An Ensemble Penalized Regression Method for Multi-ancestry Polygenic Risk Prediction

Jingning Zhang[1,*], Jianan Zhan[2], Jin Jin[3], Cheng Ma[4], Ruzhang Zhao[1], Jared O' Connell[2], Yunxuan Jiang[2], 23andMe Research Team, Bertram L. Koelsch[2], Haoyu Zhang[5,6], Nilanjan Chatterjee[1,7*]

[1] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

[2] 23andMe Inc., Sunnyvale, CA, USA

[3]Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, PA, USA

[4] Department of Statistics, University of Michigan, Ann Arbor, MI, USA

[5]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA

[6] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[7] Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

Conflicts of interest: J.Zhan, YJ, JOC, and BLK are employed by and hold stock or stock options in 23andMe, Inc.

[*]Correspondence to: Jingning Zhang (jzhan218@jhu.edu) and Nilanjan Chatterjee (nilanjan@jhu.edu)

Abstract

Great efforts are being made to develop advanced polygenic risk scores (PRS) to improve the prediction of complex traits and diseases. However, most existing PRS are primarily trained on European ancestry populations, limiting their transferability to non-European populations. In this article, we propose a novel method for generating multi-ancestry Polygenic Risk scOres based on enSemble of PEnalized Regression models (PROSPER). PROSPER integrates genome-wide association studies (GWAS) summary statistics from diverse populations to develop ancestry-specific PRS with improved predictive power for minority populations. The method uses a combination of $\mathcal{L}_1$ (lasso) and $\mathcal{L}_2$ (ridge) penalty functions, a parsimonious specification of the penalty parameters across populations, and an ensemble step to combine PRS generated across different penalty parameters. We evaluate the performance of PROSPER and other existing methods on large-scale simulated and real datasets, including those from 23andMe Inc., the Global Lipids Genetics Consortium, and All of Us. Results show that PROSPER can substantially improve multi-ancestry polygenic prediction compared to alternative methods across a wide variety of genetic architectures. In real data analyses, for example, PROSPER increased out-of-sample prediction R^2 for continuous traits by an average of 70% compared to a state-of-the-art Bayesian method (PRS-CSx) in the African ancestry population. Further, PROSPER is computationally highly scalable for the analysis of large SNP contents and many diverse populations.

Introduction

Tens of thousands of single nucleotide polymorphisms (SNP) have been mapped to human complex traits and diseases through genome-wide association studies (GWAS) [1, 2]. Though each SNP only explains a small fraction of variation of the underlying phenotype, polygenetic risk scores (PRS), which aggregate the genetic effects of many loci, can have a substantial ability to predict traits and stratify populations by underlying disease risks [3-12]. However, as existing GWAS to date have been primarily conducted in European ancestry populations (EUR) [13-16], recent studies have consistently shown that the transferability of EUR-derived PRS to non-EUR populations often is less than ideal and in particular poor for African Ancestry populations [17-22].

Despite growing efforts of conducting genetic research on minority populations [23-26], the gap in sample sizes between EUR and non-EUR populations is likely to persist in the foreseeable future. As the performance of PRS largely depends on the sample size of training GWAS [3, 27], using single ancestry methods [28-32] to generate PRS for a minority population, using data from that population alone may not achieve ideal results. To address this issue, researchers have developed methods for generating powerful PRS by borrowing information across diverse ancestry populations. For example, Weighted PRS [33] combines single-ancestry PRS generated from each population using weights that optimize performance for a target population. Bayesian methods have also been proposed that generate improved PRS for each population by jointly modeling the effect-size distribution across populations [34, 35]. Recently, our group

65 proposed a new method named CT-SLEB [22], which extends the clumping and thresholding (CT)
66 [36] method to multi-ancestry settings. The method uses an empirical-Bayes (EB) approach to
67 estimate effect sizes by borrowing information across populations and a super learning model
68 to combine PRSs under different tuning parameters. However, the optimality of the methods
69 depends on many factors, including the ability to account for heterogeneous linkage
70 disequilibrium (LD) structure across populations and the adequacy of the models for underlying
71 effect-size distribution [3, 27]. In general, our extensive simulation studies and data analyses
72 suggest that no method is uniformly the most powerful, and exploration of complementary
73 methods will often be needed to derive the optimal PRS in any given setting [22].
74
75 In this article, we propose a novel method for generating multi-ancestry Polygenic Risk scOres
76 based on an enSemble PEnalized Regression (PROSPER) using GWAS summary statistics and
77 validation datasets across diverse populations. The method incorporates $\mathcal{L}_1$ penalty functions
78 for regularizing SNP effect sizes within each population, an $\mathcal{L}_2$ penalty function for borrowing
79 information across populations, and a flexible but parsimonious specification of the underlying
80 penalty parameters to reduce computational time. Further, instead of selecting a single optimal
81 set of tuning parameters, the method combines PRS generated across different populations and
82 tuning parameters using a final ensemble regression step. We compare the predictive
83 performance of PROSPER with a wide variety of single- and multi-ancestry methods using
84 simulation datasets from our recent study[22] across five populations (EUR, African (AFR),
85 American (AMR), East Asian (EAS), and South Asian (SAS))[22]. Furthermore, we evaluate these
86 methods using a variety of real datasets from 23andMe Inc. (23andMe), the Global Lipids

Genetics Consortium (GLGC) [37], All of Us (AoU) [38], and the UK Biobank study (UKBB) [39]. Results from these analyses indicate that PROSPER is a highly promising method for generating the most powerful multi-ancestry PRS across diverse types of complex traits. Computationally, PROSPER is also exceptionally scalable compared to other advanced methods.

Results

Method overview

PRSOSPER is a method designed to improve prediction performance for PRS across distinct ancestral populations by borrowing information across ancestries (**Figure 1**). It can integrate large EUR GWAS with smaller GWAS from non-EUR populations. Ideally, individual-level tuning data are needed for all populations, because the method needs optimal parameters from single-ancestry analysis as an input; however, even when data is only available for a target population, PRSOSPER can still be performed, and the PRS will be optimized and validated towards the target population. The method can account for population-specific genetic variants, allele frequencies, and LD patterns and use computational techniques for penalized regressions for fast implementation.

PROSPER

Assuming a continuous trait, we first consider a standard linear regression model for underlying individual-level data for describing the relationship between trait values and genome-wide genetic variants across M distinct populations. Let $\boldsymbol{Y}_i$ denote the $n_i \times 1$ vector of trait values, $\boldsymbol{X}_i$ denote the $n_i \times p_i$ genotype matrix, $\boldsymbol{\beta}_i$ denote the $p_i \times 1$ vector of SNP effects, and $\boldsymbol{\epsilon}_i$ denote the $n_i \times 1$ vector of random errors for the i^{th} population. We assume underlying linear regression models of the form $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i, i = 1, \dots M$; and intend to solve the linear regression system by least square with a combination of $\mathcal{L}_1$ (lasso) [40] and $\mathcal{L}_2$ (ridge) [41] penalties in the form

$$\sum_{1\le i\le M} \frac{1}{n_i}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i)^T(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_i) + \sum_{1\le i\le M} 2\lambda_i\|\boldsymbol{\beta}_i\|_1^1 + \sum_{1\le i_1<i_2\le M} c_{i_1i_2}\left\| \boldsymbol{\beta}_{i_1}^{s_{i_1i_2}} - \boldsymbol{\beta}_{i_2}^{s_{i_1i_2}}\right\|_2^2$$

where $\lambda_i, i = 1, \dots, M$ are the population-specific tuning parameters associated with the lasso penalty; $\boldsymbol{\beta}_{i_1}^{s_{i_1i_2}}$ and $\boldsymbol{\beta}_{i_2}^{s_{i_1i_2}}$ denote the vectors of effect-sizes for SNPs for the i_1-th and i_2-th populations, respectively, restricted to the set of shared SNPs ($s_{i_1i_2}$) across the pair of the populations; and $c_{i_1i_2}, 1 \le i_1 < i_2 \le M$ are the tuning parameters associated with the ridge penalty imposing effect-size similarity across pairs of populations.

In the above, the first part, $\sum_{1\le i\le M} 2\lambda_i\|\boldsymbol{\beta}_i\|_1^1$, uses a lasso penalty. Lasso can produce sparse solution [40] and recent PRS studies that have implemented the lasso penalty in the single-ancestry setting have shown its promising performance [29, 30]. The second part, $\sum_{1\le i_1<i_2\le M} c_{i_1i_2}\left\|\boldsymbol{\beta}_{i_1}^{s_{i_1i_2}} - \boldsymbol{\beta}_{i_2}^{s_{i_1i_2}}\right\|_2^2$, uses a ridge penalty. As it has been widely shown that the causal effect sizes of SNPs tend to be correlated across populations [42, 43], we propose to use the ridge penalty to induce genetic similarity across populations. Compared to the fused lasso [44],

which uses lasso penalty for the differences, we use ridge penalty instead, which allows a small difference in SNP effects across populations rather than truncating them to zero. In addition, the ridge penalty is also computationally more efficient due to its continuous derivative. The solutions for population-specific effect size using the combined lasso and ridge penalties can be sparse.

The estimate of $\boldsymbol{\beta}_i, i = 1, \dots, M$ in the above individual-level linear regression systems can be obtained by minimizing the above least square objective function. Following the derivation of lassosum [29], a single-ancestry method for fitting the lasso model to GWAS summary statistics data, we show that the objective function for individual-level data can be approximated using GWAS summary statistics and LD reference matrices by substituting $\frac{1}{n_i}\boldsymbol{X}_i^T\boldsymbol{X}_i$ by $\boldsymbol{R}_i$, where $\boldsymbol{R}_i$ is the estimated LD matrix based on a reference sample from the i-th population , and $\frac{1}{n_i}\boldsymbol{X}_i^T\boldsymbol{y}_i$, by $\boldsymbol{r}_i$, where $\boldsymbol{r}_i$ is the GWAS summary statistics in the i-th population. Therefore, the objective function of the summary-level model can be written as

$$\sum_{1\le i\le M}(\boldsymbol{\beta}_i^T(\boldsymbol{R}_i+\delta_i\boldsymbol{I})\boldsymbol{\beta}_i-\mathbf{2}\boldsymbol{\beta}_i^T\boldsymbol{r}_i+2\lambda_i\|\boldsymbol{\beta}_i\|_1^1)+\sum_{1\le i_1<i_2\le M}c_{i_1i_2}\left\|\boldsymbol{\beta}_{i_1}^{s_{i_1i_2}}-\boldsymbol{\beta}_{i_2}^{s_{i_1i_2}}\right\|_2^2$$

where the additional tuning parameters $\delta_i, i = 1, \dots, M$, are introduced for regularization of the LD matrices across the different populations [30]. For a fixed set of tuning parameters, the above objective function can be solved using fast coordinate descent algorithms [45] by iteratively updating each element of $\boldsymbol{\beta}_i, i = 1, \dots, M$ (see the section of **Obtain PROSPER solution** in **Methods**).

Reducing tuning parameters

For the selection of tuning parameters, we assume we have access to individual-level data across the different populations which are independent of underlying GWAS from which summary statistics are generated. The above setting involves three sets of tuning parameters, $\{\delta_i\}_{i=1}^{M}$, $\{\lambda_i\}_{i=1}^{M}$, and $\{c_{i_1 i_2}\}_{1\leq i_1<i_2\leq M}$, totaling to the number of $M + M + \frac{M(M-1)}{2}$. As grid search across many combinations of tuning parameter values can be computationally intensive, we propose to reduce the search range by a series of steps. First, we use lassosum2 [30] to analyze GWAS summary statistics and tuning data from each ancestry population by itself and obtain underlying values of optimal tuning parameters, $(\delta_i^0, \lambda_i^0)$ for $i = 1, \dots, M$; if tuning data is only available for the target population, the $(\delta_i^0, \lambda_i^0)$ for non-target i can be optimized towards the target population. For fitting PROSPER, we fix $\delta_i = \delta_i^0$ for $i = 1, \dots, M$ as these are essentially used to regularize estimates of population-specific LD matrices. We note that the optimal $\{\lambda_i\}_{i=1}^{M}$ depend on sample sizes of underlying GWAS (**Supplementary Figure 1**), and thus should not be arbitrarily assumed to be equal across all populations. Considering that the optimal tuning parameters associated with the $\mathcal{L}_1$ penalty function from the single-ancestry analyses should reflect the characteristics of GWAS data, which includes underlying sparsity of effect sizes and sample sizes, we propose to specify the $\mathcal{L}_1$-tuning parameters in PROSPER as $\lambda_i = \lambda\lambda_i^0$, i.e. they are determined by the corresponding tuning parameters from the ancestry-specific analysis except for the constant multiplicative factor λ . Finally, for further computational simplification, we assume that effect sizes across all pairs of populations have a similar degree of homogeneity and thus set all $\{c_{i_1 i_2}\}_{1\leq i_1<i_2\leq M}$ to be equal to c. By using the

above assumptions, the objective function to minimize with respect to $\boldsymbol{\beta}_i, i = 1, \ldots, M$, becomes

$$\sum_{1 \le i \le M} (\boldsymbol{\beta}_i^T (\boldsymbol{R}_i + \delta_i^0 \boldsymbol{I}) \boldsymbol{\beta}_i - \mathbf{2}\boldsymbol{\beta}_i^T \boldsymbol{r}_i + 2\lambda\lambda_i^0 \|\boldsymbol{\beta}_i\|_1^1) + \sum_{1 \le i_1 < i_2 \le M} c \left\| \boldsymbol{\beta}_{i_1}^{s_{i_1 i_2}} - \boldsymbol{\beta}_{i_2}^{s_{i_1 i_2}} \right\|_2^2$$

where λ and c are the only two tuning parameters needed for lasso penalty and genetic similarity penalty, respectively.

Ensemble

Using an ensemble method to combine PRS has been shown to be promising in CT-type methods as opposed to picking an optimal threshold [22, 36]. In general, a specific form of the penalty function, or equivalently a model for prior distribution in the Bayesian framework, may not be able to adequately capture the complex nature of the underlying distribution of the SNPs across diverse populations. We conjecture that when effect size distribution is likely to be mis-specified, an ensemble method, which combines PRS across different values of tuning parameters instead of choosing one optimal set, is likely to improve prediction. Therefore, as a last step, we obtain the final PROSPER model using an ensemble method, super learning [46-48], implemented in the *SuperLearner* R package, to combine PRS generated from various tuning parameter settings and optimized using tuning data from the target population. The super learner we use here was based on three supervised learning algorithms, including lasso [40], ridge [41], and linear regression (see **Methods**).

Results

Methods comparison on simulated data

We conducted simulation analyses on continuous traits under various genetic architectures [22] to evaluate the performance of different methods that can be categorized into five groups: single-ancestry methods trained from target GWAS data (single-ancestry method), single-ancestry methods trained from EUR GWAS data (EUR PRS based method), simple multi-ancestry methods by weighting single-ancestry PRS (weighted PRS), recently published multi-ancestry methods (existing multi-ancestry methods), and our proposed method, PROSPER. Single-ancestry methods include CT [36], LDpred2 [31], and lassosum2 [30]. Existing multi-ancestry methods include PRS-CSx [34] and CT-SLEB [22]. The performance of the methods is evaluated by R^2 measured on validation samples independent of training and tuning datasets. Analyses in this and the following sections are restricted to a total of 2,586,434 SNPs, which are included in either HapMap 3 (HM3) [49] or the Multi-Ethnic Genotyping Arrays (MEGA) chips array [50]. LD reference samples for all five ancestries, EUR, AFR, AMR, EAS, and SAS, in this and the following sections, are from 1000 Genomes Project (Phase 3) [51] (1000G).

The results (**Figure 2**, **Supplementary Figure 2-6, Supplementary Table 1.1-1.5**) show that multi-ancestry methods generally exhibit superior performance compared to single-ancestry methods. Weighted PRS generated from methods modeling LD (LDpred2 and lassosum2) can lead to a noticeable improvement in performance (green bars in **Figure 2**). Notably, PROSPER

shows robust performance uniformly across different scenarios. When the sample size of the target non-EUR population is small ($N_{target} = 15\text{K}$) (**Figure 2a**), PROSPER has comparable performance with other multi-ancestry methods under a high degree of polygenicity ($p_{causal} = 0.01$). However, under the same sample size setting and lower ($p_{causal} = 0.01 \text{ and } 5 \times 10^{-4}$), PRS-CSx and CT-SLEB outperform PROSPER, with the margin of improvement increasing as the strength of negative selection decreases (strong negative selection in **Figure 2a**, mild strong negative selection in **Supplementary Figure 2a**, and no negative selection in **Supplementary Figure 3a**). When the sample size of the target population is large ($N_{target} = 80\text{K}$) (**Figure 2b**, and **Supplementary Figure 2-5 b**), PROSPER almost uniformly outperforms all other methods, particularly for the AFR population.

We further compare the computational efficiency of PROSPER in comparison to PRS-CSx, the state-of-the-art Bayesian method available for generating multi-ancestry PRS. We train PRS models for the two methods using simulated data for chromosome 22 using a single core with AMD EPYC 7702 64-Core Processors running at 2.0 GHz. We observe (**Supplementary Table 2**) that PROSPER is 37 times faster than PRS-CSx (3.0 vs. 111.1 minutes) in a two-ancestry analysis including AFR and EUR; and 88 times faster (6.8 vs. 595.8 minutes) in the analysis of all five ancestries. The memory usage for PRS-CSx is about 2.8 times smaller than PROSPER (0.78 vs. 2.24 Gb in two-ancestry analysis, and 0.84 vs. 2.35 Gb in five-ancestry analysis).

23andMe data analysis

We applied various methods to GWAS summary statistics available from the 23andMe, Inc. to predict two continuous traits, heart metabolic disease burden and height; as well as five binary traits, any cardiovascular disease (any CVD), depression, migraine diagnosis, morning person, and sing back musical note (SBMN). The datasets are available for all five ancestries, African American (AA), Latino, EAS, EUR, and SAS. The methods are tuned and validated on a set of independent individuals of the corresponding ancestry from the 23andMe participant cohort (see the section of **Real data analysis** in **Methods** for data description, and **Supplementary Table 3-4** for sample sizes used in training, tuning and validation).

From the analysis of two continuous traits (**Figure 3** and **Supplementary Table 5.1**), we observe that lassosum2 and its related methods (EUR lassosum2 and weighted lassosum2) generally perform better than CT and LDpred2, and their related methods. On the basis of the advantage of lassosum2, PROSPER further improves the performance, and for most of the settings, outperforms all alternative methods, including PRS-CSx and CT-SLEB. PROSPER demonstrates particularly remarkable improvement for both traits in AA and Latino (26.9 % relative improvement in R^2 over the second-best method on average, yellow cells in **Supplementary Table 5.2**) (first two panels in **Figure 3a-b**). For EAS and SAS, PROSPER is slightly better than other methods, except for heart metabolic disease burden of SAS (the last panel in **Figure 3a**), which has the smallest sample size (~20K).

The results from the analysis of the binary traits (**Figure 4** and **Supplementary Table 5.1**) show that PROSPER generally exhibits better performance (7.8% and 12.3% relative improvement in

logit-scale variance (see **Supplementary Notes**) over CT-SLEB and PRS-CSx, respectively, averaged across populations and traits) (blue and red cells, respectively, in **Supplementary Table 5.2**). A similar trend is observed for the analyses of AA and Latino, where PROSPER usually has the best performance (first two panels in **Figure 4a-e**). In general, no single method can uniformly outperform others. Weighted lassosum2 has outstanding performance for depression (**Figure 4b**), while PROSPER is superior for morning person (**Figure 4d**). PRS-CSx shows a slight improvement in the analysis of migraine diagnosis for EAS populations (last second panel in **Figure 4c**), and CT-SLEB performs the best in the analysis of any CVD for SAS population (last panel in **Figure 4a**).

GLGC and AoU data analysis

Considering the uncommonly huge sample sizes from 23andMe, we further applied alternative methods for the analysis of two other real datasets, GLGC and AoU. The GWAS summary statistics from GLGC for four blood lipid traits, high-density lipoprotein (HDL), low-density lipoprotein (LDL), log-transformed triglycerides (logTG), and total cholesterol (TC), are publicly downloadable and available for all five ancestries, African/Admixed African, Hispanic, EAS, EUR, and SAS (see **Methods** for data description, and **Supplementary Table 3** for sample sizes). Further, we generated GWAS summary statistics data from the AoU study for two anthropometric traits, body mass index (BMI) and height, for individuals from three ancestries, AFR, EUR, and Latino/Admixed American (see **Methods** for data description, and **Supplementary Table 3** for sample sizes). Both the blood lipid traits and anthropometric traits

have corresponding phenotype data available in the UKBB, which we use to perform tuning and validation (see the section of **Real data analysis** in **Methods** for the ancestry composition, and **Supplementary Table 4** for sample sizes). Given the limited sample sizes of genetically inferred AMR ancestry individuals in UKBB, we do not report the performance of PRS on AMR individuals in UKBB.

Results from analysis of four blood lipid traits (**Figure 5** and **Supplementary Table 6.1**) from GLGC and UKBB show that PRS generated by lasso-type methods substantially outperform alternative methods. In particular, we observe that the weighted lassosum2 always outperforms the other two weighted methods. Furthermore, our proposed method, PROSPER, shows improvement over weighted lassosum2 in both AFR and SAS (13.1% and 12.3% relative improvement in R^2, respectively, averaged across traits) (green and orange cells, respectively, in **Supplementary Table 6.2**), but not in EAS. Notably, PROSPER outperforms PRS-CSx and CT-SLEB in most scenarios (34.2% and 37.7% relative improvement in R^2, respectively, averaged across traits and ancestries) (blue and red cells, respectively, in **Supplementary Table 6.2**), with the improvement being particularly remarkable for the AFR population in which PRS development tends to be the most challenging.

The results from AoU and UKBB (**Figure 6** and **Supplementary Table 7.1**) show that PROSPER generates the most predictive PRS for the two analyzed anthropometric traits for the AFR population. It appears that Bayesian and penalized regression methods that explicitly model LD tend to outperform corresponding CT-type methods (CT, EUR CT, and weighted CT) which

303 excluded correlated SNPs. Among weighted methods, both LDpred2 and lassosum2 show major
304 improvement over the corresponding CT method. Further, for both traits, PROSPER shows
305 remarkable improvement over the best of the weighted methods and the two other advanced
306 methods, PRS-CSx and CT-SLEB (91.3% and 76.5% relative improvement in R^2, respectively,
307 averaged across the two traits) (blue and red cells, respectively, in **Supplementary Table 7.2**).
308

309 # Discussion

310

311 In this article, we propose PROSPER as a powerful method that can jointly model GWAS
312 summary statistics from multiple ancestries by an ensemble of penalized regression models to
313 improve the performance of PRS across diverse populations. We show that PROSPER is a
314 uniquely promising method for generating powerful PRS in multi-ancestry settings through
315 extensive simulation studies, analysis of real datasets across a diverse type of complex traits,
316 and considering the most recent developments of alternative methods. Computationally, the
317 method is an order of magnitude faster compared to PRS-CSx [34], an advanced Bayesian method,
318 and comparable to CT-SLEB [22], which derives the underlying PRS in closed forms. We have
319 packaged the algorithm into a command line tool based on the R programming language
320 (https://github.com/Jingning-Zhang/PROSPER).
321

322 We compare PROSPER with a number of alternative simple and advanced methods using both
323 simulated and real datasets. The simulation results show that PROSPER generally outperforms
324 other existing multi-ancestry methods when the target sample size is large (**Figure 2b**).

However, when the sample size of the target population is small (**Figure 2a**), no method performed uniformly the best. In this setting, when the degree of polygenicity is the lowest ($p_{causal} = 5 \times 10^{-4}$), CT-SLEB outperforms other methods by a noticeable margin, and PROSPER performs slightly worse than PRS-CSx. Simulations also show that in the scenario of a highly polygenic trait ($p_{causal} = 0.01$), irrespective of sample size, both weighted lassosum2 and PROSPER tend to enjoy superiority compared to all other methods. In terms of computational time and memory usage, PROSPER is an order of magnitude than PRS-CSx in a five-ancestry analysis. The memory usage for PRS-CSx is smaller than PROSPER, but both are acceptable (**Supplementary Table 2**).

We observe that for the analysis of both continuous and binary traits using 23andMe Inc. data, PROSPER demonstrates a substantial advantage over all other methods for the AA and Latino populations, which have the largest sample sizes among all minority groups. The result is consistent with the superior performance of PROSPER observed in simulation settings when the sample size of the target population is large. However, it is worth noting that even for the two other populations, EAS and SAS, which have much smaller sample sizes, PROSPER still performs the best in half of the settings (the last two panels in **Figure 3a-b** and **Figure 4a-e**). For the prediction of blood lipid traits, methods built upon the lasso penalty (lassosum2, weighted lassosum2, PROSPER) perform substantially better than all other alternative methods. Intuitively, this might result from the robustness of the heavy-tail lasso penalty function in dealing with large-effect loci that tend to be present for molecular traits, such as lipid levels (**Supplementary Table 8**), and sometimes for complex traits as well. For the analysis of two

anthropometric traits using training data from AoU, we observe that methods that explicitly model and account for LD differences (e.g. lassosum2, LDpred2, and their corresponding weighted methods) generally achieve higher predictive accuracy than CT-based methods which discard correlated SNPs. In addition, we observe major improvement in PRS performance using PROSPER over weighted lassosum2 and all other existing multi-ancestry methods. The result is consistent with what we have observed in simulation settings under extreme polygenic architectures as expected for complex traits like height and BMI. In conclusion, our results show that PROSPER is a promising method for handling complex traits of diverse genetic architectures.

PROSPER, while showing promising results in our simulations and real data analyses, does have several limitations. Specifically, when the sample size for the training sample for a target population is small, particularly for traits with low polygenicity, the method may not perform as well as some of the other existing methods (**Figure 2a**). Additionally, the use of a super learning step in PROSPER can lead to poorer performance compared to weighted lassosum2 when the sample size for the tuning dataset is not adequately large. In the analysis of lipid traits for EAS, for example, we observe lower predictive accuracy of PROSPER than weighted lassosum2 (the middle panel in **Figure 5b** and **d**). This can be attributed to overfitting in the tuning sample, as the number of tuning samples of EAS origin in the UKBB is only ~1000, while the number of PRSs combined in the super learning step is close to 500.

PROSPER and a number of other recent methods have been developed for modeling summary statistics data across discrete populations typically defined by self-reported ancestry information. However, there is an emerging need to consider the underlying continuum of genetic diversity across populations in both the development and implementational of PRS in diverse populations in the future [52]. Towards this goal, a recent method called GAUDI [53] has been proposed based on the fused lasso penalty for developing PRS in admixed population using individual-level data. While GAUDI shares similarities with PROSPER in terms of the use of the lasso-penalty function, the two methods are distinct in terms of the specification of tuning parameters and use of the ensemble step. Future studies are merited to extend PROSPER for handling data with continuous genetic ancestry information.

To conclude, we have proposed PROSPER, a statistically powerful and computationally scalable method for generating multi-ancestry PRS using GWAS summary statistics and additional tuning and validation datasets across diverse populations. While no method is uniformly powerful in all settings, we show that PROSPER is the most robust among a large variety of recent methods proposed across a wide variety of settings. As individual-level data from GWAS of diverse populations becomes increasingly available, PROSPER and other methods will require additional considerations for incorporating continuous genetic ancestry information, both global and local, into the underlying modeling framework.

Author Contribution Statement

J.Zhang and NC conceived the project. J.Zhang, J.Zhan, JJ, and HZ carried out all data analyses with supervision from NC; HZ created all simulated data and ran GWAS on simulated training data with the supervision from NC; J.Zhan, JOC, YJ run GWAS for training data from 23andMe Inc. with the supervision from BLK; RZ ran GWAS on AoU training data with the supervision from NC and HZ; J.Zhang and CM developed the PROSPER software; J.Zhang and NC drafted the manuscript, and HZ, JJ provided comments. All co-authors reviewed and approved the final version of the manuscript. The following members of the 23andMe Research Team contributed to this study: Stella Aslibekyan, Adam Auton, Elizabeth Babalola, Robert K. Bell, Jessica Bielenberg, Katarzyna Bryc, Emily Bullis, Daniella Coker, Gabriel Cuellar Partida, Devika Dhamija, Sayantan Das, Sarah L. Elson, Nicholas Eriksson, Teresa Filshtein, Alison Fitch, Kipper Fletez-Brant, Pierre Fontanillas, Will Freyman, Julie M. Granka, Karl Heilbron, Alejandro Hernandez, Barry Hicks, David A. Hinds, Ethan M. Jewett, Yunxuan Jiang, Katelyn Kukar, Alan Kwong, Keng-Han Lin, Bianca A. Llamas, Maya Lowe, Jey C. McCreight, Matthew H. McIntyre, Steven J. Micheletti, Meghan E. Moreno, Priyanka Nandakumar, Dominique T. Nguyen, Elizabeth S. Noblin, Jared O'Connell, Aaron A. Petrakovitz, G. David Poznik, Alexandra Reynoso, Morgan Schumacher, Anjali J. Shastri, Janie F. Shelton, Jingchunzi Shi, Suyash Shringarpure, Qiaojuan Jane Su, Susana A. Tat, Christophe Toukam Tchakouté, Vinh Tran, Joyce Y. Tung, Xin Wang, Wei Wang, Catherine H. Weldon, Peter Wilton, Corinna D. Wong.

Acknowledgements

We would like to thank the research participants and employees of 23andMe, Inc. for making this work possible. We want to thank Liz Noblin, Melissa J. Francis and Emily Voeglein for helping with the research collaboration agreement with Harvard T.H. Chan School of Public Health, Johns Hopkins Bloomberg School of Public Health and 23andMe, Inc. The analysis utilized the Joint High Performance Computing Exchange at Johns Hopkins Bloomberg School of Public Health. The UK Biobank data was obtained under the UK Biobank resource application 17731. This work was funded by NIH grants: R01 HG010480-01 (J.Zhang, JJ and NC), K99 CA256513-01 (HZ), U01 HG011719 (NC) and K99 HG012223 (JJ). The All of Us Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2 OD025276. In addition, the All of Us Research Program would not be possible without the partnership of its participants.

Code and Data availability

PROSPER command line tool: https://github.com/Jingning-Zhang/PROSPER

CT: https://www.cog-genomics.org/plink/1.9/

Lassosum2 and LDpred2: https://github.com/privefl/bigsnpr

PRS-CSx: https://github.com/getian107/PRScsx

CT-SLEB: https://github.com/andrewhaoyu/CTSLEB

PLINK: https://www.cog-genomics.org/plink/1.9/; https://www.cog-genomics.org/plink/2.0/

Simulated genotype data for 600K subjects from five ancestries:

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/COXHAP

The full GWAS summary statistics for the 23andMe discovery data set could be made available through 23andMe to qualified researchers under an agreement with 23andMe that protects the privacy of the 23andMe participants. Please visit https://research.23andme.com/collaborate/#dataset-access/ for more information and to apply to access the data. Participants provided informed consent and participated in the research online, under a protocol approved by the external AAHRPP-accredited IRB, Ethical & Independent Review Services.

GWAS summary level statistics for five ancestries from GLGC:

http://csg.sph.umich.edu/willer/public/glgc-lipids2021/results/ancestry_specific/

GWAS summary level statistics for three ancestries from AoU are available upon request.

Codes for simulation and data analyses in this paper: https://github.com/Jingning-Zhang/PROSPER_analysis

The full GWAS summary statistics for the 23andMe discovery data set could be made available through 23andMe to qualified researchers under an agreement with 23andMe that protects the privacy of the 23andMe participants. Please visit https://research.23andme.com/collaborate/#dataset-access/ for more information and to apply to access the data. Participants provided informed consent and volunteered to

454 participate in the research online, under a protocol approved by the external AAHRPP-

455 accredited IRB, Ethical & Independent (E&I) Review Services. As of 2022, E&I Review Services is

456 part of Salus IRB (https://www.versiticlinicaltrials.org/salusirb)

457

458 References

459 1. Buniello, A. *et al*. The NHGRI-EBI GWAS Catalog of published genome-wide association
460 studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005-D1012
461 (2019).

462 2. Visscher, P. M. *et al*. 10 years of GWAS discovery: biology, function, and translation. *The*
463 *American Journal of Human Genetics* **101**, 5-22 (2017).

464 3. Chatterjee, N. *et al*. Projecting the performance of risk prediction based on polygenic
465 analyses of genome-wide association studies. *Nat. Genet.* **45**, 400-405 (2013).

466 4. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk
467 prediction models for stratified disease prevention. *Nature Reviews Genetics* **17**, 392
468 (2016).

469 5. Sugrue, L. P. & Desikan, R. S. What are polygenic scores and why are they important?
470 *JAMA* **321**, 1820-1821 (2019).

471 6. Aragam, K. G. & Natarajan, P. Polygenic scores to assess atherosclerotic cardiovascular
472 disease risk: clinical perspectives and basic implications. *Circ. Res.* **126**, 1159-1177 (2020).

473 7. Ma, Y. & Zhou, X. Genetic prediction of complex traits with polygenic scores: a statistical
474 review. *Trends in Genetics* **37**, 995-1011 (2021).

475 8. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores.
476 *Hum. Mol. Genet.* **28**, R133-R142 (2019).

477 9. Wray, N. R. *et al*. From basic science to clinical application of polygenic risk scores: a
478 primer. *JAMA psychiatry* **78**, 101-109 (2021).

479 10. Mavaddat, N. *et al*. Polygenic risk scores for prediction of breast cancer and breast
480 cancer subtypes. *The American Journal of Human Genetics* **104**, 21-34 (2019).

11. Dikilitas, O. *et al*. Predictive utility of polygenic risk scores for coronary heart disease in three major racial and ethnic groups. *The American Journal of Human Genetics* **106**, 707-716 (2020).

12. Li, R., Chen, Y., Ritchie, M. D. & Moore, J. H. Electronic health records and polygenic risk scores for predicting disease risk. *Nature Reviews Genetics* **21**, 493-502 (2020).

13. Fatumo, S. *et al*. A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**, 243-250 (2022).

14. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161-164 (2016).

15. Peterson, R. E. *et al*. Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589-603 (2019).

16. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The missing diversity in human genetic studies. *Cell* **177**, 26-31 (2019).

17. Martin, A. R. *et al*. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584-591 (2019).

18. Cavazos, T. B. & Witte, J. S. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *Human Genetics and Genomics Advances* **2**, 100017 (2021).

19. Tanigawa, Y. *et al*. Significant sparse polygenic risk scores across 813 traits in UK Biobank. *PLoS Genetics* **18**, e1010105 (2022).

20. Duncan, L. *et al*. Analysis of polygenic risk score usage and performance in diverse human populations. *Nature communications* **10**, 1-9 (2019).

21. Duncan, L. *et al*. Analysis of polygenic risk score usage and performance in diverse human populations. *Nature communications* **10**, 1-9 (2019).

22. Zhang, H. *et al*. Novel Methods for Multi-ancestry Polygenic Prediction and their Evaluations in 3.7 Million Individuals of Diverse Ancestry. *bioRxiv* (2022).

23. Wojcik, G. L. *et al*. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514-518 (2019).

24. Mahajan, A. *et al*. Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nat. Genet.* **54**, 560-572 (2022).

512 25. Bentley, A. R. *et al*. Multi-ancestry genome-wide gene–smoking interaction study of
513 387,272 individuals identifies new loci associated with serum lipids. *Nat. Genet.* **51**, 636-
514 648 (2019).

515 26. Partanen, J. J. *et al*. Leveraging global multi-ancestry meta-analysis in the study of
516 Idiopathic Pulmonary Fibrosis genetics. *Cell Genomics* **2**, 100181 (2022).

517 27. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS genetics* **9**,
518 e1003348 (2013).

519 28. Vilhjálmsson, B. J. *et al*. Modeling linkage disequilibrium increases accuracy of polygenic
520 risk scores. *The american journal of human genetics* **97**, 576-592 (2015).

521 29. Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via
522 penalized regression on summary statistics. *Genet. Epidemiol.* **41**, 469-480 (2017).

523 30. Privé, F., Arbel, J., Aschard, H. & Vilhjálmsson, B. J. Identifying and correcting for
524 misspecifications in GWAS summary statistics and polygenic scores. *Human Genetics and*
525 *Genomics Advances* **3**, 100136 (2022).

526 31. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics*
527 **36**, 5424-5431 (2020).

528 32. Ge, T., Chen, C., Ni, Y., Feng, Y. A. & Smoller, J. W. Polygenic prediction via Bayesian
529 regression and continuous shrinkage priors. *Nature communications* **10**, 1-10 (2019).

530 33. Márquez-Luna, C., Loh, P., South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA
531 Type 2 Diabetes Consortium & Price, A. L. Multiethnic polygenic risk scores improve risk
532 prediction in diverse populations. *Genet. Epidemiol.* **41**, 811-823 (2017).

533 34. Ruan, Y. *et al*. Improving polygenic prediction in ancestrally diverse populations
534 . *Nat. Genet.* **54**, 573-580 (2022).

535 35. Cai, M. *et al*. A unified framework for cross-population trait prediction by leveraging the
536 genetic correlation of polygenic traits. *The American Journal of Human Genetics* **108**, 632-
537 655 (2021).

538 36. Privé, F., Vilhjálmsson, B. J., Aschard, H. & Blum, M. G. Making the most of clumping and
539 thresholding for polygenic scores. *The American Journal of Human Genetics* **105**, 1213-1221
540 (2019).

541 37. Graham, S. E. *et al*. The power of genetic diversity in genome-wide association studies of
542 lipids. *Nature* **600**, 675-679 (2021).

543 38. All of Us Research Program Investigators. The "All of Us" research program. *N. Engl. J.*
544 *Med.* **381**, 668-676 (2019).

545 39. Allen, N. E., Sudlow, C., Peakman, T., Collins, R. & Uk biobank. UK biobank data: come
546 and get it. *Science translational medicine* **6**, 224ed4 (2014).

547 40. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal*
548 *Statistical Society: Series B (Methodological)* **58**, 267-288 (1996).

549 41. Hoerl, A. E. & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal
550 problems. *Technometrics* **12**, 55-67 (1970).

551 42. Brown, B. C., Ye, C. J., Price, A. L., Zaitlen, N. & Asian Genetic Epidemiology Network Type
552 2 Diabetes Consortium. Transethnic genetic-correlation estimates from summary statistics.
553 *The American Journal of Human Genetics* **99**, 76-88 (2016).

554 43. Mishra, A. *et al*. Stroke genetics informs drug discovery and risk prediction across
555 ancestries. *Nature* **611**, 115-123 (2022).

556 44. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. Sparsity and smoothness via
557 the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**,
558 91-108 (2005).

559 45. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear
560 models via coordinate descent. *Journal of statistical software* **33**, 1 (2010).

561 46. Van der Laan, M. J., Polley, E. C. & Hubbard, A. E. Super learner. *Statistical applications in*
562 *genetics and molecular biology* **6** (2007).

563 47. Polley, E. C. & Van Der Laan, M. J. Super learner in prediction. (2010).

564 48. Van der Laan, M. J. & Rose, S. in *Targeted learning: causal inference for observational and*
565 *experimental data* (Springer, 2011).

566 49. International HapMap 3 Consortium. Integrating common and rare genetic variation in
567 diverse human populations. *Nature* **467**, 52 (2010).

568 50. Bien, S. A. *et al*. Strategies for enriching variant coverage in candidate disease loci on a
569 multiethnic genotyping array. *PloS one* **11**, e0167758 (2016).

570 51. 1000 Genomes Project Consortium. A global reference for human genetic variation.
571 *Nature* **526**, 68-74 (2015).

572 52. Ding, Y. *et al*. Polygenic scoring accuracy varies across the genetic ancestry continuum
573 in all human populations. *bioRxiv*, 2022.09. 28.509988 (2022).

574 53. Sun, Q. *et al*. Improving polygenic risk prediction in admixed populations by explicitly
575 modeling ancestral-specific effects via GAUDI. *bioRxiv* (2022).

Online Methods

Data preparation and formatting in PROSPER. We match SNPs and their alleles in GWAS summary statistics and genotypes of individuals for tuning and validation purposes to that in 1000G reference data (phase 3) [51]. To simplify computing huge-dimensional LD matrix, we use existing LD block information from EUR [29] to divide the whole genome, and assume the blocks to be independent. We use PLINK1.9 [54] with flag --r bin4 to compute the LD matrix within each block in each ancestry for common SNPs (MAF>0.01) either in HM3 [49] or the MEGA [50]. For SNPs not common in all populations, we only model them in the populations where they are common; if an SNP is population-specific that is only common in one population, we model it only using the lasso penalty without the genetic similarity penalty. The parameter path of the tuning parameter λ for the scale factor in lasso penalty is set to a sequence evenly spaced on a logarithmic scale from $\lambda^{\max} = \min_{1\le i\le m}\left(\frac{\max_{1\le k\le p}(|r_{ik}|)}{\lambda_i^0}\right)$ to $\lambda^{\min} = 0.001 \times \lambda^{\max}$ which is set to guarantee non-zero solutions, where r_{ik} is the GWAS summary statistics for the k-th SNP in the i-th population, and λ_i^0 is the underlying values of optimal tuning parameter λ for the i-th population. The parameter path for the tuning parameter c for the genetic similarity penalty is set to a sequence of that evenly spaced on a quad-root scale from $c^{\min} = 0.5$ to $c^{\max} = 100$. For all analyses excluding 23andMe, the length of sequences of both parameters are set to be 10, while for the analysis of 23andMe, it is set to be 5 to reduce the computation workload caused by the confidential requirements of the 23andMe dataset.

Obtain PROSPER solution. For M populations, the objective function to minimize for p_i-dimentional vector of SNP effect, $\boldsymbol{\beta}_i, i = 1, \dots, M$, is

$$\boldsymbol{L}(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m) = \sum_{1 \le i \le M} (\boldsymbol{\beta}_i^T (\boldsymbol{R}_i + \delta_i \boldsymbol{I}) \boldsymbol{\beta}_i - \mathbf{2}\boldsymbol{\beta}_i^T \boldsymbol{r}_i + 2\lambda_i \|\boldsymbol{\beta}_i\|_1^1)$$

$$+ \sum_{1 \le i_1 < i_2 \le M} c_{i_1 i_2} \left\| \boldsymbol{\beta}_{i_1}^{s_{i_1 i_2}} - \boldsymbol{\beta}_{i_2}^{s_{i_1 i_2}} \right\|_2^2$$

where $\boldsymbol{R}_i$ is an estimate of p_i-by-p_i LD matrix based on a reference sample from the i-th population, $\boldsymbol{r}_i$ is the p_i-dimentional vector of GWAS summary statistics in the i-th population, $\boldsymbol{\beta}_{i_1}^{s_{i_1 i_2}}$ and $\boldsymbol{\beta}_{i_2}^{s_{i_1 i_2}}$ denote the effect vectors for the SNPs shared across i_1-th and i_2-th populations (the set of SNPs is denoted by $s_{i_1 i_2}$); δ_i, λ_i and $c_{i_1 i_2}$ are tuning parameters as defined in above sections.

This optimization can be solved using coordinate descent algorithms by iteratively updating each element in the vectors. We take derivative for SNP k in i-th population, $k = 1, \dots, p_i$, $i = 1, \dots, M$

$$\frac{\partial \boldsymbol{L}(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_m)}{\partial \beta_{ik}}$$

$$= 2\left(1 + \delta_i + \sum_{i' \ne i, 1 \le i' \le M} c_{ii'}\right)\beta_{ik} + 2\lambda_i \frac{\partial |\beta_{ik}|}{\partial \beta_{ik}}$$

$$- 2\left(r_{ik} - \sum_{k' \ne k, 1 \le k' \le p} R_{i,k'k}\beta_{ik'} + \sum_{1 \le i' \le M, \text{s.t.} k \in S_{i,i'}} c_{ii'}\beta_{i'k}\right)$$

where β_{ik} denotes the SNP k in $\boldsymbol{\beta}_i$, r_{ik} denotes the SNP k SNP in $\boldsymbol{r}_i$, and $R_{i,k'k}$ denotes LD between the SNP k and the SNP k' in $\boldsymbol{R}_i$.

By solving $\frac{\partial \boldsymbol{L}(\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_m)}{\partial \beta_{ik}} = 0$ after the (t)-th iteration, we can get the updating rule for the $(t+1)$-th iteration

$$\beta_{ik}^{(t+1)} = \frac{\text{sign}(u_{ik}) \cdot \max\{0, |u_{ik}| - \lambda_i\}}{1 + \delta_i + \sum_{1 \le i' \le M, \text{s.t.} k \in S_{i,i'}} c_{ii'}}$$

where

$$u_{ik} = r_{ik} - \sum_{k' \ne k, 1 \le k' \le p} R_{i,k'k} \beta_{ik'}^{(t)} + \sum_{1 \le i' \le M, \text{s.t.} k \in S_{i,i'}} c_{ii'} \beta_{i'k}^{(t)}$$

Super learning. After getting PRSs for all populations under all tuning parameter settings, we further apply super learning to combine them to be trained on the tuning samples to get the final PROSPER model and tested on the validation samples. We use the function "*SuperLearner*" implemented in the R package with the same name, and include three linear prediction algorithms: lasso, ridge, and linear regression for continuous outcomes; and two prediction algorithms: lasso and linear regression for binary outcomes. We did not include ridge for binary outcomes due to the unavailability of ridge for binary outcomes in the function. For the included algorithms which have parameters: (1) in lasso, we use 100 values in lambda path calculated in the default setting in glmnet package; (2) in ridge, we use a lambda path of sequence from 1 to 20 incrementing by 0.1. We use Area under the ROC curve (AUC) as the objective function for binary outcomes and thus use the flag "method = method. AUC" in the function.

Existing PRS methods. We compare five groups of PRS methods. The first group is: single-ancestry method, which contains commonly known single-ancestry methods, including CT, LDpred2, and lassosum2, that are trained from the GWAS data from the target population. The second group is: EUR PRS based method, which is the three above single-ancestry methods trained from EUR GWAS data. The third group is: weighted PRS, which uses the weights estimated from a linear regression to combine the PRSs estimated from the corresponding single-ancestry method from all populations. The fourth group is: existing multi-ancestry methods, which includes two recently published and well-performed multi-ancestry methods, PRS-CSx and CT-SLEB. The last group is our proposed PROSPER. For all algorithms that have tuning parameters or weights, the optimal ones are determined based on predictive R^2 or AUC on tuning samples and finally evaluated on validation samples.

CT is implemented in our analysis by using r^2-cutoff of 0.1 in the clumping step and then thresholding by treating p-value-cutoff as a tuning parameter and being chosen from $5 \times 10^{-8}, 1 \times 10^{-7}, 5 \times 10^{-7}, 1 \times 10^{-6}, \ldots, 5 \times 10^{-1}, 1.0$.

LDpred2 is a PRS method that uses a spike-and-slab prior on GWAS summary statistics and modeling LD across SNPs. We implement LDpred2 by the function *"snp_ldpred2_grid"* in the R package "bigsnpr". The two tuning parameters in the algorithm include: the proportion of causal SNPs, which is chosen from a sequence of length 17 that are evenly spaced on a logarithmic scale from 10^{-4} to 1; per-SNP heritability, which is chosen from 0.7, 1, or 1.4 times the total heritability estimated by LD score regression divided by the number of causal SNPs. We fix the additional "sparse" option (for truncating small effects to zero) to FALSE.

lassosum2 is a PRS method that uses lasso regression on GWAS summary statistics for a single ancestry. We implement lassosum2 by the function *"snp_lassosum2"* in the R package "bigsnpr". The two tuning parameters in the algorithm include: tuning parameter for the lasso penalty, which is chosen from a sequence of length 20 that are evenly spaced on a logarithmic scale from $0.01 \times \max_{1 \le k \le p}(|r_k|)$ to $\max_{1 \le k \le p}(|r_k|)$; and regularization parameter for LD matrix, which is chosen from a sequence of length 10 that are evenly spaced on a cube-root scale from 0.01 to 100.

EUR PRS are the PRSs trained from EUR GWAS using the above single-ancestry methods, CT, LDpred2, and lassosum2, that are then applied to individuals of the target population. There is no need to perform tuning for them because the models have been tuned in EUR tuning samples. When computing scores for EUR PRS based method, we exclude SNPs that are not presented in the validation samples from the target population.

Weighted PRS linearly combines the corresponding single-ancestry method trained from all populations. The weights in the linear combination are estimated by a simple linear regression in the tuning samples from the target population.

PRS-CSx is a Bayesian multi-ancestry PRS method that jointly models GWAS summary statistics and LD structures across multiple populations using a continuous shrinkage prior. It has a further step to linearly combine the posterior effect-sizes estimates for EUR and the target population using weights in a simple linear regression in the tuning samples from the target population. We implement PRS-CSx using their python-based command line tool "PRS-CSx". The parameter phi was chosen from the default candidate values, $1, 10^{-2}, 10^{-4}$ and 10^{-6}. Due to the package restriction, the models are fitted with only HM3 SNPs.

676 ***CT-SLEB*** is a multi-ancestry PRS method that starts from clumping and thresholding, then uses
677 Empirical-Bayes (EB) method to estimate the coefficients of PRS, and finally combines PRS by a
678 super learning model. The three tuning parameters in the algorithm include: r^2-cutoff and base
679 size of the clumping window size used in the clumping step, which are chosen from (0.01, 0.05,
680 0.1, 0.2, 0.5) and (50kb, 100kb), respectively; and p-value cutoffs for EUR and the target
681 population, which are chosen from $5 \times 10^{-8}, 5 \times 10^{-7}, 5 \times 10^{-6}, \ldots, 5 \times 10^{-1}$ and 1.0.
682
683 **Computational time and memory usage**. The computational time and memory usage of
684 PROSPER and PRS-CSx are compared based on the analysis using simulated data on
685 chromosome 22. The analysis starts from inputting all required data into the algorithms, such as
686 summary statistics and LD reference data, and ends with outputting the final PRS coefficients
687 from the algorithms. PROSPER requires an input of optimal parameters in single-ancestry
688 analysis, so we also include the step of running the single-ancestry analysis, lassosum. The
689 analyses are performed using a single core with AMD EPYC 7702 64-Core Processors running at
690 2.0 GHz. The reported results are averaged over 10 replicates. The sample size for training
691 GWAS summary statistics is 15,000 for non-EUR populations and 100,000 for EUR population.
692 The sample size for the tuning dataset is 10,000 for each population.
693
694 **Real data analysis.** Training GWAS summary statistics are from 23andMe, GLGC, and AoU.
695 Tuning and validation individual-level data are from 23andMe and UKBB. Detailed descriptions
696 of those datasets are listed below.

23andMe Data. We analyzed two continuous traits, heart metabolic disease burden and height; and five binary traits, any CVD, depression, migraine diagnosis, morning person and SBMN, using GWAS summary statistics obtained from 23andMe Inc.. The information of individuals included in our analyses from the 23andMe participant cohort has consent and answered surveys online according to the human subject protocol reviewed and approved by Ethical & Independent Review Services, a private institutional review board (http://www.eandireview.com) as described in a previous paper [22]. Data on the seven traits are available for all five populations: AA, EAS, EUR, Latino, and SAS. The LD reference panels used for the five populations, respectively, are unrelated individuals from 1000G of AFR, EAS, EUR, AMR, and SAS origins. The tuning and validation are performed on a set of independent individuals of the corresponding ancestry from 23andMe participant cohort. Please see **Supplementary Table 3** for training sample sizes and **Supplementary Table 4** for tuning and validation sample sizes. The details of the data, including genotyping, quality control, imputation, removing related individuals, ancestry determination, and the preprocessing of GWAS, are also described in the previous paper [22]. For continuous traits, we evaluate PRS performance by the predictive R^2 of the PRS for residualized trait values obtained from regressing the traits on covariates. For binary traits, we evaluated PRS performance by the AUC by using the roc.binary function in the R package RISCA version 1.0 [55]. To compare the PRS performance for two different methods, we used the relative increase of logit-scale variance. The logit-scale variance of binary traits is converted from AUC by the formula $\sigma^2 = 2\phi^{-1}(AUC)$, where ϕ is the cumulative distribution function of the standard normal distribution.

GLGC Data. We analyzed four blood lipid traits, LDL, HDL, logTG and TC, using GWAS summary statistics computed without UKBB samples that are publicly available from GLGC (http://csg.sph.umich.edu/willer/public/glgc-lipids2021/). Detailed information about the design of the study, genotyping, quality control, and GWAS is described in Graham, S. E. *et al*. (2021) [37]. Data on the four traits are available for all five populations: admixed African or African, EAS, EUR, Hispanic, and SAS. The LD reference panels used for the five populations, respectively, are unrelated individuals from 1000G of AFR, EAS, EUR, AMR, and SAS origins. The tuning and validation are performed on UKBB individuals (as described below) from the same reference ancestry label as the LD reference panel. Please see **Supplementary Table 3** for sample sizes and the number of SNPs included in the analysis. The cleaning and preprocessing of the GWAS data are described in a previous paper [22].

AoU Data. We analyzed two anthropometric traits, BMI and height, using GWAS summary statistics trained from AoU. The information of individuals included in our analyses has been collected according to All of Us Research Program Operational Protocol (https://allofus.nih.gov/sites/default/files/aou_operational_protocol_v1.7_mar_2018.pdf). Details of the data and GWAS summary statistics are previously described[22]. Data for the two traits are available for three ancestries: AFR, Latino/Admixed American, and EUR. The LD reference panel used for the three populations, respectively, are 1000G unrelated individuals of AFR, AMR, and EUR origins. The tuning and validation are performed using UKBB individuals (as described below) from the same reference ancestry label as the LD reference panel. Please see **Supplementary Table 3** for sample sizes and the number of SNPs included in the analysis. The cleaning and preprocessing of the GWAS data are described in a previous paper [22].

UKBB data. We used UKBB data only for tuning and validation purposes. The four blood lipid traits and two anthropometric traits mentioned above have direct measurements in UKBB. The ancestry label of UKBB individuals is determined by genetically predicted ancestry, which are described in a previous paper [22]. Tuning and validation are based on R^2 of the PRS regressed on the residuals of the phenotypes adjusted by sex, age and PC1-10. Please see **Supplementary Table 4** for sample sizes. We note that for PRS we tested in UKBB validation samples, we use the ancestry labels in UKBB (AFR, AMR, EAS, EUR or SAS), instead of ancestry labels in the GWAS training data, to report the R^2 in the figures, result, and discussion sections of this paper.

References

54. Purcell, S. *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics* **81**, 559-575 (2007).

55. Chatton, A. *et al*. G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. *Scientific reports* **10**, 1-13 (2020).

Figure 1: Detailed flowchart of PROSPER. The analysis of M populations in PROSPER involves three key steps: 1. Separate single-ancestry analysis for all populations $i = 1, \dots, M$; 2. Joint analysis across populations using penalized regression; 3. Ensemble regression. In step 1, the training GWAS data is used to train lassosum2 models, and the tuning data is used to obtain the optimal tuning parameters in a single-ancestry analysis. In step 2, the training GWAS and the optimal tuning parameter values from step 1 are used to train the joint cross-population penalized regression model, and obtain solution $\boldsymbol{\beta}_{\lambda,c,i}$ for each λ and c. In step 3, the tuning data is used to train the super learning model for the ensemble of PRSs computed from the solutions in step 2, $\boldsymbol{PRS}_{\lambda,c,i} = \boldsymbol{X}\boldsymbol{\beta}_{\lambda,c,i}$. The final PRS is computed as $\boldsymbol{PRS} = \boldsymbol{X}\left(\sum w_{\lambda,c,i}\boldsymbol{\beta}_{\lambda,c,i}\right)$, where $w_{\lambda,c,i}$ are the weights from the super learning model. Refer to the "Method Overview" section in the main text for a full explanation of all notations in the flowchart.

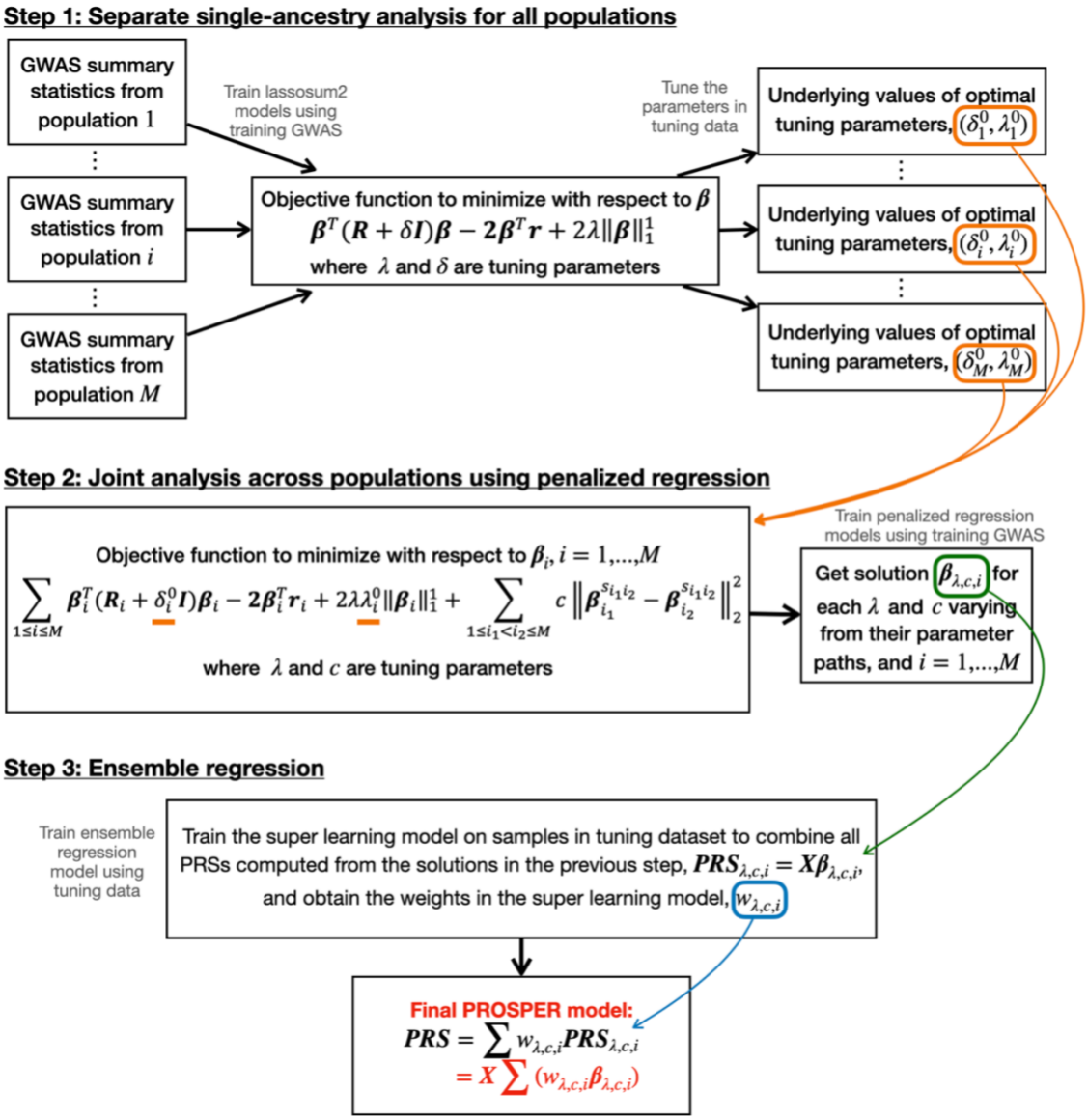
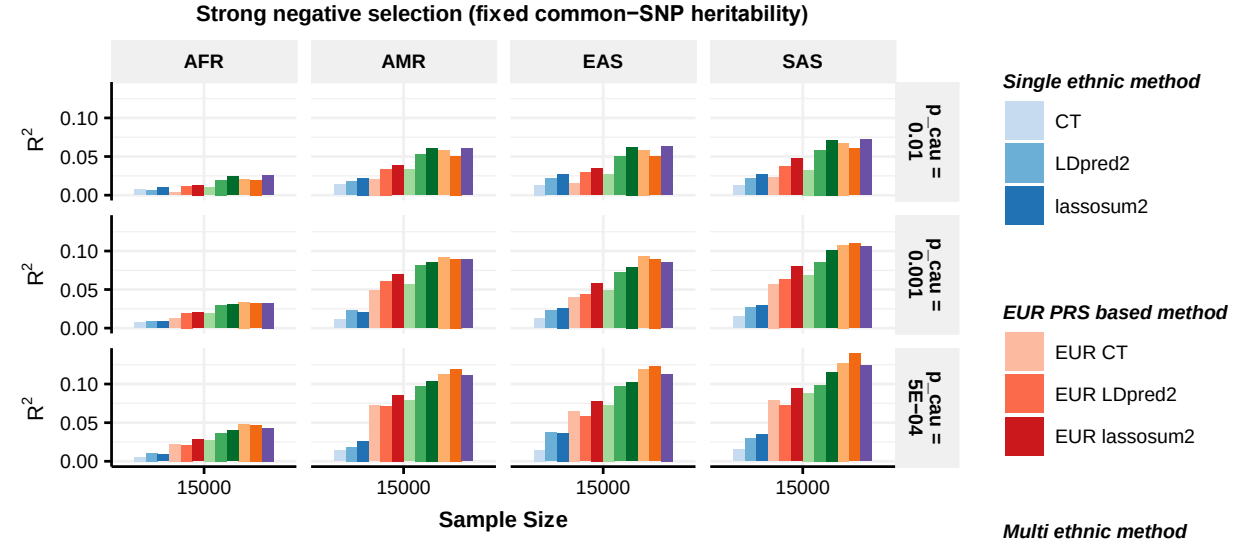


Figure 2: Performance comparison of alternative methods on simulated data generated with different sample sizes and genetic architectures under strong negative selection and fixed common-SNP heritability. Data are simulated for continuous phenotype under a strong negative selection model and three different degrees of polygenicity (top panel: $p_{causal} = 0.01$, middle panel: $p_{causal} = 0.001$, and bottom panel: $p_{causal} = 5 \times 10^{-4}$). Common SNP heritability is fixed at 0.4 across all populations, and the correlations in effect sizes for share SNPs between all pairs of populations is fixed at 0.8. The sample sizes for GWAS training data are assumed to be (**a**) 15,000, and (**b**) 80,000 for the four non-EUR target populations; and is fixed at 100,000 for the EUR population. PRS generated from all methods are tuned in 10,000 samples, and then tested in 10,000 independent samples in each target population. The PRS-CSx package is restricted to SNPs from HM3, whereas other alternative methods use SNPs from either HM3 or MEGA.

a



b

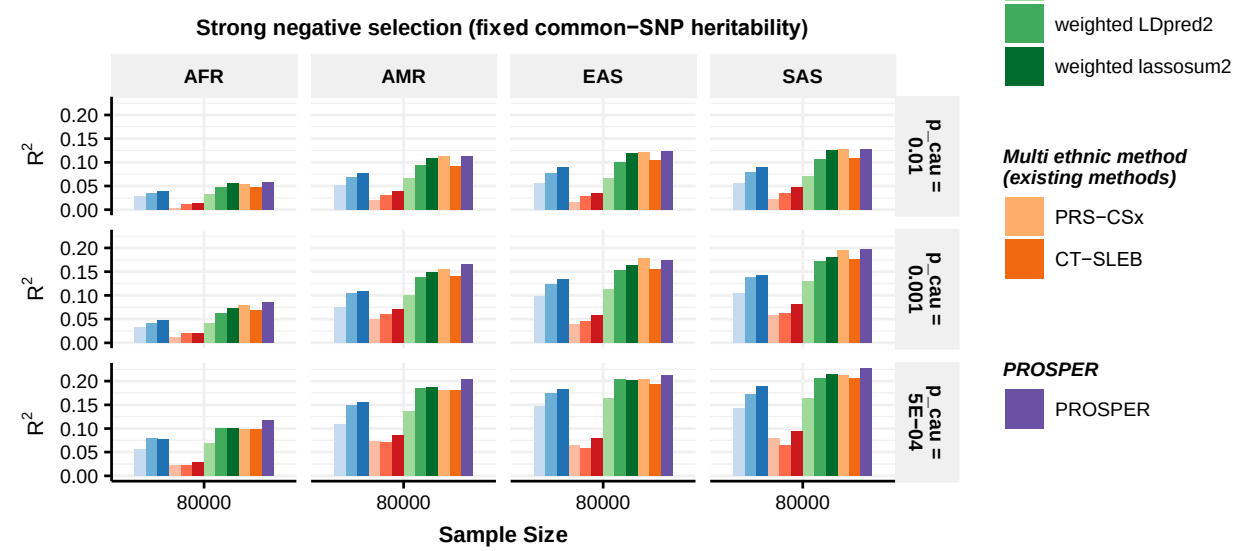


Figure 3: Performance comparison of alternative methods for prediction of two continuous traits in 23andMe. We analyzed two continuous traits, (**a**) heart metabolic disease burden and (**b**) height. PRS are trained using 23andMe data that available for five populations: African American, Latino, EAS, EUR, and SAS, and then tuned in an independent set of individuals from 23andMe of the corresponding ancestry. Performance is reported based on adjusted R^2 accounting for sex, age and PC1-5 in a held-out validation sample of individuals from 23andMe of the corresponding ancestry. The ratio of sample sizes for training, tuning and validation is roughly about 7:2:1, and detailed numbers are in **Supplementary Table 3-4**. The PRS-CSx package is restricted to SNPs from HM3, whereas other alternative methods use SNPs from either HM3 or MEGA.

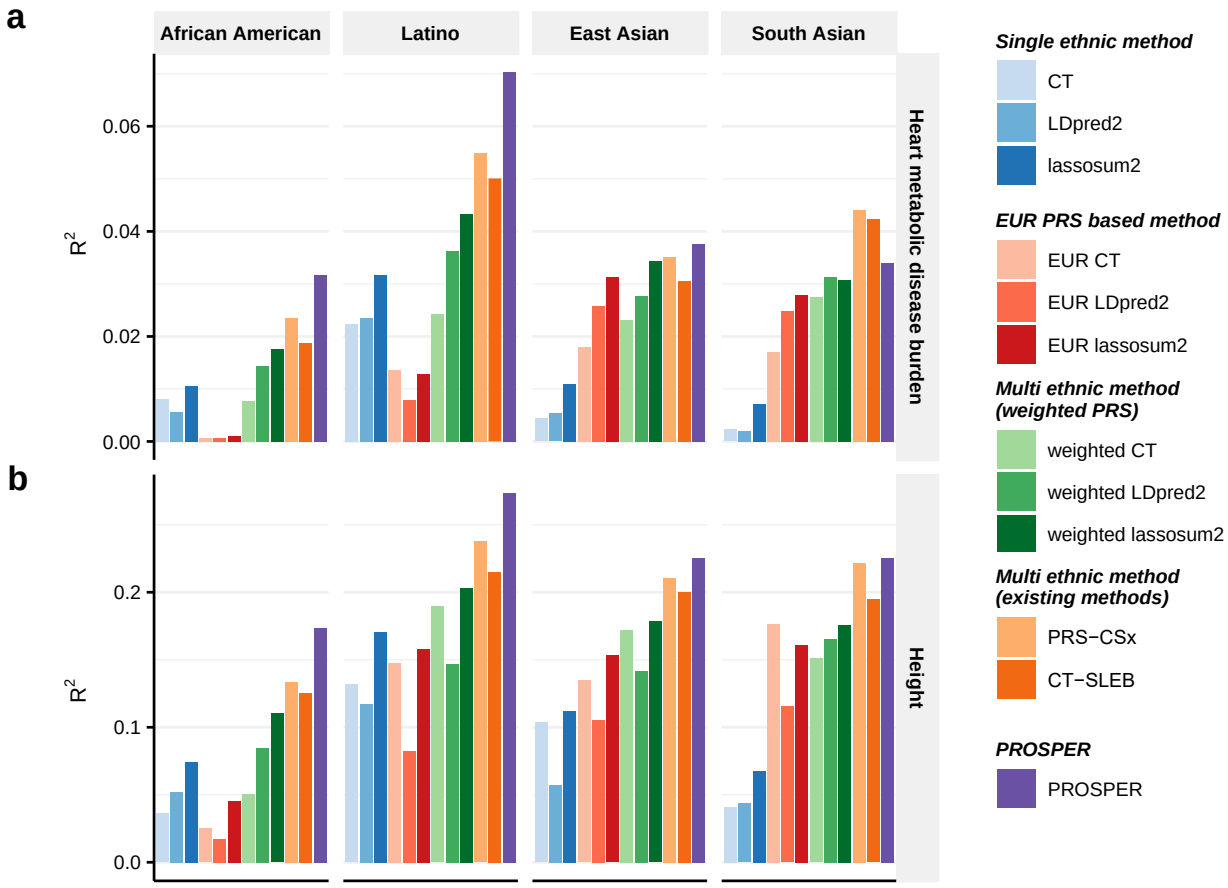


Figure 4: Performance comparison of alternative methods for prediction of five binary traits in 23andMe. We analyzed five binary traits, (**a**) any CVD, (**b**) depression, (**c**) migraine diagnosis, (**d**) morning person and (**e**) SBMN. PRS are trained using 23andMe data that available for five populations: African American, Latino, EAS, EUR, and SAS, and then tuned in an independent set of individuals from 23andMe of the corresponding ancestry. Performance is reported based on adjusted AUC accounting for sex, age, PC1-5 in a held-out validation sample of individuals from 23andMe of the corresponding ancestry. The ratio of sample sizes for training, tuning and validation is roughly about 7:2:1, and detailed numbers are in **Supplementary Table 3-4**. The PRS-CSx package is restricted to SNPs from HM3, whereas other alternative methods use SNPs from either HM3 or MEGA.

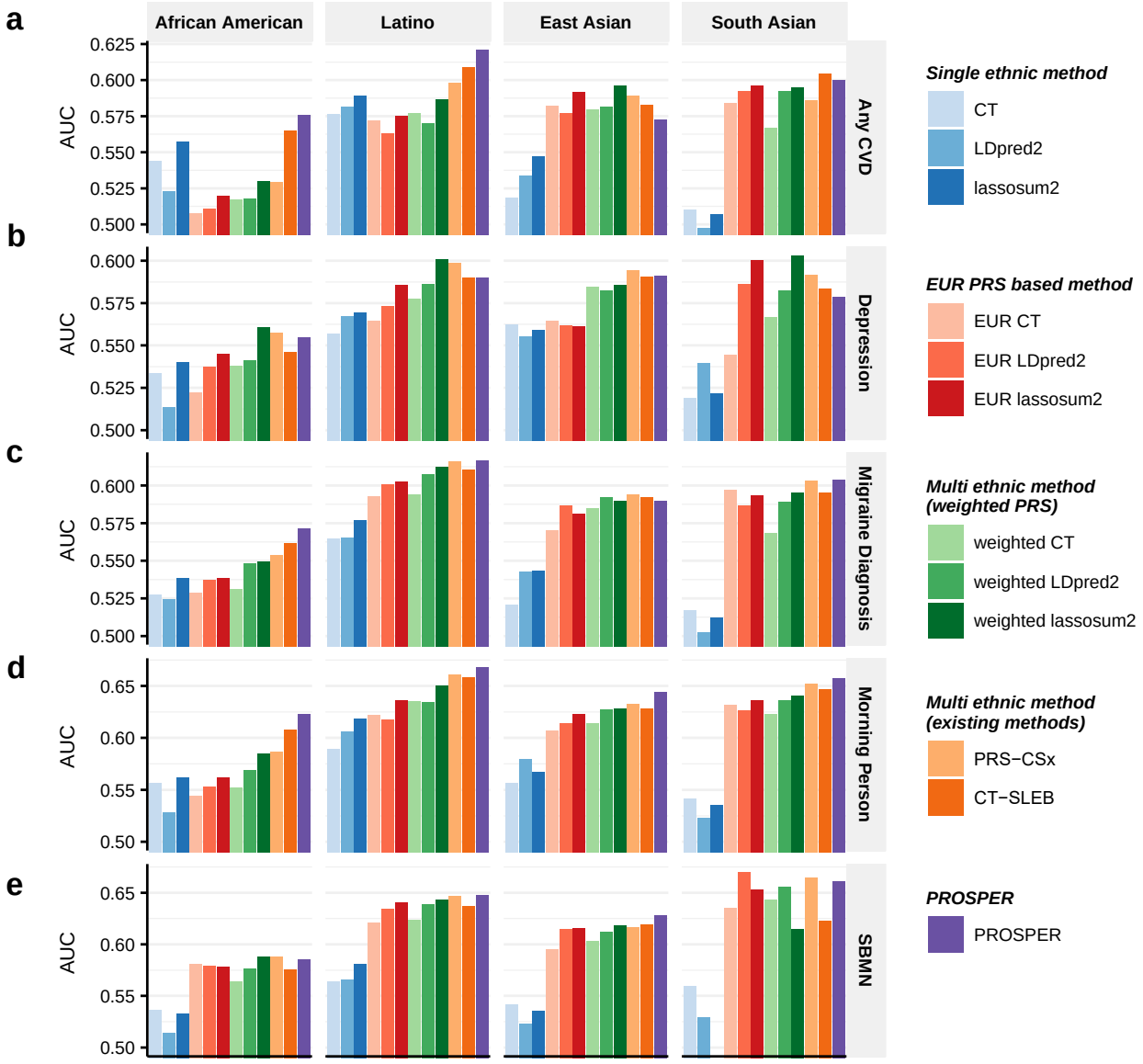


Figure 5: Performance comparison of alternative methods for prediction of four blood lipid traits (GLGC-training and UKBB-tuning/validation). We analyzed four blood lipid traits, (**a**) HDL, (**b**) LDL, (**c**) logTG and (**d**) TC. PRS are trained using GLGC data that available for five populations: admixed African or African, East Asian, European, Hispanic, and South, and then tuned in individuals from UKBB of the corresponding ancestry: AFR, EAS, EUR, AMR, and SAS (see the section of **Real data analysis** in **Methods** for ancestry composition). Performance is reported based on adjusted R^2 accounting for sex, age, PC1-10 in a held-out validation sample of individuals from UKBB of the corresponding ancestry. Sample sizes for training, tuning and validation data are in **Supplementary Table 3-4**. Results for AMR are not included due to the small sample size of genetically inferred AMR ancestry individuals in UKBB. The PRS-CSx package is restricted to SNPs from HM3, whereas other alternative methods use SNPs from either HM3 or MEGA.

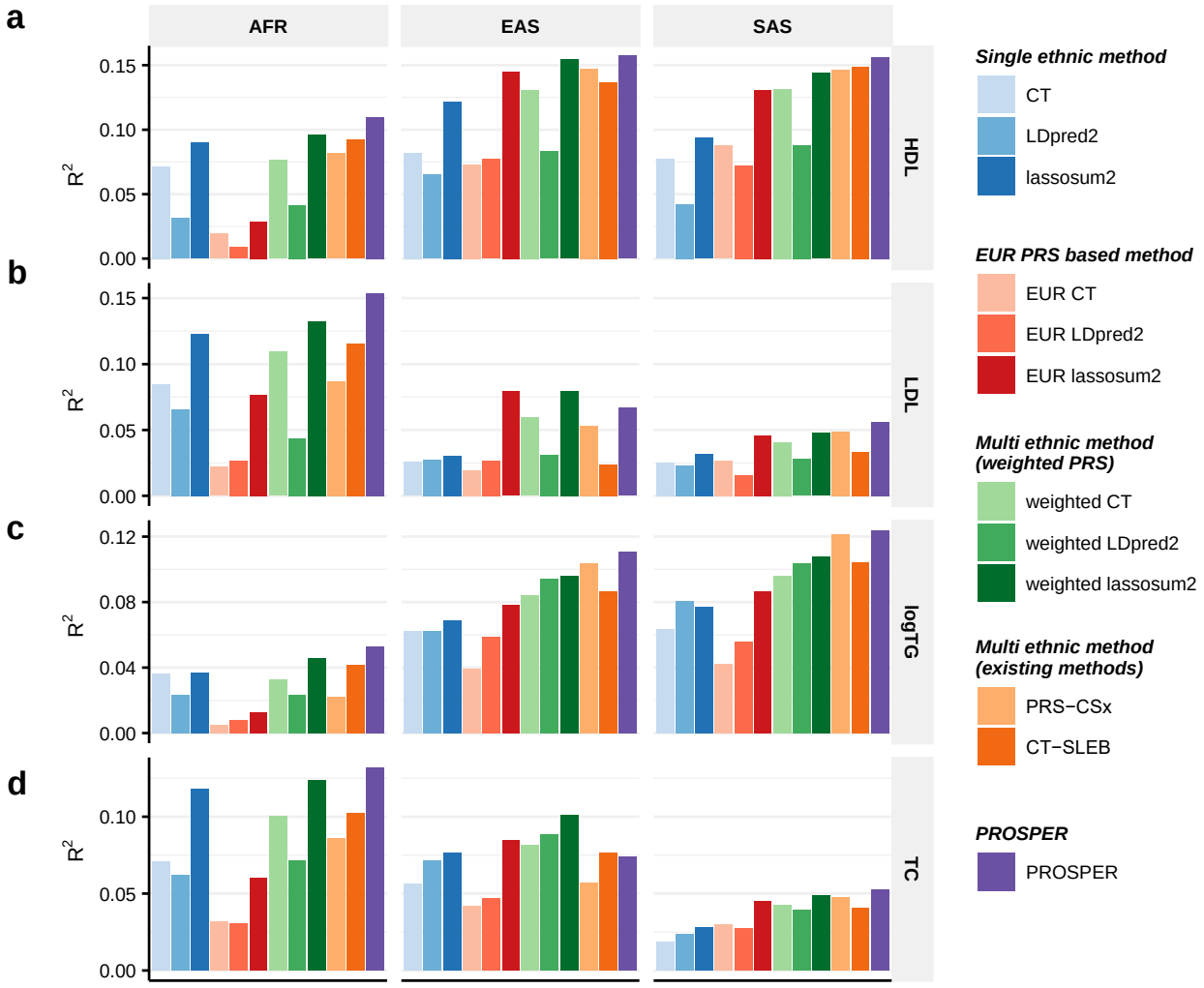
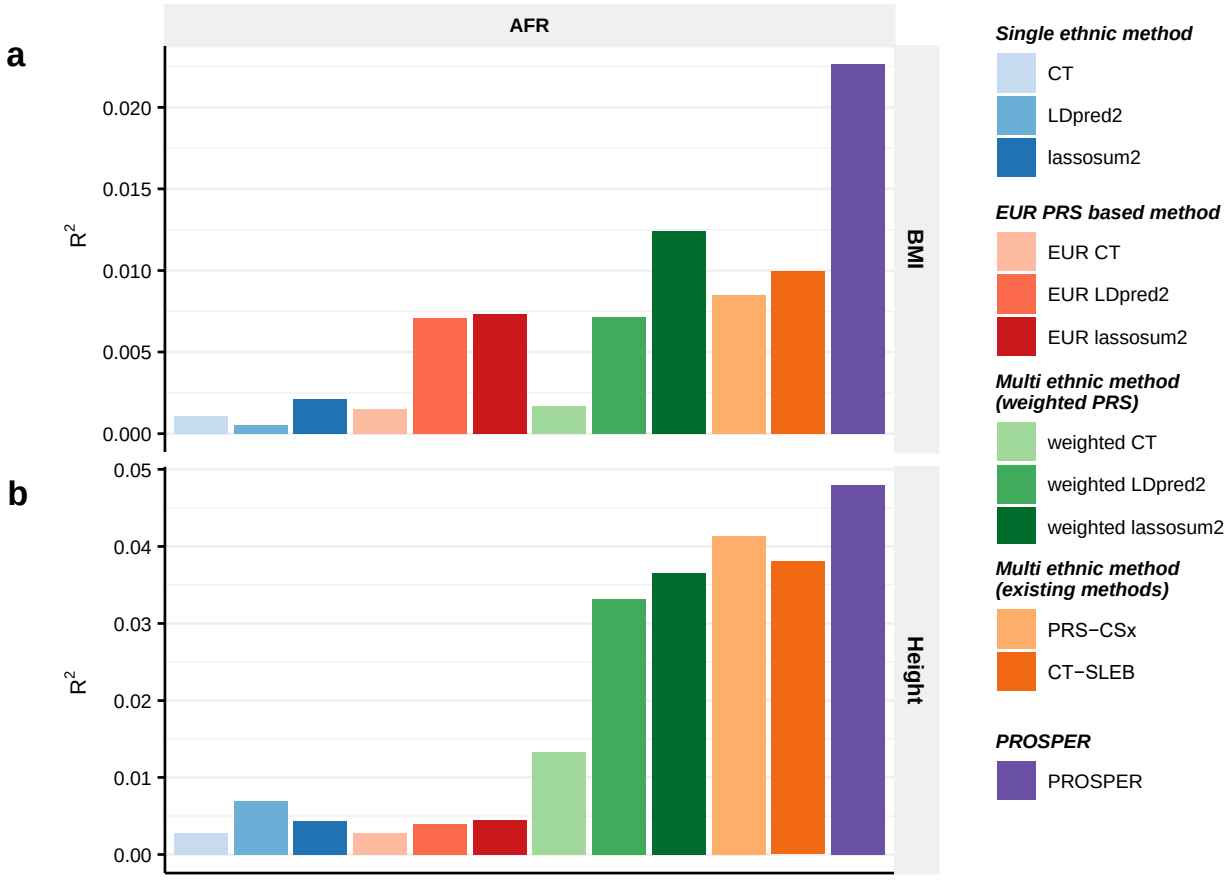
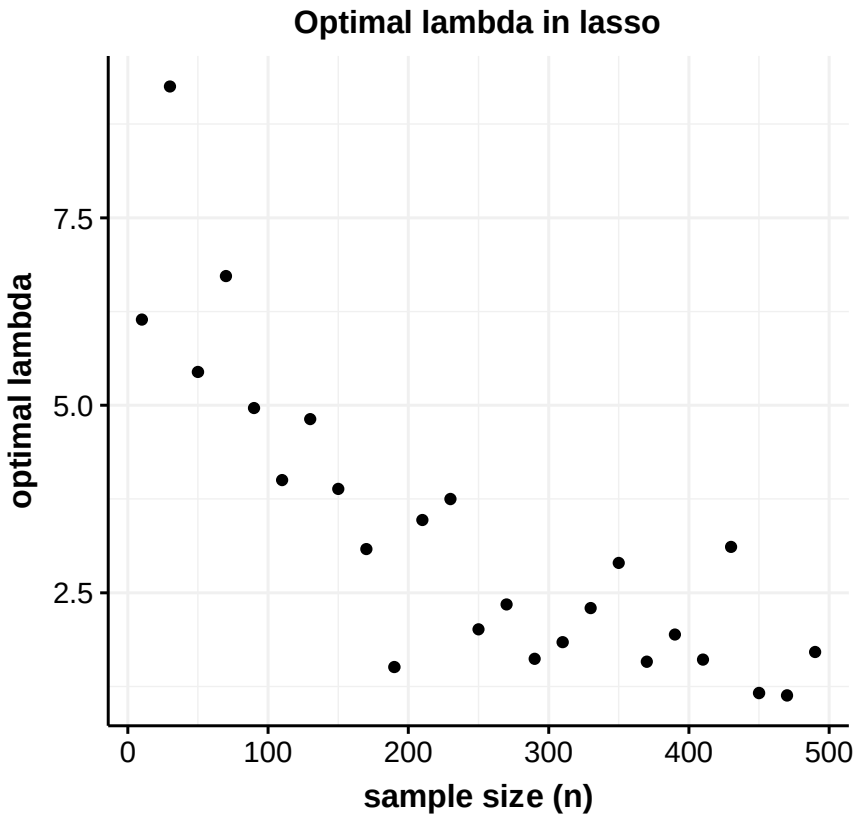


Figure 6: Performance comparison of alternative methods for prediction of two anthropometric traits (AoU-training and UKBB-tuning/validation). We analyzed two anthropometric traits, (**a**) BMI and (**b**) height. PRS are trained using AoU data that are available for three populations: African, Latino/Admixed American, and European and then tuned in individuals from UKBB of the corresponding ancestry: AFR, AMR, and EUR (see the section of **Real data analysis** in **Methods** for ancestry composition). Performance is reported based on adjusted R^2 accounting for sex, age, PC1-10 in a held-out validation sample of individuals from UKBB of the corresponding ancestry. Sample sizes for training, tuning and validation data are in **Supplementary Table 3-4**. Results for AMR are not included due to the small sample size of genetically inferred AMR ancestry individuals in UKBB. The number of SNPs analyzed in AoU analyses is much smaller than other analyses because the GWAS from AoU is on array data only (see **Supplementary Table 3** for the number of SNPs). The PRS-CSx package is restricted to SNPs from HM3, whereas other alternative methods use SNPs from either HM3 or MEGA.



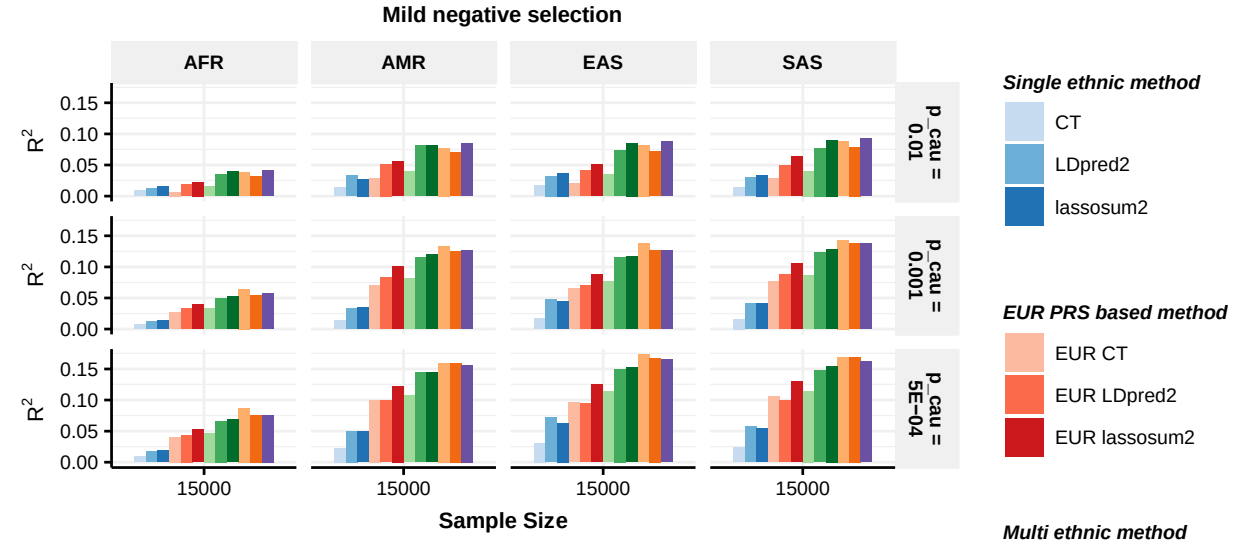
842 **Supplementary Figure 1: Optimal tuning parameter lambda in lasso.** The simulation is
843 performed for design matrix with 1000 predictors ($p = 1000$), and 5% of them are randomly
844 selected to be causal. Correlation structure of those predictors is AR1 with $\rho = 0.4$. The total
845 heritability is simulated to be 0.2.



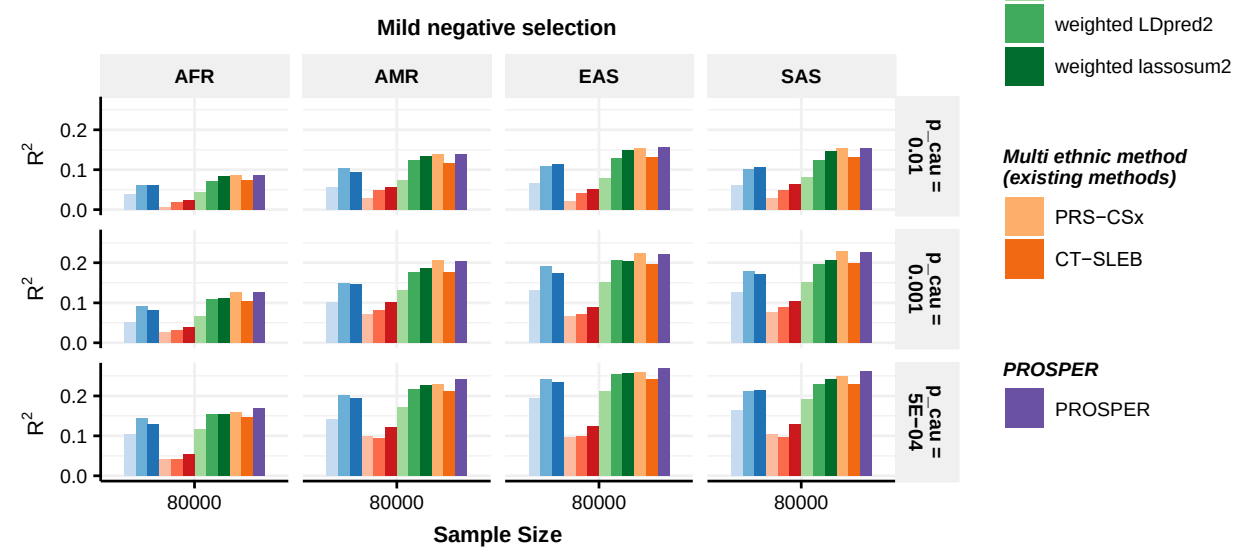
846
847

Supplementary Figure 2: Performance of alternative methods on simulated data generated with different sample sizes and different genetic architectures. Data are simulated for continuous phenotype under a mild negative selection model and three different degrees of polygenicity (top panel: $p_{causal} = 0.01$, middle panel: $p_{causal} = 0.001$, and bottom panel: $p_{causal} = 5 \times 10^{-4}$). Common SNP heritability is fixed at 0.4 across all populations, and the correlations in effect sizes for share SNPs between all pairs of populations is fixed at 0.8. The sample sizes for GWAS training data are assumed to be (**a**) 15,000, and (**b**) 80,000 for the four non-EUR target populations; and is fixed at 100,000 for the EUR population. PRS generated from all methods are tuned in 10,000 samples, and then tested in 10,000 independent samples in each target population. The PRS-CSx package is restricted to SNPs from HM3, whereas other alternative methods use SNPs from either HM3 or MEGA.

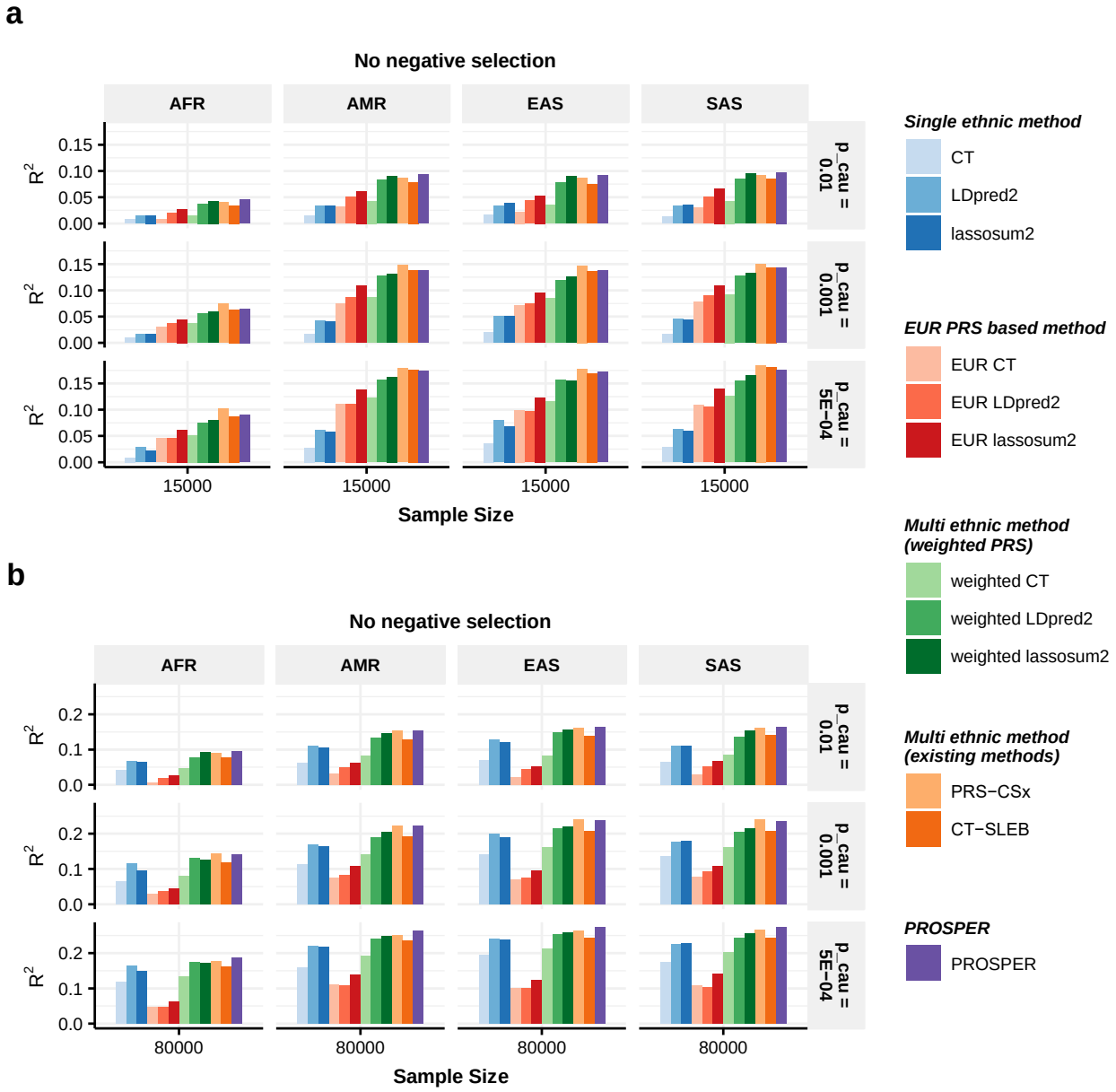
a



b

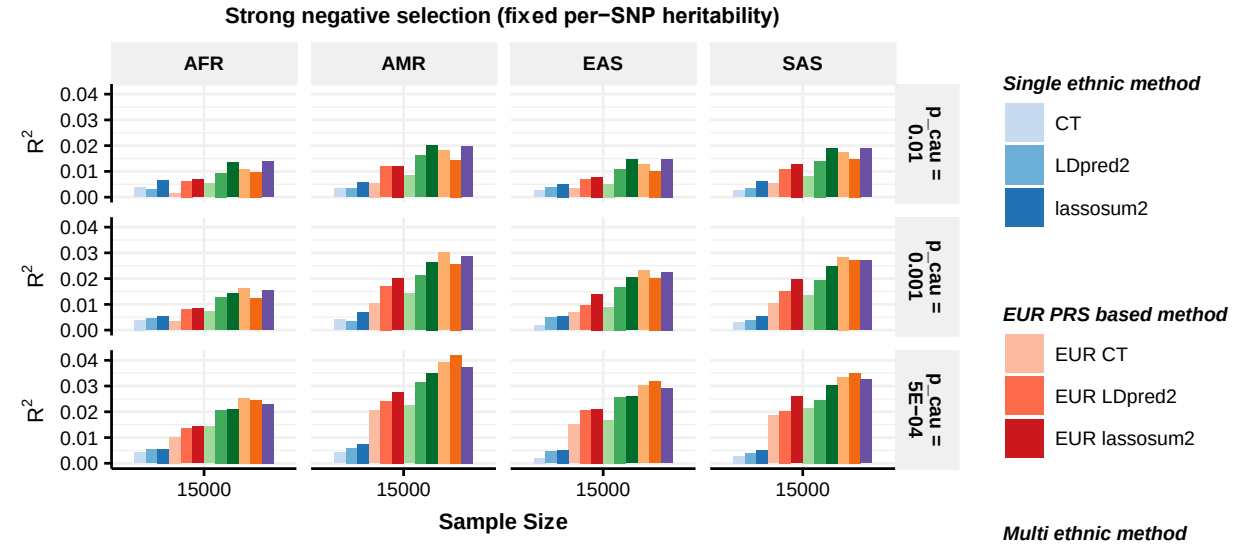


Supplementary Figure 3: Performance of alternative methods on simulated data generated with different sample sizes and different genetic architectures. Data are simulated for continuous phenotype under a no negative selection model and three different degrees of polygenicity (top panel: $p_{causal} = 0.01$, middle panel: $p_{causal} = 0.001$, and bottom panel: $p_{causal} = 5 \times 10^{-4}$). Common SNP heritability is fixed at 0.4 across all populations, and the correlations in effect sizes for share SNPs between all pairs of populations is fixed at 0.8. The sample sizes for GWAS training data are assumed to be (**a**) 15,000, and (**b**) 80,000 for the four non-EUR target populations; and is fixed at 100,000 for the EUR population. PRS generated from all methods are tuned in 10,000 samples, and then tested in 10,000 independent samples in each target population. The PRS-CSx package is restricted to SNPs from HM3, whereas other alternative methods use SNPs from either HM3 or MEGA.

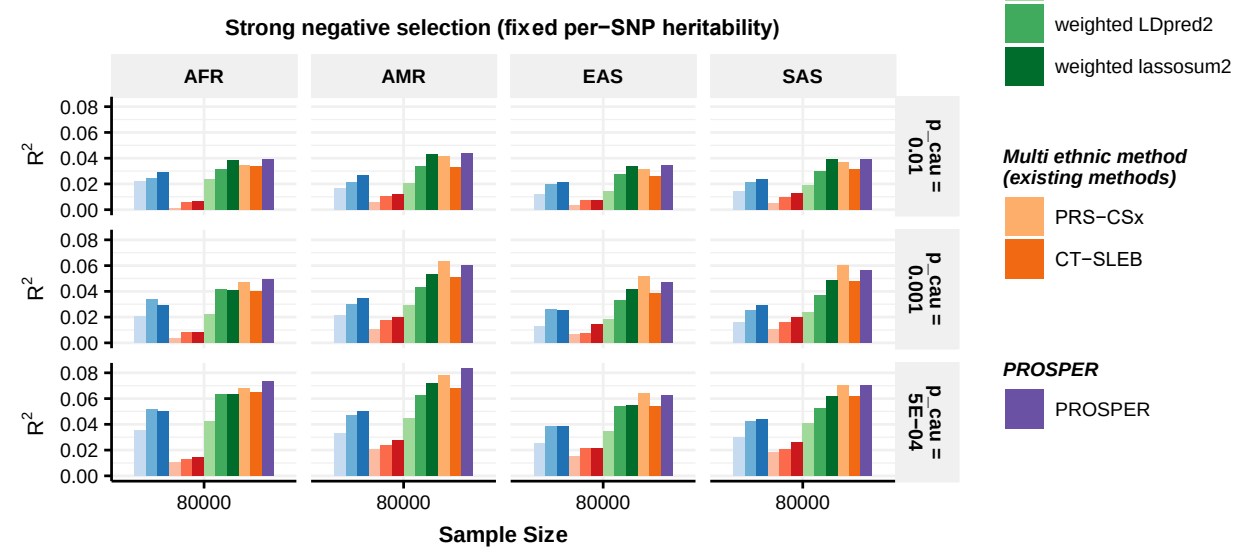


Supplementary Figure 4: Performance of alternative methods on simulated data generated with different sample sizes and different genetic architectures. Data are simulated for continuous phenotype under a strong negative selection model and three different degrees of polygenicity (top panel: $p_{causal} = 0.01$, middle panel: $p_{causal} = 0.001$, and bottom panel: $p_{causal} = 5 \times 10^{-4}$). Per-SNP heritability is assumed to be the same across all populations and thus leads to the common SNP heritability value of 0.32, 0.21, 0.16, 0.19 and 0.17 for AFR, AMR, EAS, EUR and SAS, respectively. The correlations in effect sizes for share SNPs between all pairs of populations is fixed at 0.8. The sample sizes for GWAS training data are assumed to be (**a**) 15,000, and (**b**) 80,000 for the four non-EUR target populations; and is fixed at 100,000 for the EUR population. PRS generated from all methods are tuned in 10,000 samples, and then tested in 10,000 independent samples in each target population. The PRS-CSx package is restricted to SNPs from HM3, whereas other alternative methods use SNPs from either HM3 or MEGA.

a



b



Supplementary Figure 5: Performance of alternative methods on simulated data generated with different sample sizes and different genetic architectures. Data are simulated for continuous phenotype under a strong negative selection model and three different degrees of polygenicity (top panel: $p_{causal} = 0.01$, middle panel: $p_{causal} = 0.001$, and bottom panel: $p_{causal} = 5 \times 10^{-4}$). Per-SNP heritability is assumed to be the same across all populations, and the correlations in effect sizes for share SNPs between all pairs of populations is fixed at 0.6. The sample sizes for GWAS training data are assumed to be (**a**) 15,000, and (**b**) 80,000 for the four non-EUR target populations; and is fixed at 100,000 for the EUR population. PRS generated from all methods are tuned in 10,000 samples, and then tested in 10,000 independent samples in each target population. The PRS-CSx package is restricted to SNPs from HM3, whereas other alternative methods use SNPs from either HM3 or MEGA.