BridgePRS: A powerful trans-ancestry Polygenic Risk Score method

Clive Hoggart (✉ clivehoggart@gmail.com)

Icahn School of Medicine, Mount Sinai <https://orcid.org/0000-0002-0482-0052>

Shing Choi

Icahn School of Medicine at Mount Sinai

Michael Preuss

The Charles Bronfman Institute for Personalized Medicine <https://orcid.org/0000-0001-5266-8465>

Paul O'Reilly

Icahn School of Medicine at Mount Sinai <https://orcid.org/0000-0001-7515-0845>

Brief Communication

Keywords:

Posted Date: January 25th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1272562/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

BridgePRS: A powerful trans-ancestry Polygenic Risk Score method

Clive Hoggart^{1*}, Shing Wan Choi^{1,2}, Michael Preuss³
and Paul O'Reilly^{1,2}

¹Department of Genetics and Genomic Sciences, Icahn School of Medicine, Mount Sinai, 1 Gustave L. Levy Pl, New York, 10029, NY, USA.

²MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, De Crespigny Park, Denmark Hill, London, SE5 8AF, London, UK.

³The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine, Mount Sinai, 1 Gustave L. Levy Pl, New York, 10029, NY, USA.

*Corresponding author(s). E-mail(s): clive.hoggart@mssm.edu;
Contributing authors: shingwan.choi@mssm.edu;
michael.preuss@mssm.edu; paul.oreilly@mssm.edu;

Abstract

Polygenic Risk Scores (PRS) are mostly derived using European ancestry genomewide association study (GWAS) data, resulting in substantially lower predictive power in non-European samples. We present *BridgePRS*, a novel Bayesian PRS method that leverages trans-ancestry GWAS data and compare its performance with *PRS-CSx* across 19 UK Biobank (UKB) traits in African and South Asian ancestry individuals. Validation in the UKB and the independent Mount Sinai Biome Biobank demonstrates superior performance of *BridgePRS*.

Polygenic Risk Scores (PRS) are mostly derived using European ancestry genomewide association study (GWAS) data, which results in substantially

2 BridgePRS: A powerful trans-ancestry Polygenic Risk Score method

lower predictive power when applied to non-European samples, in particular African ancestry samples [1, 2]. Despite this, PRS in non-European ancestry samples are still routinely calculated using the same European GWAS data and PRS methods as applied to European data. This takes no account for known population genetic factors affecting the data such as: marked linkage disequilibrium (LD) differences, allele frequency differences driven by genetic drift and natural selection, and GxE interactions affecting causal effect sizes [3]. Consequently, the etiological insights and clinical utility provided by PRS derived in Europeans may have limited relevance to individuals of non-European ancestries.

We introduce *BridgePRS*, a novel Bayesian PRS method that integrates trans-ancestry GWAS summary statistics to optimise PRS predictive accuracy in under-represented populations (see Methods). *BridgePRS* is most applicable to combining the information of a well-powered GWAS performed in a (discovery) population(s) not matched to the ancestry of the target sample, with a second GWAS of limited power in a (target) population that is well-matched to the ancestry of the target sample. The key modelling (Methods) is broken into two stages: (1) a PRS is trained and optimised using discovery population (e.g. European) data and a ridge regression approach that retains a large number of SNPs per locus, (2) the SNP effect sizes of this PRS are updated in a Bayesian framework by those of the target population (e.g. African) GWAS. In this way, this Bayesian ridge approach, *BridgePRS*, helps to bridge the PRS between the two populations.

The main causes of poor trans-ancestry PRS portability are differences in LD and allele frequencies between populations [3]. Differences in LD result in the best tag for a causal variant varying between populations. To account for the resultant uncertainty in the location of causal variants, *BridgePRS* averages SNP effects across putative loci, instead of selecting a single best SNP as performed by standard clumping+thresholding (C+T) PRS [4]. *BridgePRS* is first applied to the discovery population GWAS, using Bayesian modelling with zero-centred Gaussian priors, equivalent to penalised likelihood ridge regression, at putative loci. Given summary data from large GWASs in Europeans, we find that this procedure alone significantly improves predictive accuracy in African and South Asian target data compared to choosing single best SNPs at putative loci.

This first stage modelling results in multivariate Gaussian posterior distributions for SNP effect sizes at each locus. In the next stage, *BridgePRS* integrates the (smaller) target population GWAS data into the PRS by using this posterior distribution as a prior distribution for SNP effect sizes of the target population. This stage allows for different effect size estimates between the populations, accounting for differences in allele frequencies driven by drift or selection, LD and GxE interactions affecting causal effect sizes. Both stages of the modelling use conjugate prior-posterior updates, providing computationally efficient analytical solutions.

Variation in causal allele frequencies between populations can mean that causal variants with relatively low minor allele frequency in the discovery population are estimated with large errors or are missed altogether. To ameliorate this problem, PRS are derived by applying *BridgePRS* stage 1 modelling to the target population data alone (see Methods).

Each stage of the modelling is fit across a spectrum of prior parameters and criteria to select loci that contribute to the PRS, with each combination of parameters giving rise to a unique PRS. These PRS are then combined in a ridge regression fit using available genotype-phenotype test data, choosing the optimum ridge penalty parameter by cross-validation (see Methods).

BridgePRS was applied to UK Biobank (UKB) [5] samples of African and South Asian ancestry across 19 continuous biometric and biochemical traits selected to maximise heritability and samples sizes of non-European individuals, and to minimise their pairwise correlation (maximum $r^2 < 0.3$; see Methods). For each trait, UKB samples of European, African and South Asian ancestry were split into training, test and validation sets in proportions of 2/3, 1/6 and 1/6, respectively, sample sizes are shown in Supplementary Table 1. The training data were used to generate GWAS summary statistics and the test data used to select optimal parameters. Results are shown for the resultant PRS in the unseen UKB validation data. In addition, an entirely out-of-sample validation study was performed by applying PRS derived in the UK Biobank to the independent Mount Sinai BioMe Biobank (BioMe) [6] for nine overlapping traits, sample sizes for these traits for individuals of African and European ancestry are shown in Supplementary Table 2. All results are compared with the alternative trans-ancestry PRS method, *PRS-CSx* [7]. PRS were estimated using both genotyped and imputed data for both methods.

Figure 1 shows the phenotypic variance explained (R^2) by *BridgePRS* and *PRS-CSx* for these analyses. *PRS-CSx* is limited to using HapMap variants [8], therefore the PRS it estimates uses the intersection of HapMap and respectively, genotyped and imputed variants. Table 1 summarises these results in terms of mean R^2 across all traits for each ancestry and cohort and the proportion of times *BridgePRS* has higher R^2 than *PRS-CSx*. Across the eight pairs of analyses (UKB and BioMe cohorts x African and South Asian target population x genotype and imputed data) *BridgePRS* has higher average R^2 than *PRS-CSx* in all but one scenario. The difference in average R^2 is most pronounced for African ancestry samples; the one instance of higher average R^2 for *PRS-CSx* is in South Asian UKB samples using imputed data.

The relative superiority of *BridgePRS* over *PRS-CSx* can partly be explained by *PRS-CSx* restricting its PRS to HapMap variants. When we restrict *BridgePRS* to using only HapMap variants its performance relative to *PRS-CSx* in UKB samples is marginally inferior in African ancestry samples (R^2 : 0.034 s 0.035) and more so in South Asian samples (R^2 : 0.066 v 0.075). However, when applying these UKB derived PRS to the BioMe cohort, *BridgePRS* again has higher average R^2 in both populations: Africans 0.038

vs 0.027 and South Asians 0.068 vs 0.056 (see Supplementary Fig. 1 and Supplementary Table 3). For all analyses, *BridgePRS* on average performed worse with the HapMap variants than either qgenotyped or all imputed variants. However, in practice, *BridgePRS* would never be applied to this restricted SNP set since it can make use of all available SNPs, and the real-world comparison would be *BridgePRS* with either genotyped or imputed variants versus *PRS-CSx* with imputed variants, as reported in Table 1 and discussed above.

PRS-CSx has strikingly poor performance in the prediction of several traits in BioMe: height in African and South Asian ancestry samples using imputed data, mean corpuscular volume in Africans using genotype data and BMI and platelet count in South Asian ancestry samples using genotyped data.

PRS-CSx typically has many more SNPs with non-zero coefficients in its PRS than *BridgePRS*. Supplementary Tables 4 and 5 shows the number of variants included in each model and Supplementary Tables 6 and 7 shows the proportion of overlapping SNPs used in the BioMe validation. The prior used by *PRS-CSx* is similar to the LASSO, encouraging large effect sizes at single variants and small or zero effects at other local variants. Thus, *PRS-CSx*, unlike *BridgePRS*, does not average effects across loci and, therefore, may be more susceptible to missing important variants in situations with imperfect overlapping genetic variant data.

We have applied *BridgePRS* and *PRS-CSx* to powerful real genotype-phenotype data in the UK Biobank to estimate PRS for individuals of African and South Asian ancestry. The PRS have been validated in independent UKB samples and in the independent BioMe cohort of individuals recruited in the New York City area of the USA. *BridgePRS* has superior average R^2 across the traits analysed for samples of African ancestry in both the UKB and BioMe cohorts, using either genotyped or imputed data. For South Asian samples, *BridgePRS* also has superior predictive accuracy in all scenarios except when predicting into the UKB using imputed data. Therefore, according to the data utilised in this study, *BridgePRS* has superior out-of-cohort predictive accuracy relative to *PRS-CSx*, for individuals of both African and South Asian ancestry. Thus, we propose *BridgePRS* as a current leading method for deriving and applying PRS in diverse and under-represented ancestry populations.

Acknowledgements

We thank the participants in the UK Biobank and the scientists involved in the construction of this resource. This research has been conducted using the UK Biobank Resource under application 18177 (Dr O'Reilly). All participants gave full informed consent. This work was supported by a grant from the National Institute of Health (R01MH122866) to PFO. This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

Methods

Modelling

All modelling is performed at the locus level and each locus is assumed to be independent of all others. Within loci, SNP effect sizes β are modelled by a multivariate Gaussian distribution and we assume the trait y of individuals with genotype data X at the locus follows a Gaussian distribution $y \sim N(X\beta, \psi I)$. Throughout, the Gaussian distribution is parameterised by its mean and precision matrix (=inverse covariance matrix).

Below we describe *BridgePRS* methodology to derive a PRS for a target population, population 2, (in our applications Africans and South Asians) for which we have summary statistics from a relatively under-powered GWAS and GWAS summary statistics from a well powered GWAS from a different ancestral population, population 1, in our applications Europeans. We also assume we have small data sets of genotype-phenotype data from both populations.

Stage 1: PRS informed by a single population

In stage 1 modelling, we train and optimise PRS using GWAS summary statistics and test genotype-phenotype data from a single population. To determine the PRS for populations 2, this modelling stage is applied to populations 1 and 2 independently. Application to population 1 determines the prior distributions for populations 2 SNP effects used in stage 2, see below, and application to population 2 is used to identify population 2 specific effects and thus missed in population 1.

In stage 1, a zero centred conjugate Gaussian prior is assigned for the SNP effects at each locus $\beta \sim N(0, \psi\lambda I)$, where I is the identity matrix and λ is a vector of SNP specific shrinkage parameters. The use of a conjugate prior allows the posterior distribution of SNP effects to be determined analytically [9]:

$$\beta \sim N((\lambda I + X^T X)^{-1} X^T y, \psi(\lambda I + X^T X))$$

$X^T y$ can be calculated from the vector of maximum likelihood marginal effects $\hat{\beta}$ available from GWAS summary statistics by $X^T y = n\theta(1 - \theta)\hat{\beta}$, where n is the sample size and θ is the vector of allele frequencies (AF). $X^T X = n\Phi$, where Φ is the pairwise genotypic covariance which can be estimated from a reference panel representative of the population used in the GWAS. Thus, rescaling λ by n , the posterior is estimated as

$$\begin{aligned} \beta &\sim N((\lambda I + \Phi)^{-1} \theta(1 - \theta)\hat{\beta}, \psi(\lambda I + \Phi)) \\ \beta &\sim N(\tilde{\beta}, \psi\Omega) \end{aligned}$$

To accommodate the effects of natural selection we allow the prior on SNP effects to be dependent on AF such that the prior precision for the k th SNP is

$\lambda_k = \lambda_0(\theta_k(1 - \theta_k))^\alpha$ and $\alpha \in (0, 1)$ [10]. When $\alpha = 0$ AF and effect size are *a priori* independent, $\alpha = 1$ is the value implicitly assumed by many methods [11] which implies a strong assumption of larger effects at lower minor allele frequency SNPs. Multiple models are fit at each locus under priors defined by all combinations of $\alpha = (0, 0.25, 0.5, 0.75, 1)$ and $\lambda_0 = (0.05, 0.1, 0.2, 0.5, 1, 2, 5)$. Loci are ranked by the P value of their most associated SNP and assigned to subset S_k if the top SNP P value $< 10^{-k}$, values of $k = 1, \dots, 8$ are considered. Multiple genome-wide PRS are calculated for a test set of phenotype and genotype data by summing the effects across all contributing loci across all combinations of α, λ_0 and k :

$$\text{PRS}_{ijk} = \sum_{l \in S_k} X_l \tilde{\beta}_{\lambda_0^{(i)} \alpha^{(j)}}^{(l)}$$

where X_l is the genotype data at locus l , $\tilde{\beta}_{\lambda_0^{(i)} \alpha^{(j)}}^{(l)}$ is the posterior mean at locus l with prior defined by parameters $\lambda_0^{(i)}$ and $\alpha^{(j)}$, and S_k is the subset of loci with top SNP P value $< 10^{-k}$. A single PRS is calculated by a weighted sum of the PRS across all i, j and k , with weights determined by a ridge regression fit to the test data utilising leave-one-out cross-validation to select the ridge shrinkage parameter which minimises out-of-sample deviance as implemented in the R package *glmnet* [12].

Stage 2: PRS informed by stage 1

In stage 2 modelling, SNP effect sizes estimated by the application of stage 1 modelling to population 1 (eg. Europeans) are updated based on population 2 GWAS summary statistics and optimised using population 2 genotype-phenotype data. The prior used is taken as the posterior derived from the λ_0 and α prior parameters which optimise prediction in the populations 1 test data. As for stage 1 this prior is also a multivariate Gaussian. A parameter τ is added to the precision parameter of the Gaussian to control the contribution of population 1 to population 2, thus the prior is specified as $\beta_2 \sim N(\tilde{\beta}_1, \psi\tau\Omega_1)$. This is similarly a conjugate model with a Gaussian posterior [9]:

$$\begin{aligned} \beta_2 &\sim N\left((\tau\Omega_1 + \Phi_2)^{-1} \left(\tau\Omega_1\tilde{\beta}_1 + \hat{\beta}_2\theta_2(1 - \theta_2)\right), \psi(\tau\Omega_1 + \Phi_2)\right) \\ \beta_2 &\sim N(\tilde{\beta}_2, \Omega_2) \end{aligned}$$

where Φ_2 is the SNP covariance at the locus in population 2, $\hat{\beta}_2$ is the vector of marginal maximum likelihood SNP effect sizes and θ_2 is the vector of allele frequencies. Small values of τ correspond to using effect estimates close to those from population 2, as τ increases more weight is assigned to population 1, such that as $\tau \rightarrow \infty, \beta_2 \rightarrow \beta_1$. Because of differences in LD between populations we do not rank loci by the P value of a single best SNP, we aggregate information across loci by adapting the F-test. It can be shown that the

F-test in a multivariate linear regression model for the null $H_0: \beta = 0$, is well approximated by (see Appendix for proof):

$$F_{stat} = \frac{n-k}{kn\sigma^2} \beta^T X^T X \beta$$

with degrees of freedom k and $n-k$ where k is the number of covariates of β , n is the number of observations and σ^2 is the phenotypic variance. The maximum likelihood estimate and $X^T X$ are substituted by the posterior mean and precision matrix and n with $n_{eff} = n(1 + \tau)$, the effective number of observations accounting for the prior.

$$F_{Bayes} = \frac{n_{eff} - k}{k\sigma^2} \tilde{\beta}_2^T \Omega_2 \tilde{\beta}_2$$

The resultant tail probability is analogous to a P value, although the parameter estimates β and λ include prior information and, thus, it cannot be interpreted as such. Instead, for each τ , a locus with test statistic F is assigned to S_k if $F > q_k$, where q_k is the F quantile corresponding to $\text{Prob}(p < 10^{-k})$, where p are the locus-specific top SNP P values. This ranking ensures the pseudo F -statistic ranking assigns the same number of loci to each subset as the SNP P value ranking. As for the one stage single ancestry PRS multiple genome-wide PRS are constructed by:

$$\text{PRS}_{ik} = \sum_{l \in S_k} X_l \beta_{\tau_i}^{(l)}$$

where $\beta_{\tau_i}^{(l)}$ is the posterior mean at locus l with prior defined by parameters $\tau_i^{(i)}$, and S_k is the subset of loci with $F > q_k$. Models are fit for $\tau = 1, 2, 5, 10, 15, 20, 50, 100, 200, 500$ and the same P value thresholds as used in the first stage of the modelling. A single PRS is estimated via a ridge regression fit using population 2 test data as described above using glmnet.

Incomplete SNP overlap between populations 1 and 2

Quality control (QC) is performed separately in each population, see below. This results in variants included in analyses differing between populations. Thus, stage 2 analyses are performed on the intersection of variants passing QC in both populations and the prior is calculated conditional on effects of non-overlapping variants set to zero. Thus given a prior of $\beta_2 \sim N(\tilde{\beta}_1, \psi\tau\Omega_1)$, the prior on the overlapping variants is given by [9]

$$p\left(\beta_2^{(a)} \mid \beta_2^{(b)} = 0\right) = N\left(\tilde{\beta}_1^{(a)} + \left(\Omega_1^{(aa)}\right)^{-1} \Omega_1^{(ab)} \tilde{\beta}_1^{(b)}, \psi\tau\Omega_1^{(aa)}\right)$$

where a represents the overlapping variants, b the non-overlapping variants and $\Omega_1^{(aa)}$ and $\Omega_1^{(ab)}$ are the appropriate submatrices of Ω_1 .

Combining PRS

We consider three alternative models for the PRS of population 2: (1) PRS estimated using only population 2, i.e. European GWAS does not inform the PRS of population 2, (2) PRS estimated using only the two-stage European informed PRS, i.e. the population 2 GWAS is under-powered and contributes insufficient information on its own and (3) both the population 2 only PRS and the two-stage PRS contribute independent information. The estimation of models (1) and (2) are determined by a cross-validated ridge regression fit as described above using the R package glmnet [12]. Model (3) is estimated similarly by merging all single ancestry and two stage PRS and weighting by a cross-validated ridge regression fit.

The final PRS is a weighted average of these three PRS, with weights determined by the estimated marginal likelihood of each. The log-marginal likelihood of a linear regression model M_i can be approximated by [13]

$$\log p(y, X | M_i) = \frac{n}{2} \log \sigma_i^2 + \kappa$$

where σ_i^2 is the residual model variance estimated from cross-validation and κ is a constant. With equal prior weight for each of the models, the posterior model weights for models M_1 , M_2 and M_3 are given by:

$$p(M_i | y, X) = \frac{\exp \{n \log \sigma_i^2 / 2\}}{\sum_{i=1}^3 \exp \{n \log \sigma_i^2 / 2\}}$$

Definition of loci

Loci for the two-stage modelling were defined by clumping and thresholding of European GWAS summary statistics and LD structure using PLINK v1.9 [14] with the following parameters: `--clump-p1 0.01 --clump-p2 0.01 --clump-kb 1000 --clump-r2 0.01`. The P value for each locus was determined by the P value of the lead SNP of the locus in the European GWAS. The ancestry specific loci were defined similarly but used GWAS data from the appropriate ancestry (African or South Asian).

Application of *PRS-CSx*

PRS-CSx is a python based software package that integrates GWAS summary statistics and LD reference data from multiple populations to estimate population specific PRS. *PRS-CSx* applies a continuous shrinkage prior to SNP effects genome-wide in which the sparseness of the genetic architecture across populations is controlled by a parameter ϕ . *PRS-CSx* does not make inference on ϕ but instead estimates separate PRS for each value of ϕ considered. Throughout we follow the implementation described in Ruan et al [7], thus values of $\phi = (10^{-6}, 10^{-4}, 10^{-2}, 1)$ were considered. For each ϕ , *PRS-CSx* first estimates population specific PRS, eg $\text{PRS}_{\phi, EUR}$ and $\text{PRS}_{\phi, AFR}$, where $\text{PRS}_{\phi, x}$ is the

standardised PRS for population x . For each ϕ , *PRS-CSx* fits the following linear regression to the target population test data y

$$y \sim w_{\phi, EUR} \text{PRS}_{\phi, EUR} + w_{\phi, AFR} \text{PRS}_{\phi, AFR}$$

The ϕ value and the corresponding regression coefficients for the linear combination of PRS that maximise the coefficient of determination (R^2) in the target population (eg. Africans) test set were used in the validation dataset to calculate the final PRS:

$$\text{PRS}_{final} = \hat{w}_{\hat{\phi}, EUR} \text{PRS}_{\hat{\phi}, EUR} + \hat{w}_{\hat{\phi}, AFR} \text{PRS}_{\hat{\phi}, AFR}$$

Unlike *BridgePRS*, *PRS-CSx* does not use European test data to estimate non-European PRS. Therefore, to ensure both methods use the same data GWASs were performed on the European test samples using PLINK v2.0 [14] and then meta-analysed with the GWAS data from the European data METAL [15]. The meta-analysed European GWAS, the GWAS generated from the training samples of the target population (African or South Asian), and the LD reference panel generated by the authors of *PRS-CSx* were provided to *PRS-CSx*.

UK Biobank genotype and sample QC

The UK Biobank (UKB) is a prospective cohort study of around 500,000 individuals recruited across the United Kingdom during 2006-2010. The genetic data is comprised of 488,377 samples genotyped at 805,426 SNPs. Population ancestries were defined by 4-means clustering performed on the first two Principal Components (PCs) of the genotype data. The ancestry of each cluster was defined by the country of birth (field ID: 20115) of the majority of individuals in the cluster. Standard quality control (QC) procedures were then performed on each ancestry cluster independently, any SNP with minor allele frequency < 0.01 , genotype missingness > 0.02 or Hardy Weinberg Equilibrium Test P value $< 10^{-8}$ was removed. Samples with high levels of missingness or heterozygosity, with mismatching genetic-inferred and self-reported sex, or with aneuploidy of the sex chromosomes were removed as recommended by the UKB data processing team. A greedy algorithm [16] was used to remove related individuals, with kinship coefficient > 0.044 , in a way that maximised sample retention. In total, 557,369 SNPs and 387,392 individuals were retained for analysis.

Imputation

Imputed variants were extracted from imputed UK Biobank data using PLINK v2.0, converting the imputed data into hard-coded genotypes and retaining variants with the following filters: biallelic variants (`--max-alleles 2`), minor allele frequency greater than 0.001 (`-maf 0.001`), genotype missingness less than 1% (`--geno 0.01`) and a MACH info score greater than 0.8 (`--mach-r2-filter 0.8`).

Trait selection

We extracted all continuous traits from the UK Biobank and performed basic filtering, discarding samples with phenotypic values 6 standard deviation away from the mean. Traits with more than 2,000 samples of African ancestry were extracted. For each trait 300,000 European samples were extracted (retaining at least 10,000 samples for test and validation for each trait) and GWASs run on the genotype data using PLINK v2.0 using `--glm`. Sex (field ID: 31), age (field ID: 21003), genotyping batch, UK Biobank assessment centre (field ID: 54) and 40 PCs were included as covariates, with fasting time (field ID: 74) and dilution factor (field ID: 30897) also included for blood biochemical traits. LD Score regression [17] was run on the resultant summary statistics and traits were further filtered discarding those with heritability less than 1%. The remaining traits were ranked according to their heritability, and traits correlated with a more heritable trait (absolute Pearson correlation greater than 0.3) were removed resulting in 27 traits. Results are presented for 19 traits which have an R^2 in Africans of greater than 1% for at least one analysis. The sample sizes for each trait and ancestry are shown in Supplementary Table 2.

Implementation

European, African and South Asian samples were split into three independent groups: training data to construct the GWAS summary statistics; test data, to select best fitting parameters; and validation data, to calculate out-of-sample predictive accuracy. The proportion of samples allocated to each set were 2/3 training, 1/6 test and 1/6 validation.

For each trait *BridgePRS* was run with three SNP/variant panels: (1) genotyped SNPs, (2) imputed SNPs/variants and (3) the overlap of variants passing imputation QC and HapMap variants. GWASs were run separately for the training samples of European, African and East Asian ancestry for each of the 27 traits using PLINK v2.0 as described above. All PRS were calculated using two populations, African PRS used African and European GWAS data and South Asian PRS used South Asian and European GWAS data.

Application to BioMe

BioMe BioBank samples were genotyped on the Infinium Global Screening Array v1.0 (GSA) platform. Samples were removed with a population-specific heterozygosity rate of greater than ± 6 standard deviations of population-specific mean, along with a call rate of $< 95\%$. In addition, samples were removed for exhibiting persistent discordance between EHR recorded and genetic sex. Variants were removed that had a call rate $< 95\%$, a Hardy-Weinburg Equilibrium P value threshold of $p < 10^{-5}$ in African-American and European-American ancestry, or $p < 10^{-13}$ in Hispanic and South Asian ancestry.

PCA was performed; African and South Asian samples were selected by clusters on PC plots corresponding to self-reported ancestry. African samples

were selected as those with PC 1 > 0.0075, PC 2 < -0.0005 and PC 3 > -0.002. South Asian samples were selected as those with -0.01 < PC 3 < -0.004, -0.003 < PC 4 < 0.001 and PC 5 < -0.015. Supplementary Fig. 2-4 plot the top 6 PCs, with samples coloured by self-reported ancestry and show the thresholds used to select African and South Asian ancestry samples.

Imputation was performed using *IMPUTE2* [18] with the 1000G Phase3 v5 reference panel [19]. Variants were first filtered by info score > 0.3. Genotype data for the calculation of the PRS was generated for each of the two ancestry groups separately by first removing variants with minor allele frequency < 1% in the respective BioMe BioBank population and then removing one of each pair of variants with duplicate genomic position. BioMe BioBank variants were mapped onto the UKB PRS by genomic position (build 37). The number of SNPs in each PRS applied and the proportion available in the BioMe BioBank are shown Supplementary Tables 2-5. Variants were coded by their expected allele count (dosage) for the calculation of PRS.

Measure of PRS accuracy

Variance explained was calculated as

$$R^2 = 1 - \frac{\text{Var}(y \mid M_1)}{\text{Var}(y \mid M_0)}$$

where M_i is the regression model with the PRS ($i = 1$) and without ($i = 0$) and with both models including covariates for top 40 PCs, age, sex, centre and batch, fasting and dilution for the biochemical traits. Variance explained in the applications to BioMe included covariates for age, sex and the top 32 PCs. Confidence intervals were calculated by bootstrapping in the R package *boot* [20] using 10,000 replicates.

Reformulation of F-test

Without loss of generality, assume zero centred normally distributed trait data y with variance σ^2 . A linear regression is fit to this data with $n \times k$ covariate matrix X resulting in a maximum likelihood estimates $\hat{\beta}$. The F statistic is defined by the residual sum of squares of the null and alternative models (RSS_0 and RSS_1) as follows:

$$\begin{aligned} F &= \frac{n-k}{k} \left(\frac{\text{RSS}_0 - \text{RSS}_1}{\text{RSS}_1} \right) \\ &= \frac{n-k}{k} \left(\frac{y^T y - (y - X\hat{\beta})^T (y - X\hat{\beta})}{(y - X\hat{\beta})^T (y - X\hat{\beta})} \right) \\ &= \frac{n-k}{k} \left(\frac{n\sigma^2}{(y - X\hat{\beta})^T (y - X\hat{\beta})} - 1 \right) \end{aligned}$$

$$= \frac{n-k}{k} \left(\frac{n\sigma^2}{y^T y - \hat{\beta}^T X^T y - y^T X \hat{\beta} + \hat{\beta}^T X^T X \hat{\beta}} - 1 \right)$$

since $\hat{\beta} = (X^T X)^{-1} X^T y$

$$\begin{aligned} &= \frac{n-k}{k} \left(\frac{n\sigma^2}{n\sigma^2 - \hat{\beta}^T X^T X \hat{\beta}} - 1 \right) \\ &= \frac{n-k}{k} \left(\frac{\hat{\beta}^T X^T X \hat{\beta}}{n\sigma^2 - \hat{\beta}^T X^T X \hat{\beta}} \right) \\ &= \frac{n-k}{kn\sigma^2} \hat{\beta}^T X^T X \hat{\beta} \left(1 - \frac{1}{n\sigma^2} \hat{\beta}^T X^T X \hat{\beta} \right)^{-1} \end{aligned}$$

and a first order Taylor approximation gives

$$\approx \frac{n-k}{kn\sigma^2} \hat{\beta}^T X^T X \hat{\beta}$$

References

- [1] Martin, A. R. *et al.* Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics* **100** (4), 635–649 (2017).
- [2] Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications* **10** (1), 3328 (2019).
- [3] Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nature Communications* **11** (1), 3865 (2020) .
- [4] Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols* **15** (9), 2759–2772 (2020).
- [5] Sudlow, C. *et al.* UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* **12** (3), e1001779–e1001779 (2015).
- [6] Abul-Husn, N. S. & Kenny, E. E. Personalized medicine and the power of electronic health records. *Cell* **177** (1), 58–69 (2019).

- [7] Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *medRxiv* (2021). URL <https://www.medrxiv.org/content/early/2021/01/02/2020.12.27.20248738>. .
- [8] The International HapMap Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467** (7311), 52–58 (2010).
- [9] Bernardo, J. M. & Smith, A. F. M. *Bayesian Theory* (Wiley, Chichester, 1994).
- [10] Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide snps. *American journal of human genetics* **91** (6), 1011–1021 (2012).
- [11] Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. Gcta: a tool for genome-wide complex trait analysis. *American journal of human genetics* **88** (1), 76–82 (2011).
- [12] Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33** (1), 1–22 (2010).
- [13] Fong, E. & Holmes, C. C. On the marginal likelihood and cross-validation. *Biometrika* **107** (2), 489–496 (2020). .
- [14] Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4** (1) (2015). .
- [15] Willer, C. J., Li, Y. & Abecasis, G. R. Metal: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics (Oxford, England)* **26** (17), 2190–2191 (2010).
- [16] Choi, S. W. Greedyrelated: Script for greedily remove related samples, 1.2. Github (2017). URL https://zenodo.org/record/3697212#.Yd_oil3sc.
- [17] Bulik-Sullivan, B. K. *et al.* Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* **47** (3), 291–295 (2015).
- [18] Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5** (6), e1000529–e1000529 (2009).
- [19] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467** (7319), 1061–1073 (2010). .

14 BridgePRS: *A powerful trans-ancestry Polygenic Risk Score method*

[20] Cauty, A. & Ripley, B. D. *boot: Bootstrap R (S-Plus) Functions* (2021).
R package version 1.3-28.

		Genotype data			Imputed data		
		Mean R^2		<i>BridgePRS</i> > <i>PRS-CSx</i>	Mean R^2		<i>BridgePRS</i> > <i>PRS-CSx</i>
		<i>BridgePRS</i>	<i>PRS-CSx</i>		<i>BridgePRS</i>	<i>PRS-CSx</i>	
UKB (All traits)	AFR	0.041	0.029	84%	0.041	0.035	74%
	SAS	0.069	0.062	74%	0.068	0.075	21%
UKB (BioMe traits)	AFR	0.063	0.045	78%	0.063	0.053	67%
	SAS	0.084	0.073	89%	0.083	0.091	22%
BioMe	AFR	0.044	0.025	89%	0.044	0.027	67%
	SAS	0.066	0.046	67%	0.067	0.056	44%

Table 1 Mean phenotypic variance explained (R^2) of *BridgePRS* and *PRS-CSx* in the UKB and BioMe for samples of African (AFR) and South Asian (SAS) ancestry using genotype and imputed data. Also shown is the proportion of times R^2 of *BridgePRS* is greater than *PRS-CSx*. UKB results are shown across all 19 traits and the 9 overlapping BioMe traits.

Fig. 1 Phenotypic variance explained (R^2 point estimates and 95% confidence intervals) by *BridgePRS* and *PRS-CSx* in samples of African and South Asian ancestry in the UK Biobank and BioMe cohorts using genotype and imputed data. **A** African ancestry samples in UK Biobank. **B** South Asian ancestry samples in UK Biobank. **C** African ancestry samples in BioMe. **D** South Asian ancestry samples in BioMe. Neutro count=Neutrophil count, MCV=Mean corpuscular volume, Platelets=Platelet count, Retic count=Reticulocyte percentage, ALP=Alkaline phosphatase, Mono count=Monocyte count, apoA1=Apolipoprotein A, BMI=Body mass index, RDW=Red blood cell distribution width, Eos count=Eosinophil count, TG=Triglycerides, Baso %=Basophil percentage, CRP=C-reactive protein

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supp.pdf](#)