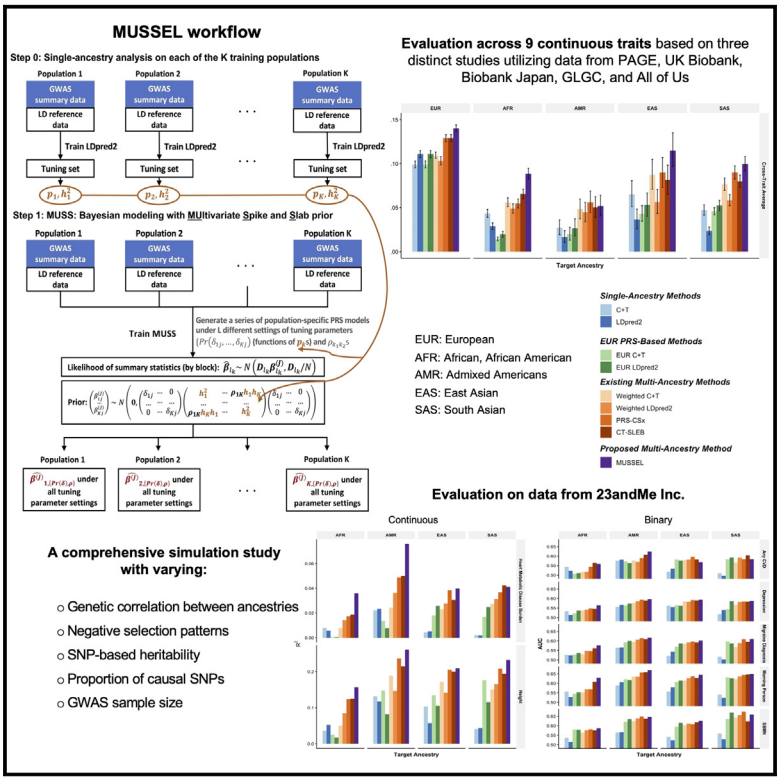
Cell Genomics

Technology

MUSSEL: Enhanced Bayesian polygenic risk prediction leveraging information across multiple ancestry groups

Graphical abstract



Highlights

- A new method for developing ancestry-specific polygenic risk prediction models
- Notable improvement in the prediction power demonstrated on non-European populations
- A command line tool, MUSSEL, was provided for method implementation
- Scope for additional improvement by ensemble of different methods

Authors

Jin Jin, Jianan Zhan, Jingning Zhang, ..., Genevieve Wojcik, Haoyu Zhang, Nilanjan Chatterjee

Correspondence

jin.jin@pennmedicine.upenn.edu (J.J.), nilanjan@jhu.edu (N.C.)

In brief

Jin et al. propose MUSSEL, a method for developing enhanced ancestry-specific polygenic risk scores via Bayesian hierarchical modeling and ensemble learning leveraging summary statistics from genome-wide association studies across multiple ancestry groups, which has the potential to reduce the performance gap in polygenic risk prediction across different ancestry populations.

Jin et al., 2024, Cell Genomics *4*, 100539
April 10, 2024
https://doi.org/10.1016/j.xgen.2024.100539

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Technology

MUSSEL: Enhanced Bayesian polygenic risk prediction leveraging information across multiple ancestry groups

Jin Jin,[1,2,15,*] Jianan Zhan,[3] Jingning Zhang,[1] Ruzhang Zhao,[1] Jared O'Connell,[3] Yunxuan Jiang,[3] 23andMe Research Team,[3] Steven Buyske,[4] Christopher Gignoux,[5] Christopher Haiman,[6] Eimear E. Kenny,[7] Charles Kooperberg,[8] Kari North,[9] Bertram L. Koelsch,[3] Genevieve Wojcik,[10,14] Haoyu Zhang,[11,12,14] and Nilanjan Chatterjee[1,13,14,*]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA
[2]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19103, USA
[3]23andMe, Inc., Sunnyvale, CA 94086, USA
[4]Department of Statistics, Rutgers University, New Brunswick, NJ 08854, USA
[5]Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA
[6]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90032, USA
[7]Icahn Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
[8]Division of Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, WA 98109, USA
[9]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27516, USA
[10]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA
[11]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
[12]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892, USA
[13]Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD 21205, USA
[14]These authors contributed equally
[15]Lead contact
*Correspondence: jin.jin@pennmedicine.upenn.edu (J.J.), nilanjan@jhu.edu (N.C.)
https://doi.org/10.1016/j.xgen.2024.100539

SUMMARY

Polygenic risk scores (PRSs) are now showing promising predictive performance on a wide variety of complex traits and diseases, but there exists a substantial performance gap across populations. We propose MUSSEL, a method for ancestry-specific polygenic prediction that borrows information in summary statistics from genome-wide association studies (GWASs) across multiple ancestry groups via Bayesian hierarchical modeling and ensemble learning. In our simulation studies and data analyses across four distinct studies, totaling 5.7 million participants with a substantial ancestral diversity, MUSSEL shows promising performance compared to alternatives. For example, MUSSEL has an average gain in prediction R^2 across 11 continuous traits of 40.2% and 49.3% compared to PRS-CSx and CT-SLEB, respectively, in the African ancestry population. The best-performing method, however, varies by GWAS sample size, target ancestry, trait architecture, and linkage disequilibrium reference samples; thus, ultimately a combination of methods may be needed to generate the most robust PRSs across diverse populations.

INTRODUCTION

Polygenic models for predicting complex traits are widely developed, utilizing summary-level association statistics from genome-wide association studies (GWASs). While being on course to translate GWAS results into clinical practice, polygenic risk scores (PRSs) encounter obstacles due to the poor predictive performance on under-represented non-European (non-EUR) ancestry populations, especially those with substantial African ancestry.[1–4] As sample sizes for GWASs in many non-EUR populations remain low for many traits, applications of PRSs often rely on EUR-based models, which underperform in other populations due in part to differences in allele frequencies, SNP effect sizes, and linkage disequilibrium (LD).[1–3,5,6]

To improve the poor performance of PRSs on non-EUR populations, several multi-ancestry methods have recently been developed to combine information from available GWAS summary statistics and LD reference data across multiple ancestry groups.[7] One simple approach is the weighted PRS,[8] which trains a linear combination of the PRS developed using single-ancestry methods (e.g., LD clumping and p value thresholding, C + T) applied separately to available GWAS data across different ancestry groups.[8] More recent methods attempt to borrow information across ancestry at the level of individual SNPs based on Bayesian methods[9,10] and penalized regressions,[11,12] or through the extension of C + T.[13] However, applications show that no single method performs uniformly the best, and their performance depends on many aspects, including

the underlying genetic architecture of the trait, the absolute and relative sample sizes across populations, and the algorithm for the estimation of LD based on the underlying reference dataset.[13]

We propose MUSSEL, a novel method for developing ancestry-specific PRS by jointly modeling ancestry-specific GWAS summary data across diverse ancestries. The method conducts Bayesian hierarchical modeling of SNP effect sizes across ancestries via a multivariate spike-and-slab prior and an ensemble learning step (MUSSEL) to seek an "optimal" combination of a series of PRSs obtained under different tuning parameter settings and across different ancestry groups. We evaluate MUSSEL and benchmark it against a variety of alternatives through large-scale simulation studies and analyses of 16 traits from four different studies: (1) the Population Architecture using Genomics and Epidemiology (PAGE) Study supplemented with data from the Biobank Japan (BBJ) and UK Biobank (UKBB); (2) Global Lipids Genetics Consortium (GLGC); (3) All of Us research program (AoU); and (4) 23andMe. These studies, with training data and additional validation samples from the UKBB study, included a total of 3.4 million European (EUR), 226,000 (226K) admixed African, African, or African American (AFR), 437K admixed Americans or Hispanic/Latino (AMR), 389K East Asian (EAS), and 56K South Asian (SAS). Results reveal the promising performance of MUSSEL for developing a robust PRS in the multi-ancestry setting and identifying a number of practical considerations for implementations that are crucial to the performance of the method.

DESIGN

MUSSEL overview

Considering that GWAS summary-level association statistics can be shared much more easily among research teams than individual-level genotype and phenotype data from GWASs, we will focus on PRS methods that can use summary statistics from the GWAS training samples. The implementation of our proposed method, MUSSEL, as well as other multi-ancestry methods to which we will compare MUSSEL, requires three (ancestry-specific) datasets from each training ancestry group: (1) GWAS summary data; (2) LD reference data; and (3) a validation (tuning + testing) dataset with genotype and phenotype data for an adequate number of individuals that are independent of GWAS samples and LD reference samples.

We now introduce MUSSEL, a novel method for enhanced ancestry-specific polygenic risk prediction based on available GWAS summary-level association statistics and LD reference data across multiple ancestry groups. MUSSEL consists of two steps (Figure 1): (1) a Bayesian modeling step (MUSS) to model the genetic correlation structure in SNP effect sizes across ancestry groups while accounting for ancestry-specific LD across SNPs; and (2) an ensemble learning (EL) step via a super learner (SL) to construct an "optimal" linear combination of a series of PRSs obtained from MUSS under different tuning-parameter settings and across all ancestry groups. Additionally, a step 0 was conducted before step 1 to obtain tuned causal SNP proportion and heritability parameters for each training ancestry group from LDpred2. These parameters will be used to specify the prior causal SNP proportions and heritability parameters in MUSS.

Step 1: MUSS: Bayesian modeling with multivariate spike-and-slab prior

MUSS tailors effect-size estimates for each ancestry group by incorporating data from other ancestry groups via Bayesian hierarchical modeling with a multivariate spike-and-slab prior on SNP effect sizes across ancestry groups. For population-specific SNPs, i.e., SNPs with minor allele frequency (MAF) > 0.01 in only one ancestry group, we assume a spike-and-slab prior as in LDpred2. For SNPs that are polymorphic across multiple populations, the between-SNP correlation is induced in two aspects: (1) we assume a SNP is causal in all those populations or none, and (2) the effect sizes for causal SNPs across populations are correlated (see STAR Methods for details). The prior specification is distinct compared to the recent method PRS-CSx[9] in two aspects: (1) the use of a multivariate spike-and-slab prior versus a continuous shrinkage prior to perform shrinkage estimation; and (2) flexible specification of genetic correlation structure across ancestry groups in MUSSEL compared to PRS-CSx, which assumes a single hyperparameter is shared across different ancestry groups and thus incorporates as fairly rigid specification of the correlation structure.

We infer posterior estimates of LD-adjusted SNP effect sizes across different ancestries via an efficient Markov chain Monte Carlo (MCMC) algorithm (STAR Methods). Multiple PRSs will be developed for each ancestry under carefully designed settings of two sets of tuning parameters: (1) the causal SNP proportion in each ancestry group, which will be used to specify the correlated prior causal probabilities across ancestry groups (STAR Methods); and (2) the between-ancestry genetic correlation in SNP effect sizes. Ancestry-specific SNP effect sizes are estimated based on MCMC with an approximation strategy previously implemented in the LDpred2 algorithm,[14] which substantially reduces the number of iterations required to reach convergence with a spike-and-slab type prior on a large number of correlated SNPs. The detailed MCMC algorithm and estimation procedure are described in STAR Methods.

Step 2: Ensemble learning via super learner

Research has shown that combining multiple C + T PRSs under different p-value thresholds[15] or combining the best ancestry-specific PRSs across multiple ancestry groups[8,9] can significantly improve predictive performances. Thus, as a second step of MUSSEL, we consider combining PRSs obtained from the MUSS step both across different tuning-parameter settings and across ancestry groups via an SL model trained on the tuning dataset. SL is an EL method for seeking an "optimal" linear combination of various base learners for prediction.[16] In our analyses, we consider three linear base learners, namely linear regression, elastic net regression,[17] and ridge regression.[18] A similar SL procedure was also implemented recently in another multi-ancestry method, CT-SLEB.[13] In our simulation studies and real data examples, we will show explicitly how much improvement in predictive power can be obtained separately through the Bayesian modeling step and the EL step. Considering that both weighted PRS and PRS-CSx construct a linear

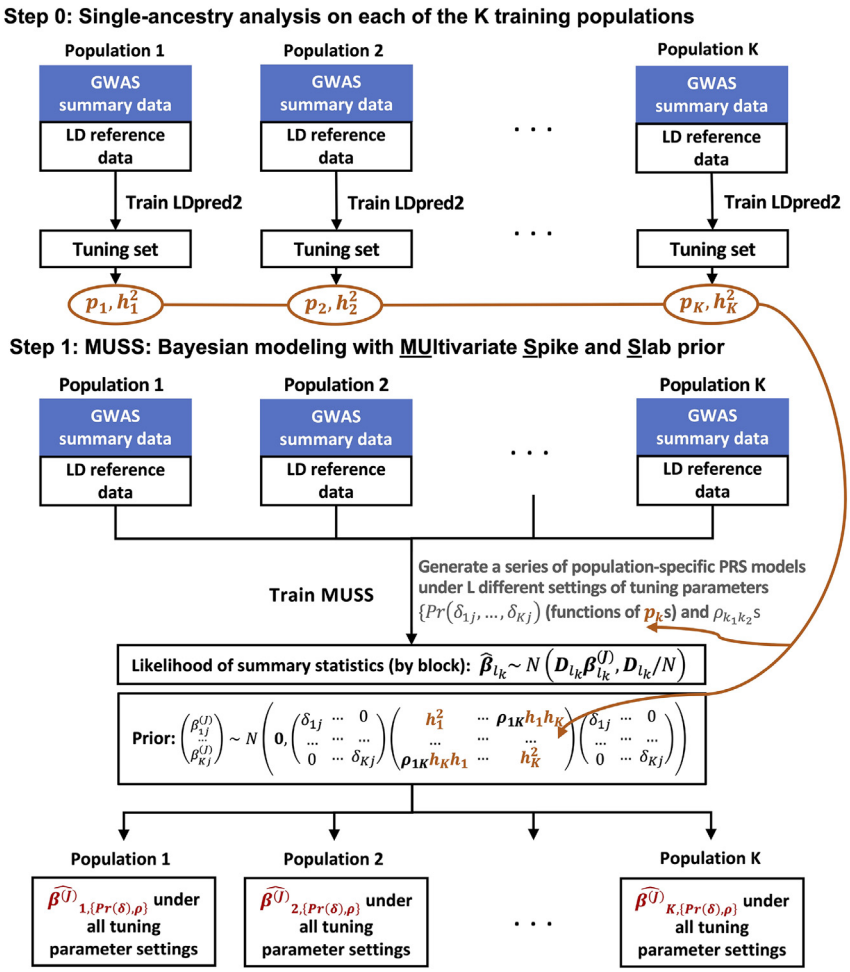


Figure 1. MUSSEL workflow

(Step 0) Apply LDpred2 to each of the K training populations (ancestry groups) to obtain estimated causal SNP proportions ($p_k, k = 1, \ldots, K$) and heritability ($h^2_k, k = 1, \ldots, K$) parameters based on the tuning set; these parameters will be used to specify the prior distributions and tuning-parameter settings for Bayesian learning with MUSS. (Step 1) MUSS: jointly model all training populations to obtain a total of (L × K) PRS models under L different tuning-parameter settings for Pr $(\delta_{1j}, \ldots, \delta_{Kj})$ (functions of p_k s) and $\rho_{k_1k_2}$ s across K training populations. (Step 2) for each target population, conduct ensemble learning (EL) via a super learner (SL) algorithm with a set of base learners (e.g., elastic net regression, ridge regression, and linear regression) to train an "optimal" linear combination of the (L × K) PRS models from the MUSS step to obtain the final MUSSEL model. The prediction performance of the final PRS derived using MUSSEL should be evaluated on an independent testing set.

combination of the best PRS for each ancestry group, we tried the same approach on our Bayesian model (MUSS) and called this alternative method "weighted MUSS." We observe on both simulated data and real data that the gain in predictive power by this linear combination strategy is mostly lower than, and sometimes comparable to, the gain by our proposed EL strategy (Figures S1–S13, "weighted MUSS" versus "MUSSEL").

RESULTS

Simulation settings

We first investigate the performance of MUSSEL and a series of existing methods under various simulated scenarios of the genetic architecture of a continuous trait and absolute and relative GWAS sample sizes across ancestry groups. This large-scale dataset, including simulated genotype and phenotype data for a total of 600,000 individuals across EUR, AFR, AMR, EAS, and SAS, was recently released by our group.[13] Detailed simulation setup is described in Zhang et al.[13] and briefly summarized in the supplemental information.

We apply eight existing approaches for comparison, which include two single-ancestry methods applied to GWAS and LD reference data from the target population: (1) C + T and (2) LDpred2; the same single-ancestry methods applied to GWAS and LD reference data for EUR: (3) EUR C + T and (4) EUR LDpred2; and three existing multi-ancestry methods applied to ancestry-specific GWAS and LD reference data for all ancestry groups: (5) weighted C + T (weighted PRS using C + T as the base method), (6) weighted LDpred2 (weighted PRS using LDpred2 as the base method), (7) PRS-CSx,[9] and (8) CT-SLEB.[13] Results from another two recently proposed multi-ancestry methods, PolyPred+[19] and XPASS,[10] on the same simulated dataset are reported in Zhang et al.[13] Table 1 provides a comparison of the various methods in terms of data requirement, similarities, and differences. Taking into account both ancestral diversity and computational efficiency, throughout the text we restrict all our analyses to the SNPs among approximately 2.0 million SNPs in HapMap 3[20] plus Multi-Ethnic Genotyping Array (MEGA)[21] that are also available in the discovery GWAS, LD reference panel, and validation (tuning + testing) samples. We assess the predictive performance of a PRS by prediction R^2, i.e., the proportion of variance of the trait explained by the PRS. The corresponding 95% bootstrap confidence intervals (CIs) are calculated based on 10,000 bootstrap samples using the Bca approach[22] implemented in the R package "boot"[23] (Figures S1–S10; Tables S1, S2, S3, S4, and S5). Results from the various methods are compared in five simulation settings: (1) fixed common SNP heritability, strong negative selection, with a genetic correlation set to $\rho = 0.8$ between any two ancestry groups (Figures 2, S1, and S2); (2) fixed per-SNP heritability, strong negative selection, $\rho = 0.8$ (Figures S3 and S4); (3) fixed per-SNP heritability, strong negative selection, with a weaker between-ancestry genetic correlation, $\rho = 0.6$ (Figures S5 and S6); (4) fixed common SNP heritability, no negative selection, $\rho = 0.8$ (Figures S7 and S8); and (5) fixed common

Table 1. Overview of the methods implemented for PRS development

Method	Required training data source[a]	Features	Tuning parameters
Single-ancestry			
C + T	target ancestry	model-free	p_t (p value threshold)
LDpred2	target ancestry	Bayesian (spike-and-slab prior)	p (causal SNP proportion), h^2 (heritability)
EUR C + T	EUR	model-free	p_t (p value threshold)
EUR LDpred2	EUR	Bayesian (spike-and-slab prior)	p (causal SNP proportion), h^2 (heritability)
Multi-ancestry			
weighted LDpred2	ancestry-specific data from each available ancestry	Bayesian (spike-and-slab prior), linear combination strategy	p, H^2, weight of each ancestry-specific PRS in the final model
PRS-CSx	ancestry-specific data from each available ancestry	Bayesian (Strawderman-Berger prior), linear combination strategy	φ (global shrinkage parameter), weight of each ancestry-specific PRS in the final model
XPASS[b]	ancestry-specific data from each available ancestry	Bayesian (bivariate normal prior), infinitesimal model	–
PolyPred+[b]	ancestry-specific data from each available ancestry	Bayesian, functional annotation, linear combination of SBayesR and PolyFun	parameters in SBayesR and PolyFun, weight of SBayesR PRS and PolyFun PRS in the final model
CT-SLEB	ancestry-specific data from each available ancestry	empirical Bayes, EL via SL	p_t (p value threshold), d (genetic distance) for C + T step, parameters in the SL
MUSSEL	ancestry-specific data from each available ancestry	Bayesian (multivariate spike-and-slab prior), EL via SL	$\Pr(\delta_1,\ldots,\delta_K), \rho_{k_1k_2}, 1 \le k_1 < k_2 \le K$, parameters in the SL

[a]All methods require three datasets to train the PRS model: (1) discovery GWAS summary data, (2) LD reference data, and (3) tuning data.
[b]Results from PolyPred+ and XPASS on all simulated and real datasets (except for PAGE + UKBB + BBJ) were reported in Zhang et al.[13]

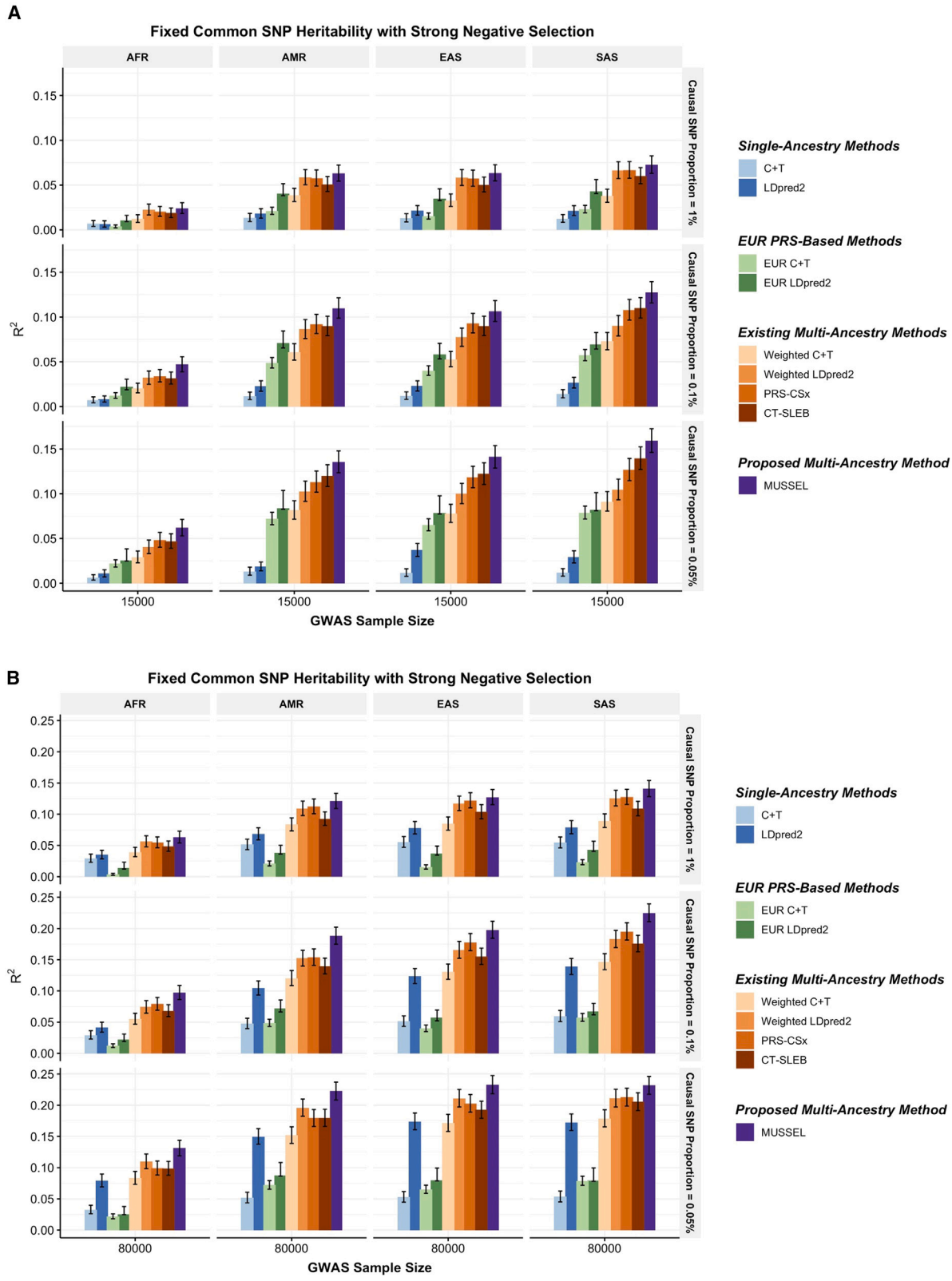
SNP heritability, mild negative selection, $\rho = 0.8$ (Figures S9 and S10).

Simulation results

The multi-ancestry methods tend to outperform the single-ancestry methods, except for weighted C + T, which performs worse than LDpred2 when GWAS sample size of the non-EUR target population becomes adequately large (Figures 2 and S1–S10). When the discovery GWAS sample size of the target non-EUR population is relatively small (N = 15,000) compared to EUR GWAS (N = 100,000), EUR PRS tends to outperform the generated PRS based on training data from the target non-EUR population; but as the GWAS sample size of the target non-EUR population increases, the prediction R^2 of LDpred2 eventually becomes substantially higher than that of EUR C + T and EUR LDpred2. Among the existing multi-ancestry methods, weighted LDpred2, PRS-CSx, and CT-SLEB perform similarly but show advantages over others in different settings: weighted LDpred2 performs well in the scenario of a large causal SNP proportion, while CT-SLEB performs similarly to PRS-CSx but shows some advantages when there is a small causal SNP proportion (0.05%) and when GWAS sample size for target non-EUR population is small. Overall, the proposed method MUSSEL outperforms these existing methods in almost all settings. This is expected, given that the SNP effect sizes were simulated under a multivariate spike-and-slab distribution as assumed in the MUSS model. The proposed EL step (in MUSSEL) and the alternative linear combination step (in weighted MUSS) only provide minimal improvement in R^2 on top of MUSS (Figures S1–S10). This may be because when the specified distribution of SNP effect sizes approximates the true distribution well, the best PRS trained for each ancestry by MUSS can already provide a high predictive power, and an additional step of combining PRSs across tuning-parameter settings and ancestry groups is unnecessary.

We also checked the computation intensity of MUSSEL in comparison to PRS-CSx (Table S6). A comparison of computation time between PRS-CSx and CT-SLEB on the same simulation dataset was reported in Zhang et al.[13] With AMD EPYC 7702 64-core processors running at 2.0 GHz using a single core, on chromosome 22 and with a total of $5 \times (K + 1)$ tuning-parameter settings, MUSSEL has an average runtime of approximately 75.9 min combining $K = 3$ ancestry groups with a total of 17,192 SNPs, 127.2 min combining $K = 4$ ancestry groups with 17,721 SNPs, and 237.4 min across $K = 5$ ancestry groups with 17,722 SNPs. Although not as fast as simpler methods such as CT-SLEB and XPASS, MUSSEL is computationally more efficient than PRS-CSx ($K = 3$: 3.8-fold; $K = 4$: 3.2-fold; $K = 5$: 2.5-fold) and thus is easier to implement than PRS-CSx, especially when four or more training populations are available to be combined.

To examine whether the performance of MUSSEL is sensitive to mis-specification of the LD matrix, we conduct a sensitivity analysis, whereby we estimate LD for each ancestry group based on a slightly mis-specified LD reference sample that contains 800 individuals from the same ancestry group and 50 individuals from each of the other four ancestry groups, totaling 200 (20%) individuals with ancestry mismatch. We repeat our analysis under the setting of having fixed common SNP heritability, a strong negative selection, and a genetic correlation of 0.8 across all pairs of ancestry groups, based on the mis-specified LD reference samples. We also apply LDpred2, EUR LDpred2, and weighted LDpred2, which may also be sensitive to ancestry mismatch between the discovery GWAS samples and LD



(legend on next page)

reference samples. Compared to the results assuming no ancestry mismatch between the discovery GWAS and LD reference data, the R^2 of LDpred2, EUR LDpred2, weighted LDpred2, and MUSSEL PRS are on average 3.3%, 5.7%, 15.1%, and 10.4% lower, respectively (Figures S14 and S15; Table S7). The amount of power loss appears to increase as the underlying causal SNP proportion decreases.

PAGE + UKBB + BBJ data analysis with validation on non-EUR individuals from PAGE

We evaluate the performance of the various methods on predicting the polygenic risk of inverse-rank normal transformed body mass index (IRNT BMI), high-density lipoprotein (HDL), and low-density lipoprotein (LDL) separately for AFR, AMR, and EAS. We collected ancestry-specific training GWAS summary data for AFR and AMR from PAGE, GWAS summary data for EAS from BBJ, and EUR GWAS summary data from UKBB. The PRSs developed by the various methods are evaluated on validation individuals of AFR, AMR, and EAS populations from PAGE. We use genotype data for 498 EUR, 659 AFR, 347 AMR, 503 EAS, and 487 SAS individuals from the 1000 Genomes Project as the LD reference data.[24]

In this set of analyses, we observe that the multi-ancestry methods tend to outperform single-ancestry methods for EUR, AFR, and AMR (Figures 3 and S11; Tables S8 and S9). For EAS, LDpred2 can reach an R^2 similar to or higher than that of EUR LDpred2 and multi-ancestry methods, which is possibly because the BBJ GWAS sample sizes for EAS are relatively large (N = 70,657–158,284). For the proposed method MUSSEL, we observe potential improvement in R^2 from both the Bayesian modeling step (MUSS versus LDpred2) and the SL step (MUSSEL versus MUSS). The linear combination strategy (weighted MUSS, Figure S11) provides a smaller or similar gain in R^2 compared to our SL strategy (MUSSEL). The relative performance of the various multi-ancestry methods varies by trait and ancestry, and no method is uniformly better than others. In some settings, MUSSEL PRS gives a lower R^2 than the PRS trained by weighted LDpred2 and PRS-CSx in some settings, such as for BMI on AFR and LDL on EAS. In general, however, the MUSSEL PRS has the best overall performance, with an average increase of 3.6% and 19.6% in R^2 compared to PRS-CSx and CT-SLEB, respectively, on non-EUR ancestries.

GLGC data analysis with validation on UKBB individuals

We apply the various methods to develop ancestry-specific PRS for four blood lipid traits, namely HDL, LDL, total cholesterol (TC), and log of triglycerides (logTG),[25] based on ancestry-specific GWAS summary data for EUR, AFR, AMR, EAS, and SAS, from the GLGC. We validate the performance of the various methods on UKBB individuals of AFR, EAS, and SAS origin separately, where the ancestry information of the UKBB validation individuals was determined based on an ancestry genetic component analysis (supplemental information).

We first use genotype data of the unrelated 1000 Genomes samples as the LD reference data.[24] We observe that the MUSSEL PRS performs the best or similarly to the best PRS (Figures 4A and S12A; Tables S10 and S11). We see a notable gain in R^2 on comparing MUSSEL PRS to weighted LDpred2 PRS (average increase: 50.7%). MUSSEL outperforms CT-SLEB in most cases (average increase in R^2: 27.1%). Although the relative performance between MUSSEL and PRS-CSx varies by ancestry and trait, MUSSEL PRS has a better overall performance, with an average increase of 19.9% in R^2 compared to PRS-CSx PRS. Similar to the results from PAGE + UKBB + BBJ analysis, MUSSEL improves on top of LDpred2 by both the Bayesian modeling step (MUSS versus LDpred2, Figure S12A) and the SL step (MUSSEL versus MUSS, Figure S12A). The PRS generated by the alternative linear combination strategy has a similar or lower R^2 than the PRS generated by our proposed EL strategy (weighted MUSS versus MUSSEL, Figure S12A).

It has been observed that LDpred2 sometimes has suboptimal performance based on the widely implemented 1000 Genomes LD reference data,[26,27] which may be due to convergence issue in the presence of inadequate LD reference sample size and/or ancestry mismatch between 1000 Genomes samples and the target population.[26] Implemented by an MCMC algorithm that utilizes computational tricks similar to those of LDpred2, MUSSEL may likewise underperform with the 1000 Genomes reference data. We therefore conduct a sensitivity analysis whereby we estimate LD based on UKBB tuning samples (10,000 EUR, 4,585 AFR, 687 AMR, 1,010 EAS, and 5,427 SAS) instead of the 1000 Genomes samples. We observe that the R^2 of MUSSEL PRS improves notably compared to using 1000 Genomes LD reference (Figure 4B; Tables S10 and S11), especially on AFR (average increase: 33.8%). The R^2 of PRS-CSx PRS has also increased but not as much as the R^2 of MUSSEL PRS. This is particularly noteworthy because PRS-CSx by default uses a much larger number of UKBB LD reference samples (375,120 EUR, 7,507 AFR, 687 AMR, 2,181 EAS, and 8,412 SAS), which also overlap with our UKBB testing samples and thus lead to potentially inflated R^2 estimates. The advantage of MUSSEL now becomes more obvious: it outperforms the existing methods in all scenarios except for HDL in EAS, where it performs slightly worse than PRS-CSx PRS. MUSSEL shows the most notable advantage on AFR, for which PRSs are typically not powerful and difficult to improve (average R^2 increase compared to the best existing method: 38.6%). Interestingly,

Figure 2. Simulation results showing performance of the PRS trained by MUSSEL and various existing methods

A fixed common SNP heritability (0.4) is assumed across all ancestries under a strong negative selection model for the relationship between SNP effect size and allele frequency. The genetic correlation in SNP effect size is set to 0.8 across all pairs of populations. The causal SNP proportion (degree of polygenicity) is set to 1.0%, 0.1%, or 0.05% (~192K, 19.2K, or 9.6K causal SNPs). We generate data for ~19 million common SNPs (MAF $\geq$ 1%) across the five ancestries but conduct analyses only on the ~2.0 million SNPs in HapMap 3 + MEGA. The PRS-CSx software only considers approximately 1.2 million HapMap 3 SNPs and, therefore, we report the performance of PRS-CSx PRSs based only on the HapMap 3 SNPs. The discovery GWAS sample size is set to (A) 15,000 or (B) 80,000 for each non-EUR ancestry, and 100,000 for EUR. A tuning set consisting of 10,000 individuals is used for parameter tuning and training the SL in CT-SLEB and MUSSEL or the linear combination model in weighted C + T, weighted LDpred2, and PRS-CSx. The reported R^2 values and the corresponding 95% bootstrap CIs are calculated based on an independent testing set of 10,000 individuals for each ancestry group.

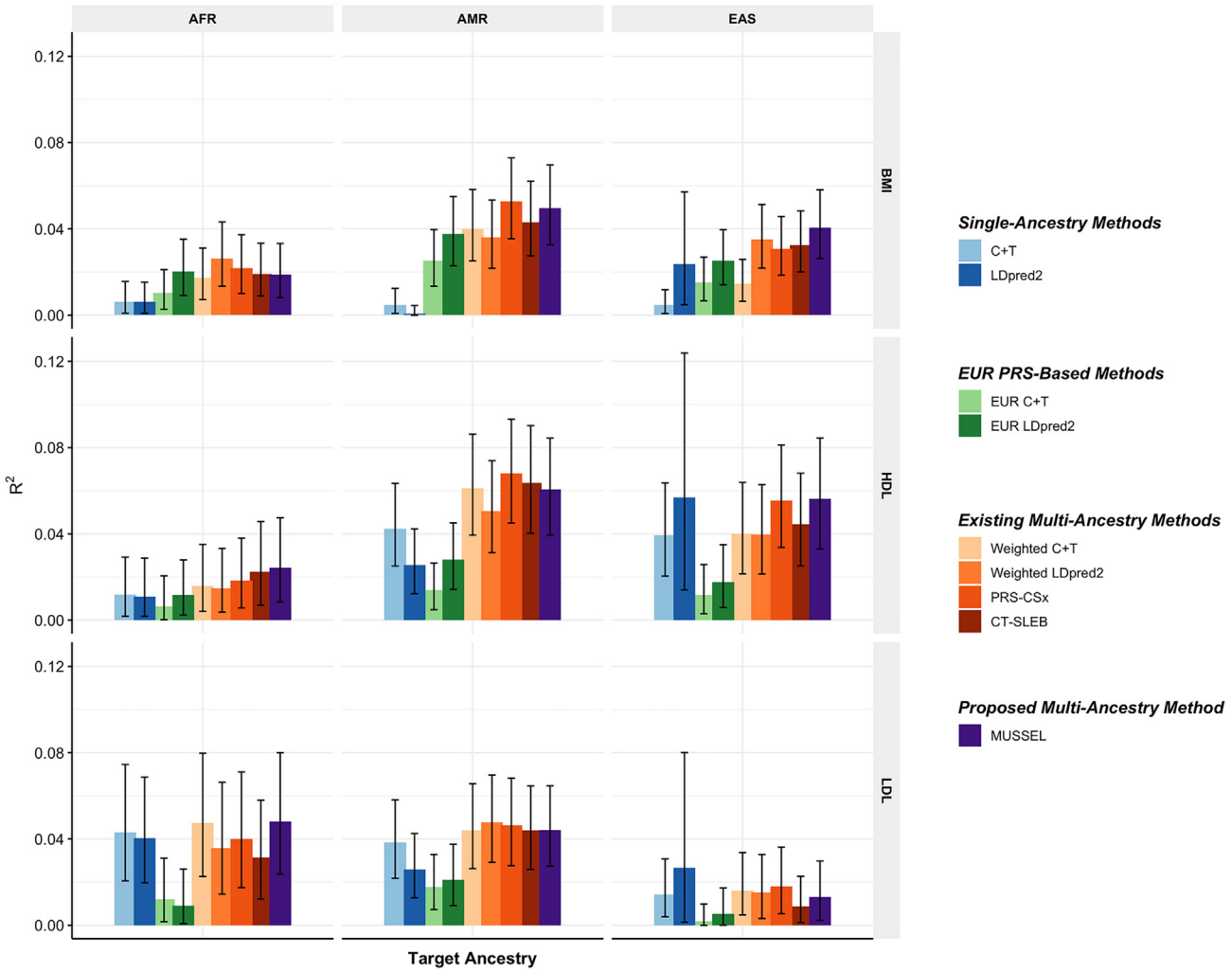


Figure 3. Prediction R^2 of the PRS trained based on GWAS summary data from PAGE + UKBB + BBJ on non-EUR validation individuals from PAGE

Discovery GWASs include GWAS from PAGE (AFR N_{GWAS} = 7,775–13,699, AMR N_{GWAS} = 13,894–17,558), BBJ (EAS N_{GWAS} = 70,657–158,284), and UKBB (EUR N_{GWAS} = 315,133–355,983). The validation dataset consists of individuals of EUR (N = 17,457–19,030), AFR (N = 7,954–8,598), EAS (N = 1,752–1,921), or SAS (N = 9,385–10,288) origin in UKBB. We used genotype data from the 1000 Genomes Project (498 EUR, 659 AFR, 347 AMR, 503 EAS, and 487 SAS) as the LD reference dataset. All methods were evaluated on the ~2.0 million SNPs that are available in HapMap 3 + MEGA, except for PRS-CSx, which is evaluated based on the HapMap 3 SNPs only, as implemented in their software. Ancestry- and trait-specific GWAS sample sizes, number of SNPs included, and validation sample sizes are summarized in Table S7. A random half of the validation individuals is used as the tuning set to tune model parameters as well as train the SL in CT-SLEB and MUSSEL or the linear combination model in weighted C + T, weighted LDpred2, and PRS-CSx. The other half of the validation set is used as the testing set to report R^2 values and the corresponding 95% bootstrap CIs for PRSs on each ancestry, after adjusting for whether the sample is from BioMe and the top ten genetic principal components for BMI, and additionally the age at lipid measurement and sex. Detailed results are reported in Table S17.

the alternative weighted MUSS approach has a similar or slightly lower R^2 than MUSSEL, but it still outperforms PRS-CSx, which utilizes the same linear combination strategy, for almost all traits and ancestry groups (Figure S12B).

AoU data analysis with validation on UKBB individuals

We also apply the various methods to develop ancestry-specific PRSs for height and BMI based on the GWAS summary data we generated from AoU for EUR, AFR, and AMR. The performance of the derived PRS is evaluated on UKBB validation samples of AFR ancestry. As in the GLGC data analysis, we first use genotype data of the unrelated 1000 Genomes samples as the LD reference data[24] (Figure 5A; Tables S12 and S13). Although no method is uniformly the best on all traits and ancestry groups, MUSSEL PRS on average has an R^2 that is 67.5% higher than that of the PRS-CSx PRS and 53.4% higher than that of the CT-SLEB PRS. MUSSEL PRS improves on top of the single-ancestry method by both the Bayesian modeling step (MUSS versus LDpred2, Figure 5A) and the SL step (MUSSEL versus MUSS, Figure S13A). The weighted MUSS PRS utilizing a linear combination strategy gives a lower R^2 than the MUSSEL PRS utilizing the EL strategy (weighted MUSS versus MUSSEL, Figure S13A).

Similar to the GLGC data analysis, we also conduct a sensitivity analysis whereby we estimate LD using the UKBB tuning

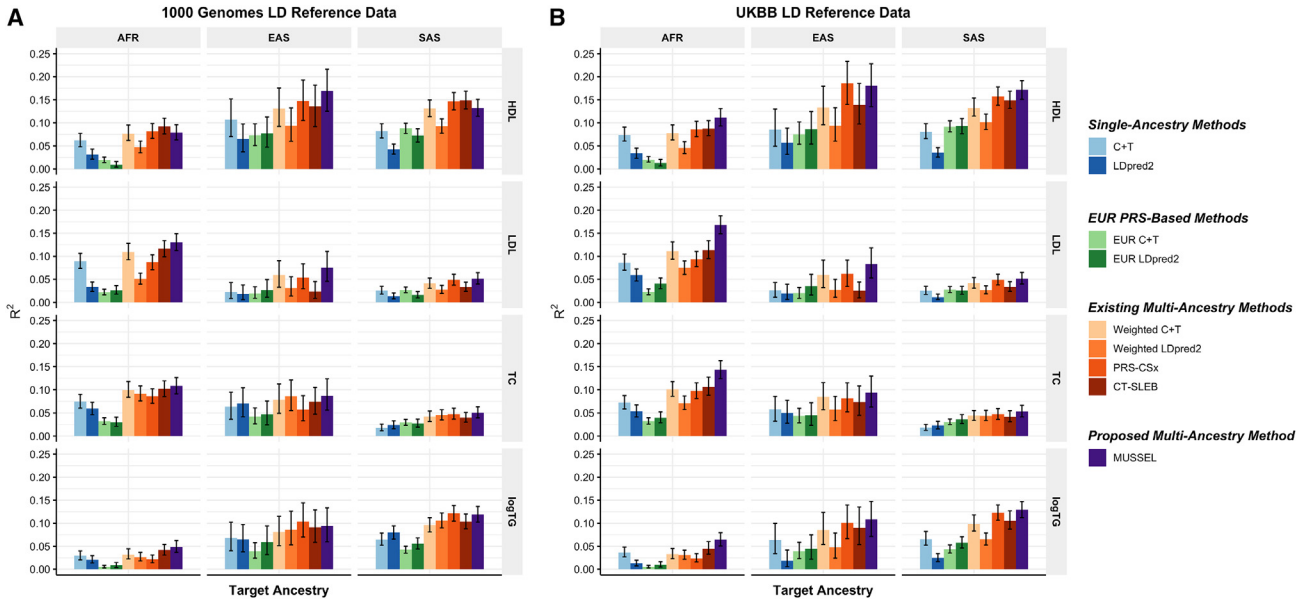


Figure 4. Prediction R^2 of the PRS trained based on GWAS summary data from GLGC on non-EUR validation individuals from UKBB

Discovery GWASs from GLGC include GWAS on EUR (N_{GWAS} = 842,660–930,671), AFR or admixed AFR (N_{GWAS} = 87,760–92,555), Hispanic/Latino (N_{GWAS} = 46,040–49,582), EAS (N_{GWAS} = 82,587–146,492), and SAS (N_{GWAS} = 33,658–34,135). The validation dataset consists of individuals of EUR (N = 17,457–19,030), AFR (N = 7,954–8,598), EAS (N = 1,752–1,921), or SAS (N = 9,385–10,288) origin in UKBB. The LD reference data are from either (A) the 1000 Genomes Project (498 EUR, 659 AFR, 347 AMR, 503 EAS, and 487 SAS), or (B) UKBB data (PRS-CSx: default UKBB LD reference data which overlap with our testing samples including 375,120 EUR, 7,507 AFR, 687 AMR, 2,181 EAS, and 8,412 SAS; all other methods: UKBB tuning samples including 10,000 EUR, 4,585 AFR, 1,010 EAS, and 5,427 SAS). The ancestry of UKBB individuals was determined by a genetic ancestry prediction approach (supplemental information). Due to the low prediction accuracy of genetic component analysis and extremely small validation sample size of UKBB AMR, prediction R^2 on UKBB AMR is unreliable and thus is not reported here. All methods were evaluated on the ~2.0 million SNPs that are available in HapMap 3 + MEGA, except for PRS-CSx, which is evaluated based on the HapMap 3 SNPs only, as implemented in their software. Ancestry- and trait-specific GWAS sample sizes, number of SNPs included, and validation sample sizes are summarized in Table S10. A random half of the validation individuals is used as the tuning set to tune model parameters as well as train the SL in CT-SLEB and MUSSEL or the linear combination model in weighted LDpred2, PRS-CSx, and weighted MUSS. The other half of the validation set is used as the testing set to report R^2 values and the corresponding 95% bootstrap CIs for each ancestry, after adjusting for age, sex, and the top ten genetic principal components. In (B), PRS-CSx and other methods do not have a fair comparison because the UKBB LD reference data provided by the PRS-CSx software ($UKBB_{PRS\text{-}CSx}$) is much larger than that for other methods, and thus the R^2 of PRS-CSx PRS may be inflated due to a large overlap between $UKBB_{PRS\text{-}CSx}$ and the UKBB testing sample. Detailed results are reported in Table S17.

samples (10,000 EUR, 4,585 AFR, 1,010 EAS, and 5,427 SAS) instead of the 1000 Genomes data. Different from the results from GLGC data analysis, no PRS has noticeably improved predictive power, even though there is a better ancestry match between the LD reference population and the target population (Figures 5B and S13B). Such results from the GLGC data analysis and the AoU data analysis suggest that for MUSSEL, the 1000 Genomes LD reference dataset may be adequate for building PRS models with relatively small discovery GWAS, such as the AoU GWAS (N = 15,364–48,332), but not so with much larger discovery GWAS, such as the GLGC GWAS (N up to 0.89 million). In other words, the ratio of the sample size of the LD reference dataset to the GWAS sample size may matter more than the sample size of the LD reference dataset itself or the population/ancestry match between datasets.

23andMe data analysis

We have collaborated with 23andMe (Sunnyvale, CA) to develop and validate PRSs for seven traits for EUR, African American (AFR), Latino (AMR), EAS, and SAS based on a large-scale dataset from 23andMe. We analyze two continuous traits, (1) heart metabolic disease burden and (2) height, and five binary traits, (3) any cardiovascular disease (CVD), (4) depression, (5) migraine diagnosis, (6) morning person, and (7) sing back musical note (SBMN). Results are summarized in Figure 6 and Tables S14 and S15. For the two continuous traits, MUSSEL shows a major advantage over the existing methods on AFR and AMR: for example, MUSSEL has a remarkable improvement over two recently proposed advanced methods that perform the best among the existing methods, PRS-CSx (average increase in R^2: 49.8%) and CT-SLEB (average increase in R^2: 47.5%). For EAS and SAS, MUSSEL performs better than all existing methods considered in all scenarios, except for heart metabolic disease burden in SAS, which has the smallest discovery GWAS (N = 20,062), where MUSSEL PRS has an R^2 value slightly lower than that of CT-SLEB PRS but higher than the R^2 value of all other PRS.

For the five binary traits, we observe a pattern similar to that of continuous traits, where MUSSEL generally performs better than or similarly to the best of the existing methods, and it shows the biggest improvement in residual area under the curve (AUC − 0.5) over existing methods on AFR (average

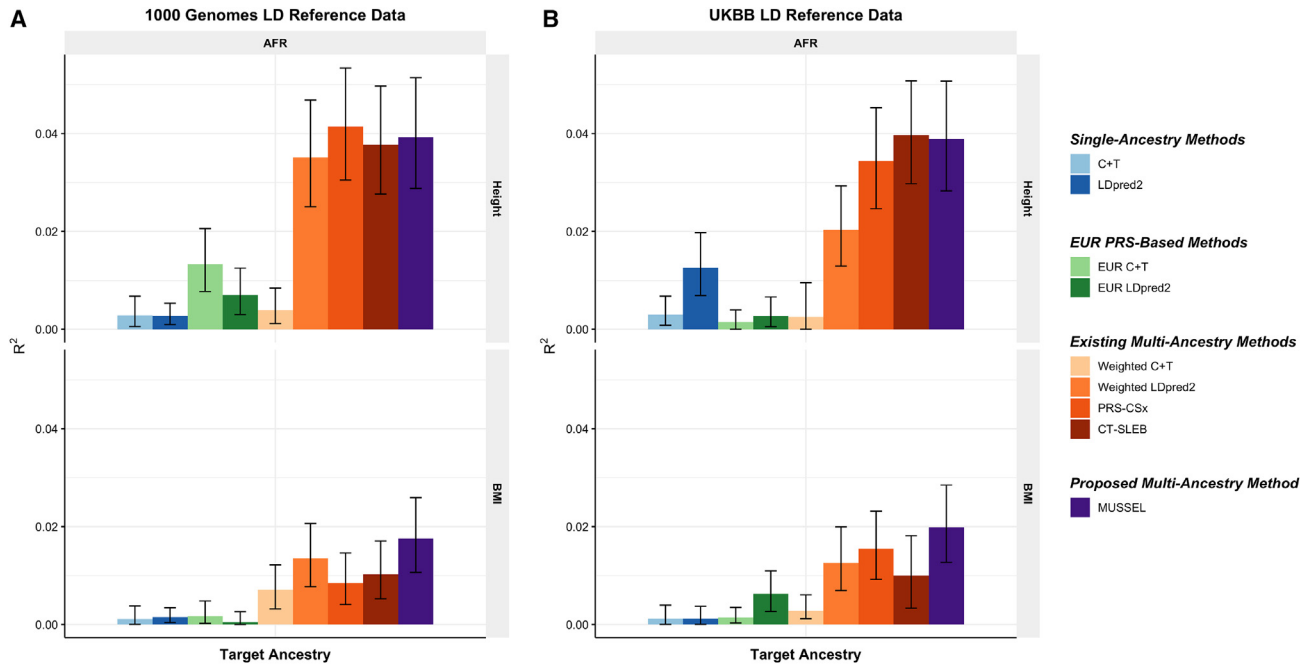


Figure 5. Prediction R^2 of the PRS trained based on GWAS summary data from AoU on non-EUR validation individuals from UKBB

Discovery GWASs from AoU include GWAS on EUR (N_{GWAS} = 48,229–48,332), AFR (N_{GWAS} = 21,514–21,550), and Hispanic/Latino (N_{GWAS} = 15,364–15,413). The validation dataset consists of individuals of AFR origin in UKBB (N = 9,026–9,042). The LD reference data are from either (A) the 1000 Genomes Project (498 EUR, 659 AFR, 347 AMR, 503 EAS, and 487 SAS) or (B) UKBB data (PRS-CSx: default UKBB LD reference data, which overlap with our testing samples including 375,120 EUR, 7,507 AFR, 687 AMR, 2,181 EAS, and 8,412 SAS; all other methods: UKBB tuning samples including 10,000 EUR, 4,585 AFR, 1,010 EAS, and 5,427 SAS). The ancestry of UKBB individuals was determined by a genetic ancestry prediction approach (supplemental information). Due to the low prediction accuracy of genetic component analysis and extremely small validation sample size of UKBB AMR, prediction R^2 on UKBB AMR is unreliable and thus is not reported here. All methods were evaluated on the ~2.0 million SNPs that are available in HapMap3 + MEGA, except for PRS-CSx, which is evaluated based on the HapMap 3 SNPs only, as implemented in their software. Ancestry- and trait-specific GWAS sample sizes, number of SNPs included, and validation sample sizes are summarized in Table S11. A random half of the validation individuals is used as the tuning set to tune model parameters as well as train the SL in CT-SLEB and MUSSEL or the linear combination model in weighted LDpred2, PRS-CSx, and weighted MUSS. The other half of the validation set is used as the testing set to report R^2 values for each ancestry, after adjusting for age, sex, and the top ten genetic principal components. Detailed 95% bootstrap CIs are reported in Table S17. In (B), PRS-CSx and other methods do not have a fair comparison because the UKBB LD reference data provided by the PRS-CSx software ($UKBB_{PRS\text{-}CSx}$) is much larger than that for other methods, and thus the R^2 of PRS-CSx may be inflated due to a large overlap between $UKBB_{PRS\text{-}CSx}$ and the UKBB testing sample.

improvement: 14.4%, Figure 6B; Tables S14 and S15). Averaged across all five traits and four non-EUR ancestry groups, MUSSEL PRS gives an (AUC − 0.5) that is 13.8% higher than that of the PRS-CSx PRS and 9.0% higher than that of the CT-SLEB PRS.

To examine the overall performance of the different methods, we further calculate the average R^2 and the corresponding 95% bootstrap CIs across all available traits in the PAGE + UKBB + BBJ, GLGC, and AoU data analyses for each ancestry group (Figure S16; Table S17). Overall, MUSSEL shows a significantly higher average R^2 than all existing methods on EUR ($p = 3.22 \times 10^{-4}$ for improvement compared to the second-best method) and AFR ($p = 7.10 \times 10^{-8}$), and a marginally significant increase in average R^2 for EAS ($p = 5.18 \times 10^{-2}$) and SAS ($p = 8.52 \times 10^{-2}$), while for AMR it has an average R^2 ($p = 0.616$) similar to that of the existing multi-ancestry methods. One of the reasons MUSSEL shows the most significant improvement on EUR and AFR is that the average R^2 for these two ancestry groups are calculated across all nine traits, with the largest total validation sample sizes that naturally lead to narrow CIs, while for EAS/SAS and AMR the average R^2 values are calculated across only seven traits and three traits, respectively. Nevertheless, we can observe a significant improvement in the average R^2 of MUSSEL for EUR and AFR and a potentially significant improvement for EAS and SAS as the number of traits and sample sizes increase, suggesting the promising gain in predictive power of MUSSEL compared to existing methods.

DISCUSSION

We propose MUSSEL, a powerful method for developing enhanced ancestry-specific PRSs integrating information from GWAS summary statistics and LD reference data across multiple ancestry groups. Based on an extension of spike-and-slab type prior,[14] MUSSEL enhances the ancestry-specific polygenic prediction by (1) borrowing information from GWAS of other ancestries via specification of a between-ancestry covariance structure in SNP effect sizes, (2) incorporating heterogeneity in LD and MAF distribution across ancestries, and (3) using an SL algorithm combining ancestry-specific PRS developed under various possible genetic architectures of the trait. We benchmark our

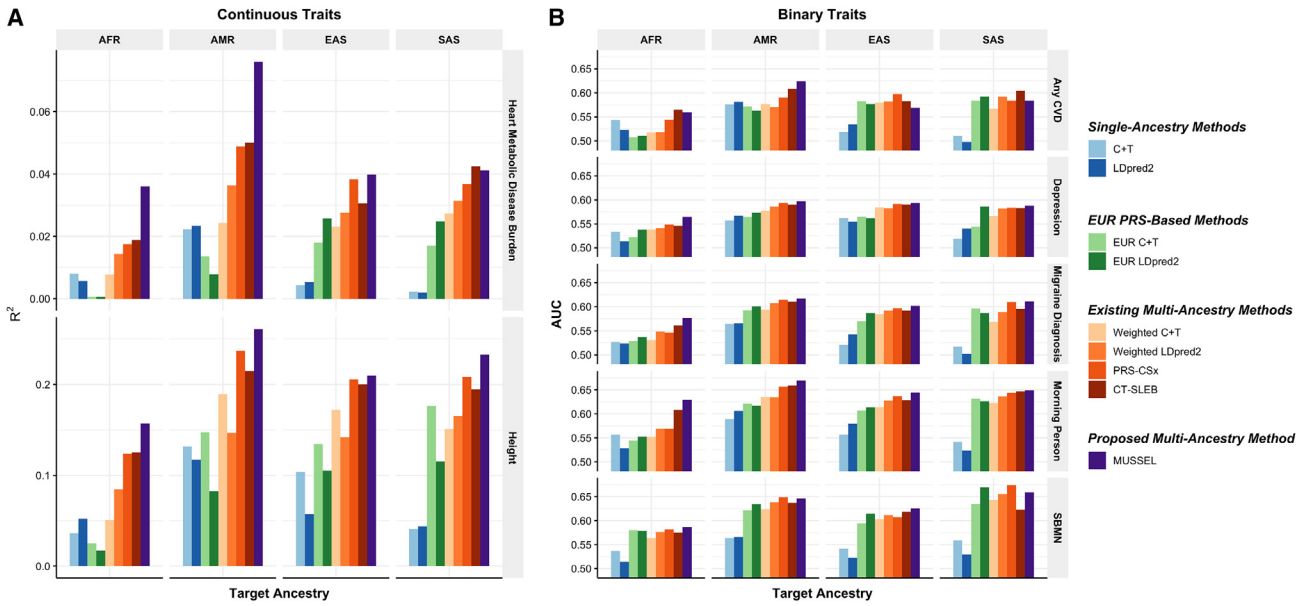


Figure 6. Prediction results on 23andMe validation individuals based on discovery GWAS from 23andMe on EUR, AFR, AMR, EAS, and SAS

The performance of the various methods is evaluated by (A) residual R^2 for two continuous traits, heart metabolic disease burden and height, and (B) residual AUC for five binary traits, any CVD, depression, migraine diagnosis, morning person, and SBMN, with LD reference data from the 1000 Genomes Project. The dataset is randomly split into 70%, 20%, and 10% for training GWAS, model tuning (tuning model parameters and training the SL in CT-SLEB and MUSSEL or the linear combination model in weighted LDpred2 and PRS-CSx), and testing (to report residual R^2 or AUC values after adjusting for the top five genetic principal components, sex, and age), respectively. All methods were evaluated on the ~2.0 million SNPs that are available in HapMap3 + MEGA, except for PRS-CSx, which is evaluated based on HapMap 3 SNPs only, as implemented in their software. Ancestry- and trait-specific GWAS sample sizes, number of SNPs included, and validation sample sizes are summarized in Table S14.

method against a wide variety of alternatives, including multiple state-of-the-art multi-ancestry methods,[8,9,13] using extensive simulation studies and data analyses. Results show that while no method is uniformly the best, MUSSEL is generally a robust method that shows close to optimal performance across a wide range of scenarios and has the potential to notably improve PRS performance in the AFR population compared to the alternative methods. While the 95% bootstrap CIs for PRS performance in individual settings can be wide and, thus, one cannot often claim superiority of one method over another with statistical significance, when we look at results across traits and studies, we observe that MUSSEL on average outperforms existing methods for optimal PRS development in the EUR, AFR, EAS, and SAS populations (Figure S16).

One important observation from the data applications is that the advantage of MUSSEL over existing methods tends to be more notable with larger GWASs accompanied by larger LD reference datasets. In the GLGC and 23andMe data analyses where the discovery GWAS sample sizes are relatively large, especially for the non-EUR populations, we can clearly observe that MUSSEL performs almost uniformly better than the existing methods. In contrast, in the PAGE + UKBB + BBJ data analysis, where the GWAS sample sizes for AFR and AMR are relatively small, MUSSEL sometimes shows a suboptimal performance. Such a trend of having more notable advantages with larger GWAS sample sizes and larger LD reference datasets exists not only when comparing MUSSEL to existing methods but also when comparing the more advanced methods, such as MUSSEL and PRS-CSx, to simpler alternatives, such as the weighted PRS method.

One key factor in implementing MUSSEL is the LD reference data. The analyses of the GLGC and AoU datasets illustrates that the sample size of the LD reference data should be sufficiently large relative to the discovery GWAS sample size to give MUSSEL an optimal performance (Figure 4; Tables S10 and S11). The performance of MUSSEL depends on estimated causal SNP proportion parameters from single-ancestry LDpred2 analysis. LDpred2 has previously been shown to underperform sometimes when using 1000 Genomes LD reference data[27] and thus could in turn affect the performance of MUSSEL. Thus, as sample sizes of the training GWAS increase, building a larger LD reference dataset than the widely used 1000 Genomes reference dataset will lead to more optimal performance.

The performance of MUSSEL is robust to modest ancestry mismatch between the discovery GWAS samples, LD reference samples, and validation samples, such as EUR in the United States (US) versus EUR in the United Kingdom (UK), as shown in the AoU data analysis. In our simulation study, we conducted a sensitivity analysis on the performance of MUSSEL given 20% ancestry mismatch between the discovery GWAS samples and LD reference samples. While the power loss of MUSSEL, as well as the LDpred-based methods, is within a reasonable range, an interesting finding is that the amount of power loss appears to increase as the underlying causal SNP proportion decreases. This suggests that, for MUSSEL and the LDpred-based

methods, ancestry mismatch between samples may be a more severe issue for those traits affected by a small number of large-effect SNPs. Ideally, the populations should be matched as closely as possible between GWAS samples, LD reference samples, and validation samples to ensure optimal performance of MUSSEL. However, if there is slight LD mis-specification, e.g., using samples from White population in the UK to estimate LD among the White population in the US, our analyses on the simulated and real datasets suggest that the power of MUSSEL may be slightly worse but still comparable.

There are several practical considerations regarding the implementation of MUSSEL and other multi-ancestry methods. First, the SL step in MUSSEL and CT-SLEB needs to be implemented with caution. We have shown by our data examples that the SL algorithm combining PRS models across various tuning parameter settings could yield additional improvement in predictive power. With a limited tuning sample, however, the SL might be overfitted in the presence of a large number of tuning-parameter settings, ultimately leading to low predictive power in an independent sample. Our analysis of the simulated data suggests that the performance of SL combining 30 different PRS models is typically stable when the effective sample size of the tuning dataset is no less than 1,000 for continuous traits. The required tuning sample size will also increase as the number of PRS models included in SL increases. Second, the advanced multi-ancestry methods, such as PRS-CSx, CT-SLEB, and MUSSEL, may not yield higher predictive power if the training GWAS sample size is too small. We expect the advanced multi-ancestry methods to outperform simpler methods when the GWAS sample size is relatively large (e.g., over 15,000 per ancestry group as in the AoU data analysis). When discovery GWAS for the target non-EUR ancestry group is relatively small (several thousand samples or fewer), a single-ancestry PRS model trained on the basis of the much larger EUR GWAS may outperform the multi-ancestry methods.

We have compared MUSSEL with a series of recent multi-ancestry methods including PRS-CSx and CT-SLEB, but there are other recently proposed methods that are worth investigating. In fact, we have implemented two other multi-ancestry methods named XPASS and PolyPred+ in our simulation study as well as GLGC, AoU, and 23andMe data analyses, with detailed results reported in Zhang et al.[13] Although computationally super-fast, XPASS, which uses a bivariate normal prior under an infinitesimal model, can only combine up to two ancestry groups, and it is always outperformed by MUSSEL (Tables S8, S9, S10, S11, S12, S13, S14, and S15). This shows the importance of including sparsity components in modeling effect-size distribution for Bayesian polygenic prediction. PolyPred+ implements a linear combination of SBayesR[28] trained separately on EUR and the target population and a PolyFun[29] PRS on EUR that additionally incorporates information from external functional annotations, and thus it is not directly comparable to the other methods. Even so, it performs worse than MUSSEL most of the time (Tables S8, S9, S10, S11, S12, S13, S14, and S15).

In our data examples, different methods show advantages in different scenarios in terms of GWAS sample size, LD reference data, the type of trait, and target ancestry. It is thus natural to consider extending our EL step from combining a series of PRSs trained within a specific type of method, such as MUSS, to those generated across different methods. MUSSEL can also be modified to enhance the performance of the PRS by borrowing information simultaneously across traits and genetically correlated traits. Two recent studies, both using simple weighting methods, have shown significant potential for cross-trait borrowing to improve PRS performance for individual traits.[30,31] There is, however, likely to be scope for additional improvement by developing formal Bayesian methods that can utilize flexible models for effect-size distribution simultaneously across ancestries and traits.

In summary, we propose a powerful method for constructing enhanced ancestry-specific PRSs combining GWAS summary data and LD reference data across multiple ancestry groups. As sample sizes of the multi-ancestry GWAS and LD reference datasets continue to increase, more advanced methods, such as MUSSEL and PRS-CSx,[9] are expected to show more and more advantages over simpler alternatives, such as the weighted methods.[8] Our large-scale simulation study and four unique data examples illustrate the relative performance of a variety of single- and multi-ancestry methods across various settings of ancestry groups, GWAS sample sizes, genetic architecture of the trait, and LD reference panel, which can serve as guidance for method implementation in future applications.

Limitations of the study

Our study has several limitations. First, the MUSS step requires two sets of tuning parameters, namely causal SNP proportion in each ancestry and between-ancestry correlation in effect sizes, the specification of which is relatively complex compared to other methods such as PRS-CSx. In the default setting of MUSS, the candidate values for genetic correlation between a pair of ancestry groups only lie between 0.7 and 0.95, while for some traits the estimated correlation can be lower.[9,25] However, given the high computational scalability of MUSS, when the number of ancestry groups is not too large ($K \leq 5$), prior information on genetic correlation can used to specify additional genetic correlation parameter settings to cover a wider range of potential genetic architectures of the trait. Second, all our analyses are based on a set of approximately 2.0 million SNPs selected on the basis of the combined content of HapMap 3 and the MEGA SNP array. While this SNP set is considered very informative for multi-ancestry genetic studies, we have previously shown that it is possible to increase PRS performance, especially in the AFR populations, by including much larger SNP contents. Future research is needed to improve scalability of the methods such as PRS-CSx and MUSSEL to datasets with larger SNP contents.

The spike-and-slab type prior in MUSSEL can be suboptimal for effect-size distribution of some traits. For example, in GLGC GWAS, we detect several top SNPs with extremely large association coefficients for all four blood lipid traits, each contributing to 0.6%–3.9% of the estimated total heritability. In this case, the Bayesian step in MUSSEL induces the same amount of shrinking on all SNPs, resulting in over-shrinkage on the few large-effect SNPs. We have considered a simple

alternative approach to compensate such overshrinkage,[32,33] whereby for each target ancestry group we first construct a "top-SNP PRS" using GWAS association coefficients of the few top SNPs for the ancestry, then combine it with the MUSSEL PRS constructed on the basis of the remaining SNPs. This approach, however, does not provide a more powerful PRS. PRS-CSx, which allows a heavy-tail Strawderman-Berger prior, while theoretically expected to be advantageous for handling such large-effect SNPs, does not show much advantage either. In the future, other heavy-tail type priors, such as the Bayesian Lasso (i.e., Laplacian),[34] Horseshoe,[35] and Bayesian Bridge,[36] are worth investigating. Another potential limitation of the method originates in the SL step: when the tuning sample is small (e.g., <1,000), the prediction algorithms utilized in SL may be overfitted in the presence of a large number of tuning parameters, ultimately leading to low predictive power in an independent sample.

CONSORTIA

The members of the 23andMe Research Team include 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, and Corinna D. Wong.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Details of MUSSEL step 1: MUSS
  - Existing methods
  - Detailed simulation setup
  - Runtimes and memory usage
  - PAGE + UKBB + BBJ data analysis with validation on non-EUR individuals from PAGE
  - GLGC data analysis with validation on UKBB individuals
  - AoU data analysis with validation on UKBB individuals
  - Predicted genetic ancestry for non-EUR individuals in UKBB
  - 23andMe data analysis
  - Calculation of the 95% bootstrap confidence intervals and p values for comparing R^2 between methods

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2024.100539.

ACKNOWLEDGMENTS

This work was supported by the following National Institutes of Health (NIH) grants: R00 HG012223 (J.J.), K99 CA256513 (H.Z.), R01 HG010480 (N.C., J.J., and J. Zhang), U01 CA249866 (N.C.), R35 HG011944 (G.L.W.), and U01 HG007419 (G.L.W.). We thank the Neale Lab and BBJ for making the GWAS summary data from UKBB and BBJ publicly available. Individual-level genotype and phenotype data for UKBB validation samples were obtained under application 17731. The PAGE Study is supported by the following NIH grants: U01 HG007419, R01 HG010297, and R01 HL151152. The All of Us Research Program is supported by the NIH, 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. We would like to thank the research participants and employees of 23andMe for making this work possible. We would like 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. We would like to thank the research participants and employees of 23andMe for making this work possible. The analyses in this paper utilized the high-performance computation Biowulf cluster at NIH, Faculty of Arts and Sciences Research Computing Cluster at Harvard University, and the Joint High Performance Computing Exchange at Johns Hopkins Bloomberg School of Public Health.

AUTHOR CONTRIBUTIONS

J.J. and N.C. developed all methods. J.J., J. Zhang, and H.Z. conducted data analyses under the supervision of N.C. H.Z. created the simulated datasets and ran GWASs on the simulated training data under the supervision of N.C. G.W. ran GWASs on the training data from the PAGE consortium. R.Z. ran GWASs on the training data from AoU under the supervision of N.C. J. Zhan, J.O., and Y.J. ran GWASs for training data from 23andMe Inc. under the supervision of B.L.K. J.J. developed the software. J.J. and N.C. drafted the manuscript, and H.Z., J. Zhang, and G.W. provided comments. All co-authors reviewed and approved the final version of the manuscript.

DECLARATION OF INTERESTS

J. Zhan, Y.J., J.O., and B.L.K. are employed by and hold stock or stock options in 23andMe, Inc.

Received: March 23, 2023
Revised: September 7, 2023
Accepted: March 14, 2024
Published: April 10, 2024

REFERENCES

1. Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., Peterson, R., and Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. Nat. Commun. *10*, 3328. https://doi.org/10.1038/s41467-019-11112-0.
2. Liu, C., Zeinomar, N., Chung, W.K., Kiryluk, K., Gharavi, A.G., Hripcsak, G., Crew, K.D., Shang, N., Khan, A., Fasel, D., et al. (2021). Generalizability of Polygenic Risk Scores for Breast Cancer Among Women With European, African, and Latinx Ancestry. JAMA Netw. Open *4*, e2119084. https://doi.org/10.1001/jamanetworkopen.2021.19084.
3. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518. https://doi.org/10.1038/s41586-019-1310-4.
4. Yu, Z., Jin, J., Tin, A., Köttgen, A., Yu, B., Chen, J., Surapaneni, A., Zhou, L., Ballantyne, C.M., Hoogeveen, R.C., et al. (2021). Polygenic Risk Scores for Kidney Function and Their Associations with Circulating Proteome, and Incident Kidney Diseases. J. Am. Soc. Nephrol. *32*, 3161–3173. https://doi.org/10.1681/ASN.2020111599.
5. Rabinowitz, J.A., Jin, J., Kahn, G., Kuo, S.I.C., Campos, A., Renteria, M., Benke, K., Wilcox, H., Ialongo, N.S., Maher, B.S., et al. (2021). Genetic propensity for risky behavior and depression and risk of lifetime suicide attempt among urban African Americans in adolescence and young adulthood. Am. J. Med. Genet. B Neuropsychiatr. Genet. *186*, 456–468. https://doi.org/10.1002/ajmg.b.32866.
6. Perkins, D.O., Olde Loohuis, L., Barbee, J., Ford, J., Jeffries, C.D., Addington, J., Bearden, C.E., Cadenhead, K.S., Cannon, T.D., Cornblatt, B.A., et al. (2020). Polygenic Risk Score Contribution to Psychosis Prediction in a Target Population of Persons at Clinical High Risk. Am. J. Psychiatr. *177*, 155–163. https://doi.org/10.1176/appi.ajp.2019.18060721.
7. Kachuri, L., Chatterjee, N., Hirbo, J., Schaid, D.J., Martin, I., Kullo, I.J., Kenny, E.E., Pasaniuc, B., Polygenic Risk Methods in Diverse Populations PRIMED Consortium Methods Working Group; Witte, J.S., and Ge, T. (2024). Principles and methods for transferring polygenic risk scores across global populations. Nat. Rev. Genet. *25*, 8–25. https://doi.org/10.1038/s41576-023-00637-2.
8. Márquez-Luna, C., Loh, P.R., and South Asian Type 2 Diabetes SAT2D Consortium; SIGMA Type 2 Diabetes Consortium; and Price, A.L. (2017). Multiethnic polygenic risk scores improve risk prediction in diverse populations. Genet. Epidemiol. *41*, 811–823. https://doi.org/10.1002/gepi.22083.
9. Ruan, Y., Lin, Y.-F., Feng, Y.-C.A., Chen, C.-Y., Lam, M., Guo, Z., Stanley Global Asia Initiatives; He, L., Sawa, A., Martin, A.R., et al. (2022). Improving polygenic prediction in ancestrally diverse populations. Nat. Genet. *54*, 573–580. https://doi.org/10.1038/s41588-022-01054-7.
10. Cai, M., Xiao, J., Zhang, S., Wan, X., Zhao, H., Chen, G., and Yang, C. (2021). A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. Am. J. Hum. Genet. *108*, 632–655. https://doi.org/10.1016/j.ajhg.2021.03.002.
11. Tian, P., Chan, T.H., Wang, Y.F., Yang, W., Yin, G., and Zhang, Y.D. (2022). Multiethnic polygenic risk prediction in diverse populations through transfer learning. Front. Genet. *13*, 906965. https://doi.org/10.3389/fgene.2022.906965.
12. Sun, Q., Rowland, B.T., Chen, J., Mikhaylova, A.V., Avery, C., Peters, U., Lundin, J., Matise, T., Buyske, S., Tao, R., et al. (2022). Improving polygenic risk prediction in admixed populations by explicitly modeling ancestral-specific effects via GAUDI. Preprint at bioRxiv. https://doi.org/10.1101/2022.10.06.511219.
13. Zhang, H., Zhan, J., Jin, J., Zhang, J., Lu, W., Zhao, R., O'Connell, J., Yu, Z., O'Connell, J., Jiang, Y., et al. (2023). A new method for ancestry polygenic prediction improves performance across diverse populations. Nat. Genet. *55*, 1757–1768. https://doi.org/10.1038/s41588-023-01501-z.
14. Privé, F., Arbel, J., and Vilhjálmsson, B.J. (2021). LDpred2: better, faster, stronger. Bioinformatics *36*, 5424–5431. https://doi.org/10.1093/bioinformatics/btaa1029.
15. Privé, F., Vilhjálmsson, B.J., Aschard, H., and Blum, M.G.B. (2019). Making the Most of Clumping and Thresholding for Polygenic Scores. Am. J. Hum. Genet. *105*, 1213–1221. https://doi.org/10.1016/j.ajhg.2019.11.001.
16. van der Laan, M.J., Polley, E.C., and Hubbard, A.E. (2007). Super learner. Stat. Appl. Genet. Mol. Biol. *6*, Article25. https://doi.org/10.2202/1544-6115.1309.
17. Zou, H., and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. J. Roy. Stat. Soc. B *67*, 301–320.
18. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Software *33*, 1–22. https://doi.org/10.18637/jss.v033.i01.
19. Weissbrod, O., Kanai, M., Shi, H., Gazal, S., Peyrot, W.J., Khera, A.V., Okada, Y., Biobank Japan Project; Martin, A.R., Finucane, H.K., and Price, A.L. (2022). Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. Nat. Genet. *54*, 450–458. https://doi.org/10.1038/s41588-022-01036-9.
20. International HapMap 3 Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58. https://doi.org/10.1038/nature09298.
21. Bien, S.A., Wojcik, G.L., Zubair, N., Gignoux, C.R., Martin, A.R., Kocarnik, J.M., Martin, L.W., Buyske, S., Haessler, J., Walker, R.W., et al. (2016). Strategies for Enriching Variant Coverage in Candidate Disease Loci on a Multiethnic Genotyping Array. PLoS One *11*, e0167758. https://doi.org/10.1371/journal.pone.0167758.
22. DiCiccio, T.J., and Efron, B. (1996). Bootstrap Confidence Intervals. Stat. Sci. *11*, 189–212. https://doi.org/10.1214/ss/1032280214.
23. Canty, A.J. (2002). Resampling methods in R: the boot package. https://journal.r-project.org/articles/RN-2002-017/RN-2002-017.pdf.
24. Siva, N. (2008). 1000 Genomes project. Nat. Biotechnol. *26*, 256. https://doi.org/10.1038/nbt0308-256b.
25. Graham, S.E., Clarke, S.L., Wu, K.H.H., Kanoni, S., Zajac, G.J.M., Ramdas, S., Surakka, I., Ntalla, I., Vedantam, S., Winkler, T.W., et al. (2021). The power of genetic diversity in genome-wide association studies of lipids. Nature *600*, 675–679. https://doi.org/10.1038/s41586-021-04064-3.
26. Privé, F., Arbel, J., Aschard, H., and Vilhjálmsson, B.J. (2022). Identifying and correcting for misspecifications in GWAS summary statistics and polygenic scores. HGG Adv. *3*, 100136.
27. Ge, T., Chen, C.Y., Ni, Y., Feng, Y.C.A., and Smoller, J.W. (2019). Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat. Commun. *10*, 1776. https://doi.org/10.1038/s41467-019-09718-5.
28. Lloyd-Jones, L.R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K.E., Wang, H., Zheng, Z., Magi, R., Esko, T., et al. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat. Commun. *10*, 5086. https://doi.org/10.1038/s41467-019-12653-0.
29. Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S., Schoech, A.P., van de Geijn, B., Reshef, Y., Márquez-Luna, C., et al. (2020). Functionally informed fine-mapping and polygenic localization of complex trait heritability. Nat. Genet. *52*, 1355–1363. https://doi.org/10.1038/s41588-020-00735-5.
30. Truong, B., Hull, L.E., Ruan, Y., Huang, Q.Q., Hornsby, W., Martin, H., van Heel, D.A., Wang, Y., Martin, A.R., Lee, S.H., and Natarajan, P. (2023). Integrative polygenic risk score improves the prediction accuracy of complex traits and diseases. Preprint at medRxiv. https://doi.org/10.1101/2023.02.21.23286110.
31. Albiñana, C., Zhu, Z., Schork, A.J., Ingason, A., Aschard, H., Brikell, I., Bulik, C.M., Petersen, L.V., Agerbo, E., Grove, J., et al. (2023). Multi-PGS enhances polygenic prediction: weighting 937 polygenic scores. Nat. Commun. *14*, 4702. https://doi.org/10.1038/s41467-023-40330-w.

32. Yang, S., and Zhou, X. (2020). Accurate and Scalable Construction of Polygenic Scores in Large Biobank Data Sets. Am. J. Hum. Genet. *106*, 679–693. https://doi.org/10.1016/j.ajhg.2020.03.013.

33. Khan, A., Turchin, M.C., Patki, A., Srinivasasainagendra, V., Shang, N., Nadukuru, R., Jones, A.C., Malolepsza, E., Dikilitas, O., Kullo, I.J., et al. (2022). Genome-wide polygenic score to predict chronic kidney disease across ancestries. Nat. Med. *28*, 1412–1420. https://doi.org/10.1038/s41591-022-01869-1.

34. Park, T., and Casella, G. (2008). The Bayesian Lasso. J. Am. Stat. Assoc. *103*, 681–686. https://doi.org/10.1198/016214508000000337.

35. Carvalho, C.M., Polson, N.G., and Scott, J.G. (2009). Handling Sparsity via the Horseshoe. In Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, D. van Dyk and M. Welling, eds. (PMLR), pp. 73–80.

36. Polson, N.G., Scott, J.G., and Windle, J. (2014). The Bayesian bridge. J. Roy. Stat. Soc. B *76*, 713–733. https://doi.org/10.1111/rssb.12042.

37. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature *518*, 197–206. https://doi.org/10.1038/nature14177.

38. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al. (2013). Discovery and refinement of loci associated with lipid levels. Nat. Genet. *45*, 1274–1283. https://doi.org/10.1038/ng.2797.

39. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. *50*, 390–400. https://doi.org/10.1038/s41588-018-0047-6.

40. Akiyama, M., Okada, Y., Kanai, M., Takahashi, A., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2017). Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nat. Genet. *49*, 1458–1467. https://doi.org/10.1038/ng.3951.

41. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience *4*, 7. https://doi.org/10.1186/s13742-015-0047-8.

42. Shaun Purcell, C.C. PLINK 2.0. https://www.cog-genomics.org/plink/2.0/.

43. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium; Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295. https://doi.org/10.1038/ng.3211.

44. Berisa, T., and Pickrell, J.K. (2016). Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics *32*, 283–285. https://doi.org/10.1093/bioinformatics/btv546.

45. Mak, T.S.H., Porsch, R.M., Choi, S.W., Zhou, X., and Sham, P.C. (2017). Polygenic scores via penalized regression on summary statistics. Genet. Epidemiol. *41*, 469–480. https://doi.org/10.1002/gepi.22050.

46. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.R., Bhatia, G., Do, R., et al. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. Am. J. Hum. Genet. *97*, 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001.

47. Zhang, H., Zhan, J., Jin, J., Zhang, J., Lu, W., Zhao, R., O'Connell, J., Yu, Z., O'Connell, J., Jiang, Y., et al. (2022). A new Method for Multi-ancestry Polygenic Prediction Improves Performance across Diverse Populations. Preprint at bioRxiv. https://doi.org/10.1101/2022.03.24.485519.

48. Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. Bioinformatics *27*, 2304–2305. https://doi.org/10.1093/bioinformatics/btr341.

49. The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74. https://doi.org/10.1038/nature15393.

50. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. *12*, e1001779. https://doi.org/10.1371/journal.pmed.1001779.

51. Friedewald, W.T., Levy, R.I., and Fredrickson, D.S. (1972). Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clin. Chem. *18*, 499–502. https://doi.org/10.1093/clinchem/18.6.499.

52. Purcell, S., and Chang, C. PLINK 2.0. URL: www.cog-genomics.org/plink/2.0/.

53. Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. R. News *2*, 18–22.

54. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Simulated data	Zhang et al.[13]	Harvard Dataverse: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/COXHAP
EUR GWAS summary data for BMI, HDL, and LDL based on UKBB samples (GWAS round 2 from the Neale Lab)	Locke et al.[37]; Willer et al.[37,38]	http://www.nealelab.is/uk-biobank
Split GWAS summary data based on 80% individuals from PAGE for BMI, HDL, and LDL stratified for AFR and AMR	Wojcik et al.[3]	Zenodo: https://doi.org/10.5281/zenodo.10800703
EAS GWAS summary data from BBJ for BMI, HDL, and LDL	Kanai et al.[39,40]	http://jenger.riken.jp/en/result
GWAS summary data from GLGC for HDL, LDL, TC, and logTG stratified for EUR, AFR, AMR, EAS, and SAS	Graham et al.[25]	https://csg.sph.umich.edu/willer/public/glgc-lipids2021/results/ancestry_specific/
GWAS summary data from All of Us for BMI and height stratified for EUR, AFR, and AMR	This paper	Harvard Dataverse: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FAWEQK
LD information used in MUSSEL for EUR, AFR, AMR, EAS, and SAS	This paper	Zenodo: https://doi.org/10.5281/zenodo.10816301
1000 Genome Phase 3	Siva[24]	https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html
Software and algorithms		
PLINK 1.9	Chang et al.[41]	https://www.cog-genomics.org/plink
PLINK 2.0	Purcelland Chang[42]	https://www.cog-genomics.org/plink/2.0/
LDpred2	Privé et al.[14]	https://privefl.github.io/bigsnpr/articles/LDpred2.html
PRS-CSx	Ruan et al.[9]	https://github.com/getian107/PRScsx
CT-SLEB	Zhang et al.[13]	https://github.com/andrewhaoyu/CTSLEB
LDSC	Bulik-Sullivan et al.[43]	https://github.com/bulik/ldsc
MUSSEL	This paper	Zenodo: https://doi.org/10.5281/zenodo.10800738

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and code should be directed to and will be fulfilled by the lead contact, Jin Jin (jin.jin@pennmedicine.upenn.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The simulated genotype and phenotype data for 600K subjects of EUR, AFR, AMR, EAS, or SAS ancestry, as well as GWAS summary statistics, and SNP information can be accessed at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/COXHAP. The EUR GWAS summary data for BMI,[37] HDL,[38] and LDL[38] based on UKBB samples (GWAS round 2) published by the Neale Lab can be downloaded at http://www.nealelab.is/uk-biobank. The EAS GWAS summary data from BBJ for BMI,[40] HDL,[39] and LDL[39] were downloaded from http://jenger.riken.jp/en/result. Split GWAS summary data generated based on 80% of individuals from PAGE for BMI, HDL, and LDL stratified for AFR and AMR, as used in the training sets in our data analysis, are deposited to Zenodo (https://doi.org/10.5281/zenodo.10800703) and are available upon request (email to Jin.Jin@Pennmedicine.upenn.edu). Stratified GWAS summary data from PAGE for BMI, HDL and LDL for AFR and AMR (not split for training/validation sets) is available on LDHub (https://ldsc.broadinstitute.org). GWAS summary data from GLGC for HDL, LDL, TC, and logTG stratified for EUR, AFR, AMR, EAS, and SAS can be downloaded

at http://csg.sph.umich.edu/willer/public/glgc-lipids2021/results/ancestry_specific/. GWAS summary data from AoU for BMI and height stratified for EUR, AFR, and AMR are available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/FAWEQK. GWAS summary data from 23andMe Inc. for top 10,000 genetic markers associated with height, morning person, and SBMN across five ancestry groups has been made available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/3NBNCV. The full GWAS summary statistics for these three traits (height, morning person, and SBMN) are available through 23andMe to qualified researchers under an agreement with 23andMe Inc. that protects the privacy of the 23andMe participants. Please visit https://research.23andme.com/collaborate/#dataset-access/for more information and to request data access. GWAS summary statistics for the other four traits (any CVD, heart metabolic disease burden, depression, and migraine) will not be made available because of 23andMe business requirements. Participants included in our 23andMe data analysis provided informed consent and participated in the research online, under a protocol approved by the external AAHRPP-accredited IRB, Ethical & Independent Review Services. 1000 Genomes Phase 3 reference data can be downloaded from https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html. Our estimated LD block matrices and other LD information used in MUSSEL for EUR, AFR, AMR, EAS, and SAS for approximately 2.0 million SNPs in HapMap 3 plus MEGA based on 1000 Genomes LD reference panel or UKBB reference panel can be downloaded from Zenodo (DOI) or on Github at https://github.com/Jin93/MUSSEL. LD block information, including the start and end positions of each block, are extracted from the "lassosum" R package and can be downloaded from https://github.com/tshmak/lassosum. Original data source for Figures 2, 3, 4, 5, and 6 in the paper is available in Tables S1, S8, S10, S12, and S14, respectively.

- PLINK 1.9: https://www.cog-genomics.org/plink. PLINK 2.0: https://www.cog-genomics.org/plink/2.0/. LDpred2: https://privefl.github.io/bigsnpr/articles/LDpred2.html. The R package "bigsnpr" (1.6.1) used in the LDpred2 pipeline is available for download on Github at https://github.com/privefl/bigsnpr. PRS-CSx: https://github.com/getian107/PRScsx. CT-SLEB: https://github.com/andrewhaoyu/CTSLEB. LD score regression: https://github.com/bulik/ldsc.
- The MUSSEL software, along with the code for conducting simulation studies and data analyses in this paper can be accessed at https://github.com/Jin93/MUSSEL and on Zenodo (https://doi.org/10.5281/zenodo.10800738).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contacts upon request.

METHOD DETAILS

Details of MUSSEL step 1: MUSS

MUSS conducts Bayesian modeling to generate ancestry-specific MUSS PRS models through joint modeling of GWAS summary data across all available ancestry groups. This step models the genetic correlation structure in SNP effect size across ancestry groups while accounting for ancestry-specific LD and allele frequency information.

Suppose we are interested in predicting the polygenic risk of some trait Y based on genotype $\{G_j, j = 1,\ldots,M_k\}$, for an individual of ancestry $k = 1,2,\ldots,K$, with M_k denoting the number of SNPs with a minor allele frequency (MAF) > 0.01 in ancestry k. For demonstration purposes, we assume the trait is continuous, but the results can be directly applied to GWAS summary-level association statistics for discrete traits in the same manner. We assume all SNPs included are biallelic, i.e., each SNP only has two alleles observed in the population. For each ancestry group k, we assume a true additive model for genetic variation, $Y_k = \sum_{j=1}^{M_k} G_j \beta_{kj}^{(J)} + \epsilon_k$, where $\beta_{kj}^{(J)}$ denotes the underlying joint effect size of $G_j, j = 1,2,\ldots,M_k$, i.e., effect size after adjusting for the effect of other SNPs, for an individual of ancestry k, and ϵ_k denotes a zero-mean random error term that includes effects of risk factors other than SNPs. Suppose we have ancestry-specific GWAS summary data, $\{(\widehat{\beta}_{kj}, \widehat{\sigma}_{kj}^2), j = 1,2,\ldots,M_k, k = 1,2,\ldots, K\}$, specifically, the marginal effect sizes of the SNPs ($\widehat{\beta}_{kj}$ s) and their corresponding standard errors ($\widehat{\sigma}_{kj}^2$ s) from one-SNP-at-a-time regressions, $y_{ki} = G_{ji}\beta_{kj} + \upsilon_{ki}, i = 1,\ldots,N_k$, for $j = 1,\ldots,M_k$ and $k = 1,\ldots,K$. Here, i, j and k are the indices of GWAS sample, SNP, and ancestry, respectively, υ_{ki} denotes a zero-mean random error term that includes effects of other risk factors and all other SNPs, β_{kj}, M_k and N_k are the true marginal SNP effect sizes, total number of SNPs, and GWAS sample size, respectively, for ancestry k. Our goal is to obtain an estimate of the joint SNP effect sizes, $\widehat{\beta_{kj}}^{(J)}$ s, to construct polygenic risk model $PRS_k = \sum_{j=1}^{M_k} G_j \widehat{\beta_{kj}}^{(J)}$ for each ancestry group k.

Our analysis is conducted on the standardized scale, where G_{kj} s are assumed to be standardized to have a zero mean and unit variance and Y_k s are assumed to have a unit variance (for continuous traits). This is reflected by rescaling the GWAS summary statistics so that the variance is equal to the inverse of the GWAS sample size. For computational scalability, we divide the whole genome into a series of independent LD blocks,[44] each containing hundreds of (up to ~2900) SNPs, and only consider the between-SNP correlation within each LD block. Such a block structure for LD matrices is considered because it yields similar predictive power as the banded-structure LD matrices accounting for LD within a 3cM genetic distance suggested by LDpred2,[14] but it is computationally more efficient and requires less memory. We estimate LD matrices for SNPs within each LD block using PLINK 2.0[42] based on LD block segmentation in Berisa and Pickrell.[44] LD block information was extracted from the R package "lassosum."[45] Note that the LD block information is available for EUR (1747 blocks, median number of SNPs per block: 816), AFR (2626 blocks, median number of SNPs per block: 716), and EAS

(1489 blocks, median number of SNPs per block: 815), but not currently available for AMR and SAS, and thus we apply the EUR LD information on AMR and SAS for now.

We denote by $\beta_{l_k}^{(J)}$ and $\widehat{\beta}_{l_k}$ the vector of true joint effect sizes and marginal effect sizes estimated from GWAS, respectively, for SNPs within a specific LD block l_k in ancestry $k = 1,2,\ldots,K$. To conduct analyses on the standardized scale, we first divide each raw effect size estimate $\widehat{\beta}_{kj}$ by $\sqrt{N_{kj}\widehat{\sigma}_{kj}^2+\widehat{\beta}_{kj}^2}$. We can then write down the likelihood of the GWAS summary statistics, $\widehat{\boldsymbol{\beta}}_{l_k} \sim N(\boldsymbol{R}_{l_k}\boldsymbol{\beta}_{l_k}^{(J)},\boldsymbol{N}_{l_k}^{1/2}\boldsymbol{R}_{l_k}\boldsymbol{N}_{l_k}^{1/2})$, where $\boldsymbol{R}_{l_k}$ denotes the LD matrix of the SNPs within the LD block l_k, and $\boldsymbol{N}_{l_k}$ is a diagonal matrix with diagonal entries being the corresponding GWAS sample sizes for SNPs within the LD block. For population-specific SNPs, i.e., SNPs with an MAF > 0.01 in only one ancestry k, we assume a spike-and-slab prior as in LDpred2, $\beta_{kj}^{(J)} \sim N(0,\delta_{kj}h_k^2),\delta_{kj} \sim \mathrm{Ber}(p_k)$, where h_k^2 denotes the per-SNP heritability, δ_{kj} is the indicator of whether SNP j is causal in ancestry k, i.e., $\delta_{kj} = 1$ if $\beta_{l_{kj}}^{(J)} \neq 0$ and 0 otherwise, and p_k is the proportion of causal SNPs in ancestry k. For SNPs that have MAF > 0.01 in all ancestry groups, we induce a prior correlation structure between $\beta_{kj}^{(J)}$ and $\beta_{k'j}^{(J)}$ for $k,k' \in \{1,2,\ldots,K\}$. The prior distribution of the joint effect size $\beta_{kj}^{(J)}$ s given δ_{kj} s is then specified as follows,

$$\begin{pmatrix} \beta_{1j}^{(J)} \\ \ldots \\ \beta_{Kj}^{(J)} \end{pmatrix} \Big| \delta_{1j},\ldots,\delta_{Kj} \sim N(\boldsymbol{0},\Delta_j\Omega_j\Delta_j),$$

where $\Delta_j = \mathrm{diag}(\delta_{1j},\ldots,\delta_{Kj})$, and $(\Omega_j)_{k,k'} = \rho_{k,k'}h_kh_{k'}$, with $\rho_{k,k'}$ denoting the genetic correlation between ancestry groups k and k'. For SNPs that have an MAF > 0.01 in only a subset of ancestries $A \subset \{1,\ldots,K\}$, similar prior distributions can be specified for SNP effect sizes within the set of ancestry groups A.

Recall that we introduce variables $\{p_k = \Pr(\delta_{kj} = 1), \forall j,k = 1,\ldots,K\}$ to denote ancestry-specific causal SNP proportions, and for ancestry-specific SNPs, we assume $\delta_{kj} \sim \mathrm{Ber}(p_k)$. Now we generalize this Bernoulli prior to a multinomial prior on $(\delta_{1j},\ldots,\delta_{Kj})^T$ for SNPs that exist in a subset of ancestry groups $A \subset \{1,\ldots,K\}$, with probabilities $\{\Pr(\delta_{1j,j\in S_j} = 1,\delta_{1j,j\notin S_j} = 0), S_j \subset A\}$ being defined as functions of $p_k,k = 1,\ldots,K$. We first focus on SNPs that only exist in two ancestry groups $A = \{k_1,k_2\}$: we set $\Pr(\delta_{k_1j} = 1, \delta_{k_2j} = 1) = \min(p_{k_1},p_{k_2})$, which reflects our assumption that if an SNP is causal in one ancestry group, it is also causal in another. We can then obtain $\Pr(\delta_{k_1j} = 1,\delta_{k_2j} = 0) = p_{k_1} - \min(p_{k_1},p_{k_2})$, $\Pr(\delta_{k_1j} = 0,\delta_{k_2j} = 1) = p_{k_2} - \min(p_{k_1},p_{k_2})$, and $\Pr(\delta_{k_1j} = 0,\delta_{k_2j} = 0) = 1 - p_{k_1} - p_{k_2} + \min(p_{k_1},p_{k_2})$. After constructing $\Pr(\delta_{k_1j},\delta_{k_2j})$ s, we then construct priors for SNPs that exist in three ancestry groups: by specifying $\Pr(\delta_{k_1j} = 1,\delta_{k_2j} = 1,\delta_{k_3j} = 1) = \min(p_{k_1},p_{k_2},p_{k_3})$, we can obtain the rest of the probabilities $\{\{\Pr(\delta_{k_1j} = a_1,\delta_{k_2j} = a_2,\delta_{k_3j} = a_3),a_1,a_2,a_3 \in \{0,1\},1 \le k_1 < k_2 < k_3 \le K\}\}$. Such specifications can be easily extended to apply to SNPs that exist in four ancestry groups, five ancestry groups, etc.

We estimate $\widehat{\beta_{kj}}^{(J)}$ s based on MCMC with an approximation strategy previously implemented in the LDpred2 algorithm,[14] which substantially reduces computation time of the algorithm. There are two sets of tuning parameters which will be estimated by grid search using a tuning dataset independent from the testing samples on which we report R^2: (1) the ancestry-specific causal SNP proportions $(p_1,\ldots,p_K)$: we fix $(p_1,\ldots,p_K)$ to either $(\tilde{p}_1,\ldots,\tilde{p}_K)$, the estimated ancestry-specific causal SNP proportions obtained from LDpred2 separately on GWAS summary data of each ancestry, or $(\tilde{p}_s,\ldots,\tilde{p}_s),s = 1,\ldots,K$, i.e., the values of all p_k s are set to the LDpred2 estimate of the causal SNP proportion in ancestry s; (2) the between-ancestry correlation parameters $\rho_{kk'}$ s: we consider two settings, i.e., either set $\rho_{kk'}$ s to all equal to $\rho = 0.7$, 0.8, 0.9, or 0.95, or set $\rho_{kk'}$ to 0.75 for any pair of ancestry groups that include AFR and 0.9 otherwise, given that correlation with AFR tends to be weaker than that among other ancestry groups. Prior to the implementation of MCMC, we further estimate the ancestry-specific heritability H_k^2 s based on GWAS summary data and LD reference data using LD score regression[43] (Table S16).

We now describe the detailed MCMC algorithm and estimation procedure. For SNPs that only exist (MAF > 0.01) in one ancestry group, the Gibbs sampler in Vilhjálmsson et al.[46] was implemented. For each SNP j that exists in all K ancestry groups, we sample $\boldsymbol{\delta}_j = (\delta_{1j},\ldots,\delta_{Kj})^T$ and $\boldsymbol{\beta}_j = (\beta_{1j},\ldots,\beta_{Kj})^T$ from

$$f(\boldsymbol{\beta}_j,\boldsymbol{\delta}_j|\widehat{\boldsymbol{\beta}},\boldsymbol{\beta}_{-j}) \approx f(\boldsymbol{\delta}_j|\widehat{\boldsymbol{\beta}}_j,\boldsymbol{\beta}_{-j})f(\boldsymbol{\beta}_j|\boldsymbol{\delta}_j,\widehat{\boldsymbol{\beta}}_j,\boldsymbol{\beta}_{-j}),$$

where $\boldsymbol{\beta}_{-j}$ denotes the joint effect sizes for the SNPs within the LD block which SNP j is in, $l_{kj},k \in \{1,\ldots,K\}$.

We first sample $\boldsymbol{\delta}_j$ from $f(\boldsymbol{\delta}_j|\widehat{\boldsymbol{\beta}}_j,\boldsymbol{\beta}_{-j})$. Here note that obtaining $f(\boldsymbol{\delta}_j|\widehat{\boldsymbol{\beta}},\boldsymbol{\beta}_{-j})$ analytically is hard, and thus we approximate it by $f(\boldsymbol{\delta}_j|\widehat{\boldsymbol{\beta}}_j,\boldsymbol{\beta}_{-j})$. For a realization of $\boldsymbol{\delta}_j,\boldsymbol{r} = (r_1,\ldots,r_K)^T$ where $r_k \in \{0,1\},\forall k$, we first derive

$$f(\boldsymbol{\delta}_j = \boldsymbol{r}|\widehat{\boldsymbol{\beta}}_j,\boldsymbol{\beta}_{-j}) = \frac{f(\widehat{\boldsymbol{\beta}}_j|\boldsymbol{\delta}_j = \boldsymbol{r},\boldsymbol{\beta}_{-j})\Pr(\boldsymbol{\delta}_j = \boldsymbol{r})}{\sum_{\boldsymbol{r}'}(\widehat{\boldsymbol{\beta}}_j|\boldsymbol{\delta}_j = \boldsymbol{r}',\boldsymbol{\beta}_{-j})\Pr(\boldsymbol{\delta}_j = \boldsymbol{r}')}.$$

We denote the numerator by $J_{\boldsymbol{r}} = f(\widehat{\boldsymbol{\beta}}_j|\boldsymbol{\delta}_j = \boldsymbol{r},\boldsymbol{\beta}_{-j})\Pr(\boldsymbol{\delta}_j = \boldsymbol{r})$, which can be derived as follows:

$$J_{\boldsymbol{r}} = \Pr(\boldsymbol{\delta}_j = \boldsymbol{r}) \int f(\widehat{\boldsymbol{\beta}}_j|\boldsymbol{\delta}_j = \boldsymbol{r},\boldsymbol{\beta}_{-j},\boldsymbol{\beta}_j)f(\boldsymbol{\beta}_j|\boldsymbol{\delta}_j = \boldsymbol{r})d\boldsymbol{\beta}_j$$

$$= \Pr(\delta_{1j} = r_1, \ldots, \delta_{Kj} = r_K) \int N\left(\widehat{\beta}_j \middle| \begin{pmatrix} \sum_{j' \neq j, j' \in I_{1j}} \boldsymbol{R}_{I_{1j}, jj'} \beta_{1j'} + \beta_{1j} r_1 \\ \cdots \\ \sum_{j' \neq j, j' \in I_{Kj}} \boldsymbol{R}_{I_{Kj}, jj'} \beta_{Kj'} + \beta_{Kj} r_K \end{pmatrix}, diag\left(\frac{1}{N_{1j}}, \ldots, \frac{1}{N_{Kj}}\right)\right) N\left(\begin{pmatrix} \beta_{1j} r_1 \\ \cdots \\ \beta_{Kj} r_K \end{pmatrix} \middle| \mathbf{0}, \Delta_j \Omega_j \Delta_j\right) d\boldsymbol{\beta}_j$$

$$= \Pr(\delta_{1j} = r_1, \ldots, \delta_{Kj} = r_K) \times N\left(\left(\tilde{\beta}_{1j,r'}, \ldots, \tilde{\beta}_{Kj,r'}\right)^T, diag\left(\frac{1}{N_{1j}}, \ldots, \frac{1}{N_{Kj}}\right) + \Delta_{j,r'} \Omega_j \Delta_{j,r'}\right),$$

where

$$\tilde{\beta}_{kj,r'} = \widehat{\beta}_{kj} - \sum_{j' \neq j, j' \in I_{kj}} \boldsymbol{R}_{I_{kj}, jj'} \beta_{kj'} + \beta_{kj} r_k,$$

$\boldsymbol{R}_{I_{kj}, jj'}$ denotes the entry in $\boldsymbol{R}_{I_{kj}}$ that corresponds to the correlation between SNPs j and j', $I_a = 1$ if $a \neq 0$ and 0 otherwise, $\Delta_{j,r'} = \text{diag}(r'_1, \ldots, r'_K)$, and $(\Omega_j)_{k,k'} = \rho_{k,k'} h_k h_{k'}$. After deriving $J_{\boldsymbol{r}}$ s, we can then sample from $f(\boldsymbol{\delta}_j = \boldsymbol{r} | \widehat{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_{-j}) = J_{\boldsymbol{r}} / \left(\sum_{\boldsymbol{r}'} J_{\boldsymbol{r}'}\right)$.

We obtain the marginal posterior mean of $\boldsymbol{\beta}_j$ after integrating out $\boldsymbol{\delta}_j$:

$$E\left(\boldsymbol{\beta}_j \middle| \widehat{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_{-j}\right) = \sum_{\boldsymbol{r}'} E\left(\boldsymbol{\beta}_j \middle| \boldsymbol{\delta}_j = \boldsymbol{r}', \widehat{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_{-j}\right) \Pr\left(\boldsymbol{\delta}_j = \boldsymbol{r}' \middle| \widehat{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_{-j}\right),$$

where

$$f\left(\boldsymbol{\beta}_j \middle| \boldsymbol{\delta}_j = \boldsymbol{r}', \widehat{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_{-j}\right) \propto f\left(\widehat{\boldsymbol{\beta}}_j \middle| \boldsymbol{\delta}_j = \boldsymbol{r}', \boldsymbol{\beta}_j, \boldsymbol{\beta}_{-j}\right) f\left(\boldsymbol{\beta}_j \middle| \boldsymbol{\delta}_j = \boldsymbol{r}'\right).$$

We can easily derive that $\boldsymbol{\beta}_j | \boldsymbol{\delta}_j = \boldsymbol{r}', \widehat{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_{-j}$ follows $N(\boldsymbol{\mu}_{j,\boldsymbol{r}'}, \boldsymbol{V}_{j,\boldsymbol{r}'})$, where

$$\boldsymbol{V}_{j,\boldsymbol{r}'} = \left(diag(N_{1j}, \ldots, N_{Kj}) + (\Delta_{j,\boldsymbol{r}'} \Omega_j \Delta_{j,\boldsymbol{r}'})^{-1}\right)^{-1},$$

$$\boldsymbol{\mu}_{j,\boldsymbol{r}'} = \boldsymbol{V}_{j,\boldsymbol{r}'} \begin{pmatrix} N_{1j} \tilde{\beta}_{1j,r'} \\ \cdots \\ N_{1j} \tilde{\beta}_{Kj,r'} \end{pmatrix}.$$

For SNPs that have an MAF > 0.01 in a subset of ancestry groups $A \subset \{1, \ldots, K\}$, similar sampling strategy can be conducted but only among ancestry groups A. In each MCMC iteration, the prior per-SNP heritability parameter is set to $h_k^2 = \frac{H_k^2}{m_k}$, where m_k denotes the number of causal SNPs ($\sum_{j=1}^{M_k} \delta_{kj}$) estimated from this iteration. The posterior estimate of $\boldsymbol{\beta}_j$ is obtained by taking the average of $E(\boldsymbol{\beta}_j | \widehat{\boldsymbol{\beta}}_j, \boldsymbol{\beta}_{-j})$ obtained from 100(K-1) MCMC iterations after a burn-in stage of 100 iterations.

Existing methods

Single-ancestry methods

LD clumping and thresholding (C + T). C + T first constructs a series of PRS by applying an LD clumping step followed by a p value filtering step with varying p value cutoffs, then selects the best performing PRS on the tuning dataset. Specifically, an LD clumping step is first conducted to exclude variants that have an absolute pairwise correlation stronger than $r^2 = 0.1$ within a genetic distance (500kb) based on an LD reference dataset. The remaining variants are then filtered by excluding the ones that have a p value larger than a significance threshold, which, in our analysis, were set to $p_t = 5 \times 10^{-8}, 1 \times 10^{-7}, 5 \times 10^{-7}, 1 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-3}, 1 \times 10^{-2}, 5 \times 10^{-2}, 1 \times 10^{-1}, 5 \times 10^{-1}$, or 1. These 16 scores were created based on these 16 different significance thresholds p_t s by calculating a weighted sum of the number of effect alleles of the selected SNPs, with weights being the effect size estimates from the discovery GWAS. C + T then selects the score with the "optimal" p value thresholds via parameter tuning with respect to the residual R^2 (for continuous traits) or residual AUC (for binary traits) on a tuning dataset that is independent of the training and testing samples. C + T was implemented using PLINK 1.90.[41]

LDpred2. LDpred2 is an LD-based Bayesian modeling approach which leverages information from GWAS summary statistics and explicitly models LD correlation structure with correlation matrices being estimated based on an external reference panel.[14,46] LDpred2 assumes a spike-and-slab prior on SNP effect sizes, i.e., each SNP has a probability p to have a non-zero causal effect $\beta_j^{(J)} \sim N(0, h_g^2)$, and a probability $(1 - p)$ to have no contribution to the phenotypic variation ($\beta_j^{(J)} = 0$). Here p and the total heritability, H^2, are treated as tuning parameters and estimated via grid search on a tuning dataset. In each iteration of MCMC, the per-SNP heritability parameter is set to $h_g^2 = H^2/m$, where m is the number of causal SNPs detected in that iteration.

We ran LDpred2 on each chromosome and GWAS of each ancestry group separately using R packages "bigsnpr" (version 1.6.1). For our analyses on the simulated datasets, PAGE + UKBB + BBJ datasets, GLGC dataset and AoU dataset, we considered the "LDpred2 grid" model, where two tuning parameters were considered: (1) causal SNP proportion p, with default candidate values 1.0×10^{-4}, 1.8×10^{-4}, 3.2×10^{-4}, 5.6×10^{-4}, 1.0×10^{-3}, 1.8×10^{-3}, 3.2×10^{-3}, 5.6×10^{-3}, 1.0×10^{-2}, 1.8×10^{-2}, 3.2×10^{-2}, 5.6×10^{-2}, 1.0×10^{-1}, 1.8×10^{-1}, 3.2×10^{-1}, 5.6×10^{-1}, and 1.0; (2) total heritability H^2, which is set to the heritability estimated by LDSC[43] multiplied by 0.7, 1, or 1.4. The "sparse" option was not considered. In our 23andMe data analysis, we considered the "LDpred2 auto" model, which estimates p and h_g^2 along with the other model parameters instead of treating them as tuning parameters and estimating them based on a grid search. The reason we considered the "auto" option instead of the "grid" option is that the "grid" option gave nonconvergent estimates under all considered tuning parameter settings. This convergence issue of the LDpred2 grid algorithm may be due to the low ratio between the 1000 Genomes reference sample size and the large discovery sample size of the 23andMe GWAS. We have discussed this issue in our GLGC data analysis as well. Note that implementation of LDpred2 in this study follows the 2021-01-11 version of the LDpred2 tutorial using version 1.6.1 of the bigsnpr R package. The LDpred2 tutorial and the bigsnpr package have been updated since then, and some issues we encountered when running LDpred2 may have been resolved in the latest version of the LDpred2 algorithm.

Multi-ancestry methods

Weighted PRS. A simple multi-ancestry method is weighted PRS, which trains an "optimal" linear combination of the effect size estimates obtained based on training data from each single ancestry. Weighted PRS was first proposed in Marquez-Luna et al.[8] to improve the performance of single ancestry C + T PRS. Suppose we have constructed C + T PRS, PRS_{EUR}, PRS_{AFR}, PRS_{AMR}, PRS_{EAS}, and PRS_{SAS}, separately based on GWAS and LD reference panel of each corresponding ancestry group. The weighted C + T PRS is then constructed as $PRS_{wP+T} = \alpha_1 PRS_{EUR} + \alpha_2 PRS_{AFR} + \alpha_3 PRS_{AMR} + \alpha_4 PRS_{EAS} + \alpha_5 PRS_{SAS}$ where α_k s are obtained by fitting a regression model on the tuning dataset. Here we apply the weighted PRS approach on either C + T ("weighted C + T") or LDpred2 ("weighted LDpred2").

PRS-CSx. "PRS-CSx"[9] is proposed as the multi-ancestry version of PRS-CS[27] which conducts Bayesian modeling followed by an additional step of constructing a linear combination of the best performing PRS trained for each ancestry. PRS-CSx assumes a continuous shrinkage prior named Strawderman-Berger prior on the ancestry-specific effect sizes. For SNPs available in more than one population, this prior induces information sharing across ancestry groups. After the Bayesian modeling step, PRS-CSx further trains a linear combination of the ancestry-specific PRS obtained from the previous step based on the tuning dataset. In all our analyses, we ran PRS-CSx with the default candidate values for the tuning parameter φ (1.0, 10^{-2}, 10^{-4}, and 10^{-6}), which is the global shrinkage parameter shared by all SNPs and all ancestries that controls the overall causal SNP proportion. The PRS-CSx software only considers approximately 1.2 million HapMap 3 SNPs and therefore we only report the performance of PRS-CSx PRS based on the HapMap 3 SNPs. We have also tried to apply PRS-CSx to HapMap 3 SNPs plus an additional 0.8 million MEGA SNPs that are also available in the 1000 Genomes reference data. But we found that, on our simulated dataset, the performance of PRS-CSx PRS using the extended HapMap 3 + MEGA SNP set is significantly worse than PRS-CSx using the HapMap 3 SNPs, and in our real data analyses, results from PRS-CSx on the two SNP sets are similar. We therefore stick to the default setting with 1.2 million HapMap 3 SNPs provided by the PRS-CSx software.

CT-SLEB. CT-SLEB is a recently proposed method for multi-ancestry PRS construction.[13] It first conducts a two-dimensional C + T between EUR GWAS and GWAS of the target population to select SNPs to be included in the target population PRS, then uses an Empirical Bayesian approach to account for genetic correlation across populations, and finally implements an SL algorithm to combine PRS generated under different p value thresholds in the C + T step. In our analyses, we implemented CT-SLEB with the default setting for p value threshold, $p_t = 5 \times 10^{-8}$, 5×10^{-7}, 5×10^{-6}, 5×10^{-5}, 5×10^{-4}, 5×10^{-3}, 5×10^{-2}, 5×10^{-1}, or 1, and a genetic distance $d = 50/r^2$ or $100/r^2$, where $r^2 = 0.01$, 0.05, 0.1, 0.2, 0.5, or 0.8.

Detailed simulation setup

We investigated the performance of MUSSEL and a series of existing methods under various simulated scenarios of genetic architecture for phenotype and GWAS sample sizes across ancestries. This large-scale, multi-ancestry simulated dataset including 600,000 individuals across EUR, AFR, AMR, EAS, and SAS origins has recently been released by our group.[47] Specifically, the genotype data was simulated using HAPGEN2 (version 2.1.2)[48] based on the genotype data of 2,504 unrelated individuals from Phase 3 1000 Genomes Project (503 EUR, 661 AFR, 347 AMR, 504 EAS, and 489 SAS).[49] We have checked and confirmed the consistency between the LD pattern in the original 1000 Genomes reference data and the LD pattern in our simulated data.[47] Approximately 19.2 million common biallelic SNPs with MAF $\geq$ 0.01 in at least one ancestry group were included. For phenotype data, genetic architectures were simulated by first selecting a random set of 1.0%, 0.1%, or 0.05% SNPs across the whole genome to be causal, that is approximately 192K, 19.2K, or 9.6K causal SNPs among 19.2 million SNPs. Under a spike and slab structure, the nonzero standardized effect sizes for the causal SNPs were then generated under various negative selection models according to a function of allele frequency, $\beta_{kj}^{(J)} \propto \{q_{kj}(1 - q_{kj})\}^{\alpha}$: (1) strong negative selection: $\alpha = 0$, (2) mild negative selection: $\alpha = 0.75$, or (3) no negative selection, $\alpha = 1$. The genetic correlation was set to $\rho = 0.8$ or 0.6 between all pairs of ancestries. Specifically, we first generated $\nu_{kj} \sim N(0, H_k^2/m_k)$ for SNPs only existing in ancestry k, with $cov(\nu_{kj}, \nu_{k'j}) = \rho H_k H_{k'}/m_k m_{k'}$ for SNPs shared between ancestries k and k', where H_k^2 and m_k denote the total heritability and the number of causal SNPs,

respectively, in ancestry k. To control the total heritability at the predefined level H_k^2 s, we set the standardized SNP effect sizes to $\beta_{kj}^{(J)} = \{q_{kj}(1-q_{kj})\}^{\alpha}\nu_{kj}\sqrt{H_k^2/\sum_{j=1}^{m_k}[\{q_{kj}(1-q_{kj})\}^{\alpha}\nu_{kj}]^2}$. Two heritability settings were considered: (1) a constant common SNP heritability 0.4 across all ancestries, and (2) a total heritability of 0.4 across all 19.2 million SNPs with a constant per-SNP heritability across ancestries, which leads to a common SNP heritability proportional to the number of common SNPs in the corresponding ancestry.

We simulated 120,000 individuals for each ancestry. For EUR, N_{GWAS} = 100,000 individuals were included in the discovery GWAS, while the remaining 20,000 individuals were evenly split into a tuning set for parameter tuning and a testing set to report prediction R^2 of the methods. For each non-EUR ancestry, N_{GWAS} individuals were included in the discovery GWAS, while two separate sets, each including 10,000 individuals, were selected randomly from the remaining (120,000 – N_{GWAS}) individuals to construct tuning and testing dataset. Although currently the non-EUR GWAS sample sizes are typically a lot smaller than EUR GWAS sample sizes, they are expected to continue growing, as there is an increasing emphasis on health equity. To mimic such real-world scenarios, we set non-EUR GWAS sample sizes to N_{GWAS} = 15,000, 45,000, 80,000, or 100,000, that gradually increase and eventually reach a similar level to the EUR GWAS sample size (100,000). For each ancestry group, the genotype data of 1000 randomly selected individuals in the discovery GWAS were used to estimate the ancestry-specific LD.

Runtimes and memory usage

We compare the computation time and memory usage of MUSSEL and PRS-CSx on chromosome 22 based on the simulated dataset (comparison between PRS-CSx and CT-SLEB on the same dataset has been reported in Zhang et al.[13]). Results from MUSSEL and PRS-CSx combining three ancestry groups (EUR, AFR, and AMR), four ancestry groups (EUR, AFR, AMR, and EAS), and five ancestry groups (EUR, AFR, AMR, EAS, and SAS) are summarized in Table S6. The training GWAS sample size is 15,000 for each non-EUR population and 100,000 for EUR population. The tuning and validation dataset each contains 10,000 individuals. All analyses were performed with AMD EPYC 7702 64-Core Processors running at 2.0 GHz. Other than the LDpred2 step which uses parallel computing with 17 cores, all other analyses were conducted using a single core. The reported computation time and memory usage are averaged over 10 replicates.

PAGE + UKBB + BBJ data analysis with validation on non-EUR individuals from PAGE

Three traits, including IRNT BMI, HDL, and LDL, that were available across PAGE, UKBB, and BBJ GWAS for EUR, AFR, AMR (Hispanic), and EAS are analyzed. Ancestry- and trait-specific GWAS sample sizes, validation sample sizes, and number of SNPs analyzed are reported in Table S8. The training GWAS datasets consist of PAGE, contributing data for AFR and AMR, UKBB, contributing data for EUR, and BBJ, contributing data for EAS. The validation datasets consist of PAGE, contributing data for the three non-EUR ancestry groups, and UKBB, contributing data for EUR. Specifically, we first collect data for a total of 43,769 PAGE individuals of AFR (N = 17,127), AMR (N = 21,995), or EAS (N = 4,647) ancestry that have data available for at least one of the three traits. For AFR and AMR that have relatively large sample sizes in PAGE, we randomly divide the samples within each ancestry group into a training dataset (80%) for conducting GWAS, a tuning dataset (10%) for tuning model parameters, and training SL in CT-SLEB and MUSSEL or the linear combination model in weighted PRS and PRS-CSx, and a testing dataset (10%) for evaluating PRS performance. For EAS which has a limited sample size in PAGE, we use all PAGE samples for external validation (tuning + testing) and obtain GWAS summary data from BBJ, which has a much larger sample size. To borrow information from large EUR GWAS, we further collect EUR GWAS summary data from UKBB[50] (N = 315,133–360,388) released by the Neale Lab. Finally, to tune the causal SNP proportion for EUR, which is required for specifying the prior causal probabilities for non-EUR ancestry groups, we further randomly select a sample of 20,000 random individuals from UKBB that do not overlap with samples in the EUR UKBB GWAS. Here the ancestry information for individuals from PAGE and UKBB is determined based on self-identified race/ethnicity.

For AFR and AMR, we conduct GWAS on individuals from the PAGE study to obtain the GWAS summary data. Specifically, we first collect a total of 17,127 AFR and 21,995 AMR from PAGE, then randomly divide the samples in each ancestry into a training set (80%) to conduct GWAS and a validation set (20%), of which 10% is used for selecting tuning parameters and training SL (tuning set), and the other 10% is used for reporting PRS performance (testing set). There was no significant difference between training and validation datasets in the distribution of the covariates adjusted for in GWAS. **PAGE GWAS**: (**1**) **IRNT BMI**. For ancestry-specific GWAS analysis on AFR and AMR, measurements of BMI outside of 6 standard deviations from the mean (based on sex and race) were removed. We first created sex-specific residuals for BMI adjusted for age, then inverse normally transformed these residuals. These inverse-normally-transformed residuals were then used in the final analysis where they were further adjusted for self-identified race/ethnicity, study, study center (for MEC and SOL only), and the top 10 genetic principal components (PCs). (**2**) **HDL**. For ancestry-specific GWAS analysis on AFR and AMR, untransformed HDL measurements were reported in mg/dL, and were adjusted for each individual's medication use by adding a constant based on the type of medication used. Details of the adjustment are described in the Supplementary Information in Wojcik et al.[3] Finally, models were adjusted by age at lipid measurement, sex, study, study center (for MEC and SOL only), self-identified race/ethnicity, and top 10 genetic PCs. (**3**) **LDL**. For ancestry-specific GWAS analysis on AFR and AMR, untransformed HDL measurements were calculated using the Friedewald Equation[51] and reported in mg/dL. The measurements were adjusted for individuals' medication use by adding a constant based on the type of medication used. Details of the calculation and adjustment are described in the Supplementary Information in Wojcik et al.[3] Participants who were pregnant at blood draw or had

fasted less than 8 h prior to lipid blood draw were excluded. Finally, models were adjusted by age at lipid measurement, sex, study, study center (for MEC and SOL only), self-identified race/ethnicity, and top 10 genetic PCs.

The PAGE individuals included in our analyses are part of the PAGE participant cohort, which were collected from Hispanic Community Health Study/Study of Latinos (HCHS/SOL), Women's Health Initiative (WHI), Multiethnic Cohort (MEC), and the Icahn School of Medicine at Mount Sinai BioMe biobank in New York City (BioMe).[3] Due to the extensive degree of admixture within and between PAGE self-identified racial/ethnic groups, individuals were not reassigned based on their genetic ancestry but remained categorized by their self-identified race/ethnicity. However, we have assigned them to ancestry groupings based on an approximation of mappings to continental-level regions for consistency with other external studies in this manuscript. Written informed consent was obtained for all participants in this study at the relevant recruitment sites. Due to the extensive degree of admixture within and between PAGE self-identified racial/ethnic groups, individuals were not reassigned based on their genetic ancestry but remained categorized by their self-identified race/ethnicity. Detailed information about genotyping, data quality control and imputation, selection of unrelated individuals, genetic principal component analysis, and phenotype harmonization are provided in the Supplementary Information in Wojcik et al.[3]

Since PAGE has a limited sample size for EAS (4,647), and thus we further collect publicly available GWAS summary data from BBJ (data availability) and use all PAGE individuals for validation on EAS. For BMI, the GWAS analysis included age, age,[2] sex, status of a series of diseases, and the top 10 genetic PCs as covariates.[40] For HDL and LDL, the GWAS analyses included age, sex, status of a series of diseases, and the top 10 genetic PCs as covariates.[39]

PAGE does not have individuals of EUR ancestry. To borrow information from the much larger EUR GWAS, we further download publicly available EUR GWAS summary data from UKBB (Data and code availability). For all three traits, the UKBB GWAS analyses include age, age,[2] inferred sex, an interaction term between age and inferred sex, an interaction term between age[2] and inferred sex, and the top 20 genetic PCs as covariates. One thing to note is that for HDL and LDL, measurements are untransformed and reported in mmol/L in UKBB, untransformed and reported in mg/dL in PAGE, and reported in mg/dL then standardized to Z score in BBJ. Although not on the same scale, the correlation in SNP effect size estimates remain the same, allowing the various GWAS summary data to be analyzed jointly. For EUR, we construct a validation dataset of 20,000 independent samples from UKBB that do not overlap with the UKBB GWAS samples. Specifically, we use the genotyping plate and well codes, which are published in the file ukb_sqc_v2.txt by UKBB and are consistent across different project applications, to identify and exclude the individuals included in the UKBB GWAS analysis by Neale Lab, and then randomly select 20,000 independent individuals from the remaining UKBB samples to conduct parameter tuning (10,000) and testing (10,000). For each ancestry group, we use unrelated samples of the same ancestry from 1000 Genomes Project as the LD reference data. For EUR, the reported prediction R^2 are adjusted for age, sex, and top 10 genetic PCs. For AFR, AMR and EAS, the R^2 for BMI are adjusted for age, sex, top 10 genetic PCs, and whether the individual is from the BioMe Biobank, and for HDL and LDL the R^2 are adjusted for age at lipid measurement, sex, top 10 genetic PCs, and whether the individual is from the BioMe Biobank.

We conduct the following quality control steps for the GWAS summary-level association statistics: (1) consistent with the procedure in our simulation study and other data analyses, we restrict our analysis to approximately 1.6 million SNPs in HapMap 3 plus MEGA that are also available in LD reference panel and validation sample; (2) we remove SNPs that have duplicated positions in GWAS or LD reference panel; (3) for EUR, we remove SNPs that have alleles "AT", "TA", "CG", or "GC" to avoid undetectable flipping strands when matching with UKBB validation data; (4) for the implementation of single-ancestry methods, we only keep common SNPs, i.e., SNPs that have ancestry-specific MAF > 0.01 in that ancestry group, and for the implementation of multi-ancestry methods we keep all SNPs that have ancestry-specific MAF > 0.01 in at least one ancestry group. The Manhattan plots and QQ plots for GWAS are reported in Figures S17–S19. No inflation is observed based on the genomic inflation factor. We estimate heritability of the three traits for EUR using LDSC[43] based on the 1000 Genomes LD reference data for EUR (Table S16).

GLGC data analysis with validation on UKBB individuals

We obtain GWAS summary data from the Global Lipids Genetics Consortium (GLGC) for four blood lipid traits including HDL, LDL, TC, and logTG[25] on five ancestry groups including EUR (N_{GWAS} = 840,018–927,975), AFR or admixed AFR (N_{GWAS} = 87,759–92,554), Hispanic (N_{GWAS} = 33,989–48,056), EAS (N_{GWAS} = 80,676–145,512), and SAS (N_{GWAS} = 33,658–34,135). Details of the study design, genotyping, quality control and GWAS are previously described.[25] We validate performance of the various methods on UKBB individuals. Specifically, we select a random set of 20,000 individuals that are of EUR origin and extracted all individuals that are of AFR (N = 9,169), EAS (N = 2,019), SAS (N = 10,853), or Hispanic/Latino (N = 785) origin. The origin of the UKBB individuals were determined by a genetic component analysis (Supplemental Information). We used 50% of the UKBB samples to tune model parameters and train the SL in CT-SLEB and MUSSEL or the linear combination model in weighted PRS and PRS-CS (tuning set), and the remaining 50% to evaluate PRS performance (testing set). The prediction of the genetic component has a low accuracy for AMR, and given the small number of identified AMR individuals (N = 785), we do not report prediction R^2 on UKBB AMR. We use genotype data of unrelated individuals from 1000 Genomes project or tuning samples from UKBB as the LD reference data.[24] Ancestry- and trait-specific GWAS sample sizes, validation sample sizes, and number of SNPs analyzed are reported in Table S10. Based on the genomic inflation factor, no inflation is observed for the various ancestry-specific GWAS. The Manhattan plots and QQ plots are reported in Zhang et al.[13] No inflation is observed given the genomic inflation factor. Heritability of the four traits in EUR is estimated using LDSC (Table S16). All GWAS summary statistics went through the same quality control steps as in PAGE + UKBB + BBJ data analysis

as well as one more step, where we further remove SNPs with a GWAS sample size less than 90% of the total GWAS sample size. The GWAS summary data from GLGC does not have information on ancestry-specific MAF, and thus we use the 1000 Genomes LD reference genotype data to calculate ancestry-specific MAF for the step where we filter out all SNPs that have MAF < 0.01 in all ancestry groups. The R^2 are adjusted for age, sex, and top 10 genetic PCs.

AoU data analysis with validation on UKBB individuals

The individuals included in our analyses are part of the All of Us participant cohort with information collected according to the All of Us Research Program Operational Protocol (https://allofus.nih.gov/sites/default/files/aou_operational_protocol_v1.7_mar_2018.pdf).

Detailed information on genotyping, ancestry determination, quality control, removal of related individuals is provided in the All Of Us Research Program Genomic Research Data Quality Report (https://www.researchallofus.org/wp-content/themes/research-hub-wordpress-theme/media/2022/06/All%20Of%20Us%20Q2%202022%20Release%20Genomic%20Quality%20Report.pdf).

On the All of Us platform, we conduct GWAS for BMI and height separately on unrelated individuals of three ancestry groups including EUR (N_{GWAS} = 48,229–48,332), admixed AFR or AFR (N_{GWAS} = 21,514–21,550), and Hispanic/Latino (N_{GWAS} = 15,364–15,413). The GWAS are adjusted for age, sex, and top 16 genetic PCs. There are only about 0.9 million SNPs in HapMap 3 + MEGA that are included in our analyses, which is due to the small number of the overlapping samples across the filtered WGS data, array data, and phenotype data. Similar to the GLGC data analysis, we validate performance of the various methods on UKBB individuals, i.e., 20,000 EUR individuals and individuals of AFR (N = 9,169) origin that are identified based on a genetic component analysis (Supplemental Information). Again, the genetic ancestry prediction accuracy for AMR is low, and considering the small number of identified AMR (N = 785), we do not report validation results on UKBB AMR. We use genotype data of unrelated individuals from 1000 Genomes project or tuning samples from UKBB as the LD reference data. Ancestry- and trait-specific GWAS sample sizes, validation sample sizes, and number of SNPs analyzed are reported in Table S12. Based on the genomic inflation factor, no inflation is observed other than height for Hispanic/Latino. The Manhattan plots and QQ plots are reported in Zhang et al. (2022).[13] Heritability of the two traits in EUR was estimated using LDSC[43] (Table S16). All GWAS summary statistics went through the same quality control steps as in the GLGC data analysis. The R^2 are adjusted for age, sex, and top 10 genetic PCs.

Predicted genetic ancestry for non-EUR individuals in UKBB

We compute genetic ancestry for all UKBB individuals that are not self-reported Whites. To balance between samples of different ancestry groups, we also include 8,000 unrelated self-reported Whites to form the set of UKBB individuals for genetic ancestry prediction. We use 2,504 unrelated individuals from 1000 Genomes Project, including 498 EUR, 659 AFR, 347 AMR, 503 EAS, and 487 SAS individuals to form the reference data for genetic ancestry prediction. We first compute the top 20 genetic principal components (PCs) for all UKBB and 1000 Genomes individuals together using PLINK 2.0 command –pca 20 allele-wts.[52] We then train a random forest classifier with 1,500 trees using the R package "randomForest"[53] based on the genetic PCs of the 1000 Genomes individuals with their true labels being provided by gnomAD[54] that can be used to capture enough ancestral information. Finally, we apply the trained random forest classifier to predict the genetic ancestry of UKBB individuals based on their genetic PCs.

23andMe data analysis

We develop and validate PRS for seven traits, including (1) heart metabolic disease burden, (2) height, (3) any cardiovascular disease (any CVD), (4) depression, (5) migraine diagnosis, (6) morning person, and (7) sing back musical note (SBMN) for EUR, African American (AFR), Latino (AMR), EAS, and SAS based on a large-scale dataset from 23andMe, Inc. We first conduct GWAS separately on the training dataset (70% samples) for each of the five ancestry groups, then apply the various methods to the generated GWAS summary-level association statistics and LD reference data from the 1000 Genomes Project. Within the remaining 30% of the samples, we use 20% to tune model parameters, train the SL in CT-SLEB and MUSSEL, and the linear combination model in weighted PRS and PRS-CSx, then validate the predictive performance of the constructed PRS on the remaining 10% samples. We observe from our analyses on the other three datasets that MUSSEL almost always outperforms the two alternative methods, MUSS and weighted MUSS, and thus for 23andMe data analysis, we only implement MUSSEL but not the two alternative methods.

All GWAS analyses on the training data from 23andMe, Inc. are performed adjusting for age, sex, and the top 5 genetic PCs. Genotype data of unrelated individuals from 1000 Genomes project is used to estimate LD matrices. Detailed information on participant inclusion, genotyping, phenotyping, data imputation and quality control, removing related individuals, ancestry determination, and GWAS analysis is provided in Zhang et al.[13] Ancestry- and trait-specific GWAS sample sizes, validation (tuning + testing) sample sizes, and the number of SNPs analyzed are reported in Table S14. Based on the genomic inflation factor, no inflation is observed for the various ancestry-specific GWAS. The Manhattan plots and QQ plots are reported in Zhang et al.[13] No inflation is observed given the genomic inflation factor. Heritability of the four traits in EUR is estimated using LDSC.[13] All GWAS summary statistics went through the same quality control steps as in PAGE + UKBB + BBJ data analysis as well as one more step where we further remove SNPs with a GWAS sample size less than 90% of the total GWAS sample size. The residual R^2 for the two continuous traits are calculated by first regressing each trait on covariates including age, sex, and the top 5 genetic PCs, and then calculating the proportion of variation of the residual explained by the PRS. The residual AUC for the five binary traits were calculated using the "roc.binary" function in the R package RISCA version 1.0171 adjusting for the same set of covariates adjusted for the continuous traits.

Calculation of the 95% bootstrap confidence intervals and p values for comparing R^2 between methods

We used bootstrapping to calculate the 95% CIs for the R^2 reported in our simulation study, the PAGE + UKBB + BBJ, GLGC, and All of Us (AoU) data analyses. We employed the R package ''boot'' with 10,000 sampling replicates on the testing dataset and used the Bca approach[22] to obtain the CIs. The R^2s and the corresponding 95% bootstrap CIs are summarized in Tables S1, S2, S3, S4, and S5 and Figures S1–S10 for simulated data, and Figures 2, 3, 4, 5, Table S17, and Figures S13–S15 for PAGE+UKBB+BBJ, GLGC, and All of Us data analyses. The 95% bootstrap CIs were not calculated in the 23andMe Inc. data analysis due to data agreement restrictions with 23andMe Inc. To compare the overall performance between different methods, we further calculate the average R^2 and the corresponding 95% bootstrap CIs across all available traits in the PAGE + UKBB + BBJ, GLGC, and AoU data analyses for each method on each ancestry group (Figure S16; Table S17). We observe that the empirical bootstrap distribution of the average R^2 values are approximately normal across all methods and all ancestry groups. We therefore calculated the p values of the paired two-sided test for the equality of average R^2 between each pair of methods based on the corresponding 95% bootstrap CI.

Cell Genomics, Volume *4*

Supplemental information

MUSSEL: Enhanced Bayesian polygenic risk prediction leveraging information across multiple ancestry groups

Jin Jin, Jianan Zhan, Jingning Zhang, Ruzhang Zhao, Jared O'Connell, Yunxuan Jiang, 23andMe Research Team, Steven Buyske, Christopher Gignoux, Christopher Haiman, Eimear E. Kenny, Charles Kooperberg, Kari North, Bertram L. Koelsch, Genevieve Wojcik, Haoyu Zhang, and Nilanjan Chatterjee

Supplemental Figures

for

MUSSEL: Enhanced Bayesian Polygenic Risk Prediction Leveraging Information across Multiple Ancestry Groups

This document includes:

Supplementary Figures S1-S19

Supplementary Figure Legends

Supplemental References

Figure S1: Simulation results showing performance of the PRS constructed by MUSSEL and various existing methods, assuming a fixed common SNP heritability (0.4) across ancestries under a strong negative selection model for the relationship between SNP effect size and allele frequency with a GWAS sample size of 15,000/45,000 for each non-EUR population, related to Figure 2. The genetic correlation in SNP effect size is set to 0.8 across all pairs of populations. The causal SNP proportion (degree of polygenicity) is set to 1.0%, 0.1%, or 0.05% (~$192K$, $19.2K$, or $9.6K$ causal SNPs). We generate data for ~19 million common SNPs (MAF≥1%) across the five ancestry groups but conduct analyses only on the ~2.0 million SNPs in HapMap 3 + MEGA. The discovery GWAS sample size is set to **(a) 15,000 or (b) 45,000** for each non-EUR ancestry, and 100,000 for EUR. A tuning set consisting of 10,000 individuals is used for parameter tuning, as well as training the SL in CT-SLEB and MUSSEL or the linear combination model in weighted C+T, weighted LDpred2, PRS-CSx, and weighted MUSS. The reported R^2 values are calculated on an independent testing set of 10,000 individuals for each ancestry group. The corresponding 95% bootstrap CIs are obtained from the same testing set based on 10,000 bootstrap samples using the Bca approach[1] implemented in the R package "boot".

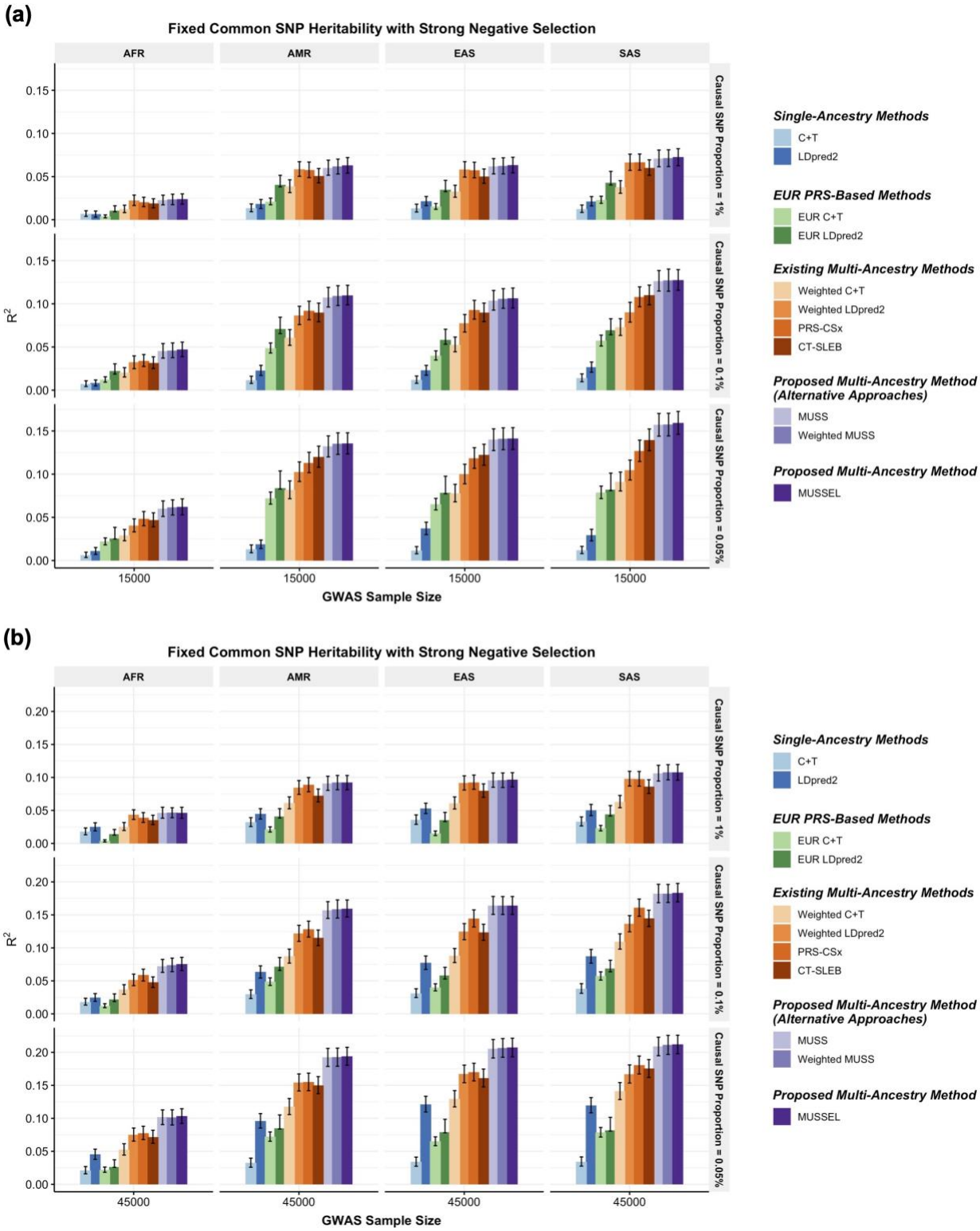

(a)
Fixed Common SNP Heritability with Strong Negative Selection
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.00
0.05
0.10
0.15
15000
GWAS Sample Size
Single-Ancestry Methods
C+T
LDpred2
EUR PRS-Based Methods
EUR C+T
EUR LDpred2
Existing Multi-Ancestry Methods
Weighted C+T
Weighted LDpred2
PRS-CSx
CT-SLEB
Proposed Multi-Ancestry Method (Alternative Approaches)
MUSS
Weighted MUSS
Proposed Multi-Ancestry Method
MUSSEL
(b)
Fixed Common SNP Heritability with Strong Negative Selection
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.00
0.05
0.10
0.15
0.20
45000
GWAS Sample Size
Single-Ancestry Methods
C+T
LDpred2
EUR PRS-Based Methods
EUR C+T
EUR LDpred2
Existing Multi-Ancestry Methods
Weighted C+T
Weighted LDpred2
PRS-CSx
CT-SLEB
Proposed Multi-Ancestry Method (Alternative Approaches)
MUSS
Weighted MUSS
Proposed Multi-Ancestry Method
MUSSEL

Figure S2: Simulation results showing performance of the PRS constructed by MUSSEL and various existing methods, assuming a fixed common SNP heritability (0.4) across ancestries under a strong negative selection model for the relationship between SNP effect size and allele frequency with a GWAS sample size of 80,000/100,000 for each non-EUR population, related to Figure 2. The genetic correlation in SNP effect size is set to 0.8 across all pairs of populations. The causal SNP proportion (degree of polygenicity) is set to 1.0%, 0.1%, or 0.05% (~$192K$, $19.2K$, or $9.6K$ causal SNPs). We generate data for ~19 million common SNPs (MAF≥1%) across the five ancestries but conduct analyses only on the ~2.0 million SNPs in HapMap 3 + MEGA. The discovery GWAS sample size is set to **(a) 80,000 or (b) 100,000** for each non-EUR ancestry, and 100,000 for EUR. A tuning set consisting of 10,000 individuals is used for parameter tuning, as well as training the SL in CT-SLEB and MUSSEL or the linear combination model in weighted C+T, weighted LDpred2, PRS-CSx, and weighted MUSS. The reported R^2 values are calculated on an independent testing set of 10,000 individuals for each ancestry group. The corresponding 95% bootstrap CIs are obtained from the same testing set based on 10,000 bootstrap samples using the Bca approach[1] implemented in the R package "boot".

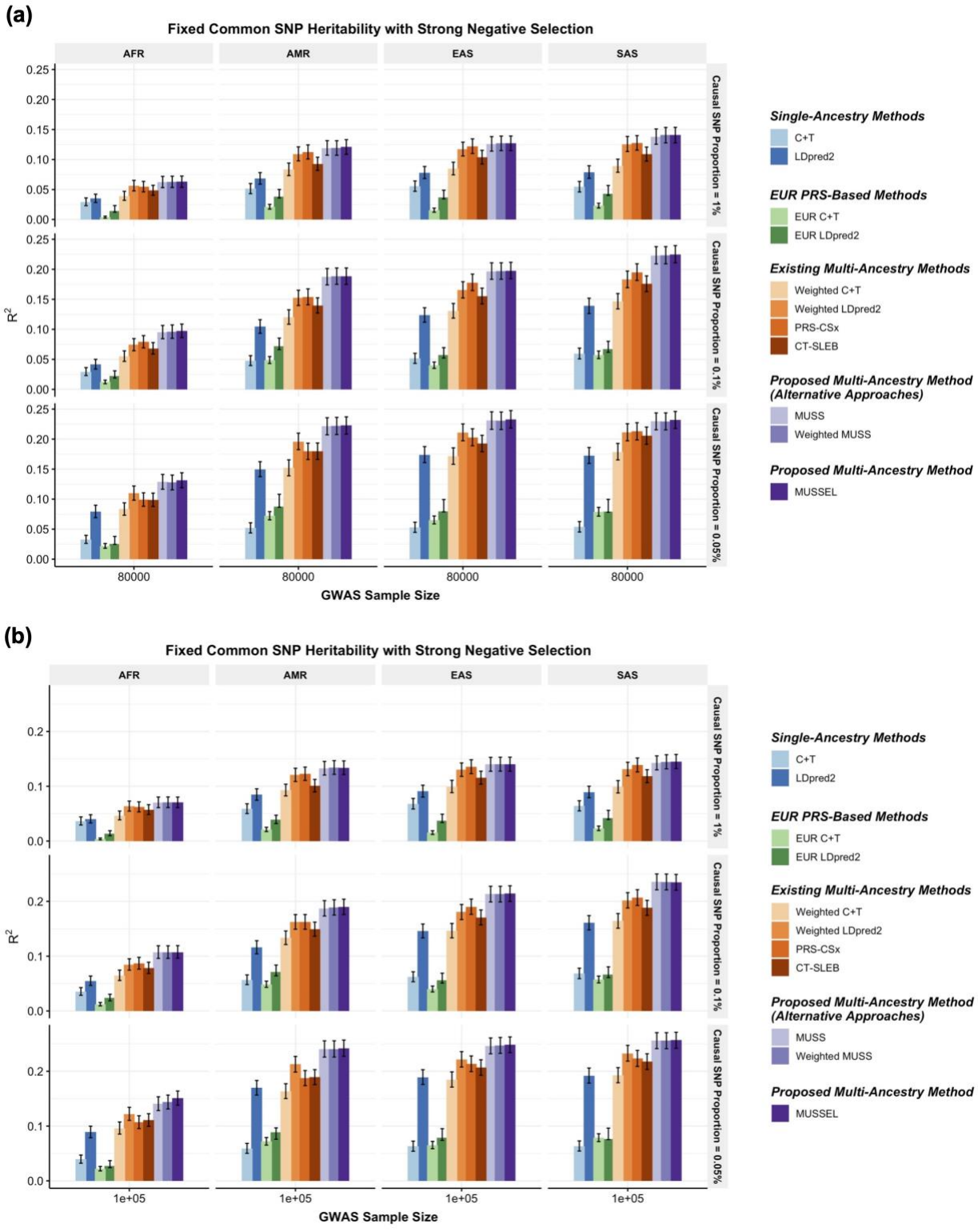

(a)
Fixed Common SNP Heritability with Strong Negative Selection
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.25
0.20
0.15
0.10
0.05
0.00
80000
GWAS Sample Size
Single-Ancestry Methods
C+T
LDpred2
EUR PRS-Based Methods
EUR C+T
EUR LDpred2
Existing Multi-Ancestry Methods
Weighted C+T
Weighted LDpred2
PRS-CSx
CT-SLEB
Proposed Multi-Ancestry Method (Alternative Approaches)
MUSS
Weighted MUSS
Proposed Multi-Ancestry Method
MUSSEL
(b)
Fixed Common SNP Heritability with Strong Negative Selection
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.2
0.1
0.0
1e+05
GWAS Sample Size
Single-Ancestry Methods
C+T
LDpred2
EUR PRS-Based Methods
EUR C+T
EUR LDpred2
Existing Multi-Ancestry Methods
Weighted C+T
Weighted LDpred2
PRS-CSx
CT-SLEB
Proposed Multi-Ancestry Method (Alternative Approaches)
MUSS
Weighted MUSS
Proposed Multi-Ancestry Method
MUSSEL

Figure S3: Simulation results showing performance of the PRS constructed by MUSSEL and various existing methods, assuming a fixed per-SNP heritability (0.4) across ancestries under a strong negative selection model for the relationship between SNP effect size and allele frequency with a GWAS sample size of 15,000/45,000 for each non-EUR population, related to Figure 2. The genetic correlation in SNP effect size is set to 0.8 across all pairs of populations. The causal SNP proportion (degree of polygenicity) is set to 1.0%, 0.1%, or 0.05% (~$192K$, $19.2K$, or $9.6K$ causal SNPs). We generate data for ~19 million common SNPs (MAF≥1%) across the five ancestries but conduct analyses only on the ~2.0 million SNPs in HapMap 3 + MEGA. The discovery GWAS sample size is set to **(a) 15,000 or (b) 45,000** for each non-EUR ancestry, and 100,000 for EUR. A tuning set consisting of 10,000 individuals is used for parameter tuning, as well as training the SL in CT-SLEB and MUSSEL or the linear combination model in weighted C+T, weighted LDpred2, PRS-CSx, and weighted MUSS. The reported R^2 values are calculated on an independent testing set of 10,000 individuals for each ancestry group. The corresponding 95% bootstrap CIs are obtained from the same testing set based on 10,000 bootstrap samples using the Bca approach[1] implemented in the R package "boot".

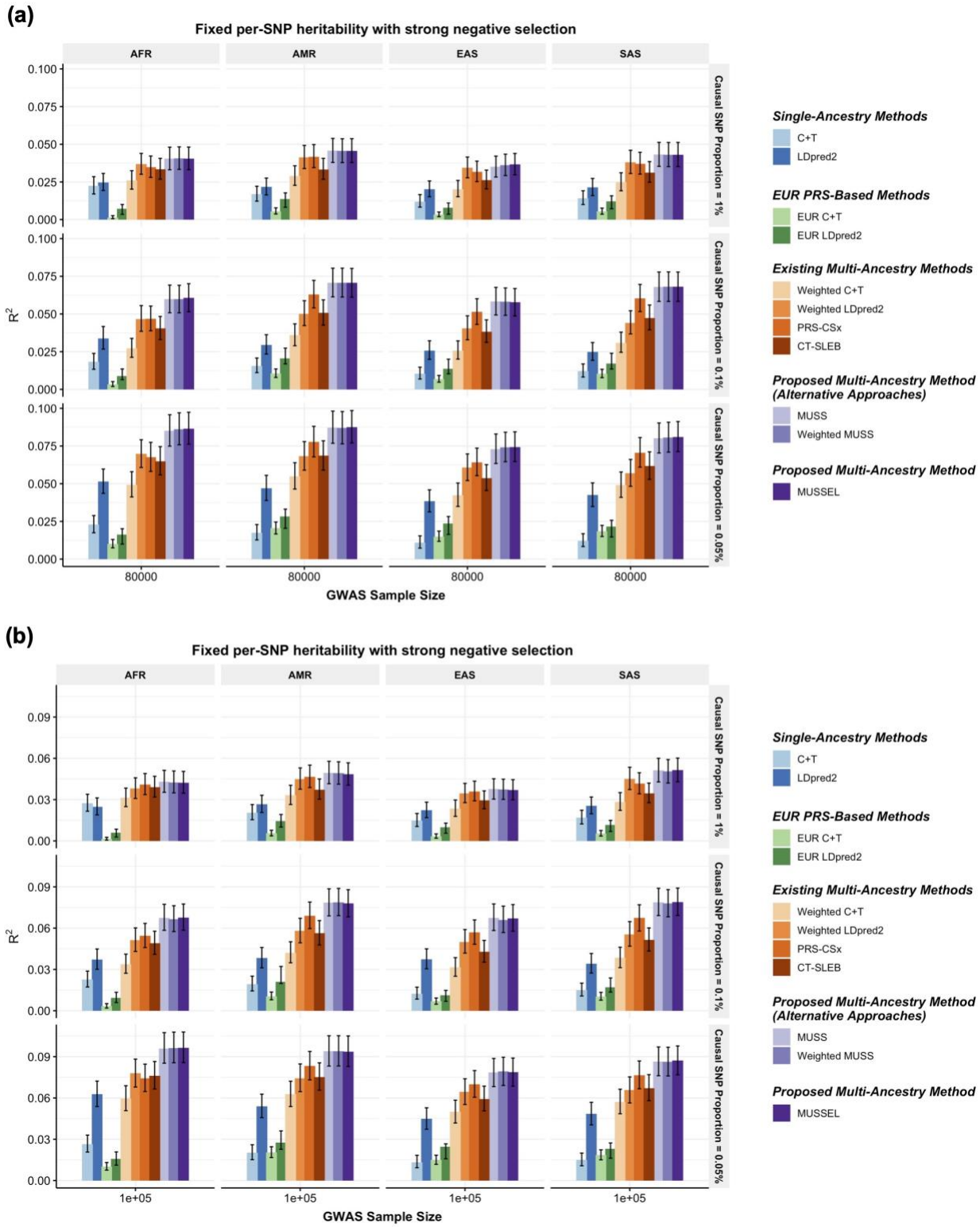

(a)
Fixed per-SNP heritability with strong negative selection
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.100
0.075
0.050
0.025
0.000
80000
GWAS Sample Size
(b)
Fixed per-SNP heritability with strong negative selection
0.09
0.06
0.03
0.00
1e+05
Single-Ancestry Methods
C+T
LDpred2
EUR PRS-Based Methods
EUR C+T
EUR LDpred2
Existing Multi-Ancestry Methods
Weighted C+T
Weighted LDpred2
PRS-CSx
CT-SLEB
Proposed Multi-Ancestry Method (Alternative Approaches)
MUSS
Weighted MUSS
Proposed Multi-Ancestry Method
MUSSEL

Figure S4: Simulation results showing performance of the PRS constructed by MUSSEL and various existing methods, assuming a fixed per-SNP heritability (0.4) across ancestries under a strong negative selection model for the relationship between SNP effect size and allele frequency with a GWAS sample size of 80,000/100,000 for each non-EUR population, related to Figure 2. The genetic correlation in SNP effect size is set to 0.8 across all pairs of populations. The causal SNP proportion (degree of polygenicity) is set to 1.0%, 0.1%, or 0.05% (~$192K$, $19.2K$, or $9.6K$ causal SNPs). We generate data for ~19 million common SNPs (MAF≥1%) across the five ancestries but conduct analyses only on the ~2.0 million SNPs in HapMap 3 + MEGA. The discovery GWAS sample size is set to **(a) 80,000 or (b) 100,000** for each non-EUR ancestry, and 100,000 for EUR. A tuning set consisting of 10,000 individuals is used for parameter tuning, as well as training the SL in CT-SLEB and MUSSEL or the linear combination model in weighted C+T, weighted LDpred2, PRS-CSx, and weighted MUSS. The reported R^2 values are calculated on an independent testing set of 10,000 individuals for each ancestry group. The corresponding 95% bootstrap CIs are obtained from the same testing set based on 10,000 bootstrap samples using the Bca approach[1] implemented in the R package "boot".

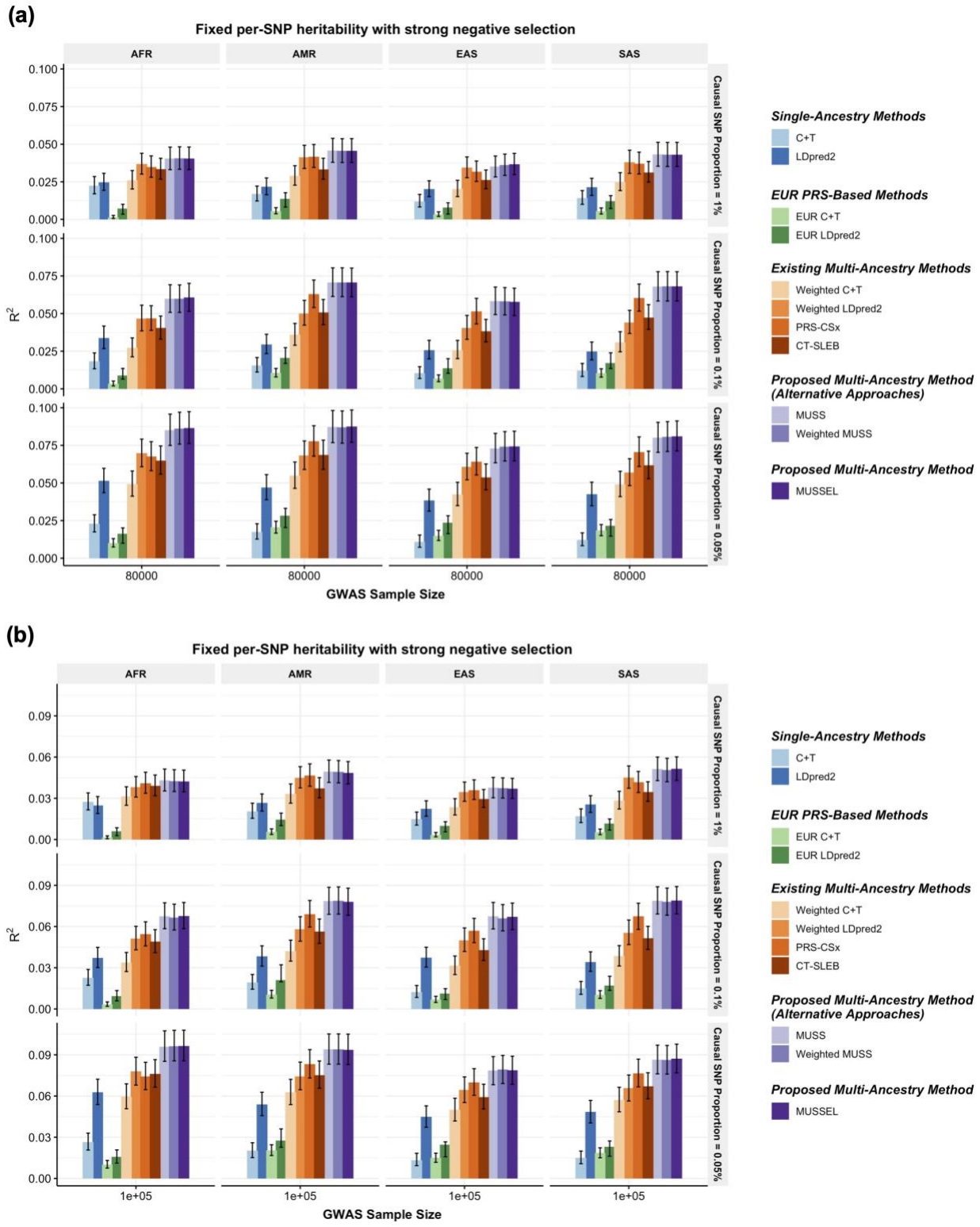

(a)
Fixed per-SNP heritability with strong negative selection
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.100
0.075
0.050
0.025
0.000
80000
GWAS Sample Size
Single-Ancestry Methods
C+T
LDpred2
EUR PRS-Based Methods
EUR C+T
EUR LDpred2
Existing Multi-Ancestry Methods
Weighted C+T
Weighted LDpred2
PRS-CSx
CT-SLEB
Proposed Multi-Ancestry Method (Alternative Approaches)
MUSS
Weighted MUSS
Proposed Multi-Ancestry Method
MUSSEL
(b)
Fixed per-SNP heritability with strong negative selection
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.09
0.06
0.03
0.00
1e+05
GWAS Sample Size
Single-Ancestry Methods
C+T
LDpred2
EUR PRS-Based Methods
EUR C+T
EUR LDpred2
Existing Multi-Ancestry Methods
Weighted C+T
Weighted LDpred2
PRS-CSx
CT-SLEB
Proposed Multi-Ancestry Method (Alternative Approaches)
MUSS
Weighted MUSS
Proposed Multi-Ancestry Method
MUSSEL

Figure S5: Simulation results showing performance of the PRS constructed by MUSSEL and various existing methods, assuming a fixed per-SNP heritability (0.4) across ancestries under a strong negative selection model for the relationship between SNP effect size and allele frequency but with weaker cross-population (0.6 across all pairs of populations), with a GWAS sample size of 15,000/45,000 for each non-EUR population, related to Figure 2. The causal SNP proportion (degree of polygenicity) is set to 1.0%, 0.1%, or 0.05% (~$192K$, $19.2K$, or $9.6K$ causal SNPs). We generate data for ~19 million common SNPs (MAF≥1%) across the five ancestries but conduct analyses only on the ~2.0 million SNPs in HapMap 3 + MEGA. The discovery GWAS sample size is set to **(a) 15,000 or (b) 45,000** for each non-EUR ancestry, and 100,000 for EUR. A tuning set consisting of 10,000 individuals is used for parameter tuning, as well as training the SL in CT-SLEB and MUSSEL or the linear combination model in weighted C+T, weighted LDpred2, PRS-CSx, and weighted MUSS. The reported R^2 values are calculated on an independent testing set of 10,000 individuals for each ancestry group. The corresponding 95% bootstrap CIs are obtained from the same testing set based on 10,000 bootstrap samples using the Bca approach[1] implemented in the R package "boot".

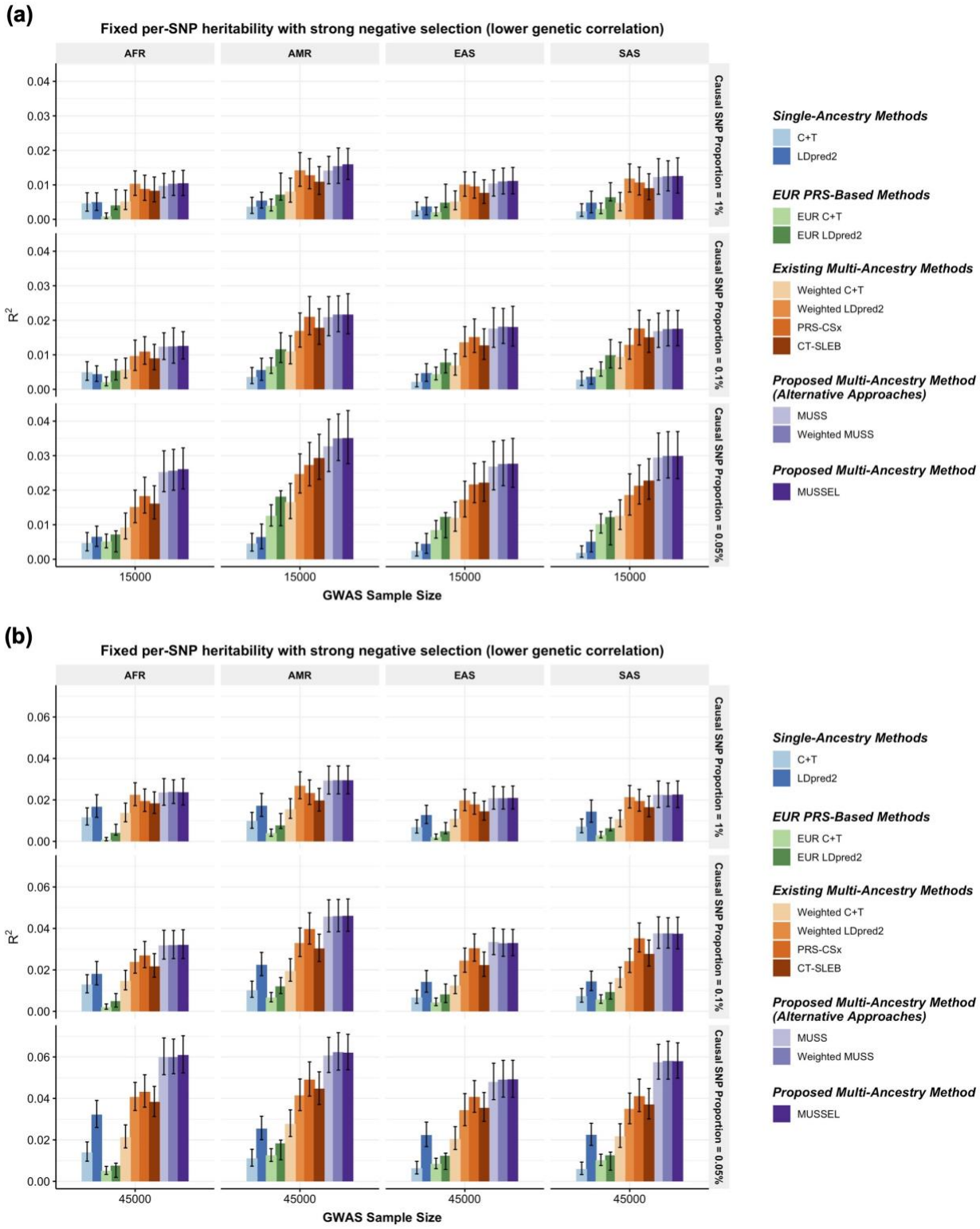

(a)
Fixed per-SNP heritability with strong negative selection (lower genetic correlation)
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R^2
0.04
0.03
0.02
0.01
0.00
15000
GWAS Sample Size
Single-Ancestry Methods
C+T
LDpred2
EUR PRS-Based Methods
EUR C+T
EUR LDpred2
Existing Multi-Ancestry Methods
Weighted C+T
Weighted LDpred2
PRS-CSx
CT-SLEB
Proposed Multi-Ancestry Method (Alternative Approaches)
MUSS
Weighted MUSS
Proposed Multi-Ancestry Method
MUSSEL
(b)
Fixed per-SNP heritability with strong negative selection (lower genetic correlation)
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R^2
0.06
0.04
0.02
0.00
45000
GWAS Sample Size
Single-Ancestry Methods
C+T
LDpred2
EUR PRS-Based Methods
EUR C+T
EUR LDpred2
Existing Multi-Ancestry Methods
Weighted C+T
Weighted LDpred2
PRS-CSx
CT-SLEB
Proposed Multi-Ancestry Method (Alternative Approaches)
MUSS
Weighted MUSS
Proposed Multi-Ancestry Method
MUSSEL

Figure S6: Simulation results showing performance of the PRS constructed by MUSSEL and various existing methods, assuming a fixed per-SNP heritability (0.4) across ancestries under a strong negative selection model for the relationship between SNP effect size and allele frequency but with weaker cross-population (0.6 across all pairs of populations), with a GWAS sample size of 80,000/100,000 for each non-EUR population, related to Figure 2. The causal SNP proportion (degree of polygenicity) is set to 1.0%, 0.1%, or 0.05% (~192K, 19.2K, or 9.6K causal SNPs). We generate data for ~19 million common SNPs (MAF≥1%) across the five ancestries but conduct analyses only on the ~2.0 million SNPs in HapMap 3 + MEGA. The discovery GWAS sample size is set to **(a) 80,000 or (b) 100,000** for each non-EUR ancestry, and 100,000 for EUR. A tuning set consisting of 10,000 individuals is used for parameter tuning, as well as training the SL in CT-SLEB and MUSSEL or the linear combination model in weighted C+T, weighted LDpred2, PRS-CSx, and weighted MUSS. The reported R^2 values are calculated on an independent testing set of 10,000 individuals for each ancestry group. The corresponding 95% bootstrap CIs are obtained from the same testing set based on 10,000 bootstrap samples using the Bca approach[1] implemented in the R package "boot".

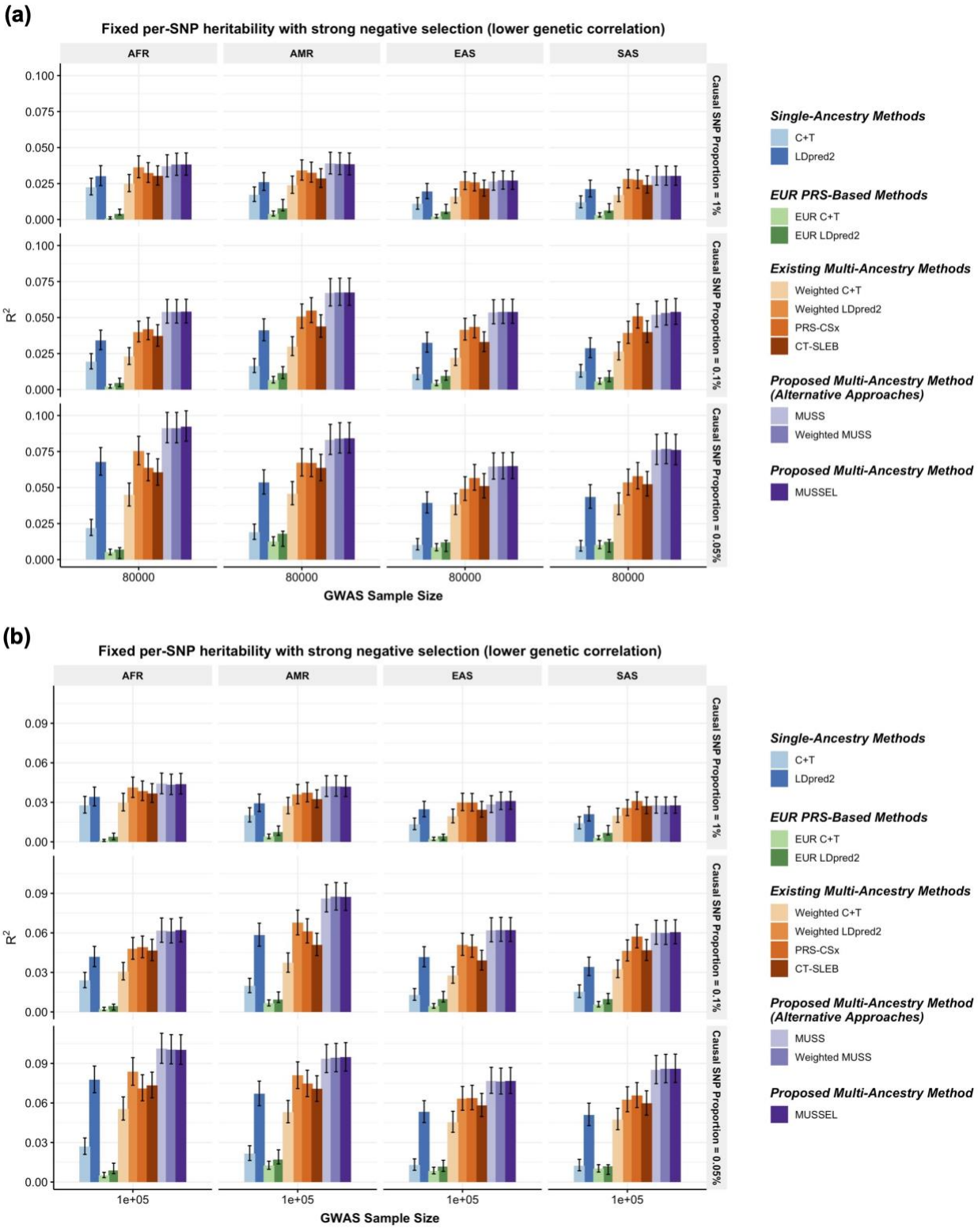

(a)
Fixed per-SNP heritability with strong negative selection (lower genetic correlation)
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.100
0.075
0.050
0.025
0.000
80000
GWAS Sample Size
Single-Ancestry Methods
C+T
LDpred2
EUR PRS-Based Methods
EUR C+T
EUR LDpred2
Existing Multi-Ancestry Methods
Weighted C+T
Weighted LDpred2
PRS-CSx
CT-SLEB
Proposed Multi-Ancestry Method (Alternative Approaches)
MUSS
Weighted MUSS
Proposed Multi-Ancestry Method
MUSSEL
(b)
Fixed per-SNP heritability with strong negative selection (lower genetic correlation)
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.09
0.06
0.03
0.00
1e+05
GWAS Sample Size
Single-Ancestry Methods
C+T
LDpred2
EUR PRS-Based Methods
EUR C+T
EUR LDpred2
Existing Multi-Ancestry Methods
Weighted C+T
Weighted LDpred2
PRS-CSx
CT-SLEB
Proposed Multi-Ancestry Method (Alternative Approaches)
MUSS
Weighted MUSS
Proposed Multi-Ancestry Method
MUSSEL

Figure S7: Simulation results showing performance of the PRS constructed by MUSSEL and various existing methods, assuming a fixed common SNP heritability (0.4) across ancestries with no negative selection for the relationship between SNP effect size and allele frequency with a GWAS sample size of 15,000/45,000 for each non-EUR population, related to Figure 2. The genetic correlation in SNP effect size is set to 0.8 across all pairs of populations. The causal SNP proportion (degree of polygenicity) is set to 1.0%, 0.1%, or 0.05% (~$192K$, $19.2K$, or $9.6K$ causal SNPs). We generate data for ~19 million common SNPs (MAF≥1%) across the five ancestries but conduct analyses only on the ~2.0 million SNPs in HapMap 3 + MEGA. The discovery GWAS sample size is set to **(a) 15,000 or (b) 45,000** for each non-EUR ancestry, and 100,000 for EUR. A tuning set consisting of 10,000 individuals is used for parameter tuning, as well as training the SL in CT-SLEB and MUSSEL or the linear combination model in weighted C+T, weighted LDpred2, PRS-CSx, and weighted MUSS. The reported R^2 values are calculated on an independent testing set of 10,000 individuals for each ancestry group. The corresponding 95% bootstrap CIs are obtained from the same testing set based on 10,000 bootstrap samples using the Bca approach[1] implemented in the R package "boot".

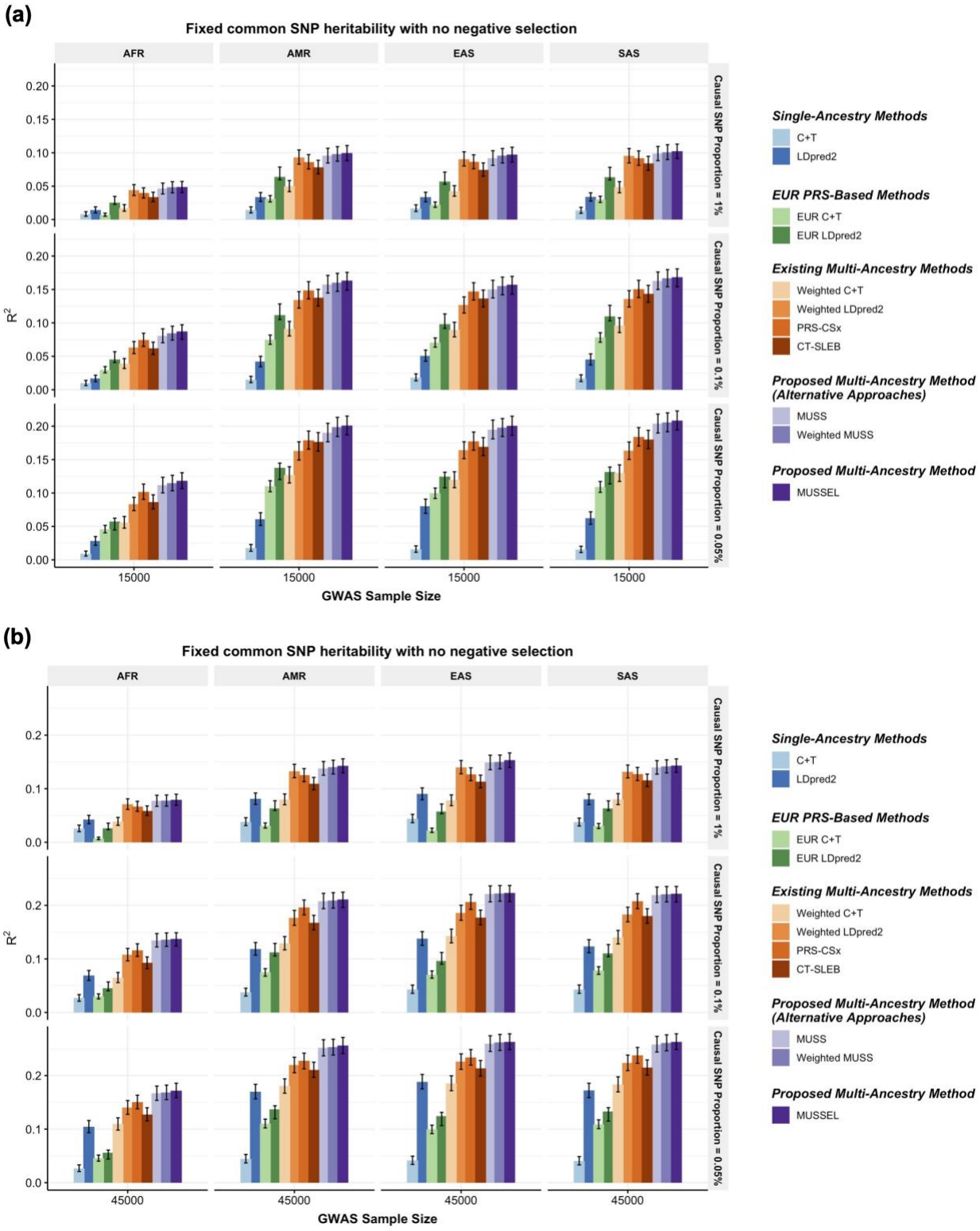

(a)
Fixed common SNP heritability with no negative selection
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.00
0.05
0.10
0.15
0.20
15000
GWAS Sample Size
Single-Ancestry Methods
C+T
LDpred2
EUR PRS-Based Methods
EUR C+T
EUR LDpred2
Existing Multi-Ancestry Methods
Weighted C+T
Weighted LDpred2
PRS-CSx
CT-SLEB
Proposed Multi-Ancestry Method (Alternative Approaches)
MUSS
Weighted MUSS
Proposed Multi-Ancestry Method
MUSSEL
(b)
Fixed common SNP heritability with no negative selection
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.0
0.1
0.2
45000
GWAS Sample Size
Single-Ancestry Methods
C+T
LDpred2
EUR PRS-Based Methods
EUR C+T
EUR LDpred2
Existing Multi-Ancestry Methods
Weighted C+T
Weighted LDpred2
PRS-CSx
CT-SLEB
Proposed Multi-Ancestry Method (Alternative Approaches)
MUSS
Weighted MUSS
Proposed Multi-Ancestry Method
MUSSEL

Figure S8: Simulation results showing performance of the PRS constructed by MUSSEL and various existing methods, assuming a fixed common SNP heritability (0.4) across ancestries with no negative selection for the relationship between SNP effect size and allele frequency with a GWAS sample size of 80,000/100,000 for each non-EUR population, related to Figure 2. The genetic correlation in SNP effect size is set to 0.8 across all pairs of populations. The causal SNP proportion (degree of polygenicity) is set to 1.0%, 0.1%, or 0.05% (~$192K$, $19.2K$, or $9.6K$ causal SNPs). We generate data for ~19 million common SNPs (MAF≥1%) across the five ancestries but conduct analyses only on the ~2.0 million SNPs in HapMap 3 + MEGA. The discovery GWAS sample size is set to **(a) 80,000 or (b) 100,000** for each non-EUR ancestry, and 100,000 for EUR. A tuning set consisting of 10,000 individuals is used for parameter tuning, as well as training the SL in CT-SLEB and MUSSEL or the linear combination model in weighted C+T, weighted LDpred2, PRS-CSx, and weighted MUSS. The reported R^2 values are calculated on an independent testing set of 10,000 individuals for each ancestry group. The corresponding 95% bootstrap CIs are obtained from the same testing set based on 10,000 bootstrap samples using the Bca approach[1] implemented in the R package "boot".

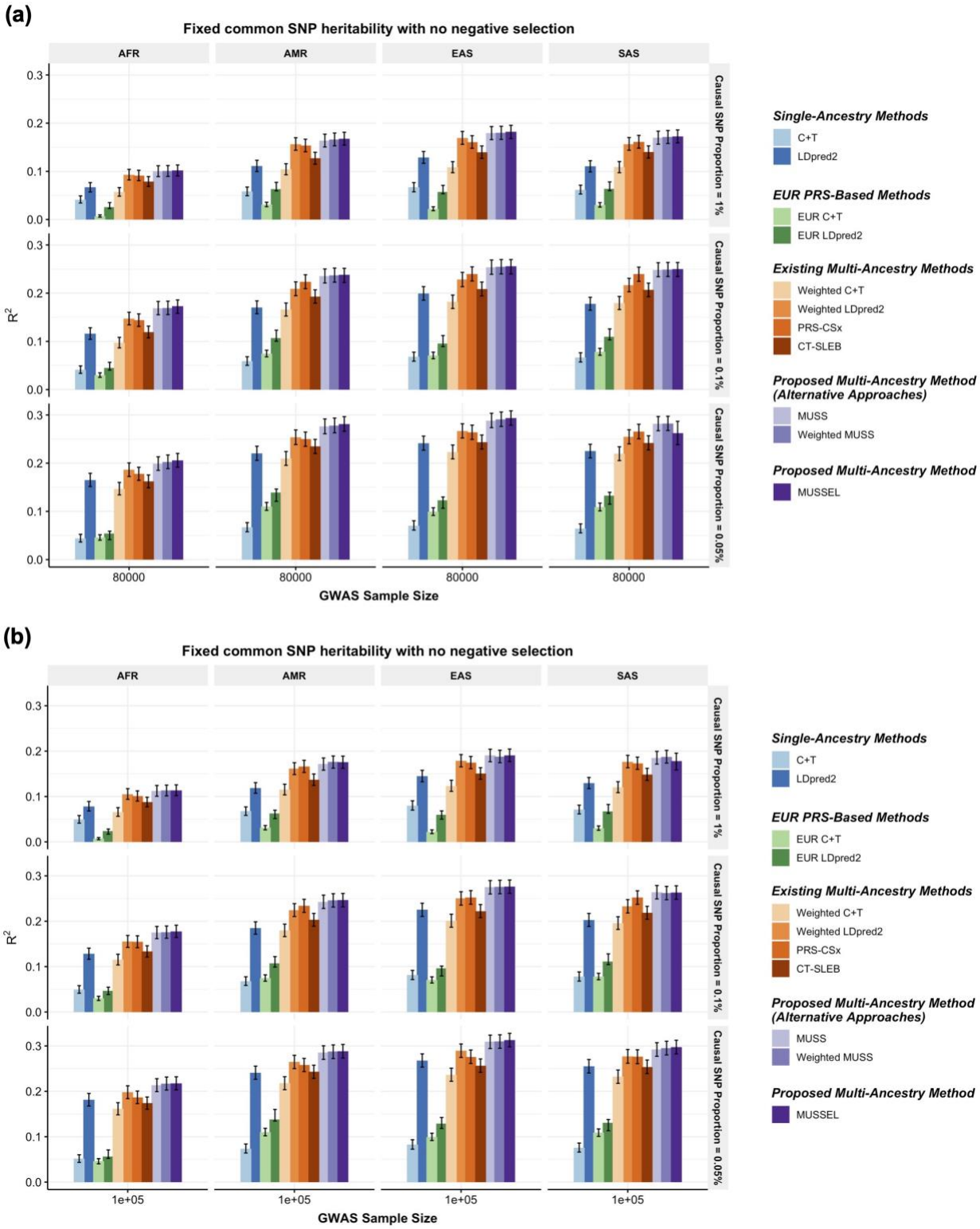

(a)
Fixed common SNP heritability with no negative selection
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.0
0.1
0.2
0.3
80000
GWAS Sample Size
Single-Ancestry Methods
C+T
LDpred2
EUR PRS-Based Methods
EUR C+T
EUR LDpred2
Existing Multi-Ancestry Methods
Weighted C+T
Weighted LDpred2
PRS-CSx
CT-SLEB
Proposed Multi-Ancestry Method (Alternative Approaches)
MUSS
Weighted MUSS
Proposed Multi-Ancestry Method
MUSSEL
(b)
Fixed common SNP heritability with no negative selection
1e+05
GWAS Sample Size

Figure S9: Simulation results showing performance of the PRS constructed by MUSSEL and various existing methods, assuming a fixed common SNP heritability (0.4) across ancestries under a mild negative selection model for the relationship between SNP effect size and allele frequency with a GWAS sample size of 15,000/45,000 for each non-EUR population, related to Figure 2. The genetic correlation in SNP effect size is set to 0.8 across all pairs of populations. The causal SNP proportion (degree of polygenicity) is set to 1.0%, 0.1%, or 0.05% (~$192K$, $19.2K$, or $9.6K$ causal SNPs). We generate data for ~19 million common SNPs (MAF≥1%) across the five ancestries but conduct analyses only on the ~2.0 million SNPs in HapMap 3 + MEGA. The discovery GWAS sample size is set to **(a) 15,000 or (b) 45,000** for each non-EUR ancestry, and 100,000 for EUR. A tuning set consisting of 10,000 individuals is used for parameter tuning, as well as training the SL in CT-SLEB and MUSSEL or the linear combination model in weighted C+T, weighted LDpred2, PRS-CSx, and weighted MUSS. The reported R^2 values are calculated on an independent testing set of 10,000 individuals for each ancestry group. The corresponding 95% bootstrap CIs are obtained from the same testing set based on 10,000 bootstrap samples using the Bca approach[1] implemented in the R package "boot".

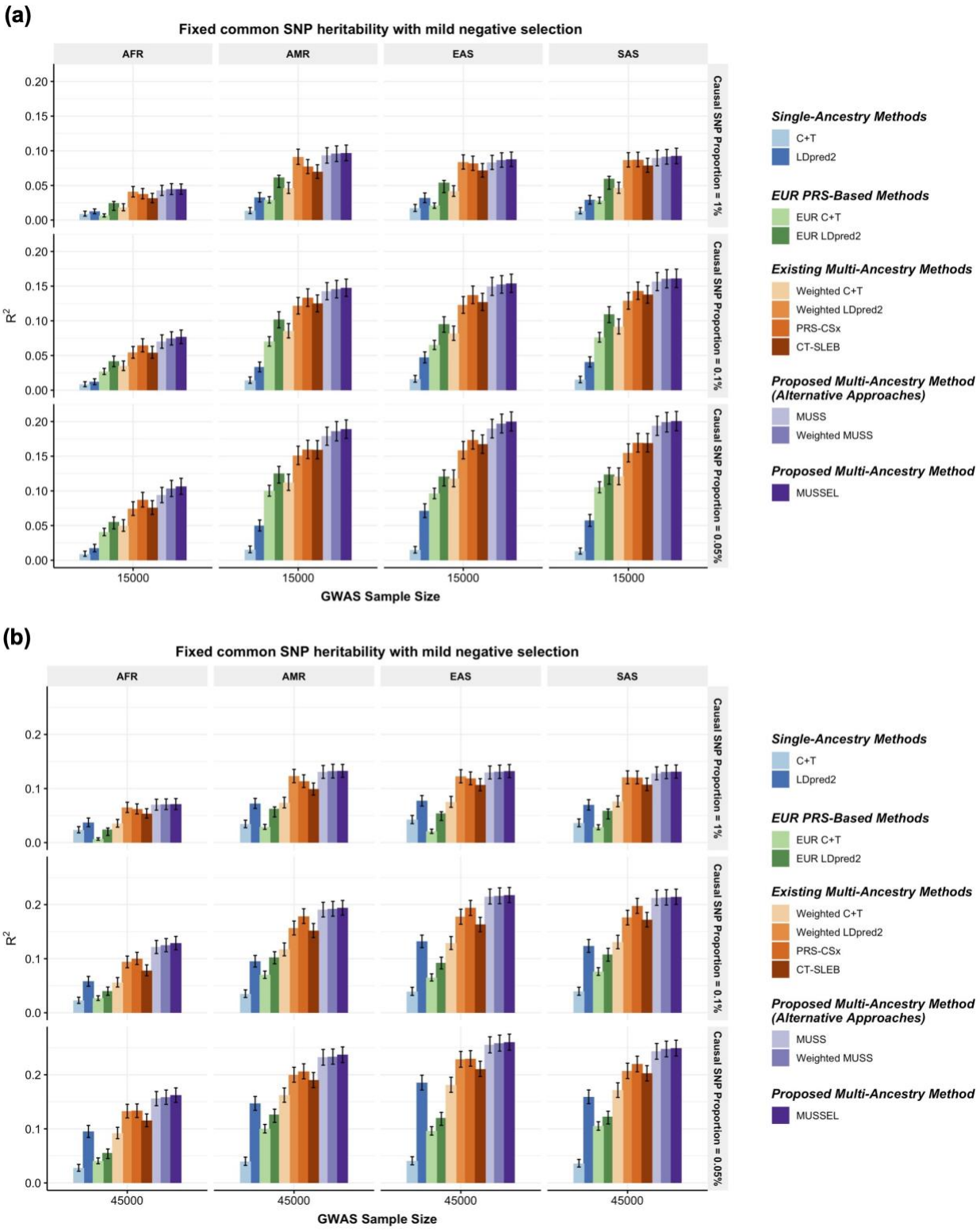

(a)
Fixed common SNP heritability with mild negative selection
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.00
0.05
0.10
0.15
0.20
15000
GWAS Sample Size
Single-Ancestry Methods
C+T
LDpred2
EUR PRS-Based Methods
EUR C+T
EUR LDpred2
Existing Multi-Ancestry Methods
Weighted C+T
Weighted LDpred2
PRS-CSx
CT-SLEB
Proposed Multi-Ancestry Method (Alternative Approaches)
MUSS
Weighted MUSS
Proposed Multi-Ancestry Method
MUSSEL
(b)
Fixed common SNP heritability with mild negative selection
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.0
0.1
0.2
45000
GWAS Sample Size
Single-Ancestry Methods
C+T
LDpred2
EUR PRS-Based Methods
EUR C+T
EUR LDpred2
Existing Multi-Ancestry Methods
Weighted C+T
Weighted LDpred2
PRS-CSx
CT-SLEB
Proposed Multi-Ancestry Method (Alternative Approaches)
MUSS
Weighted MUSS
Proposed Multi-Ancestry Method
MUSSEL

Figure S10: Simulation results showing performance of the PRS constructed by MUSSEL and various existing methods, assuming a fixed common SNP heritability (0.4) across ancestries under a mild negative selection model for the relationship between SNP effect size and allele frequency with a GWAS sample size of 80,000/100,000 for each non-EUR population, related to Figure 2. The genetic correlation in SNP effect size is set to 0.8 across all pairs of populations. The causal SNP proportion (degree of polygenicity) is set to 1.0%, 0.1%, or 0.05% (~$192K$, $19.2K$, or $9.6K$ causal SNPs). We generate data for ~19 million common SNPs (MAF≥1%) across the five ancestries but conduct analyses only on the ~2.0 million SNPs in HapMap 3 + MEGA. The discovery GWAS sample size is set to **(a) 80,000 or (b) 100,000** for each non-EUR ancestry, and 100,000 for EUR. A tuning set consisting of 10,000 individuals is used for parameter tuning, as well as training the SL in CT-SLEB and MUSSEL or the linear combination model in weighted LDpred2, PRS-CSx, and weighted MUSS. The reported R^2 values are calculated on an independent testing set of 10,000 individuals for each ancestry group. The corresponding 95% bootstrap CIs are obtained from the same testing set based on 10,000 bootstrap samples using the Bca approach[1] implemented in the R package "boot".

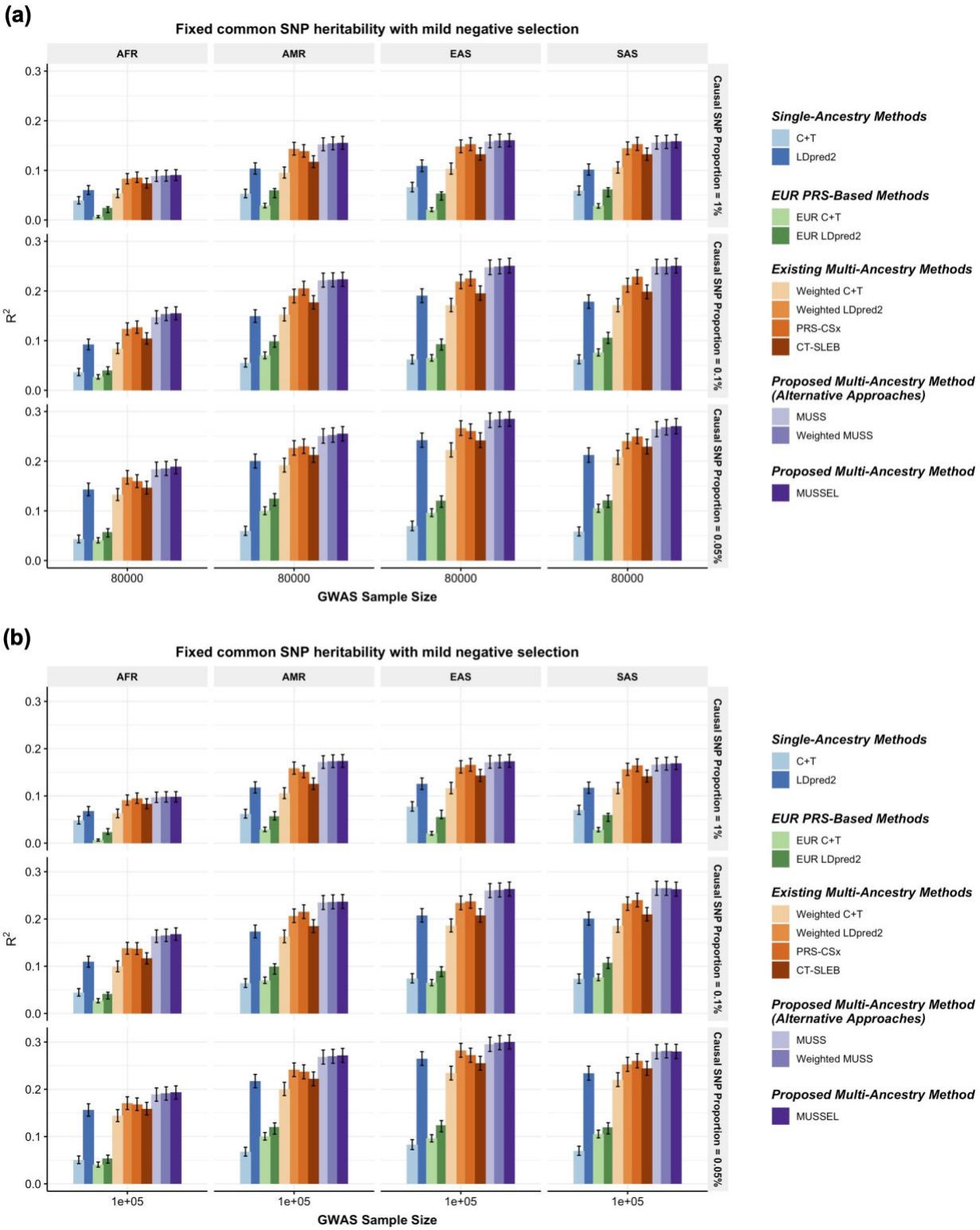

(a)
Fixed common SNP heritability with mild negative selection
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.0
0.1
0.2
0.3
80000
GWAS Sample Size
Single-Ancestry Methods
C+T
LDpred2
EUR PRS-Based Methods
EUR C+T
EUR LDpred2
Existing Multi-Ancestry Methods
Weighted C+T
Weighted LDpred2
PRS-CSx
CT-SLEB
Proposed Multi-Ancestry Method (Alternative Approaches)
MUSS
Weighted MUSS
Proposed Multi-Ancestry Method
MUSSEL
(b)
Fixed common SNP heritability with mild negative selection
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.0
0.1
0.2
0.3
1e+05
GWAS Sample Size
Single-Ancestry Methods
C+T
LDpred2
EUR PRS-Based Methods
EUR C+T
EUR LDpred2
Existing Multi-Ancestry Methods
Weighted C+T
Weighted LDpred2
PRS-CSx
CT-SLEB
Proposed Multi-Ancestry Method (Alternative Approaches)
MUSS
Weighted MUSS
Proposed Multi-Ancestry Method
MUSSEL

Figure S11: Prediction R^2 with 95% bootstrap CIs on validation individuals of AFR (N=2,015–3,428), EAS (N=2,316-4,647), and AMR ancestries (N=3,479-4,397) in PAGE based on discovery GWAS from PAGE (AFR N_{GWAS}=7,775 – 13,699, AMR N_{GWAS}=13,894 – 17,558), BBJ (EAS N_{GWAS}=70,657 – 158,284), and UKBB (EUR N_{GWAS}=315,133 – 355,983), related to Figure 3. We used genotype data from 1000 Genomes Project (498 EUR, 659 AFR, 347 AMR, 503 EAS, 487 SAS) as the LD reference dataset. All methods were evaluated on the ~2.0 million SNPs that are available in HapMap 3 + MEGA, except for PRS-CSx which is evaluated based on the HapMap 3 SNPs only, as implemented in their software. Ancestry- and trait-specific GWAS sample sizes, number of SNPs included, and validation sample sizes are summarized in Table S7. A random half of the validation individuals is used as the tuning set to tune model parameters, as well as train the SL in CT-SLEB and MUSSEL or the linear combination model in weighted C+T, weighted LDpred2, PRS-CSx, and weighted MUSS. The other half of the validation set is used as the testing set to report R^2 values for PRS on each ancestry, after adjusting for whether or not the sample is from BioMe and the top 10 genetic principal components for BMI, and additionally the age at lipid measurement and sex. The 95% bootstrap CIs of the estimated R^2 are obtained from the testing set based on 10,000 bootstrap samples using the Bca approach[1] implemented in the R package "boot". Detailed 95% bootstrap CIs are reported in Table S17.

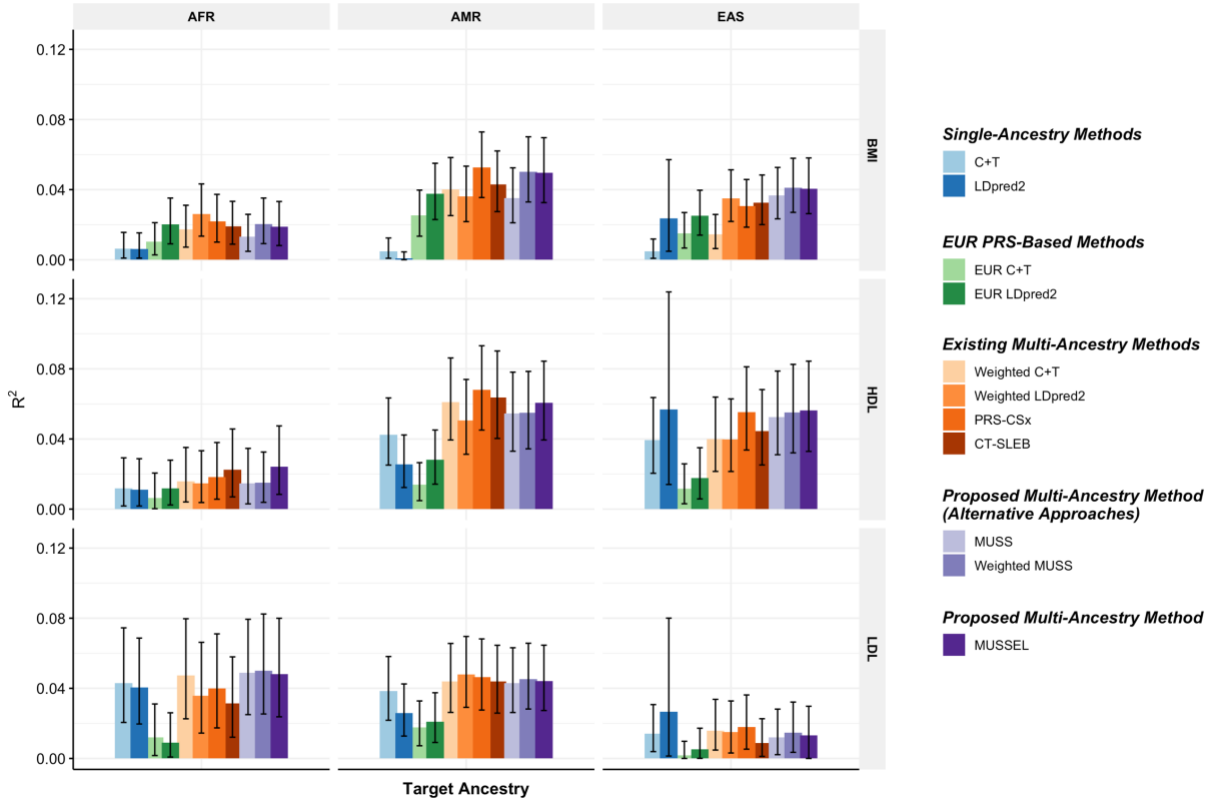


Figure S12: Prediction R^2 with 95% bootstrap CIs on UKBB validation individuals of EUR (17,457 – 19,030), AFR (7,954 – 8,598), EAS (1,752 – 1,921), and SAS (9,385 – 10,288) origin based on discovery GWAS from GLGC on EUR (N_{GWAS} =842,660 – 930,671), AFR or admixed AFR (N_{GWAS} =87,760 – 92,555), Hispanic/Latino (N_{GWAS} =46,040 – 49,582), EAS (N_{GWAS} =82,587 – 146,492), and SAS (N_{GWAS} =33,658 – 34,135). EUR (N_{GWAS} =842,660 – 930,671), AFR or admixed AFR (N_{GWAS} =87,760 – 92,555), Hispanic/Latino (N_{GWAS} =46,040 – 49,582), EAS (N_{GWAS} =82,587 – 146,492), and SAS (N_{GWAS} =33,658 – 34,135), related to Figure 4. The LD reference data is either (a) 1000 Genomes Project (498 EUR, 659 AFR, 347 AMR, 503 EAS, 487 SAS), or (b) UKBB data (PRS-CSx: default UKBB LD reference data which overlap with our testing samples including 375,120 EUR, 7,507 AFR, 687 AMR, 2,181 EAS, and 8,412 SAS; all other methods: UKBB tuning samples including 10,000 EUR, 4,585 AFR, 1,010 EAS, and 5,427 SAS). The ancestry of UKBB individuals were determined by a genetic ancestry prediction approach (Supplementary Notes). Due to the low prediction accuracy of genetic component analysis and extremely small validation sample size of UKBB AMR, prediction R^2 on UKBB AMR is unreliable and thus is not reported here. All methods were evaluated on the ~2.0 million SNPs that are available in HapMap 3 + MEGA, except for PRS-CSx which is evaluated based on the HapMap 3 SNPs only, as implemented in their software. Ancestry- and trait-specific GWAS sample sizes, number of SNPs included, and validation sample sizes are summarized in Table S9. A random half of the validation individuals is used as the tuning set to tune model parameters, as well as train the SL in CT-SLEB and MUSSEL or the linear combination model in weighted LDpred2, PRS-CSx, and weighted MUSS. The other half of the validation set is used as the testing set to report R^2 values for each ancestry. The 95% bootstrap CIs of the estimated R^2 are obtained from the testing set based on 10,000 bootstrap samples using the Bca approach[1] implemented in the R package "boot". Detailed 95% bootstrap CIs are reported in Table S17. In (b), PRS-CSx and other methods do not have a fair comparison because the UKBB LD reference data provided by the PRS-CSx software ($UKBB_{PRS\text{-}CSx}$) is much larger than that for other methods, and thus the R^2 of PRS-CSx PRS may be inflated due to a big overlap between $UKBB_{PRS\text{-}CSx}$ and the UKBB testing sample.

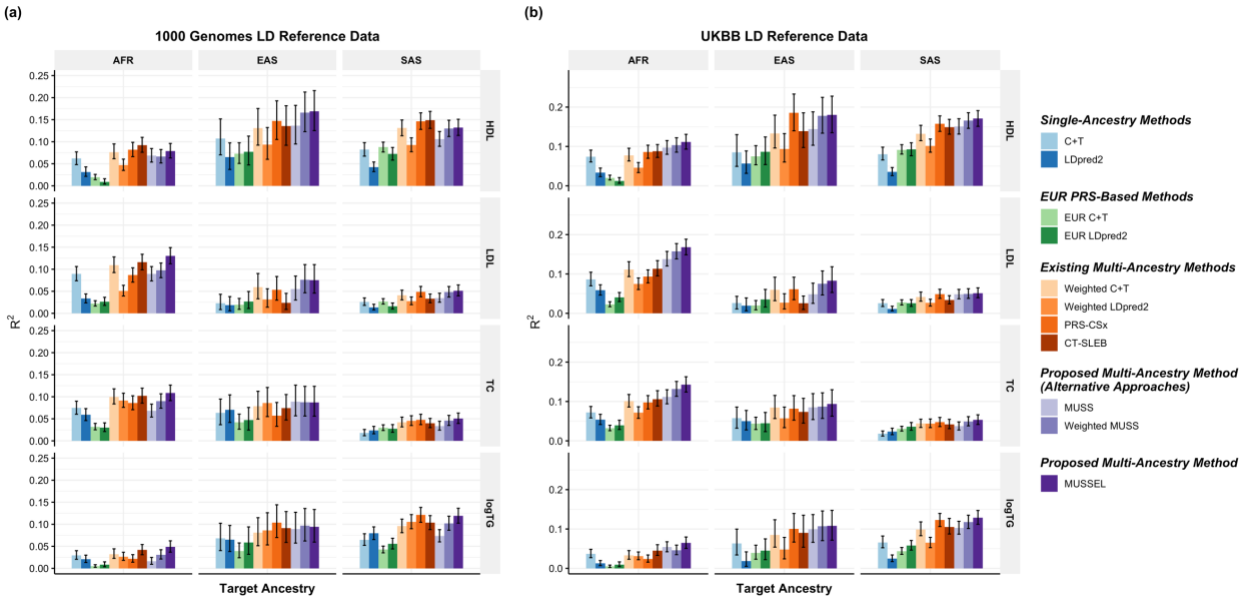


Figure S13: Prediction R^2 with 95% bootstrap CIs on UKBB validation individuals of AFR (N=9,169) origin based on discovery GWAS from AoU on EUR (N_{GWAS} =48,229 – 48,332), AFR (N_{GWAS} =21,514 – 21,550), and Hispanic/Latino (N_{GWAS} =15,364 – 15,413), related to Figure 5. The LD reference data is either **(a)** 1000 Genomes Project (498 EUR, 659 AFR, 347 AMR, 503 EAS, 487 SAS), or **(b)** UKBB data (PRS-CSx: default UKBB LD reference data which overlap with our testing samples including 375,120 EUR, 7,507 AFR, 687 AMR, 2,181 EAS, and 8,412 SAS; all other methods: UKBB tuning samples including 10,000 EUR, 4,585 AFR, 1,010 EAS, and 5,427 SAS). The ancestry of UKBB individuals were determined by a genetic ancestry prediction approach (Supplementary Notes). Due to the low prediction accuracy of genetic component analysis and extremely small validation sample size of UKBB AMR, prediction R^2 on UKBB AMR is unreliable and thus is not reported here. All methods were evaluated on the ~2.0 million SNPs that are available in HapMap3 + MEGA, except for PRS-CSx which is evaluated based on the HapMap 3 SNPs only, as implemented in their software. Ancestry- and trait-specific sample sizes of GWAS, number of SNPs included, and validation sample sizes are summarized in Table S11. A random half of the validation individuals is used as the tuning set to tune model parameters, as well as train the SL in CT-SLEB and MUSSEL or the linear combination model in weighted LDpred2, PRS-CSx, and weighted MUSS. The other half of the validation set is used as the testing set to report R^2 values for each ancestry, after adjusting for age, sex, and the top 10 genetic principal components. The 95% bootstrap CIs of the estimated R^2 are obtained from the testing set based on 10,000 bootstrap samples using the Bca approach[1] implemented in the R package "boot". Detailed 95% bootstrap CIs are reported in Table S17. In (b), PRS-CSx and other methods do not have a fair comparison because the UKBB LD reference data provided by the PRS-CSx software ($UKBB_{PRS\text{-}CSx}$) is much larger than that for other methods, and thus the R^2 of PRS-CSx may be inflated due to a big overlap between $UKBB_{PRS\text{-}CSx}$ and the UKBB testing sample.

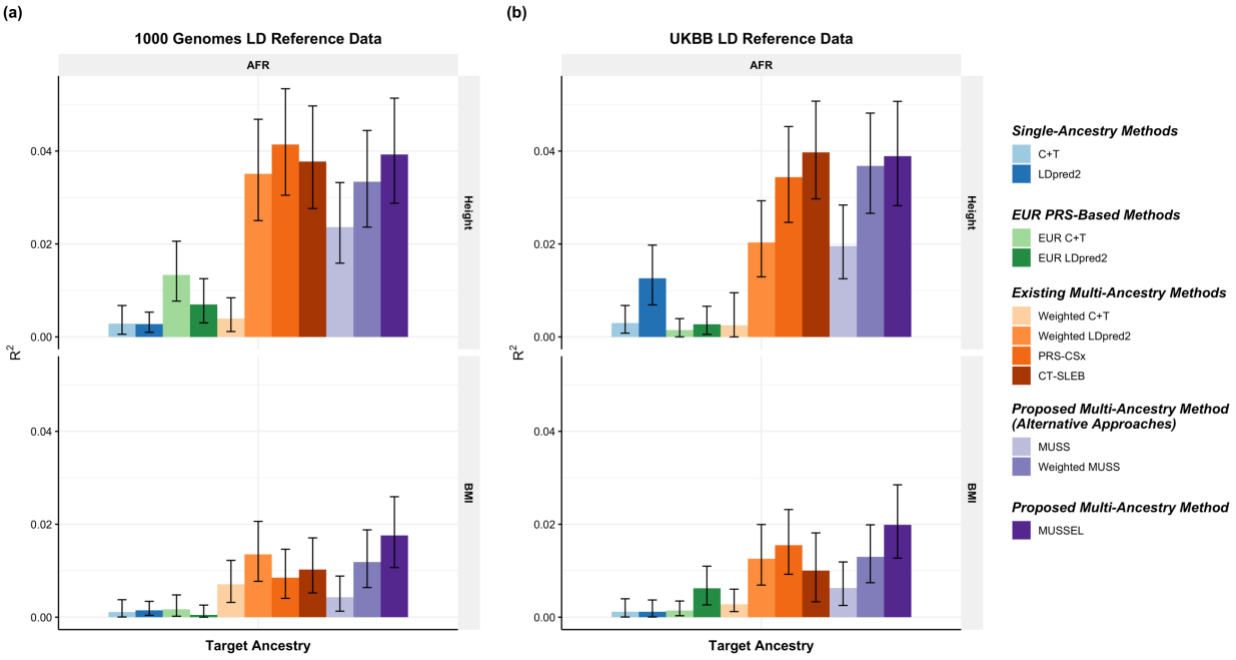


Figure S14: Simulation results with 20% ancestry mis-specification in the LD reference sample, assuming a fixed common SNP heritability (0.4) across ancestries under a strong negative selection model for the relationship between SNP effect size and allele frequency, related to Figure 2. The LD matrix for each ancestry group is estimated based on a slightly mis-specified LD reference sample that contains 800 individuals from the same ancestry group and 50 individuals from each of the other four ancestry groups, totaling 200 individuals with ancestry mismatch. The genetic correlation in SNP effect size is set to 0.8 across all pairs of populations. The causal SNP proportion (degree of polygenicity) is set to 1.0%, 0.1%, or 0.05% (~$192K$, $19.2K$, or $9.6K$ causal SNPs). We generate data for ~19 million common SNPs (MAF≥1%) across the five ancestry groups but conduct analyses only on the ~2.0 million SNPs in HapMap 3 + MEGA. The discovery GWAS sample size is set to **(a) 15,000 or (b) 45,000** for each non-EUR ancestry, and 100,000 for EUR. A tuning set consisting of 10,000 individuals is used for parameter tuning, as well as training the SL in CT-SLEB and MUSSEL or the linear combination model in weighted C+T, weighted LDpred2, PRS-CSx, and weighted MUSS. The reported R^2 values are calculated on an independent testing set of 10,000 individuals for each ancestry group.

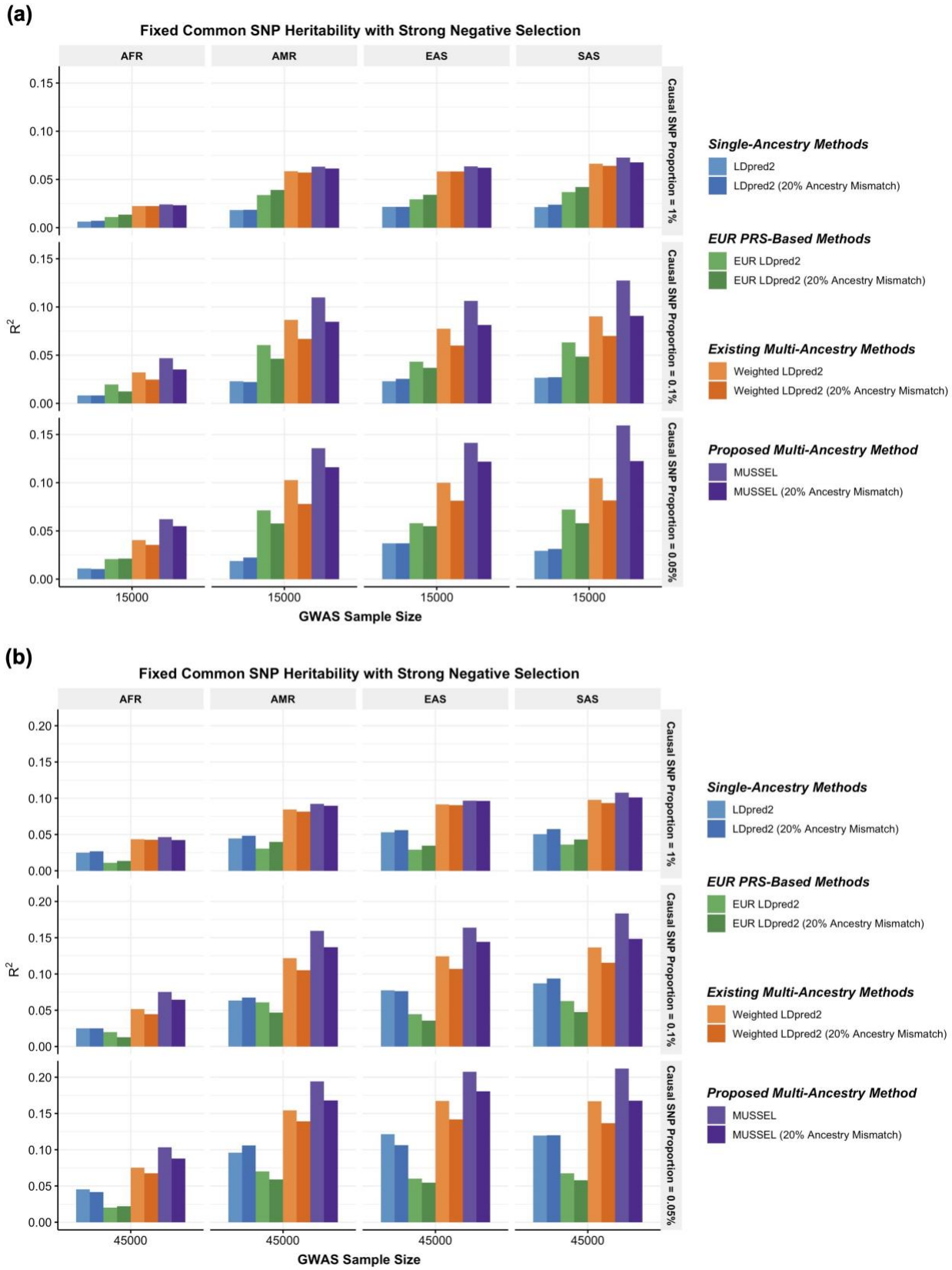
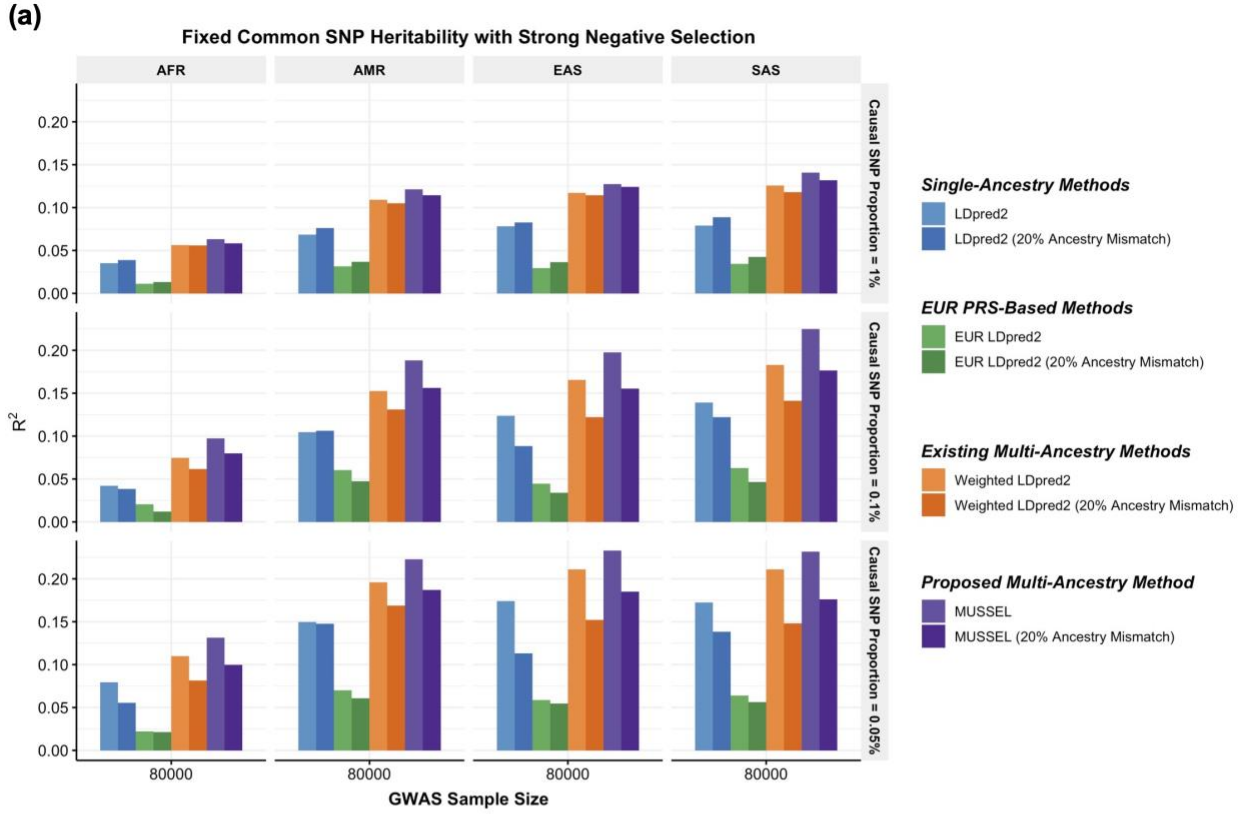

(a)
Fixed Common SNP Heritability with Strong Negative Selection
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.15
0.10
0.05
0.00
15000
GWAS Sample Size
Single-Ancestry Methods
LDpred2
LDpred2 (20% Ancestry Mismatch)
EUR PRS-Based Methods
EUR LDpred2
EUR LDpred2 (20% Ancestry Mismatch)
Existing Multi-Ancestry Methods
Weighted LDpred2
Weighted LDpred2 (20% Ancestry Mismatch)
Proposed Multi-Ancestry Method
MUSSEL
MUSSEL (20% Ancestry Mismatch)
(b)
Fixed Common SNP Heritability with Strong Negative Selection
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.20
0.15
0.10
0.05
0.00
45000
GWAS Sample Size
Single-Ancestry Methods
LDpred2
LDpred2 (20% Ancestry Mismatch)
EUR PRS-Based Methods
EUR LDpred2
EUR LDpred2 (20% Ancestry Mismatch)
Existing Multi-Ancestry Methods
Weighted LDpred2
Weighted LDpred2 (20% Ancestry Mismatch)
Proposed Multi-Ancestry Method
MUSSEL
MUSSEL (20% Ancestry Mismatch)

Figure S15: Simulation results with 20% ancestry mis-specification in the LD reference sample, assuming a fixed common SNP heritability (0.4) across ancestries under a strong negative selection model for the relationship between SNP effect size and allele frequency, related to Figure 2. The LD matrix for each ancestry group is estimated based on a slightly mis-specified LD reference sample that contains 800 individuals from the same ancestry group and 50 individuals from each of the other four ancestry groups, totaling 200 individuals with ancestry mismatch. The genetic correlation in SNP effect size is set to 0.8 across all pairs of populations. The causal SNP proportion (degree of polygenicity) is set to 1.0%, 0.1%, or 0.05% (~$192K$, $19.2K$, or $9.6K$ causal SNPs). We generate data for ~19 million common SNPs (MAF≥1%) across the five ancestry groups but conduct analyses only on the ~2.0 million SNPs in HapMap 3 + MEGA. The discovery GWAS sample size is set to **(a) 80,000 or (b) 100,000** for each non-EUR ancestry, and 100,000 for EUR. A tuning set consisting of 10,000 individuals is used for parameter tuning, as well as training the SL in CT-SLEB and MUSSEL or the linear combination model in weighted C+T, weighted LDpred2, PRS-CSx, and weighted MUSS. The reported R^2 values are calculated on an independent testing set of 10,000 individuals for each ancestry group.

(a)


Fixed Common SNP Heritability with Strong Negative Selection
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.00
0.05
0.10
0.15
0.20
80000
GWAS Sample Size
Single-Ancestry Methods
LDpred2
LDpred2 (20% Ancestry Mismatch)
EUR PRS-Based Methods
EUR LDpred2
EUR LDpred2 (20% Ancestry Mismatch)
Existing Multi-Ancestry Methods
Weighted LDpred2
Weighted LDpred2 (20% Ancestry Mismatch)
Proposed Multi-Ancestry Method
MUSSEL
MUSSEL (20% Ancestry Mismatch)

(b)

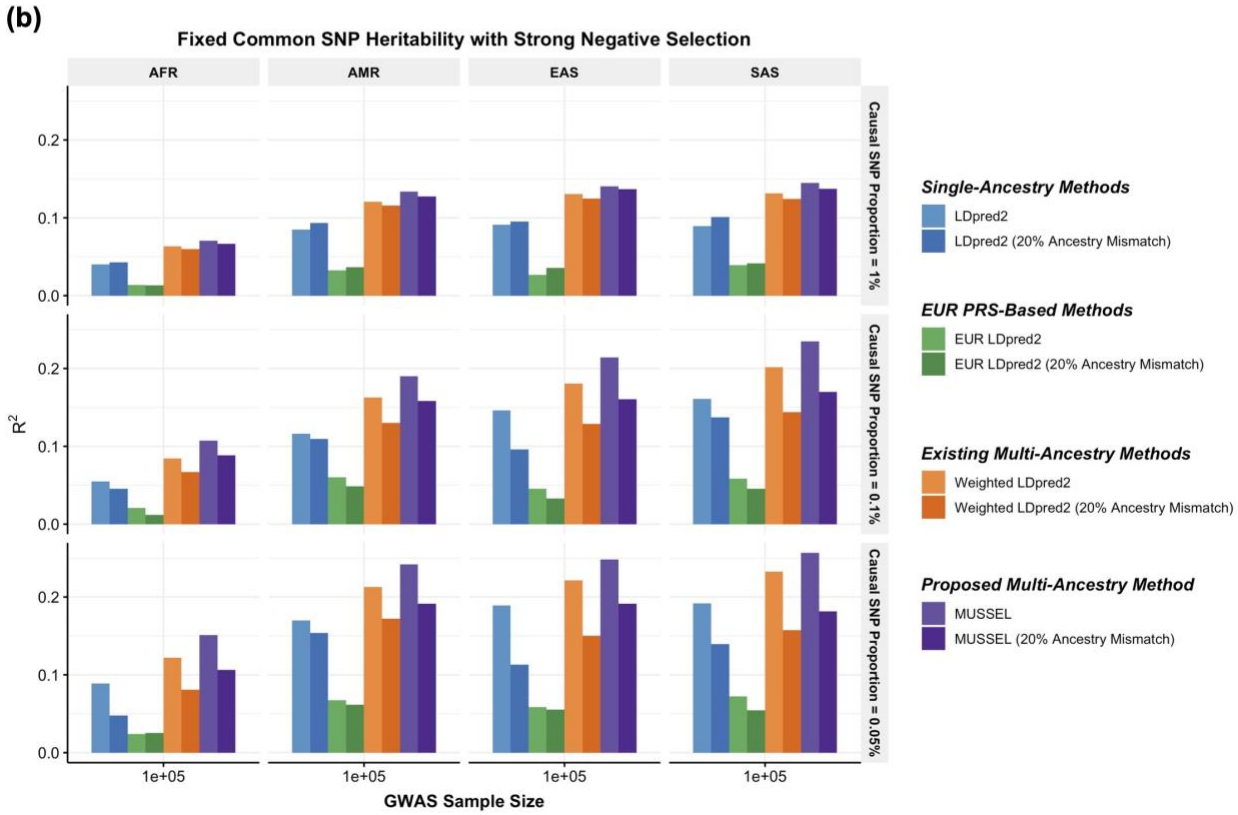

Fixed Common SNP Heritability with Strong Negative Selection
AFR
AMR
EAS
SAS
Causal SNP Proportion = 1%
Causal SNP Proportion = 0.1%
Causal SNP Proportion = 0.05%
R2
0.0
0.1
0.2
1e+05
GWAS Sample Size
Single-Ancestry Methods
LDpred2
LDpred2 (20% Ancestry Mismatch)
EUR PRS-Based Methods
EUR LDpred2
EUR LDpred2 (20% Ancestry Mismatch)
Existing Multi-Ancestry Methods
Weighted LDpred2
Weighted LDpred2 (20% Ancestry Mismatch)
Proposed Multi-Ancestry Method
MUSSEL
MUSSEL (20% Ancestry Mismatch)

Figure S16: Average R^2 and the 95% bootstrap CIs calculated by ancestry group across all traits available in the PAGE + UKBB + BBJ, GLGC, and AoU data analyses, related to Figures 3 - 6. For EUR and AFR, the calculations were conducted across all nine traits in the three data analyses; for EAS and SAS, the calculations were conducted across the seven traits in the PAGE + UKBB + BBJ and GLGC analyses; and for AMR (Hispanic/Latino), the calculations were conducted across the three traits in the PAGE + UKBB + BBJ analysis. To account for the effect of validation sample size on R^2, we calculated the weighted average of R^2, where weights are proportional to the validation sample sizes for different traits. Detailed results are summarized in Table S17.

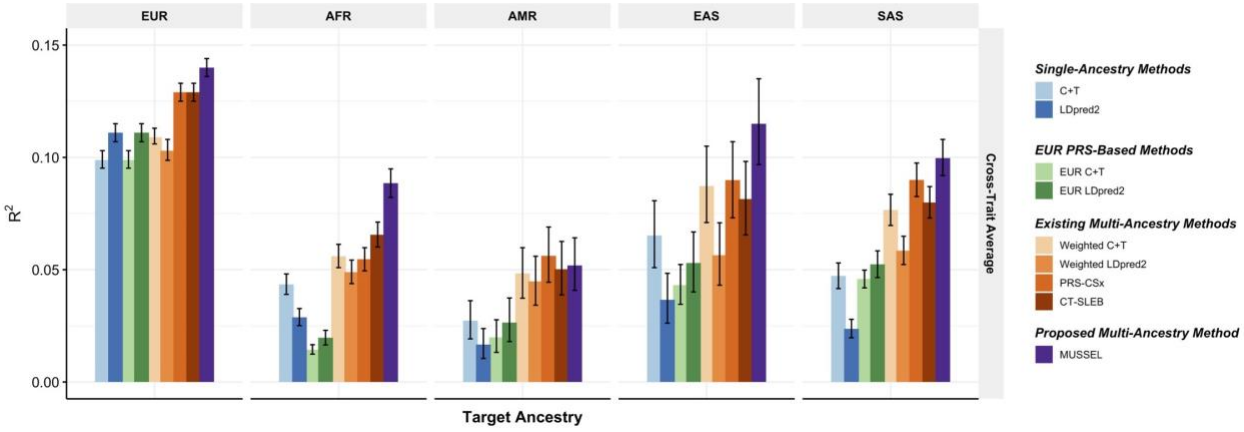
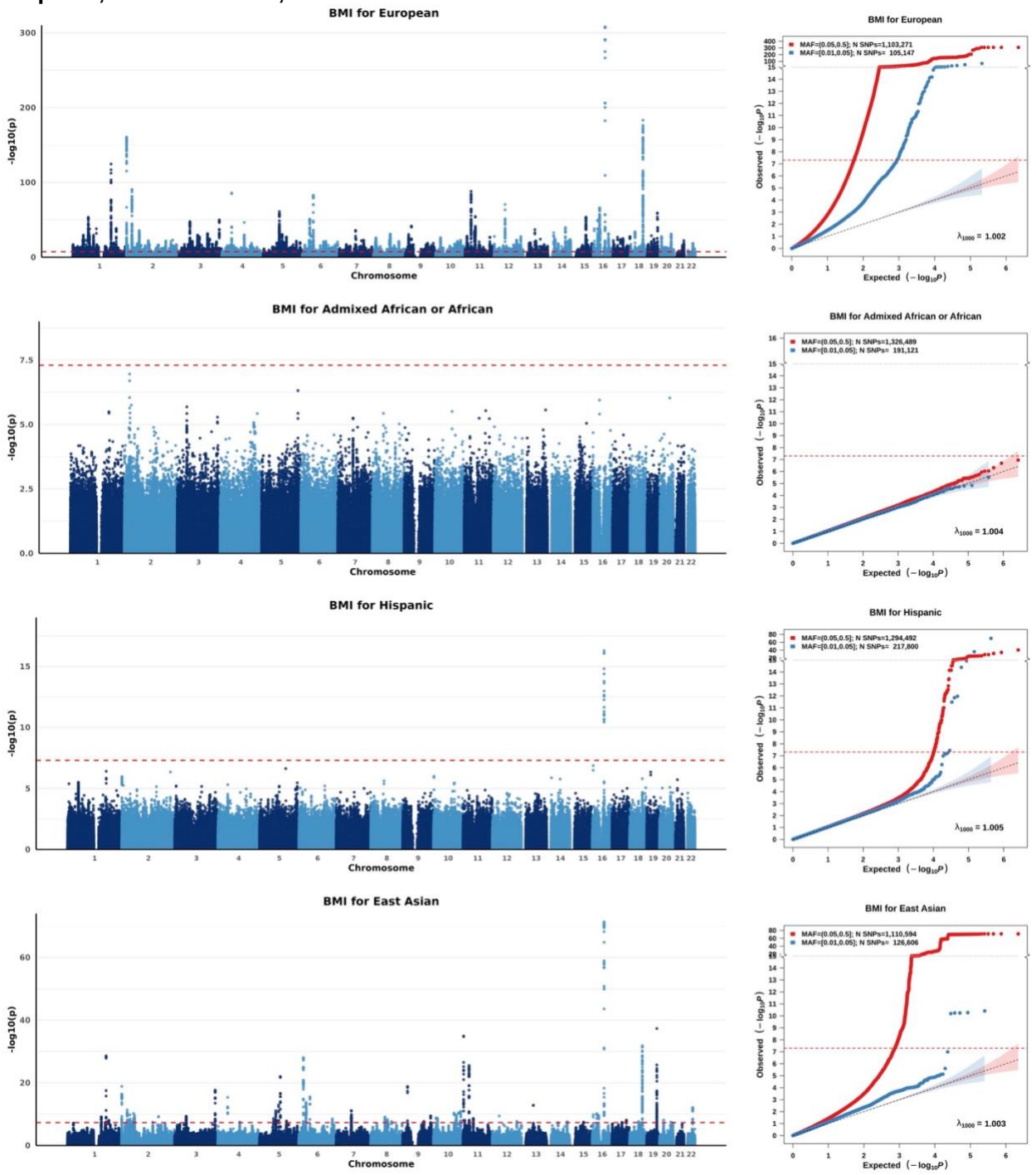
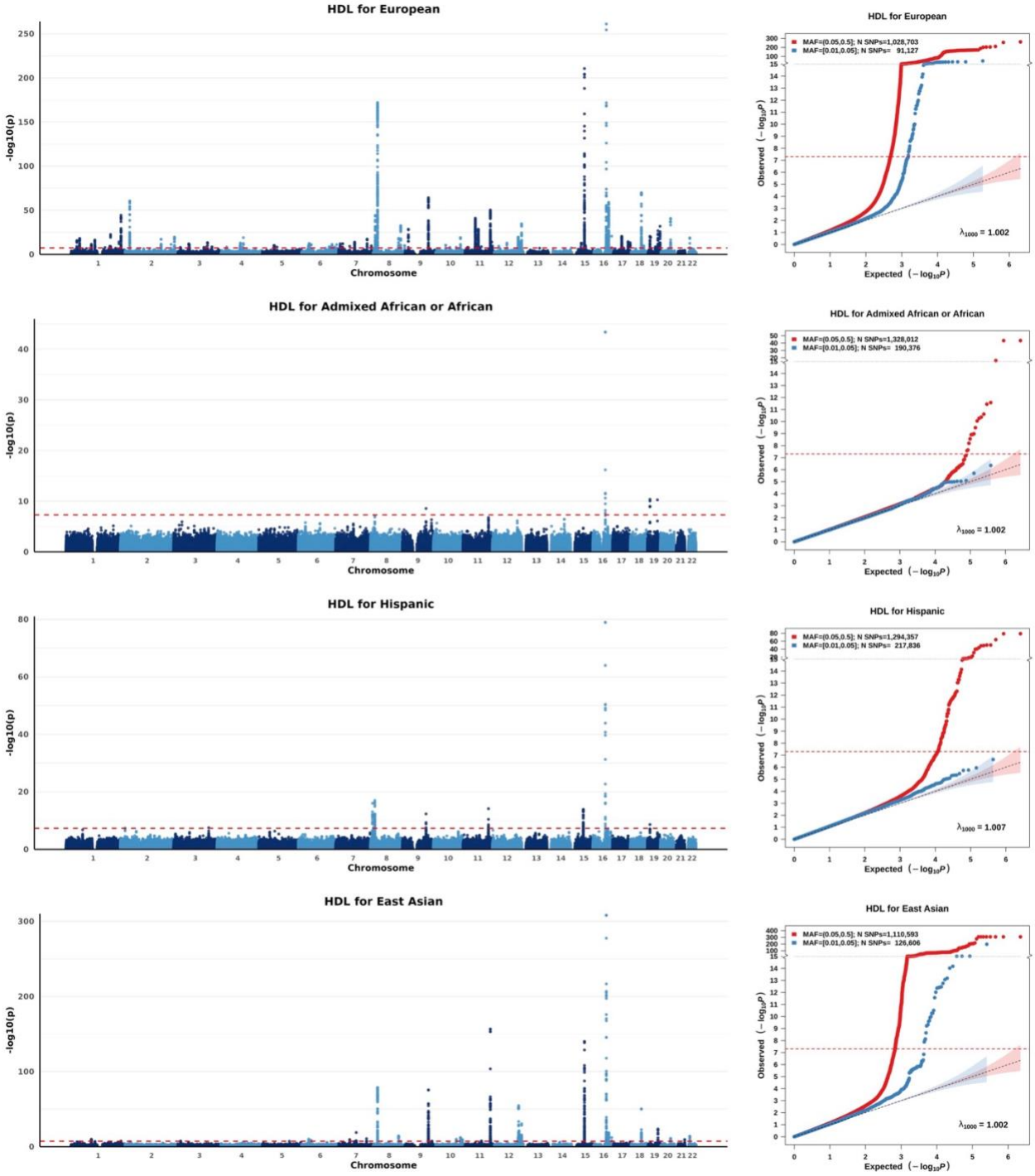


Figure S17: Manhattan plot and QQ plot[1] based on the GWAS summary-level association statistics from PAGE for BMI in four populations: European, Admixed African or African, Hispanic, and East Asian, related to STAR Methods.



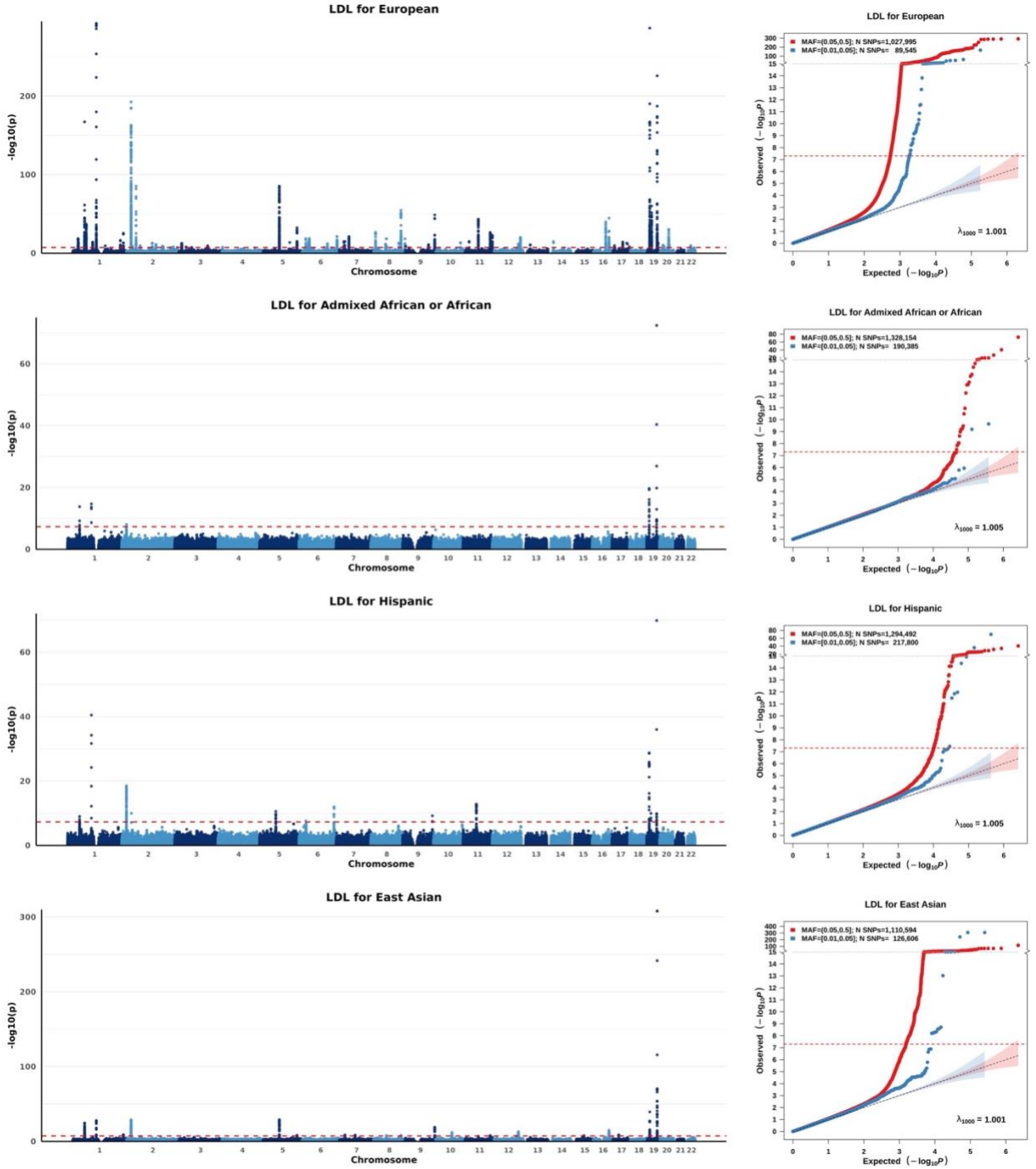
[1] For continuous traits, λ_{1000} scales the genomic inflation factor λ to a study with 1000 subjects using $\lambda_{1000} = 1 + 1000(\lambda - 1)/N$, where N is the total sample size. For binary traits, λ_{1000} scales λ to a study with 1000 cases and 1000 controls using $\lambda_{1000} = 1 + 1000(\lambda - 1)(\frac{1}{N_{case}} + \frac{1}{N_{control}})$.

Figure S18: Manhattan plot and QQ plot[1] based on the GWAS summary-level association statistics from PAGE for high-density lipoprotein (HDL) in four populations: European, Admixed African or African, Hispanic, and East Asian, related to STAR Methods.



[1] For continuous traits, λ_{1000} scales the genomic inflation factor λ to a study with 1000 subjects using $\lambda_{1000} = 1 + 1000(\lambda - 1)/N$, where N is the total sample size. For binary traits, λ_{1000} scales λ to a study with 1000 cases and 1000 controls using $\lambda_{1000} = 1 + 1000(\lambda - 1)(\frac{1}{N_{case}} + \frac{1}{N_{control}})$.

Figure S19: Manhattan plot and QQ plot[1] based on the GWAS summary-level association statistics from PAGE for low-density lipoprotein (LDL) in four populations: European, Admixed African or African, Hispanic, and East Asian, related to STAR Methods.



[1] For continuous traits, λ_{1000} scales the genomic inflation factor λ to a study with 1000 subjects using $\lambda_{1000} = 1 + 1000(\lambda - 1)/N$, where N is the total sample size. For binary traits, λ_{1000} scales λ to a study with 1000 cases and 1000 controls using $\lambda_{1000} = 1 + 1000(\lambda - 1)(\frac{1}{N_{case}} + \frac{1}{N_{control}})$.

Supplemental References

1. DiCiccio TJ, Efron B. Bootstrap Confidence Intervals. Statistical Science 1996;11:189-212.
2. Zhang, H., Zhan, J., Jin, J., Zhang, J., Ahearn, T.U., Yu, Z., O'Connell, J., Jiang, Y., Chen, T., Garcia-Closas, M., et al. (2023). A new method for multiancestry polygenic prediction improves performance across diverse populations. Nature Genetics 55(10), 1757-1768. 10.1038/s41588-023-01501-z.