

Supplementary Information for *Improving Polygenic Prediction in Ancestrally Diverse Populations*

Yunfeng Ruan^{1,2}, Yen-Feng Lin^{3,4,5}, Yen-Chen Anne Feng^{1,6,7,8,9}, Chia-Yen Chen¹⁰, Max Lam^{1,8,11,12,13},
Zhenglin Guo¹, Stanley Global Asia Initiatives, Lin He², Akira Sawa^{14,15}, Alicia R. Martin^{1,8},
Shengying Qin^{2,*}, Hailiang Huang^{1,8,16,*}, Tian Ge^{1,6,7,17,*}

1 Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA

2 Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Shanghai Jiao Tong University, Shanghai, China

3 Center for Neuropsychiatric Research, National Health Research Institutes, Miaoli, Taiwan

4 Department of Public Health & Medical Humanities, School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan

5 Institute of Behavioral Medicine, College of Medicine, National Cheng Kung University, Tainan, Taiwan

6 Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

7 Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

8 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA

9 Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei, Taiwan

10 Translational Biology, Biogen Inc., Cambridge, MA, USA

11 Division of Psychiatry Research, The Zucker Hillside Hospital, Northwell Health, Glen Oaks, NY, USA

12 Research Division, Institute of Mental Health Singapore, Singapore, Singapore

13 Human Genetics, Genome Institute of Singapore, Singapore, Singapore

14 Departments of Psychiatry, Neuroscience, and Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD, USA

15 Department of Mental Health, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, USA

16 Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

17 Center for Precision Psychiatry, Massachusetts General Hospital, Boston, MA, USA

*Email: chinsir@sjtu.edu.cn (S.Q.); hhuang@broadinstitute.org (H.H.); tge1@mgh.harvard.edu (T.G.)

SUPPLEMENTARY METHODS

PRS-CSx employs the following Bayesian high-dimensional regression framework for K populations:

$$\mathbf{y}_k = \mathbf{X}_k \boldsymbol{\beta}_k + \boldsymbol{\epsilon}_k, \quad \boldsymbol{\epsilon}_k \sim \text{MVN}(\mathbf{0}, \sigma_k^2 \mathbf{I}), \quad p(\sigma_k^2) \propto \sigma_k^{-2}, \quad k = 1, 2, \dots, K,$$

where, for each population k , \mathbf{y}_k is a vector of standardized phenotypes (zero mean and unit variance) from N_k individuals, \mathbf{X}_k is an $N_k \times M$ matrix of standardized genotypes (each column has zero mean and unit variance), $\boldsymbol{\beta}_k$ is a vector of SNP effect sizes, $\boldsymbol{\epsilon}_k$ is a vector of normally distributed non-genetic effects with variance σ_k^2 , and \mathbf{I} is an identity matrix. For SNP j in population k , a continuous shrinkage prior is placed on its effect size β_{jk} , which can be represented as global-local scale mixtures of normals:

$$\beta_{jk} \sim \text{N}\left(0, \frac{\sigma_k^2}{N_k} \psi_j\right), \quad \psi_j \sim G(a, \delta_j), \quad \delta_j \sim G(b, \phi),$$

where ϕ is a global shrinkage parameter shared across all SNPs that models the overall sparseness of the genetic architecture, and ψ_j is a local, SNP-specific shrinkage parameter that is adaptive to marginal GWAS associations. Note that both ϕ and ψ_j do not depend on k , and thus the continuous shrinkage prior is shared across populations.

The full conditional distributions for unknown model parameters are analytically tractable. Let $\text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$; $G(\zeta, \eta)$ and $iG(\zeta, \eta)$ denote the gamma distribution and inverse gamma distribution, respectively, with probability density functions

$$f_G(x; \zeta, \eta) = \frac{\eta^\zeta}{\Gamma(\zeta)} x^{\zeta-1} \exp(-\eta x), \quad f_{iG}(x; \zeta, \eta) = \frac{\eta^\zeta}{\Gamma(\zeta)} x^{-\zeta-1} \exp\left(-\frac{\eta}{x}\right), \quad x > 0, \quad \zeta > 0, \quad \eta > 0,$$

where $\Gamma(\cdot)$ is the gamma function; and $\text{giG}(\lambda, \rho, \chi)$ denote the three-parameter generalized inverse Gaussian distribution with density function

$$f_{\text{giG}}(x; \lambda, \rho, \chi) = \frac{(\rho/\chi)^{\lambda/2}}{2K_\lambda(\sqrt{\rho\chi})} x^{\lambda-1} \exp\left\{-\frac{1}{2}\left(\rho x + \frac{\chi}{x}\right)\right\}, \quad x > 0, \quad \rho > 0, \quad \chi > 0,$$

where K_λ is the modified Bessel function of the second kind. In addition, let $\widehat{\boldsymbol{\beta}}_k = \mathbf{X}_k^T \mathbf{y}_k / N_k$ denote the marginal least squares effect size estimates from the GWAS summary statistics for population k , $\mathbf{D}_k = \mathbf{X}_k^T \mathbf{X}_k / N_k$ denote the LD matrix for population k , $\boldsymbol{\Psi} = \text{diag}\{\psi_1, \psi_2, \dots, \psi_M\}$, where M is the total number of unique SNPs across populations, and k_j denote the number of populations in which SNP j is present. The Gibbs sampler for the PRS-CSx model involves the following steps in each Markov Chain Monte Carlo (MCMC) iteration:

- Update $\boldsymbol{\beta}_k$:

$$[\boldsymbol{\beta}_k \mid \sigma_k^2, \boldsymbol{\Psi}, \widehat{\boldsymbol{\beta}}_k, \mathbf{D}_k] \sim \text{MVN}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad \boldsymbol{\mu}_k = \frac{N_k}{\sigma_k^2} \boldsymbol{\Sigma}_k \widehat{\boldsymbol{\beta}}_k, \quad \boldsymbol{\Sigma}_k = \frac{\sigma_k^2}{N_k} (\mathbf{D}_k + \boldsymbol{\Psi}^{-1})^{-1},$$

- Update σ_k^2 :

$$[\sigma_k^2 \mid \boldsymbol{\beta}_k, \boldsymbol{\Psi}, \widehat{\boldsymbol{\beta}}_k, \mathbf{D}_k] \sim iG\left(\frac{N_k + M}{2}, \frac{N_k}{2} [1 - 2\widehat{\boldsymbol{\beta}}_k^T \boldsymbol{\beta}_k + \boldsymbol{\beta}_k^T (\mathbf{D}_k + \boldsymbol{\Psi}^{-1}) \boldsymbol{\beta}_k]\right),$$

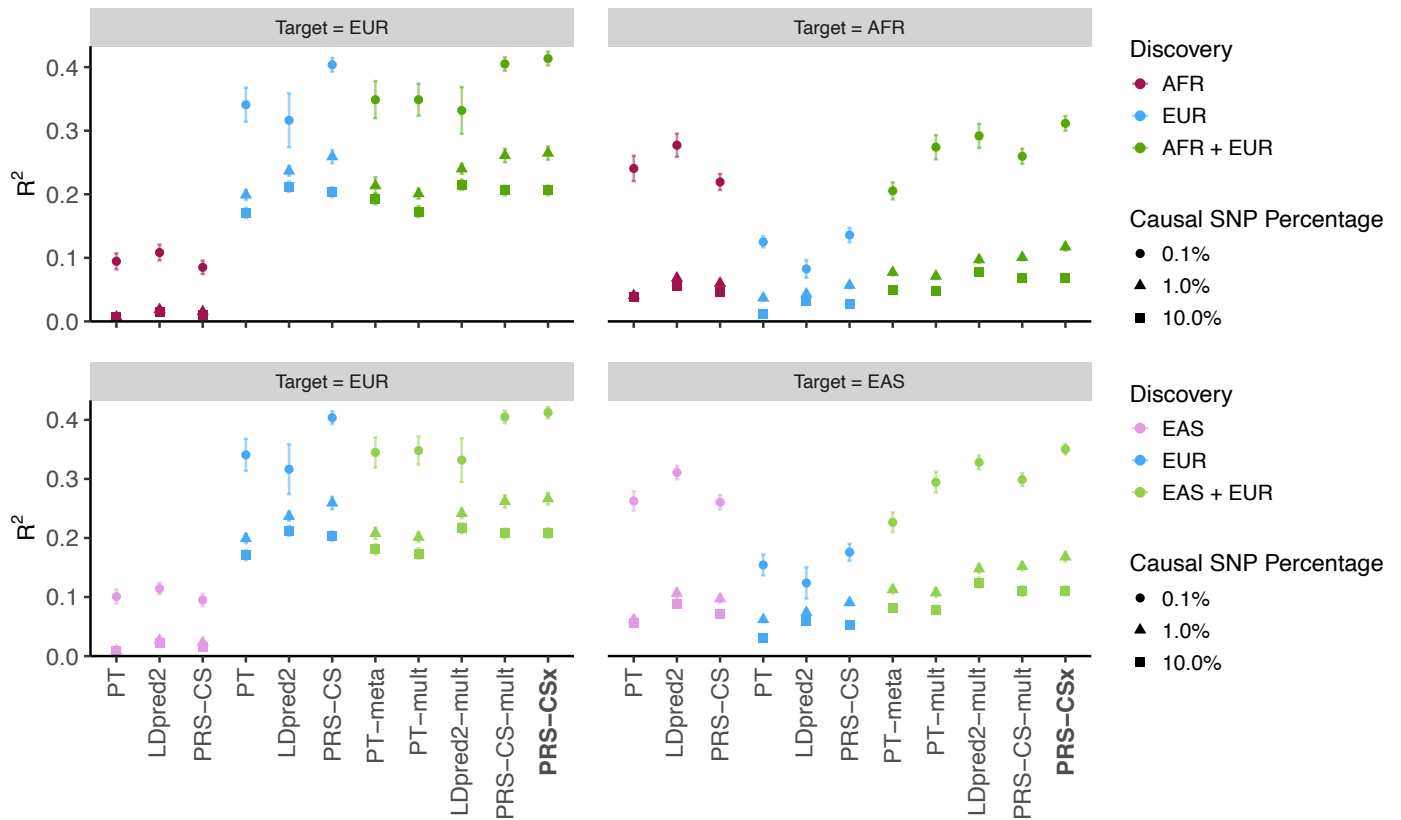
- Update ψ_j :

$$[\psi_j | \beta_{jk}, \sigma_k^2, \delta_j] \sim \text{giG}\left(a - \frac{k_j}{2}, 2\delta_j, \sum_k \frac{N_k}{\sigma_k^2} \beta_{jk}^2\right),$$

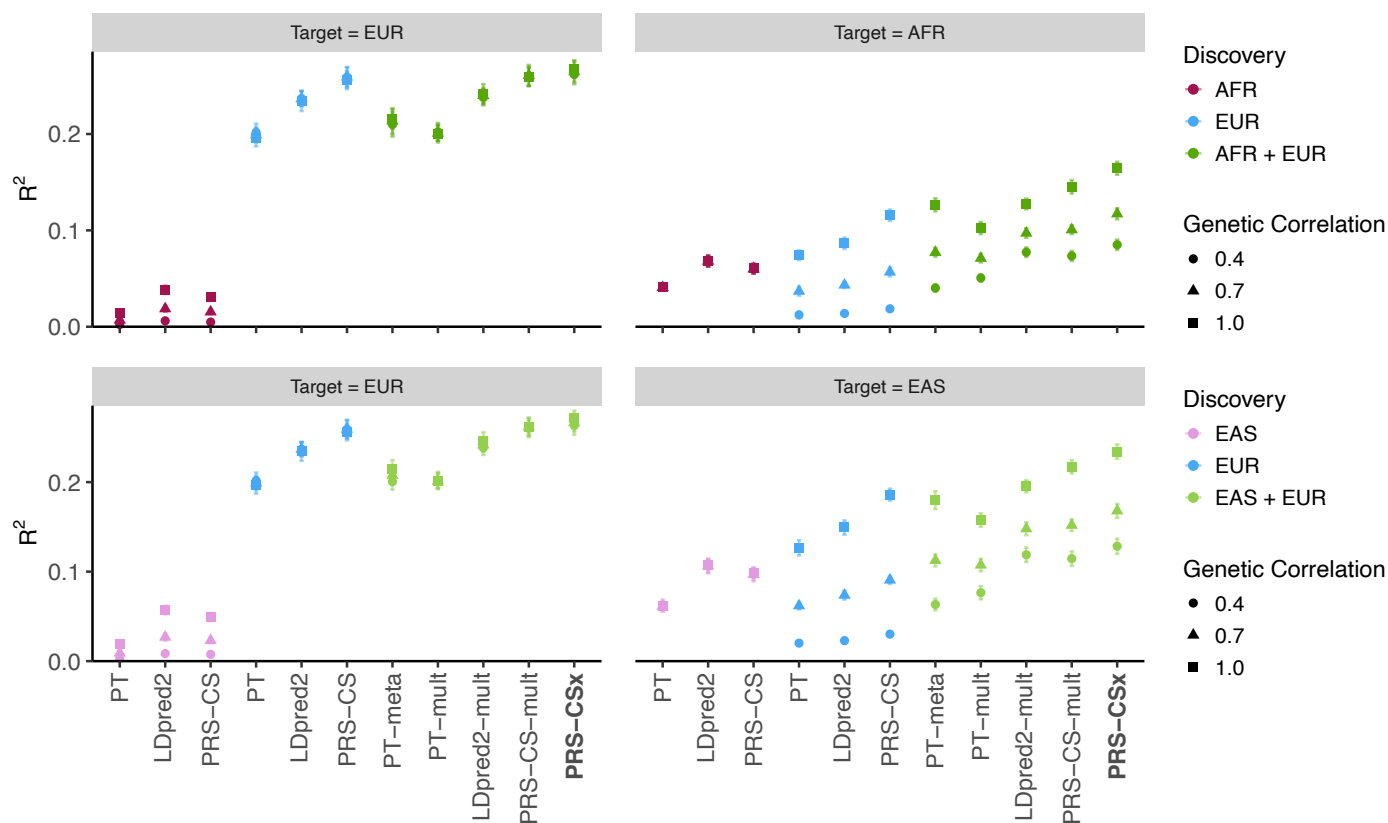
- Update δ_j :

$$[\delta_j | \psi_j] \sim G(a + b, \psi_j + \phi).$$

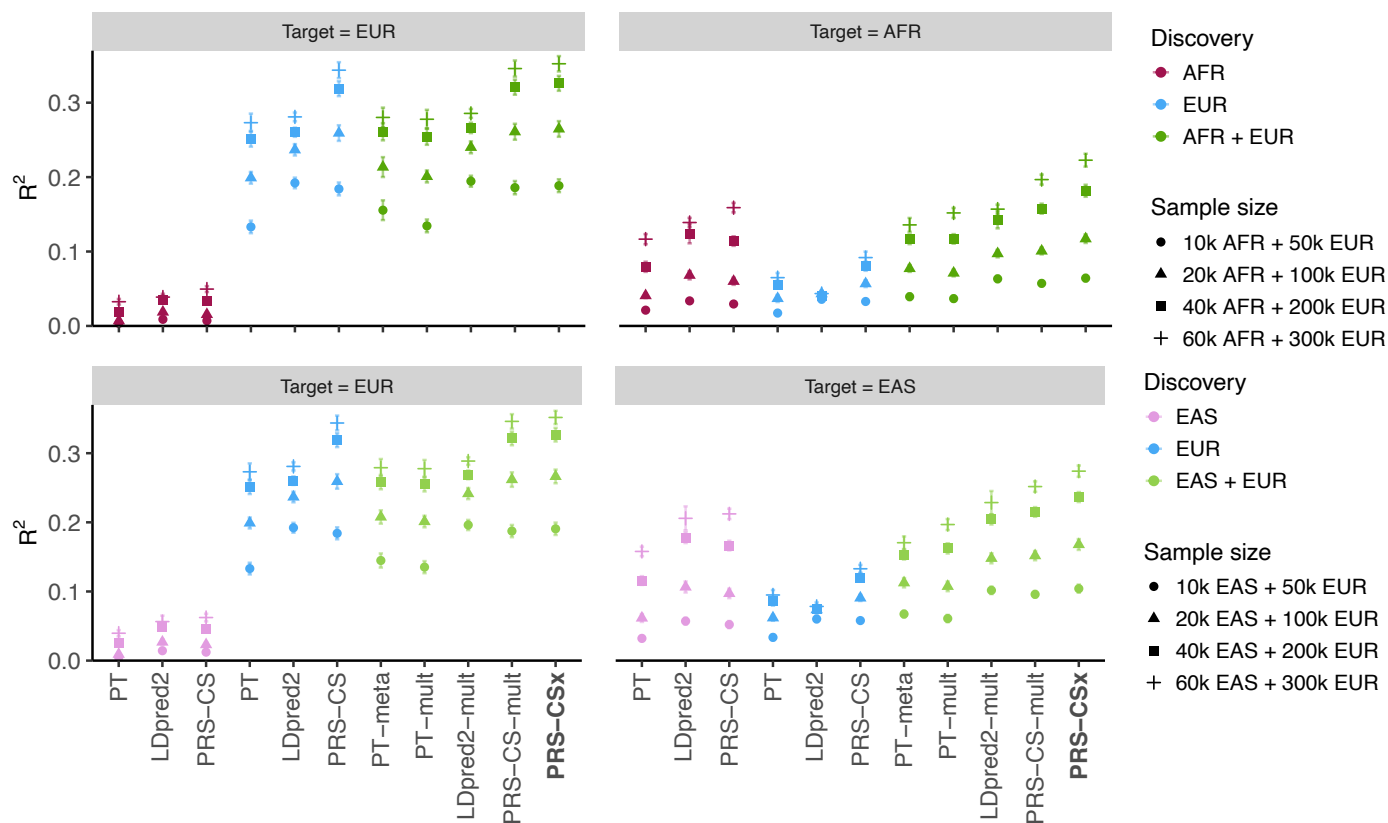
SUPPLEMENTARY FIGURES



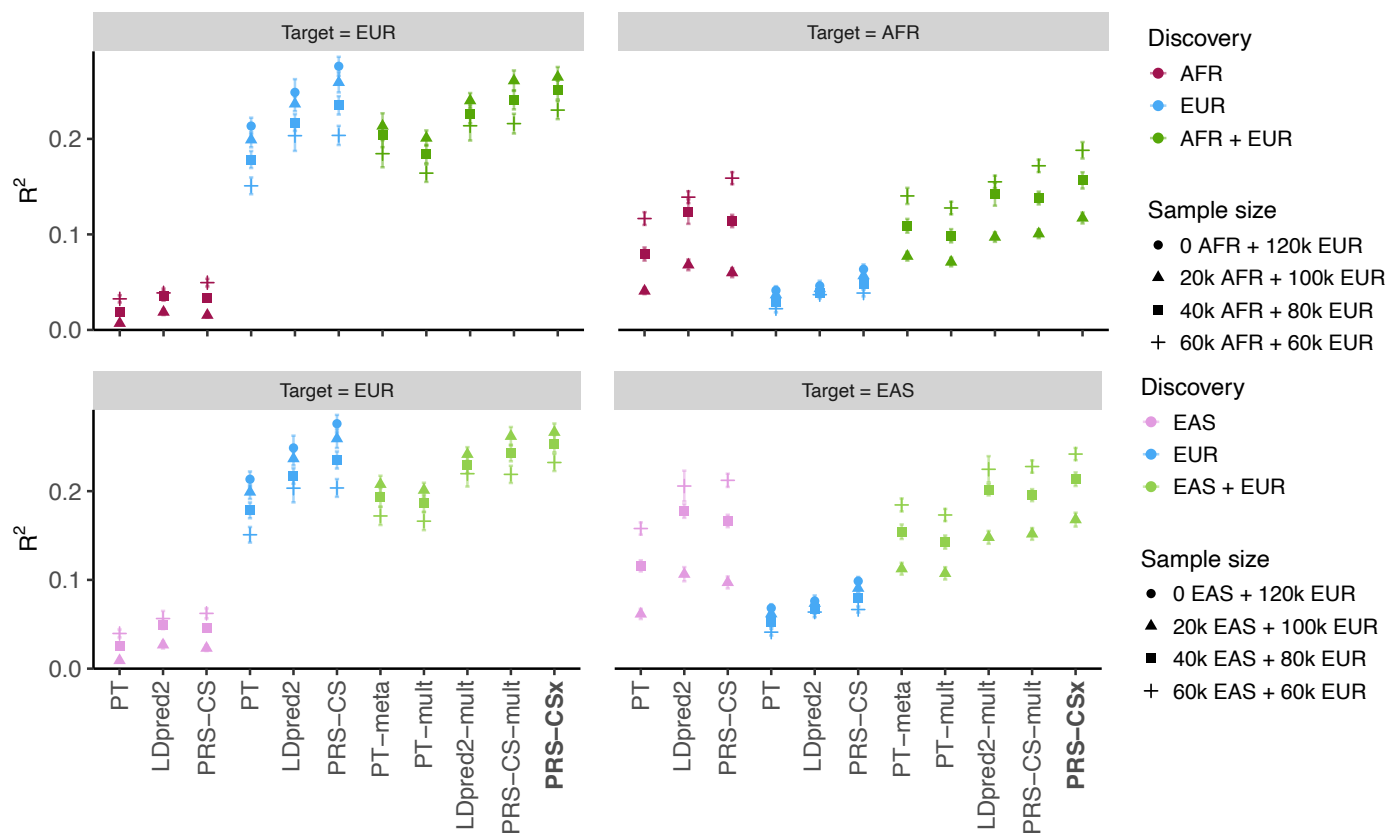
Supplementary Figure 1: Prediction accuracy of different polygenic prediction methods across different genetic architectures. Phenotypes were simulated using 0.1%, 1% or 10% of randomly sampled causal variants (shared across populations), a cross-population genetic correlation of 0.7, and SNP heritability of 50%. PRS were trained using 100K EUR samples and 20K non-EUR (EAS or AFR) samples. Numerical results are reported in Supplementary Table 2.



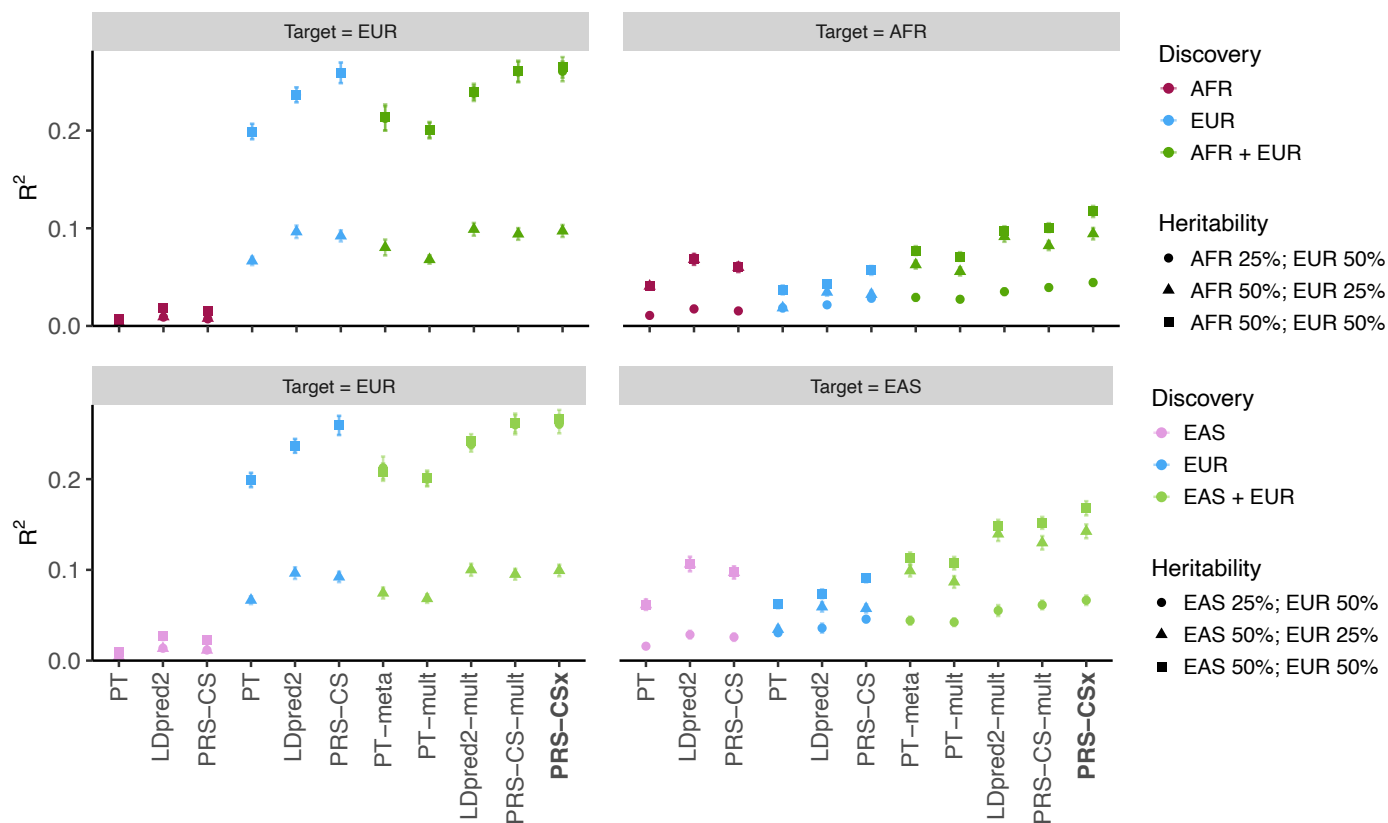
Supplementary Figure 2: Prediction accuracy of different polygenic prediction methods across different cross-population genetic correlations. Phenotypes were simulated using 1% of randomly sampled causal variants (shared across populations), a cross-population genetic correlation of 0.4, 0.7 or 1.0, and SNP heritability of 50%. PRS were trained using 100K EUR samples and 20K non-EUR (EAS or AFR) samples. Numerical results are reported in Supplementary Table 3.



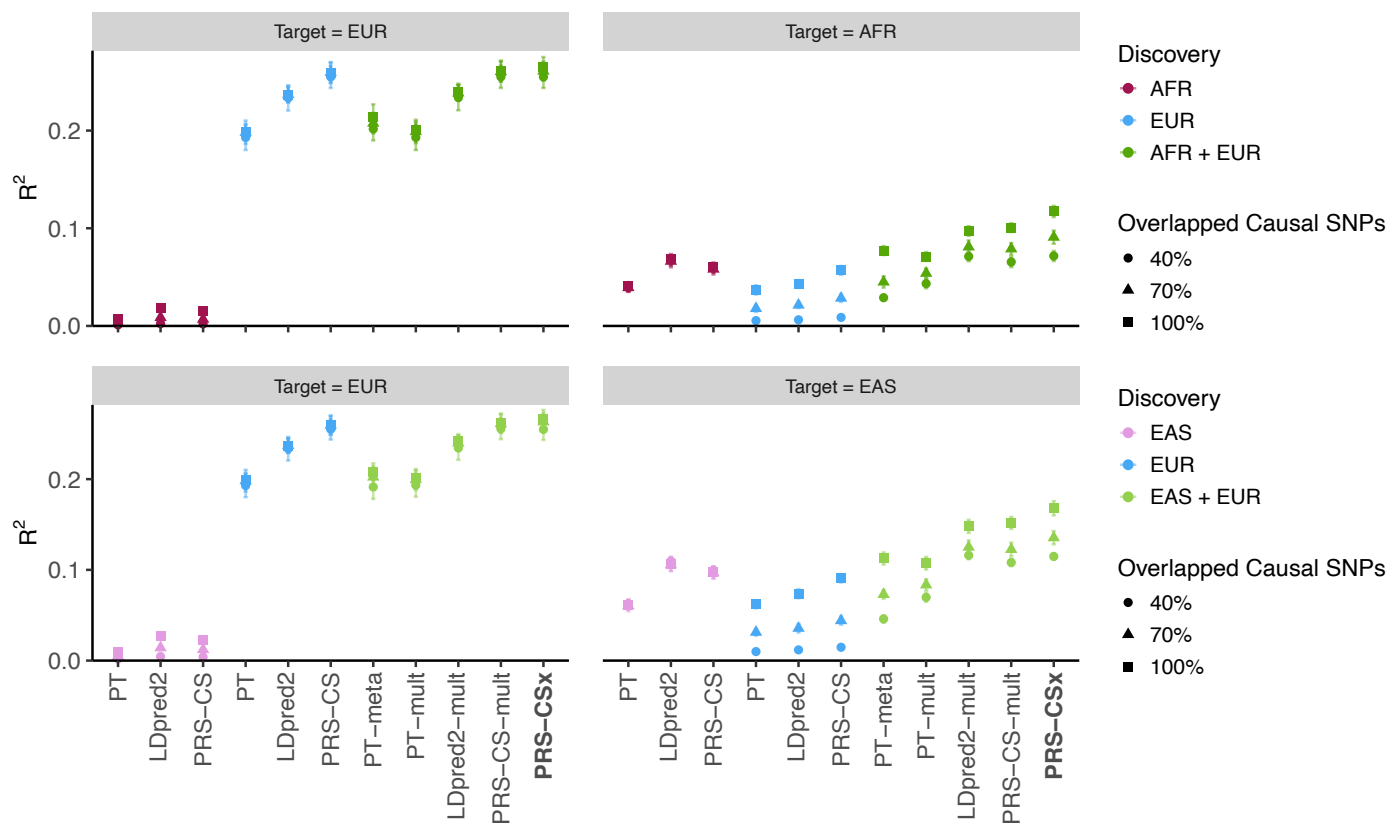
Supplementary Figure 3: Prediction accuracy of different polygenic prediction methods across different discovery GWAS sample sizes. Phenotypes were simulated using 1% of randomly sampled causal variants (shared across populations), a cross-population genetic correlation of 0.7, and SNP heritability of 50%. PRS were trained using 50K EUR and 10K non-EUR (EAS or AFR) samples, 100K EUR and 20K non-EUR samples, 200K EUR and 40K non-EUR samples, or 300K EUR and 60K non-EUR samples. Numerical results are reported in Supplementary Table 4.



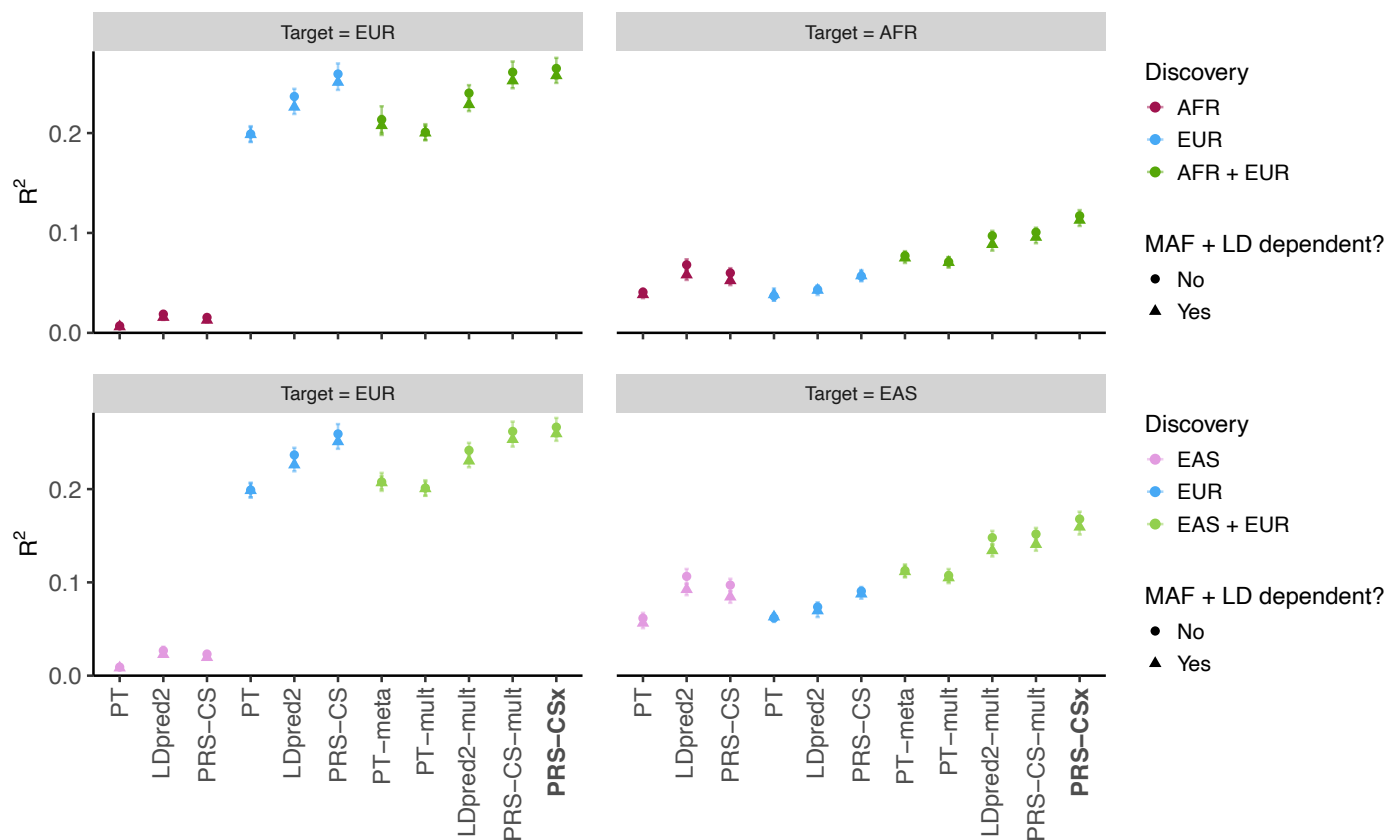
Supplementary Figure 4: Prediction accuracy of different polygenic prediction methods across different ratios of EUR vs. non-EUR GWAS sample sizes. Phenotypes were simulated using 1% of randomly sampled causal variants (shared across populations), a cross-population genetic correlation of 0.7, and SNP heritability of 50%. PRS were trained using 120K EUR samples without non-EUR samples, 100K EUR and 20K non-EUR (EAS or AFR) samples, 80K EUR and 40K non-EUR samples, or 60K EUR and 60K non-EUR samples. Numerical results are reported in Supplementary Table 5.



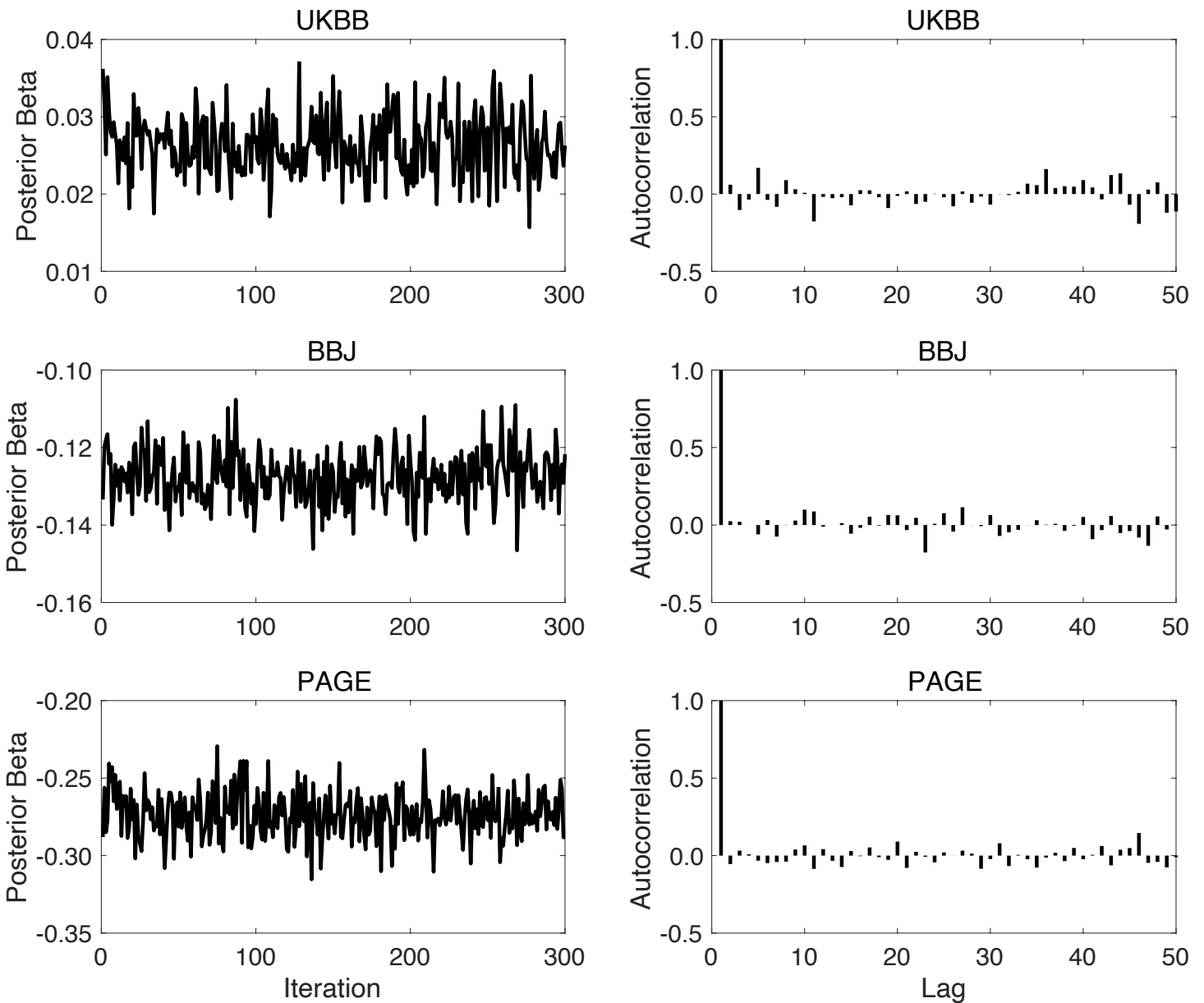
Supplementary Figure 5: Prediction accuracy of different polygenic prediction methods across different SNP heritability. Phenotypes were simulated using 1% of randomly sampled causal variants (shared across populations) and a cross-population genetic correlation of 0.7. SNP heritability was fixed at 50% in each population, 50% in the EUR population and 25% in the non-EUR population, or 25% in the EUR population and 50% in the non-EUR population. PRS were trained using 100K EUR samples and 20K non-EUR (EAS or AFR) samples. Numerical results are reported in Supplementary Table 6.



Supplementary Figure 6: Prediction accuracy of different polygenic prediction methods across different proportions of shared causal variants between populations. Phenotypes were simulated using 1% of randomly sampled causal variants. 100%, 70% or 40% of the causal variants were shared across populations. Shared causal variants had a cross-population genetic correlation of 0.7. SNP heritability was fixed at 50%. PRS were trained using 100K EUR samples and 20K non-EUR (EAS or AFR) samples. Numerical results are reported in Supplementary Table 7.



Supplementary Figure 7: Prediction accuracy of different polygenic prediction methods when SNP effect sizes are minor allele frequency (MAF) and LD dependent. Phenotypes were simulated using 1% of randomly sampled causal variants (shared across populations), a cross-population genetic correlation of 0.7, and SNP heritability of 50%. SNP effect sizes were dependent on MAF and LD scores such that SNPs with lower MAF and located in lower LD regions tended to have larger effect sizes. PRS were trained using 100K EUR samples and 20K non-EUR (EAS or AFR) samples. Numerical results are reported in Supplementary Table 8.



Supplementary Figure 8: Trace plots and autocorrelation functions (ACFs) for assessing the convergence and mixing of the Gibbs sampler used in PRS-CSx. Left panels: Trace plots, after discarding the burn-in iterations and thinning the Markov chain by a factor of 5, for the posterior effects of rs7412 on low-density lipoprotein cholesterol when integrating UKBB, BBJ and PAGE GWAS summary statistics using PRS-CSx. Right panels: The autocorrelation functions (ACFs) for the traces shown on the left.