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

Yunfeng Ruan[1,2], Yen-Chen Anne Feng[1,3,4,5], Chia-Yen Chen[6], Max Lam[1,5,7,8,9], Stanley Global Asia Initiatives, Akira Sawa[10,11], Alicia R. Martin[1,5], Shengying Qin[2,*], Hailiang Huang[1,5,12,*], Tian Ge[1,3,4,*]

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 Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
4 Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
5 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
6 Translational Biology, Biogen Inc., Cambridge, MA, USA
7 Division of Psychiatry Research, The Zucker Hillside Hospital, Northwell Health, Glen Oaks, NY, USA
8 Research Division, Institute of Mental Health Singapore, Singapore, Singapore
9 Human Genetics, Genome Institute of Singapore, Singapore, Singapore
10 Departments of Psychiatry, Neuroscience, and Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD, USA
11 Department of Mental Health, Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, USA
12 Department of Medicine, Massachusetts General Hospital, Harvard Medical School, 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 with coupled continuous shrinkage priors on the SNP effect sizes across populations:

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

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

where, for each population k, $\boldsymbol{y}_k$ is a vector of standardized phenotypes (zero mean and unit variance) from N_k individuals, $\boldsymbol{X}_k$ is an $N_k \times M_k$ 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, $\boldsymbol{I}$ is an identify matrix, ϕ is a global shrinkage parameter shared across SNPs, and ψ_j is a local, SNP-specific shrinkage parameter.

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

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

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

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

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

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

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

- Update σ_k^2:

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

- Update ψ_j:

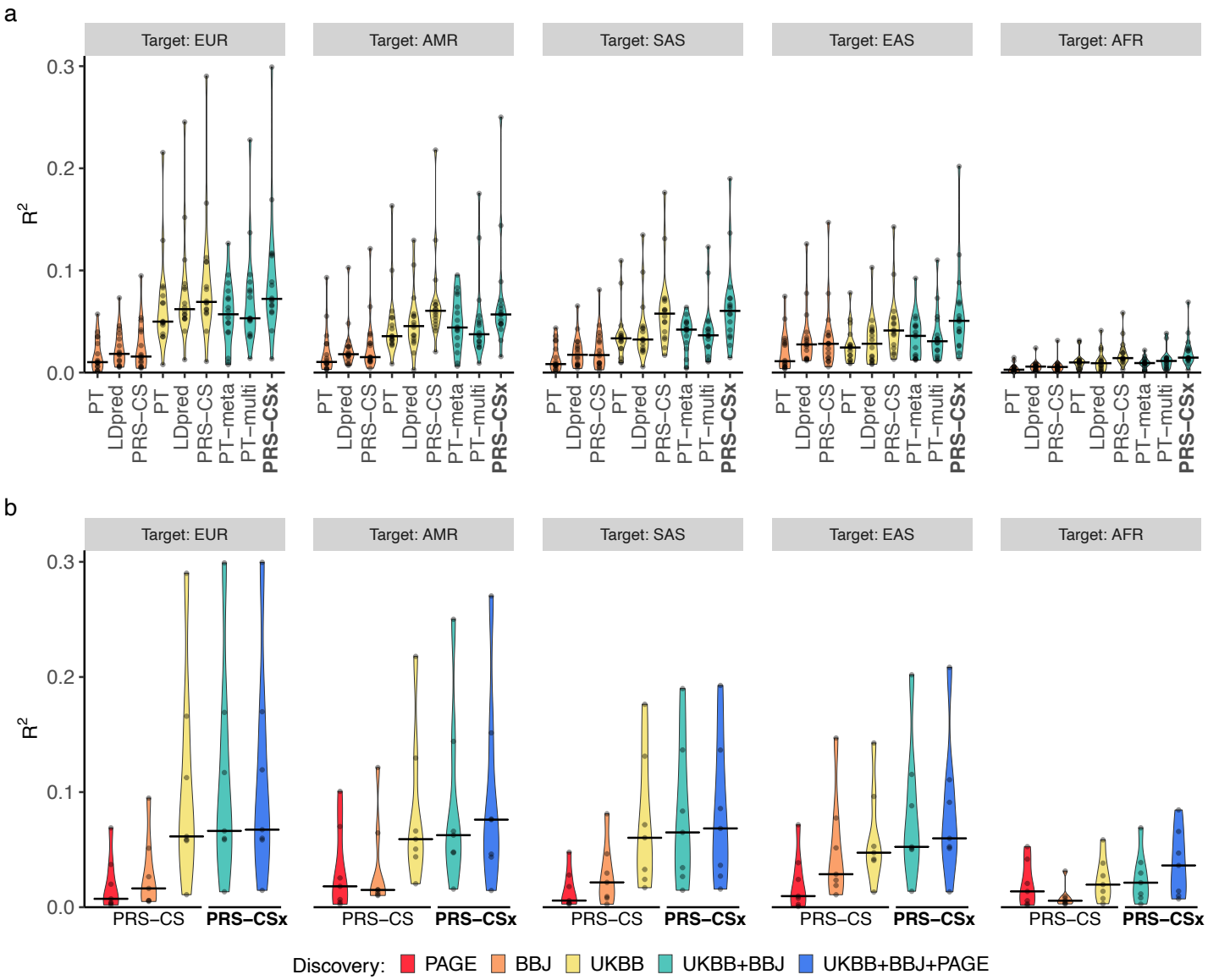
$$\left[\,\psi_j \,\middle|\, \beta_{jk}, \sigma_k^2, \delta_j\,\right] \sim \mathrm{giG}\left(a - \frac{k_j}{2}, 2\delta_j, \sum\nolimits_k \frac{N_k}{\sigma_k^2}\beta_{jk}^2\right),$$

- Update δ_j:

$$[\,\delta_j \mid \psi_j\,] \sim \mathrm{G}(a+b, \psi_j + \phi).$$

In this work, we used pre-calculated 1000 Genomes (1KG) Project Phase 3 LD reference panels for African (AFR), East Asian (EAS) and European (EUR) populations. The genome was partitioned into 2,582, 1,445 and 1,703 largely independent LD blocks for AFR, EAS and EUR, respectively, and LD matrices were calculated for HapMap3 variants with minor allele frequency (MAF) >0.01 using 1KG super-population samples (EUR *N*=503; EAS *N*=504; AFR *N*=661).

SUPPLEMENTARY FIGURES



Supplementary Figure 1: Prediction accuracy of quantitative traits from UKBB, BBJ and PAGE. (a) Prediction accuracy of single-discovery and multi-discovery polygenic prediction methods across 16 traits in five different target populations when using UKBB and BBJ summary statistics as the discovery dataset. (b) Prediction accuracy of PRS-CS and PRS-CSx across 7 traits in different target populations when using UKBB, BBJ and PAGE summary statistics as the discovery dataset. Each point shows the prediction R^2 of a trait, averaged across 100 random splits of the target samples into validation and testing datasets. Crossbar indicates the median R^2 across traits.