One score to rule them all: regularized ensemble polygenic risk prediction with GWAS summary statistics

Zijie Zhao[1,*], Stephen Dorn[1,*], Yuchang Wu[1], Xiaoyu Yang[1], Jin Jin[2], Qiongshi Lu[1,3,#]

[1] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI
[2] Department of Biostatistics, Epidemiology and Bioinformatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA
[3] Department of Statistics, University of Wisconsin-Madison, Madison, WI

* These authors contributed equally to this work

To whom correspondence should be addressed:
Dr. Qiongshi Lu
qlu@biostat.wisc.edu

Abstract

Ensemble learning has been increasingly popular for boosting the predictive power of polygenic risk scores (PRS), with almost every recent multi-ancestry PRS approach employing ensemble learning as a final step. Existing ensemble approaches rely on individual-level data for model training, which severely limits their real-world applications, especially in non-European populations without sufficient genomic samples. Here, we introduce a statistical framework to construct regularized ensemble PRS, which allows us to combine a large number of candidate PRS models using only summary statistics from genome-wide association studies. We demonstrate its robust and substantial improvement over many existing PRS models in both within- and cross-ancestry applications. We believe this is truly “one score to rule them all” due to its capability to continuously combine newly developed PRS models with existing models to improve prediction performance, which makes it a universal approach that should always be employed in future PRS applications.

Introduction

With the continued success of genome-wide association studies (GWAS)[1,2] and increasingly accessible summary statistics from these studies, genetic prediction efforts have generally focused on creating polygenic risk scores (PRS) that combine individually negligible but collectively substantial effects from millions of single nucleotide polymorphisms (SNPs) in GWAS summary data. Over the years, PRS methodology has evolved through improved model designs[3-9], integration of functional genomic annotations[8,10-12], and multi-ancestry/multi-trait joint modeling[12-14]. Due to the diverse genetic architecture of complex traits and moderate signal-to-noise ratio at current GWAS sample sizes, not all PRS models perform equally well, and no single method consistently outperforms others[15-17]. Ensemble learning is a strategy that trains a machine learning model to combine multiple learning algorithms for better predictive performance. In recent years, researchers have applied ensemble learning to integrate multiple PRS into an aggregated score with improved performance compared to any single PRS model[15,18-20]. Earlier work used linear regression or penalized regression to develop ensemble PRS[18,19]. More recently, super learning has been introduced as an omnibus approach for ensemble PRS construction[21-23]. It employs an "ensemble of ensemble" modeling strategy to achieve additional prediction gains from various PRS model designs and ensemble techniques[24,25]. In particular, these approaches have proven effective in multi-ancestry PRS applications[21-23,26,27], leveraging the many PRS models optimized for each ancestry respectively to improve the predictive performance in the target population. Due to its apparent effectiveness, almost every recent PRS method employs ensemble learning in some way.

Unsurprisingly, ensemble learning is data-demanding – ensemble model training requires a holdout dataset (or multiple datasets in the case of super learning) independent from GWAS samples. This creates a major hurdle for ensemble PRS application. Often, in practice, a summary-level GWAS dataset is all there is for PRS model training, which is especially true in applications in non-European ancestries. Inevitably, researchers who wish to employ ensemble learning need to partition the valuable testing dataset, which is often small in size if it even exists, leading to insufficient PRS benchmarking and reduced statistical power in downstream applications[18,19,28]. This problem is further exacerbated in applications of super learning, where finer data partitioning is required to train a multi-level ensemble model. To avoid the need for individual-level holdout datasets, we recently introduced an approach (PUMAS) to fine-tune PRS[29] and obtain a linear combination of multiple PRS models using only GWAS summary statistics[30]. Although it was an important proof of concept, the simple linear combination approach may yield problematic results when combining a large number of PRS models. An ensemble PRS approach that requires only GWAS summary statistics and can employ more advanced ensemble learning strategies to further improve PRS predictive utility is thus naturally desired.

Here, we address these challenges by introducing two summary statistics-based ensemble learning techniques, which we have incorporated into the PUMAS-ensemble software suite[30]. Our elastic net (PUMAS-EN) ensemble learning approach simultaneously conducts ensemble model training, fine-tuning, and benchmarking based on a large number of input PRS models. We also introduce a super learning approach (PUMAS-SL) to combine multiple regularized ensemble PRS. Through extensive simulations and analysis of many datasets, we demonstrate the robust and superior performance of our approach in both within and cross-ancestry PRS applications. Most importantly, we showcase the capability of our approach to combine a large number of PRS models, which makes it a universal framework that can continue to evolve by always including the most recent PRS models in the literature.

Results

Method overview

We first provide an overview of our workflow and will delve into statistical details in the **Methods** section. If a dataset with individual-level genotype and phenotype information is available, the conventional strategy for fitting and evaluating ensemble PRS models is to split the full sample into independent subsets for ensemble model training and benchmarking (**Figure 1A**). If the ensemble model has tuning parameters, e.g., regularization parameters in penalized regressions, the dataset for model training needs to be further divided so that a subset can be used for fine-tuning of hyperparameters. However, since an individual-level dataset with sufficient samples is often unavailable in practice, we design and employ an ensemble learning process following a similar modeling framework but requiring only GWAS summary statistics. In PUMAS-ensemble, we partition the full GWAS summary statistics dataset into three down-sampled summary statistics datasets for training, ensemble training, and testing, respectively (**Figure 1B**). With these down-sampled summary statistics datasets, we can train multiple PRS models by various methods, apply the two ensemble learning approaches based on elastic net and super learning to integrate the various PRS models, and finally, benchmark the performance of the constructed ensemble PRS. Only GWAS summary statistics and linkage disequilibrium (LD) reference data are required in this framework. We note that this is a general framework that allows researchers to choose and combine any set of PRS methods for improved polygenic prediction. For illustration, we considered eight commonly implemented PRS methods for most of our analyses in this study[4-9,31,32] (**Supplementary Table 1**). Details of method implementation are summarized in **Methods**.

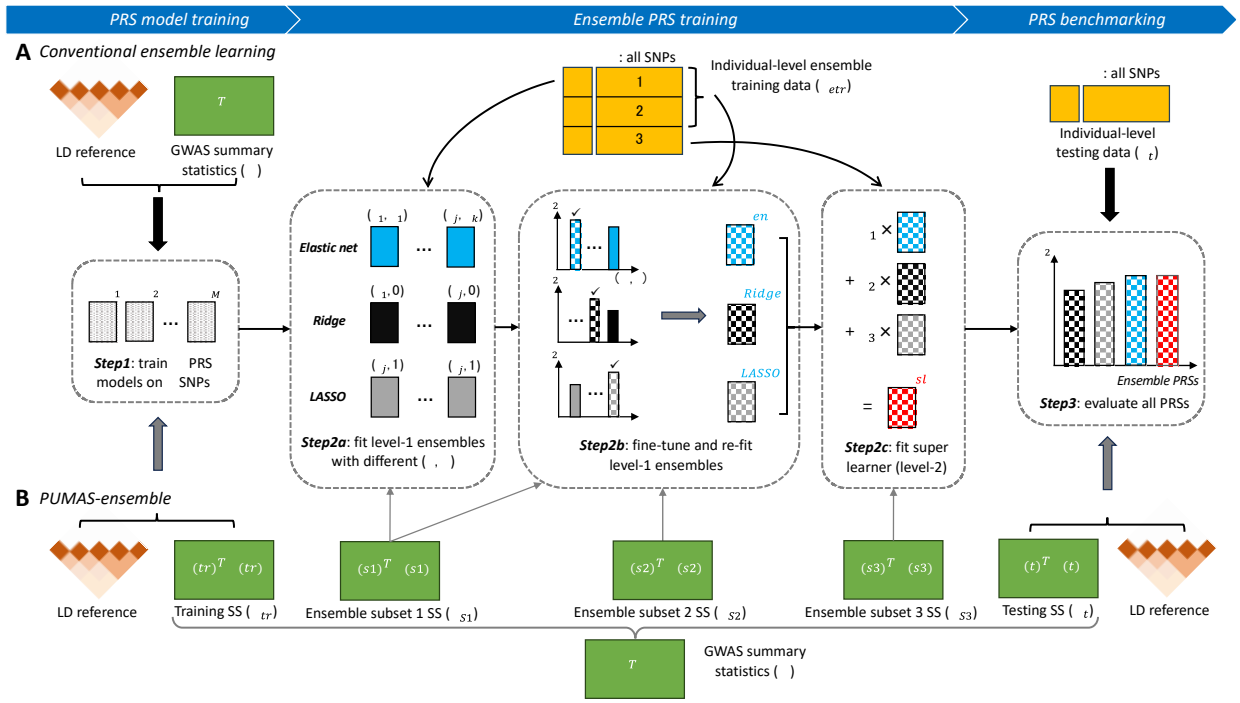

Figure 1. Workflow of summary-statistics-based ensemble learning. (**A**) Conventional ensemble learning approaches require individual-level genotype and phenotype data (orange boxes) to train ensemble learning models and evaluate predictive performance. (**B**) The proposed PUMAS-ensemble approach can follow the same procedure without the need for individual-level data. It leverages a resampling strategy to partition the full GWAS summary statistics into multiple sub-sampled summary datasets for different analytical aims.

Simulation studies

We randomly selected 100,000 independent individuals of European descent in UK Biobank (UKB)[33] and utilized their genotype data on 944,547 HapMap 3 SNPs in our simulation study. To emulate diverse genetic architectures, continuous trait values were generated based on different levels of heritability and different numbers of causal variants (**Methods**). We compared the performance of three ensemble strategies: linear regression, elastic net, and super learning, using either the conventional individual-level data-based ensemble learning and summary-level data-based PUMAS-ensemble. For conventional ensemble learning, we partitioned the full individual-level dataset into three subsets for single PRS training (N_{tr} =60,000), ensemble training (N_{en} =30,000), and testing (N_t =10,000), respectively. Similarly, for PUMAS-ensemble, we performed GWAS to obtain summary statistics based on the full sample, and then subsampled three sets of summary statistics to train single PRS models (N_{tr}=60,000), combine them to obtain the ensemble PRS (N_{en}=30,000), and evaluate performance (N_t=10,000). A total of 110 single PRS models were combined by each ensemble PRS method. For both individual-level and summary-level analyses, we performed 4-fold Monte Carlo cross-validation (MCCV)[29,30] and reported the average predictive R^2.

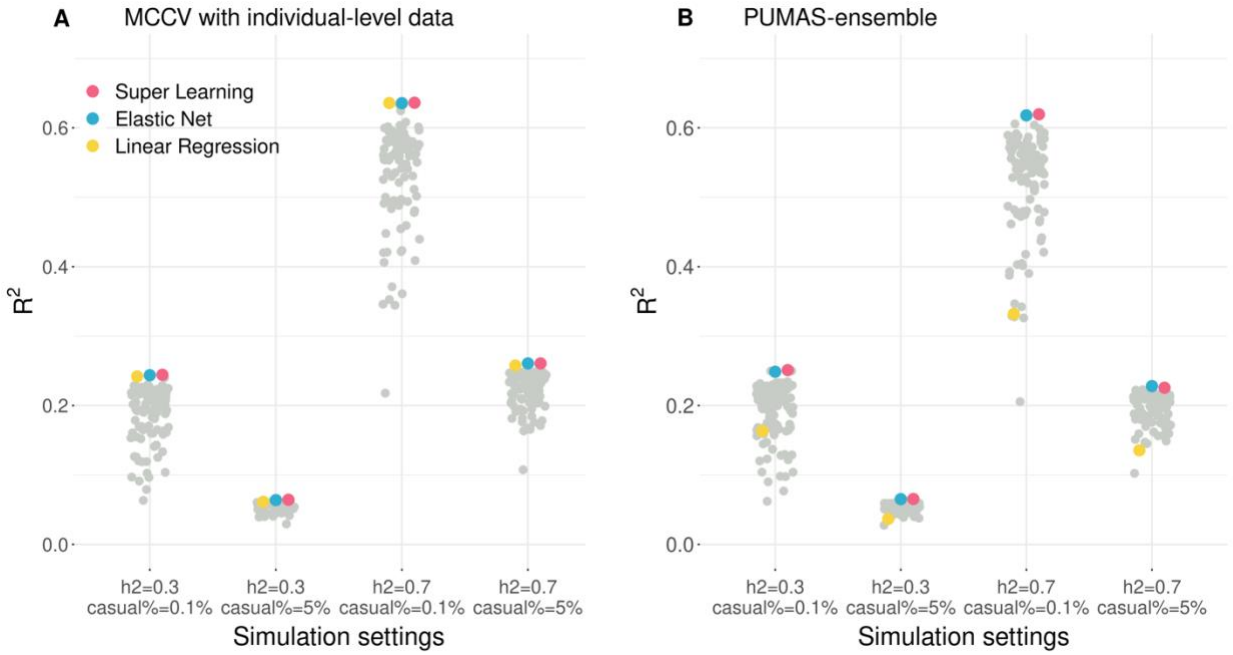

Figure 2. Performance of various ensemble PRS strategies on simulated data. (**A**) Performance of ensemble PRS trained on individual-level data. Prediction accuracy was estimated based on 4-fold MCCV. (**B**) Performance of PUMAS-ensemble PRS quantified by summary-statistics-based 4-fold MCCV. Ensemble PRS models, i.e., linear regression, elastic net, and super learning, are highlighted. Single PRS models are shown in gray. X-axis: simulation settings; Y-axis: predictive R^2; h2: heritability; causal%: proportion of causal variants. Detailed simulation results are summarized in **Supplementary Table 2**.

Both the elastic net-based and super learning-based ensemble PRS approaches (i.e., PUMAS-EN and PUMAS-SL) showed superior predictive performance compared to the single PRS models under all simulation settings (**Figure 2**; **Supplementary Table 2**). Using the median accuracy of single PRS models as the baseline, PUMAS-EN and PUMAS-SL improved predictive R^2 by 11.9%-22.0% and 10.7%-23.2%, respectively, across the various simulation settings, demonstrating robust and nearly identical performance. Notably, the performance of the summary-statistics-based ensemble learning was highly consistent with the performance of the conventional individual-level data-based ensemble learning when utilizing elastic net or super

learning, but not when utilizing linear regression ensemble learning without regularization which tend to be affected by substantial collinearity among PRS models. These results highlight the importance of statistical regularization in summary-statistics-based ensemble learning, especially when aggregating a large number of single PRS models. Because of this observation, we will focus on elastic net and super learning in following analyses.

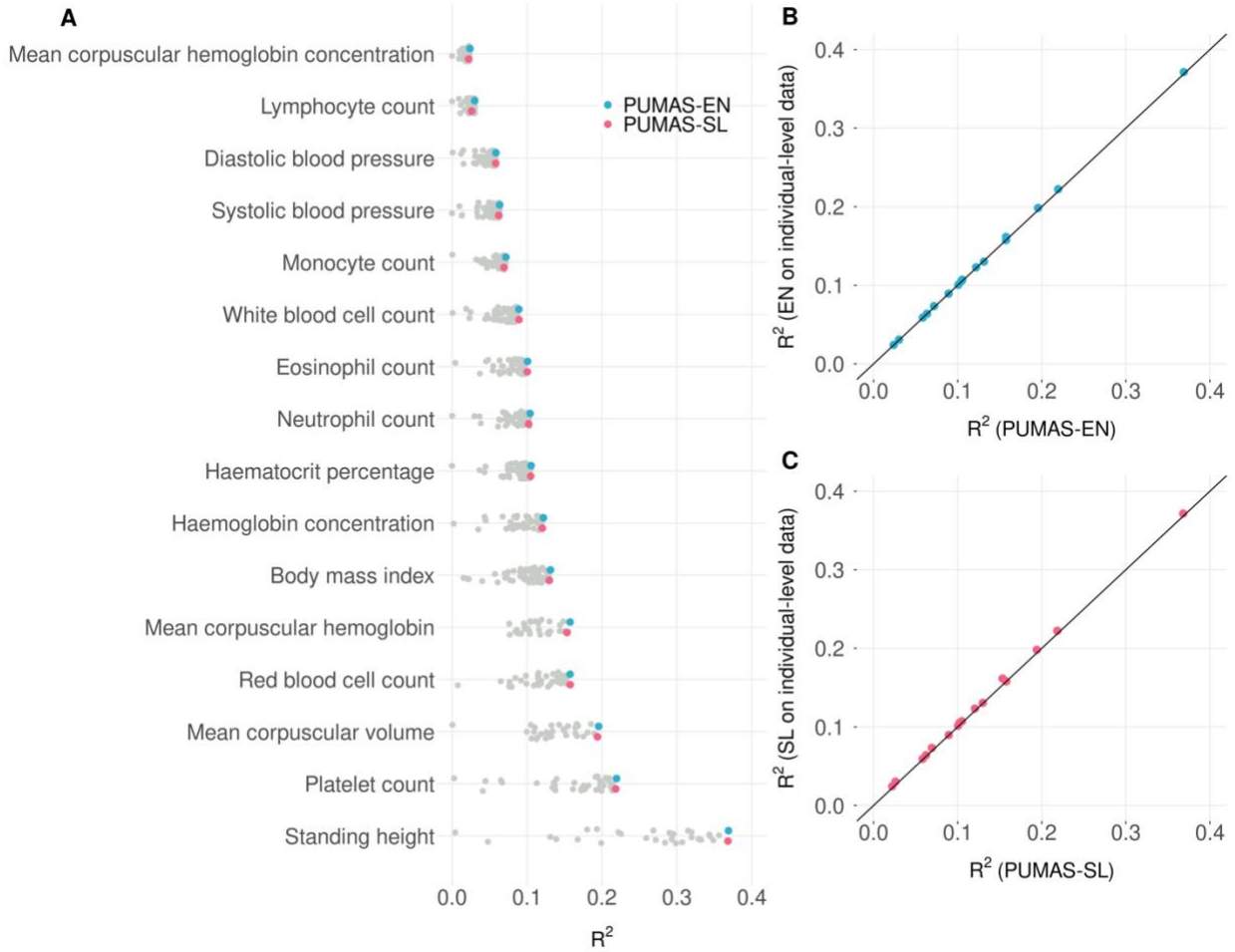

Figure 3. Ensemble PRS prediction for 16 traits in UKB. (**A**) Comparison of PUMAS-ensemble PRS with single PRS models on the holdout UKB dataset (N=38,521). Single PRS models are colored in gray while PUMAS-EN and PUMAS-SL scores are highlighted in blue and red, respectively. Y-axis: trait names; X-axis: predictive R^2. (**B**-**C**) Comparison of summary-statistics-based ensemble PRS with elastic net (panel B) and super learning (panel C) PRS trained on holdout UKB data. Y-axis: R^2 of conventional ensemble PRS; X-axis: R^2 of PUMAS-ensemble PRS. The diagonal line indicates equal R^2 by PUMAS-ensemble PRS and conventional ensemble PRS. Details of the UKB GWAS summary statistics used for PRS training are summarized in **Supplementary Table 3**. Predictive performance of all models and traits is reported in **Supplementary Table 4**.

Ensemble PRS outperforms single PRS models in UKB

Next, we applied PUMAS-ensemble to 16 complex traits in UKB (**Supplementary Table 3**) and compared its prediction accuracy to the conventional ensemble learning strategy based on individual-level data. Specifically, we trained conventional ensemble PRS on three quarters of a holdout UKB dataset (N=38,521) and benchmarked all PRS models on the remaining quarter of samples (**Methods**). Similar to what we observed in simulations, both elastic net and super learning ensemble approaches showed consistently higher predictive R^2 compared to single PRS

models (**Figure 3A**; **Supplementary Table 4**). Elastic net and super learning respectively achieved 20.3% and 17.7% average increase in predictive R^2 across 16 traits compared to the median performance of the 110 single PRS models. Compared to the tuning-free PRS models such as LDpred2-auto[6] and PRS-CS-auto[5], PUMAS-EN had a 10.0% and 16.5% average increase in R^2, respectively, across the 16 traits. Similarly, PUMAS-SL had a 7.6% and 13.8% average increase in R^2, respectively. This suggests that the substantial performance gain of PUMAS-ensemble is achieved by ensemble learning of multiple PRS models rather than improved fine-tuning of single PRS models. In addition, the performance of PUMAS-ensemble PRS was almost indistinguishable from the ensemble PRS trained based on individual-level data (**Figure 3B**-**C**). However, in practice, individual-level holdout datasets often have smaller sample sizes when they do exist. When we reduced the size of individual-level ensemble training data to N=500, PUMAS-EN and PUMAS-SL outperformed elastic net and super learning by an average R^2 increase of 4.8% and 7.2%, respectively (**Methods**; **Supplementary Table 4**). These findings suggest that GWAS summary-level data alone is sufficient for building powerful ensemble PRS in real-world applications, and can even outperform conventional ensemble learning when the individual-level holdout dataset has limited sample size.

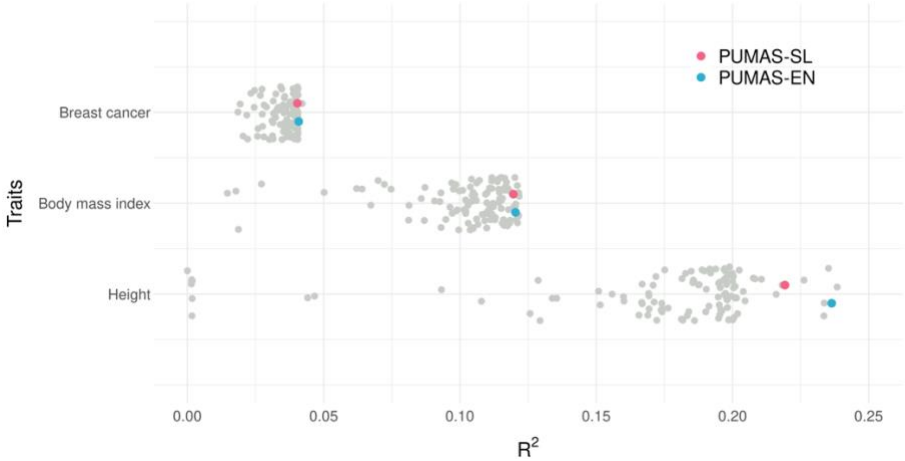

Figure 4. Evaluation of PUMAS-ensemble PRS in AllofUs. Predictive performance of PUMAS-EN, PUMAS-SL, and single PRS models are shown in blue, red, and gray, respectively. Y-axis: trait names; X-axis: predictive R^2. Detailed results are summarized in **Supplementary Table 6**.

PUMAS-ensemble PRS demonstrates robust out-of-sample performance in AllofUs

Additionally, we investigated the out-of-sample performance of ensemble PRS. We applied PUMAS-ensemble to build PRS for height, body mass index, and breast cancer using well-powered and publicly available GWAS summary-level datasets[34-36] and evaluated their prediction accuracy in the All of Us Research Program[37] (AllofUs) (**Methods**; **Supplementary Table 5**). We compared PUMAS-EN, PUMAS-SL, and single PRS models on independent AllofUs participants of European ancestry and reported prediction R^2 (liability scale for breast cancer). We observed consistent performance between PUMAS-EN and PUMAS-SL (**Figure 4**; **Supplementary Table 6**). Elastic net ensemble PRS improved prediction accuracy by 24.2%, 9.8%, and 7.7% respectively for height, body mass index, and breast cancer, compared to the median R^2 of the 110 single PRS models. Super learning showed similar R^2 gains of 23.3%, 8.3%, and 6.5%, respectively.

Figure 5. Cross-ancestral performance of PUMAS-ensemble PRS in AllofUs. (A) Workflow for cross-ancestry PUMAS-ensemble application. **(B-E)** We trained PUMAS-EN and PUMAS-SL PRS models for **(B)** height, **(C)** body mass index, **(D)** diastolic blood pressure, and **(E)** systolic blood pressure, combining single EUR PRS and EAS PRS, i.e., PRS models trained based on either EUR GWAS from UKB or EAS GWAS from BBJ. Model performance was evaluated on EAS participants in AllofUs. The R^2 distribution of single PRS models trained based on EUR and EAS GWAS are shown in yellow and green, respectively. R^2 of PUMAS-EN and PUMAS-SL are highlighted as blue and red dashed lines. Y-axis: number of PRS models; X-axis: predictive R^2. Full results for all models and traits are reported in **Supplementary Table 8**.

Cross-ancestry ensemble PRS improves polygenic prediction accuracy on East Asian individuals in AllofUs

Next, we showcased the benefit of employing our ensemble PRS strategy in a cross-ancestral risk prediction setting. We extended PUMAS-ensemble to construct optimal ensemble scores for participants of East Asian (EAS) descent in AllofUs by aggregating PRS models trained using European (EUR) GWAS summary data from UKB and EAS GWAS summary data from Biobank Japan (BBJ)[38-40] (**Methods**; **Figure 5A**). We then compared PUMAS-EN and PUMAS-SL with the single EUR and EAS PRS models on four traits that are well-powered in both UKB GWAS and BBJ GWAS: height, body mass index, diastolic blood pressure, and systolic blood pressure (**Supplementary Table 7**). The ensemble approach generates the best-performing PRS model

for all traits analyzed, with markedly improved prediction accuracy comparing to single-ancestry PRS models (**Figure 5B-E**; **Supplementary Table 8**). Across the four traits, predictive R^2 of PUMAS-EN is 60.7% and 230.1% higher on average than the median R^2 of the various EAS and EUR PRS models (58.8% and 227.0% for PUMAS-SL), respectively. Between the two proposed ensemble strategies, PUMAS-EN and PUMAS-SL have quite consistent performance with similar average R^2 across the four traits.

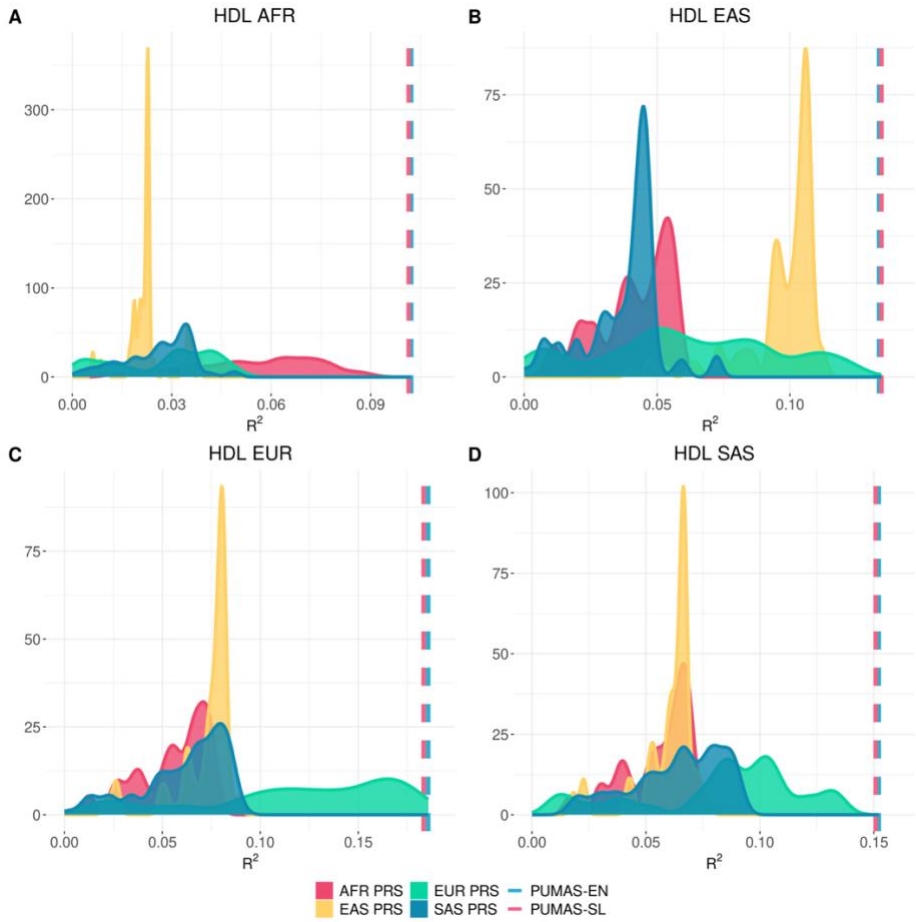

Figure 6. Cross-ancestry performance of PUMAS-ensemble PRS on HDL cholesterol. (A-D) We trained PUMAS-EN and PUMAS-SL PRS models for HDL cholesterol for ancestries **(A)** AFR, **(B)** EAS, **(C)** EUR, and **(D)** SAS, combining single PRS models trained based on ancestry-specific GWAS from GLGC. Model performance was evaluated on the highlighted ancestries participants in UKB. The R^2 distribution of single PRS models trained based on AFR, EAS, EUR, and SAS ancestries are highlighted in pink, yellow, green, and blue, respectively. The R^2 of PUMAS-EN and PUMAS-SL PRS are marked by dashed lines. Y-axis: number of PRS models; X-axis: predictive R^2. Full results for all models and traits are reported in **Supplementary Table 10**.

Cross-ancestry ensemble PRS of blood lipid traits for multiple ancestries

Finally, we further evaluated the performance of our ensemble PRS method on blood lipid traits across four ancestries on validation individuals from UKB. We utilized ancestry-stratified GWAS summary data from the Global Lipids Genetics Consortium[41] (GLGC) for high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), log-transformed triglycerides (logTG), and total cholesterol (TC), across four ancestries, including African (AFR), EAS, EUR, and South Asian (SAS). For each ancestry, we used PUMAS-ensemble to generate scores for individuals of a matching ancestry in UKB by aggregating models trained using subsampled GLGC summary data from a matching ancestry and full GLGC summary data from the other three ancestries

(**Supplementary Figure 1**). The consortium also provides data for Admixed American and Hispanic/Latino population (AMR); however, we did not include them in our analysis because two traits were flagged during our GWAS QC process and there is limited sample size of AMR ancestry individuals in UKB (**Methods, Supplementary Table 9**). We compared PUMAS-EN and PUMAS-SL with PRS methods of single ancestry (**Figure 6** and **Supplementary Table 10**). For all traits, the ensemble approach outperforms the median R^2 of single PRS models of the same ancestry by an average of 305.6% for PUMAS-EN (median of 55.7%) and 314.3% for PUMAS-SL (median of 59.5%). Compared to the median R^2 of single PRS models of European ancestry, PUMAS-EN outperforms by an average of 128.2% (median of 101.4%) and PUMAS-SL by 129.2% (median of 105.7%) (**Supplementary Table 11; Supplementary Figure 2**). This highlights the ability of our summary data-based ensemble PRS approach in cross-ancestry applications to improve over single PRS models and to leverage the large sample size of European GWAS to enhance prediction accuracy in diverse non-European populations**.**

Discussion

Ensemble learning can effectively combine multiple PRS models and improve risk prediction accuracy, but it is a data-demanding task that is often impossible to implement in practice due to the lack of adequately large individual-level holdout datasets. In this study, we introduced two summary-statistics-based ensemble learning approaches based on elastic net and super learning under the PUMAS-ensemble framework. Our proposed approaches employ statistical regularization to allow adaptive integration of a large number of single PRS models that may be highly correlated. We demonstrate that PUMAS-ensemble PRS closely approximates the ensemble PRS trained based on individual-level holdout data and show its superior performance compared to single PRS models in both within-ancestry and cross-ancestry applications.

Our work brings several key advancements to the field. First, PUMAS-ensemble is the only method in the literature that performs PRS ensemble learning on GWAS summary statistics, bypassing the stringent data requirement in existing approaches and fully exploiting the widely available summary-level GWAS data resources without compromising predictive performance. Second, our approach employs sophisticated regularization, allowing researchers to combine possibly hundreds of PRS models without acquiring additional holdout samples. Importantly, this strategy can build ensemble scores for non-European ancestries by combining a variety of ancestry-specific PRS models. While cross-ancestry ensemble learning has proven effective in improving upon single PRS models across several recent studies, existing strategies[21-23,26,27] rely on non-European data at the individual level which can be close to impossible to obtain in practice. Our approach removes this critical constraint in data requirement which is a major step towards reducing disparity in genomic medicine[42,43]. Third, PRS method development is a crowded research field. When a new PRS method is introduced, it is common to see incremental gains in predictive accuracy over existing approaches. Our method shows a consistent 10-20% R^2 improvement in within-ancestry applications and an R^2 gain of as high as 300% in cross-ancestry predictions. This is a substantial improvement. Further, perhaps the most important feature of our approach is its ability to continuously evolve. We did not introduce just another PRS model. This is a powerful framework that can always incorporate everyone's favorite PRS models, including future models once they become available. This highlights summary statistics-based ensemble learning as a crucial direction for future PRS development, and is also why we believe we may have found the "one score to rule them all". Every future PRS method should consider this strategy to combine the cumulative wisdom from many existing models with new methodological innovations. Summary statistics-based ensemble learning is the core technique that makes this possible.

There are still some important future directions. In this study, we included several PRS methods for illustration. Additional models researchers can consider in their analysis include methods that introduce new statistical designs[28,44], employ non-parametric modeling[45,46], leverage multi-trait integration[12-14,20,47,48], or incorporate biological information[10,11,49]. We also did not consider non-linear models for either PRS construction or ensemble learning[20,21,50]. Additionally, it would be meaningful work to extend all recent multi-ancestry PRS methods that use ensemble learning on holdout samples[21-23,26,27] to the summary statistics-based version using PUMAS-ensemble. Finally, it remains an open question how LD mismatch[44], population admixture[51], and ancestry continuum[52] should be modeled.

In conclusion, our study presents a highly innovative and data-efficient statistical framework for PRS ensemble learning. We highlight its capability of combining, and thus surpassing, all existing (and future) PRS models. PUMAS-ensemble is a versatile tool that can bring immediate benefits to the many applications of PRS which will no doubt greatly facilitate future studies.

Methods

An overview of summary-statistics-based PRS ensemble learning

Conventionally, PRS ensemble learning requires a summary-level GWAS dataset for single PRS model training and an independent individual-level dataset for ensemble model training. If the ensemble model contains tuning parameters, such as regularization parameters in penalized regressions, the individual-level dataset needs to be first partitioned for model tuning and evaluation, and eventually combined again for fitting the best ensemble model (e.g., training-validation split or cross-validation). Such a procedure is straightforward when the required individual-level dataset exists. In practice, however, it is much more common that only GWAS summary statistics are available. Therefore, we extend a flexible summary-statistics-based cross-validation approach we previously introduced[30] to train and evaluate elastic net and super learning ensemble PRS models using only GWAS summary statistics and LD reference data as inputs. We first provide an overview of the PUMAS-ensemble framework. Under an additive genetic model, the relationship between a trait Y and genotype $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p)$ can be quantified as:

$$Y = \mathbf{X}\boldsymbol{\beta} + \epsilon,$$

where p is the number of SNPs, $\boldsymbol{\beta} \in \mathbb{R}^p$ denote true SNP effects, and ϵ denotes distributed random error terms independent from $\mathbf{X}$ with zero mean and finite and positive variance σ_e^2. We then define ensemble PRS as a linear combination of many PRS models:

$$Z_{ensemble} = \mathbf{Zb} = \mathbf{Xwb},$$

where $\mathbf{b} = [b_1\ b_2 \dots\ b_M]^T$ is the vector of ensemble weights for M PRS models, $\mathbf{Z} = [Z_1\ Z_2 \dots Z_M] = \mathbf{Xw}$ is the PRS matrix of M models with corresponding SNP weights $\mathbf{w} = [\mathbf{w}_1\ \mathbf{w}_2 \dots\ \mathbf{w}_M]$. The key objective is to obtain the ensemble weights $\mathbf{b}$ for optimal PRS performance.

Assume that we have GWAS summary-level data of sample size N and external reference genotype data for LD estimation, in order to train and evaluate ensemble PRS models, we need to partition the full GWAS summary statistics into three subsets for PRS model training, ensemble learning, and benchmarking, respectively. We have shown in earlier work[29,30] that given the observed summary-level data $\mathbf{x}^T\mathbf{y}$ from the full GWAS dataset, the conditional distribution of summary statistics for a subset of GWAS samples $\mathbf{x}^{(s)^T}\mathbf{y}^{(s)}$ with sample size $N^{(s)} < N$ is:

$$\mathbf{x}^{(s)^T}\mathbf{y}^{(s)} | \mathbf{x}^T\mathbf{y} \sim \mathbf{N}\left(\frac{N^{(s)}}{N}\mathbf{x}^T\mathbf{y}, \frac{\left(N - N^{(s)}\right)N^{(s)}}{N}\widehat{\boldsymbol{\Sigma}}\right),$$

$$\mathbf{x}^T\mathbf{y} = \left[\mathbf{x}_1^T\mathbf{y} \dots \mathbf{x}_p^T\mathbf{y}\right]^{\mathbf{T}} = \left[N\hat{\beta}_1\hat{\sigma}_1^2 \dots N\hat{\beta}_p\hat{\sigma}_p^2\right]^{\mathbf{T}},$$

where $\widehat{\mathbf{\Sigma}}$ is the observed covariance matrix of $\mathbf{x}^T\mathbf{y}$, and $\hat{\beta}_j$ and $\hat{\sigma}_j^2$ are the marginal effect size estimate and MAF-based variance estimator, respectively, for SNP j, $j = 1,2,\dots,p$. We have shown that $\widehat{\mathbf{\Sigma}}$ can be obtained based on GWAS summary statistics and LD reference data, and an iterative subsampling scheme can be used to partition the full summary statistics into three independent subsets of summary statistics[30]:

$$\mathbf{x}^{(t)^T}\mathbf{y}^{(t)}|\mathbf{x}^T\mathbf{y} \sim \mathbf{N}\left(\frac{N^{(t)}}{N}\mathbf{x}^T\mathbf{y}, \frac{\left(N - N^{(t)}\right)N^{(t)}}{N}\widehat{\mathbf{\Sigma}}\right),$$

$$\mathbf{x}^{(etr)^T}\mathbf{y}^{(etr)}|\mathbf{x}^T\mathbf{y} - \mathbf{x}^{(t)^T}\mathbf{y}^{(t)} \sim \mathbf{N}\left(\frac{N^{(etr)}}{N - N^{(t)}}\left(\mathbf{x}^T\mathbf{y} - \mathbf{x}^{(t)^T}\mathbf{y}^{(t)}\right), \frac{\left(N - N^{(t)} - N^{(etr)}\right)N^{(etr)}}{N - N^{(t)}}\widehat{\mathbf{\Sigma}}\right),$$

$$\mathbf{x}^{(tr)^T}\mathbf{y}^{(tr)} = \mathbf{x}^T\mathbf{y} - \mathbf{x}^{(t)^T}\mathbf{y}^{(t)} - \mathbf{x}^{(etr)^T}\mathbf{y}^{(etr)},$$

where $\mathbf{x}^{(tr)^T}\mathbf{y}^{(tr)}$, $\mathbf{x}^{(etr)^T}\mathbf{y}^{(etr)}$, and $\mathbf{x}^{(t)^T}\mathbf{y}^{(t)}$ correspond to the summary statistics for PRS model training, ensemble learning, and benchmarking (testing), respectively. Details of PRS model fitting and training of ensemble models are described in later sections.

Once we obtain the ensemble weights $\mathbf{b}$, we can evaluate prediction accuracy of the ensemble model by calculating its predictive R^2 on the testing summary-level data as[30]:

$$R^2_{ensemble} = \frac{\left[\frac{1}{N^{(t)}}\mathbf{b}^T\mathbf{w}^T\mathbf{x}^{(t)^T}\mathbf{y}^{(t)}\right]^2}{N \max_j\left(SE\left(\hat{\beta}_j\right)^2\hat{\sigma}_j^2\right)\mathbf{b}^T\widehat{\mathbf{\Sigma}}_{\mathbf{Z}}\mathbf{b}},$$

where $\widehat{\mathbf{\Sigma}}_{\mathbf{Z}}$ is the estimated covariance matrix of M PRS models, which can be approximated using the LD reference data. Finally, we repeat the above procedures K times (i.e., K-fold MCCV) to ensure robust performance of ensemble PRS. Note that if the goal is to evaluate ensemble PRS without accessing external, individual-level validation datasets, PUMAS-ensemble can report the average prediction accuracy of ensemble PRS as $\bar{R}^2_{ensemble}$ across K folds. When the analytical aim is to produce ensemble PRS for maximal out-of-sample prediction accuracy, $\mathbf{x}^{(t)^T}\mathbf{y}^{(t)}$ and $\mathbf{x}^{(etr)^T}\mathbf{y}^{(etr)}$ can be combined to calculate SNP weights for the ensemble model as $\mathbf{w}_{ensemble} = \mathbf{w}\bar{\mathbf{b}}$, where $\bar{\mathbf{b}}$ is the vector of average ensemble weights across K folds. In the following two sections, we will introduce the summary-statistics-based ensemble model fitting based on elastic net and super learning.

PUMAS-EN: elastic net ensemble PRS based on summary statistics

Our earlier work has shown that combining multiple fine-tuned PRS models with linear regression can lead to better predictive performance than single PRS[30]. While classic linear regression serves as a proof-of-concept example for ensemble PRS, it is often of interest to aggregate as many PRS models as possible for achieving maximal gain in prediction accuracy. However, as demonstrated in our simulation study, summary-statistics-based linear regression becomes highly unstable and hinders performance of ensemble score when many highly correlated PRS models are included. To address multicollinearity and further improve ensemble score, we introduce PUMAS-EN which adaptively integrates a large number of PRS models via elastic net[53] using GWAS summary statistics. Another advantage of this elastic net model is that it relieves PUMAS-ensemble from conducting fine-tuning for each PRS method prior to ensemble learning and can directly combine all PRS across various methods and tuning parameter settings.

PUMAS-EN has two tuning parameters, λ and α, where λ controls the overall shrinkage of each PRS's coefficient in the ensemble model and α allocates the relative contribution of L1 and L2

penalty terms. To obtain elastic net coefficient estimates for $\mathbf{b}$, the ensemble weights for M single PRS models, PUMAS-EN minimizes the following objective function:

$$\mathbf{b}_{\mathbf{e}n} = \underset{\mathbf{b}}{\operatorname{argmin}} \frac{1}{2n} \|\mathbf{y} - \mathbf{xwb}\|^2 + \alpha\lambda\|\mathbf{b}\|_1 + \frac{1-\alpha}{2}\lambda\|\mathbf{b}\|^2,$$

where n is sample size of elastic net training data, and $\mathbf{xw}$ are standardized PRS with mean of 0 and variance of 1. Following derivations from lassosum[4] and recent PRS frameworks[12,22,23] that utilized penalized regression, we show that $\mathbf{b}_{\mathbf{e}n}$ can be estimated using only GWAS summary statistics and an external LD reference data. Since the objective function is not continuously convex, we use the coordinate descent[54] algorithm to iteratively update ensemble weight for PRS model m, $m = 1,2,\ldots,M$:

$$\frac{\partial \mathcal{L}}{\partial b_m} = -\frac{1}{n}\mathbf{w}_{,m}^T\mathbf{x}^T\mathbf{y} + \frac{1}{n}\mathbf{w}_{,m}^T\mathbf{x}^T\mathbf{x}\mathbf{w}_{,m}b_m + \frac{1}{n}\mathbf{w}_{,m}^T\mathbf{x}^T\left(\mathbf{xwb} - \mathbf{xw}_{,m}b_m\right) + \lambda(1-\alpha)b_m + \lambda\alpha\frac{b_m}{|b_m|},$$

where $\mathbf{w}_{,m}$ is the m-th column of $\mathbf{w}$ that represents SNP weights in the m-th PRS model. By setting $\frac{\partial \mathcal{L}}{\partial b_m}$ to zero, we obtain the formula for iteratively updating b_m:

$$b_m^{(t+1)} = \begin{cases} \dfrac{sign\left(\frac{1}{n}\mathbf{w}_{,m}^T\mathbf{x}^T\mathbf{y} - u_m\right)\left(\left|\frac{1}{n}\mathbf{w}_{,m}^T\mathbf{x}^T\mathbf{y} - u_m\right| - \lambda\alpha\right)}{\frac{1}{n}\mathbf{w}_{,m}^T\mathbf{x}^T\mathbf{x}\mathbf{w}_{,m} + \lambda(1-\alpha)} & , \quad \text{if} \left|\frac{1}{n}w_{,m}^T x^T y - u_m\right| > \lambda\alpha \\ 0 & , \quad \text{if} \left|\frac{1}{n}w_{,m}^T x^T y - u_m\right| \le \lambda\alpha \end{cases},$$

where $u_m = \frac{1}{n}\mathbf{w}_{,m}^T\mathbf{x}^T\left(\mathbf{xwb}^{(t)} - \mathbf{xw}_{,m}b_m^{(t)}\right)$ utilizes the PRS weights estimated from the t-th iteration $b_m^{(t)}$.

It is clear that only the terms $\mathbf{x}^T\mathbf{y}$ and $\mathbf{x}^T\mathbf{x}$ are required for fitting elastic net ensemble model. To fine-tune hyperparameters λ and α, we further partition the ensemble training summary statistics $\mathbf{x}^{(etr)^T}\mathbf{y}^{(etr)}$ into two independent subsets. We fit multiple elastic net models with different combinations of α and λ on the first subset, select the "optimal" combination based on their performance on the second subset, and eventually train the fine-tuned elastic net model with the selected hyperparameter values on the entire ensemble training summary statistics $\mathbf{x}^{(etr)^T}\mathbf{y}^{(etr)}$. Throughout this study, we consider hyperparameter settings $\alpha = 0, 0.25, 0.5, 0.75, 1$ and $\lambda = 10^{\psi}$ where ψ includes 51 numbers evenly spaced in $[-5,0]$. For model fitting, we initialize $b_m^{(0)}$ at zero and iteratively update $b_m^{(t)}$ until the algorithm converges (i.e., $max\left|b_m^{(t)} - b_m^{(t-1)}\right| < 0.001$) or when reaching the maximum number of iterations (default: 10^4).

PUMAS-SL: super learning PRS based on summary statistics

Super learning[24] is essentially a two-level ensemble learning approach that trains an optimal weighted combination of multiple machine learning models such as elastic net[53], ridge regression[55], and LASSO[56]. It can be used to train an "ensemble of ensemble" model combining a large set of baseline PRS models for optimized polygenic risk prediction. All existing super learning strategies for PRS training require individual-level data as input. Here, we introduce PUMAS-SL, which trains a super learning model using only GWAS summary statistics.

First, we define super learning PRS as a linear combination of level-one ensemble PRS models:

$$Z_{sl} = \mathbf{XwB\gamma}$$

where $\mathbf{B} = [\mathbf{b}_{en}\ \mathbf{b}_{ridge}\ \mathbf{b}_{LASSO}]$ denotes the matrix of level-one ensemble weights, i.e., the weights of various PRS models in different ensemble models (elastic net, ridge regression, LASSO), and

$\boldsymbol{\gamma} = [\gamma_{en}\ \gamma_{ridge}\ \gamma_{LASSO}]^T$ denotes the level-two ensemble weights for the various ensemble models in the super learning PRS. Both $\mathbf{B}$ and $\boldsymbol{\gamma}$ are parameters of interests that need to be estimated. Using our subsampling scheme for summary statistics, we partition the ensemble training summary statistics $\mathbf{x}^{(etr)^T}\mathbf{y}^{(etr)}$ into two independent subsets denoted by $\mathbf{x}^{(s1)^T}\mathbf{y}^{(s1)}$ and $\mathbf{x}^{(s2)^T}\mathbf{y}^{(s2)}$ to train $\mathbf{B}$ and $\boldsymbol{\gamma}$, respectively. We fit the level-one ensemble models including ridge regression, LASSO, and elastic net on the first subset:

$$\mathbf{b}_{ridge} = \left(\mathbf{w}^T\mathbf{x}^{(s1)^T}\mathbf{x}^{(s1)}\mathbf{w} + \lambda\mathbf{I}_M\right)^{-1}\mathbf{w}^T\mathbf{x}^{(s1)^T}\mathbf{y}^{(s1)},$$

where $\mathbf{b}_{LASSO}$ and $\mathbf{b}_{en}$ can be estimated using the coordinate descent algorithm introduced earlier. To improve the stability of the super learning model, we further apply a K-fold summary-statistics-based MCCV on $\mathbf{x}^{(s1)^T}\mathbf{y}^{(s1)}$ and divide it into two independent subsets for training and fine-tuning level-one ensemble models, respectively. We then combine elastic net, ridge, and LASSO ensemble models through a level-two linear regression on $\mathbf{x}^{(s2)^T}\mathbf{y}^{(s2)}$ from the second subset:

$$\boldsymbol{\gamma} = \left(\mathbf{B}^T\mathbf{w}^T\mathbf{x}^{(s2)^T}\mathbf{x}^{(s2)}\mathbf{w}\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{w}^T\mathbf{x}^{(s2)^T}\mathbf{y}^{(s2)}.$$

Taking $\mathbf{B}$ and $\boldsymbol{\gamma}$ together, we can obtain a super learning PRS model based only on GWAS summary statistics. For out-of-sample prediction, PUMAS-SL outputs SNP weights for super learning PRS model as $\mathbf{w}_{sl} = \mathbf{w}\bar{\mathbf{B}}\bar{\boldsymbol{\gamma}}$, where $\bar{\mathbf{B}}$ and $\bar{\boldsymbol{\gamma}}$ are estimated level-one and level-two ensemble weights averaged across K-fold MCCV.

Details of PRS model training

We trained and combined PRS based on eight methods including lassosum[4], LDpred2[6], PRS-CS[5], MegaPRS[8], SBayesR[7], DBSLMM[9], Vilma[31], SBLUP[32] throughout this study. Single PRS models were trained separately on each chromosome in parallel, except for the ones trained by MegaPRS and lassosum. We constructed LD reference data for PRS methods that do not provide such datasets throughout this study. For EUR (i.e., simulation, UKB, and AllofUs) and EAS PRS analyses (i.e., AllofUs), we used a UKB genotype dataset consisting of N=1000 randomly selected individuals of EUR ancestry and another UKB EAS genotype dataset (N=500) as the corresponding LD reference datasets, respectively. When training ancestry-specific PRS for blood lipid traits in UKB, we randomly picked 500 samples from the UKB testing dataset of matched ancestry to generate LD data for each of EUR, AFR, EAS, and SAS populations. We used HapMap 3 SNPs in all our analyses. For the rest of this section, we outline and briefly introduce each single PRS method considered in our study.

Lassosum estimates LASSO coefficients by jointly modeling SNPs in LD using estimated marginal SNP effect sizes from GWAS summary statistics. Lassosum has two tuning parameters s and λ, where s regulates the sparsity of LD blocks and λ is the regularization parameter in LASSO that shrinks SNP effects towards zero. We trained lassosum models with $s = 0.2, 0.5, 0.9$ and $\lambda = 0.005, 0.01$ using the R package 'lassosum' (v0.4.5).

LDpred2 is a Bayesian PRS method that adaptively shrinks SNP effects while accounting for LD. LDpred2 employs two versions assuming two distinct prior distributions for SNP effects: LDpred2-inf, which is a tuning-free model based on the 'infinitesimal model', and LDpred2-grid, which assumes a spike-and-slab prior distribution with hyperparameters p representing the true proportion of causal variants and total heritability h^2. In addition, LDpred2-auto is an empirical Bayes approach that avoids the need for hyperparameter tuning by estimating p and h^2 along with other parameters during model fitting[6]. We included LDpred2-grid and LDpred2-auto models in ensemble PRS in all our analyses. We trained LDpred2 models using the R package 'bigsnpr' (v1.9.11) with $p = 0.001, 0.01, 0.1$ and low heritability $0.1 \cdot h^2, 0.3 \cdot h^2$, where h^2 is the heritability

estimated by LD-score regression[57], both sparse and non-sparse models for LDpred2-grid, and more stringent LD shrinkage (shrink_corr = 0.5) for LDpred2-auto. We adopted both lower heritability value and stronger LD shrinkage to improve LDpred2 model convergence following recent improvement made to LDpred2[58].

PRS-CS places a continuous shrinkage prior on SNP effect size distribution, as opposed to the spike-and-slab prior in LDpred2. It includes a global shrinkage factor ϕ which uniformly shrinks SNP effects throughout the genome. Alternatively, ϕ can be adaptively learned from the GWAS data by a fully Bayesian approach (PRS-CS-auto). We fitted PRS-CS models using 1000 Genomes Project EUR LD matrices in simulation and UKB LD matrices for real data analysis. All LD reference data were provided by the PRS-CS software (v1.0.0).

SBayesR employs a mixture of point mass at zero and three normal distributions with different variance parameters as the prior distribution for SNP effects, representing SNP effect sizes of different magnitudes. SBayesR does not require hyperparameter tuning because all hyperparameter values are pre-specified. We fitted SBayesR models using the GCTB software (v2.04.3) and the sparse UKB LD matrices for HapMap 3 SNPs provided by GCTB. SBayesR was not included in EAS data analysis in AllofUs due to the lack of LD matrices for EAS.

Vilma is another recently developed Bayesian approach with a more flexible normal mixture prior than SBayesR. It can be applied to model summary statistics from multiple traits and different genetic ancestries. The number of component normal distributions in its mixture prior is the only tuning parameter in Vilma; like SBayesR, it recommends a default value (i.e., 81) for this parameter. We fitted Vilma models using its software provided on GitHub.

MegaPRS is a flexible Bayesian framework that can employ multiple prior specifications such as LASSO, ridge, Bolt (i.e., a mixture of two Gaussian distributions), and BayesR (i.e., a mixture of three Gaussian distributions and a point mass). We fitted MegaPRS models using the LDAK software (v5.2) with the recommended BayesR prior specification which include 84 pairs of tuning parameters that determine the relative weights of component Gaussian distributions. LDAK-thin[59] was used for per-predictor heritability estimation. To improve robustness of ensemble score, we only included MegaPRS models with no greater than 10 predictors that failed to converge.

DBSLMM first conducts LD clumping and thresholding to partition SNPs into a large-effect group and a small-effect group. It then fits a linear mixed effects model (i.e., large fixed effects and small random effects) to obtain updated SNP weights using summary statistics while accounting for LD. DBSLMM is a computationally efficient approach that has one tuning parameter, the p-value threshold in LD clumping and thresholding. We fitted DBSLMM models using fine-tuned p-value cutoff determined by summary-statistics-based parameter tuning implemented in the DBSLMM software (v0.3).

SBLUP bases its framework on a linear mixed model where SNP effects are assumed to be random and normally distributed, thus making it conceptually equivalently to LDpred-inf. SBLUP uses GWAS summary statistics as input and does not have hyperparameters. We trained SBLUP model using the GCTA software (v1.93.0).

Simulation studies based on UKB genotype data

We conducted simulations using UKB genotype data imputed to the Haplotype Reference Consortium panel. We kept samples of European ancestry and removed genetic variants with MAF below 0.01, imputation R^2 below 0.9, Hardy-Weinberg equilibrium test p-value below 1e-6,

or missing genotype call rate greater than 2%. We further extracted variants in the HapMap 3 SNP list and the LD reference data for European ancestry from Phase 3 of the 1000 Genomes Project. 377,509 samples and 944,547 variants remained after quality control. Then, we randomly selected 100,000 samples to form the simulation dataset with their genotype and randomly selected 1,000 samples to form the LD reference dataset. To generate phenotypic values, we simulated SNP effect sizes from a spike-and-slab distribution, i.e., $\beta_j \sim (1-p)\delta_0 + p\mathrm{N}(0, \frac{h^2}{Mp})$, as assumed in LDpred2, where p is the proportion of causal variants, δ_0 denotes a point mass at 0, h^2 is the total heritability of the phenotype, and M is the total number of SNPs. We considered four simulation settings with distinct genetic architectures and heritability by setting p to 0.1% or 5% and h^2 to 0.3 or 0.7. To simulate trait values, we randomly selected causal variants across the genome, computed the "true PRS" by aggregating SNP allele counts weighted by true effect sizes, and added gaussian noises scaled according to trait heritability. We fitted marginal linear regressions using PLINK[60] to obtain GWAS summary statistics in each simulation setting.

We compared ensemble PRS constructed by PUMAS-ensemble and by 4-fold MCCV using individual-level data. To implement 4-fold MCCV, in each fold, we randomly selected 60% of the samples as the training dataset (N=60,000), 30% as the ensemble training dataset (N=30,000), and the remaining 10% as the testing dataset (N=10,000). We fitted GWAS and PRS models on the training data, calculated ensemble weights from elastic net and super learning on the ensemble training data, and finally, evaluated PRS model performance on the testing data. For elastic net, we divided ensemble training data into two halves and fitted elastic net models using the R package 'glmnet' (v4.1)[54] with a prespecified set of tuning parameters on the first subset. Then we benchmarked the performance of each model on the second subset and re-trained the most predictive elastic net model on the entire ensemble training data. We used the R package 'SuperLearner' (v2.0)[24] to train super learning PRS on the ensemble training dataset. As a comparison, we applied PUMAS-ensemble to implement 4-fold summary-statistics-based MCCV. We partitioned the full summary-level data into three independent sets of summary statistics for PRS training (N=60,000), ensemble learning (N=30,000), and PRS benchmarking (N=10,000), respectively. For PUMAS-EN, we trained fine-tuned ensemble model by dividing the ensemble training summary statistics into two halves and following the PUMAS-EN model fitting strategy we introduced earlier. For PUMAS-SL, we partitioned the ensemble training summary-level data to two subsets for training level-one (2/3; N=40,000) and level-two (1/3; N=20,000) ensemble weights, respectively. An additional 50%-50% training-testing split on the level-one ensemble training subset was applied to train fine-tuned LASSO, ridge, and elastic net regression models. Throughout simulation, we used European LD reference data from Phase 3 of the 1000 Genomes Project[61] to subsample summary statistics and evaluate PRS performance. The holdout UKB LD genotype data (N=1,000) was used as the LD reference for PRS model training, except for PRS-CS[5] and SBayesR[7], where software-provided UKB LD matrices were used. Finally, for both approaches, we calculated and reported average R^2 for each PRS model across 4 folds.

Evaluating ensemble PRS in UKB

We compared the ensemble PRS constructed by PUMAS-ensemble with those constructed by training-testing split based on individual-level data for 16 quantitative phenotypes on UKB individuals of EUR descent. The list of UKB phenotypes and detailed sample size information are summarized in **Supplementary Table 3**. We reserved approximately 10% of the UKB samples with non-missing phenotypic values (N=38,521) as the holdout set for individual-level ensemble PRS training and PRS benchmarking. We then obtain GWAS summary statistics for each trait by performing linear regression analysis adjusting for sex, age polynomials to the power of two, interactions between sex and age polynomials, and top 20 genetic principal components via Hail

(v0.2.57)[62] on the remaining samples. A smaller subset of the holdout data (N=1,000) was used as LD reference for PRS model training. Prior to evaluating each PRS model, we regressed out covariate effects from phenotypes in the holdout dataset. Predictive R^2 was reported for each PRS model.

To fit ensemble PRS based on individual-level data, we randomly partitioned the holdout dataset into two subsets for ensemble model training (3/4 of samples) and PRS benchmarking (1/4 of samples), respectively. We trained single PRS methods based on full GWAS summary statistics, fitted elastic net and super learning PRS on the ensemble training subset, and calculated predictive R^2 for all PRS on the benchmarking subset. We trained fine-tuned elastic net model and super learning model on the ensemble training subset following the same protocols described in our simulation study but without cross-validation. For comparison, we used PUMAS-ensemble to build ensemble PRS models by 4-fold MCCV. Within each fold of MCCV, we partitioned full GWAS summary statistics into training (70%) and ensemble training (30%) summary statistics. Model fitting for PUMAS-EN and PUMAS-SL follows the exact procedures described in PUMAS-ensemble simulation. Then, we computed and evaluated PUMAS-ensemble PRS on the benchmarking subset. The same subset of UKB holdout data (N=1,000) was used for subsampling summary statistics and ensemble model training for PUMAS-ensemble. Finally, we sought to compare PUMAS-ensemble with conventional ensemble learning under a common scenario where the individual-level data for ensemble model training is limited. To mimic this real-world setting, we randomly selected 500 samples from the ensemble model training subset to train elastic net and super learning ensemble PRS models and assessed their performance on the same PRS benchmarking subset.

External validation of PUMAS-ensemble PRS in AllofUs

We compared PUMAS-EN and PUMAS-SL with single PRS models for height, body mass index, and breast cancer on AllofUs data[37]. AllofUs is a nationally representative cohort in the USA with whole-genome-sequencing (WGS) data. We kept independent samples of European descent. Genetic ancestry inferred from the principal components and sample relatedness were provided in AllofUs. We removed samples with extreme phenotypic values and nonbinary biological sex. We computed BMI as weight/(height^2) after extracting standing height in meters and body weight in kilograms. If a participant had multiple entries for a given phenotype, only the latest measurement was kept. Detailed phenotypic information for these three traits is summarized in **Supplementary Table 5**.

We trained ensemble PRS models using publicly available GWAS summary statistics for height (N=1,597,374)[35], body mass index (N=795,640)[36], and breast cancer (N=247,173)[34]. For breast cancer, we transformed logistic regression association statistics to linear scale before subsampling summary statistics[29]. For each trait, we partitioned the full summary statistics into two subsets for PRS training (70%) and ensemble model fitting (30%), respectively. SNP weights for elastic net and super learning PRS were obtained following the same analytical procedures used in our simulation study. We computed PRS for AllofUs samples using genotype data for HapMap 3 SNPs extracted from WGS data via PLINK1.9[60]. To evaluate PRS, we computed R^2 on the observed scale for height and body mass index and R^2 on the liability scale for breast cancer by adjusting for its case-control ratio in AllofUs[63]. Covariates including biological sex, age polynomials to the power of two, interactions between sex and age polynomials, and top 16 genetic principal components were regressed out from both the phenotype and PRS in advance. All PRS calculation and evaluation were conducted in the AllofUs cloud analysis environment using the v7 data release.

Cross-ancestry ensemble PRS for EAS samples in AllofUs

We benchmarked PUMAS-ensemble against single PRS models based on GWAS summary statistics of European and East Asian ancestral populations, respectively, on East Asian samples in AllofUs[37]. We considered four continuous traits including height, body mass index, diastolic blood pressure, and systolic blood pressure and used independent EAS samples based on genetic principal components and genetic relatedness for PRS evaluation. Samples with extreme phenotypic values and nonbinary biological sex were removed from the analysis. Body mass index was imputed for EAS samples in AllofUs following the same procedures described in the previous section. Detailed data information is summarized in **Supplementary Table 7**.

We trained PUMAS-EN and PUMAS-SL using published EAS GWAS from BBJ[38-40] and in-house EUR GWAS from UKB[30]. Sample size for each BBJ and UKB GWAS summary-level dataset is summarized in **Supplementary Table 7**. We fitted EUR PRS models using various PRS methods based on the full UKB GWAS summary statistics. Then, we applied PUMAS-ensemble to partition full BBJ summary statistics into two subsets for EAS PRS training (70%) and ensemble model fitting (30%), respectively. We used the same UKB EUR LD reference data (N=1,000) in UKB data analysis for EUR PRS training and a random subset of UKB EAS samples (N=500) for EAS PRS model fitting. The assignment of EAS ancestry for non-European UKB participants was described in our earlier work[26]. SNP weights for PUMAS-EN and PUMAS-SL were obtained based on the same procedures outlined in our simulation study. We computed and evaluated EAS PRS, EUR PRS, PUMAS-EN PRS, and PUMAS-SL PRS models on the testing dataset. The same set of covariates considered in the AllofUs data analysis in the previous section have also been adjusted for prior to computing PRS R^2.

Cross-ancestry ensemble PRS of blood lipid traits for four ancestries in UKB

We compared PUMAS-ensemble against single PRS models based on GWAS summary statistics from GLGC of AFR, EAS, EUR, and SAS ancestral populations[41] on independent samples in UKB. We considered four blood lipid traits HDL, LDL, logTG, and TC and used independent UKB samples with genetically predicted AFR, EAS, EUR, and SAS ancestries[22]. GLGC additionally provides GWAS summary statistics of AMR ancestry, which we excluded due to our standard QC pipeline flagging sample size issues on chromosomes 13 through 22 for both LDL and logTG traits. An additional reason for excluding AMR samples was their limited samples in the independent UKB data.

For each ancestry and trait pair, we trained PUMAS-EN and PUMAS-SL using ancestry-specific GLGC summary statistics. Sample size for the GLGC summary statistics for each ancestry and trait pair is available in **Supplementary Table 9**. For each ancestry and trait, we applied PUMAS-ensemble to partition full summary statistics into two subsets for PRS training (70%) and ensemble model fitting (30%). We fitted single PRS models for the given trait for all ancestries using the same PRS methods described in previous sections which we later include in the ensemble learning. We used UKB LD reference data from a matching ancestry for each ancestry PRS training and model fitting. PUMAS-EN and PUMAS-SL SNP weights were obtained with similar procedures as the simulation study. Covariates including biological sex, age polynomials to the power of two, interactions between sex and age polynomials, and the top 10 genetic principal components were regressed out from both the phenotype and PRS.

Code availability

PUMAS-ensemble software is freely available at https://github.com/qlu-lab/PUMAS.

Competing interests

The authors declare no competing interests.

Author Contributions

Z.Z. and Q.L. conceived and designed the study.
Z.Z. developed the statistical framework.
Z.Z. and S.D. performed the statistical analysis.
Y.W. and S.D. performed AllofUs data analysis.
X.Y. assisted in UKB simulation analysis.
Z.Z. and S.D. implemented the software.
Q.L. and J.J. advised on statistical and genetic issues.
Z.Z., S.D., and Q.L. wrote the manuscript.
All authors contributed to manuscript editing and approved the manuscript.

Acknowledgment

The authors gratefully acknowledge research support from National Institutes of Health (NIH) grant R21 AG085162, and support from the University of Wisconsin-Madison Office of the Chancellor and the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation (WARF). This research has been conducted using the UK Biobank Resource under Applications 42148 and 17731. This study makes use of summary statistics from GWAS consortia. We thank GWAS investigators for providing publicly accessible GWAS summary statistics. This research uses data from AllofUs. The All of Us Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2 OD025276. In addition, the All of Us Research Program would not be possible without the partnership of its participants.

References

1. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9 (2010).
2. International Schizophrenia, C. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748-52 (2009).
3. Vilhjálmsson, B.J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet* **97**, 576-92 (2015).
4. Mak, T.S.H., Porsch, R.M., Choi, S.W., Zhou, X. & Sham, P.C. Polygenic scores via penalized regression on summary statistics. *Genet Epidemiol* **41**, 469-480 (2017).
5. Ge, T., Chen, C.Y., Ni, Y., Feng, Y.A. & Smoller, J.W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun* **10**, 1776 (2019).
6. Privé, F., Arbel, J. & Vilhjálmsson, B.J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424-5431 (2020).
7. Lloyd-Jones, L.R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat Commun* **10**, 5086 (2019).
8. Zhang, Q., Privé, F., Vilhjálmsson, B. & Speed, D. Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nature Communications* **12**, 4192 (2021).
9. Yang, S. & Zhou, X. Accurate and Scalable Construction of Polygenic Scores in Large Biobank Data Sets. *Am J Hum Genet* **106**, 679-693 (2020).
10. Hu, Y. *et al.* Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Comput Biol* **13**, e1005589 (2017).
11. Márquez-Luna, C. *et al.* Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *Nature Communications* **12**, 6052 (2021).
12. Chen, T.-H., Chatterjee, N., Landi, M.T. & Shi, J. A penalized regression framework for building polygenic risk models based on summary statistics from genome-wide association studies and incorporating external information. *Journal of the American Statistical Association*, 1-19 (2020).
13. Hu, Y. *et al.* Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLoS Genet* **13**, e1006836 (2017).
14. Maier, R.M. *et al.* Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nat Commun* **9**, 989 (2018).
15. Monti, R. *et al.* Evaluation of polygenic scoring methods in five biobanks shows larger variation between biobanks than methods and finds benefits of ensemble learning. *Am J Hum Genet* **111**, 1431-1447 (2024).
16. Wang, Y., Tsuo, K., Kanai, M., Neale, B.M. & Martin, A.R. Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. *Annu Rev Biomed Data Sci* **5**, 293-320 (2022).
17. Ni, G. *et al.* A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts. *Biol Psychiatry* **90**, 611-620 (2021).
18. Pain, O. *et al.* Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLOS Genetics* **17**, e1009021 (2021).
19. Yang, S. & Zhou, X. PGS-server: accuracy, robustness and transferability of polygenic score methods for biobank scale studies. *Brief Bioinform* **23**(2022).

20. Albiñana, C. *et al.* Multi-PGS enhances polygenic prediction by combining 937 polygenic scores. *Nature Communications* **14**, 4702 (2023).
21. Zhang, H. *et al.* A new method for multiancestry polygenic prediction improves performance across diverse populations. *Nature Genetics* **55**, 1757-1768 (2023).
22. Zhang, J. *et al.* An ensemble penalized regression method for multi-ancestry polygenic risk prediction. *Nature Communications* **15**, 3238 (2024).
23. Jin, J. *et al.* MUSSEL: Enhanced Bayesian polygenic risk prediction leveraging information across multiple ancestry groups. *Cell Genomics* **4**(2024).
24. van der Laan, M.J., Polley, E.C. & Hubbard, A.E. Super learner. *Stat Appl Genet Mol Biol* **6**, Article25 (2007).
25. Naimi, A.I. & Balzer, L.B. Stacked generalization: an introduction to super learning. *Eur J Epidemiol* **33**, 459-464 (2018).
26. Miao, J. *et al.* Quantifying portable genetic effects and improving cross-ancestry genetic prediction with GWAS summary statistics. *Nat Commun* **14**, 832 (2023).
27. Ruan, Y. *et al.* Improving polygenic prediction in ancestrally diverse populations. *Nature Genetics* **54**, 573-580 (2022).
28. Chen, T., Zhang, H., Mazumder, R. & Lin, X. Fast and scalable ensemble learning method for versatile polygenic risk prediction. *Proceedings of the National Academy of Sciences* **121**, e2403210121 (2024).
29. Zhao, Z. *et al.* PUMAS: fine-tuning polygenic risk scores with GWAS summary statistics. *Genome Biol* **22**, 257 (2021).
30. Zhao, Z. *et al.* Optimizing and benchmarking polygenic risk scores with GWAS summary statistics. *Genome Biology* **25**, 260 (2024).
31. Spence, J.P., Sinnott-Armstrong, N., Assimes, T.L. & Pritchard, J.K. A flexible modeling and inference framework for estimating variant effect sizes from GWAS summary statistics. *bioRxiv*, 2022.04.18.488696 (2022).
32. Robinson, M.R. *et al.* Genetic evidence of assortative mating in humans. *Nature Human Behaviour* **1**, 0016 (2017).
33. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
34. Zhang, H. *et al.* Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat Genet* **52**, 572-581 (2020).
35. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* (2022).
36. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641-3649 (2018).
37. Bick, A.G. *et al.* Genomic data in the All of Us Research Program. *Nature* **627**, 340-346 (2024).
38. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nature Genetics* **50**, 390-400 (2018).
39. Akiyama, M. *et al.* Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nature Genetics* **49**, 1458-1467 (2017).
40. Akiyama, M. *et al.* Characterizing rare and low-frequency height-associated variants in the Japanese population. *Nature Communications* **10**, 4393 (2019).

41. Graham, S.E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675-679 (2021).
42. Martin, A.R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am J Hum Genet* **100**, 635-649 (2017).
43. Martin, A.R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics* **51**, 584-591 (2019).
44. Dun, Y., Chatterjee, N., Jin, J. & Nishimura, A. A Robust Bayesian Method for Building Polygenic Risk Scores using Projected Summary Statistics and Bridge Prior. *arXiv preprint arXiv:2401.15014* (2024).
45. Zhou, G. & Zhao, H. A fast and robust Bayesian nonparametric method for prediction of complex traits using summary statistics. *PLoS Genet* **17**, e1009697 (2021).
46. Chun, S. *et al.* Non-parametric Polygenic Risk Prediction via Partitioned GWAS Summary Statistics. *Am J Hum Genet* **107**, 46-59 (2020).
47. Xu, C., Ganesh, S.K. & Zhou, X. mtPGS: Leverage multiple correlated traits for accurate polygenic score construction. *The American Journal of Human Genetics* **110**, 1673-1689 (2023).
48. Truong, B. *et al.* Integrative polygenic risk score improves the prediction accuracy of complex traits and diseases. *Cell Genomics* **4**(2024).
49. Zheng, Z. *et al.* Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. *Nature Genetics* **56**, 767-777 (2024).
50. Elgart, M. *et al.* Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. *Communications Biology* **5**, 856 (2022).
51. Sun, Q. *et al.* Improving polygenic risk prediction in admixed populations by explicitly modeling ancestral-differential effects via GAUDI. *Nature communications* **15**, 1016 (2024).
52. Ruan, Y. *et al.* Leveraging genetic ancestry continuum information to interpolate PRS for admixed populations. *medRxiv*, 2024.11. 09.24316996 (2024).
53. Zou, H. & Hastie, T. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **67**, 301-320 (2005).
54. Friedman, J.H., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**, 1 - 22 (2010).
55. Hoerl, A.E. & Kennard, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55-67 (1970).
56. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267-288 (1996).
57. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
58. Privé, F., Arbel, J., Aschard, H. & Vilhjálmsson, B.J. Identifying and correcting for misspecifications in GWAS summary statistics and polygenic scores. *Human Genetics and Genomics Advances* **3**, 100136 (2022).
59. Speed, D., Hemani, G., Johnson, M.R. & Balding, D.J. Improved heritability estimation from genome-wide SNPs. *American journal of human genetics* **91**, 1011-1021 (2012).
60. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).

61. McVean, G.A. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
62. Poterba, T. *et al.* The Scalable Variant Call Representation: Enabling Genetic Analysis Beyond One Million Genomes. *bioRxiv*, 2024.01.09.574205 (2024).
63. Lee, S.H., Goddard, M.E., Wray, N.R. & Visscher, P.M. A Better Coefficient of Determination for Genetic Profile Analysis. *Genetic Epidemiology* **36**, 214-224 (2012).