Zhao *et al. Genome Biology* (2021) 22:257
https://doi.org/10.1186/s13059-021-02479-9

Genome Biology

METHOD Open Access

PUMAS: fine-tuning polygenic risk scores with GWAS summary statistics

Zijie Zhao[1†], Yanyao Yi[2†], Jie Song[2], Yuchang Wu[1], Xiaoyuan Zhong[3], Yupei Lin[3], Timothy J. Hohman[4,5], Jason Fletcher[6,7,8] and Qiongshi Lu[1,2,8*]

* Correspondence: qlu@biostat.wisc.edu
† Zijie Zhao and Yanyao Yi contributed equally to this work.
[1] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53703, USA
[2] Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA
Full list of author information is available at the end of the article

Abstract

Polygenic risk scores (PRSs) have wide applications in human genetics research, but often include tuning parameters which are difficult to optimize in practice due to limited access to individual-level data. Here, we introduce PUMAS, a novel method to fine-tune PRS models using summary statistics from genome-wide association studies (GWASs). Through extensive simulations, external validations, and analysis of 65 traits, we demonstrate that PUMAS can perform various model-tuning procedures using GWAS summary statistics and effectively benchmark and optimize PRS models under diverse genetic architecture. Furthermore, we show that fine-tuned PRSs will significantly improve statistical power in downstream association analysis.

Keywords: GWAS, Polygenic risk score, Model tuning, Summary statistics

Background

Accurate prediction of complex traits with genetic data is a major goal in human genetics research and precision medicine [1]. In the past decade, advancements in genotyping and imputation techniques have greatly accelerated discoveries in genome-wide association studies (GWASs) for numerous complex diseases and traits [2]. These data have also enabled statistical learning applications that leverage genome-wide data in genetic risk prediction [3–8]. However, despite these advances, it remains challenging to access, store, and process individual-level genetic data at a large scale due to privacy concerns and high computational burden. With increasingly accessible GWAS summary statistics for a variety of complex traits [9], polygenic risk scores (PRSs) that use marginal association statistics as input enjoy great popularity and have had success in diverse applications [10–12].

With great popularity, there also come great challenges. Prediction accuracy of PRS remains moderate for most phenotypes [13]. Methods have been developed to improve PRS performance by explicitly modeling linkage disequilibrium (LD) [14], incorporating functional annotations and pleiotropy [15, 16], and improving effect estimates through statistical shrinkage [17]. Notably, most PRS models have tuning parameters, including the *p*-value threshold in traditional PRS, the penalty strength in penalized regression models, and the proportion of causal variants in LDpred [14]. Tuning parameters are very common in predictive modeling. When properly selected, these parameters add flexibility to the model and improve prediction accuracy. This is a well-understood problem with a rich literature—a well-known solution is cross-validation [18]. However, most model-tuning methods require individual-level genetic data either as the training dataset or as a validation dataset independent from both the input GWAS and the testing samples. In practice, these data rarely exist, especially when PRS is generated using GWAS summary statistics in the public domain. This has created a significant gap between current conventions in PRS construction and optimal methodologies. Without a method to fine-tune models using summary statistics, it is challenging to benchmark and optimize PRS, thus limiting its clinical utility.

We introduce PUMAS (Parameter-tuning Using Marginal Association Statistics), a novel method to fine-tune PRS models using GWAS summary data. As a general framework, PUMAS can conduct a variety of model-tuning procedures on PRS, including training-testing data split, cross-validation, and repeated learning. Through extensive simulations on realistic genetic architecture, we demonstrate that the performance of PUMAS is as good as methods based on individual-level data. Additionally, we apply PUMAS to GWAS traits with distinct types of genetic architecture and validate our results using well-powered external datasets. Furthermore, we systematically benchmark and optimize PRS for numerous diseases and traits and showcase the immediate benefits of fine-tuned PRSs in downstream applications.

Results

Method overview

Here, we outline the PUMAS framework. Detailed derivations and technical discussions are included in the "Methods" section. There are two key steps in our proposed model-tuning framework (Fig. 1). First, we sample marginal association statistics for a subset of individuals based on the complete GWAS summary data (Eqs. (7) and (12), the "Methods" section). Using this approach, we can generate summary statistics for independent training and validation sets without actually partitioning the samples. Second, we propose an approach to evaluate the predictive performance (e.g., predictive R^2) of PRS using summary statistics in the validation set so that we can select the best model based on its superior performance (Eq. (20), the "Methods" section). These two steps together make it possible to select the best-performing model with only one set of GWAS summary statistics as input.

Simulation results

We conducted simulations using genotype data from the Wellcome Trust Case Control Consortium (WTCCC) to investigate if PUMAS can achieve similar performance compared to classic model-tuning procedures. A total of 15,567 individuals and 322,235 genetic variants were included in the simulation after quality control. We simulated phenotype data with a heritability of 0.5, varying assumptions on SNP effects and proportion of causal variants (the "Methods" section). We used these data to calculate marginal association statistics and ranked SNPs based on association *p*-values. Next, we applied PUMAS to perform 4-fold repeated learning on marginal association statistics

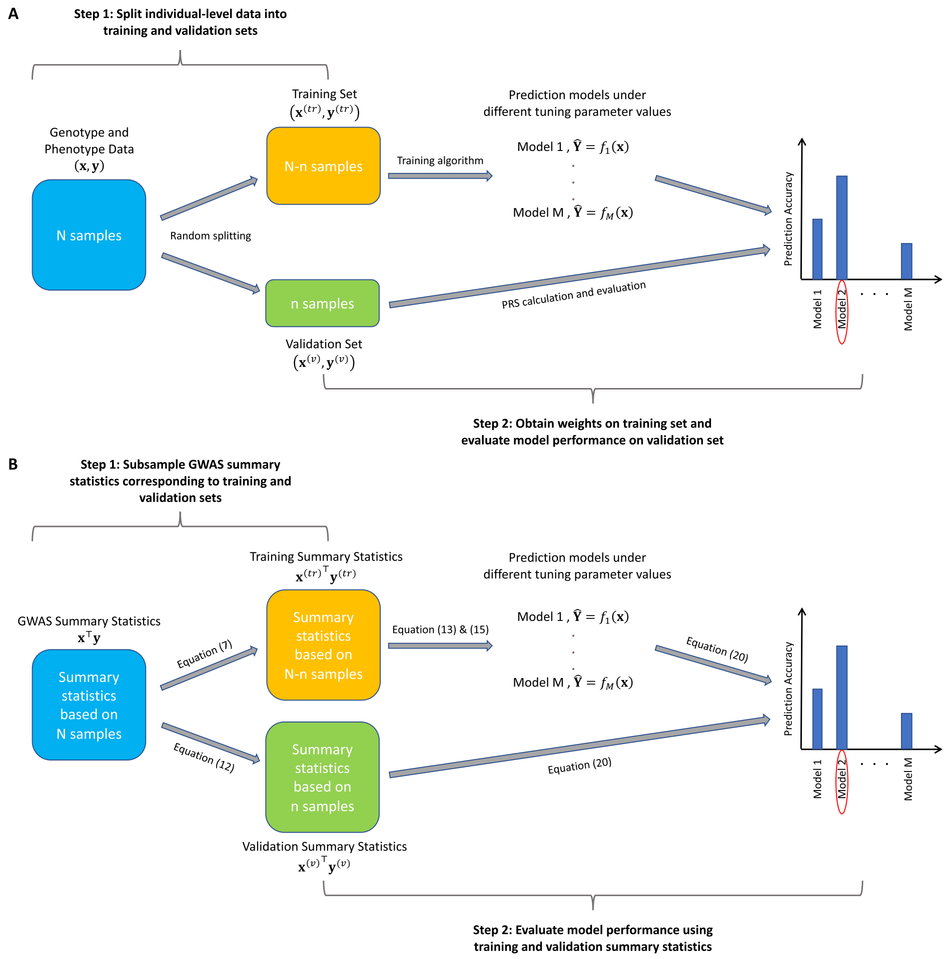
A
Step 1: Split individual-level data into training and validation sets
Genotype and Phenotype Data
$(\mathbf{x}, \mathbf{y})$
N samples
Random splitting
Training Set
$(\mathbf{x}^{(tr)}, \mathbf{y}^{(tr)})$
N-n samples
Training algorithm
Prediction models under different tuning parameter values
Model 1 , $\hat{\mathbf{Y}} = f_1(\mathbf{x})$
Model M , $\hat{\mathbf{Y}} = f_M(\mathbf{x})$
n samples
PRS calculation and evaluation
Validation Set
$(\mathbf{x}^{(v)}, \mathbf{y}^{(v)})$
Prediction Accuracy
Model 1
Model 2
Model M
Step 2: Obtain weights on training set and evaluate model performance on validation set
B
Step 1: Subsample GWAS summary statistics corresponding to training and validation sets
GWAS Summary Statistics
$\mathbf{x}^{\mathrm{T}}\mathbf{y}$
Summary statistics based on N samples
Equation (7)
Equation (12)
Training Summary Statistics
${\mathbf{x}^{(tr)}}^{\mathrm{T}}\mathbf{y}^{(tr)}$
Summary statistics based on N-n samples
Equation (13) & (15)
Prediction models under different tuning parameter values
Model 1 , $\hat{\mathbf{Y}} = f_1(\mathbf{x})$
Model M , $\hat{\mathbf{Y}} = f_M(\mathbf{x})$
Equation (20)
Summary statistics based on n samples
Equation (20)
Validation Summary Statistics
${\mathbf{x}^{(v)}}^{\mathrm{T}}\mathbf{y}^{(v)}$
Prediction Accuracy
Model 1
Model 2
Model M
Step 2: Evaluate model performance using training and validation summary statistics

Fig. 1 A workflow of model-tuning strategies. **A** Traditional approaches split individual-level data into training and validation subsets to fine-tune prediction models. **B** Our method directly generates training and validation summary statistics without using individual-level information and use simulated summary statistics as input to select the best model

and selected the optimal number of SNPs to include in the prediction model by maximizing the average R^2 across folds. Additionally, we implemented a traditional repeated learning approach with the same simulated individual-level data as a reference. Details about our implementation of PUMAS and repeated learning are described in the "Methods" section. Overall, these two approaches yielded equivalent results on both quantitative and binary traits (Fig. 2 and Additional file 1: Fig. S1-S9; Additional file 2 and 3: Table. S1-S2). Across all simulation settings, our summary statistics-based approach showed nearly identical results compared to a state-of-the-art model-tuning approach based on individual-level data and could effectively select the optimal tuning parameter (i.e., number of SNPs in the PRS).

PUMAS effectively fine-tunes PRS models based on genetic architecture

Next, we demonstrate our method's performance using a gold-standard approach—we apply PUMAS to the summary statistics from well-powered GWASs to select the optimal *p*-value cutoffs in PRS models and validate their performance on large independent

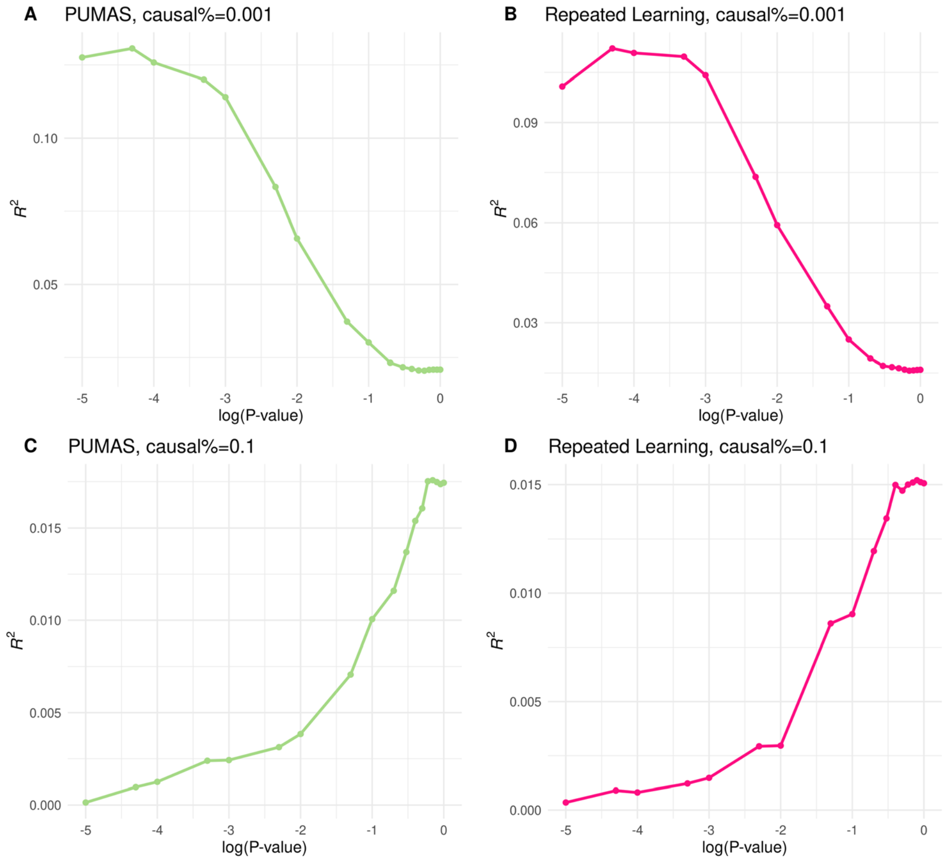

Fig. 2 Comparison of PUMAS and repeated learning. **A**, **C** Model tuning results based on PUMAS. **B**, **D** Results of repeated learning with individual-level data as input. The proportion of causal variants was set to be 0.001 in **A** and **B** and 0.1 in **C** and **D**. The *X*-axis shows the log-transformed *p*-value thresholds. The *Y*-axis shows the predictive performance quantified by average R^2 across four folds. Parameter *α* was set to be 0 in this simulation (the "Methods" section). Results for other settings are summarized in Additional file 1: Fig. S1-S9

cohorts. First, we applied PUMAS to a recent GWAS of educational attainment (EA) conducted by the Social Science Genetic Association Consortium (N = 742,903) [19]. 4775 samples with European ancestry in the National Longitudinal Study of Adolescent to Adult Health (Add Health) [20] and 10,214 European samples in the Health and Retirement Study (HRS) [21] were used as two independent validation sets to assess the predictive performance of EA PRS. We used GWAS of Alzheimer's disease (AD) as a second example. We applied PUMAS to the stage-1 summary statistics from the 2013 study conducted by the International Genomics of Alzheimer's Project (IGAP; N = 54,162) [22] to optimize PRS models for AD. These PRSs were then evaluated on summary-level data of 7050 independent samples [23] from the Alzheimer's Disease Genetics Consortium (ADGC) and individual-level data of 355,583 samples in the UK Biobank with a family history-based proxy phenotype for AD (the "Methods" section) [24].

Our summary statistics-based analyses showed highly consistent results compared with external validations (Fig. 3; Additional file 4: Table. S3). Our analysis clearly suggested that a model with a large number of SNPs tends to be more predictive for EA, a pattern validated in both Add Health and HRS cohorts. The EA PRS based on *p*-value cutoffs of 0.8, 0.8, and 0.7 were the most predictive models suggested by PUMAS, HRS, and Add Health cohorts, respectively. Results on AD were also consistent between PUMAS and external validations. The optimal p-value cutoffs suggested by PUMAS, ADGC validation, and UK Biobank validation were 5e−6, 1e−9, and 1e−10, respectively. PRS models based on p-value cutoffs more stringent than 1e−5 showed good predictive performance in two validation sets for AD. Notably, as more SNPs are included in the model, the predictive performance of PRS sharply declines. Our model-tuning results based on GWAS summary statistics accurately predicted this pattern. Additionally, since we used an AD-proxy phenotype in the UK Biobank, the reduced predictive R^2 is expected. But the trend of predictive performance remained consistent with the validation result in case-control data from the ADGC.

EA is known to be extremely polygenic—more than 1200 independent genetic associations have been identified for EA to date [19]. AD has a very different genetic architecture compared to EA. The *APOE* locus has an unusually large effect on AD risk [25]. In addition to *APOE*, about 30 independent loci have been implicated in AD GWASs [26]. Our method correctly suggested that the EA PRS would perform better if more SNPs are in the model (87,985 SNPs were included) while a substantially sparser model with 31 SNPs would yield better predictive performance for AD. These results showcased our method's ability to adaptively choose the optimal tuning parameter for traits with different patterns of genetic architecture. These results also highlighted the importance of model tuning. An AD PRS based on an arbitrary p-value cutoff of 0.01 can have a 5-fold reduction in predictive R^2 compared to the fine-tuned PRS.

Some technical considerations

We discuss two unique technical issues that may arise in summary statistics-based model tuning. First, sample sizes for different SNPs in a GWAS meta-analysis may vary due to technical differences across cohorts. However, it is not uncommon for a GWAS to only report the maximum sample size. Here, we investigate the robustness of PUMAS when the sample size is mis-specified. We use two GWAS datasets that provided accurate sample size for each SNP: summary statistics for low-density lipoprotein (LDL) cholesterol from the Global Lipids Genetics Consortium (GLGC; N = 188,577) [27] and the same EA GWAS summary statistics we have described before. We compared PUMAS results based on four different approaches. The first approach uses the accurate sample size reported in the summary statistics ("original"). The second approach removes SNPs with sample size below the 30% quantile of its distribution and uses the accurate sample size for the remaining SNPs ("QCed"). The third and fourth approaches apply the maximum or minimum sample size to all SNPs ("Uniform large/small N"). For the "original" and "QCed" approaches where precise sample size is available for each SNP, we assigned 25% of the minimal N value as the sample size for the validation dataset and used the remaining samples of each SNP in the training subset. Overall, PUMAS results showed consistent patterns under these four scenarios (Fig. 4A; Additional file 5: Table. S4). Although the R^2 estimates can inflate or deflate if the sample size is mis-specified, the optimal p-value cutoffs selected by PUMAS remained stable. Thus, PUMAS can still select the best-performing model even if accurate sample size information is unavailable. In practice, performing quality control to remove SNPs with outlier sample size may make the R^2 estimates most interpretable.

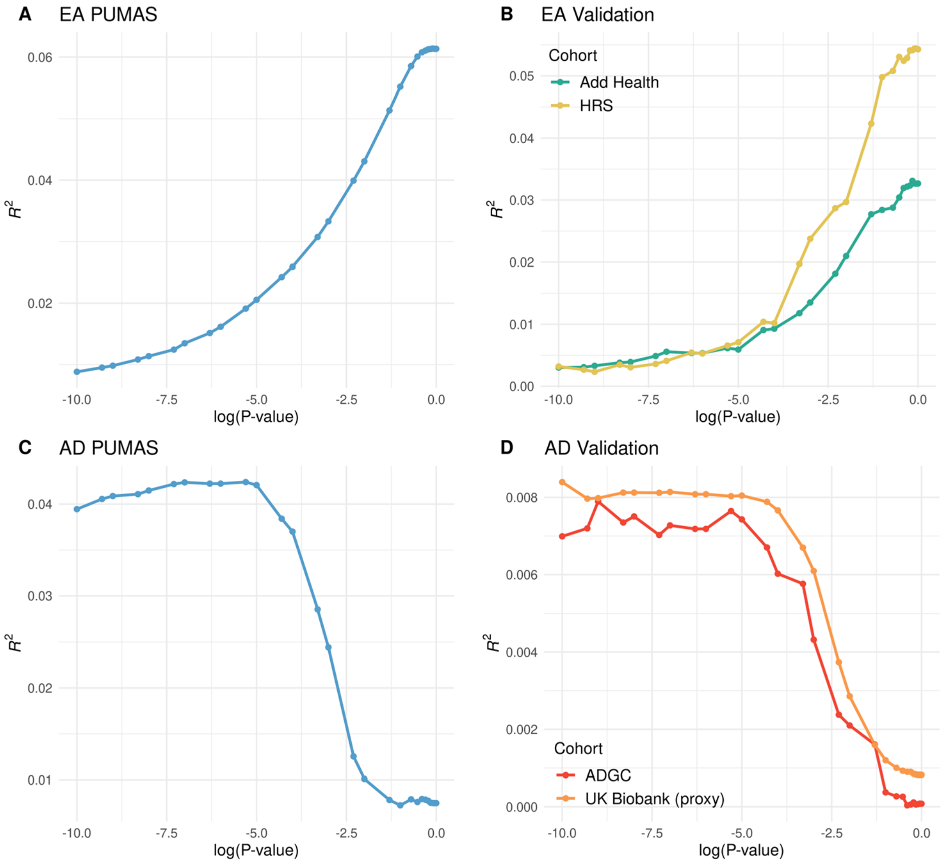

Fig. 3 Model-tuning performance on real GWAS data. **A** PUMAS performance on the EA training set. **B** Prediction performance on two validation sets for EA. **C** PUMAS performance on the AD training set. **D** Prediction performance on two validation sets for AD. The *X*-axis shows the log-transformed *p*-value cutoffs in PRS which is the tuning parameter of interest. The *Y*-axis indicates predictive R^2. EA educational attainment, AD Alzheimer's disease

The second issue is to see if PUMAS can be applied to clumped GWAS summary statistics. In PRS applications, it is a common practice to clump the data by removing SNPs in strong LD with the most significant SNP in a region. However, since *p*-values based on the full sample have been used during LD clumping, directly applying the same model-tuning methods to clumped data may lead to information leak and overfitting. We applied PUMAS to clumped summary statistics of the IGAP 2013 AD GWAS (Additional file 1: Fig. S10). The model-tuning results in PUMAS were completely inconsistent with the optimal models in external validation (Fig. 4B; Additional file 6: Table. S5), confirming that PUMAS should not be applied to clumped data. However, we note that the predictive curves were very similar in external validations no matter if pruned or clumped data were used as input. Therefore, in practice, it may be plausible to apply PUMAS to pruned GWAS summary data and obtain the optimal *p*-value threshold. This way, *p*-values based on the complete sample will not influence the model-tuning procedure. Then, we can apply this selected *p*-value cutoff with clumped GWAS summary statistics to calculate PRS.

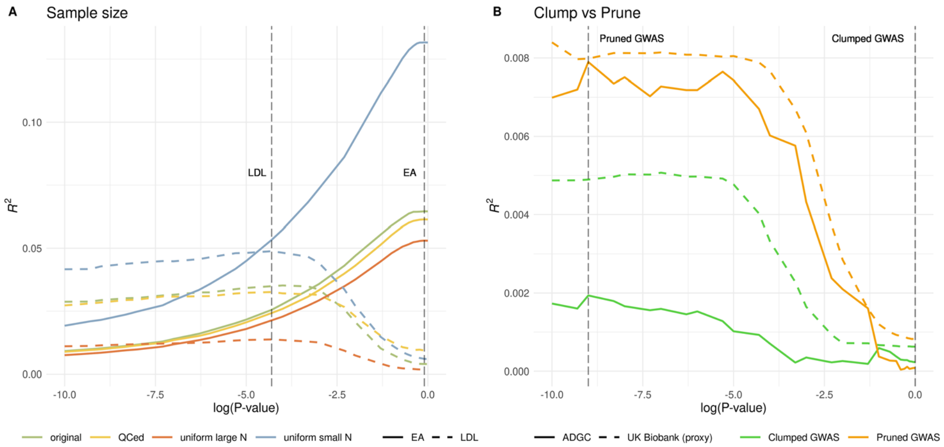

Fig. 4 Technical issues involving sample size and LD clumping. **A** PUMAS results on LDL cholesterol and EA with various sample size specifications. The two gray dashed lines represent the optimal *p*-value cutoffs selected by the "QCed" setting for LDL and EA, respectively. **B** Predictive performance on external validation for AD PRS based on pruned and clumped summary statistics. Two gray dashed lines mark the optimal *p*-value cutoffs inferred by PUMAS on pruned and clumped summary statistics. LDL low-density lipoprotein, EA educational attainment, AD Alzheimer's disease

Benchmarking and optimizing PRS for 65 diseases and traits

Next, we apply PUMAS to provide an atlas of optimized PRSs for complex diseases and traits (Fig. 5). In total, we analyzed 65 GWASs with available summary statistics and documented each trait's optimal *p*-value cutoff and predictive R^2 (Additional file 7: Table. S6). The average gain in predictive R^2 with our method is 0.0106 (205.6% improvement) and 0.0034 (62.5% improvement) compared to PRSs with *p*-value cutoffs of 0.01 and 1, respectively (Additional file 1: Fig. S11 and Additional file 8: Table. S7). We annotated the traits into five categories: behavioral/social, metabolic/cardiovascular, psychiatric/neurological, immune, and others. Most behavioral/social traits and psychiatric/neurological disorders had optimal *p*-value cutoffs between 0.1 and 1 which is consistent with their extreme polygenic genetic architecture. The exceptions include alcoholism (drinks per week), smoking behavior (cigarettes per day), and AD. PRSs with fewer SNPs showed superior performance for these traits. Among immune diseases, systemic lupus erythematosus, primary biliary cirrhosis, rheumatoid arthritis, multiple sclerosis, and eczema all favored a sparse model, while the optimal PRSs for inflammatory bowel diseases and celiac disease had substantially more SNPs. We also note that molecular traits such as blood lipids and 25-hydorxyvitamin D favored sparse PRS models, possibly due to stronger genetic effects and more homogeneous genetic mechanisms. These results also shed light on the differences in the predictive power of diverse types of diseases and traits. PRSs for height, systemic lupus erythematosus, inflammatory bowel diseases, and schizophrenia showed substantially better predictive performance, while the R^2 for most behavioral/social traits remained moderate despite the large sample size in those studies. We also investigated the computational efficiency of our approach in real GWAS applications. Using only one CPU, PUMAS has an average computation time of 8.3 s and maximum of 38.23 s in the analyses of 65 traits (Additional file 1: Fig. S12; Additional file 9: Table. S8), showing computationally scalable performance.

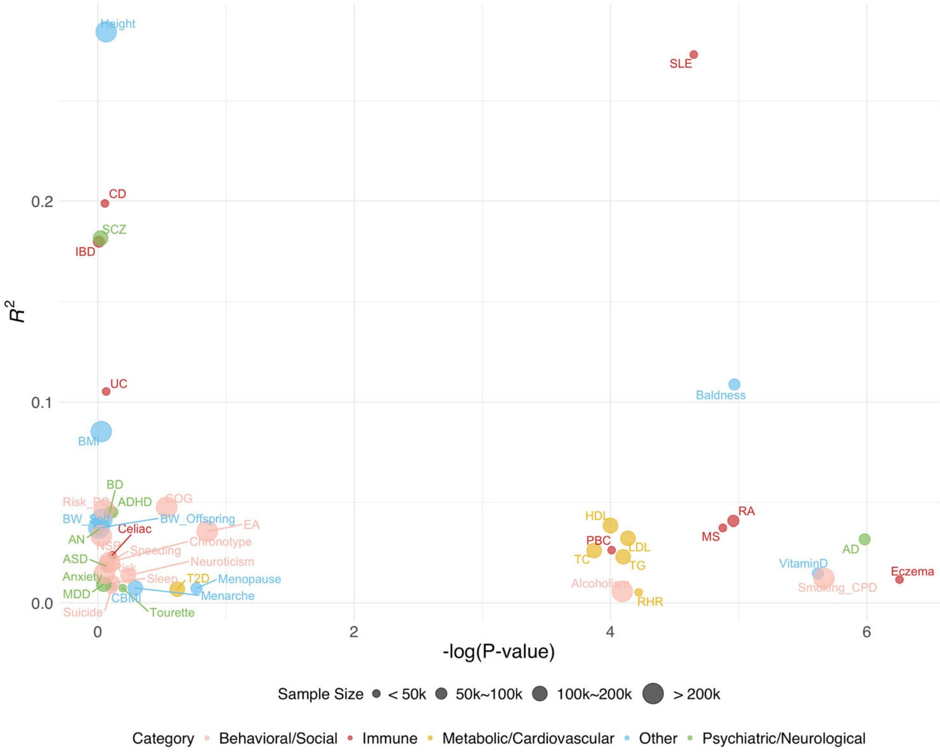

Fig. 5 An atlas of optimized PRSs for complex diseases and traits. 45 diseases/traits with optimized R^2 > 0.005 are included in the figure. Each circle represents a disease or trait. The size of circles indicates the sample size of the study; colors mark the five trait categories. The *X*-axis indicates the negative log-transformed *p*-value cutoff in PRS which is also the tuning parameter of interest. The *Y*-axis indicates the optimal R^2. Information on all diseases and traits is summarized in Additional file 7: Table. S6

Identifying neuroimaging associations for AD

Finally, we demonstrate that fine-tuned PRS will lead to power gain in association analysis. We generated PRSs for 211 neuroimaging traits based on two recent studies conducted using samples from the UK Biobank (N = 17,706 and 19,629 for diffusion tensor imaging traits and regional volume phenotypes, respectively) [28, 29]. We optimized PRS for each imaging trait using PUMAS (Additional file 10: Table. S9). For comparison, we also generated PRSs for all traits using an arbitrary p-value cutoff of 0.01. We applied the BADGERS [30] approach to test associations between 211 neuroimaging trait PRSs with AD in two large, independent AD datasets: the 2019 IGAP GWAS for AD (N = 63,926) [26] and the UK Biobank-based GWAS with a proxy phenotype for AD (N = 318,773) [24, 26]. Samples used in the neuroimaging GWAS were removed from the AD-proxy GWAS to avoid overfitting of PRS models (the "Methods" section; Additional file 1: Fig. S13 and Additional file 11: Table. S10). Association results in two AD datasets were meta-analyzed to improve statistical power.

The complete association results of 211 neuroimaging traits with AD are summarized in Additional file 12: Table. S11. Using fine-tuned PRSs, we identified 2 significant associations with AD under a stringent Bonferroni correction for multiple testing: fornix (cres)/stria terminalis mode of anisotropy (p = 1.7E−05) and axial diffusivities (p = 2.7E−05) whereby genetic risk for worse white matter integrity in the fornix was associated with risk of AD. No significant associations were identified using PRSs with an arbitrary *p*-value cutoff (Fig. 6A). Association *p*-values based on optimized PRSs were significantly lower than those based on arbitrary PRSs (p = 0.03; two-sample Kolmogorov-Smirnov test). Additionally, effect size estimates for top associations were consistent in two independent AD GWASs (Fig. 6B). Although the effect sizes in two AD studies were not at the same scale due to the difference in AD phenotype definition, effect estimates showed strong concordance between two independent analyses (correlation = 0.84). The fornix is a critical white matter tract projecting from the medial temporal lobe where pathology begins in AD; thus, it is unsurprising that microstructural changes in the fornix measured with diffusion tensor imaging are observed in mild cognitive impairment and AD [31–33]. Furthermore, as a negative control, we applied the same analysis to a well-powered breast cancer GWAS (N = 228,951) [34]. Results for fine-tuned PRSs and arbitrary PRSs were consistent with the expectation under the null (Additional file 1: Fig. S14). No significant associations were identified. These findings demonstrated that our model-tuning approach can increase the statistical power in PRS association analysis.

Discussion

Fine-tuning PRS models with GWAS summary statistics has long been considered an impossible task. In this work, we introduced a statistical framework to solve this challenging problem. First, using GWAS summary data as input, PUMAS simulates training and validation summary statistics without accessing individual-level information. Then, PUMAS evaluates and optimizes PRS models on the simulated validation summary statistics. Both steps in the PUMAS framework are statistically rigorous, computationally efficient, and highly novel. Through simulations and analysis of real GWAS data with diverse genetic architecture, we demonstrated that PUMAS can effectively conduct sophisticated model-tuning tasks using GWAS summary statistics. We also showed that

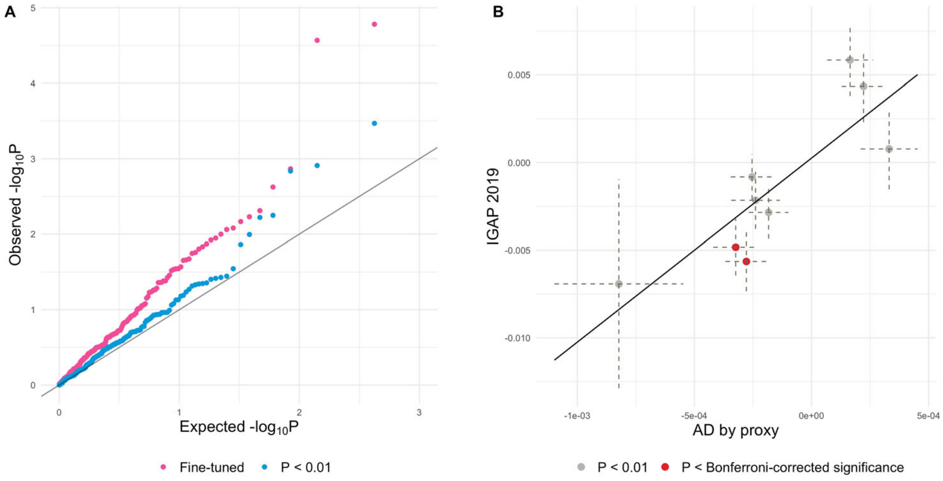

Fig. 6 Identifying neuroimaging trait PRSs associated with AD. **A** QQ plot for the associations between 211 neuroimaging trait PRSs and AD. *p*-values were based on the meta-analysis of IGAP 2019 GWAS and the UK Biobank with a proxy AD phenotype. **B** Effect size estimates for top associations. Imaging trait PRSs that reached a *p*-value < 0.01 in the meta-analysis are shown in the plot. *X*-axis: effect sizes of imaging trait PRSs on the AD-proxy phenotype in the UK Biobank; *Y*-axis: effect sizes on AD in the IGAP 2019 GWAS. Imaging traits whose *p*-value achieved Bonferroni-corrected significance in the meta-analysis are highlighted in red. The dashed lines represent the standard error of effect size estimates

optimizing PRSs improves the statistical power in downstream association analysis and identified neuroimaging traits significantly associated with AD.

This work will bring multiple advances to the field. First, it is no longer necessary to leave one dataset out in the GWAS for model-tuning purpose. With PUMAS, researchers can safely use effect size estimates from the largest available GWAS for PRS model training, which will lead to improved prediction accuracy. Second, when an independent validation set is not available, most studies in the literature select tuning parameters using one of the two strategies. Some studies fine-tune PRSs on testing samples that are used again in downstream applications, creating an overfitting problem, while other studies use a subset of testing samples to tune the model, reducing the sample size and power in the testing data. PUMAS allows researchers to apply fine-tuned PRS models to the full testing samples, thus avoiding overfitting and improving statistical power. Third, selecting the optimal tuning parameter is not the only application of PUMAS. Given a PRS model, our method allows researchers to calculate cross-validated predictive accuracy, providing a systematic approach to benchmark model performance without requiring external samples.

Our proposed framework has some limitations. First, our analyses so far have only focused on a classic PRS model with pruned SNPs and a varying p-value cutoff that needs to be tuned. Despite the simplicity, it remains one of the most widely used PRS models in the field. However, more sophisticated PRS methods have emerged [35–39]. Future work will focus on generalizing PUMAS to fine-tune parameters in other PRS models and benchmarking the performance of all models for different traits. Second, although we demonstrated that PUMAS can also select the optimal tuning parameters for PRS of binary traits, the approximated R^2 metric is less interpretable. A future direction is to explore other metrics (e.g., AUC) to quantify prediction performance for binary traits. Third, our method assumes that GWAS is performed on independent samples. It is an open question whether PUMAS can be directly applied to family-based GWAS results [40–43].

Our results have provided strong evidence that it is possible to fine-tune PRS models with GWAS summary data. This new approach, in conjunction with widely available GWAS summary statistics, will have a long-lasting impact on future PRS model development and genetic prediction applications.

Methods

Step 1: Subsampling GWAS association statistics

Step 1-a: Specify sampling distribution of summary statistics

We assume the quantitative trait Y follows a linear model:

$$Y = \mathbf{X}\boldsymbol{\beta} + \epsilon \tag{1}$$

where $\mathbf{X} = (X_1, \ldots, X_p)$ denotes the random vector of p SNP; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{T}$ is a p-dimensional vector representing fixed SNP effect sizes; ϵ is the error term following a normal distribution with zero mean. Let $\mathbf{y}$ and $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ denote the observed $N \times 1$ phenotypic and $N \times p$ genotypic data of N independent individuals. For simplicity, we assume $\mathbf{y}$ and $\mathbf{x}_j$'s are centered. The summary association statistics in GWAS are obtained from the marginal linear regressions. Then, for $j = 1, \ldots, p$, we can denote the regression coefficients and their standard errors as follows:

$$\hat{\beta}_j = \left(\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j\right)^{-1}\left(\mathbf{x}_j^{\mathrm{T}}\mathbf{y}\right) \tag{2}$$

$$\frac{\mathrm{SE}(\hat{\beta}_j) = \sqrt{\widehat{\varepsilon_j}^{\top}\widehat{\varepsilon_j}}}{(N-1)\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_j} \quad (3)$$

where ϵ_j is the error term for the marginal linear regression of phenotype on the jth SNP and $\widehat{\varepsilon_j} = \mathbf{y} - \mathbf{x}_j\hat{\beta}_j$ is the observed residual from the jth marginal regression. If we have access to the full data set, most model-tuning approaches involve randomly sampling a subset of $N - n$ individuals as the training set, i.e., $\mathbf{y}^{(tr)}$ and $\mathbf{x}^{(tr)}$. Naturally, the remaining subset of n individuals will be the validation dataset denoted as $\mathbf{y}^{(\nu)}$ and $\mathbf{x}^{(\nu)}$. When only the summary statistics file based on the full dataset is provided, the traditional model-tuning approaches cannot be implemented. Instead, we propose a method to generate marginal summary statistics for the training and validating datasets from summary statistics of the full dataset. By central limit theorem, as sample size $N \rightarrow \infty$, we have

$$\mathbf{x}^{\mathrm{T}}\mathbf{y} \sim \mathbf{N}\left(N\mathrm{E}\left(\mathbf{X}^{\mathrm{T}}Y\right), N\mathrm{Var}\left(\mathbf{X}^{\mathrm{T}}Y\right)\right) \tag{4}$$

$$\mathbf{x}^{(tr)^{\mathrm{T}}}\mathbf{y}^{(tr)} \sim \mathbf{N}\left((N-n)\mathrm{E}\left(\mathbf{X}^{T}Y\right), (N-n)\mathrm{Var}\left(\mathbf{X}^{T}Y\right)\right) \tag{5}$$

$$\mathbf{x}^{(\nu)^{\mathrm{T}}}\mathbf{y}^{(\nu)} \sim \mathbf{N}\left(n\mathrm{E}\left(\mathbf{X}^{T}Y\right), n\mathrm{Var}\left(\mathbf{X}^{T}Y\right)\right) \tag{6}$$

where $\mathbf{x}^{T}\mathbf{y}$ is the observed $p \times 1$ summary statistics for N individuals and p genetic markers that can be directly calculated or approximated from the full GWAS summary statistics, and $\mathbf{x}^{(tr)\top}\mathbf{y}^{(tr)}$ and $\mathbf{x}^{(\nu)\top}\mathbf{y}^{(\nu)}$ represent summary statistics for two partitions of full GWAS samples, which are the training set and validation set, respectively. In the following derivation, we use superscripts (tr) and (ν) to indicate whether any summary statistics are computed from $\mathbf{x}^{(tr)\top}\mathbf{y}^{(tr)}$ or $\mathbf{x}^{(\nu)\top}\mathbf{y}^{(\nu)}$. It can be shown that

$$\mathbf{x}^{(tr)^{\mathrm{T}}}\mathbf{y}^{(tr)}|\mathbf{x}^{\mathrm{T}}\mathbf{y} \sim \mathbf{N}\left(\frac{(N-n)}{N}\mathbf{x}^{\mathrm{T}}\mathbf{y}, \frac{(N-n)}{N}\boldsymbol{\Sigma}\right) \quad (7)$$

where $\cdot \mid \cdot$ denotes the conditional distribution and $\boldsymbol{\Sigma}$ is the observed covariance matrix of $\mathbf{x}^{\top}\mathbf{y}$ from the GWAS data. A detailed derivation of this conditional distribution is included in Additional file 1. Note that until now our framework does not depend on the assumption of linkage equilibrium.

Step 1-b: Estimate covariance matrix of summary statistics

To subsample summary statistics, we now estimate the covariance matrix of $\mathbf{x}^{\top}\mathbf{y}$. Under simple scenarios where the SNPs are independent (i.e., GWAS summary statistics is pruned), $\boldsymbol{\Sigma}$ is a symmetric matrix whose diagonal and non-diagonal elements can be denoted as

$$\Sigma_j = \beta_j^2 \mathrm{Var}(X_j^2) + E(\epsilon_j^2)E(X_j^2) \tag{8}$$

$$\Sigma_{ji} = \beta_j\beta_i E(X_j^2)E(X_i^2) \tag{9}$$

For $j = 1, \ldots, p$ and $i \neq j$. Here, $E(\epsilon_j^2)$ can be estimated by the mean squared error in marginal regressions, which can be further approximated by $N[\mathrm{SE}(\widehat{\beta}_j)]^2 E(X_j^2)$. In addition, each SNP's effect size (i.e., β_j) is typically very small in GWAS and $E(X_j^2)$ only depends on each SNP's minor allele frequency (MAF) which is commonly provided in GWAS summary statistics or can be estimated from a reference panel such as the 1000 Genomes Project [44]. Taken together, $\mathbf{\Sigma}$ can be estimated with

$$\hat{\Sigma}_j = N\left[\mathrm{SE}(\hat{\beta}_j)\hat{\sigma}_j^2\right]^2 \tag{10}$$

$$\hat{\Sigma}_{ji} = \hat{\beta}_j\hat{\beta}_i\hat{\sigma}_j^2\hat{\sigma}_i^2 \tag{11}$$

where $\hat{\sigma}_j^2$ is an MAF-based estimator of $E(X_j^2)$.

Step 1-c: Partition summary statistics for training and validation sets

After generating $\mathbf{x}^{(tr)\top}\mathbf{y}^{(tr)}$ terms as described above from the conditional distribution, we can obtain the validating or testing summary statistics by

$$\mathbf{x}^{(v)^{\mathrm{T}}}\mathbf{y}^{(v)} = \mathbf{x}^{\mathrm{T}}\mathbf{y}\text{-}\mathbf{x}^{(tr)^{\mathrm{T}}}\mathbf{y}^{(tr)} \tag{12}$$

Consequently, subsampled GWAS summary statistics for the training set can be estimated by

$$\hat{\beta}_j^{(tr)} = \left[(N-n)\hat{\sigma}_j^2\right]^{-1}\mathbf{x}^{(tr)^{\mathrm{T}}}\mathbf{y}^{(tr)} \tag{13}$$

$$\mathrm{SE}\left(\widehat{\beta}_j^{(tr)}\right) = \sqrt{\frac{N}{N-n}}\mathrm{SE}\left(\widehat{\beta}_j\right) \tag{14}$$

Step 2: Evaluate model performance using GWAS summary data

Step 2-a: Calculate PRS prediction accuracy with summary statistics

Being able to generate summary statistics for the training and validation datasets resolves a critical issue in model tuning. However, challenges remain in evaluating PRS performance on the testing or validation set without individual-level data. Almost all the PRS approaches in the literature use a linear prediction model as follows:

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{w} \tag{15}$$

where $\mathbf{w}^{\top} = (w_1, \ldots, w_p)$ is the weight for SNPs in PRS. In a traditional PRS, marginal regression coefficients from GWAS are used as the weight values, i.e., $\mathbf{w} = \hat{\boldsymbol{\beta}}$, while in other PRS models the weight can be more sophisticated. Here, we demonstrate how to calculate R^2, a commonly used metric to quantify PRS predictive performance, from subsampled GWAS summary data, but our method can be extended to other metrics (e.g., AUC [45]) as well. R^2 on the validation dataset $(\mathbf{y}^{(v)}, \mathbf{x}^{(v)})$ can be calculated as

$$R^2 = \frac{\left(\sum_{i=1}^n y_i^{(v)}\hat{y}_i^{(v)} - n\bar{\mathbf{y}}^{(v)}\bar{\hat{\mathbf{y}}}^{(v)}\right)^2}{\sum_{i=1}^n \left(y_i^{(v)} - \bar{\mathbf{y}}^{(v)}\right)^2 \sum_{i=1}^n \left(\hat{y}_i^{(v)} - \bar{\hat{\mathbf{y}}}^{(v)}\right)^2} \tag{16}$$

where $\hat{\mathbf{y}}^{(v)} = \mathbf{x}^{(v)}\mathbf{w}$ and $\overline{\hat{\mathbf{y}}^{(v)}}$ is the sample mean of $\hat{\mathbf{y}}^{(v)}$. If the SNPs are pruned, it can be shown that the empirical variance of $\hat{\mathbf{Y}}$ can be approximated by

$$\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\hat{y}_i^{(\nu)}-\bar{\hat{\mathbf{y}}}^{(\nu)}\right)^2 \approx \sum\nolimits_{j=1}^{p} w_j^2\hat{\sigma}_j^2 \tag{17}$$

Although empirical variance of Y does not affect model tuning, it affects the scale of R^2 and is thus critical for interpreting the results. This term can be approximated by

$$\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(y_i^{(\nu)}-\bar{\mathbf{y}}^{(\nu)}\right)^2 = \beta_j^2 E\left(X_j^2\right) + E\left(\epsilon_j^2\right), \forall j \tag{18}$$

Although $\mathrm{Var}(Y)$ is always greater than $E\left(\epsilon_j^2\right)$ for any j, the gap between these two is negligible in real GWAS due to the small effect size of each individual SNP. Thus, a simple estimator for $\mathrm{Var}(Y)$ can be

$$\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(y_i^{(\nu)}-\bar{\mathbf{y}}^{(\nu)}\right)^2 \approx \ \max\nolimits_j\left[\frac{1}{N}\hat{\varepsilon}_j^T\hat{\varepsilon}_j\right] \approx N \ \ \max\nolimits_j\left[\mathrm{SE}\left(\hat{\beta}_j\right)^2\hat{\sigma}_j^2\right] \tag{19}$$

Additionally, since we assumed data to be centered, the mean values in the numerator can be dropped. Taken together, R^2 can be estimated as

$$R^2 \approx \frac{\left(\frac{1}{n}\sum_{j=1}^{p} w_j\mathbf{x}_j^{(\nu)^{\mathrm{T}}}\mathbf{y}^{(\nu)}\right)^2}{N \ \ \max_j \left[\mathrm{SE}\left(\hat{\beta}_j\right)^2\hat{\sigma}_j^2\right]\sum_{j=1}^{p} w_j^2\hat{\sigma}_j^2} \tag{20}$$

In practice, we use the 90% quantile of $\frac{\hat{\varepsilon}_j^{\mathrm{T}}\hat{\varepsilon}_j}{N-1}$, $j = 1, 2, \ldots p$, as a more robust estimator for $\mathrm{Var}(Y)$.

Step 2-b: Model-tuning strategies

So far, we have introduced strategies to subsample association statistics on training and validation sets and evaluate model performance using GWAS summary statistics. Combining these two key steps, we will be able to perform model tuning using GWAS summary data. Suppose a PRS model uses GWAS marginal estimates $\hat{\boldsymbol{\beta}}$ as input and generates SNP weights $w_j\left(\hat{\boldsymbol{\beta}}, \lambda\right)$ for each SNP. The goal is to find the optimal value of tuning parameter λ that maximizes the predictive accuracy. In the simple setting we introduced above, we will generate summary statistics for training and validation datasets. After specifying a tuning parameter λ, SNP weights in PRS can be trained by applying the model to the training summary statistics. Then, the prediction accuracy R^2 on the validation summary statistics will be a function of λ. Therefore, we can select λ so that it maximizes model performance.

$$\hat{\lambda} = \mathrm{argmax}_{\lambda}\left(R^2(\lambda)\right) \tag{21}$$

More generally, if the goal is to compare different models, both the summary statistics subsampling and performance evaluation steps remain unchanged. In this case, R^2 will be a function of the model and we can choose the best-performing model by optimizing R^2

$$\hat{m} = \ \mathrm{arg\ max}_{m=1,2\ldots,M}\left(R^2(\mathrm{model\,m})\right) \tag{22}$$

Furthermore, this framework can be used to conduct various types of model-tuning procedures. What we have laid out above is the simple training-validation data split approach. If one is interested in applying repeated learning, they can simply repeat the procedure (i.e., resampling training/validation datasets and evaluating R^2 on the validation set) K times. The average R^2 across K folds can be used to select the best model. Similarly, if K-fold cross-validation needs to be implemented, we can first independently simulate $K - 1$ sets of training subsample $\mathbf{x}^{(tr,k)\top}\mathbf{y}^{(tr,k)}$ with sample size $\frac{N}{K}$. Then, we can obtain the K^{th} subsample by

$$\mathbf{x}^{(K,tr)^T}\mathbf{y}^{(K,tr)} = \mathbf{x}^{\mathrm{T}}\mathbf{y}\text{-}\sum\nolimits_{k=1}^{K-1}\mathbf{x}^{(tr,k)^{\mathrm{T}}}\mathbf{y}^{(tr,k)} \tag{23}$$

Finally, rotate each one of the K subsamples as a validation sample and the rest as a training sample, and use the average R^2 to select the best model. Taken together, PUMAS is a general framework that can perform a variety of model-tuning tasks.

Simulation settings

We conducted simulations using real genotype data from WTCCC. The WTCCC dataset contains 15,918 samples with 393,273 genotyped SNPs across the whole genome. We removed SNPs that are not available in 1000 Genomes Project Phase III European samples from the simulations since 1000 Genomes data were used as the LD reference panel. We excluded individuals with genotype missingness rate higher than 0.01 and removed SNPs that satisfy any of the following conditions: (i) having minor allele frequency less than 0.01, (ii) having missing rate higher than 0.01 among all subjects, and (iii) having p-value from the Hardy-Weinberg equilibrium test lower than 1e−6. After quality control, 322,235 variants and 15,567 samples remained in the analyses. We first simulated effect sizes β_j from a normal distribution $N(0, \frac{h^2}{Mp})$ where h^2 is the heritability (fixed at 0.5), M is the total number of SNPs, and p is the proportion of causal variants. We chose two values of p (i.e., 0.001 and 0.1) to represent sparser and more polygenic genetic architecture. Following the LDAK paper [46], we then replaced the raw effect sizes by $\beta_j^* = \beta_j[2p_j(1-p_j)]^\alpha$ where $\alpha = -2, -1, 0, 1, 2$ to better evaluate the performance of PUMAS under various genetic architecture. Thus, in total, we conducted simulations under 10 different settings. In each setting, causal SNPs were randomly selected across the genome and the effect sizes of non-causal SNPs were set to be 0. Using these simulated effect sizes, we generated continuous trait values in GCTA [47]. We then performed marginal linear regressions and obtained GWAS summary statistics using PLINK [48]. These summary statistics were used as input for PUMAS.

We compared PUMAS with repeated learning (i.e., Monte Carlo cross-validation). Instead of partitioning N samples into k non-overlapping folds, which is what a k-fold cross-validation does, repeated learning randomly selects $\frac{N(k-1)}{k}$ samples to form the training dataset and evaluates the model performance on the remaining $\frac{N}{k}$ samples. This procedure is then repeated k times to obtain an averaged prediction accuracy (i.e., R^2) across k folds of analysis for each prediction model. Here, we implemented a 4-fold repeated learning approach. In each fold, we randomly select 75% of WTCCC samples (i.e., $\frac{3}{4} \times 15567 \approx 11675$) to perform GWAS, and evaluate the predictive performance of PRS on the remaining 25% of individuals (i.e., 15567 – 11675 = 3892). We repeated this process 4 times and reported the average predictive R^2 for each PRS model with different p-value cutoffs. For comparison, we implemented PUMAS in a similar fashion.

Based on the GWAS summary data computed from all WTCCC samples, PUMAS generates a set of summary statistics for 75% of samples as the training data for PRS and evaluates the predictive performance of PRS on the corresponding validation summary statistics (i.e., summary statistics for the remaining 25% of samples). We repeated this procedure 4 times and reported the average R^2 for each PRS model. In this simulation, we consider PRS models with p-value cutoffs ranging from 1e−5 to 1.

To show that PUMAS can be applied to binary traits, we conducted additional simulations under settings described above. For each simulation setting, we kept the same SNP effects, heritability, and proportion of causal variants. However, instead of generating quantitative phenotypes, we simulated binary phenotypes using a population prevalence of 50% and case-control ratio of 1:1 in GCTA [47]. We performed marginal logistic regressions in PLINK to obtain GWAS summary statistics. Like the simulation for quantitative traits, we compared the performance between PUMAS and 4-fold repeated learning using individual-level data for binary traits. We used the area under the ROC curve (AUC) as the metric to assess prediction accuracy in repeated learning.

Finally, to investigate the accuracy of our approximation for the covariance of $\mathbf{x}^T\mathbf{y}$, we performed additional simulations in WTCCC. We implemented a total of 8 settings with different sample sizes, heritability, and proportion of causal variants. We simulated quantitative trait values and performed marginal linear regression to obtain GWAS summary statistics. Then, we calculated and compared diagonal elements Σ_j and $\hat{\Sigma}_j$, and off-diagonal elements Σ_{ji} and $\hat{\Sigma}_{ji}$, respectively. Note that the theoretical values of elements in the covariance matrix are shown in Eqs. (8–9), while the approximated values are shown in Eqs. (10–11). We assessed the approximation of both diagonal and off-diagonal elements. We found a very high correlation between $\hat{\Sigma}_j$ and Σ_j (greater than 0.99 in all of 8 simulation settings) and negligible off-diagonal elements (Additional file 1: Fig. S15; Additional file 13: Table. S12), which justifies our approximation procedure.

GWAS data

GWAS summary statistics on EA was shared to us by Dr. Aysu Okbay. In this dataset, samples from Add Health, HRS, 23&me, and Wisconsin Longitudinal Study were excluded (N = 742,903; number of SNPs = 10,824,042). Imputed genotype data for Add Health (N = 9974; number of SNPs = 9,664,514) and HRS (N = 15,567; number of SNPs = 18,144,468) were accessed through dbGap (phs001367 and phs000428) and the EA phenotypes were defined following the SSGAC GWAS [19]. Both Add Health and HRS genotype data were imputed using 1000 Genome Project data as reference. The comprehensive data cleaning procedure is documented on the Add Health website at https://addhealth.cpc.unc.edu/wp-content/uploads/docs/user_guides/AH_GWAS_QC.pdf. For HRS, SNPs with imputation quality score < 0.8 were removed from the dataset. After matching samples with accessible phenotypic information, 4775 Add Health samples and 10,214 HRS samples with self-reported European ancestry were used to validate EA PRS. The IGAP 2013 AD GWAS (N = 54,162; number of SNPs = 7,055,881) dataset was accessed through the IGAP website (http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php). GWAS summary statistics (number of SNPs = 9,037,014) for 7050 ADGC samples can be accessed through the NIAGADS database

(NG00076). Predictive performance on ADGC samples was assessed using summary statistics-based R^2. Following a recent paper, we constructed the AD-proxy phenotype in the UK Biobank based on each sample's AD status, AD history of parents, whether parents are still alive, and parental age (or age at death) [24]. Imputed genotype data (N = 355,583; number of SNPs = 9,605,099) were accessed through the UKB. The UKB genotype data was imputed to the Haplotype Reference Consortium reference. We removed samples who are not of European ancestry and SNPs with minor allele frequency < 0.01 or imputation R^2 < 0.9. In addition, we applied PUMAS to benchmark PRS performance on 65 GWASs. Details on these studies are summarized in Additional files 7, 8, and 9: Table. S6-S8.

For all PUMAS analysis throughout the paper, we first extracted SNPs intersected with the 1000 Genome Phase III data of European ancestry [44]. Then, we pruned GWAS summary statistics by a LD-block window size of 100 variants, a step size of 5 variants to shift windows, and a pairwise LD (i.e., r^2) threshold of 0.1 using PLINK [48]. We used samples of European ancestry in the 1000 Genome Project Phase III as the reference panel to estimate LD. For GWASs that do not report MAF in the summary statistics, we estimated MAF from 1000 Genome project European samples. In addition, for the analysis of EA and AD, we also intersected GWAS summary statistics with SNPs in the validation set before LD-pruning. A p-value grid was used to search for the optimal p-value cutoff (Additional file 4: Table. S3).

Identifying neuroimaging traits associated with AD

GWAS results for imaging traits were accessed from https://med.sites.unc.edu/bigs2/data/. The IGAP 2019 AD GWAS summary statistics was accessed via NIAGADS (NG00075). We constructed the AD-proxy phenotype in the UK Biobank following a recent paper [24]. To avoid sample overlap between GWASs, we inferred individuals in the UK Biobank who have undergone brain MRI scans and removed them from the AD-proxy GWAS. All individuals who have visited at least one of the UKB imaging centers were removed from the analysis. 318,773 independent samples remained after removing imaging samples from the data. We performed GWAS with the first 12 principal components [49], age, sex, genotyping array, and assessment center as covariates. To test if our approach to remove overlapping samples between neuroimaging GWAS and the AD-proxy analysis was effective, we used cross-trait LD score regression to estimate the intercepts between 211 imaging traits and the AD-proxy GWAS (Additional file 1: Fig. S13) [50]. BADGERS software was used to conduct the imaging trait PRS-AD association analysis [30]. Meta-analysis was conducted using the sample size-weighted approach [51].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02479-9.

Additional file 1: Supplementary notes and figures.

Additional file 2: Table S1. Compare model tuning techniques in WTCCC simulation with continuous phenotype.

Additional file 3: Table S2. Compare model tuning techniques in WTCCC simulation with binary phenotype.

Additional file 4: Table S3. R^2 based on PUMAS and external validations for EA and AD.

Additional file 5: Table S4. PUMAS results on LDL cholesterol and EA with various sample size specifications.

Additional file 6: Table S5. Predictive performance on external validation for AD PRS based on pruned and clumped summary statistics.

Additional file 7: Table S6. Summary of optimized 65 GWAS summary statistics.

Additional file 8: Table S7. Summary of improvement of optimized 65 GWAS summary statistics comparing to PRS with arbitrary *p*-value cutoff.

Additional file 9: Table S8. Computation time for analysis of 65 publicly available GWAS summary statistics.

Additional file 10: Table S9. Summary of fine-tuned PRSs for UK Biobank imaging traits.

Additional file 11: Table S10. Cross-trait LD score regression intercept estimates for 211 imaging traits with UKBB AD-proxy GWAS.

Additional file 12: Table S11. Complete association results between 211 neuroimaging trait PRSs and AD.

Additional file 13: Table S12. Simulation settings for investigating the estimation accuracy of covariance of summary statistics.

Additional file 14. Review history.

Acknowledgements
We thank Dr. Aysu Okbay for generously sharing the EA summary statistics excluding 23andMe, Add Health, HRS, and WLS samples. This study makes use of summary statistics from many GWAS consortia. We thank many GWAS investigators for providing publicly accessible GWAS summary statistics.

Review history
The review history is available as Additional file 14.

Peer review information
Anahita Bishop was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
Q.L. conceived and designed the study. Z.Z., Y.Y., and Q.L. developed the statistical framework. Z.Z., Y.Y., and J.S. performed the statistical analysis. Y.L. implemented the software. J.F. assisted in Add Health and HRS data preparation and interpretation. Y.W. and X.Z. assisted in UK Biobank data processing and analysis. T.J.H helped interpret neuroimaging risk factors for Alzheimer's disease. Q.L. advised on statistical and genetic issues. Z.Z., Y.Y., and Q.L. wrote the manuscript. All authors contributed in manuscript editing and approved the final manuscript.

Funding
This project was supported by the Clinical and Translational Science Award (CTSA) program, through the NIH National Center for Advancing Translational Sciences (NCATS), grant UL1TR000427; National Institute on Aging grant R21 AG067092; and by 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. This research uses data from Add Health and HRS. The Add Health Study is supported by Eunice Kennedy Shriver National Institute of Child Health and Human Development Grant P01HD31921 and GWAS Grants R01HD073342 and R01HD060726, with cooperative funding from 23 other federal agencies and foundations. The HRS is supported by the National Institute on Aging Grants U01AG009740, RC2AG036495, and RC4AG039029 and is conducted by the University of Michigan. No direct support was received from grant P01-HD31921, R01HD073342, R01HD060726, U01AG009740, RC2AG036495, and RC4AG039029 for this analysis. Fletcher acknowledges research support from NIA (P30 AG17266).

Availability of data and materials
The Add Health study data and HRS study data were accessed through dbGap (http://dbgap.ncbi.nlm.nih.gov) with accession codes phs001367 and phs000428 [20, 21]. The UKB data were downloaded from UK Biobank Resource (https://www.ukbiobank.ac.uk) under application number 42148 [52].
The PUMAS software is available at https://github.com/qlu-lab/PUMAS [53]. The PUMAS source code used in this study is deposited at https://zenodo.org/record/5202800#.YRmC3y2cbRZ with DOI: 10.5281/zenodo.5202800. The PUMAS package and source code are under MIT license.
This study used existing software and tools for data analysis. LDSC can be downloaded from https://github.com/bulik/ldsc [54]. GCTA can be downloaded from https://cnsgenomics.com/software/gcta/#Download [47]. PLINK version 1.9 can be downloaded from https://www.cog-genomics.org/plink/1.9/ [48]. PRSice-2 can be downloaded from https://www.prsice.info [55].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details

[1]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53703, USA. [2]Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA. [3]University of Wisconsin-Madison, Madison, WI, USA. [4]Vanderbilt Memory and Alzheimer's Center, Vanderbilt University Medical Center, Vanderbilt University School of Medicine, Nashville, TN, USA. [5]Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, TN, USA. [6]La Follette School of Public Affairs, University of Wisconsin-Madison, Madison, WI, USA. [7]Department of Sociology, University of Wisconsin-Madison, Madison, WI, USA. [8]Center for Demography of Health and Aging, University of Wisconsin-Madison, Madison, WI, USA.

Received: 1 April 2021 Accepted: 25 August 2021
Published online: 06 September 2021

References

1. 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. https://doi.org/10.1038/nrg.2016.27.
2. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008;9(5):356–69. https://doi.org/10.1038/nrg2344.
3. 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. https://doi.org/10.1016/j.ajhg.2013.05.002.
4. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. 2013; 9(2):e1003264. https://doi.org/10.1371/journal.pgen.1003264.
5. Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. Genome Res. 2014;24(9):1550–7. https://doi.org/10.1101/gr.169375.113.
6. 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. https://doi.org/10.1080/01621459.2014.908778.
7. Li C, Yang C, Gelernter J, Zhao H. Improving genetic risk prediction by leveraging pleiotropy. Hum Genet. 2014;133(5): 639–50. https://doi.org/10.1007/s00439-013-1401-5.
8. Maier R, Moser G, Chen GB, Ripke S, Cross-Disorder Working Group of the Psychiatric Genomics C, Coryell W, et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. Am J Hum Genet. 2015;96(2):283–94. https://doi.org/10.1016/j.ajhg.2014.12.006.
9. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nat Rev Genet. 2017; 18(2):117–27. https://doi.org/10.1038/nrg.2016.142.
10. 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. https://doi.org/10.1038/s41588-018-0183-z.
11. Weiner DJ, Wigdor EM, Ripke S, Walters RK, Kosmicki JA, Grove J, et al. Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. Nature genetics. 2017;49(7): 978–85. https://doi.org/10.1038/ng.3863.
12. 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. https://doi.org/10.1038/nature08185.
13. Schrodi SJ, Mukherjee S, Shan Y, Tromp G, Sninsky JJ, Callear AP, et al. Genetic-based prediction of disease traits: prediction is very difficult, especially about the future. Front Genet. 2014;5:162.
14. Vilhjalmsson BJ, Yang J, Finucane HK, Gusev A, Lindstrom S, Ripke S, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am J Hum Genet. 2015;97(4):576–92. https://doi.org/10.1016/j.ajhg.2015.09.001.
15. 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. https://doi.org/10.1371/journal.pcbi.1005589.
16. 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. https://doi.org/10.1371/journal.pgen.1006836.
17. 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. https://doi.org/10.1002/gepi.22050.
18. Zhang P. Model selection via multifold cross validation. Ann Stat. 1993;21(1):299–313.
19. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet. 2018;50(8):1112–21. https://doi.org/10.1038/s41588-018-0147-3.
20. Harris KM, Halpern CT, Whitsel EA, Hussey JM, Killeya-Jones LA, Tabor J, et al. Cohort profile: the national longitudinal study of adolescent to adult health (Add Health). Int J Epidemiol. 2019;48(5):1415–1415k. https://doi.org/10.1093/ije/dyz115.
21. Sonnega A, Faul JD, Ofstedal MB, Langa KM, Phillips JW, Weir DR. Cohort profile: the Health and Retirement Study (HRS). Int J Epidemiol. 2014;43(2):576–85. https://doi.org/10.1093/ije/dyu067.
22. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013;45(12):1452–8. https://doi.org/10.1038/ng.2802.
23. Hu Y, Li M, Lu Q, Weng H, Wang J, Zekavat SM, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. Nature genetics. 2019;51(3):568–76. https://doi.org/10.1038/s41588-019-0345-7.
24. Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet. 2019;51:404-13.
25. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small G, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science (New York, NY). 1993;261(5123):921–3.

26. 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. https://doi.org/10.1038/s41588-019-0358-2.
27. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013;45(11):1274–83.
28. Zhao B, Luo T, Li T, Li Y, Zhang J, Shan Y, et al. GWAS of 19,629 individuals identifies novel genetic variants for regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. bioRxiv. 2019:586339. https://doi.org/10.1038/s41588-019-0516-6.
29. Zhao B, Zhang J, Ibrahim JG, Luo T, Santelli RC, Li Y, et al. Large-scale GWAS reveals genetic architecture of brain white matter microstructure and genetic overlap with cognitive and mental health traits (*n*= 17,706). BioRxiv. 2019;288555. https://doi.org/10.1038/s41380-019-0569-z.
30. Yan D, Hu B, Darst BF, Mukherjee S, Kunkle BW, Deming Y, et al. Biobank-wide association scan identifies risk factors for late-onset Alzheimer's disease and endophenotypes. bioRxiv. 2018:468306.
31. Shim G, Choi KY, Kim D, Suh SI, Lee S, Jeong HG, et al. Predicting neurocognitive function with hippocampal volumes and DTI metrics in patients with Alzheimer's dementia and mild cognitive impairment. Brain and Behavior. 2017;7(9): e00766.
32. Ji F, Pasternak O, Ng KK, Chong JSX, Liu S, Zhang L, et al. White matter microstructural abnormalities and default network degeneration are associated with early memory deficit in Alzheimer's disease continuum. Sci Rep. 2019;9(1): 4749. https://doi.org/10.1038/s41598-019-41363-2.
33. Mayo CD, Mazerolle EL, Ritchie L, Fisk JD, Gawryluk JR. Alzheimer's Disease Neuroimaging I. Longitudinal changes in microstructural white matter metrics in Alzheimer's disease. Neuroimage Clin. 2017;13:330–8. https://doi.org/10.1016/j.nicl.2016.12.012.
34. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017;551(7678):92–4. https://doi.org/10.1038/nature24284.
35. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. Bioinformatics. 2020;36(22-23):5424–31.
36. 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.
37. 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. https://doi.org/10.1038/s41467-019-09718-5.
38. 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.
39. 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. J Am Stat Assoc. 2021;116(533):133-43.
40. 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. https://doi.org/10.1038/s41467-020-16829-x.
41. Wu Y, Zhong X, Lin Y, Zhao Z, Chen J, Zheng B, et al. Estimating genetic nurture with summary statistics of multigenerational genome-wide association studies. Proc Natl Acad Sci U S A. 2021;118(25):e2023184118.
42. Huang K, Wu Y, Shin J, Zheng Y, Siahpirani AF, Lin Y, et al. Transcriptome-wide transmission disequilibrium analysis identifies novel risk genes for autism spectrum disorder. PLoS Genet. 2021;17(2):e1009309. https://doi.org/10.1371/journal.pgen.1009309.
43. Howe LJ, Nivard MG, Morris TT, Hansen AF, Rasheed H, Cho Y, et al. Within-sibship GWAS improve estimates of direct genetic effects. bioRxiv. 2021:2021.03.05.433935.
44. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65. https://doi.org/10.1038/nature11632.
45. Song L, Liu A, Shi J. 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.
46. 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. https://doi.org/10.1016/j.ajhg.2012.10.010.
47. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011; 88(1):76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.
48. 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. Am J Hum Genet. 2007;81(3):559–75. https://doi.org/10.1086/519795.
49. Abraham G, Qiu Y, Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics. 2017;33(17):2776–8. https://doi.org/10.1093/bioinformatics/btx299.
50. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015;47(11):1236–41. https://doi.org/10.1038/ng.3406.
51. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190–1. https://doi.org/10.1093/bioinformatics/btq340.
52. 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.
53. Zhao Z, Yi Y, Song J, Wu Y, Zhong X, Lin Y, et al. Fine-tuning polygenic risk scores with GWAS summary statistics. Github: https://github.com/qlu-lab/PUMAS; 2021.
54. 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. https://doi.org/10.1038/ng.3211.
55. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience. 2019;8(7):giz082.

Publisher's Note