrix} \sigma_1^2\tilde{\mathbf{K}}_1 & \delta\tilde{\mathbf{K}}_{12} \\ \delta\tilde{\mathbf{K}}_{12}^T & \sigma_2^2\tilde{\mathbf{K}}_2 \end{bmatrix}^T\right. \\
&+ \begin{bmatrix} \sigma_\varepsilon^2\mathbf{M}_1 & 0 \\ 0 & \sigma_\xi^2\mathbf{M}_2 \end{bmatrix}\begin{bmatrix} \sigma_\varepsilon^2\mathbf{M}_1 & 0 \\ 0 & \sigma_\xi^2\mathbf{M}_2 \end{bmatrix}^T\Big] \\
&+ \text{tr}\left[-2\begin{bmatrix} \tilde{\mathbf{y}}_1\tilde{\mathbf{y}}_1^T & \tilde{\mathbf{y}}_1\tilde{\mathbf{y}}_2^T \\ \tilde{\mathbf{y}}_2\tilde{\mathbf{y}}_1^T & \tilde{\mathbf{y}}_2\tilde{\mathbf{y}}_2^T \end{bmatrix}\right. \\
&\times \begin{bmatrix} \sigma_1^2\tilde{\mathbf{K}}_1 & \delta\tilde{\mathbf{K}}_{12} \\ \delta\tilde{\mathbf{K}}_{12}^T & \sigma_2^2\tilde{\mathbf{K}}_2 \end{bmatrix} - 2\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T\begin{bmatrix} \sigma_\varepsilon^2\mathbf{M}_1 & 0 \\ 0 & \sigma_\xi^2\mathbf{M}_2 \end{bmatrix} \\
&+ 2\begin{bmatrix} \sigma_1^2\tilde{\mathbf{K}}_1 & \delta\tilde{\mathbf{K}}_{12} \\ \delta\tilde{\mathbf{K}}_{12}^T & \sigma_2^2\tilde{\mathbf{K}}_2 \end{bmatrix}\begin{bmatrix} \sigma_\varepsilon^2\mathbf{M}_1 & 0 \\ 0 & \sigma_\xi^2\mathbf{M}_2 \end{bmatrix} \\
&= \tilde{\mathbf{y}}^T\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T\tilde{\mathbf{y}} + (\sigma_1^2)^2\text{tr}\left[\tilde{\mathbf{K}}_1^2\right] + (\sigma_2^2)^2\text{tr}\left[\tilde{\mathbf{K}}_2^2\right] + 2\delta^2\text{tr}\left[\tilde{\mathbf{K}}_{12}\tilde{\mathbf{K}}_{12}^T\right] \\
&+ (\sigma_\varepsilon^2)^2\text{tr}[\mathbf{M}_1] + (\sigma_\xi^2)^2\text{tr}[\mathbf{M}_2] - 2\sigma_1^2\tilde{\mathbf{y}}_1^T\tilde{\mathbf{K}}_1\tilde{\mathbf{y}}_1 - 4\delta\tilde{\mathbf{y}}_1^T\tilde{\mathbf{K}}_{12}\tilde{\mathbf{y}}_2 \\
&- 2\sigma_2^2\tilde{\mathbf{y}}_2^T\tilde{\mathbf{K}}_2\tilde{\mathbf{y}}_2 - 2\sigma_\varepsilon^2\tilde{\mathbf{y}}_1^T\tilde{\mathbf{y}}_1 - 2\sigma_\xi^2\tilde{\mathbf{y}}^T\tilde{\mathbf{y}}_2 \\
&+ 2\sigma_1^2\sigma_\varepsilon^2\text{tr}\left[\tilde{\mathbf{K}}_1\right] + 2\sigma_2^2\sigma_\xi^2\text{tr}\left[\tilde{\mathbf{K}}_2\right].
\end{aligned}$$

Taking derivatives of the objective function with respect to $\boldsymbol{\theta} = \{\sigma_1^2, \sigma_2^2, \delta, \sigma_\varepsilon^2, \sigma_\xi^2\}$ and setting them to zero leads to estimating Equation 5.

Derivation of posterior mean (Equation 8)

Given the estimates of parameters $\{\widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \widehat{\delta}, \widehat{\sigma}_\varepsilon^2, \widehat{\sigma}_\xi^2\}$ and fixed effects $\widehat{\boldsymbol{\omega}}$, the posterior means of $\boldsymbol{\beta}$ are given as:

$$\begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}\Big|\mathbf{Z}_1, \mathbf{X}_1, \mathbf{y}_1, \mathbf{Z}_2, \mathbf{X}_2, \mathbf{y}_2 \sim \mathcal{N}(\widehat{\boldsymbol{\mu}}^{\mathbf{XPA}}, \boldsymbol{\Sigma}), \quad \text{(Equation A2)}$$

with

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \dfrac{\mathbf{X}_1^T\mathbf{X}_1}{\widehat{\sigma}_\varepsilon^2} & 0 \\ 0 & \dfrac{\mathbf{X}_2^T\mathbf{X}_2}{\widehat{\sigma}_\xi^2} \end{bmatrix} + \left(\widehat{\boldsymbol{\Sigma}}_\beta^{-1} \otimes \mathbf{I}_p\right),$$

$$\widehat{\boldsymbol{\mu}} \equiv \begin{bmatrix} \widehat{\boldsymbol{\mu}}_1^{\mathbf{XPA}} \\ \widehat{\boldsymbol{\mu}}_2^{\mathbf{XPA}} \end{bmatrix} = \boldsymbol{\Sigma}\begin{bmatrix} \dfrac{1}{\widehat{\sigma}_\varepsilon^2}\mathbf{X}_1^T\left(\mathbf{y}_1 - \mathbf{Z}_1\widehat{\boldsymbol{\omega}}_1\right) \\ \dfrac{1}{\widehat{\sigma}_\xi^2}\mathbf{X}_2^T\left(\mathbf{y}_2 - \mathbf{Z}_2\widehat{\boldsymbol{\omega}}_2\right) \end{bmatrix},$$

where $\widehat{\boldsymbol{\mu}}_1^{\mathbf{XPA}}$ and $\widehat{\boldsymbol{\mu}}_2^{\mathbf{XPA}}$ are the posterior means of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, respectively. Note that directly working on this form of posterior mean $\boldsymbol{\mu}^{\mathbf{XPA}}$ requires solving a $2p \times 2p$ linear system, which is intractable since p is in the order of millions. Here, we derive an equivalent form that computes $\boldsymbol{\mu}^{\mathbf{XPA}}$ by solving only an $(n_1+n_2)\times(n_1+n_2)$ linear system:

$$\boldsymbol{\mu}^{\mathbf{XPA}} = \boldsymbol{\Lambda}^{-1}\left(\begin{bmatrix}\widehat{\sigma}_{\varepsilon}^2 & 0\\ 0 & \widehat{\sigma}_{\xi}^2\end{bmatrix}\otimes\mathbf{I}_p\right)^{-1}\mathbf{X}^T(\mathbf{y}-\mathbf{Z}\widehat{\boldsymbol{\omega}})$$
$$= \left[\begin{bmatrix}\mathbf{X}_1^T\mathbf{X}_1 & 0\\ 0 & \mathbf{X}_2^T\mathbf{X}_2\end{bmatrix} + \widehat{\boldsymbol{\Delta}}\otimes\mathbf{I}_p\right]^{-1}\mathbf{X}^T(\mathbf{y}-\mathbf{Z}\widehat{\boldsymbol{\omega}}),$$
$$= \left[\mathbf{X}^T\mathbf{X} + \widehat{\boldsymbol{\Delta}}\otimes\mathbf{I}_p\right]^{-1}\mathbf{X}^T(\mathbf{y}-\mathbf{Z}\widehat{\boldsymbol{\omega}})$$
$$= \left(\widehat{\boldsymbol{\Delta}}^{-1}\otimes\mathbf{I}_p\right)\mathbf{X}^T\left(\mathbf{I}_{n_1+n_2} + \mathbf{X}\left(\widehat{\boldsymbol{\Delta}}^{-1}\otimes\mathbf{I}_p\right)\mathbf{X}^T\right)^{-1}(\mathbf{y}-\mathbf{Z}\widehat{\boldsymbol{\omega}}),$$

(Equation A3)

where $\widehat{\boldsymbol{\Delta}} = \begin{bmatrix}\widehat{\sigma}_{\varepsilon}^2 & 0\\ 0 & \widehat{\sigma}_{\xi}^2\end{bmatrix}\begin{bmatrix}\widehat{\sigma}_1^2 & \widehat{\delta}\\ \widehat{\delta} & \widehat{\sigma}_2^2\end{bmatrix}^{-1} = \begin{bmatrix}\frac{\widehat{\sigma}_{\varepsilon}^2\widehat{\sigma}_2^2}{\widehat{\sigma}_1^2\widehat{\sigma}_2^2-\widehat{\delta}^2} & -\frac{\widehat{\sigma}_{\varepsilon}^2\widehat{\delta}}{\widehat{\sigma}_1^2\widehat{\sigma}_2^2-\widehat{\delta}^2}\\ -\frac{\widehat{\sigma}_{\xi}^2\widehat{\delta}}{\widehat{\sigma}_1^2\widehat{\sigma}_2^2-\widehat{\delta}^2} & \frac{\widehat{\sigma}_{\xi}^2\widehat{\sigma}_1^2}{\widehat{\sigma}_1^2\widehat{\sigma}_2^2-\widehat{\delta}^2}\end{bmatrix}$ and the last equation is granted by the Woodbury matrix identity. XPA computes the posterior mean of the target population by:

$$\widehat{\boldsymbol{\mu}}_1^{\mathbf{XPA}} = \begin{bmatrix}\widehat{\sigma}_1^2\mathbf{X}_1\\ \widehat{\delta}\mathbf{X}_2\end{bmatrix}^T\begin{bmatrix}\widehat{\sigma}_1^2\mathbf{X}_1\mathbf{X}_1^T + \widehat{\sigma}_{\varepsilon}^2\mathbf{I}_{n1} & \widehat{\delta}\mathbf{X}_1\mathbf{X}_2^T\\ \widehat{\delta}\mathbf{X}_2\mathbf{X}_1^T & \widehat{\sigma}_2^2\mathbf{X}_2\mathbf{X}_2^T + \widehat{\sigma}_{\xi}^2\mathbf{I}_{n2}\end{bmatrix}^{-1}$$
$$\times\begin{bmatrix}\mathbf{y}_1-\mathbf{Z}_1\widehat{\boldsymbol{\omega}}_1\\ \mathbf{y}_2-\mathbf{Z}_2\widehat{\boldsymbol{\omega}}_2\end{bmatrix} \equiv \widehat{\Omega}^{-1}(\mathbf{y}-\mathbf{Z}\widehat{\boldsymbol{\omega}}),$$

Computation of posterior means and fixed effects with the CG algorithm

The fixed effects and posterior means are given as

$$\widehat{\boldsymbol{\omega}} \equiv \begin{bmatrix}\widehat{\boldsymbol{\omega}}_1\\ \widehat{\boldsymbol{\omega}}_2\end{bmatrix} = \left(\mathbf{Z}^T\widehat{\Omega}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^T\widehat{\Omega}^{-1}\mathbf{y},$$

and

$$\widehat{\boldsymbol{\mu}}_1^{\mathbf{XPA}} = \begin{bmatrix}\widehat{\sigma}_1^2\mathbf{X}_1\\ \widehat{\delta}\mathbf{X}_2\end{bmatrix}^T\underbrace{\begin{bmatrix}\widehat{\sigma}_1^2\mathbf{X}_1\mathbf{X}_1^T+\widehat{\sigma}_{\varepsilon}^2\mathbf{I}_{n1} & \widehat{\delta}\mathbf{X}_1\mathbf{X}_2^T\\ \widehat{\delta}\mathbf{X}_2\mathbf{X}_1^T & \widehat{\sigma}_2^2\mathbf{X}_2\mathbf{X}_2^T+\widehat{\sigma}_{\xi}^2\mathbf{I}_{n2}\end{bmatrix}^{-1}}_{\widehat{\Omega}^{-1}\in\mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}}\begin{bmatrix}\mathbf{y}_1-\mathbf{Z}_1\widehat{\boldsymbol{\omega}}_1\\ \mathbf{y}_2-\mathbf{Z}_2\widehat{\boldsymbol{\omega}}_2\end{bmatrix}$$
$$= \begin{bmatrix}\widehat{\sigma}_1^2\mathbf{X}_1\\ \widehat{\delta}\mathbf{X}_2\end{bmatrix}^T(\widehat{\Omega}^{-1}\mathbf{y}-\widehat{\Omega}^{-1}\mathbf{Z}\widehat{\boldsymbol{\omega}}),$$

respectively. Both quantities involve solving the linear systems $\mathbf{U} = \widehat{\Omega}^{-1}\mathbf{Z}$ and $\mathbf{v} = \widehat{\Omega}^{-1}\mathbf{y}$. Therefore, we can compute the estimates of fixed effects and the posterior means at the same time while only solving the linear systems once. Specifically, we first construct a working matrix $\mathbf{W} = [\mathbf{y}, \mathbf{Z}]$ and apply the conjugate gradient approach to solve the combined linear systems $[\mathbf{v}, \mathbf{U}] = \widehat{\Omega}^{-1}\mathbf{W}$, as summarized in Algorithm 1.

At each iteration of the CG algorithm, we need to compute

$$\widehat{\Omega}\mathbf{P}_j = \widehat{\Omega}\begin{bmatrix}\mathbf{P}_{1j}\\ \mathbf{P}_{2j}\end{bmatrix} = \begin{bmatrix}\widehat{\sigma}_1^2\mathbf{X}_1\mathbf{X}_1^T\mathbf{P}_{1j} + \widehat{\sigma}_{\varepsilon}^2\mathbf{P}_{1j} + \widehat{\delta}\mathbf{X}_1\mathbf{X}_2^T\mathbf{P}_{2j}\\ \widehat{\sigma}_2^2\mathbf{X}_2\mathbf{X}_2^T\mathbf{P}_{2j} + \widehat{\sigma}_{\xi}^2\mathbf{P}_{2j} + \widehat{\delta}\mathbf{X}_2\mathbf{X}_1^T\mathbf{P}_{1j}\end{bmatrix},$$

where $\mathbf{P}_{1j}\in\mathbb{R}^p$ and $\mathbf{P}_{2j}\in\mathbb{R}^p$ are column vectors. With entries of $\mathbf{X}_1$ and $\mathbf{X}_2$ decoded from the Hash table, we first compute $\mathbf{X}_1^T\mathbf{P}_{1j}$ and $\mathbf{X}_2^T\mathbf{P}_{2j}$ and then multiply $\mathbf{X}_1$ and $\mathbf{X}_2$ on their left to obtain the corresponding terms. This operation is highly efficient since it only involves matrix-vector multiplication. The final time complexity of the CG algorithm is $\mathcal{O}(p(n_1+n_2)(c_1+c_2+1)\sqrt{\kappa})$, where κ is the condition number of $\widehat{\Omega}$. Because κ is usually small, the CG algorithm offers substantial computational improvement in solving the linear system.

Assumptions for XPASS model

We consider the datasets $\{\mathbf{z}_1, \mathbf{G}_1', \mathbf{Z}_1'\}$ and $\{\mathbf{z}_2, \mathbf{G}_2', \mathbf{Z}_1'\}$ from the two populations. The vectors $\mathbf{z}_1 = [z_{11}, \ldots, z_{1j}, \ldots, z_{1p}]^T\in\mathbb{R}^p$ and $\mathbf{z}_2 = [z_{21}, \ldots, z_{2j}, \ldots, z_{2p}]^T\in\mathbb{R}^p$ contain the z-scores derived from the two populations, where $z_{1j} = \frac{(\mathbf{x}_{1j}^T\mathbf{x}_{1j})^{-1}\mathbf{x}_{1j}^T\mathbf{y}_1}{\sqrt{\widehat{\sigma}_{1j}^2(\mathbf{x}_{1j}^T\mathbf{x}_{1j})^{-1}}}$ and $z_{2j} = \frac{(\mathbf{x}_{2j}^T\mathbf{x}_{2j})^{-1}\mathbf{x}_{2j}^T\mathbf{y}_2}{\sqrt{\widehat{\sigma}_{2j}^2(\mathbf{x}_{2j}^T\mathbf{x}_{2j})^{-1}}}$ and $\widehat{\sigma}_{1j}^2$ and $\widehat{\sigma}_{2j}^2$ are the residual variance of regressing $\mathbf{y}_1$ on $\mathbf{x}_{1j}$ and $\mathbf{y}_2$ on $\mathbf{x}_{2j}$, respectively. Following Vilhjálmsson et al.,[29] we assume the z-scores are derived from GWAS datasets with phenotype vectors $\mathbf{y}_1$ and $\mathbf{y}_2$ standardized to have mean of zero and standard deviation of one. The first few PCs of the reference genotypes from the two populations are given in $\mathbf{Z}_1'\in\mathbb{R}^{m_1\times c_{1'}}$ and $\mathbf{Z}_2'\in\mathbb{R}^{m_2\times c_{2'}}$. Similar to XPA, we first standardize the reference genotype matrices $\mathbf{G}_1'$ and $\mathbf{G}_2'$ to obtain the corresponding $\mathbf{X}_1'$ and $\mathbf{X}_2'$ that have column means zero and variances $1/p$.

Parameter estimation in XPASS

To derive the normal equations using the summary-level datasets, we start by eliminating the σ_{ε} and σ_{ξ} in Equation 5, which leads to

$$\begin{bmatrix}\mathrm{tr}\left(\tilde{\mathbf{K}}_1^2\right) - \frac{\mathrm{tr}^2\left(\tilde{\mathbf{K}}_1\right)}{n_1} & 0 & 0\\ 0 & \mathrm{tr}\left(\tilde{\mathbf{K}}_2^2\right) - \frac{\mathrm{tr}^2\left(\tilde{\mathbf{K}}_2\right)}{n_2} & 0\\ 0 & 0 & \mathrm{tr}\left(\tilde{\mathbf{K}}_{12}\tilde{\mathbf{K}}_{12}^T\right)\end{bmatrix}$$
$$\times\begin{bmatrix}\sigma_1^2\\ \sigma_2^2\\ \delta\end{bmatrix} = \begin{bmatrix}\tilde{\mathbf{y}}_1^T\tilde{\mathbf{K}}_1\tilde{\mathbf{y}}_1 - \frac{\tilde{\mathbf{y}}_1^T\tilde{\mathbf{y}}_1\mathrm{tr}\left(\tilde{\mathbf{K}}_1\right)}{n_1}\\ \tilde{\mathbf{y}}_2^T\tilde{\mathbf{K}}_2\tilde{\mathbf{y}}_2 - \frac{\tilde{\mathbf{y}}_2^T\tilde{\mathbf{y}}_2\mathrm{tr}\left(\tilde{\mathbf{K}}_2\right)}{n_2}\\ \tilde{\mathbf{y}}_1^T\tilde{\mathbf{K}}_{12}\tilde{\mathbf{y}}_2\end{bmatrix}.$$

Algorithm 1. Conjugate Gradient algorithm for solving $[\mathbf{v}, \mathbf{U}] = \hat{\boldsymbol{\Omega}}^{-1}\mathbf{W}$

Data: $\mathbf{X}_1, \mathbf{X}_2, \mathbf{W} = [\mathbf{y}, \mathbf{Z}], \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\delta}, \hat{\sigma}_\epsilon^2, \hat{\sigma}_\xi^2$
Result: $\mathbf{v} = \hat{\Omega}^{-1}\mathbf{y}, \mathbf{U} = \hat{\Omega}^{-1}\mathbf{Z}$
initialize $\{\mathbf{v}, \mathbf{U}\}$;
$\mathbf{R} \leftarrow \mathbf{W} - \hat{\Omega}[\mathbf{v}, \mathbf{U}]$;
$\mathbf{P} \leftarrow \mathbf{R}$;
$rs_{old} \leftarrow diag(\mathbf{R}^T\mathbf{R})$;
while $\sqrt{\max(rs_{old})} < 5 \times 10^{-4}$ do
  for $j = 1, ..., c_1 + c_2 + 1$ *as column index* do
    $(\hat{\Omega}\mathbf{P}_j) \leftarrow \hat{\Omega}\mathbf{P}_j$;
    $\alpha \leftarrow \frac{rs_{old,j}}{\mathbf{P}_j^T(\hat{\Omega}\mathbf{P}_j)}$;
    $[\mathbf{v}, \mathbf{U}]_j \leftarrow [\mathbf{v}, \mathbf{U}]_j + \alpha\mathbf{P}_j$;
    $\mathbf{R} \leftarrow \mathbf{R} - \alpha(\hat{\Omega}\mathbf{P}_j)$;
    $rs_{new,j} \leftarrow \mathbf{R}_j^T\mathbf{R}_j$;
    $\gamma \leftarrow \frac{rs_{new,j}}{rs_{old,j}}$;
    $\mathbf{P}_j \leftarrow \mathbf{R}_j + \gamma\mathbf{P}_j$;
  end
  $rs_{old} \leftarrow rs_{new}$;
end

By dividing the three equations by n_1^2, n_2^2 and n_1n_2, the estimating Equation 5 became:

$$\begin{bmatrix} \frac{\mathrm{tr}\left(\tilde{\mathbf{K}}_1^2\right) - \frac{\mathrm{tr}^2\left(\tilde{\mathbf{K}}_1\right)}{n_1}}{n_1^2} & 0 & 0 \\ 0 & \frac{\mathrm{tr}\left(\tilde{\mathbf{K}}_2^2\right) - \frac{\mathrm{tr}^2\left(\tilde{\mathbf{K}}_2\right)}{n_2}}{n_2^2} & 0 \\ 0 & 0 & \frac{\mathrm{tr}\left(\tilde{\mathbf{K}}_{12}\tilde{\mathbf{K}}_{12}^T\right)}{n_1n_2} \end{bmatrix} \times \begin{bmatrix} \sigma_1^2 \\ \sigma_2^2 \\ \delta \end{bmatrix} = \begin{bmatrix} \frac{\tilde{\mathbf{y}}_1^T\tilde{\mathbf{K}}_1\tilde{\mathbf{y}}_1}{n_1^2} - \frac{\tilde{\mathbf{y}}_1^T\tilde{\mathbf{y}}_1\mathrm{tr}\left(\tilde{\mathbf{K}}_1\right)}{n_1^3} \\ \frac{\tilde{\mathbf{y}}_2^T\tilde{\mathbf{K}}_2\tilde{\mathbf{y}}_2}{n_2^2} - \frac{\tilde{\mathbf{y}}_2^T\tilde{\mathbf{y}}_2\mathrm{tr}\left(\tilde{\mathbf{K}}_2\right)}{n_2^3} \\ \frac{\tilde{\mathbf{y}}_1^T\tilde{\mathbf{K}}_{12}\tilde{\mathbf{y}}_2}{n_1n_2} \end{bmatrix}. \quad \text{(Equation A4)}$$

Note that the summary statistic of the j-th SNP $z_{1j} = \frac{(x_{1j}^Tx_{1j})^{-1}x_{1j}^Ty}{\sqrt{\widehat{\sigma_j^2}(x_{1j}^Tx_{1j})^{-1}}} = \frac{x_{1j}^Ty}{\sqrt{\widehat{\sigma_j^2}\frac{n}{p}}} \approx \frac{x_{1j}^Ty}{\sqrt{\sigma_{y_1}^2\frac{n}{p}}}$, where the first equation is derived from the assumption that x_j has mean 0 and variance $1/p$, and the approximation is granted by the assumption that each SNP only contributes a small proportion of the total phenotypic variance. Similarly, $z_{2j} \approx \frac{x_{2j}^Ty}{\sqrt{\sigma_{y_2}^2\frac{n}{p}}}$. Therefore, we can approximate the terms in the right-hand side of Equation A4 by

$$\frac{\tilde{\mathbf{y}}_1^T\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_1^T\tilde{\mathbf{y}}_1}{n_1^2} = \sum_{j=1}^{p}\left(\frac{\tilde{\mathbf{x}}_{1j}^T\tilde{\mathbf{y}}_1}{n_1}\right)^2 \approx \sum_{j=1}^{p}\left(z_{1j}\sqrt{\frac{\sigma_{y_1}^2}{n_1p}}\right)^2 = \frac{\sigma_{y_1}^2}{n_1p}\sum_{j=1}^{p}z_{1j}^2,$$

$$\frac{\tilde{\mathbf{y}}_1^T\tilde{\mathbf{y}}_1\mathrm{tr}\left(\tilde{\mathbf{K}}_1\right)}{n_1^3} \approx \frac{n_1^2\sigma_{y_1}^2}{n_1^3} = \frac{\sigma_{y_1}^2}{n_1},$$

$$\frac{\tilde{\mathbf{y}}_2^T\tilde{\mathbf{X}}_2\tilde{\mathbf{X}}_2^T\tilde{\mathbf{y}}_2}{n_2^2} = \sum_{j=1}^{p}\left(\frac{\tilde{\mathbf{x}}_{2j}^T\tilde{\mathbf{y}}_2}{n_2}\right)^2 \approx \sum_{j=1}^{p}\left(z_{2j}\sqrt{\frac{\sigma_{y_2}^2}{n_2p}}\right)^2 = \frac{\sigma_{y_2}^2}{n_2p}\sum_{j=1}^{p}z_{2j}^2,$$

$$\frac{\tilde{\mathbf{y}}_2^T\tilde{\mathbf{y}}_2\mathrm{tr}\left(\tilde{\mathbf{K}}_2\right)}{n_2^3} \approx \frac{n_2^2\sigma_{y_2}^2}{n_2^3} = \frac{\sigma_{y_2}^2}{n_2},$$

$$\frac{\tilde{\mathbf{y}}_1^T\tilde{\mathbf{X}}_1\tilde{\mathbf{X}}_2^T\tilde{\mathbf{y}}_2}{n_1n_2} = \sum_{j=1}^{p}\left(\frac{\tilde{\mathbf{y}}_1^T\tilde{\mathbf{x}}_{1j}\tilde{\mathbf{x}}_{2j}^T\tilde{\mathbf{y}}_2}{\sqrt{n_1n_2}}\right)^2 \approx \frac{\sigma_{y_1}\sigma_{y_2}}{\sqrt{n_1n_2}p}\sum_{j=1}^{p}z_{1j}z_{2j}.$$

In the left-hand side of Equation A4, $\tilde{\mathbf{y}}_1$ and $\tilde{\mathbf{y}}_2$ are not involved, and the terms involving $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2$ can be approximated using the reference genotypes $\mathbf{X}_1'$ and $\mathbf{X}_2'$. For example, $\frac{\mathrm{tr}(\tilde{\mathbf{K}}_1^2) - \frac{\mathrm{tr}^2(\tilde{\mathbf{K}}_1)}{n_1}}{n_1^2}$ can be approximated by $\frac{\mathrm{tr}(\tilde{\mathbf{K}}_1'^2) - \frac{\mathrm{tr}^2(\tilde{\mathbf{K}}_{\prime 1})}{m_1}}{m_1^2}$, where $\tilde{\mathbf{K}}_1' = \mathbf{M}\prime_1\mathbf{X}\prime_1(\mathbf{M}\prime_1\mathbf{X}\prime_1)^T$ and $\mathbf{M}\prime_1 = \mathbf{I}_{m_1} - \mathbf{Z}\prime_1(\mathbf{Z}_1'^T\mathbf{Z}\prime_1)^{-1}\mathbf{Z}_1'^T$. Other terms in the left-hand side can be approximated in the same way. Using these approximations, the final normal equation is given as

$$\begin{bmatrix} \frac{\mathrm{tr}\left(\tilde{\mathbf{K}}_1'^2\right) - \frac{\mathrm{tr}^2\left(\tilde{\mathbf{K}}\prime_1\right)}{m_1}}{m_1^2} & 0 & 0 \\ 0 & \frac{\mathrm{tr}\left(\tilde{\mathbf{K}}_2'^2\right) - \frac{\mathrm{tr}^2\left(\tilde{\mathbf{K}}\prime_2\right)}{m_2}}{m_2^2} & 0 \\ 0 & 0 & \frac{\mathrm{tr}\left(\tilde{\mathbf{K}}\prime_{12}\mathbf{K}_{12}'^T\right)}{m_1m_2} \end{bmatrix} \times \begin{bmatrix} h_1^2 \\ h_2^2 \\ h_{12} \end{bmatrix} = \begin{bmatrix} \frac{1}{p}\sum_{j=1}^{p}\frac{z_{1j}^2 - 1}{n_1} \\ \frac{1}{p}\sum_{j=1}^{p}\frac{z_{2j}^2 - 1}{n_2} \\ \frac{1}{p}\sum_{j=1}^{p}\frac{z_{1j}z_{2j}}{\sqrt{n_1n_2}} \end{bmatrix}, \quad \text{(Equation A5)}$$

where $\tilde{\mathbf{K}}_2' = \mathbf{M}\prime_2\mathbf{X}\prime_2(\mathbf{M}\prime_2\mathbf{X}\prime_2)^T$, $\tilde{\mathbf{K}}_{12}' = \mathbf{M}\prime_1\mathbf{X}\prime_1(\mathbf{M}\prime_2\mathbf{X}\prime_2)^T$, $\mathbf{M}\prime_2 = \mathbf{I}_{m_2} - \mathbf{Z}\prime_2(\mathbf{Z}_2'^T\mathbf{Z}\prime_2)^{-1}\mathbf{Z}_2'^T$, $h_1^2 := \frac{\sigma_1^2}{\sigma_1^2+\sigma_\xi^2}$ and $h_2^2 := \frac{\sigma_2^2}{\sigma_2^2+\sigma_\xi^2}$ are heritabilities for the two populations, respectively,

and $h_{12} := \rho h_1 h_2$ is the co-heritability between two populations. Note that all terms in Equation A5 depend only on the z-scores and reference panels. Solving Equation A5 leads to the parameter estimates $\{\widehat{h}_1^2, \widehat{h}_2^2, \widehat{h}_{12}\}$.

Given the parameter estimates $\{\widehat{h}_1, \widehat{h}_2, \widehat{h}_{12}\}$, the genetic correlation can be estimated by $\widehat{\rho} = \widehat{h}_{12}/\widehat{h}_1\widehat{h}_2$. To test for $H_0 : \rho = 0$ using summary statistics, we estimate the standard error of $\widehat{\rho}$ by applying the jackknife resampling method, with a leaving-one-chromosome-out strategy.

Derivation of Equations 9 and 10

With the model in Equation 3, the estimates of fixed effects and the posterior means are given as

$$\begin{bmatrix}\widehat{\boldsymbol{\gamma}}_1\\ \widehat{\boldsymbol{\gamma}}_2\end{bmatrix} = \left(\begin{bmatrix}n_1\mathbf{R}_{l1} & 0\\ 0 & n_2\mathbf{R}_{l2}\end{bmatrix} - \begin{bmatrix}n_1\mathbf{R}_{ls1} & 0\\ 0 & n_2\mathbf{R}_{ls2}\end{bmatrix}\left(\begin{bmatrix}\frac{\widehat{h}_1^2}{1-\widehat{h}_1^2} & \frac{\widetilde{h}_{12}^2}{1-\widetilde{h}_2^2}\\ \frac{\widehat{h}_{12}^2}{1-\widehat{h}_1^2} & \frac{\widetilde{h}_2^2}{1-\widetilde{h}_2^2}\end{bmatrix}^{-1} \otimes \mathbf{I}_p + \begin{bmatrix}n_1\mathbf{R}_1 & 0\\ 0 & n_2\mathbf{R}_2\end{bmatrix}\right)^{-1}\begin{bmatrix}n_1\mathbf{R}_{sl1} & 0\\ 0 & n_2\mathbf{R}_{sl2}\end{bmatrix}\right)^{-1}$$

$$\left(\begin{bmatrix}\mathbf{X}_{l1}^T\mathbf{y}_{l1}\\ \mathbf{X}_{l2}^T\mathbf{y}_{l2}\end{bmatrix} - \begin{bmatrix}n_1\mathbf{R}_{ls1} & 0\\ 0 & n_2\mathbf{R}_{ls2}\end{bmatrix}\left(\begin{bmatrix}\frac{\widehat{h}_1^2}{1-\widehat{h}_1^2} & \frac{\widetilde{h}_{12}^2}{1-\widetilde{h}_2^2}\\ \frac{\widehat{h}_{12}^2}{1-\widehat{h}_1^2} & \frac{\widetilde{h}_2^2}{1-\widetilde{h}_2^2}\end{bmatrix}^{-1} \otimes \mathbf{I}_p + \begin{bmatrix}n_1\mathbf{R}_1 & 0\\ 0 & n_2\mathbf{R}_2\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{X}_1^T\mathbf{y}_1\\ \mathbf{X}_2^T\mathbf{y}_2\end{bmatrix}\right),$$

$$\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix} = \left(\begin{bmatrix}\frac{\widehat{h}_1^2}{1-\widehat{h}_1^2} & \frac{\widetilde{h}_{12}^2}{1-\widetilde{h}_2^2}\\ \frac{\widehat{h}_{12}^2}{1-\widehat{h}_1^2} & \frac{\widetilde{h}_2^2}{1-\widetilde{h}_2^2}\end{bmatrix}^{-1} \otimes \mathbf{I}_p + \begin{bmatrix}n_1\mathbf{R}_1 & 0\\ 0 & n_2\mathbf{R}_2\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{X}_1 & 0\\ 0 & \mathbf{X}_2\end{bmatrix}^T\begin{bmatrix}\mathbf{y}_1 - \mathbf{Z}_1\widehat{\boldsymbol{\omega}}_1 - \mathbf{X}_{l1}\widehat{\boldsymbol{\gamma}}_1\\ \mathbf{y}_2 - \mathbf{Z}_2\widehat{\boldsymbol{\omega}}_2 - \mathbf{X}_{l2}\widehat{\boldsymbol{\gamma}}_2\end{bmatrix},$$

(Equation A6)

By taking the advantage of the Woodbury matrix identity to compute the matrix inversion of $\widehat{\Omega}^{-1}$, we obtain

where $\mathbf{R}_1 = \mathbf{X}_1^T\mathbf{X}_1/n_1$ and $\mathbf{R}_2 = \mathbf{X}_2^T\mathbf{X}_2/n_2$ are the LD matrices of all SNPs, $\mathbf{R}_{l1} = \mathbf{X}_{l1}^T\mathbf{X}_{l1}/n_1$ and $\mathbf{R}_{l2} = \mathbf{X}_{l2}^T\mathbf{X}_{l2}/n_2$ are the LD matrices of large-effect SNPs, and $\mathbf{R}_{ls1} = \mathbf{R}_{sl1}^T = \mathbf{X}_{l1}^T\mathbf{X}_1/n_1$ and $\mathbf{R}_{ls2} = \mathbf{R}_{sl2}^T = \mathbf{X}_{l2}^T\mathbf{X}_2/n_2$ are the SNP correlation matrices between large-effect SNPs and all SNPs from the two populations, respectively.

With the reference genotype matrices, we can approximate the LD matrices in Equation A6 with $\mathbf{R}_1 \approx \widehat{\mathbf{R}}_1 = \mathbf{X}_1'^T\mathbf{X}_1'/m_1$, $\mathbf{R}_2 \approx \widehat{\mathbf{R}}_2 = \mathbf{X}_2'^T\mathbf{X}_2'/m_2$, $\mathbf{R}_{ls1} \approx \widehat{\mathbf{R}}_{ls1} = \mathbf{X}_{l1}'^T\mathbf{X}_1'/m_1$, and $\mathbf{R}_{ls2} \approx \widehat{\mathbf{R}}_{ls2} = \mathbf{X}_{l2}'^T\mathbf{X}_2'/m_2$. In addition, the

$$\begin{bmatrix}\widehat{\boldsymbol{\gamma}}_1\\ \widehat{\boldsymbol{\gamma}}_2\end{bmatrix} = \left(\begin{bmatrix}\mathbf{X}_{l1} & 0\\ 0 & \mathbf{X}_{l2}\end{bmatrix}^T\widehat{\Omega}^{-1}\begin{bmatrix}\mathbf{X}_{l1} & 0\\ 0 & \mathbf{X}_{l2}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{X}_{l1} & 0\\ 0 & \mathbf{X}_{l2}\end{bmatrix}^T\widehat{\Omega}^{-1}\mathbf{y},$$

$$\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix} = \left(\begin{bmatrix}\frac{\widehat{h}_1^2}{1-\widehat{h}_1^2} & \frac{\widetilde{h}_{12}^2}{1-\widetilde{h}_2^2}\\ \frac{\widehat{h}_{12}^2}{1-\widehat{h}_1^2} & \frac{\widetilde{h}_2^2}{1-\widetilde{h}_2^2}\end{bmatrix}^{-1} \otimes \mathbf{I}_p + \begin{bmatrix}\mathbf{X}_1^T\mathbf{X}_1 & 0\\ 0 & \mathbf{X}_2^T\mathbf{X}_2\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{X}_1 & 0\\ 0 & \mathbf{X}_2\end{bmatrix}^T\begin{bmatrix}\mathbf{y}_1 - \mathbf{Z}_1\widehat{\boldsymbol{\omega}}_1 - \mathbf{X}_{l1}\widehat{\boldsymbol{\gamma}}_1\\ \mathbf{y}_2 - \mathbf{Z}_2\widehat{\boldsymbol{\omega}}_2 - \mathbf{X}_{l2}\widehat{\boldsymbol{\gamma}}_2\end{bmatrix}.$$

terms involving $\mathbf{y}$ can be approximated by $\mathbf{X}_{l1}^T\mathbf{y}_{l1} \approx \sqrt{\frac{n_1}{p}}\mathbf{z}_{l1}$, $\mathbf{X}_{l2}^T\mathbf{y}_{l2} \approx \sqrt{\frac{n_2}{p}}\mathbf{z}_{l2}$, $\mathbf{X}_1^T\mathbf{y}_1 \approx \sqrt{\frac{n_1}{p}}\mathbf{z}_1$, and $\mathbf{X}_2^T\mathbf{y}_2 \approx \sqrt{\frac{n_2}{p}}\mathbf{z}_2$, where $\mathbf{z}_{l1} \in \mathbb{R}^{l_1}$ and $\mathbf{z}_{l2} \in \mathbb{R}^{l_2}$ are sub-vectors of $\mathbf{z}_1$ and $\mathbf{z}_2$ corresponding to the large-effect SNPs. Replacing the relevant terms with corresponding approximations leads to Equations 9 and 10.

Data and code availability

The publicly available GWAS summary statistics for estimating genetic correlations were obtained from the links summarized in Table S2. Access to genome-wide summary statistics from the Chinese data has to be approved by the contact authors (see Table S2). Researchers who wish to gain access to the data are required to submit the data request by sending an email to the contact authors. GWAS summary statistics from other studies can be accessed using the links provided in Table S2. The UKBB data are available through the UK Biobank Access Management System. The IPM BioMe biobank data are accessible on dbGap with accession number phs000925.v1.p1. The C++ software for XPA and the R package for XPASS are available online.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.03.002.

Acknowledgments

This work is supported in part by National Key R&D Program of China (2020YFA0713900), Hong Kong Research Grant Council (12301417, 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.

Declaration of interests

The authors declare no competing interests.

Received: October 25, 2020
Accepted: March 1, 2021
Published: March 25, 2021

Web resources

ANNOVAR, http://annovar.openbioinformatics.org/en/latest/
BOLT-LMM, https://alkesgroup.broadinstitute.org/BOLT-LMM
C++ software for XPA, https://github.com/YangLabHKUST/XPA
dbGaP, https://www.ncbi.nlm.nih.gov/gap
GCTA, https://cnsgenomics.com/software/gcta/#Overview
GCTB, https://cnsgenomics.com/software/gctb/#Overview
glmnetPlus, https://github.com/junyangq/glmnetPlus
lassosum, https://github.com/tshmak/lassosum
LDSC, https://github.com/bulik/ldsc
LDpred, https://github.com/bvilhjal/ldpred
PLINK, https://www.cog-genomics.org/plink/
Popcorn, https://github.com/brielin/Popcorn
R package for XPASS, https://github.com/YangLabHKUST/XPASS
snpnet, https://github.com/junyangq/snpnet
UKBB, https://www.ukbiobank.ac.uk

References

1. Torkamani, A., Wineinger, N.E., and Topol, E.J. (2018). The personal and clinical utility of polygenic risk scores. Nat. Rev. Genet. *19*, 581–590.
2. Abul-Husn, N.S., Manickam, K., Jones, L.K., Wright, E.A., Hartzel, D.N., Gonzaga-Jauregui, C., O'Dushlaine, C., Leader, J.B., Lester Kirchner, H., Lindbuchler, D.M., et al. (2016). Genetic identification of familial hypercholesterolemia within a single U.S. health care system. Science *354*, aaf7000.
3. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *50*, 1219–1224.
4. Craig, J.E., Han, X., Qassim, A., Hassall, M., Cooke Bailey, J.N., Kinzy, T.G., Khawaja, A.P., An, J., Marshall, H., Gharahkhani, P., et al.; NEIGHBORHOOD consortium; and UK Biobank Eye and Vision Consortium (2020). Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression. Nat. Genet. *52*, 160–166.
5. Bustamante, C.D., Burchard, E.G., and De la Vega, F.M. (2011). Genomics for the world. Nature *475*, 163–165.
6. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161–164.
7. Need, A.C., and Goldstein, D.B. (2009). Next generation disparities in human genomics: concerns and remedies. Trends Genet. *25*, 489–494.
8. Mills, M.C., and Rahal, C. (2020). The GWAS Diversity Monitor tracks diversity by disease in real time. Nat. Genet. *52*, 242–243.
9. Lewis, C.M., and Vassos, E. (2020). Polygenic risk scores: from research tools to clinical instruments. Genome Med. *12*, 44.
10. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518.
11. Peterson, R.E., Kuchenbaecker, K., Walters, R.K., Chen, C.Y., Popejoy, A.B., Periyasamy, S., Lam, M., Iyegbe, C., Strawbridge, R.J., Brick, L., et al. (2019). Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations. Cell *179*, 589–603.
12. Lam, M., Chen, C.Y., Li, Z., Martin, A.R., Bryois, J., Ma, X., Gaspar, H., Ikeda, M., Benyamin, B., Brown, B.C., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium; Indonesia Schizophrenia Consortium; and Genetic REsearch on schizophreniA neTwork-China and the Netherlands (GREAT-CN) (2019). Comparative genetic architectures of schizophrenia in East Asian and European populations. Nat. Genet. *51*, 1670–1678.

13. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591.
14. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. Am. J. Hum. Genet. *100*, 635–649.
15. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
16. Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T., et al.; International Multiple Sclerosis Genetics Consortium; and International IBD Genetics Consortium (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. *47*, 979–986.
17. Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D., Thompson, D., Ballinger, D.G., Struewing, J.P., Morrison, J., Field, H., Luben, R., et al.; SEARCH collaborators; kConFab; and AOCS Management Group (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. Nature *447*, 1087–1093.
18. Mahajan, A., Go, M.J., Zhang, W., Below, J.E., Gaulton, K.J., Ferreira, T., Horikoshi, M., Johnson, A.D., Ng, M.C., Prokopenko, I., et al.; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium; Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium; Mexican American Type 2 Diabetes (MAT2D) Consortium; and Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples (T2D-GENES) Consortium (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat. Genet. *46*, 234–244.
19. Waters, K.M., Stram, D.O., Hassanein, M.T., Le Marchand, L., Wilkens, L.R., Maskarinec, G., Monroe, K.R., Kolonel, L.N., Altshuler, D., Henderson, B.E., and Haiman, C.A. (2010). Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups. PLoS Genet. *6*, e1001078.
20. McGuire, A.L., Gabriel, S., Tishkoff, S.A., Wonkam, A., Chakravarti, A., Furlong, E.E.M., Treutlein, B., Meissner, A., Chang, H.Y., López-Bigas, N., et al. (2020). The road ahead in genetics and genomics. Nat. Rev. Genet. *21*, 581–596.
21. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.
22. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. *33*, 1–22.
23. Habier, D., Fernando, R.L., Kizilkaya, K., and Garrick, D.J. (2011). Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics *12*, 186.
24. Erbe, M., Hayes, B.J., Matukumalli, L.K., Goswami, S., Bowman, P.J., Reich, C.M., Mason, B.A., and Goddard, M.E. (2012). Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. *95*, 4114–4129.
25. Zeng, J., de Vlaming, R., Wu, Y., Robinson, M.R., Lloyd-Jones, L.R., Yengo, L., Yap, C.X., Xue, A., Sidorenko, J., McRae, A.F., et al. (2018). Signatures of negative selection in the genetic architecture of human complex traits. Nat. Genet. *50*, 746–753.
26. Qian, J., Tanigawa, Y., Du, W., Aguirre, M., Chang, C., Tibshirani, R., Rivas, M.A., and Hastie, T. (2020). A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. PLoS Genet. *16*, e1009141.
27. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. *88*, 294–305.
28. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., Sklar, P.; and International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.
29. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.R., Bhatia, G., Do, R., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet. *97*, 576–592.
30. Mak, T.S.H., Porsch, R.M., Choi, S.W., Zhou, X., and Sham, P.C. (2017). Polygenic scores via penalized regression on summary statistics. Genet. Epidemiol. *41*, 469–480.
31. Turley, P., Walters, R.K., Maghzian, O., Okbay, A., Lee, J.J., Fontana, M.A., Nguyen-Viet, T.A., Wedow, R., Zacher, M., Furlotte, N.A., et al.; 23andMe Research Team; and Social Science Genetic Association Consortium (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat. Genet. *50*, 229–237.
32. Coram, M.A., Fang, H., Candille, S.I., Assimes, T.L., and Tang, H. (2017). Leveraging multi-ethnic evidence for risk assessment of quantitative traits in minority populations. Am. J. Hum. Genet. *101*, 218–226.
33. Akiyama, M., Okada, Y., Kanai, M., Takahashi, A., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2017). Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nat. Genet. *49*, 1458–1467.
34. Akiyama, M., Ishigaki, K., Sakaue, S., Momozawa, Y., Horikoshi, M., Hirata, M., Matsuda, K., Ikegawa, S., Takahashi, A., Kanai, M., et al. (2019). Characterizing rare and low-frequency height-associated variants in the Japanese population. Nat. Commun. *10*, 4393.
35. Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N.L., and Yu, W. (2010). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am. J. Hum. Genet. *87*, 325–340.
36. Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. *47*, 284–290.
37. Wu, Y., and Sankararaman, S. (2018). A scalable estimator of SNP heritability for biobank-scale data. Bioinformatics *34*, i187–i194.

38. Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2012). Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. *91*, 1011–1021.
39. Zhou, X. (2017). A unified framework for variance component estimation with summary statistics in genome-wide association studies. Ann. Appl. Stat. *11*, 2027–2051.
40. Lu, Q., Li, B., Ou, D., Erlendsdottir, M., Powles, R.L., Jiang, T., Hu, Y., Chang, D., Jin, C., Dai, W., et al. (2017). A powerful approach to estimating annotation-stratified genetic covariance via GWAS summary statistics. Am. J. Hum. Genet. *101*, 939–964.
41. Berisa, T., and Pickrell, J.K. (2016). Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics *32*, 283–285.
42. Yang, S., and Zhou, X. (2020). Accurate and scalable construction of polygenic scores in large biobank data sets. Am. J. Hum. Genet. *106*, 679–693.
43. Lloyd-Jones, L.R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K.E., Wang, H., Zheng, Z., Magi, R., Esko, T., et al. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat. Commun. *10*, 5086.
44. Yao, X., Tang, S., Bian, B., Wu, X., Chen, G., and Wang, C.-C. (2017). Improved phylogenetic resolution for Y-chromosome Haplogroup O2a1c-002611. Sci. Rep. *7*, 1146.
45. Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods *10*, 5–6.
46. Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.
47. Liu, S., Huang, S., Chen, F., Zhao, L., Yuan, Y., Francis, S.S., Fang, L., Li, Z., Lin, L., Liu, R., et al. (2018). Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and Chinese population history. Cell *175*, 347–359.e14.
48. Chen, J., Zheng, H., Bei, J.-X., Sun, L., Jia, W.H., Li, T., Zhang, F., Seielstad, M., Zeng, Y.-X., Zhang, X., and Liu, J. (2009). Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. Am. J. Hum. Genet. *85*, 775–785.
49. Xu, S., Yin, X., Li, S., Jin, W., Lou, H., Yang, L., Gong, X., Wang, H., Shen, Y., Pan, X., et al. (2009). Genomic dissection of population substructure of Han Chinese and its implication in association studies. Am. J. Hum. Genet. *85*, 762–774.
50. Fuchsberger, C., Abecasis, G.R., and Hinds, D.A. (2015). minimac2: faster genotype imputation. Bioinformatics *31*, 782–784.
51. Jiang, J., Li, C., Paul, D., Yang, C., and Zhao, H. (2016). On high-dimensional misspecified mixed model analysis in genome-wide association study. Ann. Statist. *44*, 2127–2160.
52. Nelis, M., Esko, T., Mägi, R., Zimprich, F., Zimprich, A., Toncheva, D., Karachanak, S., Piskácková, T., Balascák, I., Peltonen, L., et al. (2009). Genetic structure of Europeans: a view from the North-East. PLoS ONE *4*, e5472.
53. Lee, S.H., Clark, S., and van der Werf, J.H.J. (2017). Estimation of genomic prediction accuracy from reference populations with varying degrees of relationship. PLoS ONE *12*, e0189775.
54. 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. Nat. Rev. Genet. *20*, 567–581.
55. Truong, B., Zhou, X., Shin, J., Li, J., van der Werf, J.H.J., Le, T.D., and Lee, S.H. (2020). Efficient polygenic risk scores for biobank scale data by exploiting phenotypes from inferred relatives. Nat. Commun. *11*, 3074.
56. Li, C., Yang, C., Gelernter, J., and Zhao, H. (2014). Improving genetic risk prediction by leveraging pleiotropy. Hum. Genet. *133*, 639–650.
57. Maier, R., Moser, G., Chen, G.-B., Ripke, S., Coryell, W., Potash, J.B., Scheftner, W.A., Shi, J., Weissman, M.M., Hultman, C.M., et al.; Cross-Disorder Working Group of the Psychiatric Genomics Consortium (2015). Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. Am. J. Hum. Genet. *96*, 283–294.
58. Maier, R.M., Zhu, Z., Lee, S.H., Trzaskowski, M., Ruderfer, D.M., Stahl, E.A., Ripke, S., Wray, N.R., Yang, J., Visscher, P.M., and Robinson, M.R. (2018). Improving genetic prediction by leveraging genetic correlations among human diseases and traits. Nat. Commun. *9*, 989.
59. Weissbrod, O., Flint, J., and Rosset, S. (2018). Estimating SNP-based heritability and genetic correlation in case-control studies directly and with summary statistics. Am. J. Hum. Genet. *103*, 89–99.
60. Yang, L., Neale, B.M., Liu, L., Lee, S.H., Wray, N.R., Ji, N., Li, H., Qian, Q., Wang, D., Li, J., et al.; Psychiatric GWAS Consortium: ADHD Subgroup (2013). Polygenic transmission and complex neuro developmental network for attention deficit hyperactivity disorder: genome-wide association study of both common and rare variants. Am. J. Med. Genet. B. Neuropsychiatr. Genet. *162B*, 419–430.
61. Speed, D., Cai, N., Johnson, M.R., Nejentsev, S., Balding, D.J.; and UCLEB Consortium (2017). Reevaluation of SNP heritability in complex human traits. Nat. Genet. *49*, 986–992.
62. Speed, D., Holmes, J., and Balding, D.J. (2020). Evaluating and improving heritability models using summary statistics. Nat. Genet. *52*, 458–462.
63. Turchin, M.C., Chiang, C.W., Palmer, C.D., Sankararaman, S., Reich, D., Hirschhorn, J.N.; and Genetic Investigation of ANthropometric Traits (GIANT) Consortium (2012). Evidence of widespread selection on standing variation in Europe at height-associated SNPs. Nat. Genet. *44*, 1015–1019.
64. Cai, M., Chen, L.S., Liu, J., and Yang, C. (2020). IGREX for quantifying the impact of genetically regulated expression on phenotypes. NAR Genom Bioinform *2*, a010.
65. Farh, K.K., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature *518*, 337–343.
66. Shi, X., Chai, X., Yang, Y., Cheng, Q., Jiao, Y., Chen, H., Huang, J., Yang, C., and Liu, J. (2020). A tissue-specific collaborative mixed model for jointly analyzing multiple tissues in transcriptome-wide association studies. Nucleic Acids Res. *48*, e109, e109.
67. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K., et al.; ReproGen Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; and RACI Consortium (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.
68. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H.,

Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

69. Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat. Genet. *45*, 124–130.
70. Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573.
71. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al.; Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.
72. Ming, J., Dai, M., Cai, M., Wan, X., Liu, J., and Yang, C. (2018). LSMM: a statistical approach to integrating functional annotations with genome-wide association studies. Bioinformatics *34*, 2788–2796.
73. Ming, J., Wang, T., and Yang, C. (2020). LPM: a latent probit model to characterize the relationship among complex traits using summary statistics from multiple GWASs and functional annotations. Bioinformatics *36*, 2506–2514.
74. Hu, Y., Lu, Q., Powles, R., Yao, X., Yang, C., Fang, F., Xu, X., and Zhao, H. (2017). Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS Comput. Biol. *13*, e1005589.
75. Marquez-Luna, C., Gazal, S., Loh, P.-R., Kim, S.S., Furlotte, N., Auton, A., Price, A.L.; and 23andMe Research Team (2020). LDpred-funct: incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. bioRxiv. https://doi.org/10.1101/375337.
76. Amariuta, T., Ishigaki, K., Sugishita, H., Ohta, T., Koido, M., Dey, K.K., Matsuda, K., Murakami, Y., Price, A.L., Kawakami, E., et al. (2020). Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. Nat. Genet. *52*, 1346–1354.

The American Journal of Human Genetics, Volume *108*

Supplemental information

A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits

Mingxuan Cai, Jiashun Xiao, Shunkang Zhang, Xiang Wan, Hongyu Zhao, Gang Chen, and Can Yang

1 Supplementary Figures

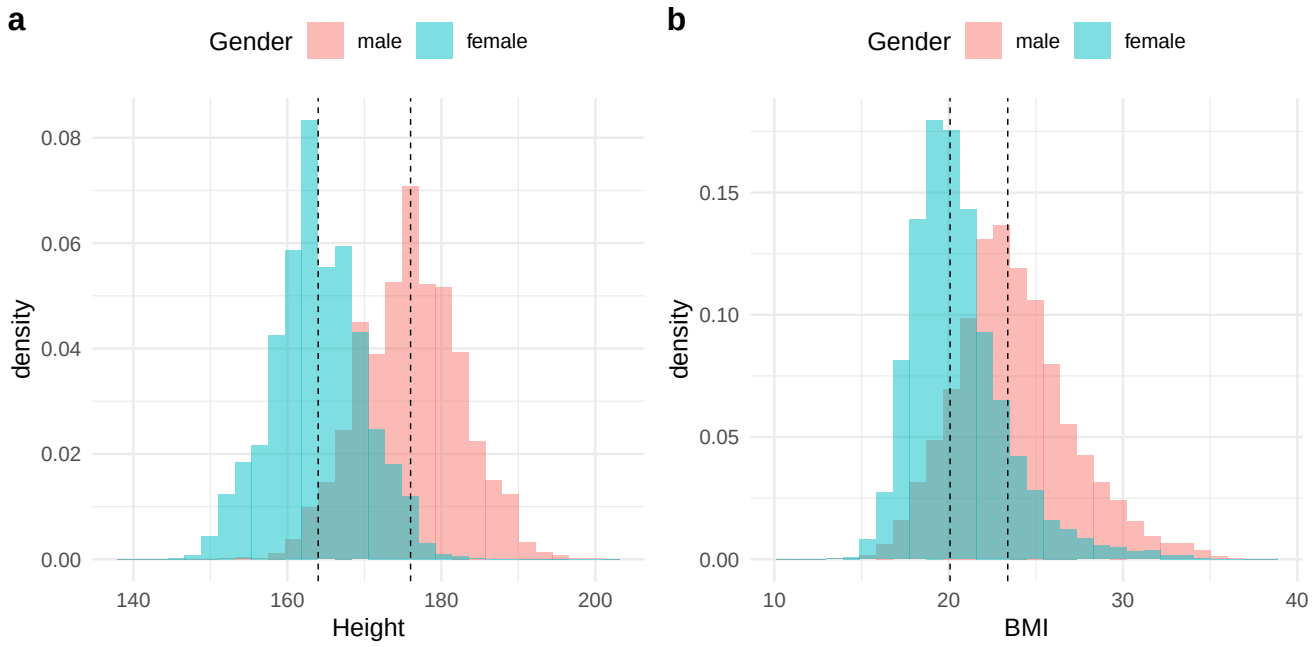

Figure S1: The distribution of height (**a**) and BMI (**b**) in Chinese dataset.

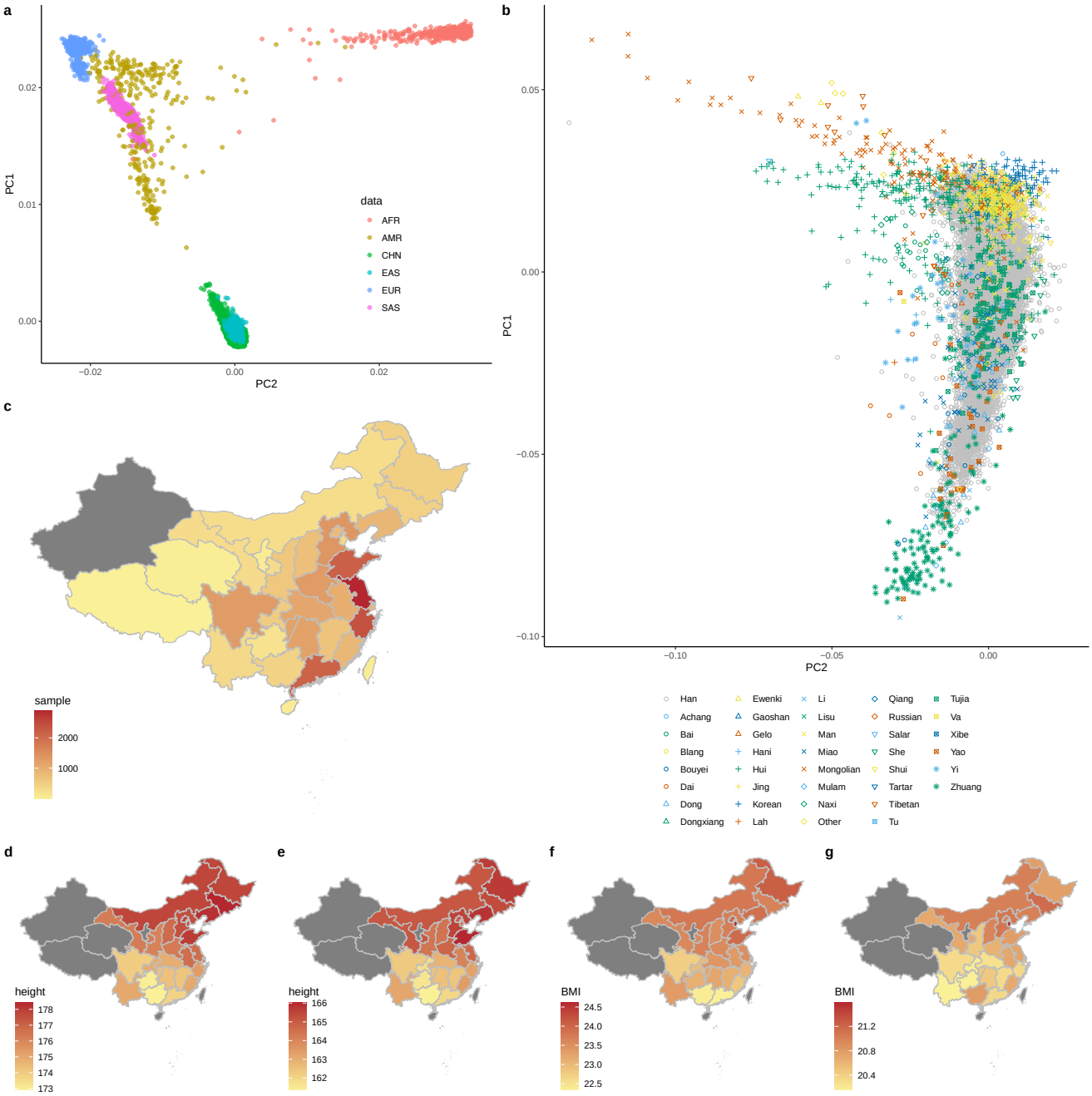

Figure S2: **a**, PCA of the combined samples from Chinese cohort and the 1000 Genomes Project. Chinese are genetically closest to East Asians in the 1000 Genomes project. **b**, PCA of Chinese participants only. The first two principal components reflect the longitudinal and latitudinal differentiation behind Chinese genetic structure. **c**, Distribution of genotyped individuals by province. The majority of genotyped samples are from the southeastern area of China. **d-g**, Average phenotypic values of male height (**d**), female height (**e**), male BMI (**f**) and female BMI (**g**) for provinces with more than 50 samples. Four administrative divisions Xinjiang, Tibet, Qinghai and Ningxia are shown in grey because their sample sizes are less than 50.

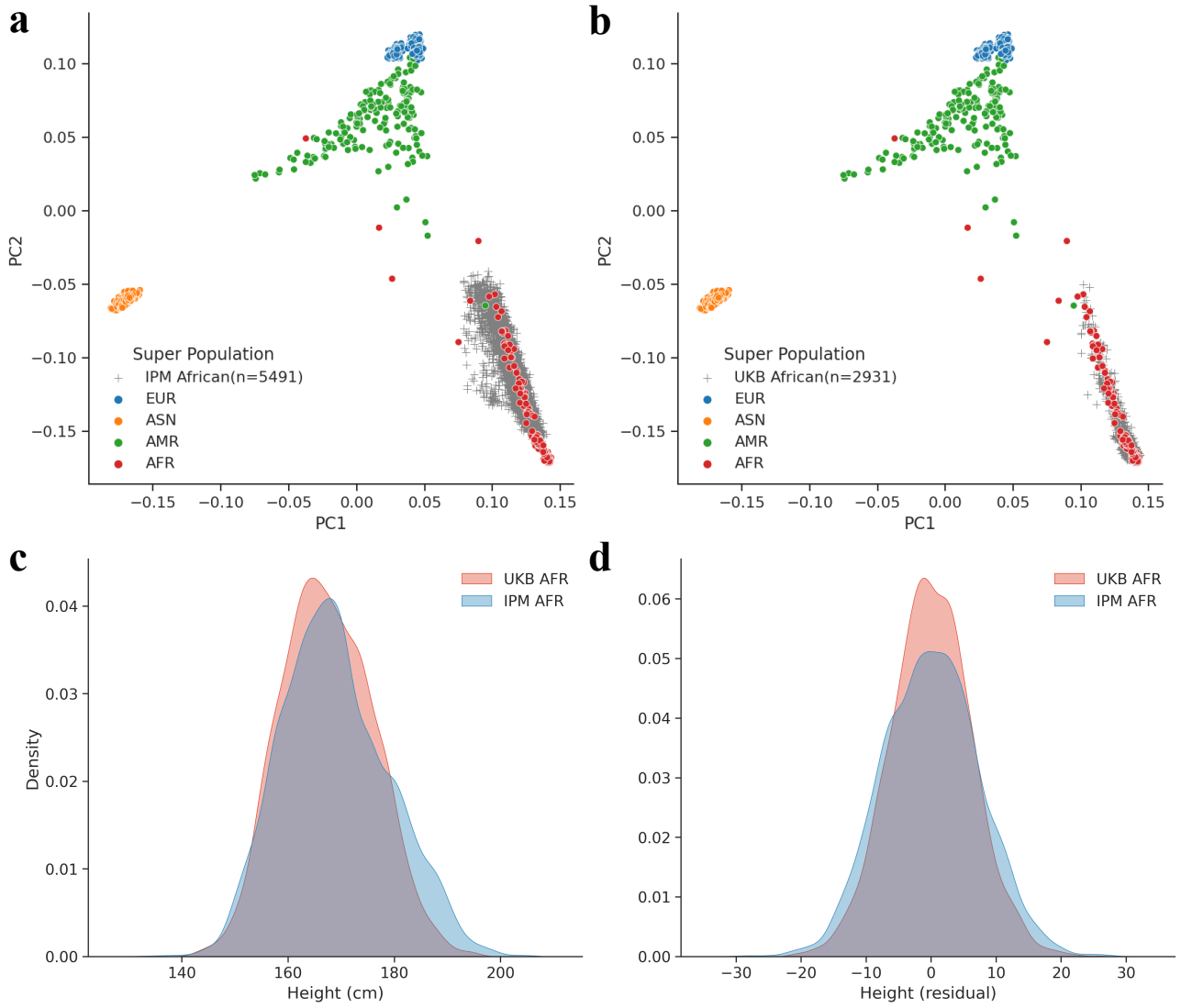

Figure S3: The PCA projection of IPM Afican participants (a) and UKBB African participants (b) to the 1000 Genome Project dataset. Kernel density estimation (KDE) of height (c) and its residual (d).

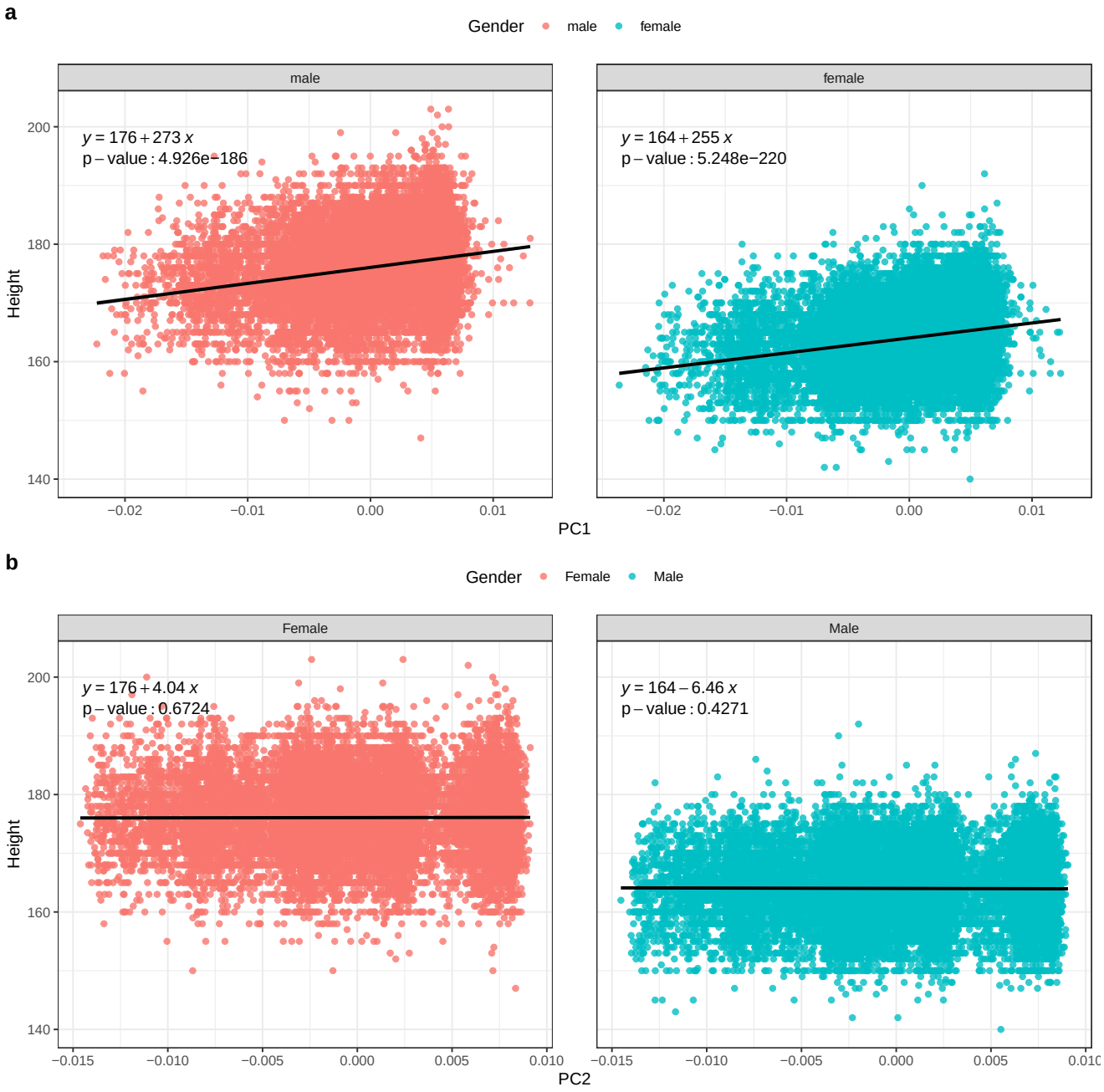

Figure S4: The relationship between height and first two principal components in Chinese dataset. (**a**) Height against the first principal component grouped by sex; (**b**) Height against the second principal component grouped by sex. The black lines represent the fitted regression between height and corresponding PCs. There is an increasing trend of height along the gradient of the first PC.

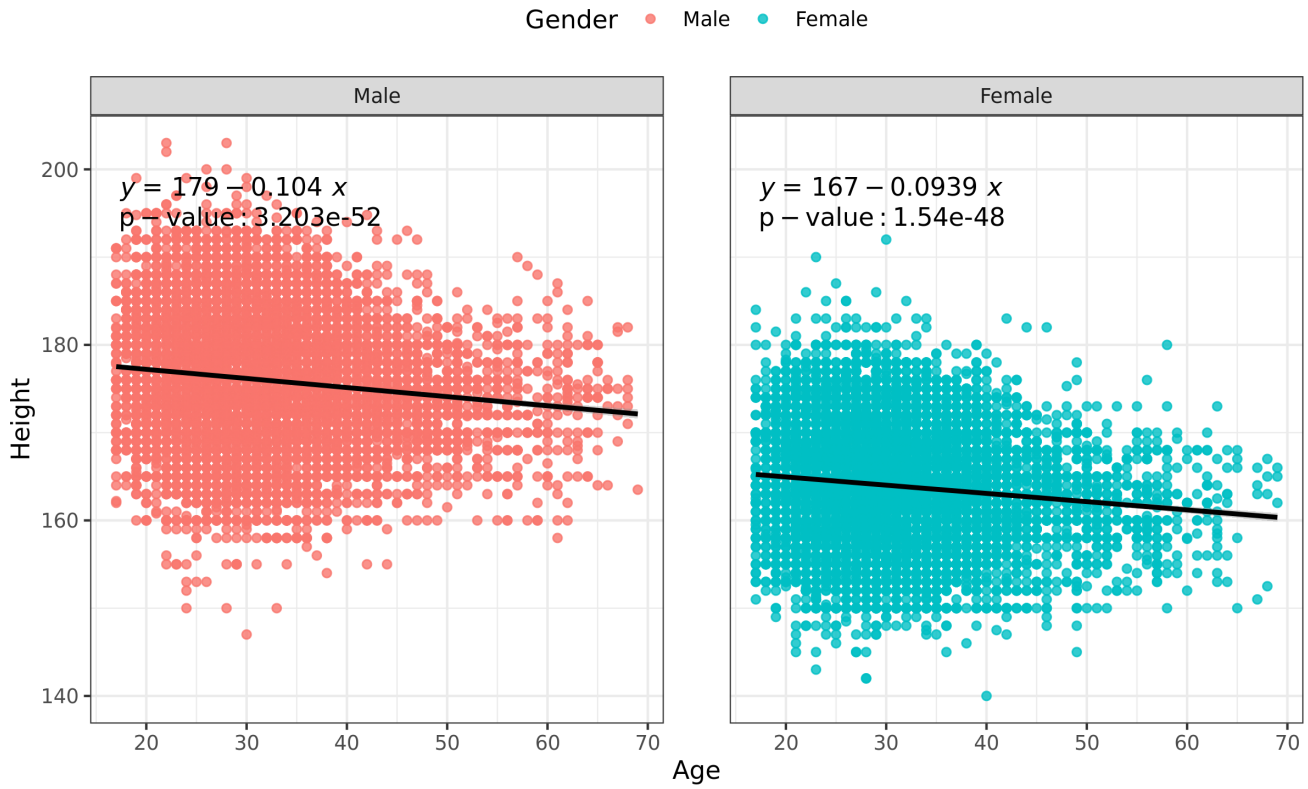

Figure S5: The relationship between height and age in Chinese dataset. (**a**) Height against age for males; (**b**) Height against age for females. The black lines represent the fitted regression between height and age. There is an decreasing trend of height for older people.

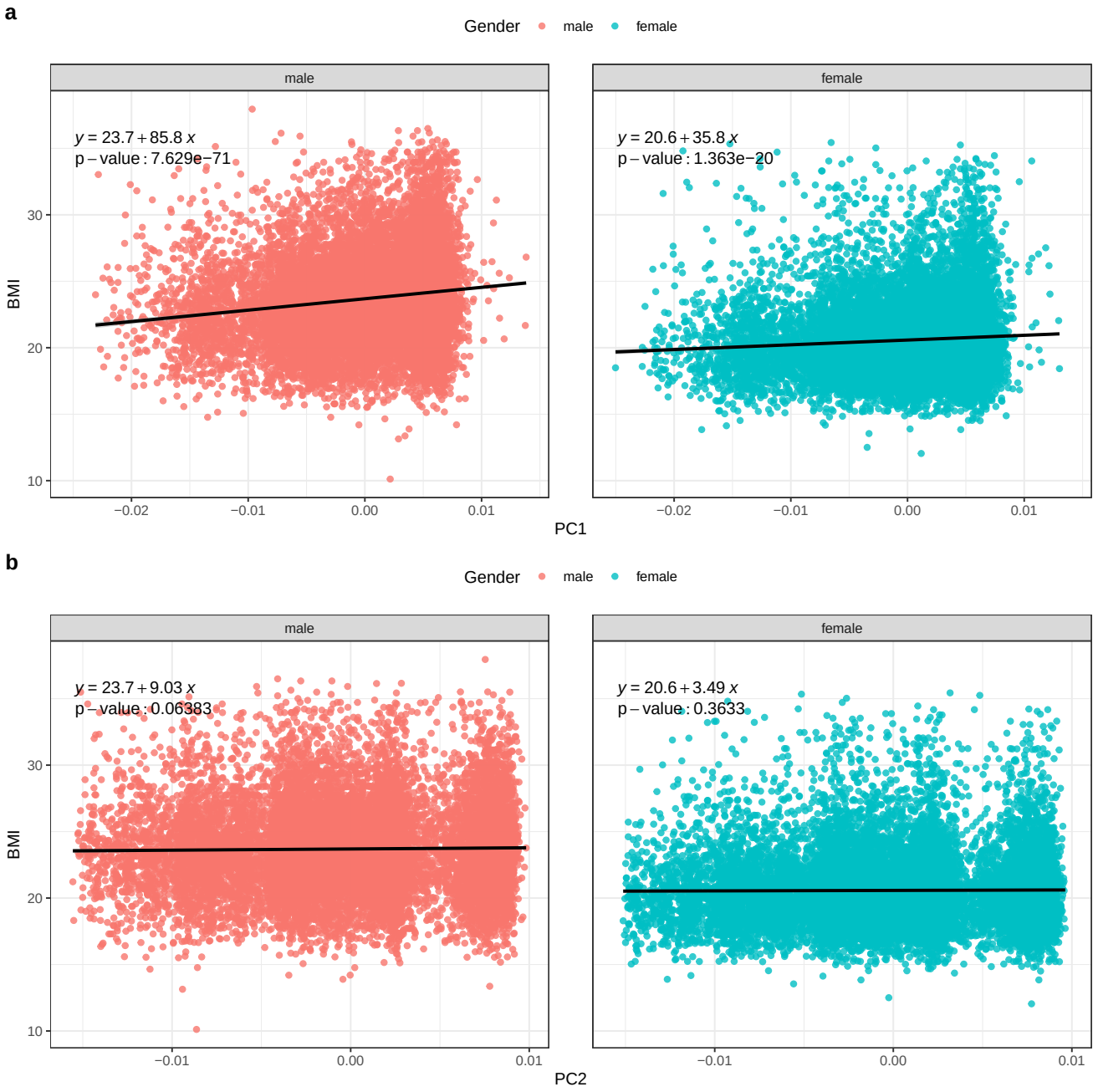

Figure S6: The relationship between BMI and first two principal components in Chinese dataset. (**a**) BMI against the first principal component grouped by sex; (**b**) BMI against the second principal component grouped by sex. The black lines represent the fitted regression between BMI and corresponding PCs. There is an increasing trend of BMI along the gradient of the first PC.

Figure S7: The relationship between BMI and age in Chinese dataset. (**a**) BMI against age for males; (**b**) BMI against age for females. The black lines represent the fitted regression between BMI and age. There is an increasing trend of BMI for older people.

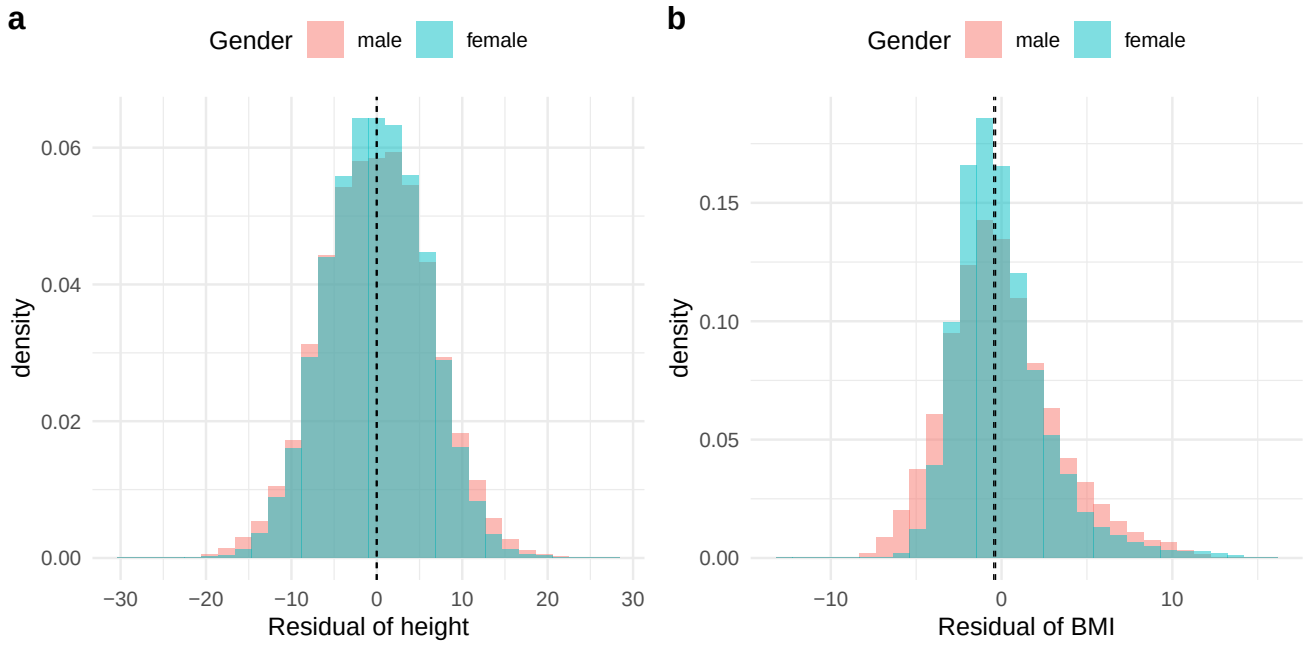

Figure S8: The distribution of residuals of height (**a**) and BMI (**b**) after adjusting for covariates in Chinese dataset.

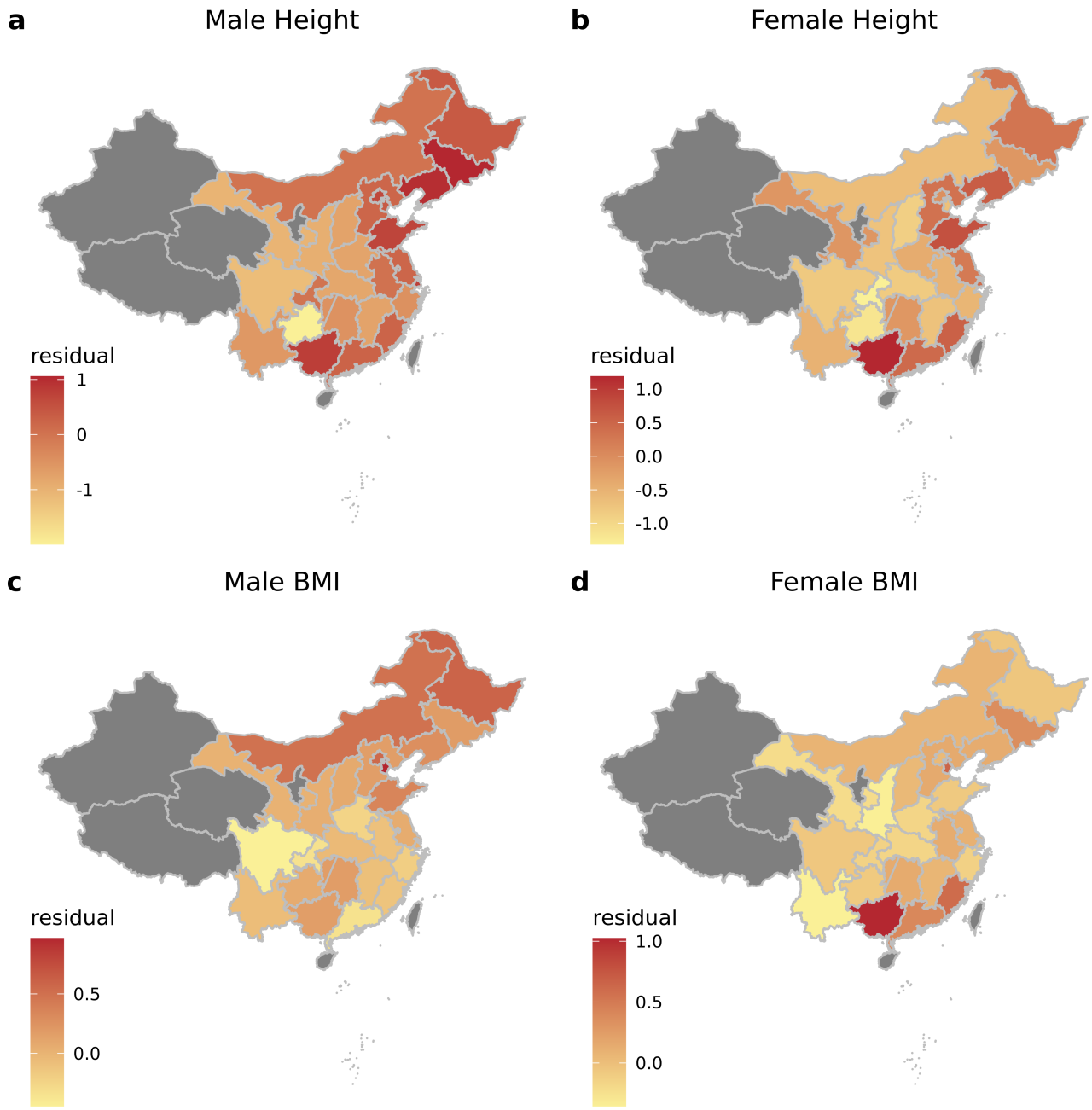

Figure S9: Average residual values after adjusting for the covariates of male height (**a**), female height (**b**), male BMI (**c**) and female BMI (**d**) in the provinces with more than 50 samples.

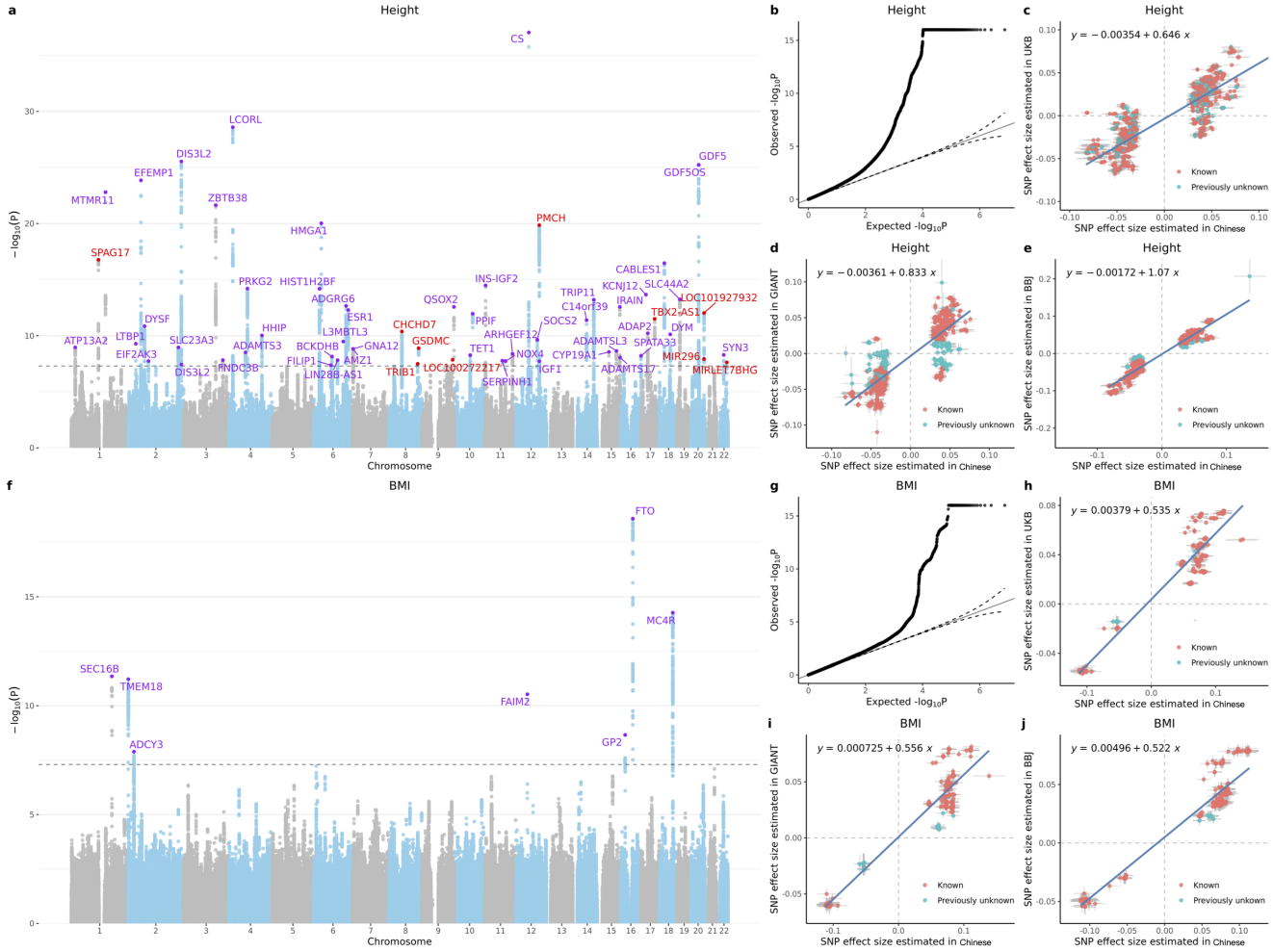

Figure S10: Manhattan plots of Chinese height (**a**) and BMI (**f**). The x axis shows chromosomal position, and the y axis shows significance on the $-\log_{10}$ scale. The dashed line marks the threshold for genome-wide significance (p-value$= 5 \times 10^{-8}$). Previously unknown associations are highlighted with purple dots, with the nearest gene names printed in purple. Known associations are highlighted with red dots, with the nearest gene names in red text. QQ plots of Chinese height (b) and BMI (f). **c-e** Comparison of the effect sizes for the genome-wide significant SNPs identified from the GWAS of Chinese height versus those identified in previous studies. **h-j** Comparison of the effect sizes for the genome-wide significant SNPs identified from the GWAS of Chinese BMI versus those identified in previous studies.

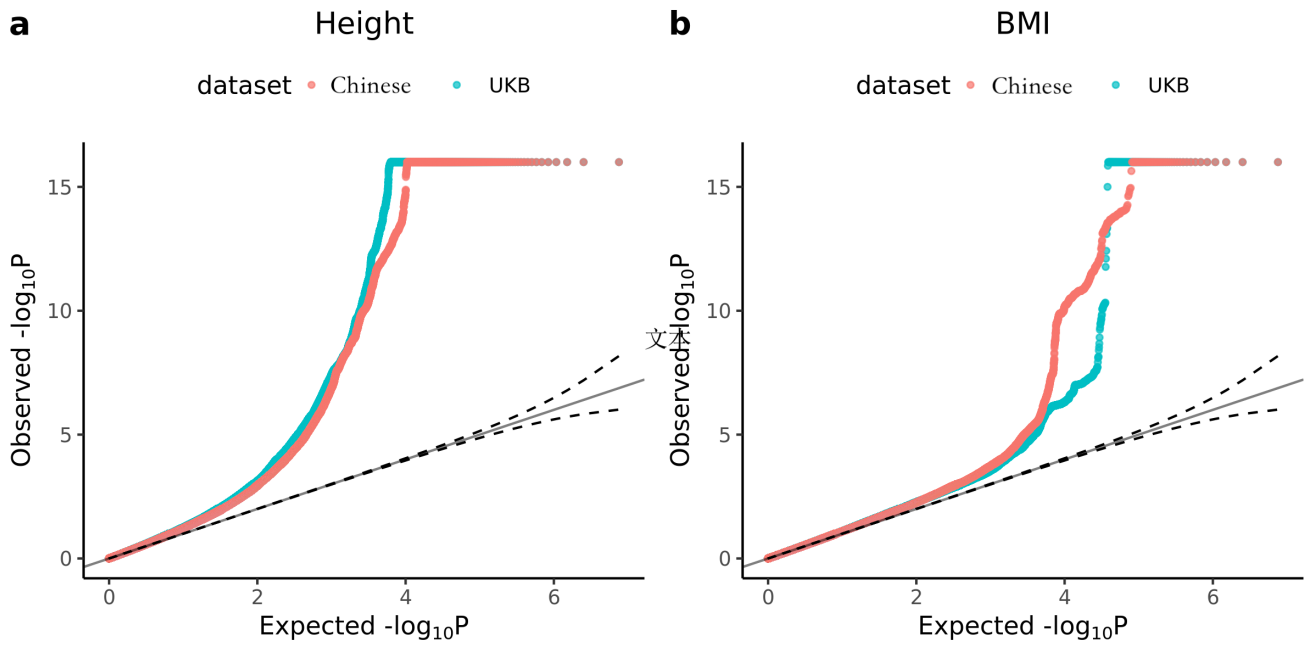

Figure S11: The Q-Q plot of GWAS p-values in height (**a**) and BMI (**b**) derived from Chinese dataset and 33,000 UKBB samples. We used the BOLT-LMM v2.3.2 to test for associations between phenotypes and SNPs. For Chinese population, we included age, sex and first 10 principal components as covariates. For UKBB, we used the top 20 principal components, age, squared age, sex, genotyping arrays and sequencing platforms as covariates.

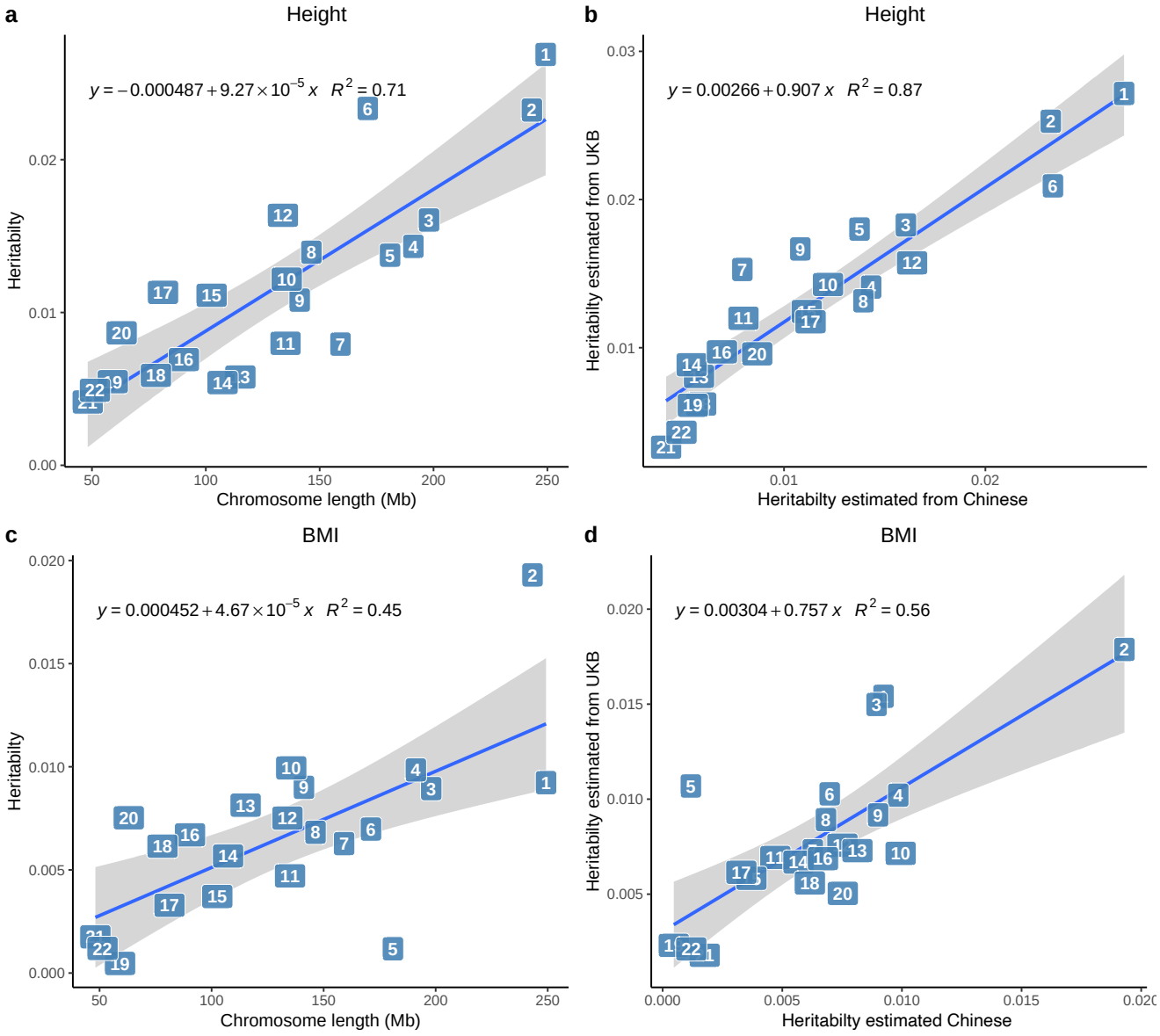

Figure S12: (**a**) The heritability of height against chromosome length in million base pair. (**b**) The chromosome heritabilities of height estimated from Chinese cohort against those estimated from UKBB. (**c**) The heritability of BMI against chromosome length in million base pair. (**d**) The chromosome heritabilities of BMI estimated from Chinese cohort against those estimated from UKBB.

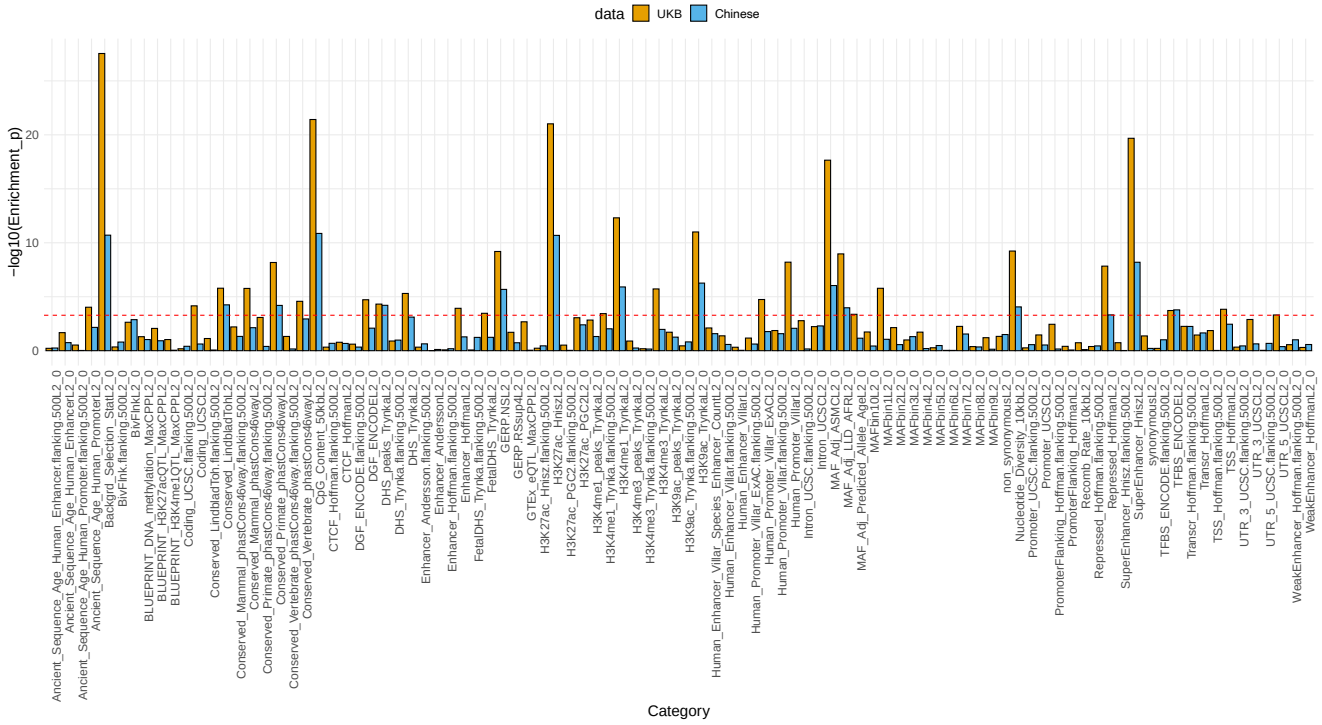

Figure S13: Enrichment of heritability for height in 95 functional annotations. The dashed line represents the significance threshold after Bonferroni correction (0.05/95). The LDSC software v1.0.0 was used to identify the heritability enrichment for the genome partitions in baseline model [1]. We used the LD scores provided by the Alkes Price group (`https://data.broadinstitute.org/alkesgroup/LDSCORE/`) in the analyses of Chinese cohort and UKBB.

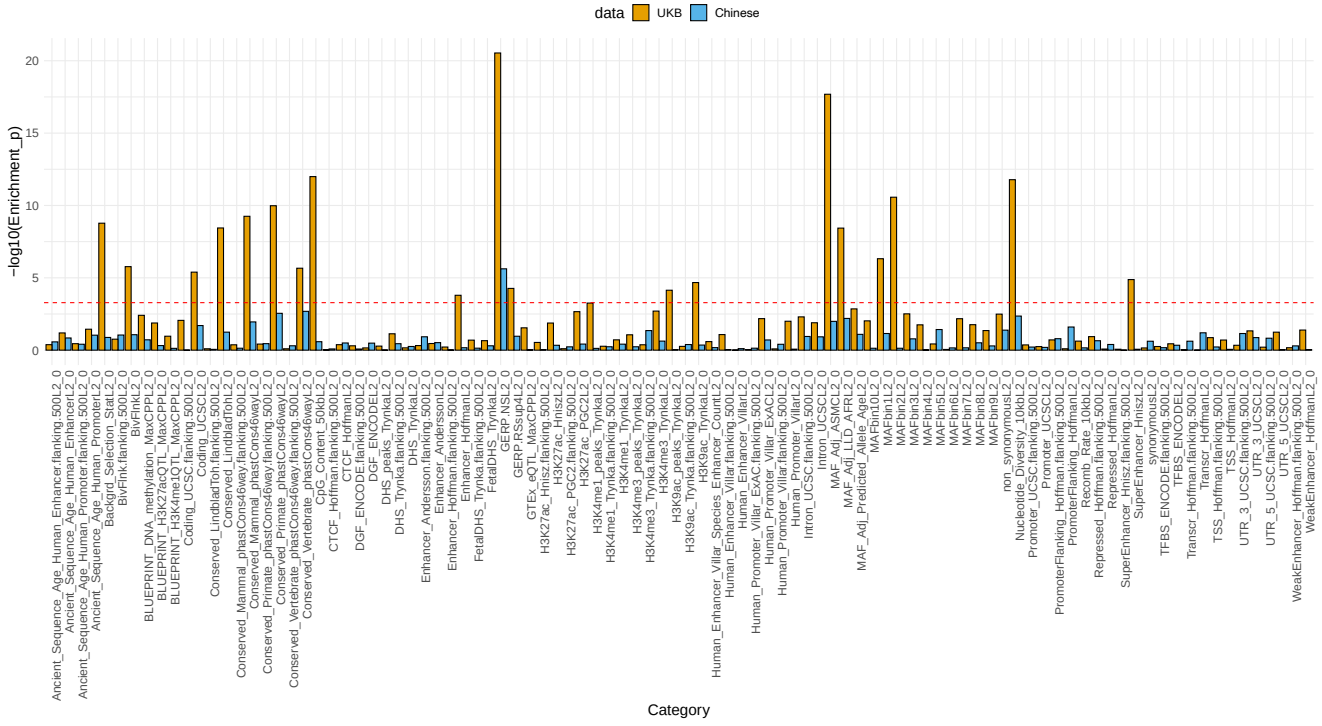

Figure S14: Enrichment of heritability for BMI in 95 functional annotations. The dashed line represents the significance threshold after Bonferroni correction (0.05/95).

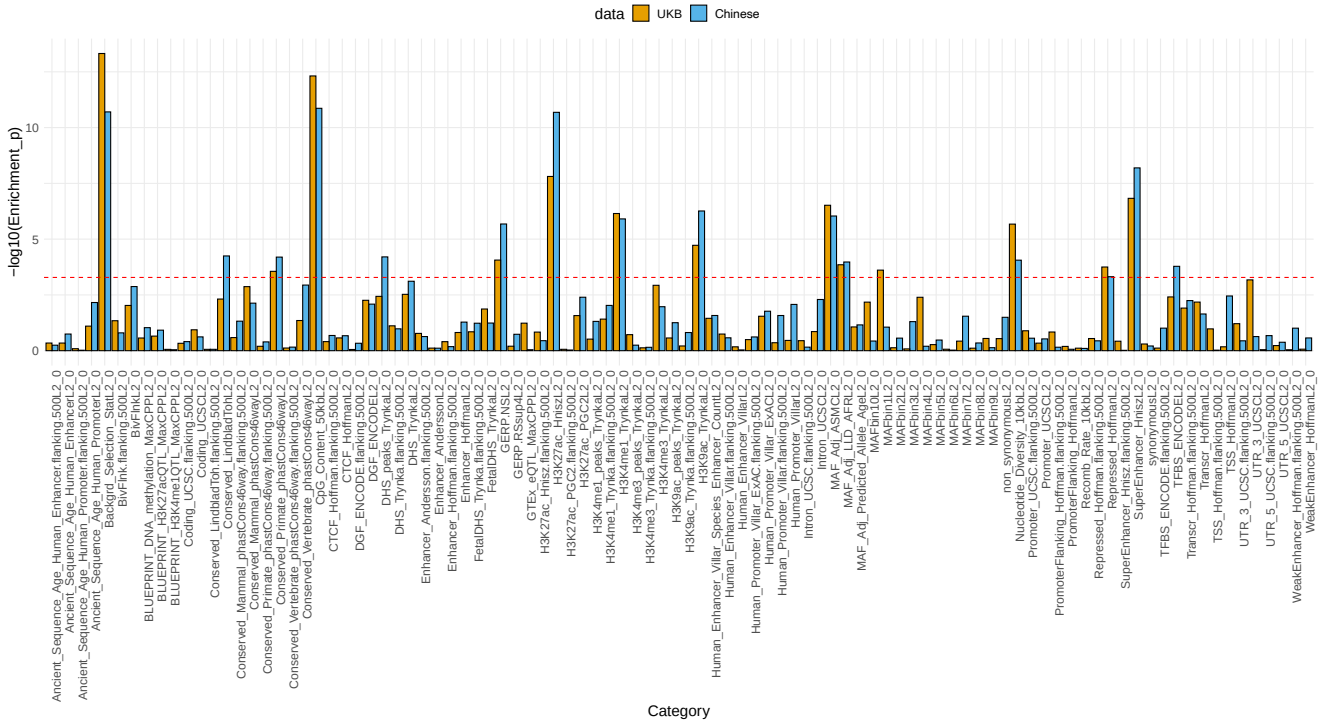

Figure S15: Enrichment of heritability for height in 95 functional annotations. We have randomly subsampled 20,000 individuals from UKBB to make the sample size comparable with Chinese cohort. The dashed line represents the significance threshold after Bonferroni correction (0.05/95).

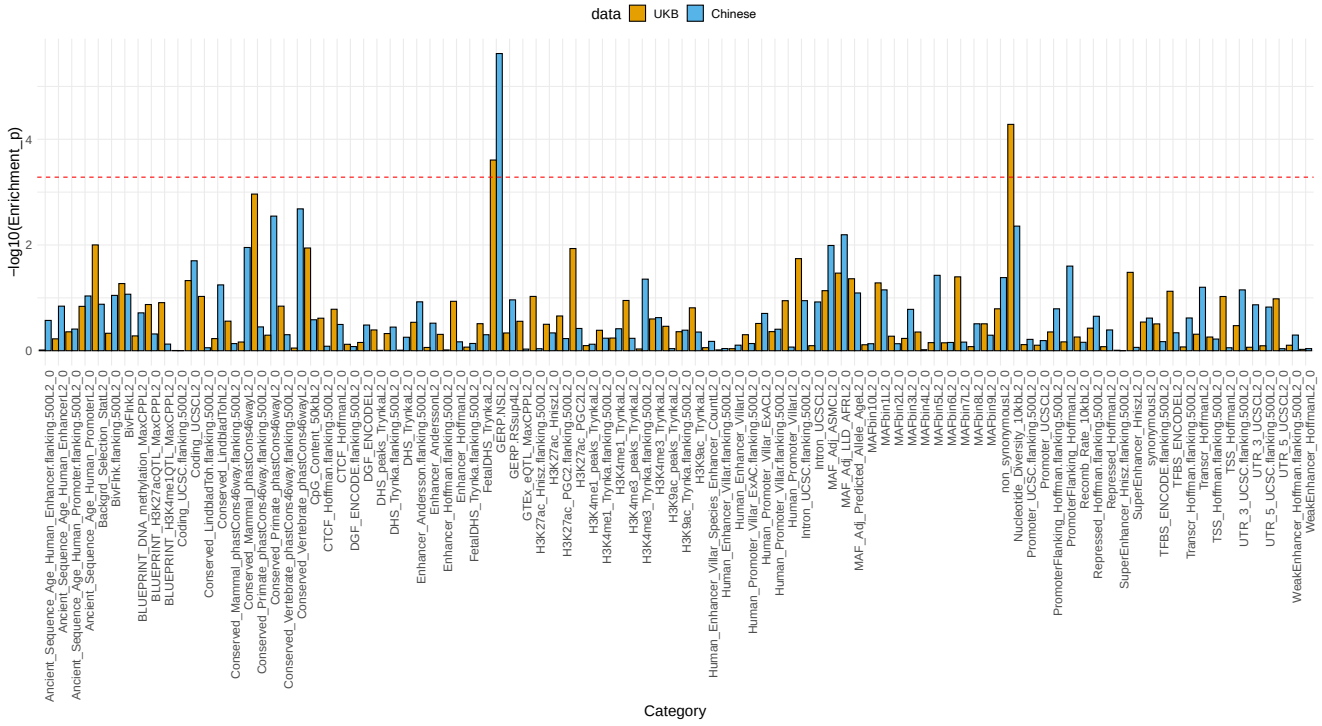

Figure S16: Enrichment of heritability for BMI in 95 functional annotations. We have randomly subsampled 20,000 individuals from UKBB to make the sample size comparable with Chinese cohort. The dashed line represents the significance threshold after Bonferroni correction (0.05/95).

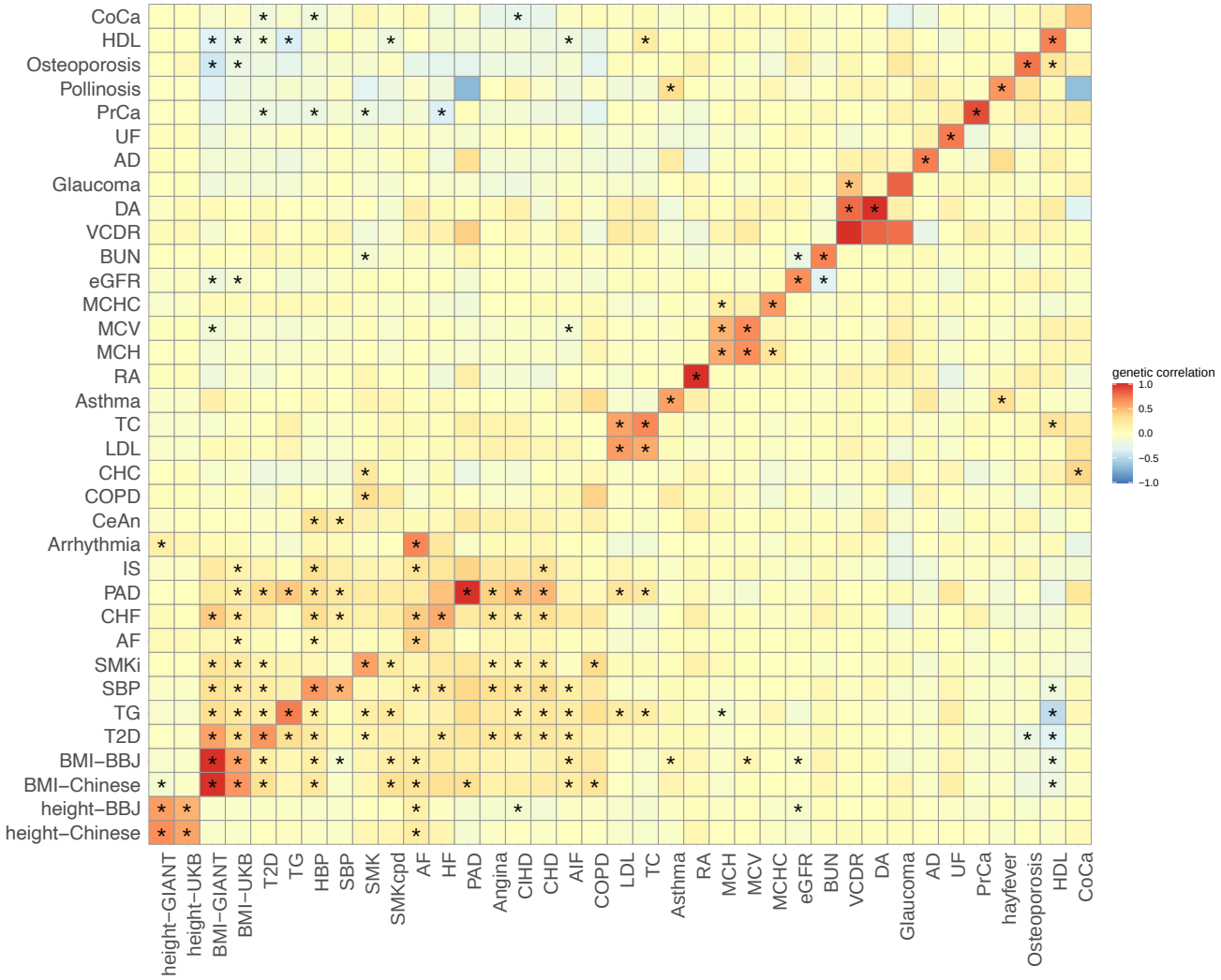

Figure S17: Trans-ancestry genetic correlations estimated by XPA. Positive correlations are colored in red. Negative correlations are colored in blue. Genetic correlations that are significantly different from zero after Bonferroni correction for the 1295 tests (p-value$< 0.05/1295$) are marked with asterisk.

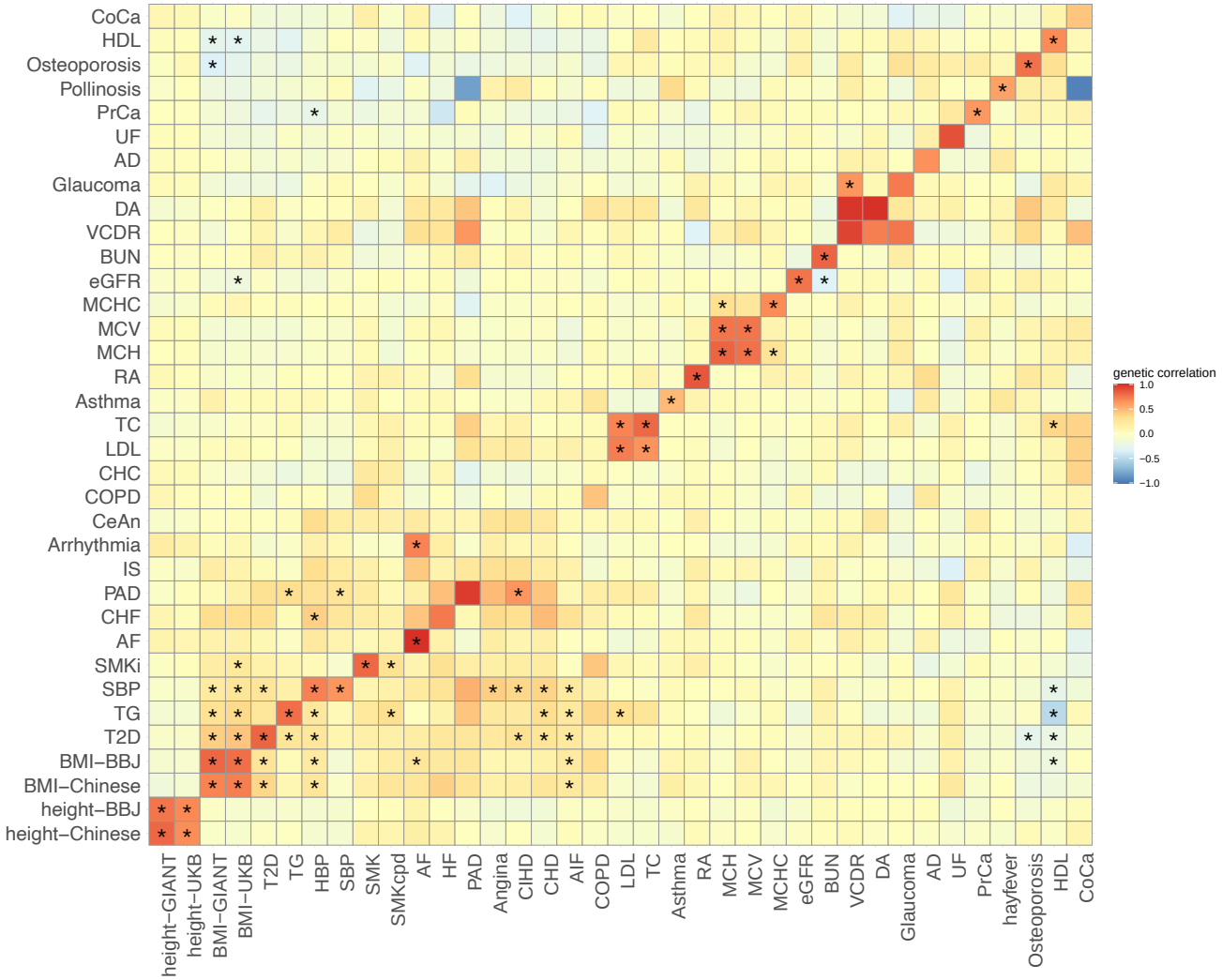

Figure S18: Trans-ancestry genetic correlations estimated by popcorn. Positive correlations are colored in red. Negative correlations are colored in blue. Genetic correlations that are significantly different from zero after Bonferroni correction for the $37 \times 35 = 1295$ tests (p-value$< 0.05/1295$) are marked with asterisk.

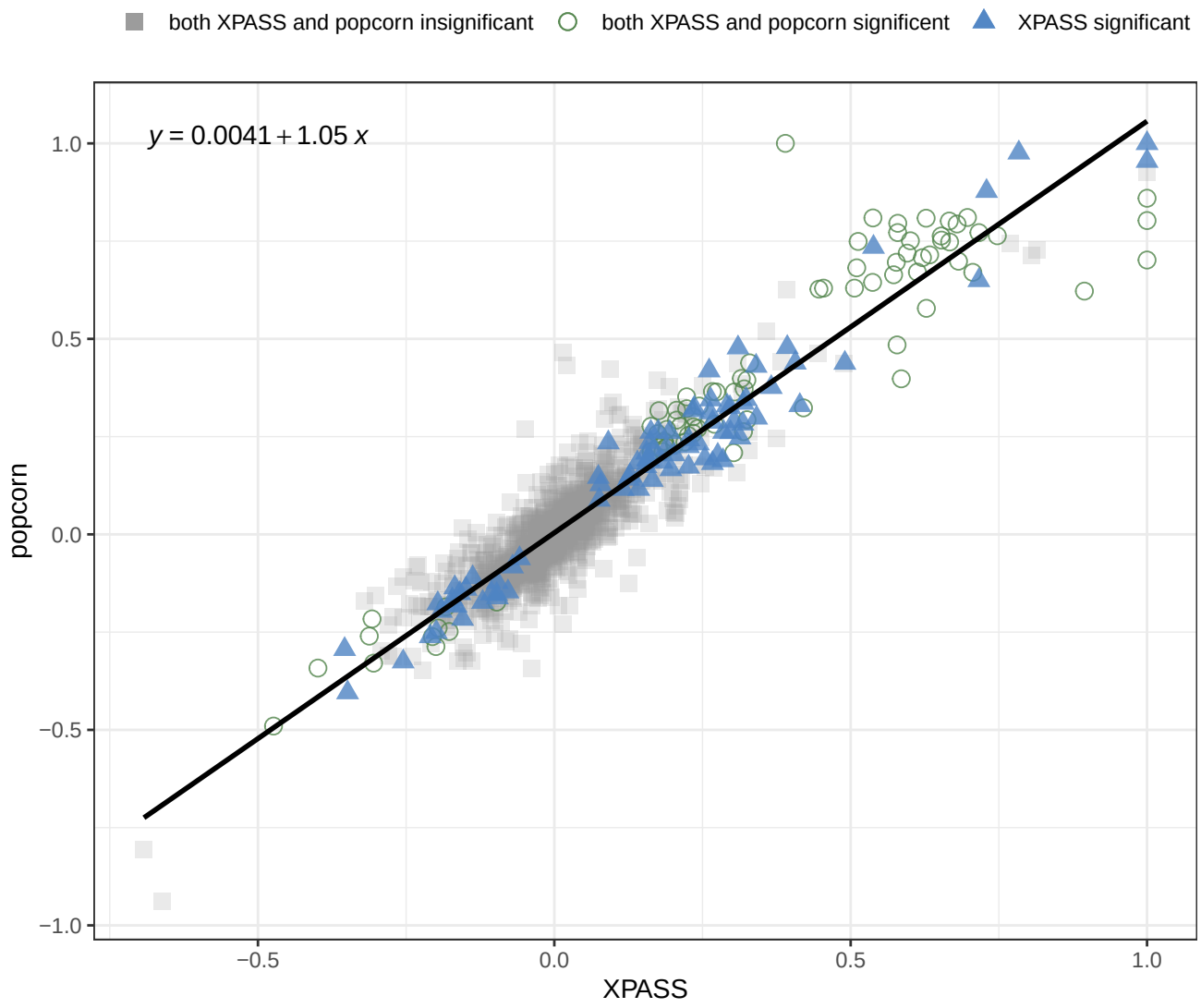

Figure S19: Genetic correlation estimates generated by popcorn versus those generated by XPASS. A regression line between the two sets of estimates is added.

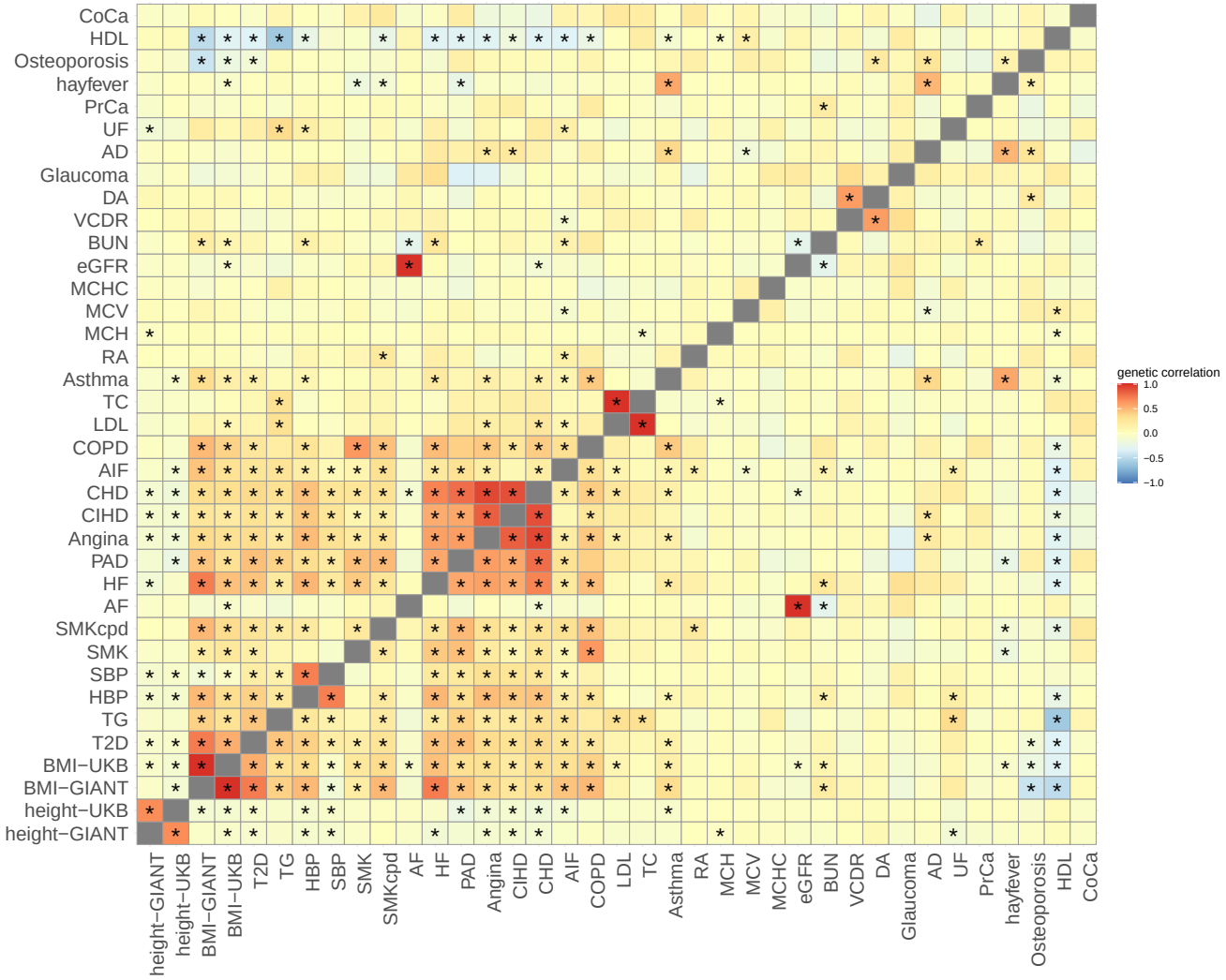

Figure S20: Genetic correlation of 37 traits in Europeans estimated by GNOVA. Positive correlations are colored in red. Negative correlations are colored in blue. Genetic correlations that are significantly different from zero after Bonferroni correction for the $(37 \times 36/2) = 666$ tests (p-value$< 0.05/666$) are marked with asterisk.

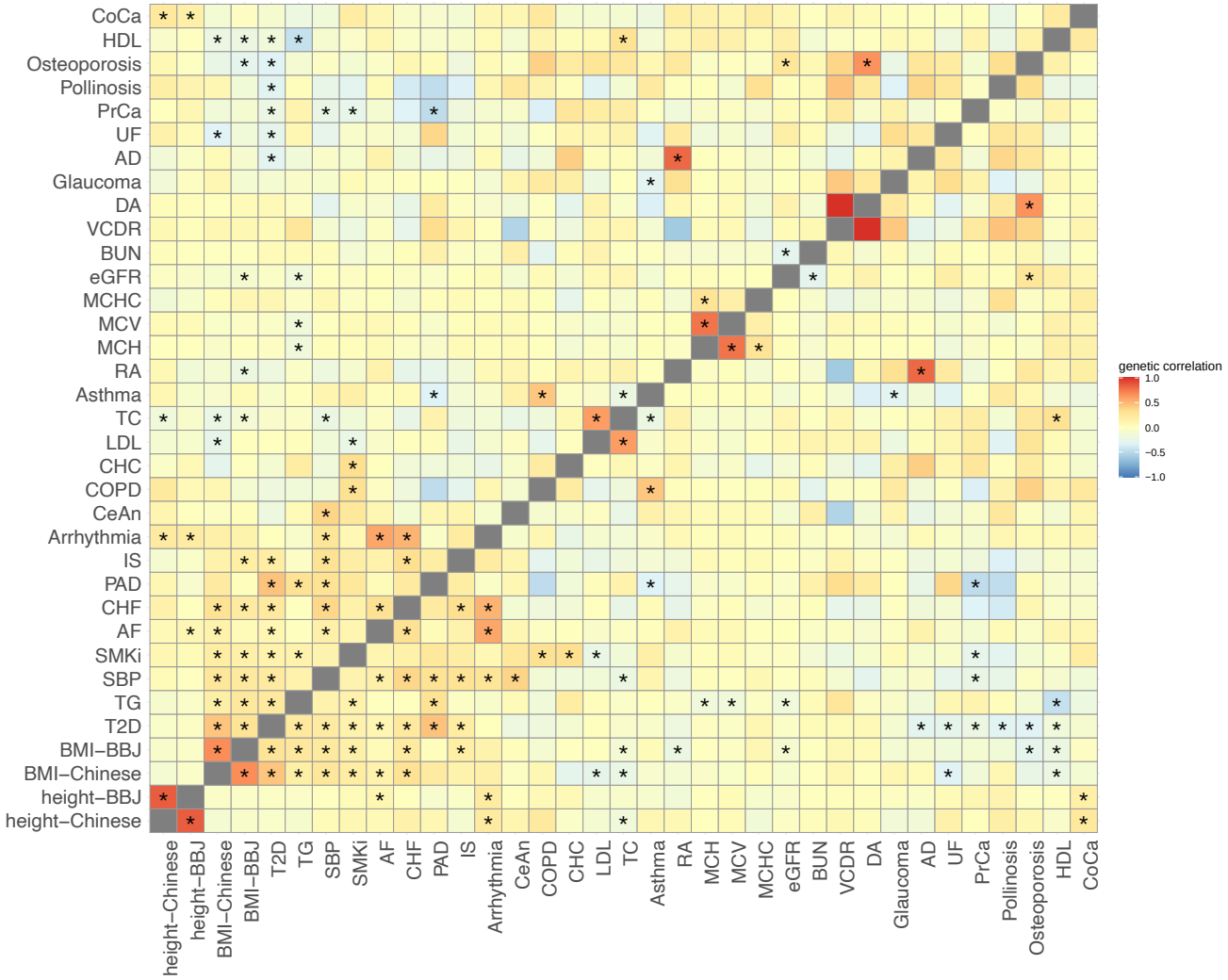

Figure S21: Genetic correlations of 35 traits in East Asians estimated by GNOVA. Positive correlations are colored in red. Negative correlations are colored in blue. Genetic correlations that are significantly different from zero after Bonferroni correction for the $(35 \times 34/2) = 595$ tests (p-value$< 0.05/595$) are marked with asterisk.

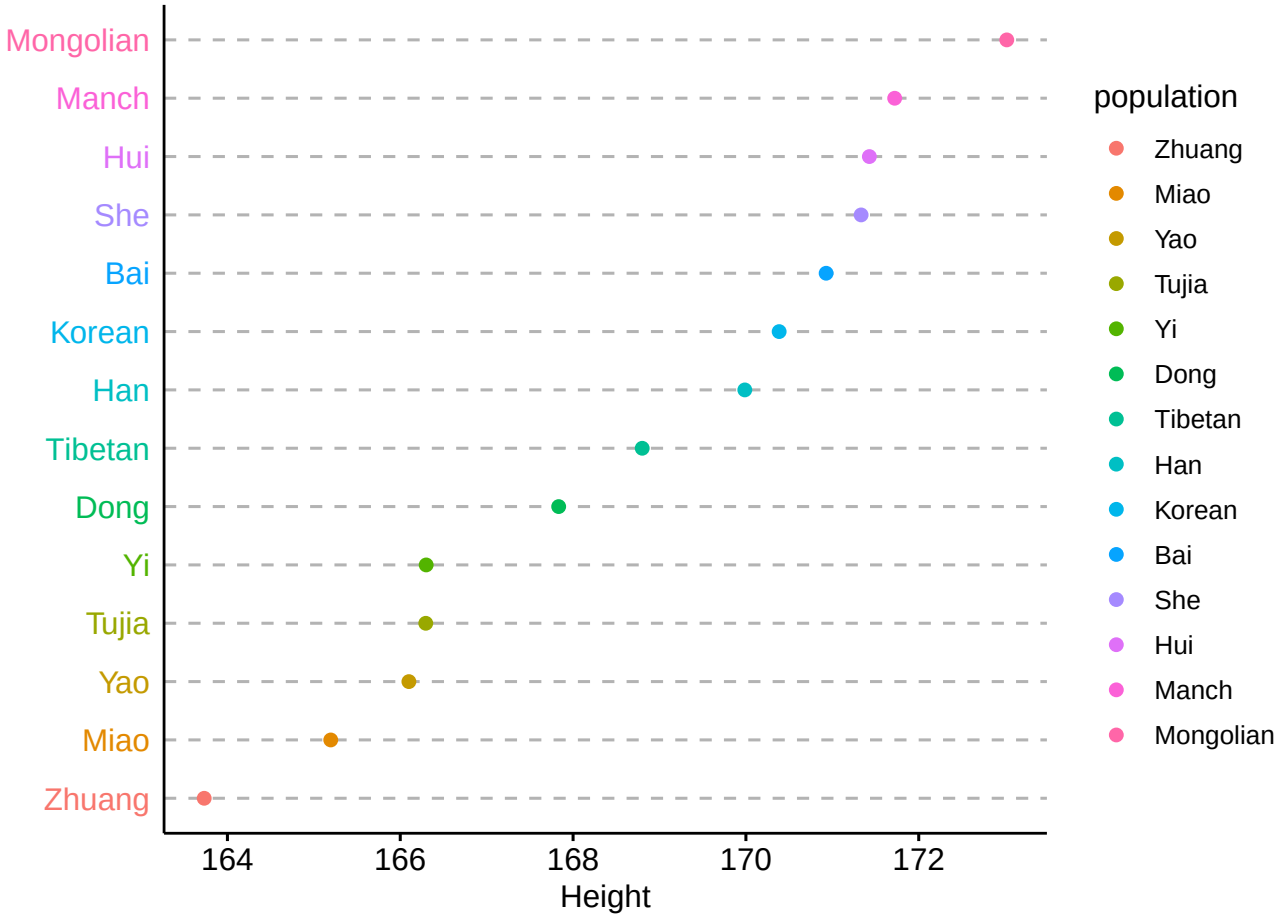

Figure S22: The average height among minority ethnic groups with ≥ 50 samples.

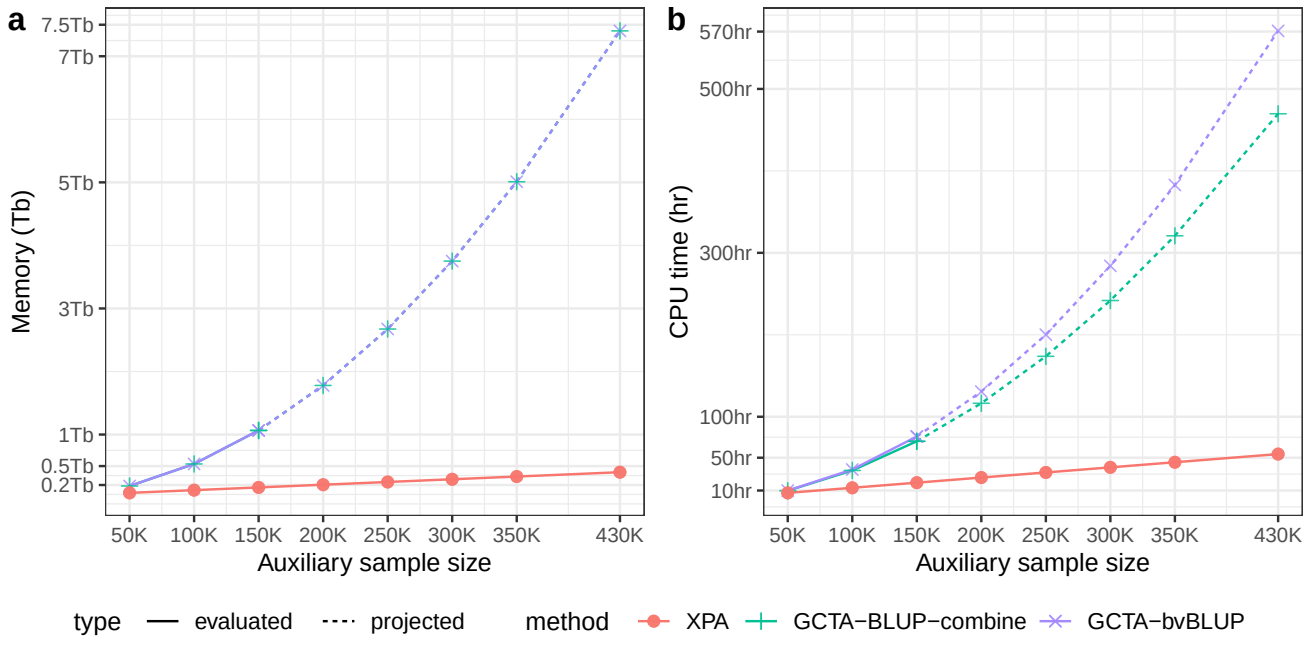

Figure S23: Memory usage (a) and CPU time (b) for XPA, GCTA-BLUP-combine, and GCTA-bvBLUP are shown for increasing auxiliary sample sizes when combining Chinese cohort and UKBB data to construct PRS for height. XPA used only 54.5 hours (including 9 hours for loading data, 3 hours for estimating variance components, and 42.5 hours for computing the posterior means and estimating fixed effects) and 385Gb to analyze all 430K Chinese and UKBB samples. In contrast, GCTA-bvBLUP required 1.07Tb when only 150K UKBB samples were included in the analysis, reaching the memory limit of our server. We note that the memory requirement exceeding this value is also infeasible for most high performance computational platforms. Therefore, we projected its CPU time and memory by fitting a quadratic curve using the recorded values. Our projection suggests that it would cost 570.8 hours and 7.5 Tb memory for GCTA-bvBLUP to integrate all 430K UKBB samples. Note that the memory of a node is often about 512Gb at Yale high-performance computing server, and the maximum memory of a node at the Hong Kong University of Science and Technology is about 1.5 Tb. Given above observations, we believe that XPA has advantage over GCTA-bvBLUP in practice as it can leverage the bio-bank scale dataset from the European population to construct more accurate PRS in minor populations. Both computational time and memory cost of XPA were linear to the auxiliary sample size, which was consistent with our observations in the simulation study. We evaluated all approaches with 32 CPU threads on the platform of Intel Xeon Gold 6152 CPU.

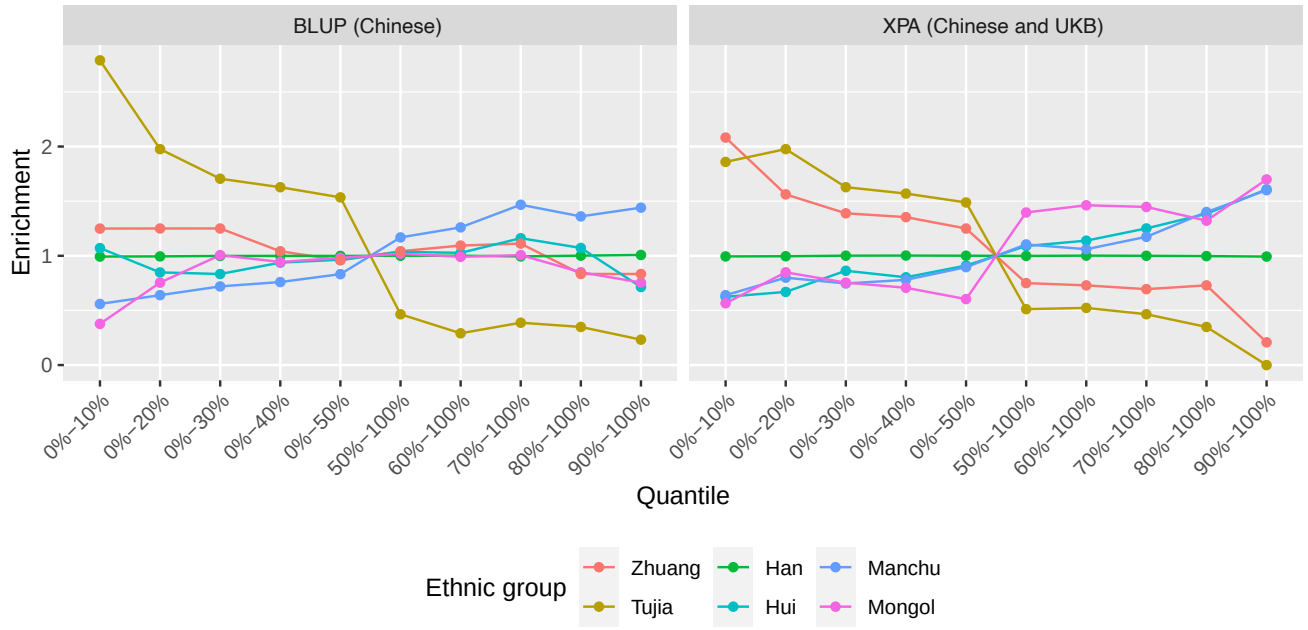

Figure S24: Proportion enrichment of ethnic groups in the top and bottom PRS quantiles. XPA successfully prioritizes the heights of the five minor ethnic groups with more than 50 samples in the test set, whilst BLUP can only predict Tujia and Manchu.

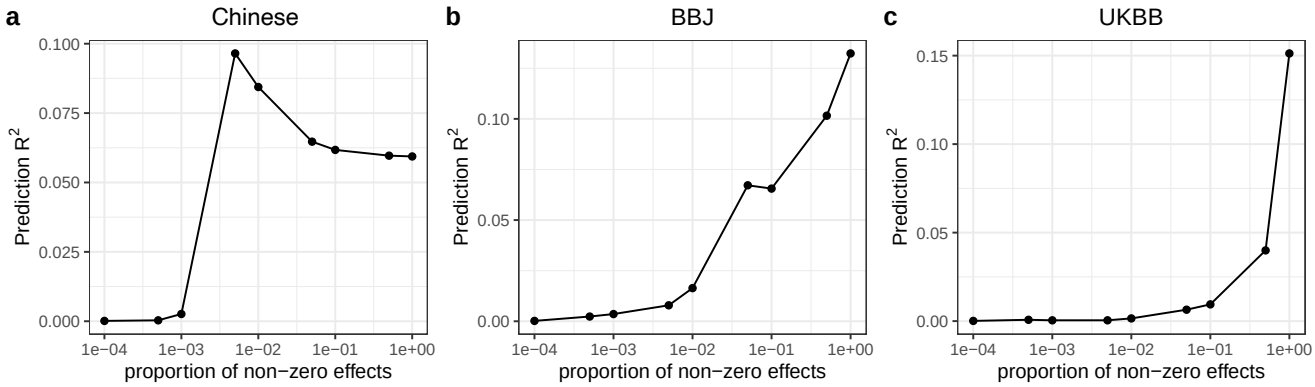

Figure S25: Tuning the proportion of non-zero effects in LDpred for height: (a) Chinese, (b) BBJ and (c) UKBB.

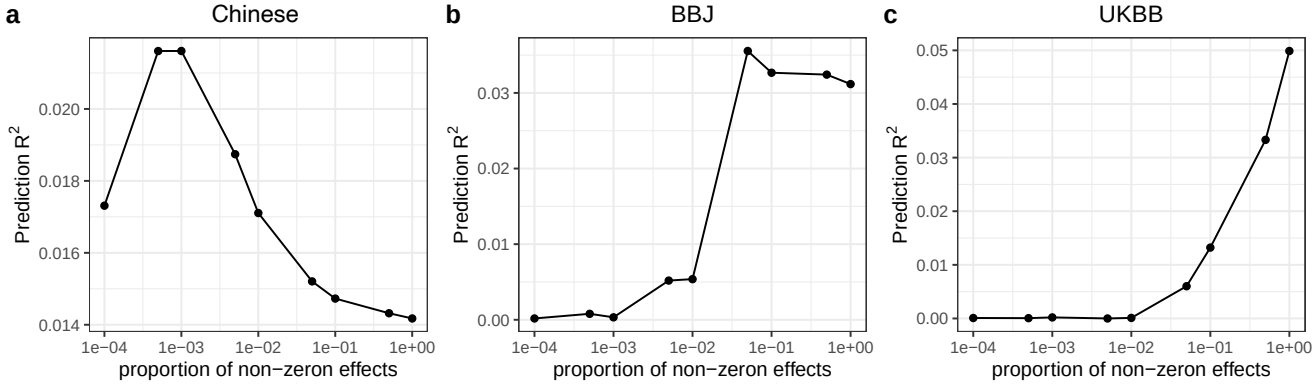

Figure S26: Tuning the proportion of non-zero effects in LDpred for BMI: (a) Chinese, (b) BBJ and (c) UKBB.

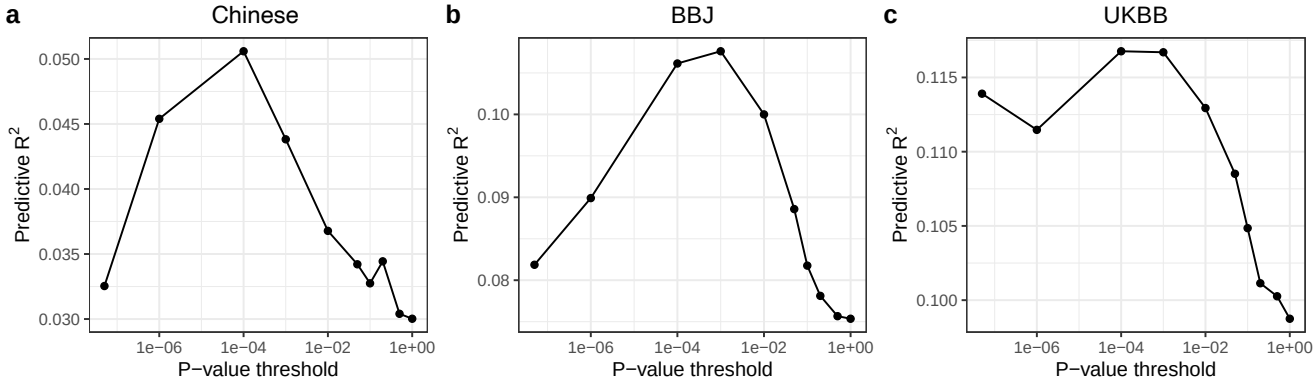

Figure S27: Tuning the p-value threshold for height: (a) Chinese, (b) BBJ and (c) UKBB.

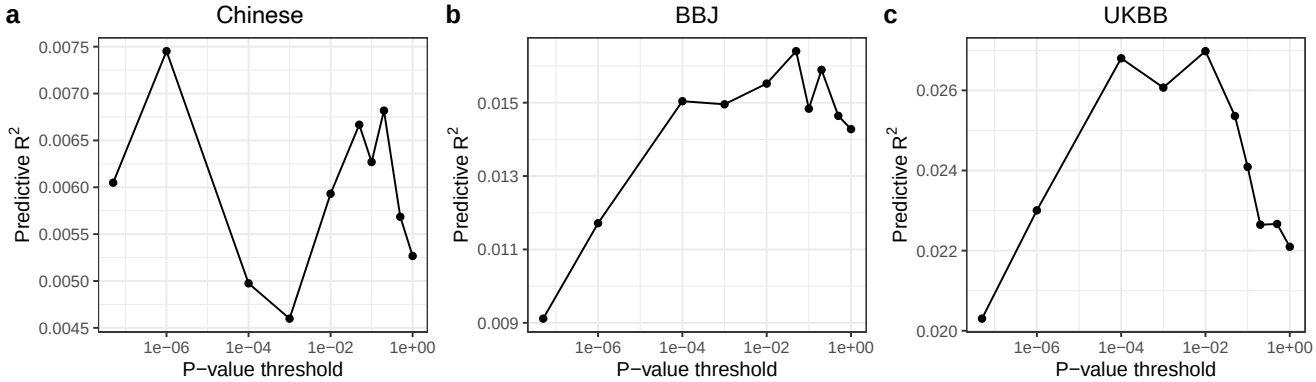

Figure S28: Tuning the p-value threshold for BMI: (a) Chinese, (b) BBJ and (c) UKBB.

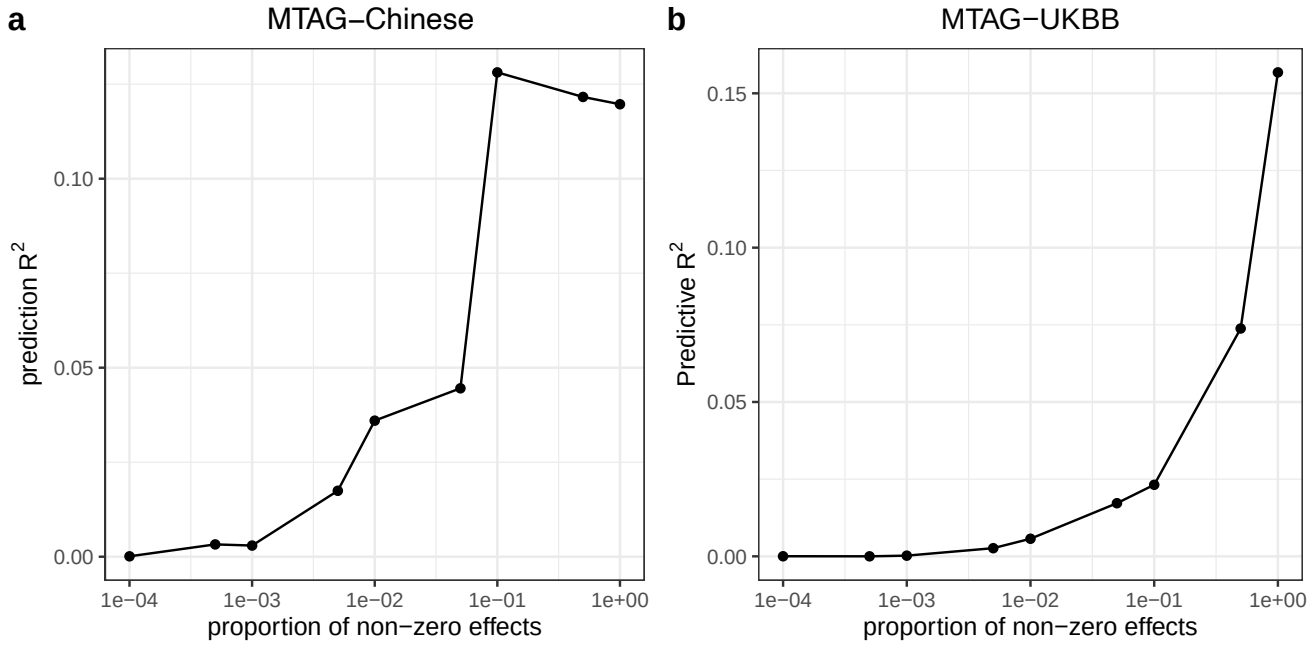

Figure S29: Tuning the proportion of non-zero effects in LDpred for height: (a) MTAG-Chinese, (b) MTAG-UKBB.

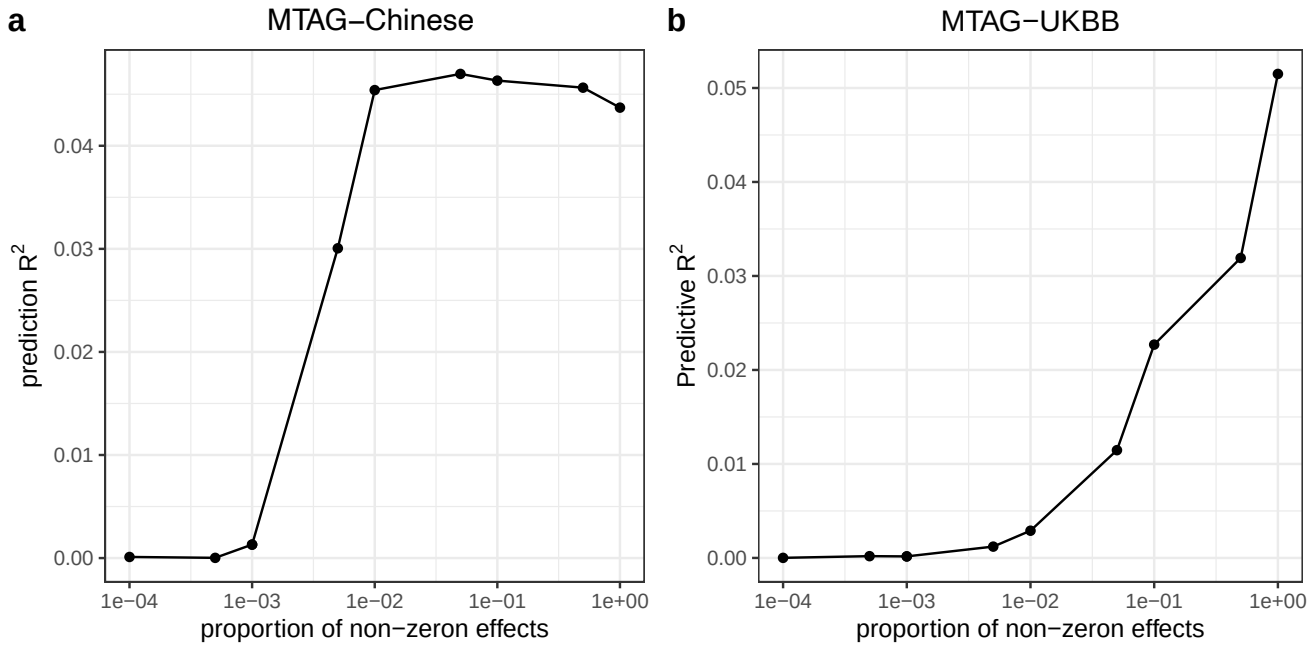

Figure S30: Tuning the proportion of non-zero effects in LDpred for BMI: (a) MTAG-Chinese, (b) MTAG-UKBB.

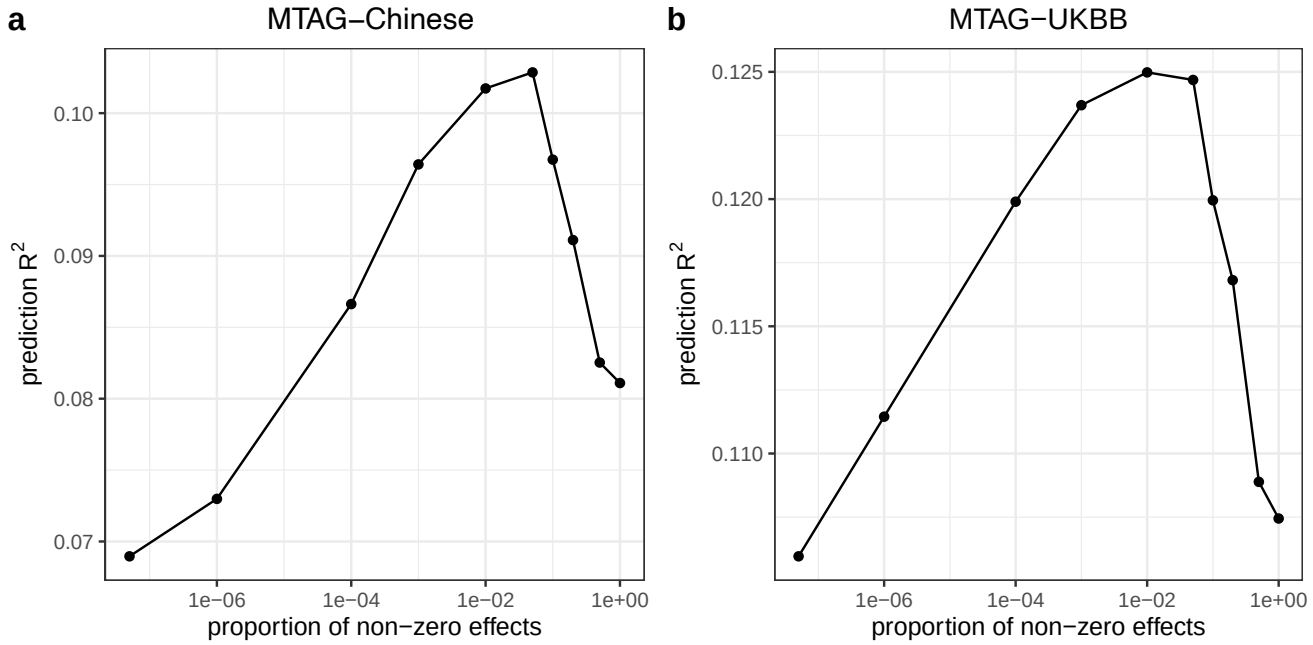

Figure S31: Tuning the p-value threshold for height: (a) MTAG-Chinese, (b) MTAG-UKBB.

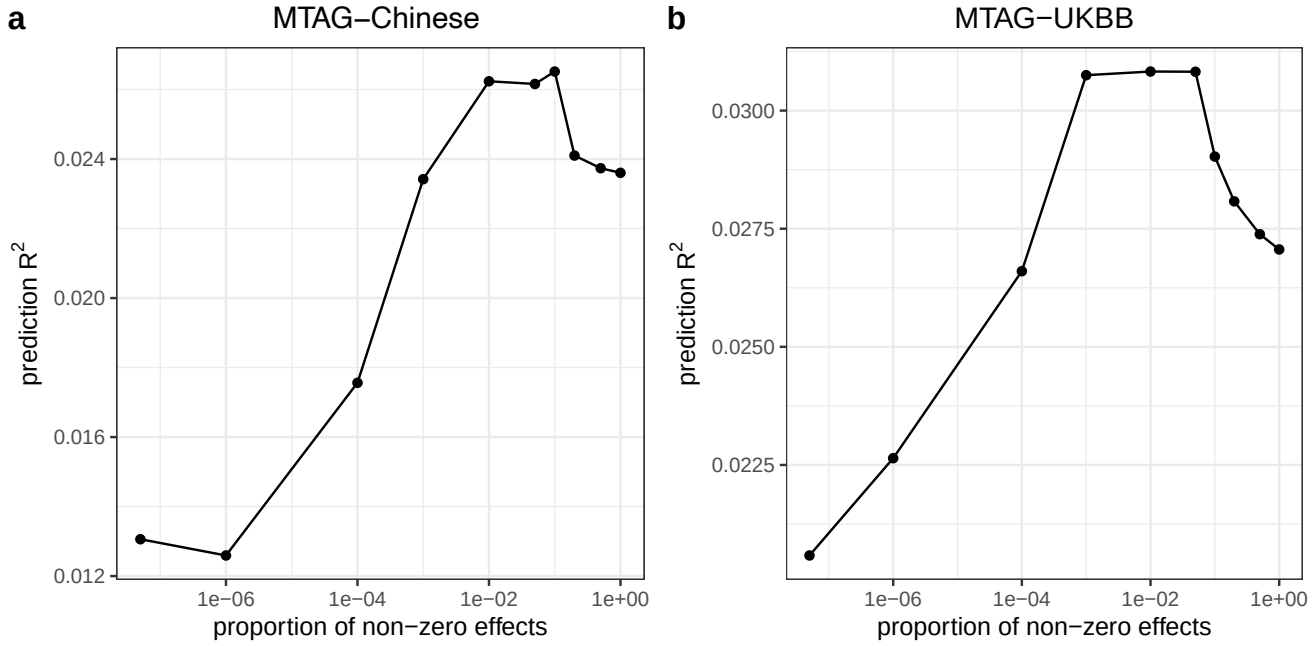

Figure S32: Tuning the p-value threshold for BMI: (a) MTAG-Chinese, (b) MTAG-UKBB.

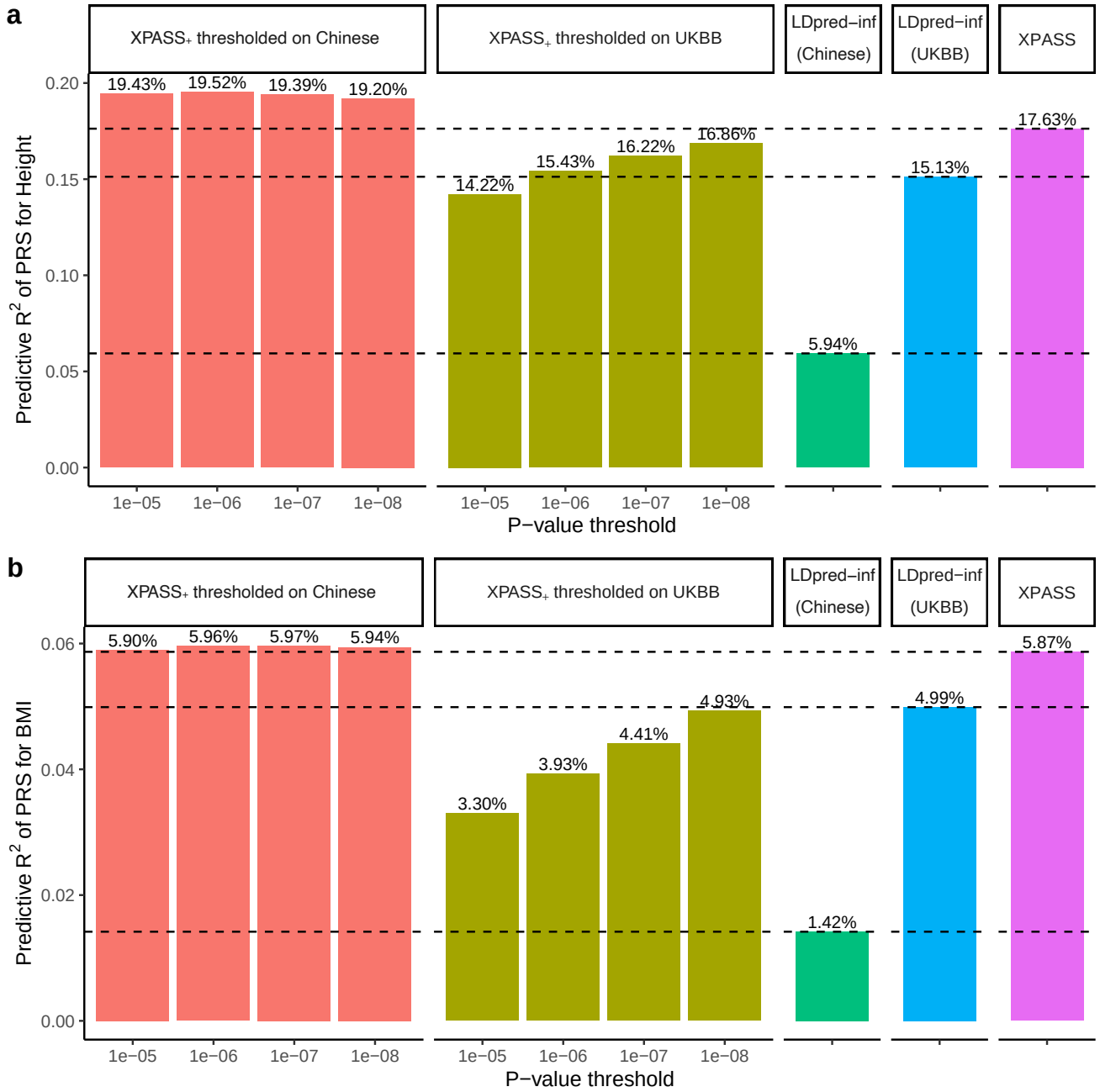

Figure S33: Predictive R^2 of height (a) and BMI (b) when XPASS$_+$ was applied with different p-value thresholds. The SNPs to be included in the covariates were selected by applying the p-value threshold to either the Chinese data or the UKBB data. The LD threshold was set as r2 = 0.1. The predictive R^2 of LDpred-inf (Chinese), LDpred-inf (UKBB) and XPASS were also shown as reference. When the P+T procedure was applied to the target dataset, including the selected SNPs as fixed effects further improved the prediction accuracy. In contrast, when the P+T procedure was applied to the auxiliary dataset, the predictive R^2 decreased as the p-value threshold increases. This observation suggests that when the pre-selected SNPs are specific to the target population, XPASS$_+$ can effectively utilize these signals to improve prediction accuracy.

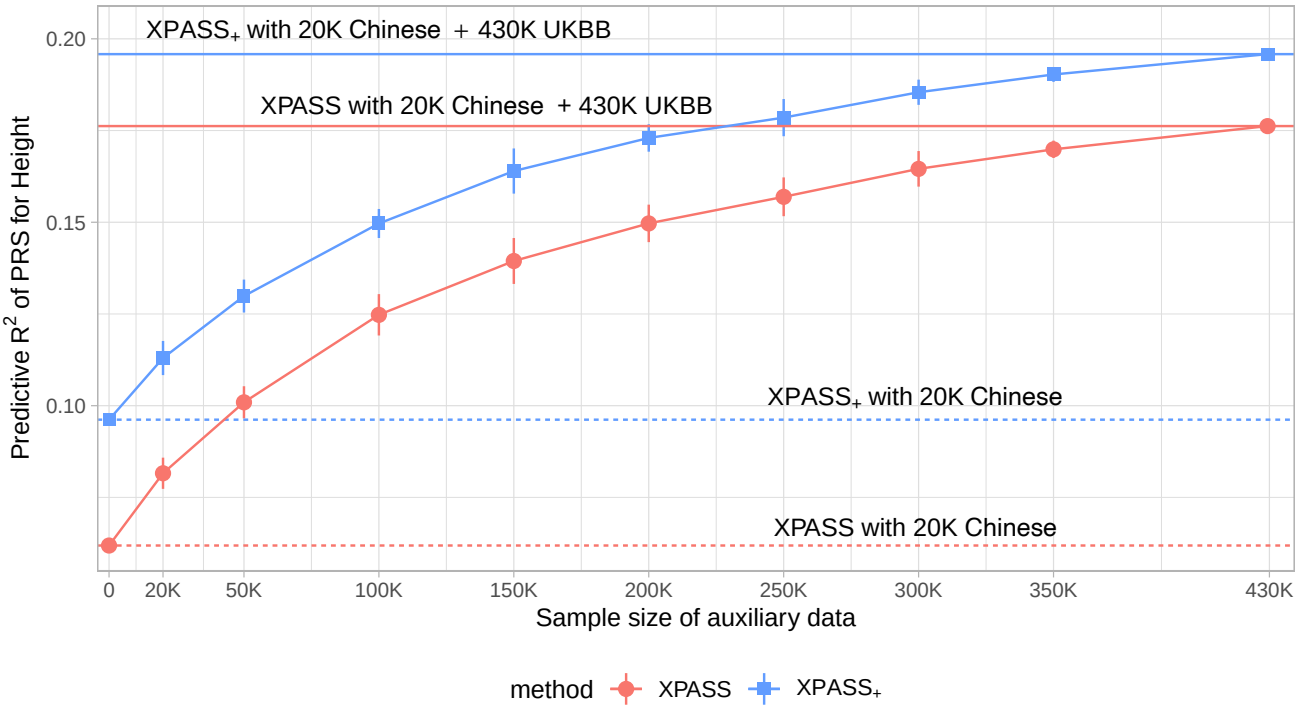

Figure S34: Influence of the auxiliary sample size on the prediction performance of XPASS and XPASS$_+$ for predicting height. We trained XPASS and XPASS$_+$ by integrating $21,069$ Chinese training samples with $20,000 \sim 300,000$ random subsamples drawn from UKBB, where samples from UKBB could be viewed as the auxiliary dataset. The results are summarized from 10 replications. Dashed horizontal lines mark the results obtained by training with only Chinese cohort using XPASS (red) or XPASS$_+$ (blue). Solid horizontal lines in mark the results obtained by combining 20K Chinese samples and all 430K UKBB samples using XPASS (red) or XPASS$_+$ (blue).

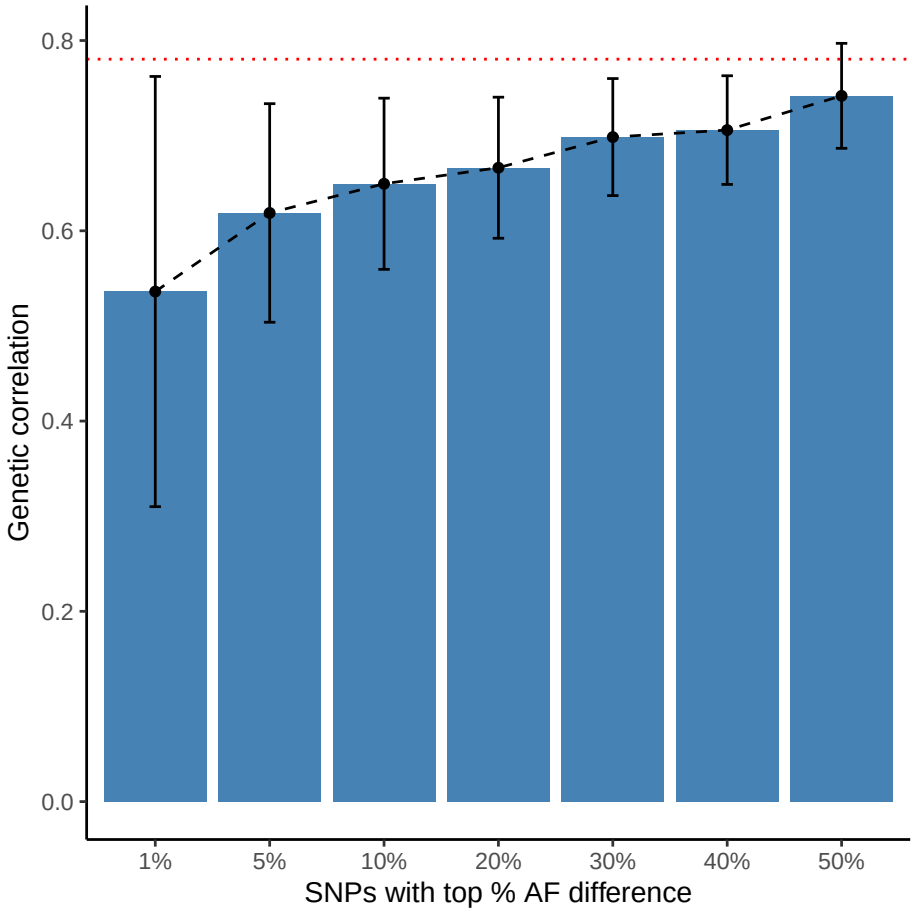

Figure S35: Genetic correlation of height estimated using the top 1% $\sim$ 50% SNPs sorted by the frequency (AF) difference between EAS and EUR. The AF difference is measured by $\frac{|f_1 - f2|}{\sqrt{2(f_1(1-f_1))+2(f_2(1-f_2))}}$. The red dotted line represents the genetic correlation estimated using all SNPs.

2 Supplementary Tables

		1%	5%	10%
Height	h^2_{1A}	0.747% (0.408%)	2.103% (0.721%)	4.244% (1.086%)
	Enrichment	2.135 (1.147)	1.205 (0.405)	1.213 (0.295)
	Predictive R^2	17.18%	16.87%	16.63%
BMI	h^2_{1A}	-0.301%	0.370% (0.466%)	1.395% (0.700%)
	Enrichment	-	0.451 (0.566)	0.845 (0.414)
	Predictive R^2	5.86%	5.68%	5.62%

Table S1: Application of the extended model to the Chinese and UKBB data. Estimated heritability and the enrichment of heritability explained by the 'heterogeneous' SNPs in Chinese samples are summarized in the table, with the corresponding standard errors given in the parentheses. The heritability explained by the 'heterogeneous' SNPs is computed as $h^2_{1A} = \hat{\sigma}^2_{1A}/(\hat{\sigma}^2_{1A} + \hat{\sigma}^2_{1B} + \hat{\sigma}^2_{\epsilon})$, and its enrichment is obtained as $(\hat{\sigma}^2_{1A}/p_A)/((\hat{\sigma}^2_{1A} + \hat{\sigma}^2_{1B})/p)$, where p_A and p are the number of SNPs in $\mathcal{A}$ and the total number of SNPs, respectively. The standard errors are obtained by applying the Jackknife approach with approximately independent LD blocks derived from the EAS population. Top 1%, 5% and 10% SNPs with highest diff_j were considered as 'heterogeneous' SNPs. The predictive R^2 were also computed for corresponding partition strategies.

Trait name	Full name	sample size (case+control)	paper link
RA-EAS	Rheumatoid Arthritis	4,873+17,642	https://www.nature.com/articles/nature12873?message-global=remove
RA-EUR	Rheumatoid Arthritis	14,361+43,923	https://www.nature.com/articles/nature12873?message-global=remove
T2D-EAS	Type 2 Diabetes	36,614+155,150	http://jenger.riken.jp:8080/pheno/Type_2_Diabetes
T2D-EUR	Type 2 Diabetes	459324	https://www.nature.com/articles/s41588-018-0144-6
BMI-BBJ	Body Mass Index	158284	https://www.nature.com/articles/ng.3951
BMI-Chinese	Body Mass Index	29147	http://www.sribd.cn/index.php/cn/%E7%A7%91%E5%AD%A6%E7%A0%94%E7%A9%B6/%E8%BD%AF%E4%BB%B6%E4%B8%8E%E6%95%B0%E6%8D%AE.html?layout=edit&id=322
BMI-UKB	Body Mass Index	457824	https://www.nature.com/articles/s41588-018-0144-6
BMI-GIANT	Body Mass Index	485,648∼795,640	https://academic.oup.com/hmg/article/27/20/3641/5067845
height-BBJ	height	159095	https://www.nature.com/articles/s41467-019-12276-5#Ack1
height-Chinese	height	32921	http://www.sribd.cn/index.php/cn/%E7%A7%91%E5%AD%A6%E7%A0%94%E7%A9%B6/%E8%BD%AF%E4%BB%B6%E4%B8%8E%E6%95%B0%E6%8D%AE.html?layout=edit&id=322
height-UKB	height	458303	https://www.nature.com/articles/s41588-018-0144-6
height-Giant	height	50,003∼253,280	https://www.nature.com/articles/ng.3097
HDL-EAS	High-density-lipoprotein cholesterol	70657	https://www.nature.com/articles/s41588-018-0047-6
HDL-EUR	High-density-lipoprotein cholesterol	188577	https://www.nature.com/articles/ng.2797
LDL-EAS	Low-density-lipoprotein cholesterol	72866	https://www.nature.com/articles/s41588-018-0047-6
LDL-EUR	Low-density-lipoprotein cholesterol	188577	https://www.nature.com/articles/ng.2797
MCH-EAS	Mean corpuscular hemoglobin concentration	108054	https://www.nature.com/articles/s41588-018-0047-6
MCH-UKB1+BIL	Mean corpuscular hemoglobin	132224	https://www.cell.com/cell/references/S0092-8674(16)31463-5
MCHC-EAS	Mean corpuscular hemoglobin concentration	108728	https://www.nature.com/articles/s41588-018-0047-6
MCHC-UKB1+BIL	Mean corpuscular hemoglobin concentration	132586	https://www.cell.com/cell/references/S0092-8674(16)31463-5
MCV-EAS	Mean corpuscular volume	108256	https://www.nature.com/articles/s41588-018-0047-6
MCV-UKB1+BIL	Mean corpuscular volume	132353	https://www.cell.com/cell/references/S0092-8674(16)31463-5
SysBP-EAS	Systolic blood pressure	136597	https://www.nature.com/articles/s41588-018-0047-6
SysBP-EUR	Systolic blood pressure	422771	https://www.nature.com/articles/s41588-018-0144-6
TC -EAS	Total cholesterol	128305	https://www.nature.com/articles/s41588-018-0047-6
TC-EUR	Total cholesterol	188577	https://www.nature.com/articles/ng.2797
TG-EAS	Triglyceride	105597	https://www.nature.com/articles/s41588-018-0047-6
TG-EUR	Triglyceride	188577	https://www.nature.com/articles/ng.2797
eGFR-EAS	Estimated glomerular filtration rate	143658	https://www.nature.com/articles/s41588-018-0047-6
eGFR-EUR	Estimated glomerular filtration rate	480698	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6377354/
BUN-EAS	Blood urea nitrogen	139818	https://www.nature.com/articles/s41588-018-0047-6
BUN-EUR	Blood urea nitrogen	480698	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6377354/
AF-BBJ	Atrial Fibrillation	8,180+28,612	https://www.nature.com/articles/ng.3842
AF-UKB	Atrial Fibrillation	10,986+233,904	https://www.gov.uk/government/publications/atrial-fibrillation-prevalence-estimates-for-local-populations
VCDR-EAS	vertical cup-disc ratio	8373	https://www.nature.com/articles/s41467-017-01913-6
VCDR-EUR	vertical cup-disc ratio	23899	https://www.nature.com/articles/s41467-017-01913-6
DA-EAS	Disc Area	7307	https://www.nature.com/articles/s41467-017-01913-6
DA-EUR	Disc Area	22504	https://www.nature.com/articles/s41467-017-01913-6
SMKi-BBJ	Smoking Initiation	83,810+81,626	https://www.nature.com/articles/s41562-019-0557-y
SMK-UKB	Current tobacco smoking	386150	https://www.nature.com/articles/s41588-019-0481-0
SMKcpd-UKB	Cigarettes per day	90143	https://www.nature.com/articles/s41588-019-0481-0
Angina-UKB	Angina	12114+373585	https://www.nature.com/articles/s41588-019-0481-0
CIHD-UKB	Chronic ischemic heart disease	14456+286335	https://www.nature.com/articles/s41588-019-0481-0
CHD-finngen	Major coronary heart disease event	10157+351037	http://www.nealelab.is/uk-biobank/
AIF-UKB	Alcohol intake frequency	360726	http://www.nealelab.is/uk-biobank/
HBP-UKB	High blood pressure	103381+282318	https://www.nature.com/articles/s41588-019-0481-0
hayfever-UKB	Hayfever, allergic rhinitis or eczema	83407+277120	http://www.nealelab.is/uk-biobank/
Asthma-UKB	Asthma	41633+318894	http://www.nealelab.is/uk-biobank/
glaucoma-UKB	Glaucoma-ICD10(H40)	1715+359479	http://www.nealelab.is/uk-biobank/
HF-UKB	heart failure	6504+387652	https://doi.org/10.1161/CIRCULATIONAHA.118.035774
CoCa-UKB	malignant neoplasm of colon-IDC10(C18)	2226+358968	http://www.nealelab.is/uk-biobank/
COPD-UKB	Other chronic obstructive pulmonary disease-ICD90(J44)	1531+359663	http://www.nealelab.is/uk-biobank/
Osteoporosis-UKB	Osteoporosis	5736+355405	http://www.nealelab.is/uk-biobank/
PAD_UKB	Peripheral artery disease	1230+359964	http://www.nealelab.is/uk-biobank/
PrCa-finngen	Prostate cancer	6321+160699	http://www.nealelab.is/uk-biobank/
UF-UKB	Uterine fibroids	5514+188639	http://www.nealelab.is/uk-biobank/
AD-UKB	Atopic dermatitis	9321+351820	http://www.nealelab.is/uk-biobank/
Arrhythmia-BBJ	Arrhythmia	17861+194592	https://www.nature.com/articles/s41588-020-0640-3
Asthma-BBJ	Asthma	8216+201592	https://www.nature.com/articles/s41588-020-0640-3
Cataract-BBJ	Cataract	24622+187831	https://www.nature.com/articles/s41588-020-0640-3
CHC_BBJ	Chronic hepatitis C	5794+206659	https://www.nature.com/articles/s41588-020-0640-3
CHF_BBJ	Congestive heart failure	9413+203040	https://www.nature.com/articles/s41588-020-0640-3
CoCa_BBJ	Colorectal cancer	7062+195745	https://www.nature.com/articles/s41588-020-0640-3
COPD_BBJ	Chronic obstructive pulmonary disease	3315+201592	https://www.nature.com/articles/s41588-020-0640-3
Glaucoma_BBJ	Glaucoma	5761+206692	https://www.nature.com/articles/s41588-020-0640-3
IS_BBJ	Ischemic stroke	17671+192383	https://www.nature.com/articles/s41588-020-0640-3
Osteoporosis_BBJ	Osteoporosis	7788+204665	https://www.nature.com/articles/s41588-020-0640-3
PAD_BBJ	Peripheral artery disease	3593+208860	https://www.nature.com/articles/s41588-020-0640-3
Pollinosis_BBJ	Pollinosis	5746+206707	https://www.nature.com/articles/s41588-020-0640-3
PrCa_BBJ	Prostate cancer	5408+103939	https://www.nature.com/articles/s41588-020-0640-3
UF_BBJ	Uterine fibroids	5954+95010	https://www.nature.com/articles/s41588-020-0640-3
AD_BBJ	Atopic dermatitis	2385+209651	https://www.nature.com/articles/s41588-020-0640-3

Table S2: Sources of 37 traits from EUR and 35 traits from EAS.

3 Supplementary Note

3.1 Sample quality control of Chinese cohort

We first removed non-Chinese and individuals without height records. We also excluded the related individuals with genetic relatedness exceeding 0.025 to ensure that heritability estimation and PRS construction are not influenced by related individuals. Only individuals with reported age between 16 and 70, and height between 130 cm and 220 cm were retained. Individuals with the genotyping rate less than 5% were also removed. Next, we excluded SNPs with one or more of the following properties: minor allele frequency less than 1%; missing genotypes in more than 5% of the samples; Hardy-Weinberg equilibrium (HWE) p-value below 0.0001. Finally, we took the overlap of SNPs between the Chinese dataset and the UKBB dataset. After these steps, we had $32,921$ individuals with $3,776,575$ SNPs for GWAS and the individual-level PRS analysis. We computed the genetic relatedness matrix (GRM) based on genome-wide genotype data, and then performed a randomized approximation of principal component analysis using plink v2.00 to extract the first 10 principal components for GWAS and cross-population analysis.

For BMI, we further removed individuals with extreme BMI values (larger than 38 or less than 10). This step results in $29,147$ participants with $3,777,871$ SNPs for GWAS and the individual-level PRS analysis. We conducted approximated PCA using plink v2.00 on these genotypes and used the first 10 principal components in data analysis.

3.2 Sample quality control of UKBB data

The full UKBB data were downloaded from `https://www.ukbiobank.ac.uk`. We first extracted the European whites who have reported their height and age. Then the relatives were removed by a genetic relatedness threshold 0.025. Only the individuals with reported height between 130 cm and 220 cm were retained. Individuals with genotyping rate less than 5% were also removed. SNPs were removed if at least one of the following is satisfied: minor allele frequency less than 1%; missing genotypes in more than 5% of the samples; Hardy-Weinberg equilibrium (HWE) p-value below 0.0001. Finally, we took the overlap of SNPs between the Chinese dataset and the UKBB dataset. At the end, we had $429,312$ individuals with $3,776,575$ SNPs for GWAS and the individual-level PRS analysis. Using plink v2.00, the approximate PCA was carried out on these genotypes and the first 20 principal components were included as covariates for data analysis.

For BMI, the same QC steps were applied, resulting in $428,846$ samples with $3,777,871$ SNPs for GWAS and the individual-level PRS analysis. We conducted approximate PCA using plink v2.00 on these genotypes and used the first 20 principal components for data analysis.

3.3 Sample quality control of IPM data

To obtain the ancestries of samples from IPM, we first projected their genotypes to the PC coordinates derived from the 1000 Genomes Project. The samples with the coordinate of the first PC > 0.09 and the coordinate of the second PC < -0.1 were identified as Africans, roughly corresponding to the boundary of African ancestry suggested by the AFR from the 1000 Genomes Project. We applied the same threshold to remove the ancestry outliers in the self-reported Africans

from the UKBB dataset. For both datasets, samples that have phenotypic values more than 4 standard errors away from the mean phenotypic values were identified as outliers and excluded from the analysis.

SNP-level (Rsq score$\geq$0.3) and genotype-per-participant-level (genotype probability$\geq$0.9) filters were used to exclude poorly imputed variants. Genotype QC was performed in PLINK V2.0 after excluding SNPs with a high missing call rate ($\geq$5%), a low minor allele frequency ($\leq$0.01) and deviation from Hardy-Weinberg equilibrium (p-value$\leq 1 \times 10^{-6}$). After phenotype and genotype quality control process (with details given in the Supplementary Note), we first merged two African datasets together, leading to 8,422 confirmed African samples with a total of 2,690,737 overlapping SNPs. Then we randomly selected 1K samples as testing data and used the remaining 7.4K samples as training data.

3.4 Height and BMI associations in Chinese population

To analyze the PRS performance in multi-ancestry datasets, we have collected more than 30k Chinese samples. Here, to study the genetic basis of height and BMI in Chinese population, we conducted GWAS to identify associations from 3.7 million SNPs in the Chinese population. Covariates including age, sex, and first 10 principal components were incorporated in the linear mixed model. Using LD score regression (LDSC) [1], we observed genomic control factor $\lambda_{gc} = 1.20$ and LDSC intercept= 1.026 with standard error (SE=0.014) for height, $\lambda_{gc} = 1.10$ and intercept= 0.998 with (SE=0.012) for BMI, respectively. Considering the polygenicity and the sample size, these statistics suggested no evidence of inflation in our GWAS analysis (Q-Q plots in Figure S10b and g, and Figure S11). After adjusting for the covariates, the residuals of both BMI and height show no correlation with either sexual or geometric factors, suggesting the confounding factors were well-controlled (Figure S8 and S9).

We used the BOLT-LMM v2.3.2 to test for associations between phenotypes and SNPs. We first identified the genome-wide significant SNPs using the p-value threshold 5×10^{-8}. Next, we conducted LD clumping on the significant SNPs using PLINK v2.0 with the LD threshold of 0.1 and clumping radius of one million base pairs. The nearly independent index SNPs were then annotated by the ANNOVAR software [2].

The GWAS identified 58 and 7 genome-wide significant loci (i.e., with leading SNP p-value$<$ 5×10^{-8}) for height and BMI, respectively (Figure S10a and f). Among the 58 height associated loci, 50 loci were previously known, and 36 of them were reported in EAS [3]. The eight novel loci include three intragenic ones (*TBX2-AS1*, *LOC101927932* and *GSDMC*), one located in the exonic area of gene *MIRLET7BHG* and six at the intergenic regions with nearby genes *SPAG17*, *PMCH*, *MIR296*, *TRIB1*, *CHCHD7* and *LOC100272217*. All the seven loci of BMI were previously reported and six of them were found in EAS [3].

To validate the associations identified from the Chinese data, we considered the summary-statistics datasets released from UKBB, the GIANT consortia [4, 5] and BBJ [6, 7] as validation. We compared the effect sizes of the genome-wide significant SNPs in our discovery study with those from the validation studies. For height associated SNPs (Figure S10c-e), all the effects in BBJ were in the same direction with Chinese cohort. In contrast, a number of SNP effects showed opposite directions between EAS and EUR. Besides, the slopes obtained by regressing the effect sizes of the Chinese data on those from the other studies were higher for EAS than for EUR (1.07 for BBJ

compared with 0.65 for UKBB and 0.83 for GIANT), suggesting a more similar genetic architecture within the EAS population and attenuated sharing of genetic basis between EAS and EUR. For BMI (Figure S10h-j), the effect sizes were consistent in directions across all studies, with similar slopes in regression analysis for all non-Chinese populations (0.54 for UKBB, 0.56 for GIANT and 0.52 for BBJ).

By partitioning the genome by chromosomes, we found the heritability of height explained by a chromosome was largely proportional to the chromosome length (Figure S12), consistent with previous studies conducted in EUR [4]. We further conducted a heritability enrichment analysis using the baseline model in the stratified LDSC. We found that all the significantly enriched functional regions in EAS are also enriched in Europeans (Figure S13 and S14). By subsampling the UKBB to the same sample size with Chinese, the enrichment patterns are very similar for the Chinese and UKBB datasets (Figure S15 and S16). The comparative study of GWAS results suggest that the genetic architectures of height and BMI are largely overlapped between EAS and EUR.

3.5 PRS performance in different ethnic groups of the Chinese population

Because the Chinese population is comprised of individuals from various ethnic backgrounds (Supplementary Fig.S1), the PRS performance may also vary across ethnic groups. To study the behavior of PRS in different minority-ethnic groups, we computed the enrichment of each ethnic group in different PRS-defined groups as the ratio between the proportion of an ethnic group in each PRS quantiles to its proportion in the whole test set. The results from six ethnic groups with more than 50 samples in the testing dataset are summarized in Fig.S24. For all the PRS models, there is no enrichment for Han Chinese as the ratio is nearly one across the PRS range. For the PRS derived by BLUP using the Chinese data only, Tujia and Manchu were enriched in the bottom and top quantiles, respectively. This is consistent with the relative height of these two ethnic groups in the population (Supplementary Fig.S22). However, BLUP failed to stratify the other ethnic groups based on the Chinese training data. By incorporating the UKBB dataset in training, XPA not only stratified the Tujia and Manchu people, but also captured the enrichment of Mongols (the highest group) and Hui people (the third highest group) in the top quantile and Zhuang people (the shortest group) in the bottom quantile. These results suggest that the PRS derived by XPA can effectively stratify the subgroups in Chinese population, despite their different ethnic backgrounds.

3.6 Trans-ancestry genetic correlations estimated by XPASS

The success of XPA and XPASS relies on the robust estimate of trans-ancestry genetic correlation. In addition to risk prediction, the trans-ancestry genetic correlation has the value of representing the shared genetic basis between populations.

Here, we applied XPASS to estimate trans-ancestry genetic correlations for a wide spectrum of complex phenotypes, including complex traits/diseases as well as cellular and organismal phenotypes, to provide a global picture of genetic architecture shared between EAS and EUR. Our analysis includes 37 traits from EUR and 35 traits from EAS, where 28 of them are matched pairs (Figure

S17). We also estimated the pair-wise genetic correlations of the phenotypes within each population using GNOVA [8] (Figure S20 and S21). We used the individuals from the 1000 Genomes project as external reference panels. For Europeans, 417 independent samples with $1,313,833$ SNPs were used for constructing the reference panel. For East Asians, 337 independent samples with $1,209,411$ SNPs were used in analysis. Because the sets of variants vary across studies, we only considered the SNPs from the third phase of the International HapMap project phase 3 (HapMap3), resulting in $850,000$ SNPs on average included for estimating the genetic correlation after overlapping procedure. For XPASS, we included the first 5 and 20 principal components as covariates for EAS and EUR reference panels, respectively. The summary statistics of GWAS used in the analysis are summarized in Supplementary Table 1.

Out of the the 28 matched traits, XPASS identified 27 traits that are significantly correlated between the two populations (p-value $< 0.05/28$). Six traits, including type-2 diabetes (T2D), systolic blood pressure (SBP), low-density lipoprotein (LDL), mean corpuscular hemoglobin (MCH), Disc Area (DA) and Glaucoma, were highly correlated between EAS and EUR ($\rho \geq 0.9$). The estimated glomerular filtration rate (eGFR) had the lowest genetic correlation. We estimated the trans-ancestry correlation of height as 0.67 (SE=0.018) and BMI as 0.63 (SE=0.034), consistent with previous findings [9].

Among all 1,295 trans-ancestry pairs of traits, 171 were significantly correlated after Bonferroni correction (p-value $< 0.05/1521$), suggesting pervasive shared genetic basis between the two populations. In particular, multiple pairs of traits strongly correlated within EUR largely remain between EAS and EUR. Examples include positive genetic correlations between triglyceride levels (TG) and T2D, BMI and heart-related diseases, and BMI and smoking behaviors as well as negative genetic correlations between height and chronic ischemic heart disease (CIHD), high-density lipoprotein (HDL) and TG, and eGFR and BMI [10].

We compared the estimates generated by XPASS with those generated by popcorn [11] and summarized the results in Figure S19. We found that the estimated correlations were highly consistent between XPASS and popcorn. Besides, XPASS identified 164 pairs of significantly correlated traits in total, including all 81 significant correlations reported by popcorn.

3.7 Extended Variance Component Model for Accounting for Allele Frequency Difference

To assess the effect of allele frequency difference on the prediction accuracy, we extended the XPASS model to include an additional genetic component that captures the effects of SNPs with large allele frequency differences across populations. From our real data analysis, we did not observe significant enrichment of heritability among these SNPs. As a result, we did not obtain a better PRS by modeling the effect sizes of these SNPs as an additional variance component in the extended model.

We first partitioned the p SNPs into two disjoint sets according to the frequency difference $\text{diff}_j = \frac{|\mathbf{f}_{1,j} - \mathbf{f}_{2,j}|}{\sqrt{2\mathbf{f}_{1,j}(1-\mathbf{f}_{1,j}) + 2\mathbf{f}_{2,j}(1-\mathbf{f}_{2,j})}}$. The set $\mathcal{A}$ included all the 'heterogeneous' SNPs with large allele differences and the set $\mathcal{B}$ contains the remaining SNPs that are not in $\mathcal{A}$. Let $\mathbf{X}_1^A \in \mathbb{R}^{n_1 \times p_A}$ and $\mathbf{X}_2^A \in \mathbb{R}^{n_2 \times p_A}$ denote the standardized genotype matrices of 'heterogeneous' SNPs and $\mathbf{X}_1^B \in \mathbb{R}^{n_1 \times p_B}$ and $\mathbf{X}_2^B \in \mathbb{R}^{n_2 \times p_B}$ denote the standardized genotype matrices of the remaining SNPs for populations one and two, respectively. We related the phenotypes and genotypes using the extended linear

models:

$$\mathbf{y}_1 = \mathbf{Z}_1\boldsymbol{\omega}_1 + \mathbf{X}_1^A\boldsymbol{\beta}_1^A + \mathbf{X}_1^B\boldsymbol{\beta}_1^B + \boldsymbol{\epsilon},$$
$$\mathbf{y}_2 = \mathbf{Z}_2\boldsymbol{\omega}_2 + \mathbf{X}_2^A\boldsymbol{\beta}_2^A + \mathbf{X}_2^B\boldsymbol{\beta}_2^B + \boldsymbol{\xi},$$

where $\boldsymbol{\omega}_1 \in \mathbb{R}^{c_1}$ and $\boldsymbol{\omega}_2 \in \mathbb{R}^{c_2}$ are fixed effects of covariates, $\boldsymbol{\beta}_1^A = [\beta_{1,1}^A, \beta_{1,2}^A, \ldots, \beta_{1,p_A}^A]^T \in \mathbb{R}^{p_A}$ and $\boldsymbol{\beta}_2^A = [\beta_{2,1}^A, \beta_{2,2}^A, \ldots, \beta_{2,p_A}^A]^T \in \mathbb{R}^{p_A}$ are vectors collecting the effect sizes of the 'heterogeneous' SNPs from the two populations, $\boldsymbol{\beta}_1^B = [\beta_{1,1}^B, \beta_{1,2}^B, \ldots, \beta_{1,p_B}^B]^T \in \mathbb{R}^{p_B}$ and $\boldsymbol{\beta}_2^B = [\beta_{2,1}^B, \beta_{2,2}^B, \ldots, \beta_{2,p_B}^B]^T \in \mathbb{R}^{p_B}$ are vectors collecting the effect sizes of the remaining SNPs from the two populations, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2\mathbf{I}_{n_1})$ and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma_\xi^2\mathbf{I}_{n_2})$ are independent errors. We considered probabilistic structures to the SNPs in sets $\mathcal{A}$ and $\mathcal{B}$ as

$$\begin{pmatrix}\beta_{1,j}^A \\ \beta_{2,j}^A\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_{1A}^2 & \rho_A\sigma_{1A}\sigma_{2A} \\ \rho_A\sigma_{1A}\sigma_{2A} & \sigma_{2A}^2\end{pmatrix}\right), \quad j = 1, \ldots, p_A,$$
$$\begin{pmatrix}\beta_{1,j}^B \\ \beta_{2,j}^B\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_{1B}^2 & \rho_B\sigma_{1B}\sigma_{2B} \\ \rho_B\sigma_{1B}\sigma_{2B} & \sigma_{2B}^2\end{pmatrix}\right), \quad j = 1, \ldots, p_B,$$

respectively, where σ_{1A}^2 and σ_{2A}^2 are the variance components of the 'heterogeneous' SNP effects in the two populations, respectively, ρ_A is the trans-ancestry genetic correlation of the 'heterogeneous' SNP effects, σ_{1B}^2 and σ_{2B}^2 are the variance components of the remaining SNP effects in the two populations, respectively, and ρ_B is the trans-ancestry genetic correlation of the remaining SNP effects. With this flexible statistical structure of genetic effects, the variance and genetic correlation of 'heterogeneous' SNPs are allowed to be different from the remaining SNPs. We can estimate the parameters and obtain the posterior means of $\boldsymbol{\beta}^A$ and $\boldsymbol{\beta}^B$ using GWAS summary statistics in the similar way as in XPASS. To evaluate the impact of the 'heterogeneous' SNPs on prediction performance, we applied this extended model to the height and BMI datasets. We estimated the heritability explained by the two components, evaluated the enrichment of heritabilities of the 'heterogeneous' component, and constructed PRS from the extended model. Recall that the predictive R^2 of the original XPASS model is 17.63%. As summarized in Table S1, we observed neither significant enrichment of heritability in the 'heterogeneous' SNPs, nor improvement of prediction performance when these SNPs were introduced as an additional component in the model. Our results suggest that modeling the effect sizes of SNPs with large allele frequency difference may not be the key to improve PRS.

References

[1] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, Benjamin M Neale, Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295, 2015.

[2] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, 38(16):e164–e164, 2010.

[3] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2019.

[4] Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Zoltán Kutalik, Najaf Amin, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173–1186, 2014.

[5] Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015.

[6] Masato Akiyama, Yukinori Okada, Masahiro Kanai, Atsushi Takahashi, Yukihide Momozawa, Masashi Ikeda, Nakao Iwata, Shiro Ikegawa, Makoto Hirata, Koichi Matsuda, et al. Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nature genetics*, 49(10):1458–1467, 2017.

[7] Masato Akiyama, Kazuyoshi Ishigaki, Saori Sakaue, Yukihide Momozawa, Momoko Horikoshi, Makoto Hirata, Koichi Matsuda, Shiro Ikegawa, Atsushi Takahashi, Masahiro Kanai, et al. Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nature communications*, 10(1):1–11, 2019.

[8] Qiongshi Lu, Boyang Li, Derek Ou, Margret Erlendsdottir, Ryan L Powles, Tony Jiang, Yiming Hu, David Chang, Chentian Jin, Wei Dai, et al. A powerful approach to estimating annotation-stratified genetic covariance via GWAS summary statistics. *The American Journal of Human Genetics*, 101(6):939–964, 2017.

[9] Alicia R Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, 51(4):584–591, 2019.

[10] Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, Laramie Duncan, John RB Perry, Nick Patterson, Elise B Robinson, et al. An atlas of genetic correlations across human diseases and traits. *Nature genetics*, 47(11):1236–1241, 2015.

[11] Brielin C Brown, Chun Jimmie Ye, Alkes L Price, Noah Zaitlen, Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, et al. Transethnic genetic-correlation estimates from summary statistics. *The American Journal of Human Genetics*, 99(1):76–88, 2016.