Optimizing and benchmarking polygenic risk scores with GWAS summary statistics

Zijie Zhao[1], Tim Gruenloh[1], Meiyi Yan[2], Yixuan Wu[1], Zhongxuan Sun[1], Jiacheng Miao[1], Yuchang Wu[1,3], Jie Song[2], Qiongshi Lu[1,2,3,#]

[1] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, WI

[2] Department of Statistics, University of Wisconsin-Madison, WI

[3] Center for Demography of Health and Aging, University of Wisconsin-Madison, Madison, WI

To whom correspondence should be addressed:

Dr. Qiongshi Lu (qlu@biostat.wisc.edu)

Keywords: genome-wide association study; GWAS summary statistics; polygenic risk score; PRS model-tuning; PRS benchmark; ensemble learning

Abstract

Background

Polygenic risk score (PRS) is a major research topic in human genetics. However, a significant gap exists between PRS methodology and applications in practice due to often unavailable individual-level data for various PRS tasks including model fine-tuning, benchmarking, and ensemble learning.

Results

We introduce an innovative statistical framework to optimize and benchmark PRS models using summary statistics of genome-wide association studies. This framework builds upon our previous work and can fine-tune virtually all existing PRS models while accounting for linkage disequilibrium. In addition, we provide an ensemble learning strategy named PUMAS-ensemble to combine multiple PRS models into an ensemble score without requiring external data for model fitting. Through extensive simulations and analysis of many complex traits in the UK Biobank, we demonstrate that this approach closely approximates gold-standard analytical strategies based on external validation, and substantially outperforms state-of-the-art PRS methods.

Conclusions

Our method is a powerful and general modeling technique that can continue to combine the best-performing PRS methods out there through ensemble learning and could become an integral component for all future PRS applications.

Background

Genetic risk prediction is a main focus in human genetics research and a key step towards precision medicine(1-3). Continued success in genome-wide association studies (GWAS) in the past decade has facilitated the development of polygenic risk scores (PRS) that aggregate the effects of millions of single nucleotide polymorphisms (SNPs) for many complex traits(4-6). Compared to earlier statistical methods that require individual-level data for model training(7-10), PRS which only relies on GWAS summary data is much more generally applicable due to the wide availability of GWAS summary statistics. Although earlier PRS models struggled to produce accurate prediction results, recent and more sophisticated PRS methods have achieved substantially improved prediction accuracy through statistical regularization and biological data integration(11-17). In numerous studies, PRS has shown promising performance in stratifying disease risk and great potential in informing early lifestyle changes or medical interventions(18-21).

Despite the progress, several lingering challenges create a significant gap between PRS methodology and applications. A main recurring issue we highlight (and address) throughout the paper is that PRS modelers often assume the existence of independent individual-level datasets that can be used for additional model tuning. But in practice, GWAS summary statistics are used for PRS model training, meaning that conventional sample splitting schemes cannot be used. Additional datasets that are independent from both training and testing samples also rarely exist. This suggests that model-tuning samples will have to come from the precious testing dataset which inevitably reduces the sample size and statistical power in downstream applications.

This disconnection between impractical method requirements and limited data availability can lead to a variety of problems. For example, many PRS methods have tuning parameters that could substantially swing model performance when not chosen properly(12-15, 22-24). Conventionally, these parameters need to be fine-tuned on a separate dataset with individual-level genotypes and phenotypes. Although some recent methods employ fully Bayesian or empirical Bayesian techniques to bypass model fine-tuning(25-27), these hyperparameter-free PRS do not always outperform fine-tuned models, trading predictive accuracy for computational feasibility(28, 29). Second, no PRS method universally outperforms all other approaches. The empirical performance of a PRS model depends on GWAS sample size, genetic architecture of the phenotype, quality of GWAS summary statistics, and heterogeneity between training and testing samples(30-33). Thus, it is of great interest to systematically and impartially benchmark various PRS methods for each trait, ideally in an independent dataset(11, 30, 34). Third, several recent studies have employed ensemble learning which combines multiple PRS models via another regression(28, 29) and showed improved PRS accuracy in both within- and cross-ancestry prediction applications(35-37). This brute-force approach has shown superior performance compared to any single PRS method but is data-demanding – the second level regression model needs to be fit on a separate dataset. Finally, we note that it may be of interest to combine all these tasks in practice, e.g., benchmarking an ensemble learner that combines multiple PRS models which all need to be tuned separately, which really seems like an impossible task.

In this paper, we seek a solution to these problems. We base our statistical framework on PUMAS, a method we recently introduced to perform Monte Carlo cross-validation (MCCV) using GWAS summary statistics(38). We have shown that PUMAS can effectively fine-tune PRS models with

clumped SNPs(39) and similar approaches have since been adopted in other applications(40-42). Here, we first demonstrate that PUMAS can now fine-tune and benchmark state-of-the-art PRS models without SNP pruning. Second, we introduce an extension to the PUMAS framework named PUMAS-ensemble which is an innovative strategy to perform ensemble learning using GWAS summary data alone. Taken together, we showcase a sophisticated statistical framework for fine-tuning, benchmarking, and combining PRS models using GWAS summary statistics as input. We demonstrate the performance of our approach through extensive simulations and analysis of 21 complex traits in UK Biobank (UKB). On average, the PUMAS-ensemble ensemble PRS achieves a 8.93% relative gain in predictive R^2 compared to LDpred2-auto and a 17.68% gain compared to PRS-CS-auto, respectively. We also apply our method to 31 well-powered GWAS with publicly available summary statistics and provide a catalog of ensemble PRS with benchmarked predictive performance.

Results

Method Overview

First, we present an overview of the PUMAS-ensemble workflow. Statistical details and technical discussions are presented in the **Methods** section. For illustration, first we assume individual-level data is available. In this case, we would divide the samples into 4 independent sets for PRS training, model fine-tuning, constructing ensemble PRS, and benchmarking model performance,

respectively (**Figure 1A**). The main goal of our new approach is to mimic this procedure when only summary statistics are available. Using PUMAS, we could sample marginal association statistics for a subset of individuals in the GWAS(38). Doing this repeatedly, we could divide the full GWAS summary data to corresponding training, tuning, ensemble learning, and testing summary statistics (**Figure 1B**). Using these four sets of sub-sampled summary statistics, we train a series of PRS models, fine-tune each PRS model to select the besting tuning parameters, apply PUMAS-ensemble to combine PRS models through linear regression, and finally evaluate the predictive performance of PRS models. The entire procedure only requires GWAS summary statistics and linkage disequilibrium (LD) references as input.

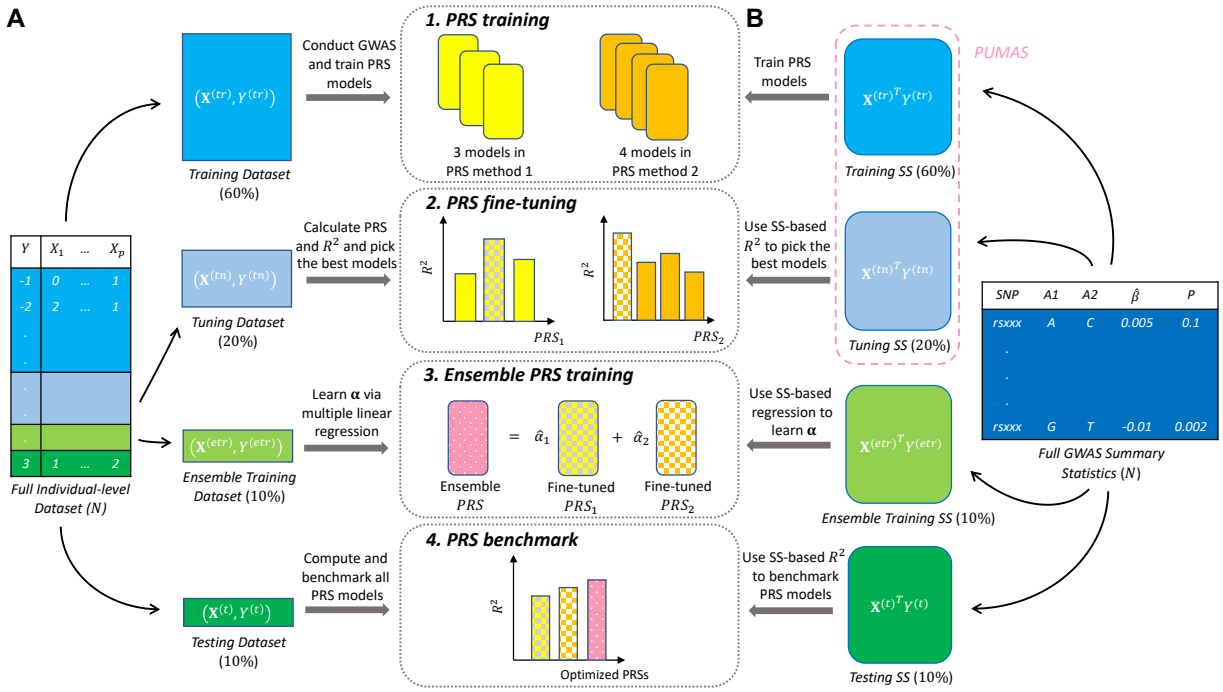

Figure 1. Workflow of PRS construction and evaluation. (**A**) Conventional approach divides the entire individual-level dataset to different subset of samples for each of 4 stages of PRS analysis. (**B**) PUMAS-ensemble directly partitions the full summary-level data to corresponding summary statistics for different analytical purposes.

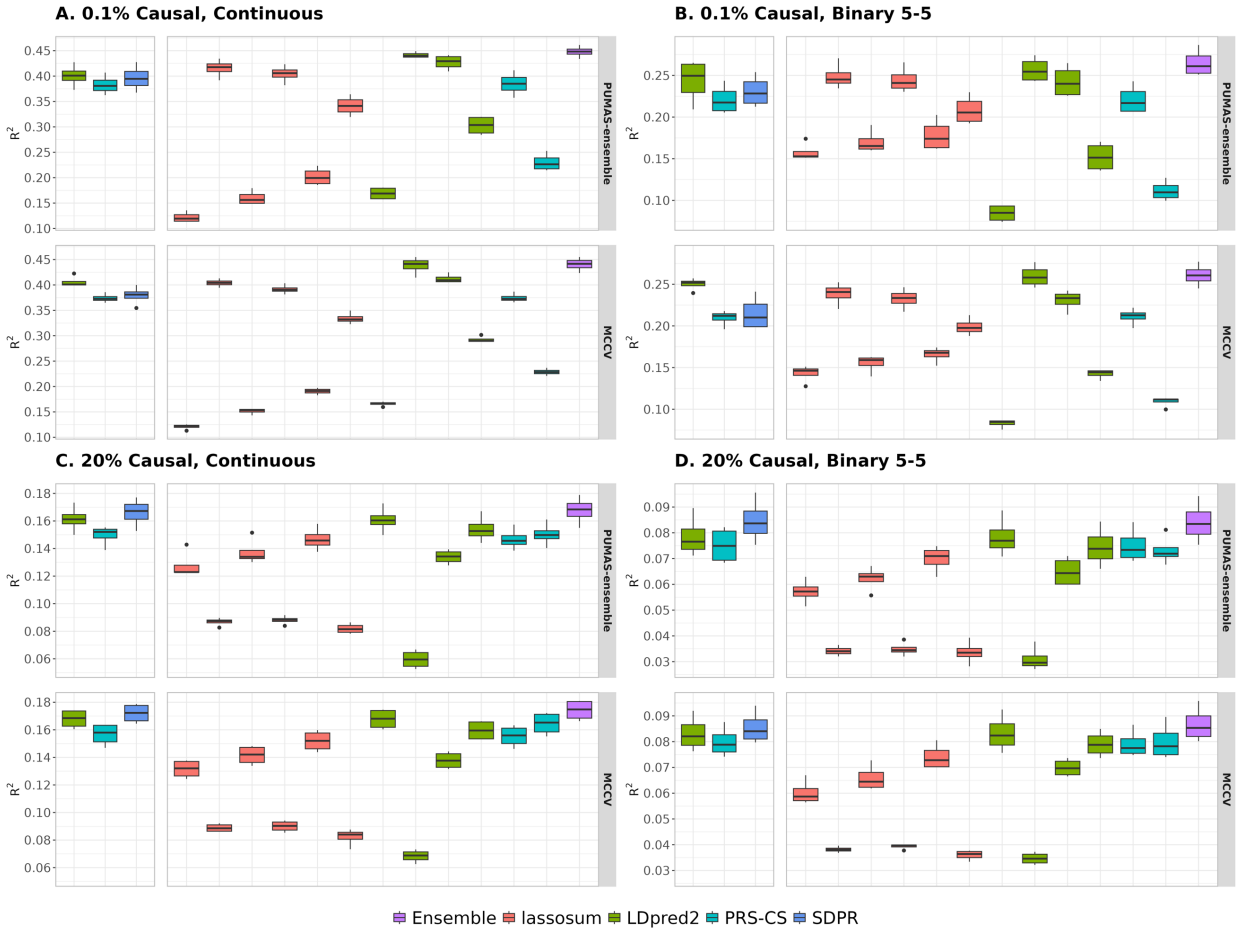

Figure 2. Comparison of PUMAS-ensemble and MCCV in UKB simulation. (**A** and **C**) Simulation results for quantitative traits. (**B** and **D**) Simulation results for binary traits with balanced case-control ratio. Proportion of causal variants is 0.1% in **A** and **B**, and 20% in **C** and **D**. The heritability is set to be 0.5 in all panels. Models that do not require fine-tuning are shown on the left side of each panel. Y-axis: predictive R^2 across 4 repeats of MCCV; X-axis (left to right): tuning-free models: LDpred2-auto (green box), PRS-CS-auto (blue box), and SDPR (dark-blue box). lassosum models (red boxes) with tuning parameter settings: s=0.2 and λ=0.005, s=0.2 and λ=0.01, s=0.5 and λ=0.005, s=0.5 and λ=0.01, s=0.9 and λ=0.005, s=0.9 and λ=0.01. LDpred2 models (green boxes): non-infinitesimal with p=0.1, non-infinitesimal with p=0.01, non-infinitesimal with p=0.001, and infinitesimal model. PRS-CS (blue boxes): ϕ=0.01 and 0.0001.

Finally, the purple box shows the results of ensemble PRS. Results for remaining simulation settings are summarized in **Additional file 1**: **Fig. S1**-**S10** and **Additional file 2**, **3**, **4**, **5**, **6**, and **7**: **Table. S1**-**S6**.

Simulation results

We performed simulations using imputed genotype data from UKB to demonstrate that PUMAS and PUMAS-ensemble can fine-tune, combine, and benchmark PRS models. We included 100,000 independent individuals of European descent and 944,547 HapMap3 SNPs in the analysis. We simulated phenotypes with heritability of 0.2, 0.5, and 0.8 and randomly assigned causal variants under sparse and polygenic settings to mimic different types of genetic architecture (**Methods**). We performed GWAS and obtained marginal association statistics. We then implemented PUMAS and PUMAS-ensemble to conduct a 4-fold MCCV to train, optimize, and evaluate lassosum, PRS-CS, LDpred2, and an ensemble PRS which combines all three methods and SDPR(22, 25, 26, 43). For comparison, we also implemented a MCCV procedure using individual-level UKB data. We partitioned the UKB dataset into 4 mutually exclusive datasets. We used datasets 1 and 2 to train and fine-tune each PRS method, then used the third dataset to fit a regression to combine multiple PRS. We evaluated each PRS method in the fourth dataset and reported PRS prediction accuracy quantified by R^2. We describe implementation details of both summary-statistics-based and individual-level-data-based MCCV in **Methods**.

Overall, we observed highly consistent results between PUMAS/ PUMAS-ensemble and MCCV for both quantitative and binary phenotypes (**Figure 2**; **Additional file 1**: **Fig. S1**-**S7**; **Additional file 2**, **3**, **4**, and **5**: **Table. S1**-**S4**). In addition, summary statistics-based approaches can closely

approximate R^2 values obtained from model-tuning and benchmarking techniques using individual-level data. PUMAS-ensemble also constructed scores that were highly concordant with ensemble PRS built from individual-level data which universally outperformed all PRS models used as input. During the revision, we also added simulation analysis for MegaPRS(40) which also yields similar ensemble score performance (**Additional file 1**: **Fig. S8**-**S10**; **Additional file 6** and **7**: **Table. S5**-**S6**). Computation-wise, PUMAS's subsampling step can execute in parallel for each chromosome and our benchmarking results suggest that PUMAS/PUMAS-ensemble are scalable for GWAS summary statistics including millions of SNPs (**Additional file 8**: **Table. S7**).

We conducted two additional analyses to demonstrate the validity of PUMAS subsampling framework. First, we benchmarked PUMAS against the subsampling approach implemented in MegaPRS(40), which uses 'pseudo summary statistics' for parameter tuning, and the gold standard approach MCCV based on individual-level data (**Methods**). We found consistent subsampling results from both approaches compared to MCCV (**Additional file 1**: **Fig. S11**; **Additional file 9**: **Table. S8**). In addition, while PUMAS's subsampling framework assumes weak individual SNP effects, we observed robust performance of PUMAS under extremely sparse genetic architecture with large SNP effects (**Additional file 1**: **Fig. S12**; **Additional file 10**: **Table. S9**). These results highlight the robustness of PUMAS's summary statistics subsampling scheme under different genetic architecture.

PUMAS can fine-tune and benchmark PRS methods

Next, we demonstrate that PUMAS effectively fine-tunes PRS models and performs accordantly with the gold standard external validation approach based on individual-level data. We applied PUMAS to 16 quantitative traits, 4 diseases, and 1 ordinal trait in UKB(44) (**Additional file 11** and **12**: **Table. S10**-**S11**)**.** After quality control, the UKB dataset contained 375,064 independent individuals and 1,030,187 SNPs (**Methods**). We applied a 9-to-1 data split to hold out 10% of the samples for external validation, and performed GWAS for all traits using 90% of the samples. We applied 4-fold MCCV implemented in PUMAS to train and fine-tune three PRS models (i.e., LDpred2, lassosum, and PRS-CS which have been demonstrated to achieve high prediction accuracy in a recent benchmark study(22, 25, 26, 29)) using only summary statistics. For external validation, we trained PRS models using the full summary statistics and calculated PRS prediction accuracy on the holdout dataset. We report the best tuning parameters for LDpred2, lassosum, and PRS-CS and corresponding R^2 obtained from both PUMAS and external validation.

Our summary-statistics-based approach showed highly consistent model-tuning performance for all analyzed traits compared to external validation (**Figure 3**, **Additional file 1**: **Fig. S13**-**S33**; **Additional file 13** and **14**: **Table. S12**-**S13**). Among 21 traits, PUMAS and external validation selected the same best tuning parameters 21, 18, and 11 times for lassosum, LDpred2, and PRS-CS, respectively. When the model tuning results were different between PUMAS and external validation, both approaches still selected models with very similar prediction accuracy. Indeed, PUMAS provided precise R^2 estimates for all models compared to external validation, advocating the use of our summary-statistics-based approach for PRS model benchmarking. In addition, it is noteworthy that empirical and full Bayesian approaches (i.e., LDpred2-auto and PRS-CS-auto) did not always outperform other fine-tuned PRS models even within the UKB cohort, demonstrating the necessity of PRS model tuning for optimizing out-of-sample prediction.

We also observed that the parameter-tuning results are accordant with the analyzed traits' genetic architecture. For both height and monocyte count, PUMAS accurately selected the best tuning parameters based on external validation (**Figure 3A-B**), but the selected models were not the same between these two traits. Height is known to be extremely polygenic with more than 12,000 independent GWAS signals in the latest GWAS(45). In comparison, fewer loci have been found to significantly associate with monocyte count(46). Our model-tuning results suggest that polygenic prediction models fit best for height (e.g., LDpred2-Infinitesimal and PRS-CS with $\phi = 0.01$) while sparser PRS models with stronger regularization (e.g., PRS-CS with $\phi = 0.0001$) provide better prediction accuracy for monocyte count.

Finally, PUMAS can also effectively estimate predictive R^2 for binary traits (**Figure 3C-D**). To calculate interpretable R^2 for binary outcomes, PUMAS first transforms GWAS summary statistics obtained from logistic regressions to the linear regression scale, and then computes R^2 on the observed scale(47-49). To show that such transformation is valid, we trained two sets of PRS models using both transformed and original logistic regression summary statistics for 4 disease traits and observed nearly identical PRS performance between two approaches (**Additional file 1**: **Fig. S29-S32**; **Additional file 14**: **Table. S13**). Details in the implementation of binary trait analysis and summary statistics transformation are presented in **Methods**.

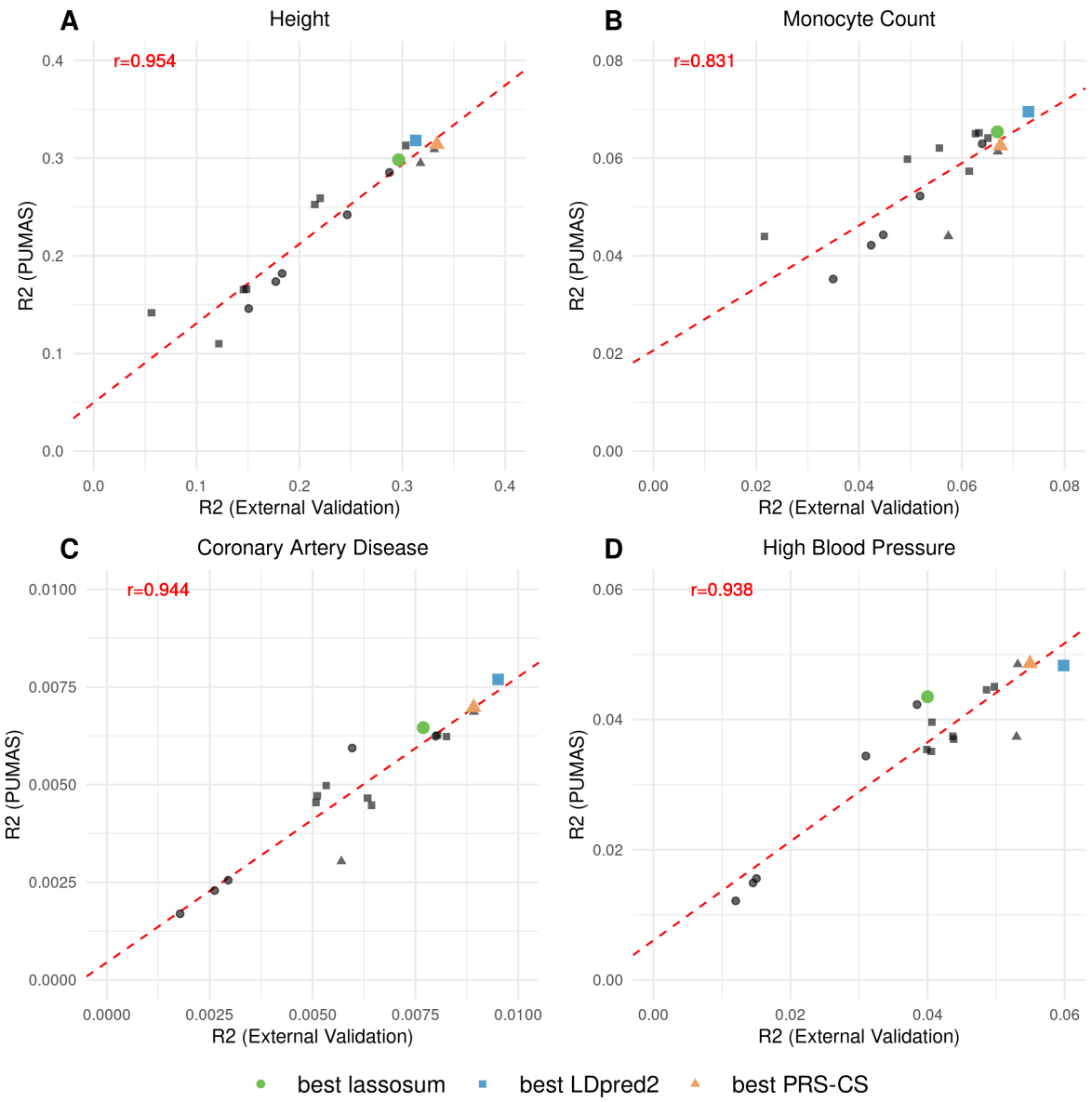

Figure 3. Comparing PUMAS results with external validation. Four panels show the model-tuning results for (**A**) height, (**B**) monocyte count, (**C**) coronary artery disease, and (**D**) high blood pressure. Y-axis: average predictive R^2 across 4-fold replications from PUMAS; X-axis: predictive R^2 evaluated by external validation on the holdout dataset. Each data points represents a PRS model with different tuning parameters and the shape of data points indicate three different PRS methods: LDpred2, PRS-CS, and lassosum. The best tuning parameter setting suggested by PUMAS for each PRS method is highlighted and colored. The dashed red line is fitted regression line between PRS R^2 from PUMAS and external validation. Pearson correlations between two sets of results are shown in each panel. Detailed model-tuning

results for all 21 traits are summarized in **Additional file 1**: **Fig. S13**-**S33** and **Additional file 13** and **14**: **Table. S12**-**S13**.

Ensemble learning via PUMAS-ensemble substantially improves PRS prediction accuracy

Here we apply PUMAS-ensemble, the ensemble learning extension of PUMAS, to UKB traits and show that ensemble PRS has superior prediction accuracy compared to each PRS method and our summary statistics-based approach is comparable to ensemble learning results based on individual-level data. We constructed linearly combined scores of lassosum, PRS-CS, LDpred2, and SDPR. Using individual-level data, we split the 10% UKB holdout dataset into two equally sized subsets. We fitted a multiple regression on the first holdout set to aggregate the best-performing PRS models trained and tuned from GWAS summary statistics, and then evaluated the ensemble score's prediction accuracy using the second holdout set. For comparison, we implemented PUMAS-ensemble to conduct 4-fold MCCV to perform ensemble learning using summary statistics alone and assess its performance on the second holdout set.

Our approach showed almost identical performance compared to individual-level data results (**Figure 4A**), showcasing PUMAS-ensemble's ability to benchmark and construct ensemble PRS without requiring additional datasets. In addition, ensemble PRS achieved the highest prediction accuracy compared with four input PRS models for all traits except diastolic blood pressure (**Additional file 1**: **Fig. S34**; **Additional file 15**: **Table. S14**). The ensemble PRS using individual-level data as input had an average 19.67% and 10.76% relative gain in R^2 compared to PRS-CS-auto and LDpred2-auto while the PUMAS-ensemble ensemble PRS delivered a similar 17.68%

and 8.93% R^2 increase respectively (**Figure 4B**), highlighting the substantial gain in prediction accuracy from ensemble learning.

We sought to explore some properties of PUMAS-ensemble in real-world settings. A comparison between PUMAS and PUMAS-ensemble suggests that although ensemble learning requires additional splitting of data and reduces training sample size for individual PRS models, ensemble scores trained on smaller GWAS training subsets outperforms individual fine-tuned PRS, highlighting the benefit of ensemble learning (**Methods**; **Additional file 16**: **Table. S15**). Furthermore, we conducted sensitivity analyses to investigate the effect of LD misspecification on PUMAS-ensemble (**Methods**). We observed reduced ensemble PRS performance from PUMAS-ensemble when the LD reference data mismatches the ancestral population of GWAS samples (**Additional file 1**: **Fig. S35-S36**; **Additional file 17**: **Table. S16**).

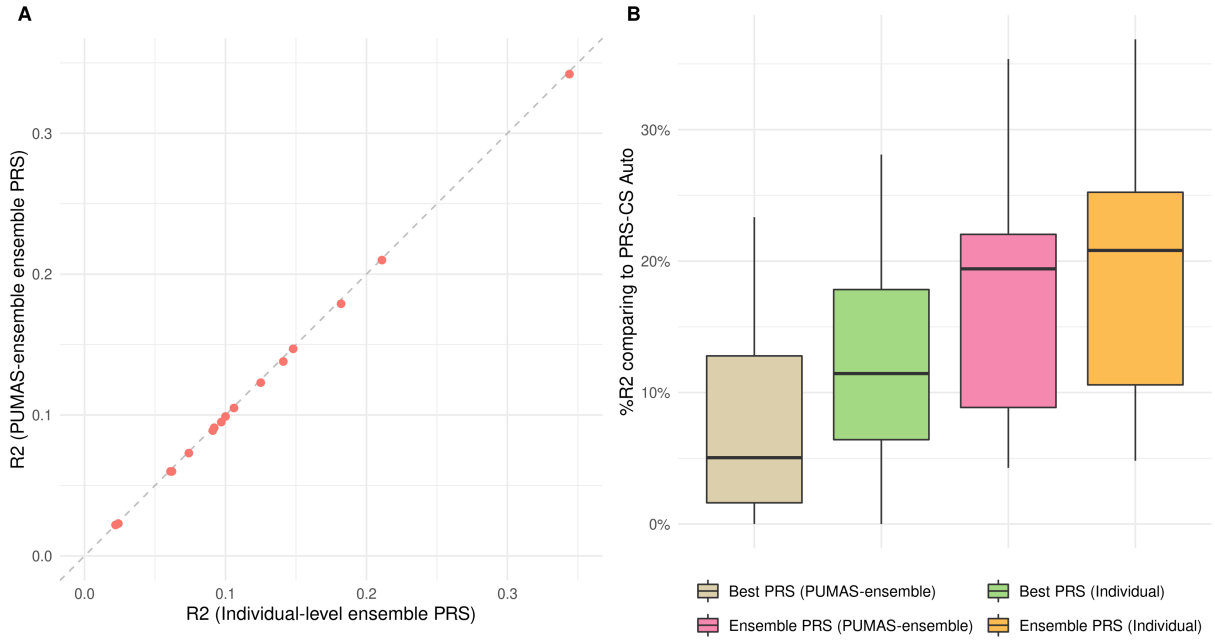

Figure 4. Constructing ensemble PRS for UKB traits. (**A**) Comparing two sets of ensemble PRS obtained from PUMAS-ensemble and individual-level data. The gray dashed line is the diagonal line. (**B**) Comparing ensemble PRS with input PRS methods. Y-axis: relative percentage increase in R^2 compared to PRS-CS-auto; X-axis: 4 sets of PRS models, including the best single PRS suggested by PUMAS, the best single PRS selected based on the first individua-level holdout set, the ensemble PRS obtained from PUMAS-ensemble, and the ensemble PRS trained from individual-level data. All R^2 values were computed using the second half of holdout dataset.

Constructing and benchmarking ensemble PRS for 31 complex traits

Finally, we applied PUMAS-ensemble to provide a comprehensive catalog of ensemble PRS for 31 publicly available GWAS summary statistics with varying sample size and genetic architecture. The detailed information and selecting criteria for GWAS summary-level data are summarized in **Methods** and **Additional file 18**: **Table. S17**. We employed extensive quality controls to pinpoint and calibrate misspecifications in GWAS summary statistics following a recent study(31) (**Additional file 19**: **Table. S18**). We also transformed logistic summary statistics to linear scale to produce interpretable R^2 for binary traits(47-49). For each trait, we reported prediction accuracy of the best performing PRS model and ensemble PRS. The full results of the PRS catalog are presented in **Additional file 20**: **Table. S19**. The predictive performance of ensemble PRS is correlated with estimated trait heritability, and the predictive R^2 ranged from 3E-4 to 0.213 across 31 traits, showing highly diverse predictive performance of genetic risk prediction. We also note that ensemble PRS improved predictive R^2 for every trait in the analysis with an median increase of 26.36% compared to PRS-CS-auto (**Additional file 1**: **Fig. S37**). Among 31 complex diseases

and traits, we observed the highest prediction improvement for rheumatoid arthritis (113.7%), Alzheimer's disease (93.0%, 94.3%, and 109.5% on three datasets), and post-traumatic stress disorder (70.7%).

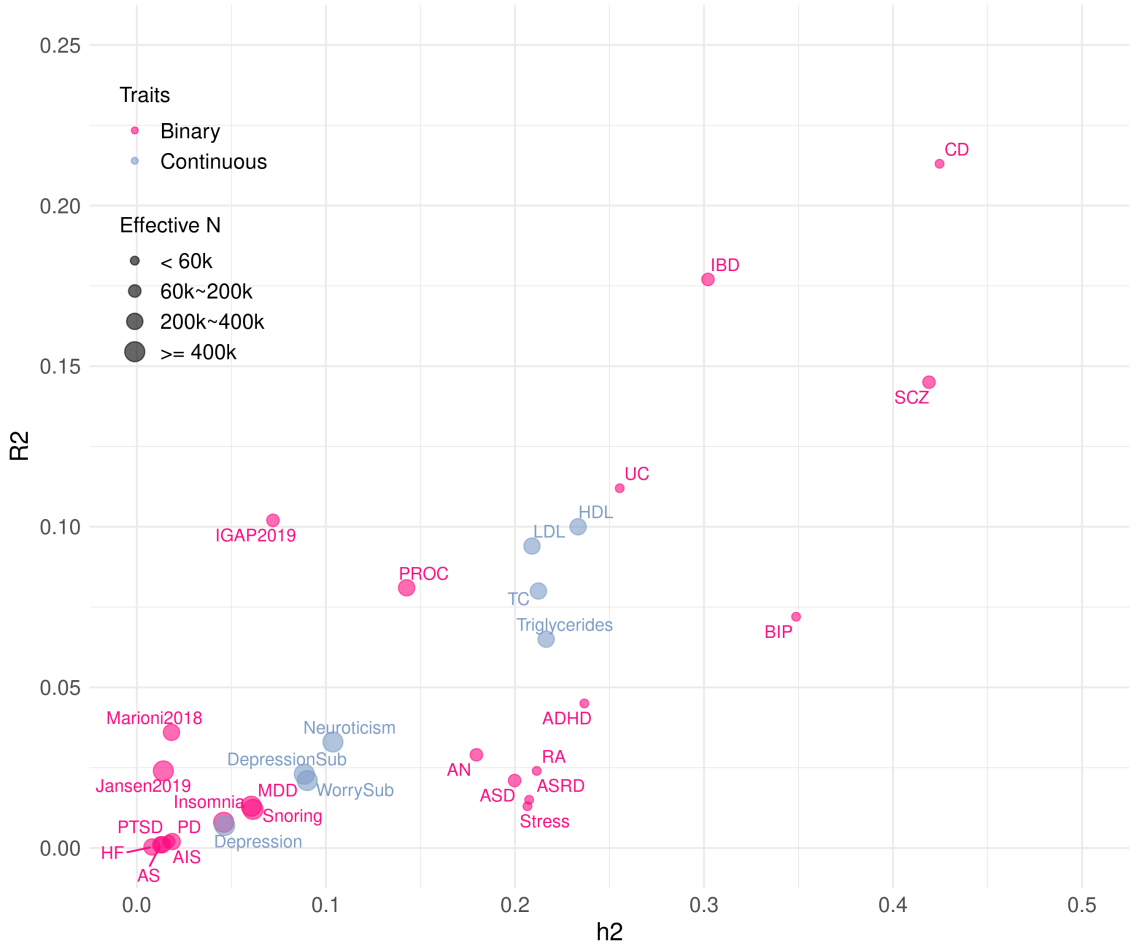

Figure 5. An ensemble PRS catalog for 31 complex traits. Y-axis: average predictive R^2 of PUMAS-ensemble ensemble PRS; X-axis: heritability estimates from LD score regression(50). Size of data points indicates the effective sample size of each GWAS. Binary traits and continuous traits are highlighted with different colors. Detailed PRS benchmark results are presented in **Additional file 20**: **Table. S19**.

Another observation is that the ensemble PRS R^2 exceeded the estimated trait heritability for all three Alzheimer's disease GWAS. To demonstrate that this is not an artifact from overestimating predictive R^2, we conducted additional analysis (**Methods**) using IGAP 2019 Alzheimer's GWAS summary statistics(51) and compared our results with external validation based on 2,600 Alzheimer's disease cases and 5,200 healthy controls in UKB (**Additional file 21**: **Table. S20**). The R^2 of AD PRS obtained from external validation also exceeded estimated heritability (h^2=0.072, SE=0.012) and the results were consistent with PUMAS R^2 estimation (**Additional file 1**: **Fig. S38**; **Additional file 22**: **Table. S21**). We hypothesized that this is driven by the *APOE* region which contributes an unusually large fraction of AD risk(52-54). Indeed, after removing 383 SNPs in the *APOE* region from IGAP 2019 AD summary statistics (**Methods**), we observed a steep decline in R^2 for both external validation and PUMAS. Both R^2 values became substantially lower than the estimated h^2 of 0.066 without *APOE* region (SE=0.009; **Additional file 22**: **Table. S21**).

Discussion

Fine-tuning and benchmarking PRS models are challenging tasks due to the need of external individual-level datasets that are independent from the input GWAS. In this work, we extended our PUMAS approach to incorporate LD and fine-tune state-of-the-art PRS methods. In both simulations and analysis of UKB traits, we observed high concordance between PUMAS and results based on external validation using holdout samples. In addition, we presented a novel

framework named PUMAS-ensemble to perform ensemble learning and create combined PRS using only GWAS summary statistics. We showed that ensemble PRS created by PUMAS-ensemble closely approximates scores built from holdout samples. Further, these ensemble scores substantially outperformed state-of-the-art PRS methods for complex traits we analyzed in the study. Finally, we applied PUMAS-ensemble to a collection of publicly available GWAS summary statistics and provided a comprehensive catalog of benchmarked and optimized PRS.

Our work presents several major advances that will impact future PRS applications. First, our method fills an important gap between PRS methodological research and its real-world applications. Currently, many PRS methods still have tuning parameters and grid search on external individual-level datasets remains the most common technique for fine-tuning these models. In practice, this kind of data can either be impossible to obtain, or need to be split from testing samples which could hurt statistical power in PRS applications(32). Our method provides a universal solution to PRS model fine-tuning. It is also noteworthy that some recent PRS methods such as MegaPRS(40) can also conduct model fine-tuning. MegaPRS bases its framework on GWAS z-scores and uses the LD matrix for summary statistics subsampling. On the other hand, PUMAS uses unstandardized SNP effect estimates and standard errors as inputs, and also considers GWAS regression residual variances in addition to LD for summary statistics partitioning. In practice, directly modeling GWAS standard errors and regression residual variances can be crucial when handling meta-analytic GWAS summary statistics(31) and when the trait of interest has a sparse genetic architecture. Second, model benchmarking is another major challenge in the field which conventionally relies on external validation data. Comprehensive and unbiased benchmarking allows researchers to compare the effectiveness of different PRS methods for particular traits of interest, and importantly, estimate PRS predictive accuracy without using testing samples. We note that although some advanced PRS approaches

do not require model fine-tuning anymore, no existing methods could benchmark model performance using a single set of GWAS summary data, which is crucial for model selection, power calculation, and study design. Our approach now provides a solution to this problem. Third, the ensemble learning approach which combines multiple predictive models through a second level regression has been viewed as a highly effective but data-demanding approach(28, 29, 33). A major advance in this study is the introduction of PUMAS-ensemble which allows ensemble learning on GWAS summary statistics. We note that this approach not only showcased a substantial gain over existing PRS methods, but is generally applicable to future PRS developments. If a future PRS approach shows promising improvements compared to older methods, that new approach can also be incorporated into the ensemble PRS. In our view, PUMAS-ensemble is not a competing approach for any existing PRS model, but instead is a flexible and general modeling technique that combines the best-performing methods out there and should be applied to all future PRS applications. It is important to note that our implemented software allows users to specify how they wish to split the GWAS summary statistics into subsets. In practice, we recommend that researchers customize data partitioning of summary statistics tailored towards their analytical needs.

Our study has several limitations. First, we have constrained most statistical analysis in this study to the European ancestral population. PRS is known to transfer poorly in terms of prediction accuracy for non-European populations which could exacerbate the disparity in genomic medicine between ancestral groups(55, 56). While we conducted sensitivity analysis to demonstrate less robust ensemble model training due to LD misspecification, it is an important future direction to systematically optimize and benchmark PRS for diverse ancestral populations which would require incorporation of multiple sets of ancestry-specific GWAS and LD references. Although we did not extensively explore this topic in this paper, our recent work introduced parallel ideas to

tackle the challenges in multi-ancestry genetic risk prediction(42). Second, we did not investigate the effect of assortative mating on PUMAS and PUMAS-ensemble in this study. Assortative mating is known to affect LD structure in human genome, bias heritability estimation(57), and affect PRS accuracy(58). The extent to which assortative mating influences our results requires further investigation. Third, analyses in this study were limited to GWAS summary statistics computed from independent samples. It remains to be investigated whether application of these approaches will be affected if the input GWAS summary statistics were obtained from linear mixed models with related samples or family-based designs(59-61). Future work will focus on developing statistical methods to correct for sample relatedness or demonstrate robustness to these issues. That said, we expect PRS model-tuning to remain valid even with sample relatedness since the inflation in R^2 should be uniform across various tuning parameter settings, although biases may be introduced to the predictive R^2 which could affect benchmarking efforts. Fourth, PUMAS/PUMAS-ensemble uses R^2 on the observed scale(49) to evaluate PRS accuracy for binary traits but AUC is adopted for classification more frequently. Although we have shown in an earlier work(38) that AUC and R^2 demonstrated highly consistent performance for PRS model fine-tuning, it remains future work to incorporate summary-statistics-based AUC estimator(62) into the PUMAS framework. Furthermore, our current analyses focused only on PRS derived from lassosum, PRS-CS, LDpred2, SDPR, and MegaPRS based on HapMap3 SNPs. While it serves to support the superiority of ensemble PRS as a proof of concept, more genetic variants and more PRS methods need to be jointly modeled and evaluated in the future, including scores that leverage auxiliary information including functional annotation(13, 14) or multiple phenotypes(15, 17, 63). Particularly for multi-trait PRS models, extending PUMAS to conduct multi-GWAS subsampling while modeling sample overlap between these GWAS summary statistics may be necessary. Finally, collinearity among PRS models could arise when using multiple regression to combine a large number of scores since some PRS methods tend to yield similar results.

Therefore, another future direction is to incorporate variable selection strategies into our ensemble learning framework, including penalized regression that has been employed in ensemble models based on individual-level data(35-37). It also remains an interesting but challenging task to fit non-linear ensemble PRS models using only GWAS summary statistics for incorporating machine learning ensemble methods such as XGBoost(64) used in Multi-PGS(63).

Conclusions

We presented a sophisticated statistical framework to fine-tune, combine, and benchmark PRS methods using only GWAS summary statistics. This is a statistically novel and computationally efficient approach with flexible implementation that can handle a variety of applications. We have demonstrated its performance through careful and comprehensive analyses, and we argue that this framework presents highly innovative and generally applicable features that should become the default in many future PRS studies.

Methods

Sampling distribution of summary statistics

We adopt a commonly used linear model framework to quantify the relationship between a quantitative trait and SNP genotypes:

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

Here, Y denotes the trait, $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p)$ denotes the genotypes of p SNPs, $\boldsymbol{\beta} \in \mathbb{R}^p$ denotes their true effect sizes, and ϵ denotes the random error that is independent from $\mathbf{X}$ and follows a normal distribution with mean zero and some variance σ_e^2. Let $\mathbf{y}$ and $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ denote the observed values for Y and $\mathbf{X}$ from N independent individuals. For simplicity, we assume both $\mathbf{y}$ and $\mathbf{x}_j$ ($j = 1, \dots, p$) are centered. Then, GWAS summary statistics can be denoted as:

$$\hat{\beta}_j = (\mathbf{x}_j^T \mathbf{x}_j)^{-1} (\mathbf{x}_j^T \mathbf{y}) \tag{1}$$

$$SE(\hat{\beta}_j) = \sqrt{\frac{\hat{\boldsymbol{\epsilon}}_j^T \hat{\boldsymbol{\epsilon}}_j}{(N-1)\mathbf{x}_j^T \mathbf{x}_j}} \tag{2}$$

where $\hat{\boldsymbol{\epsilon}}_j = \mathbf{y} - \mathbf{x}_j \hat{\beta}_j$ are the residuals from the marginal linear regression between the trait and the j-th SNP. To train, fine-tune, combine, and benchmark PRS models, independent datasets are required to avoid overfitting. We have previously proposed a flexible statistical framework to generate training and fine-tuning datasets when only GWAS summary statistics are available(38). Here, we generalize this statistical framework in two different directions. First, we allow our method to incorporate LD information. We note that this extension is similar to some recent work built on our initial PUMAS paper(40, 42). Second, we allow the method to partition full GWAS summary statistics into more than two datasets for various analytical purposes. Let $y^{(s)}$ and $x^{(s)}$ denote phenotype and genotype data for any arbitrary subset of N individuals with sample size $N^{(s)}$. When N is large enough, we have previously shown that by central limit theorem(38):

$$\mathbf{x}^T\mathbf{y} \sim \mathbf{N}\big(NE(\mathbf{X}^TY), NVar(\mathbf{X}^TY)\big)$$

$$\mathbf{x}^{(s)^T}\mathbf{y}^{(s)} \sim \mathbf{N}(N^{(s)}E(\mathbf{X}^TY), N^{(s)}Var(\mathbf{X}^TY))$$

where $\mathbf{X}^TY = (X_1Y, \ldots, X_pY)^T$. Then, given the observed summary-level data from GWAS, the conditional distribution of summary statistics of a subset of GWAS samples is

$$\mathbf{x}^{(s)^T}\mathbf{y}^{(s)}|\mathbf{x}^T\mathbf{y} \sim \mathbf{N}\left(\frac{N^{(s)}}{N}\mathbf{x}^T\mathbf{y}, \frac{\left(N - N^{(s)}\right)N^{(s)}}{N}\widehat{\mathbf{\Sigma}}\right) \quad (3)$$

where $\widehat{\mathbf{\Sigma}}$ is the observed variance-covariance matrix for $\mathbf{X}^TY$. To subsample summary statistics $\mathbf{x}^{(s)^T}\mathbf{y}^{(s)}$, we need to estimate $\mathbf{x}^T\mathbf{y}$ and $\widehat{\mathbf{\Sigma}}$ first. Recall formula *(1)* for marginal regression coefficient estimation, $\mathbf{x}_j^T\mathbf{y}$ can be calculated using $\hat{\beta}_j$ and $\mathbf{x}_j^T\mathbf{x}_j$ which is proportional to SNP variance and can be estimated by minor allele frequency (MAF) reported from GWAS or imputed from LD reference panel. On the other hand, deriving $\mathbf{\Sigma}$ is more complicated and we discuss how $\widehat{\mathbf{\Sigma}}$ is estimated using summary statistics and an LD reference panel in the following section.

Estimate variance-covariance matrix of summary statistics

Let D denote the SNP correlation matrix and d_{jk} denote the correlation between the j-th and the k-th SNPs. Let $\mathbf{\Sigma}$ be the true covariance matrix of summary statistics with diagonal and off-diagonal elements denoted as $\mathbf{\Sigma}_j$ and $\mathbf{\Sigma}_{jk}$, respectively. For convenience, we write $Y = \mathbf{X\beta} + \epsilon = X_1\beta_1 + \ldots + X_p\beta_p + \epsilon = X_j\beta_j + \epsilon_j$, where $\epsilon_j = \sum_{i:i\neq j} X_i\beta_i + \epsilon$. Then the diagonal terms of the $\mathbf{\Sigma}$ can be written as

$$
\begin{aligned}
\mathbf{\Sigma}_j = \quad & Var(X_j Y) \\
= \quad & Var[X_j(X_j\beta_j + \epsilon_j)] \\
= \quad & \beta_j^2 Var(X_j^2) + Var(X_j\epsilon_j) + 2\beta_j Cov(X_j^2, X_j\epsilon_j) \\
= \quad & \beta_j^2 Var(X_j^2) + Var[X_j(\sum_{i:i\neq j} X_i\,\beta_i + \epsilon)] + 2\beta_j Cov(X_j^2, X_j\epsilon_j)
\end{aligned}
$$

We partition all SNPs in the genome into 2 sets. Let S_1 be the index set that contains all SNPs that are independent from the j-th SNP and S_2 be the set with all remaining SNPs that are in LD with the j-th SNP. Then we can further expand $\mathbf{\Sigma}_j$ by

$$
\begin{aligned}
\mathbf{\Sigma}_j = \quad & \beta_j^2 Var(X_j^2) + Var[X_j(\sum_{g\in S_1} X_g\beta_g + \epsilon)] + Var[X_j(\sum_{g'\in S_2} X_{g'}\beta_{g'})] + \\
& 2Cov[X_j(\sum_{g\in S_1} X_g\beta_g + \epsilon), X_j(\sum_{g'\in S_2} X_{g'}\beta_{g'})] + 2\beta_j Cov(X_j^2, X_j\epsilon_j) \\
= \quad & \beta_j^2 Var(X_j^2) + Var[X_j(\sum_{g\in S_1} X_g\beta_g + \epsilon)] + \sum_{g'\in S_2} \beta_{g'}^2\, Var(X_j X_{g'}) + \\
& 2\sum_{g_1'\in S_2, g_2'\in S_2, g_1'\neq g_2'} \beta_{g_1'}\beta_{g_2'} Cov(X_j X_{g_1'}, X_j X_{g_2'}) + 2\sum_{g'\in S_2} \beta_{g'} Cov[X_j\epsilon, X_j X_{g'}] + \\
& 2\sum_{g\in S_1}\sum_{g'\in S_2} \beta_g\beta_{g'} Cov[X_j X_g, X_j X_{g'}] + 2\beta_j Cov(X_j^2, X_j\epsilon_j)
\end{aligned}
$$

We can simplify $\mathbf{\Sigma}_j$ based on two commonly made assumptions. First, any given SNP should be in linkage equilibrium with the vast majority of SNPs in the genome. Therefore, we can safely assert $|S_1| \gg |S_2|$. Second, each individual SNP's effect on the phenotype is typically very small such that the products of any effect sizes are negligible in practice. Taken together, we can reduce the expansion of $\mathbf{\Sigma}_j$ by discarding SNPs in S_2 which eventually allows us to treat X_j and ϵ_j as independent in practice:

$$
\begin{aligned}
\mathbf{\Sigma}_j \approx \quad & Var[X_j(\sum_{g\in S_1} X_g\beta_g + \epsilon)] \\
\approx \quad & Var[X_j\epsilon_j] \\
= \quad & E(X_j^2\epsilon_j^2) - [E(X_j\epsilon_j)]^2 \\
\approx \quad & E(X_j^2)E(\epsilon_j^2)
\end{aligned}
$$

Note that $E(X_j^2)$ can be easily approximated using an MAF-based estimator, denoted as $\hat{\sigma}_j^2$, that may be obtained either from the full GWAS summary statistics or the LD reference data. For $E(\epsilon_j^2)$, we can estimate its value by standard error of effect size estimation from GWAS summary data using formula *(2)*. In this way we can obtain an estimator of $\mathbf{\Sigma}_j$ as

$$\hat{\mathbf{\Sigma}}_j = N[SE(\hat{\beta}_j)\hat{\sigma}_j^2]^2 \qquad (4)$$

To estimate off-diagonal terms $\mathbf{\Sigma}_{jk}$, we now write $Y = \mathbf{X\beta} + \epsilon = X_1\beta_1 + \ldots + X_p\beta_p + \epsilon = X_j\beta_j + X_k\beta_k + \epsilon_{jk}$, where $\epsilon_{jk} = \sum_{i:i\notin\{j,k\}} X_i\beta_i + \epsilon$. Under the same assumption where the magnitude of SNP effects is very small, we can simplify $\mathbf{\Sigma}_{jk}$ by:

$$\begin{aligned}
\mathbf{\Sigma}_{jk} = \;& Cov[X_j(X_j\beta_j + X_k\beta_k + \epsilon_{jk}), X_k(X_j\beta_j + X_k\beta_k + \epsilon_{jk})] \\
= \;& Cov(X_j^2\beta_j, X_jX_k\beta_j) + Cov(X_jX_k\beta_k, X_jX_k\beta_j) + Cov(X_j\epsilon_{jk}, X_jX_k\beta_j) + \\
& Cov(X_j^2\beta_j, X_k^2\beta_k) + Cov(X_jX_k\beta_k, X_k^2\beta_k) + Cov(X_j\epsilon_{jk}, X_k^2\beta_k) + \\
& Cov(X_j^2\beta_j, X_k\epsilon_{jk}) + Cov(X_jX_k\beta_k, X_k\epsilon_{jk}) + Cov(X_j\epsilon_{jk}, X_k\epsilon_{jk}) \\
\approx \;& Cov(X_j\epsilon_{jk}, X_jX_k\beta_j) + Cov(X_j\epsilon_{jk}, X_k^2\beta_k) + Cov(X_j^2\beta_j, X_k\epsilon_{jk}) + \\
& Cov(X_jX_k\beta_k, X_k\epsilon_{jk}) + Cov(X_j\epsilon_{jk}, X_k\epsilon_{jk}) \\
\approx \;& Cov(X_j\epsilon_{jk}, X_k\epsilon_{jk})
\end{aligned}$$

In a similar fashion, we further partition all SNPs in the genome other than the j-th and the k-th SNP into two sets. Let S_3 denote the collection of SNPs that are independent from both the j-th and the k-th SNPs, and S_4 includes the remaining SNPs that are in LD with either the j-th or the k-th SNP. Based on a similar rationale, we can safely assume that $|S_3| \gg |S_4|$. Then, by ignoring SNPs in S_4 and thus treating X_j and X_k as being independent from ϵ_{jk}, we express $\mathbf{\Sigma}_{jk}$ as:

$$
\begin{aligned}
Cov(X_j\epsilon_{jk}, X_k\epsilon_{jk}) = \quad & Cov[X_j(\sum_{l\in S_3} X_l\beta_l + \epsilon), X_k(\sum_{l\in S_3} X_l\beta_l + \epsilon)] + Cov[X_j(\sum_{l'\in S_4} X_{l'}\beta_{l'}), X_k(\sum_{l\in S_3} X_l\beta_l + \epsilon)] \\
& + Cov[X_j(\sum_{l\in S_3} X_l\beta_l + \epsilon), X_k(\sum_{l'\in S_4} X_{l'}\beta_{l'})] + Cov[X_j(\sum_{l'\in S_4} X_{l'}\beta_{l'}), X_k(\sum_{l'\in S_4} X_{l'}\beta_{l'})] \\
\approx \quad & Cov[X_j(\sum_{l\in S_3} X_l\beta_l + \epsilon), X_k(\sum_{l\in S_3} X_l\beta_l + \epsilon)] \\
\approx \quad & Cov[X_j\epsilon_{jk}, X_i\epsilon_{jk}] \\
\approx \quad & E(X_jX_k)E(\epsilon_{jk}^2)
\end{aligned}
$$

where $E(X_jX_k)$ can be directly estimated by the LD correlation matrix and MAF-based SNP variance estimator. For $E(\epsilon_{jk}^2)$, it is the residual variance from a two-SNP regression model and should be smaller than both $E(\epsilon_j^2)$ and $E(\epsilon_k^2)$. In practice, we can approximate it by the smaller value between $\frac{\hat{\boldsymbol{\epsilon}}_j^T\hat{\boldsymbol{\epsilon}}_j}{N-1}$ and $\frac{\hat{\boldsymbol{\epsilon}}_k^T\hat{\boldsymbol{\epsilon}}_k}{N-1}$. Therefore, the numerical approximation for $\boldsymbol{\Sigma}_{jk}$ becomes

$$
\hat{\boldsymbol{\Sigma}}_{jk} = Nd_{jk}\hat{\sigma}_j\hat{\sigma}_k \min\{SE(\hat{\beta}_j)\hat{\sigma}_j, SE(\hat{\beta}_k)\hat{\sigma}_k\}^2 \tag{5}
$$

Now we can then generate summary statistic from the multivariate normal distribution in formula *(3)*. Note that our earlier subsampling framework is a special case where SNPs are independent and its only difference with the current method is the estimation of $\hat{\boldsymbol{\Sigma}}_{jk}$. In the next session we will discuss how to subsample summary statistics efficiently from a multivariate normal distribution.

Strategy for subsampling summary statistics

Next, we discuss how to partition full GWAS summary statistics into K independent subsets of GWAS samples, denoted as $\mathbf{x}^{(1)^T}\mathbf{y}^{(1)}, \dots, \mathbf{x}^{(K)^T}\mathbf{y}^{(K)}$ for $K > 2$. When $K = 2$, formula *(3)* can be directly applied to divide GWAS summary statistics into two independent sets. Otherwise, let

$N^{(1)}, \ldots, N^{(K)}$ denote the corresponding sample size for each subset of individuals and $N = \sum_{s=1}^{K} N^{(s)}$. By formula *(3)*, we can subsample $\mathbf{x}^{(1)^T}\mathbf{y}^{(1)}$ from $\mathbf{x}^T\mathbf{y}$ observed in the complete GWAS summary data. After that, we calculate summary statistics excluding $N^{(1)}$ individuals from the first subset as $\mathbf{x}^{(-1)^T}\mathbf{y}^{(-1)} = \mathbf{x}^T\mathbf{y} - \mathbf{x}^{(1)^T}\mathbf{y}^{(1)}$. This technique of combining or subtracting independent sets of summary statistics has been commonly utilized by methods such as METAL and Metasubtract(65, 66). To generate summary statistics for any following subset numbered $t+1$ (i.e., $\mathbf{x}^{(t+1)^T}\mathbf{y}^{(t+1)}$) for $t = 1, \ldots, K-2$, we update the conditional distribution in *(3)* with the new "full" GWAS summary statistics and correspondent total sample size:

$$\mathbf{x}^{(t+1)^T}\mathbf{y}^{(t+1)} | \mathbf{x}^{(-t)^T}\mathbf{y}^{(-t)} \sim \mathbf{N}\left(\frac{N^{(t+1)}}{N - \sum_{s=1}^{t} N^{(s)}}\mathbf{x}^{(-t)^T}\mathbf{y}^{(-t)}, \frac{\left(N - \sum_{s=1}^{t+1} N^{(s)}\right)N^{(t+1)}}{N - \sum_{s=1}^{t} N^{(s)}}\widehat{\mathbf{\Sigma}}\right) \qquad (6)$$

where $\mathbf{x}^{(-t)^T}\mathbf{y}^{(-t)}$ represents summary statistics excluding first t subsets of individuals. This subsampling strategy guarantees that every subset is independent from each other and avoids overfitting when $K > 2$. Finally, for the last subset K, we can directly calculate its summary statistics by $\mathbf{x}^{(K)^T}\mathbf{y}^{(K)} = \mathbf{x}^T\mathbf{y} - \sum_{s=1}^{K-1} \mathbf{x}^{(s)^T}\mathbf{y}^{(s)}$. Together, this is a flexible framework for generating summary statistics and can be used for various types of PRS analyses as we discuss in later sections.

It is a difficult task to subsample summary statistics for all SNPs in the genome simultaneously given the large dimension of genotype and imputed data. Even if PRS modeling is restricted to HapMap3 SNPs, it remains challenging to subsample $\mathbf{x}^{(s)^T}\mathbf{y}^{(s)}$ for more than one million SNPs altogether(26). To efficiently generate data, we partition the whole genome into approximately independent LD blocks and subsample summary statistics for SNPs in each LD block separately(67, 68). Then $\widehat{\mathbf{\Sigma}}$ becomes a sparse block-diagonal matrix, i.e., $\widehat{\mathbf{\Sigma}} = diag(\widehat{\mathbf{\Sigma}}_{D_i})$. Within

each LD block, the empirical SNP correlation matrix may not always be positive-definite and thus making it impossible to randomly generate data from that LD block. A straightforward remedy is to conduct eigen decomposition for any $\widehat{\mathbf{\Sigma}}_{D_i}$ that is negative definite, manually change negative eigenvalues to 0's, and obtain an approximation of $\widehat{\mathbf{\Sigma}}_{D_i}$ that is positive semi-definite. Note that this may not be the best approach and other methods for estimating LD blocks can also be applied(69, 70).

Evaluate predictive performance of PRS

Here, we generalize the summary-statistics-based PRS evaluation scheme proposed in our previous work to incorporate LD. We denote PRS as a weighted sum of allele counts across many SNPs:

$$\hat{Y} = \mathbf{X}\boldsymbol{\omega}$$

where $\boldsymbol{\omega} \in \mathbb{R}^p$ is a vector of SNP weights, which can be marginal regression coefficients from GWAS or post-hoc effect size estimates. If individual-level data is available, then R^2 evaluated on any holdout dataset $\left(\mathbf{y}^{(s)}, \mathbf{x}^{(s)}\right)$ can be calculated as

$$R^2_{individual} = \frac{\left[Cov\left(\mathbf{y}^{(s)}, \hat{\mathbf{y}}^{(s)}\right)\right]^2}{Var(\mathbf{y}^{(s)})Var(\hat{\mathbf{y}}^{(s)})} = \frac{\left(\sum_{i=1}^{N^{(s)}} y_i^{(s)}\hat{y}_i^{(s)} - N^{(s)}\overline{\mathbf{y}^{(s)}}\,\overline{\hat{\mathbf{y}}^{(s)}}\right)^2}{\sum_{i=1}^{N^{(s)}}\left(y_i^{(s)} - \overline{\mathbf{y}^{(s)}}\right)^2 \sum_{i=1}^{N^{(s)}}\left(\hat{y}_i^{(s)} - \overline{\hat{\mathbf{y}}^{(s)}}\right)^2}$$

where $\hat{y}_i$ is the PRS for the i-th person, $\overline{\mathbf{y}^{(s)}}$ is the mean phenotypic value, and $\overline{\hat{\mathbf{y}}^{(s)}}$ is the mean PRS value in holdout dataset s. On the other hand, we have shown that when only summary statistics of the holdout dataset is available and SNPs are independent, $R^2_{individual}$ can be approximated by(38):

$$\frac{1}{N^{(s)}}\sum_{i=1}^{N^{(s)}}\left(\hat{y}_i^{(s)}-\overline{\hat{\mathbf{y}}^{(s)}}\right)^2 \approx \sum\nolimits_{j=1}^{p} w_j^2\hat{\sigma}_j^2$$

$$\frac{1}{N^{(s)}}\sum_{i=1}^{N^{(s)}}\left(y_i^{(s)}-\overline{\mathbf{y}^{(s)}}\right)^2 \approx \max_j\left[\frac{\hat{\boldsymbol{\epsilon}}_j^{\top}\hat{\boldsymbol{\epsilon}}_j}{N-1}\right] \approx N\max_j\{SE\left(\hat{\beta}_j\right)^2\hat{\sigma}_j^2\}$$

$$\hat{R}_{noLD}^2 \approx \frac{\left(\frac{1}{N^{(s)}}\sum_{j=1}^{p} w_j\mathbf{x}_j^{(s)\top}\mathbf{y}^{(s)}\right)^2}{N\max_j\{SE\left(\hat{\beta}_j\right)^2\hat{\sigma}_j^2\}\sum_{j=1}^{p} w_j^2\,\hat{\sigma}_j^2}$$

given that $\mathbf{x}^{(s)}$, $\mathbf{y}^{(s)}$, and $\hat{\mathbf{y}}^{(s)}$ are centered. In practice, we use the 90% quantile instead of $\max_j\{SE\left(\hat{\beta}_j\right)^2\hat{\sigma}_j^2\}$ to get a robust estimate of $Var\left(\mathbf{y}^{(s)}\right)$. When LD is present, the approximations for $Cov\left(\mathbf{y}^{(s)},\hat{\mathbf{y}}^{(s)}\right)$ and $Var\left(\mathbf{y}^{(s)}\right)$ remain the same. For $Var\left(\hat{\mathbf{y}}^{(s)}\right)$, it can now be approximated by $\boldsymbol{\omega}^T Var\left(\mathbf{x}^{(s)}\right)\boldsymbol{\omega}$, with $Var\left(\mathbf{x}^{(s)}\right)$ estimated using the LD correlation matrix and MAF calculated from the reference panel. Taken together, we have

$$\hat{R}_{LD}^2 = \frac{\left(\frac{1}{N^{(s)}}\sum_{j=1}^{p}\omega_j\,\mathbf{x}_j^{(s)^T}\mathbf{y}^{(s)}\right)^2}{N\max_j\{SE\left(\hat{\beta}_j\right)^2\hat{\sigma}_j^2\}\left(\sum_{j=1}^{p}\sum_{k\neq j} w_j^2\hat{\sigma}_j^2 + w_j w_k d_{jk}\hat{\sigma}_j\hat{\sigma}_k\right)} \qquad (7)$$

Note that similar versions of this formula have been tested and applied in the literature(22, 40, 41). In practice, we can directly calculate PRS on the LD reference genotype data and use the sample variance of PRS to replace $\sum_{j=1}^{p}\sum_{k\neq j} w_j^2\hat{\sigma}_j^2 + w_j w_k d_{jk}\hat{\sigma}_j\hat{\sigma}_k$ for optimal computational efficiency.

The PUMAS framework

Given the flexible framework we introduced for subsampling GWAS summary data and evaluating PRS based on summary statistics, PUMAS becomes a special case where the entire GWAS summary-level data is partitioned into a training and a tuning dataset, denoted as $\mathbf{x}^{(tr)^T}\mathbf{y}^{(tr)}$ and $\mathbf{x}^{(tn)^T}\mathbf{y}^{(tn)}$. PUMAS first draws $\mathbf{x}^{(tn)^T}\mathbf{y}^{(tn)}$ from *(3)* and then calculates $\mathbf{x}^{(tr)^T}\mathbf{y}^{(tr)}$ by $\mathbf{x}^T\mathbf{y} - \mathbf{x}^{(tn)^T}\mathbf{y}^{(tn)}$. For each SNP, the marginal effect size and its standard error from the training set can be calculated as

$$\hat{\beta}_j^{(tr)} = [N^{(tr)}\,\hat{\sigma}_j^2]^{-1}\mathbf{x}_j^{(tr)^T}\mathbf{y}^{(tr)}$$

$$SE(\hat{\beta}_j^{(tr)}) = \sqrt{\frac{N}{N^{(tr)}}}SE(\hat{\beta}_j)$$

Then these summary statistics from the training dataset can be used to train any PRS methods that use GWAS summary statistics as input. R^2 of the PRS model assessed on the fine-tuning dataset can be approximated by replacing $\mathbf{x}^{(s)^T}\mathbf{y}^{(s)}$ with $\mathbf{x}^{(tn)^T}\mathbf{y}^{(tn)}$ and changing the corresponding sample size in formula *(7)*. This procedure can be repeated k times to implement a k-fold Monte Carlo cross-validation (MCCV) to select the best-performing tuning parameter. When there is a set of tuning parameters $\boldsymbol{\lambda}$ in a PRS framework, that is, $\hat{Y}(\lambda) = \mathbf{X}\boldsymbol{\omega}(\lambda)$**,** PUMAS chooses the optimal tuning parameter $\hat{\lambda}$ by

$$\hat{\lambda} = \underset{\lambda\in\boldsymbol{\lambda}}{\operatorname{argmax}}\ \bar{R}_{LD}^2(\lambda)$$

where $\bar{R}_{LD}^2$ denotes the mean $\hat{R}_{LD}^2$ across k-fold MCCV. This cross-validation technique also applies to models that are hyperparameter-free or fine-tuned in advance. When the goal is to pick the best PRS model among a total of M PRS methods, the best model $\hat{m}$ can be selected by

$$\hat{m} = \underset{m=1,2,\ldots,M}{\operatorname{argmax}}\ \bar{R}_{LD}^2\left(m, \hat{\lambda}_m\right)$$

where $\hat{\lambda}_m$ is the besting tuning parameter for PRS framework m.

Combining multiple PRSs with PUMAS-ensemble

Next, we introduce PUMAS-ensemble, an extension of PUMAS that applies ensemble learning to combine multiple PRS using GWAS summary statistics. To do this, PUMAS-ensemble further partitions the full GWAS association results to 4 independent sets of summary statistics corresponding to training ($\mathbf{x}^{(tr)^T}\mathbf{y}^{(tr)}$), tuning ($\mathbf{x}^{(tn)^T}\mathbf{y}^{(tn)}$), ensemble training ($\mathbf{x}^{(etr)^T}\mathbf{y}^{(etr)}$), and testing ($\mathbf{x}^{(t)^T}\mathbf{y}^{(t)}$) summary statistics. Using formula *(6)*, we subsample summary statistics iteratively and compute $\mathbf{x}^{(tr)^T}\mathbf{y}^{(tr)} = \mathbf{x}^T\mathbf{y} - \mathbf{x}^{(tn)^T}\mathbf{y}^{(tn)} - \mathbf{x}^{(etr)^T}\mathbf{y}^{(etr)} - \mathbf{x}^{(t)^T}\mathbf{y}^{(t)}$. Like PUMAS, PUMAS-ensemble first conducts k-fold MCCV using training and tuning summary statistics to pick the best tuning parameter for each PRS method. Then, it trains each optimal PRS model's weight on the ensemble training data and evaluates the combined PRS on the testing summary statistics. A straightforward and intuitive way of combining PRS is through multiple linear regression. However, if individual-level genotype and phenotype data is not available, we cannot fit the regression in the conventional way. Below we illustrate how to calculate regression coefficients using summary-level data alone. We define the multiple linear regression model on the ensemble training dataset as:

$$Y^{(etr)} = \alpha_1 \times \hat{Y}_1^{(etr)} + \alpha_2 \times \hat{Y}_2^{(etr)} + \cdots + \alpha_M \times \hat{Y}_M^{(etr)} + \epsilon_{prs}$$

where $\boldsymbol{\alpha} = [\alpha_1\ \alpha_2 \ldots\ \alpha_M]^T$ are PRS weights for M PRS methods. We also define

$$\mathbf{z} = \left[\hat{\mathbf{y}}_1^{(etr)}\ \hat{\mathbf{y}}_2^{(etr)} \ldots\ \hat{\mathbf{y}}_M^{(etr)}\right] = \mathbf{x}^{(etr)}\mathbf{W}$$

as the observed PRS matrix with dimension $N^{(etr)} \times M$, and $\mathbf{W} = [\mathbf{w}_1\ \mathbf{w}_2 \ldots\ \mathbf{w}_M]$ are a $p \times M$ SNP weights matrix for p SNPs from M methods. To obtain the least squares estimator of $\boldsymbol{\alpha}$, that

is $\widehat{\boldsymbol{\alpha}} = \left(\mathbf{z}^{\mathrm{T}}\mathbf{z}\right)^{-1}\mathbf{z}^{\mathrm{T}}\mathbf{y}^{(etr)}$, we need to estimate $\mathbf{z}^{\mathrm{T}}\mathbf{z}$ and $\mathbf{z}^{\mathrm{T}}\mathbf{y}^{(etr)}$ separately. In fact, under the assumption that genotype and phenotype are both centered, we can show that

$$\mathbf{z}^T\mathbf{z} \approx N^{(etr)} \cdot \widehat{\boldsymbol{\Sigma}}_{\mathbf{z}} \tag{8}$$

$$\mathbf{z}^T\mathbf{y}^{(etr)} \approx \mathbf{W}^T\mathbf{x}^{(etr)^T}\mathbf{y}^{(etr)} \tag{9}$$

where $\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}$ is the empirical covariance matrix of the PRS matrix $\mathbf{z}$. In practice, we can estimate $\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}$ by calculating PRSs and their sample covariance matrix on a reference LD genotype dataset or approximate it by computing $\mathbf{W}^T\mathbf{D}\mathbf{W}$. Taken *(8)* and *(9)* together, we can estimate PRS weights using only summary statistics. Then we take the average PRS weights across k folds, i.e., $\overline{\boldsymbol{\alpha}} = \frac{1}{k}\sum_{j=1}^{k}\widehat{\boldsymbol{\alpha}}_j$, and report it as the PRS weight to combine optimized PRSs. Finally, we modify equation *(7)* to calculate predictive R^2 for ensemble PRS on the testing summary-level data:

$$\hat{R}^2_{ensemble} = \frac{\left[\frac{1}{N^{(t)}}\widehat{\boldsymbol{\alpha}}^T\mathbf{W}^T\mathbf{x}^{(t)^T}\mathbf{y}^{(t)}\right]^2}{N\max_j\{SE\left(\hat{\beta}_j\right)^2\hat{\sigma}_j^2\}\,\widehat{\boldsymbol{\alpha}}^T\widehat{\boldsymbol{\Sigma}}_{\mathbf{z}}\widehat{\boldsymbol{\alpha}}} \tag{10}$$

In the end, PUMAS-ensemble reports the average prediction accuracy of ensemble PRS across k folds. Note that PUMAS-ensemble can benchmark all PRS models in addition to the ensemble PRS on the testing summary statistics since it is independent from training and tuning datasets. Given sufficient data for model training, the ensemble learning model should always outperform fine-tuned PRS identified through grid-search. This is because grid search result is a special case in the ensemble learning, with the weight set to be 1 for a particular PRS and 0 for all other PRS models. Instead, the ensemble learning approach fits a regression to identify optimal weight values to maximize the predictive performance of ensemble score. Taken together, PUMAS-ensemble is a highly flexible framework to train, fine-tune, combine, and evaluate PRS models based on GWAS summary statistics.

Binary phenotypes

There are two challenges when applying PUMAS and PUMAS-ensemble to binary phenotypes. First, summary statistics obtained from logistic regression frameworks violate the linear regression model assumption in our derivation. Therefore equations *(3)* and *(6)* are not directly appliable to subsampling summary statistics for binary traits because $\mathbf{X}^T Y$ calculation is non-trivial for log odds ratios. Second, squared Pearson correlation between a binary outcome and PRS using logistic regression coefficients as input is less interpretable and rarely reported. On the other hand, area under the ROC curve (AUC) is often the preferred metric to quantify PRS accuracy for binary outcome. AUC calculation based on summary statistics has been developed but is not yet generalized to handle whole genome data, making it difficult to evaluate more sophisticated PRS methods that leverage contributions from millions of SNPs when individual-level data is not accessible(71). Here we propose a simple solution that allows us to apply PUMAS and PUMAS-ensemble to binary phenotypes and report interpretable R^2. For binary traits, R^2 on the observed scale (i.e., R^2_{obs}) has been defined and discussed in the literature as an alternative metric for evaluating PRS prediction accuracy(49). R^2_{obs} is the squared correlation between PRS and 0-1 status where PRS uses effect sizes estimated from linear probability model (LPM, i.e., linear regression between the binary response and SNP allele counts) as inputs(72). If GWAS summary-level data is acquired from linear probability model, then PUMAS and PUMAS-ensemble can be directly applied to calculate R^2_{obs} for binary traits(60). When LPM summary statistics are not available, since a single SNP has very weak effect on the phenotypic outcome in practice, we can still safely approximate LMP coefficient estimations using Z-score from logistic

regression(47, 48). Specifically, we can calculate $\hat{\beta}_{j,LPM} \approx Z_{j,logistic} \times \sqrt{\frac{v(1-v)}{X_j^T X_j}}$ where $Z_{j,logistic}$ is Z-score for the j-th SNP from logistic summary statistics and v is the sample prevalence. Then, we can use $\hat{\beta}_{j,LPM}$ and correspondent standard error $SE(\hat{\beta}_{j,LPM}) \approx \sqrt{\frac{v(1-v)}{X_j^T X_j}}$ to apply PUMAS and PUMAS-ensemble to dichotomous phenotypes. Eventually, if it is preferred to transform R^2_{obs} to R^2 on the liability scale ($R^2_{liability}$) which can be comparable across different studies and phenotypes, such transformation has been developed using sample and population prevalence(49).

Sample size imputation

In this section, we discuss how to handle sample size misspecification in GWAS summary statistics when applying our approach. Sample size misspecification is common in published GWAS datasets since many studies often do not report SNP-specific sample size and only provide a maximum sample size for the entire study. This is sub-optimal for PRS training if variant-level samples sizes differ substantially (e.g., in meta-analysis). A recent study has extensively investigated sample size misspecification in marginal association statistics and observed consistently decreased PRS prediction accuracy when the issue is not properly addressed(31). For PUMAS and PUMAS-ensemble, incorrect sample sizes will both affect the quality of subsampled summary statistics and bias the estimation of predictive R^2. To address this issue, we employed the approach proposed in Privé et al. to impute and conduct quality control on variant-specific sample size(31). Specifically, when the summary-level data does not provide sample size information for each SNP, we first impute sample size and remove SNPs with imputed

sample size smaller than 70% and larger than 110% of reported maximum sample size. For summary statistics that provides per-SNP sample sizes, we simply removed variants with sample size smaller than 70% of the largest sample size. On the other hand, to make sure formula *(7)* and *(10)* work for summary statistics with varying SNP-specific sample sizes, we enforce all summary statistics other than training summary statistics to have the same sample size for every SNP. We achieve this by subsampling all other summary statistics first where we can specify subset size and calculate $\mathbf{x}^{(tr)^T}\mathbf{y}^{(tr)}$ at last.

PRS training

We trained lassosum, PRS-CS, and LDpred2 models for all PRS analyses in this study(22, 25, 26). lassosum is a penalized regression framework that trains lasso regression coefficients for SNPs in each LD blocks with tuning parameters s and λ, where s controls the sparsity of LD matrix and λ is the penalty term that regularizes shrinkage of effect sizes. PRS-CS and LDpred2 are both Bayesian PRS frameworks with different prior assumptions for the SNP effect size distribution. PRS-CS has a global shrinkage parameter ϕ that uniformly shrinks its continuous prior distribution for each SNP and includes a fully Bayesian approach that automatically learns ϕ during model fitting. LDpred2 is an extension of LDpred that places a point normal prior on SNP effects based on tuning parameter p that represents the proportion of causal variants in the genome (LDpred non-inf and LDpred2_grid) or a univariate normal prior on all SNPs that doesn't require model-tuning (LDpred/LDpred2-Inf)(12). Like PRS-CS, LDpred2 can also employ an empirical Bayesian approach to optimize p on the training summary statistics. For implementation, we trained PRS-CS (v1.0.0) models using UKB European LD reference for simulation study and

1000 Genomes European LD reference for real data analysis. We followed PGS server pipeline to implement lassosum (R package 'lassosum' v0.4.5) and LDpred2 (R package 'bigsnpr' v1.9.11)(22, 28, 73). Due to larger computational burden, we implemented LDpred2 on each chromosome separately and only used the estimated heritability from LD-score regression as the tuning parameter h^2 in LDpred2(50). For real data analysis in UKB we constructed both non-sparse and sparse versions of LDpred2 models. We employed more shrinkage on LDpred2-auto model (shrink_corr = 0.5) and LDpred2_grid models (low_h^2=0.1*h^2) when analyzing publicly available GWAS summary statistics to ensure model convergence. The best tuning parameter for lassosum was obtained through grid search. For LDpred2 and PRS-CS, we compared grid search with empirical Bayesian models to find the best parameter.

In addition, we trained SDPR(43) and MegaPRS(40) models for all ensemble PRS analyses and PUMAS-ensemble simulation, respectively. SDPR is a recently developed Bayesian nonparametric PRS model that is computationally efficient and tuning-free. We fitted SDPR models using its latest v0.9.1 release on github with provided 1000 Genomes Project EUR LD reference. MegaPRS is a flexible Bayesian PRS framework that can employ multiple different prior specifications. The MegaPRS model with BayesR prior includes 84 sets of tuning parameters that determine the relative weights of various Gaussian components. We fitted MegaPRS models using the LDAK software (v5.2) with LDAK-thin heritability estimation(74). We only trained PRS models on HapMap3 SNPs in all analyses throughout this study.

Simulation settings

We conducted simulations using UKB genotype data(44) imputed to the Haplotype Reference Consortium reference. We removed samples who are not of European ancestry and 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 HapMap3 SNP list and 1000 Genomes Project Phase III LD reference data for European ancestry from PRS-CS. 377,509 samples and 944,547 variants remained after quality control. Then, we randomly selected 100,000 samples to be the training dataset and 1,000 samples as the LD genotype reference for our summary-statistics-based approach. To generate trait values, we simulated true effect sizes from a point normal distribution, i.e., $\beta_j \sim (1-p)\delta_0 + p\mathrm{N}(0, \frac{h^2}{Mp})$ where p is the proportion of causal variants, δ_0 is point mass at 0, h^2 is the total heritability of the phenotype, and M is the total number of SNPs(7, 12). We did not simulate associations between SNP true effects on the allelic scale and MAF since previous analysis has shown minimal difference in performance between PUMAS and PRS validation using individual-level data(38, 74). We chose p to be 0.1% and 20% corresponding to sparse and polygenic genetic models, and $h^2 = 0.2, 0.5, 0.8$ to create a total of 6 simulation settings with various types of genetic architecture. Within each setting, we randomly selected causal variants across the whole genome. Then we simulated quantitative traits by adding up the SNP allele counts weighted by their true effect sizes and randomly generated gaussian noises scaled based on trait heritability. We fitted marginal linear regression in PLINK to obtain GWAS summary statistics in each setting(75).

We compared PUMAS-ensemble with 4-fold MCCV. To implement 4-fold MCCV, in each fold we randomly selected 60% of all samples to form the training dataset (N=60,000), 20% as the tuning dataset (N=20,000), 10% as the ensemble training dataset (N=10,000), and the remaining 10% as the testing dataset (N=10,000). We conducted GWAS on the training data and used summary

statistics to train PRS models, fine-tuned PRS methods on the tuning data, obtained optimized PRSs' weights in the ensemble score by fitting multiple linear regression on the ensemble training data, and finally evaluated each PRS model's predictive R^2 on the testing data. For PUMAS-ensemble, we first used all samples (N=100,000) to fit marginal linear regression and obtained the full summary statistics. In a similar fashion, we partitioned the full summary statistics to training summary data (N=60,000), tuning summary data (N=20,000), ensemble learning summary data (N=10,000) and testing summary data (N=10,000) for corresponding PRS analysis. Similarly, we compared PUMAS with 4-fold MCCV by using only the training and tuning summary-level and individual-level data for two approaches, respectively. We included lassosum, LDpred2, and PRS-CS in all simulation analysis, and added SDPR and MegaPRS in PUMAS-ensemble simulations. In all simulations, we used 1000 Genomes Project European LD dataset provided by the PRS-CS software to subsample summary statistics. lassosum, LDpred2, and MegaPRS model training used the holdout UKB LD genotype data (N=1,000) as the LD reference. We implemented SDPR, lassosum with $s = 0.2, 0.5, 0.9$ and $\lambda = 0.005, 0.01$, PRS-CS with $\phi = 0.0001, 0.01, auto$, and LDpred2 with $p = 0.001, 0.01, 0.1, auto$, and the infinitesimal model. For MegaPRS, to ensure robust model convergence in all simulation settings, we included 9 models with distinct tuning parameters {p1,p2,p3,p4} = {0.99,0.01,0,0}, {0.95,0.05,0,0}, {0.9,0.1,0,0}, {0.8,0.1,0.05,0.05}, {0.7,0.1,0.1,0.1}, {0.6,0.2,0.1,0.1}, {0.5,0.2,0.2,0.1}, {0.4,0.2,0.2,0.2}, {0,0,0,1}. We repeated this procedure four times and calculated average R^2 to pick the best set of tuning parameters for both approaches.

We conducted additional simulations to demonstrate that PUMAS and PUMAS-ensemble can be applied to binary traits. For each setting in the quantitative simulation study, we dichotomized the continuous phenotype (i.e., true liability value under a liability threshold model) using either the median or 90% quantile to acquire balanced (5-to-5) and unbalanced (1-to-9) case-control ratios.

Therefore, we have a total of 12 binary simulation settings. We fitted logistic regressions in PLINK to obtain GWAS summary statistics in each setting and transformed logistic regression summary statistics to the linear scale(47, 48, 75). We then compared PUMAS/PUMAS-ensemble with MCCV using R^2 computed on the observed scale (i.e., R^2 between PRS and 0-1 status).

We conducted two additional simulation analyses for PUMAS's subsampling scheme. First, we investigated the similarity between PUMAS and MegaPRS under two simulation settings with heritability of 0.5 from six quantitative trait simulation scenarios described above. Since MegaPRS only uses z-scores as its input and output, we focused on z-scores in this simulation. As a benchmark, we compared both approaches with MCCV where we randomly selected a subset of individuals and obtained SNP association statistics from regression analysis. For each approach, we subsampled SNP z-scores based on 75% of samples (total N = 100,000) and repeated this procedure 100 times. We summarized the results for randomly selected 5 causal variants and 5 non-causal variants in each simulation. Second, we investigated the robustness of PUMAS under an extremely sparse simulation setting. We followed the same simulation strategy and set number of causal variants to be 10 on chromosome 1 and heritability to be 0.1. For both approaches, we subsampled summary statistics based on 75% of individuals and repeated this procedure 100 times. We compared the distribution of summary statistics generated from PUMAS and MCCV for each causal SNP.

UKB data analysis

We applied our approach to 16 quantitative traits, 4 diseases, and 1 ordinal trait in UKB. The list of UKB phenotypes is presented in **Additional file 11** and **12**: **Table. S10**-**S11**. The imputed UKB genotype data consists of 375,064 independent individuals of European ancestry and 1,030,187 variants after quality control. We used Hail (v0.2.57) to perform linear regression for quantitative and ordinal traits while adjusting for sex, age polynomials to the power of two, interactions between sex and age polynomials, and top 20 principal components. For 4 disease outcomes, we obtained GWAS summary statistics via regenie (v3.0.3) accounting for sex, age polynomials to the power of 3, interactions between sex and age polynomials, and top 10 principal components as recommended(76).

We compared PUMAS with external validation using a holdout subset of UKB samples. For external validation of quantitative traits, we randomly selected 38,521 samples with non-missing phenotypic measurements for all traits to form the holdout dataset. The remaining samples for each phenotype were used as training data. In this way, we implemented an approximately 9-to-1 training-testing split. Similarly for each binary and ordinal outcome, we continued to employ a 9-to-1 sample partition while matching the case-control ratio between the training and holdout datasets. Detailed sample size information for all traits is included in **Additional file 11** and **12**: **Table. S10**-**S11**. Then, we conducted GWAS on the training data and obtained summary statistics. For quantitative and ordinal traits, we computed and evaluated PRS models on the entire holdout set and reported predictive R^2 between PRS and phenotypes with covariates regressed out. For disease traits, we constructed PRS models and calculated R^2 on the observed scale using both linear probability model summary statistics and logistic model summary statistics. For all phenotypes, the holdout set of quantitative traits (N=38,521) was also used as LD reference data for PRS model training. For comparison, we applied PUMAS to partition the same GWAS summary-level data used in MCCV to 75% training summary statistics and 25% tuning summary

statistics. We used the holdout dataset (N=38,521) for summary statistics subsampling(67) and as the LD reference for lassosum and LDpred2 model training. We estimated variance of PRS models based on a smaller subset (N=1,000) of the holdout data when evaluating PRS performance. This procedure was repeated 4 times and we reported the average R^2 for each PRS model. In all analyses, we implemented lassosum with $s = 0.2, 0.5, 0.9$ and $\lambda = 0.005, 0.01$, PRS-CS with $\phi = 0.0001, 0.01, auto$, LDpred2 with $p = 0.001, 0.01, 0.1, auto$ and the infinitesimal model.

Next, we compared PUMAS-ensemble with the training-testing split approach for ensemble learning on the holdout dataset. For PUMAS-ensemble, we partitioned full GWAS summary statistics into training (60%), tuning (20%), and ensemble training (10%) summary statistics to train PRS models based on a grid of tuning parameters, select the best tuning parameter setting for each PRS method, and fit a second level regression to obtain regression weights for fine-tuned PRS models. We then randomly partitioned the holdout dataset into two equally sized subsets. We used PUMAS-ensemble to obtain PRS models' regression weights and then constructed and evaluated the ensemble PRS on the second half of the holdout set. PRS models with negative weights were removed from linear combination. In comparison, for the training-testing split approach based on individual-level data, we used the first half of the holdout set to fit multiple linear regression to obtain regression coefficients for SDPR and fine-tuned lassosum, LDpred2, and PRS-CS scores. Then we computed and evaluated the ensemble PRS models on the second half of the holdout data. In all analyses, we trained SDPR, lassosum with $s = 0.2, 0.9$ and $\lambda = 0.001, 0.01, 0.1$, PRS-CS with $\phi = 0.0001, 0.01, auto$, LDpred2 with $p = 0.001, 0.01, 0.1, auto$ and the infinitesimal model. As a secondary analysis, we compared performance of PUMAS (70% training, 20% tuning, 10% testing) and PUMAS-ensemble (50% training, 20% tuning, 20% ensemble learning, 10% testing) on 16 quantitative traits in UKB. We benchmarked the best PRS

model chosen by PUMAS and the ensemble score trained by PUMAS-ensemble on the second half of UKB holdout dataset.

To investigate how sensitive PUMAS-ensemble is to LD misspecification, we repeated PUMAS-ensemble analysis on 16 complex traits in UKB with different LD references. Previously, we used UKB LD panel; in this sensitivity analysis, we explored the impact of using 1000 Genomes Project Phase III European samples (1KG EUR) and East Asian samples (1KG EAS) as the LD reference panel, while keeping everything else unchanged. The 1KG LD reference data were prepared from our earlier work(42). We trained ensemble scores by PUMAS-ensemble using different LD reference panels and evaluated these scores on UKB holdout dataset. Results based on individual-level data were used as a benchmark of performance. In addition, we meta-analyzed UKB and Biobank Japan(77-79) (BBJ) GWAS summary statistics for 16 complex traits using METAL(65) and applied PUMAS-ensemble using either 1KG EUR or 1KG EAS data as the LD reference. Sample size information for BBJ GWAS summary statistics is included in **Additional file 17**: **Table. S16**. Similarly, we compared ensemble scores from PUMAS-ensemble and individual-level ensemble learning on the UKB holdout dataset. lassosum with $s = 0.2, 0.9$ and $\lambda = 0.001, 0.01, 0.1$, PRS-CS with $\phi = 0.0001, 0.01, auto$, and LDpred2 with $p = 0.001, 0.01, 0.1, auto$ and the infinitesimal model were considered for ensemble PRS training in this analysis.

Building a catalog of PUMAS-ensemble ensemble scores

We applied PUMAS-ensemble to a collection of publicly available GWAS summary statistics. We selected complex diseases and traits with a minimal case sample size of 5,000 and a total sample

size of 50,000, respectively. We excluded studies that performed GWAS on related samples and retained traits with significant heritability estimation (p-value below 0.05) from LD score regression(50). In the end, we obtained a list of 31 GWAS summary statistics including 23 binary outcomes and 8 complex traits as summarized in **Additional file 18**: **Table. S17**. For each summary statistics, we kept HapMap 3 SNPs that passed a series of quality control criteria listed in **Additional file 19**: **Table. S18**, including transformation of logistic summary statistics and imputation of per-SNP sample size. Then we applied PUMAS-ensemble to each phenotype to implement 4-fold MCCV by partitioning the summary statistics to training (60%), tuning (20%), ensemble training (10%), and testing (10%) datasets. We used 1000 Genomes Project Phase III European samples as the LD panel for summary statistics subsampling, PRS model fitting and benchmarking. We implemented SDPR, lassosum with $s = 0.2, 0.5, 0.9$ and $\lambda = 0.005, 0.01$, PRS-CS with $\phi = 0.0001, 0.01, auto$, LDpred2 with $p = 0.001, 0.01, 0.1, auto$ and the infinitesimal model. We reported average predictive R^2 of ensemble PRS, the best single PRS model, PRS-CS-auto and LDpred2-auto on the testing summary statistics.

We conducted additional analysis to investigate the validity of predictive R^2 of ensemble PRS for Alzheimer's disease. We used IGAP 2019 Alzheimer's GWAS summary statistics to train PRS models and included 2,600 Alzheimer's disease cases of European ancestry from the UKB cohort in the external validation dataset(51). The data fields used for Alzheimer's cases extraction are presented in **Additional file 21**: **Table. S20**. We randomly selected 5,200 independent UKB samples not diagnosed with Alzheimer's disease to use as healthy controls to match the case-control ratio in the IGAP 2019 study. Together, we obtained a UKB external validation dataset with 7,800 samples in total. We applied PUMAS to IGAP 2019 GWAS summary-level data and compared its performance with external validation. We compared R^2 from both approaches with and without removing the *APOE* region from GWAS summary statistics. We excluded the *APOE*

region from PRS analysis by removing variants between base pairs 45,116,911 and 46,318,605 (hg19) on chromosome 19.

Declarations

Ethics approval

Not applicable.

Availability of data and materials

The UKB data were downloaded from UK Biobank Resource (https://www.ukbiobank.ac.uk) under application number 42148(44). PUMAS/ PUMAS-ensemble software is freely available at https://github.com/qlu-lab/PUMAS(80). The source code for PUMAS/ PUMAS-ensemble used in this study is deposited at https://zenodo.org/records/13826837 with DOI: 10.5281/zenodo.13826837. The PUMAS/ PUMAS-ensemble package is under MIT license.

Competing interests

The authors declare no competing interests.

Funding

The authors gratefully acknowledge research support from National Institutes of Health (NIH) grants U01 HG012039 and R21 AG067092, 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). We also acknowledge use of the facilities of the Center for Demography of Health and Aging at the University of Wisconsin-Madison, funded by NIA Center Grant P30 AG017266.

Author's contributions

Z.Z. and Q.L. conceived and design the study.

Z.Z. developed the statistical framework.

Z.Z., T.G., and M.Y. performed statistical analyses.

Z.Z. and Y.W. wrote the software.

S.Z. assisted in preparing and curating summary statistics.

J.M. assisted in developing ensemble PRS approach.

J.M. and Y.W. assisted in UKB data preparation.

J.S. assisted in developing statistical method for subsampling summary statistics.

Q.L. advised on statistical and genetic issues.

Z.Z. and Q.L. wrote the manuscript.

All authors contributed to manuscript editing and approved the manuscript.

Acknowledgements

We thank members of the Social Genomics Working Group at University of Wisconsin for helpful comments. This research has been conducted using the UK Biobank Resource under Application 42148. This study makes use of summary statistics from many GWAS consortia. We thank many GWAS investigators for providing publicly available GWAS summary-level datasets.

Supplementary information

Additional file 1:

Supplementary figures.

Additional file 2: Table S1

PUMAS-ensemble quantitative simulation results in UKB.

Additional file 3: Table S2

PUMAS-ensemble binary simulation results in UKB.

Additional file 4: Table S3

PUMAS quantitative simulation results in UKB.

Additional file 5: Table S4

PUMAS binary simulation results in UKB.

Additional file 6: Table S5

PUMAS-ensemble quantitative simulation results in UKB (including MegaPRS).

Additional file 7: Table S6

PUMAS-ensemble binary simulation results in UKB (including MegaPRS).

Additional file 8: Table S7

Benchmark computational requirement at different stages of PUMAS-ensemble.

Additional file 9: Table S8

Simulated Z scores from different approaches in quantitative trait simulation.

Additional file 10: Table S9

Simulated summary statistics from PUMAS and MCCV under extremely sparse simulation setting.

Additional file 11: Table S10

Description of UKB quantitative traits.

Additional file 12: Table S11

Description of UKB binary and ordinal traits.

Additional file 13: Table S12

Compare PRS prediction accuracy between PUMAS and external validation for quantitative traits in UKB.

Additional file 14: Table S13

Compare PRS prediction accuracy between PUMAS and external validation for binary traits in UKB.

Additional file 15: Table S14

Compare PRS prediction accuracy between PUMAS-ensemble and external validation in UKB.

Additional file 16: Table S15

Compare fine-tuned PRS and ensemble PRS based on different data partitioning on UKB holdout data.

Additional file 17: Table S16

Compare PRS prediction accuracy between PUMAS-ensemble (with LD misspecification) and external validation in UKB.

Additional file 18: Table S17

Description of 31 GWAS summary statistics used in PUMAS-ensemble catalog.

Additional file 19: Table S18

PUMAS-ensemble catalog GWAS summary statistics quality controls.

Additional file 20: Table S19

PRS prediction accuracy in PUMAS-ensemble catalog.

Additional file 21: Table S20

UKB data fields for AD cases extraction.

Additional file 22: Table S21

IGAP 2019 AD PRS prediction accuracy.

References

1. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018;19(9):581-90.
2. Chatterjee N, Shi J, Garcia-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat Rev Genet. 2016;17(7):392-406.
3. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Med. 2020;12(1):44.
4. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 2007;17(10):1520-8.
5. International Schizophrenia C, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009;460(7256):748-52.
6. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565-9.
7. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. 2013;9(2):e1003264.
8. Minnier J, Yuan M, Liu JS, Cai T. Risk Classification with an Adaptive Naive Bayes Kernel Machine Model. J Am Stat Assoc. 2015;110(509):393-404.
9. Wei Z, Wang W, Bradfield J, Li J, Cardinale C, Frackelton E, et al. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. Am J Hum Genet. 2013;92(6):1008-12.
10. Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 2014;24(9):1550-7.
11. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet. 2013;14(7):507-15.
12. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am J Hum Genet. 2015;97(4):576-92.
13. Hu Y, Lu Q, Powles R, Yao X, Yang C, Fang F, et al. Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS Comput Biol. 2017;13(6):e1005589.
14. Márquez-Luna C, Gazal S, Loh P-R, Kim SS, Furlotte N, Auton A, et al. Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. Nature Communications. 2021;12(1):6052.
15. Hu Y, Lu Q, Liu W, Zhang Y, Li M, Zhao H. Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. PLoS Genet. 2017;13(6):e1006836.
16. Maier RM, Zhu Z, Lee SH, Trzaskowski M, Ruderfer DM, Stahl EA, et al. Improving genetic prediction by leveraging genetic correlations among human diseases and traits. Nat Commun. 2018;9(1):989.
17. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet. 2018;50(2):229-37.

18. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50(9):1219-24.
19. Meisner A, Kundu P, Zhang YD, Lan LV, Kim S, Ghandwani D, et al. Combined Utility of 25 Disease and Risk Factor Polygenic Risk Scores for Stratifying Risk of All-Cause Mortality. Am J Hum Genet. 2020;107(3):418-31.
20. Hao L, Kraft P, Berriz GF, Hynes ED, Koch C, Korategere V Kumar P, et al. Development of a clinical polygenic risk score assay and reporting workflow. Nature Medicine. 2022;28(5):1006-13.
21. Kulm S, Marderstein A, Mezey J, Elemento O. A systematic framework for assessing the clinical impact of polygenic risk scores. medRxiv. 2021:2020.04.06.20055574.
22. Mak TSH, Porsch RM, Choi SW, Zhou X, Sham PC. Polygenic scores via penalized regression on summary statistics. Genet Epidemiol. 2017;41(6):469-80.
23. Chen T-H, Chatterjee N, Landi MT, 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. 2020:1-19.
24. Chung W, Chen J, Turman C, Lindstrom S, Zhu Z, Loh PR, et al. Efficient cross-trait penalized regression increases prediction accuracy in large cohorts using secondary phenotypes. Nat Commun. 2019;10(1):569.
25. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. Bioinformatics. 2020;36(22-23):5424-31.
26. Ge T, Chen CY, Ni Y, Feng YA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun. 2019;10(1):1776.
27. Lloyd-Jones LR, Zeng J, Sidorenko J, Yengo L, Moser G, Kemper KE, et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat Commun. 2019;10(1):5086.
28. Yang S, Zhou X. PGS-server: accuracy, robustness and transferability of polygenic score methods for biobank scale studies. Brief Bioinform. 2022;23(2).
29. Pain O, Glanville KP, Hagenaars SP, Selzam S, Fürtjes AE, Gaspar HA, et al. Evaluation of polygenic prediction methodology within a reference-standardized framework. PLOS Genetics. 2021;17(5):e1009021.
30. Wang Y, Tsuo K, Kanai M, Neale BM, Martin AR. Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. Annu Rev Biomed Data Sci. 2022;5:293-320.
31. Privé F, Arbel J, Aschard H, Vilhjálmsson BJ. Identifying and correcting for misspecifications in GWAS summary statistics and polygenic scores. Human Genetics and Genomics Advances. 2022;3(4):100136.
32. Ni G, Zeng J, Revez JA, Wang Y, Zheng Z, Ge T, et al. A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts. Biol Psychiatry. 2021;90(9):611-20.
33. Ruan Y, Lin Y-F, Feng Y-CA, Chen C-Y, Lam M, Guo Z, et al. Improving polygenic prediction in ancestrally diverse populations. Nature Genetics. 2022;54(5):573-80.
34. Ma Y, Zhou X. Genetic prediction of complex traits with polygenic scores: a statistical review. Trends in Genetics. 2021;37(11):995-1011.
35. Jin J, Zhan J, Zhang J, Zhao R, O'Connell J, Jiang Y, et al. MUSSEL: Enhanced Bayesian polygenic risk prediction leveraging information across multiple ancestry groups. Cell Genomics. 2024;4(4).
36. Zhang H, Zhan J, Jin J, Zhang J, Lu W, Zhao R, et al. A new method for multiancestry polygenic prediction improves performance across diverse populations. Nature Genetics. 2023;55(10):1757-68.

37. Zhang J, Zhan J, Jin J, Ma C, Zhao R, O'Connell J, et al. An ensemble penalized regression method for multi-ancestry polygenic risk prediction. Nature Communications. 2024;15(1):3238.
38. Zhao Z, Yi Y, Song J, Wu Y, Zhong X, Lin Y, et al. PUMAS: fine-tuning polygenic risk scores with GWAS summary statistics. Genome Biol. 2021;22(1):257.
39. Picard RR, Cook RD. Cross-Validation of Regression Models. Journal of the American Statistical Association. 1984;79(387):575-83.
40. Zhang Q, Privé F, Vilhjálmsson B, Speed D. Improved genetic prediction of complex traits from individual-level data or summary statistics. Nature Communications. 2021;12(1):4192.
41. Yang S, Zhou X. Accurate and Scalable Construction of Polygenic Scores in Large Biobank Data Sets. Am J Hum Genet. 2020;106(5):679-93.
42. Miao J, Guo H, Song G, Zhao Z, Hou L, Lu Q. Quantifying portable genetic effects and improving cross-ancestry genetic prediction with GWAS summary statistics. Nature Communications. 2023;14(1):832.
43. Zhou G, Zhao H. A fast and robust Bayesian nonparametric method for prediction of complex traits using summary statistics. PLoS Genet. 2021;17(7):e1009697.
44. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203-9.
45. Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, et al. A saturated map of common genetic variants associated with human height. Nature. 2022.
46. Vuckovic D, Bao EL, Akbari P, Lareau CA, Mousas A, Jiang T, et al. The Polygenic and Monogenic Basis of Blood Traits and Diseases. Cell. 2020;182(5):1214-31.e11.
47. Grotzinger AD, Rhemtulla M, de Vlaming R, Ritchie SJ, Mallard TT, Hill WD, et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. Nature Human Behaviour. 2019;3(5):513-25.
48. Matti P, Peter D, Chris CAS. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. The Annals of Applied Statistics. 2013;7(1):369-90.
49. Lee SH, Goddard ME, Wray NR, Visscher PM. A Better Coefficient of Determination for Genetic Profile Analysis. Genetic Epidemiology. 2012;36(3):214-24.
50. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47(3):291-5.
51. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. Nat Genet. 2019;51(3):414-30.
52. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, et al. Gene Dose of Apolipoprotein E Type 4 Allele and the Risk of Alzheimer's Disease in Late Onset Families. Science. 1993;261(5123):921-3.
53. Bellenguez C, Küçükali F, Jansen IE, Kleineidam L, Moreno-Grau S, Amin N, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. Nature Genetics. 2022;54(4):412-36.
54. de Rojas I, Moreno-Grau S, Tesi N, Grenier-Boley B, Andrade V, Jansen IE, et al. Common variants in Alzheimer's disease and risk stratification by polygenic risk scores. Nature Communications. 2021;12(1):3417.
55. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet. 2017;100(4):635-49.
56. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51(4):584-91.

57. Border R, O'Rourke S, de Candia T, Goddard ME, Visscher PM, Yengo L, et al. Assortative mating biases marker-based heritability estimators. Nature Communications. 2022;13(1):660.
58. Privé F, Albiñana C, Arbel J, Pasaniuc B, Vilhjálmsson BJ. Inferring disease architecture and predictive ability with LDpred2-auto. The American Journal of Human Genetics. 2023;110(12):2042-55.
59. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet. 2018;50(9):1335-41.
60. Loh P-R, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nature Genetics. 2015;47(3):284-90.
61. Truong B, Zhou X, Shin J, Li J, van der Werf JHJ, Le TD, et al. Efficient polygenic risk scores for biobank scale data by exploiting phenotypes from inferred relatives. Nat Commun. 2020;11(1):3074.
62. Song L, Liu A, Shi J, Consortium MGoS. SummaryAUC: a tool for evaluating the performance of polygenic risk prediction models in validation datasets with only summary level statistics. Bioinformatics. 2019;35(20):4038-44.
63. Albiñana C, Zhu Z, Schork AJ, Ingason A, Aschard H, Brikell I, et al. Multi-PGS enhances polygenic prediction by combining 937 polygenic scores. Nature Communications. 2023;14(1):4702.
64. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016.
65. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190-1.
66. Nolte IM. Metasubtract: an R-package to analytically produce leave-one-out meta-analysis GWAS summary statistics. Bioinformatics. 2020;36(16):4521-2.
67. Berisa T, Pickrell JK. Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics. 2016;32(2):283-5.
68. Zhang Y, Lu Q, Ye Y, Huang K, Liu W, Wu Y, et al. SUPERGNOVA: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. Genome Biology. 2021;22(1):262.
69. Spence JP, Sinnott-Armstrong N, Assimes TL, Pritchard JK. A flexible modeling and inference framework for estimating variant effect sizes from GWAS summary statistics. bioRxiv. 2022:2022.04.18.488696.
70. Xiang Z, Matthew S. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. The Annals of Applied Statistics. 2017;11(3):1561-92.
71. Song L, Liu A, Consortium M, Shi J, Gejman V, Sanders R, et al. SummaryAUC: a tool for evaluating the performance of polygenic risk prediction models in validation datasets with only summary level statistics. Bioinformatics. 2019.
72. Amemiya T. Some Theorems in the Linear Probability Model. International Economic Review. 1977;18(3):645-50.
73. Privé F, Aschard H, Ziyatdinov A, Blum MGB. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. Bioinformatics. 2018;34(16):2781-7.
74. Speed D, Hemani G, Johnson MR, Balding DJ. Improved heritability estimation from genome-wide SNPs. American journal of human genetics. 2012;91(6):1011-21.
75. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007;81(3):559-75.

76. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, et al. Computationally efficient whole-genome regression for quantitative and binary traits. Nature Genetics. 2021;53(7):1097-103.
77. Kanai M, Akiyama M, Takahashi A, Matoba N, Momozawa Y, Ikeda M, et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nature Genetics. 2018;50(3):390-400.
78. Akiyama M, Ishigaki K, Sakaue S, Momozawa Y, Horikoshi M, Hirata M, et al. Characterizing rare and low-frequency height-associated variants in the Japanese population. Nature Communications. 2019;10(1):4393.
79. Akiyama M, Okada Y, Kanai M, Takahashi A, Momozawa Y, Ikeda M, et al. Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nature Genetics. 2017;49(10):1458-67.
80. Zhao Z, Gruenloh T, Yan M, Wu Y, Sun Z, Miao J, et al. Optimizing and benchmarking polygenic risk scores with GWAS summary statistics. Github: https://github.com/qlu-lab/PUMAS; 2024.