Novel Methods for Multi-ancestry Polygenic Prediction and their Evaluations in 3.7 Million Individuals of Diverse Ancestry

Haoyu Zhang[1,4], Jianan Zhan[2], Jin Jin[3], Jingning Zhang[3], Thomas U. Ahearn[4], Zhi Yu[5], Jared O' Connell[2], Yunxuan Jiang[2], Tony Chen[1], 23andMe Research Team, Montserrat Garcia-Closas[4], Xihong Lin[1,5,6], Bertram L. Koelsch[2], Nilanjan Chatterjee[3,7]

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

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

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

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

[5] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[6] Department of Statistics, Harvard University, Cambridge, MA, USA

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

Conflicts of interest: J.Z., Y.J., J.O., and B.L.K. are employed by and hold stock or stock options in 23andMe, Inc.

Correspondence to: Haoyu Zhang (haoyuzhang@hsph.harvard.edu) and Nilanjan Chatterjee (nilanjan@jhu.edu)

Abstract

Polygenic risk scores are becoming increasingly predictive of complex traits, but subpar performance in non-European populations raises concerns about their potential clinical applications. We develop a powerful and scalable method to calculate PRS using GWAS summary statistics from multi-ancestry training samples by integrating multiple techniques, including clumping and thresholding, empirical Bayes and super learning. We evaluate the performance of the proposed method and a variety of alternatives using large-scale simulated GWAS on ~19 million common variants and large 23andMe Inc. datasets, including up to 800K individuals from four non-European populations, across seven complex traits. Results show that the proposed method can substantially improve the performance of PRS in non-European populations relative to simple alternatives and has comparable or superior performance relative to a recent method that requires a higher order of computational time. Further, our simulation studies provide novel insights to sample size requirements and the effect of SNP density on multi-ancestry risk prediction.

Introduction

Genome-wide association studies (GWAS) have identified tens of thousands of single nucleotide polymorphisms (SNPs) associated with complex traits and diseases[1,2]. Polygenetic risk scores (PRSs), which summarize the combined effect of individual SNPs, have the potential to improve risk stratification for various diseases and conditions[3–13]. However, GWAS to date have mostly been conducted in populations predominately comprised of European (EUR) origin individuals[14]. Consequently, the PRS generated from these studies tends to underperform in non-EUR populations, particularly in African (AFR) ancestry populations[9,15–18]. The lack of representation of non-EUR populations in PRS research has thus raised concerns that the use of current PRS for clinical applications may exacerbate health inequities[19–21].

In addition to the critical importance of addressing inequalities in representation of non-European population in genetic research, there is also an important need to develop statistical methods that leverage genetic data across populations to develop better performing PRS. Most of the PRS methods to date have been developed to analyze data from a single ancestry group[22–31], and subsequently, their performance was primarily evaluated for risk prediction in EUR populations[4–6,8–11]. While the same methods can also be used to build PRS in non-European populations, the resulting PRS, irrespective of the methods, tend to have limited performance due to limited sample sizes of the training datasets compared to sample sizes in European populations[15,20]. Some studies have conducted meta-analyses of GWAS across diverse populations to develop an underlying multi-ancestry PRS[32–34]. While such an approach may lead to a single PRS that performs more “equally” across diverse groups, it does not account for heterogeneity across populations and thus is not optimal for deriving the best PRS possible for each of the underlying populations.

Recent methods have focused on developing more optimal PRS in non-European populations by combining available GWAS from the target population of interest with “borrowed” information from larger GWAS in the EUR populations. One such study developed PRS in separate populations and then combined the PRS by optimally

weighting them to maximize performance in the target population[35]. More recent studies have proposed Bayesian methods that leverage multivariate priors for effect-size distribution to borrow information across populations; however, empirical studies showed the improvements in PRS performance from this approach was generally modest compared to simpler weighting methods[35,36]. Irrespective of these methods, methods for building PRS leveraging multi-ancestry datasets remain limited. Both theoretical and empirical studies have indicated that the optimal method for building PRS depends on multiple factors[22,36,37], including sample size, heritability and effect-size distribution and thus exploration of alternative methods with complementary advantages are needed to build optimal PRS is any given setting. Moreover, and perhaps more importantly, evaluation of the scope of multi-ancestry methods for building improved PRS remains quite limited to date due to the lack of suitably large GWAS for various non-EUR populations, especially of African origin, where risk prediction remains the most challenging.

In this paper, we propose a computationally simple and powerful method for generating PRSs using GWAS across diverse ancestry population. The method, which we refer to as CT-SLEB, is a model-free approach and combines the strength of multiple techniques, including a two-dimensional extension of the popular clumping and thresholding (CT) method[22,23], a super-learning (SL) model for combining multiple PRS and an empirical-Bayes (EB) approach to effect-size estimation. We compare the performance of the proposed method with a variety of alternatives based on large-scale simulated GWAS across five ancestry groups. In addition, we develop and validate population-specific PRS for seven complex traits using GWAS data from 23andMe, Inc. across Europeans (average $N \approx$ 2.47 million), African Americans (average $N \approx$ 117K), Latino (average $N \approx$413K), East Asians (average $N \approx$ 96K) and South Asians (average $N \approx$ 26K). Both simulation studies and empirical data analyses indicate that CT-SLEB is a highly scalable and powerful method for generating PRS for non-EUR populations. Further, our simulation studies and evaluation of various methods in the very large 23andMe datasets provide insights into the future yield of multi-ancestry PRSs as GWAS in diverse populations continues to grow.

Results

Method overview

CT-SLEB is a method designed to generate multi-ancestry PRSs that incorporate existing large GWAS from EUR populations and smaller GWAS from non-EUR populations. The method has three key steps (**Figure 1, Supplementary Figure 1**): 1. Clumping and Thresholding (CT) for selecting SNPs to be included in a PRS for the target population; 2. Empirical-Bayes (EB) method for estimating the coefficients of the SNPs; 3. Super-learning (SL) model to combine a series of PRSs generated under different SNP selection thresholds. The method requires three independent datasets: (1) GWAS summary statistics from training datasets across EUR and non-EUR populations; (2) a tuning dataset for the target population to find optimal model parameters; and (3) a validation dataset for the target population to report the final prediction performance. While this report assumes that individual-level data are available for model tuning and validation, summary-statistics-based methods[38,39] could also be used in these steps.

Two-dimensional Clumping and Thresholding

In step one, CT-SLEB uses two-dimensional clumping and thresholding on GWAS summary-statistics data to incorporate SNPs with either shared effects across the EUR and the target populations or population-specific effects in the target population (**Figure 1a**). Each SNP is assigned to the following two groups based on the p-value from the EUR and the target population: 1. SNPs with a p-value smaller in the EUR population; 2. SNPs with a p-value smaller in the target population or those which exist only in the target population. SNPs in the first group are ranked based on EUR p-value (smallest to largest) and then clumped using linkage disequilibrium (LD) estimates from the EUR reference sample. SNPs in the second group are ranked based on the target population p-value and clumped using LD estimates from the target population reference samples. Then the clumped SNPs from the two groups are combined as a candidate set for the next step. In the thresholding step, p-value thresholds are varied over a two-dimensional set of grid points. Each dimension corresponds to the threshold used for

the p-value obtained from one of the populations. At any given combination of thresholds, a SNP may be included in the target population PRS if the corresponding p-value from either the EUR or the target population achieves the corresponding threshold.

Empirical-Bayes Estimation of Effect Sizes

Since the effect sizes of SNPs are expected to be correlated across populations[40,41], we propose an EB method to efficiently estimate effect sizes for SNPs to be included in the PRSs (**Figure 1b**). Based on the selected SNP set from the CT step, we first estimate an underlying "prior" covariance matrix of effect sizes between the EUR population and the target population. Then, we estimate the effect size for each SNP in the target population based on the corresponding posterior mean, which weighs the effect-size estimate available from each population based on the bias-variance trade-off (**Methods**).

Super Learning

Previous research has shown that combining PRSs under different p-value thresholds can efficiently increase the prediction performance[25]. Thus, as a final step, we propose a super-learning model to predict the outcome using PRSs generated under different tuning parameters as training variables (**Figure 1c**). The super-learning model is a linear combination of different predictors based on multiple supervised learning algorithms[42–45]. The set of prediction algorithms can be self-designed or chosen from classical prediction algorithms. We choose Lasso[46], ridge regression[47], and neural networks[48] as three different candidate models in the implementation. We train the super-learning model on the tuning dataset and evaluate the performance of the final PRS using the independent validation dataset.

*Design of Simulation Studies***.**

We use simulation studies to compare eight methods across three broad categories: 1. *single ancestry* methods that only use the training and tuning data from the target population; 2. *EUR PR*S, which are generated using single ancestry methods to EUR

only GWAS; 3. *multi-ancestry* methods that use the training data from both the EUR and the target population. Single ancestry methods include two representative PRS methods: CT[22,23] and LDpred2[24,31]. EUR PRSs are generated based on CT and LDpred2. The multi-ancestry methods include 1. Weighted-PRS method that applies CT separately on the EUR and the target population and derives an optimal linear combination of the two. 2. PRS-CSx[49], which uses a Bayesian framework to calculate the posterior mean of effect sizes for the EUR and the target population, and then further derives an optimal linear combination of the two using a tuning dataset. 3. CT-SLEB, the proposed method. For CT-SLEB, we generate candidate SNP sets for PRS for each target population by applying the CT step across the EUR and the target population. However, for estimating effect sizes for any target population using the EB method, we combine GWAS summary-statistics data from either two ancestries (the EUR and the target population) or all five ancestries (**Supplementary Figure 1**). For computational efficiency, most analyses are restricted to ~2.8 million SNPs included in Hapmap3 (HM3)[50], or the Multi-Ethnic Genotyping Arrays (MEGA)[51] chips array, or both. However, the PRS-CSx method is currently implemented with only ~1.3 million HM3 SNPs in the provided software and thus the application of the method in our analysis is also restricted to only the HM3 SNPs.

Simulation Study Results

Results from simulation studies **(Figure 2** and **Supplementary Figure 2-6)** show that generally multi-ancestry methods lead to the most predictive PRSs in different settings. When the training data sample size for the target population is small (**Figure 2a, Supplementary Figure 2a, 3-6 a-b**), PRSs derived from the single ancestry methods perform poorly compared to EUR-based PRS. On the other hand, when the training sample size for the target population is large (**Figure 2b, Supplementary Figure 2b, 3-6 c-d**), PRSs generated by the single ancestry methods can outperform EUR PRS. PRS generated from the multi-ancestry methods can lead to substantial improvement in either setting.

Among multi-ancestry methods, we observe that both CT-SLEB and PRS-CSx can lead to improvement over the weighted PRS method. Between the two methods, none is uniformly superior to the other across all scenarios considered. When the sample size for the target population is relatively small (N = 15K), the PRS-CSx often outperforms CT-SLEB when the degree of polygenicity is the highest ($p_{causal} = 0.01$) . On the other hand, in the same sample size setting, CT-SLEB often outperforms PRS-CSx, by a notable margin when the degree of polygenicity is the lowest ($p_{causal} = 5 \times 10^{-4}$). When the sample size for the target population is larger (N=45K-100K), the difference between the two methods decreases, but in several scenarios, significant advantages of the CT-SLEB for lower polygenic setting remains but not vice versa (see **Figures 2b, Supplementary Figure 2a-b, 5 b-d and 6 b-d).** CT-SLEB, when implemented with EB estimation across all five ancestries, outperforms all alternative approaches across all scenarios considered. Under different simulation settings, the number of SNPs used by CT-SLEB ranged from 549K to 933K, while PRS-CSx retained all HM3 SNPs (**Supplementary Table 1**). Further, in a comparison of runtime using data on chromosome 22 to construct PRS for AFR (**Methods, Supplementary Table 2**), we observed runtime of CT-SLEB is on average almost 40 times faster than that of PRS-CSx (5.74 vs. 213.13 mins) using a single core with Intel E5-26840v4 CPU.

Unequal predictive performance of PRS across different populations has been considered as a barrier to ethical implementation of the technology in healthcare. Thus, we examined how large the sample size one may need for training GWAS in various minority populations to bridge the gap in the performance of PRS in comparison to that of the EUR population. Results indicate that when effect sizes for shared causal SNPs are similar across populations (genetic correlation=0.8), the gap can be mostly eliminated for all populations except AFR when the sample size reaches between a quarter to half of that of the EUR population (**Figure 3**, **Supplementary Figure 7**). For the AFR population, however, the sample size requirement can dramatically vary depending on the underlying genetic architecture of the traits. If the common SNP heritability is assumed to be the same for the AFR population as that of the other populations, then the sample size requirement for the AFR population is dauntingly

large because of substantially smaller per-SNP heritability (**Figure 3a-b**, **Supplementary Figure 7a-b**). If we allow the per-SNP heritability to remain the same across populations, but heritability to vary proportionately to the number of common variants, then the sample size requirement for the AFR population is similar to those for the other minority populations (**Figure 3c-d**, **Supplementary Figure 7c)**.

A major advantage of CT-SLEB over PRS-CSx is its computational scalability of the former method to handle much larger number of SNPs. Thus we use CT-SLEB to investigate the effect of SNP density on PRS performance by considering three different SNP sets to be used for PRS building: (1) ~1.3 million SNPs represented in HM3[50] (2) ~2.8 million SNPs that include all HM3 SNPs and additional SNPs represented in the MEGA array (3) All ~19 million common SNPs included in the 1000 Genomes Project (Phase 3)[52] which were used to generate the traits in our simulation studies. We observe that in general performances of PRS in various US minority populations can be substantially enhanced by inclusion of SNPs in denser panels, and the benefit due to denser panels is more enhanced when the sample size for the target population is larger and in settings where the proportions of causal SNPs are smaller (**Figure 4 and Supplementary Figure 8**).

23andMe data analysis results

We develop and validate population-specific PRS for seven complex traits using GWAS data from 23andMe, Inc. (**Methods, Supplementary Table 3**). We conduct GWAS using a training dataset for each population adjusting for principal component (PC) 1-5, sex and age following standard quality control (**Methods**). The Manhattan plots and QQ plots for GWAS are shown in **Supplementary Figures 9-15**, and no inflation is observed given the genomic inflation factor (**Supplementary Table 4**). We estimate heritability for the seven traits in the EUR population using LD-score regression[53] (**Supplementary Table 5, Methods**).

Results for heart metabolic disease burden and height (**Figure 4, Supplementary Table 7**) show a similar pattern as our simulation studies. The CT-SLEB and PRS-CSx

methods generally lead to the best performing PRS across different populations. Compared to the best performing European or single ancestry PRS, the relative gain is often large, especially for the African American (AA) population. The weighted method did not perform well for the AA population, but it substantially improved performance compared to each component PRS (EUR and single ancestry) for other populations. Among CT-SLEB and PRS-CSx, both of which perform notably better than the weighted method, the former tends to outperform the latter by a modest margin for heart metabolic disease burden, while the converse is true for height. We also observe that even with the best performing method and large sample sizes across all populations, a significant gap remains for the performance of the PRSs in non-EUR populations compared to those in the EUR population (**Supplementary Table 7**).

We also observe similar trends for the analyses of 23andMe data for the five binary traits: any cardiovascular disease (any CVD), depression, migraine diagnosis, morning person, and sing back musical note (SBMN) (**Figure 5, Supplementary Table 7**). For most settings, CT-SLEB and PRS-CSx often produce the best performing PRS and often lead to substantial improvement over best EUR PRS, single ancestry PRS, or weighted PRS. For CVD, which is the clinically most relevant trait for risk prediction and preventive intervention, we observe that CT-SLEB tends to outperform PRS-CSx by a notable margin except for the EAS population. We also observe that for the AA population, which are particularly underrepresented in genetic research, CT-SLEB outperforms PRS-CSx by a notable margin for several traits (e.g., CVD and morning person). In contrast, PRS-CSx outperforms CT-SLEB by a significant margin for predicting migraine diagnosis and SBMN in the SAS population. Similar to the continuous traits, we also observe that even with best performing methods and substantially large GWAS in a number of non-EUR populations, major gap often remains for the performance of PRS in these populations compared to those for the EUR population.

Discussion

In summary, we have proposed CT-SLEB as a powerful and computationally scalable method to generate optimal PRS across ancestrally distinct groups by utilizing GWAS across diverse populations. We compare the performance of CT-SLEB with those from a variety of both simple and complex methods, in large-scale simulation studies and very large datasets. Results indicate that while there is not a uniformly best performing method across all scenarios, the CT-SLEB method remains optimal or close to optimal in a wide variety of settings. Computationally, CT-SLEB is an order of magnitude faster than a recently proposed Bayesian method, PRS-CSx[49], and can more easily handle much larger SNP contents and additional populations.

A unique contribution of our study is the evaluation of a variety of PRS methodologies in the unprecedented large and diverse settings of the 23andMe, Inc. GWAS datasets. Our results provide important insights into the future yield of emerging large multi-ancestry GWAS. Adult height is often used as a model to explore the genetic architecture of complex traits and the potential for polygenic prediction. We observe that the standard CT method, when trained in ~2 million EUR individuals, leads to a PRS for the underlying population with a prediction R^2 of approximately 0.276. Application of LD-score regression to the same 23andMe data leads to an estimated GWAS heritability (the optimal R^2 for a PRS) of height of 0.395, indicating that the PRS has achieved about 69.8% (0.276/0.395) of its maximum potential in the 23andMe EUR population. We observe, however, that even with the best method and large sample size of the GWAS (N_{Latino}~350K and N_{AA} ~100K), the prediction accuracy of height PRS for non-EUR populations fell substantially short compared to that of the EUR population (Relative R^2~0.67 for East Asians, Latinos and South Asians and ~0.33 for AA compared to that of EUR).

We also observe similar patterns for other traits, including disease outcomes for which risk prediction is of most interest. For CVD, for example, the CT method, when trained in a sample of ~700K cases and ~1.3 million controls from the EUR population, produces a PRS that by itself has a prediction accuracy of the area under the ROC curve (AUC) as 0.65. For other populations, in some of which the sample size is

considerably large ($N_{case}/N_{control}$ =32K/66K for AA and $N_{case}/N_{control}$ =84K/270K for Latino) but still much smaller than that of the EUR population, the AUCs for best performing PRSs are close to 60% or lower. Further, the sample size is not the only driving factor for differential performances of PRS across populations. For example, the performance of the best performing CVD-PRS for the Latino and South Asian populations are very similar even though the sample size for the later population is much smaller. Collectively, these and additional results from simulation studies, indicate that bridging the gap between PRS performance across populations will require much more parity in the sample size of the underlying GWAS.

Both our simulation studies and data analyses indicate that no single PRS method is expected to be uniformly most powerful in all settings. In general, the optimal method for generating PRS will depend on the nature of the underlying multivariate effect-size distribution of the traits across different populations. While Bayesian methods, in principle, can generate the optimal PRS under correct specification of underlying effect-size distribution[24,31], modeling of effect-size distribution in multi-ancestry settings can be challenging. The CT method and their extensions, on the other hand, while they do not require strong modeling assumptions about effect-size distribution, they do not optimally incorporate LD among SNPs. We advocate that in future applications, researchers consider generating and evaluating a variety of PRS obtained from complementary methods. As different PRS may contain some orthogonal information, at the end, instead of choosing one best PRS, the best strategy could be to combine them using a final super learning step.

Our study has several limitations. While sample sizes for 23andMe datasets are extremely large, the power of genetic risk prediction is likely to have been blunted in this population, compared to other settings, due to the presence of a higher level of environmental heterogeneity. For example, a recent study[54] reported achieving prediction R^2 for height of ~41% for the EUR individuals within the UK Biobank using a PRS developed on ~1.1 million individuals from the UK Biobank (N=400K) and 23andMe (N=700K). In comparison, the PRS prediction R^2 for height we could achieve

within 23andMe EUR population is only ~30% despite doubling the sample size of the training dataset. We, however, also note that the estimate of heritability in 23andMe ($h^2_{SNP} = 0.395$) is substantially smaller than those previously reported[55,56] based on the UK Biobank ($h^2_{SNP} \sim 0.5 - 0.7$). When we compare the results across the two studies using prediction R^2 relative to the underlying heritability of the respective populations, we do see a significant gain in performance due to the increased sample size of the current study. Thus, we believe that while caution is needed to extrapolate 23andMe study results to other populations, the relative performance of PRSs we observe across different methods and different ancestry groups within this population is likely to be generalizable to other settings.

While we have compared the performance of the proposed method relative to a variety of alternatives, several additional methods not included in our analysis merit attention. These include the XPASS method[57], which uses a bivariate normal prior for effect sizes for shared variants across a pair of population and allows additional components for incorporating of population-specific SNPs. The method is also shown to improve the performance of PRS in a minority population by borrowing information from a larger GWAS from a majority population. The PRS-CSx method, which we did include in our comparison, is more flexible and likely to be a more powerful Bayesian method as it allows non-normal effect-size distribution and incorporation of data from more than two populations. Other available methods include Polypred[58], which extends the weighted PRS method incorporating functional annotation information. The method, however, is not directly comparable to the other methods we considered in our analysis, which do not incorporate any functional annotation data. Future studies are needed to explore how functional annotation data, including those from recent multi-ancestry omic studies[59], can be optimally incorporated in alternative advanced methods including PRS-CSx and CT-SLEB. Another limitation of the proposed method is that it is primarily designed to generate PRS across diverse populations which can be considered ancestrally distinct. But for many populations, such as the African and Hispanic origin population in the US, are highly admixed in nature and for these populations the

development or/and clinical reporting of PRS can be potentially improved by explicitly taking into consideration individual-level estimates admixture proportions.

In conclusion, we have proposed a novel and computationally scalable method for generating powerful PRS using data from GWAS studies in diverse populations. Further, our simulation studies and data analysis across multiple traits involving large 23andMe Inc. studies provide unique insight into what is likely to emerge from future GWAS in diverse populations for years to come.

Author contribution

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

Acknowledgements

We would like to thank the research participants and employees of 23andMe, Inc for making this work possible. We want to thank Liz Noblin, Melissa J. Francis and Emily Voeglein for helping with the research collaboration agreement with Harvard T.H. Chan School of Public Health, Johns Hopkins Bloomberg School of Public Health and 23andMe, Inc. The analysis utilized the high-performance computation Biowulf cluster at National Institutes of Health, USA, 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. This work was funded by NIH grants: K99 CA256513-01 (H. Z.), R01 HG010480-01 (N. C., J. J. and J. Z.) and U01HG011724 (N. C.)

Code and Data availability

CT-SLEB package and tutorial: https://github.com/andrewhaoyu/CTSLEB

P + T: https://www.cog-genomics.org/plink/1.9/

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

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

LDSC: https://github.com/bulik/ldsc

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

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

References

1. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
2. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
3. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).
4. Khera, A. v. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
5. Rao, A. S. & Knowles, J. W. Polygenic risk scores in coronary artery disease. *Curr. Opin. Cardiol.* **34**, 435–440 (2019).
6. Mavaddat, N. *et al.* Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am. J. Hum. Genet.* **104**, 21–34 (2019).
7. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* **28**, R133–R142 (2019).
8. Jia, G. *et al.* Evaluating the Utility of Polygenic Risk Scores in Identifying High-Risk Individuals for Eight Common Cancers. *JNCI Cancer Spectr.* **4**, (2020).
9. Dikilitas, O. *et al.* Predictive Utility of Polygenic Risk Scores for Coronary Heart Disease in Three Major Racial and Ethnic Groups. *Am. J. Hum. Genet.* **106**, 707–716 (2020).
10. Li, R., Chen, Y., Ritchie, M. D. & Moore, J. H. Electronic health records and polygenic risk scores for predicting disease risk. *Nat. Rev. Genet.* **21**, 493–502 (2020).
11. Zhang, H. *et al.* Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat. Genet.* **52**, 572–581 (2020).
12. Graff, R. E. *et al.* Cross-cancer evaluation of polygenic risk scores for 16 cancer types in two large cohorts. *Nat. Commun.* **12**, (2021).
13. Wray, N. R. *et al.* From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. *JAMA Psychiatry* **78**, 101–109 (2021).
14. Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**, 243–250 (2022).
15. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 1–9 (2019).
16. Liu, C. *et al.* Generalizability of Polygenic Risk Scores for Breast Cancer Among Women With European, African, and Latinx Ancestry. *JAMA Netw. Open* **4**, e2119084–e2119084 (2021).
17. Du, Z. *et al.* Evaluating Polygenic Risk Scores for Breast Cancer in Women of African Ancestry. *J. Natl. Cancer. Inst.* **113**, 1168–1176 (2021).
18. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).

19. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).
20. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
21. Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.* **11**, 1–9 (2020).
22. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**, 1520–1528 (2007).
23. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
24. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
25. Privé, F., Vilhjálmsson, B. J., Aschard, H. & Blum, M. G. B. Making the Most of Clumping and Thresholding for Polygenic Scores. *Am. J. Hum. Genet.* **105**, 1213–1221 (2019).
26. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 1–11 (2019).
27. Newcombe, P. J., Nelson, C. P., Samani, N. J. & Dudbridge, F. A flexible and parallelizable approach to genome-wide polygenic risk scores. *Genet. Epidemiol.* **43**, 730–741 (2019).
28. Ge, T., Chen, C. Y., Ni, Y., Feng, Y. C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1–10 (2019).
29. Song, S., Jiang, W., Hou, L. & Zhao, H. Leveraging effect size distributions to improve polygenic risk scores derived from summary statistics of genome-wide association studies. *PLoS Comput. Biol.* **16**, e1007565 (2020).
30. Zhou, G. & Zhao, H. A fast and robust Bayesian nonparametric method for prediction of complex traits using summary statistics. *PLoS Genet.* **17**, (2021).
31. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2021).
32. Koyama, S. *et al.* Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease. *Nat. Genet.* **52**, 1169–1177 (2020).
33. Sakaue, S. *et al.* Trans-biobank analysis with 676,000 individuals elucidates the association of polygenic risk scores of complex traits with human lifespan. *Nat. Med.* **26**, 542–548 (2020).
34. Agbaedeng, T. A. *et al.* Polygenic risk score and coronary artery disease: A meta-analysis of 979,286 participant data. *Atherosclerosis* **333**, 48–55 (2021).
35. Márquez-Luna, C. *et al.* Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* **41**, 811–823 (2017).
36. Dudbridge, F. & Wray, N. R. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet.* **9**, e1003348 (2013).

37. Chatterjee, N. *et al.* Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* **45**, 400–405 (2013).
38. Song, L. *et al.* SummaryAUC: a tool for evaluating the performance of polygenic risk prediction models in validation datasets with only summary level statistics. *Bioinformatics* **35**, 4038–4044 (2019).
39. Zhao, Z. *et al.* PUMAS: fine-tuning polygenic risk scores with GWAS summary statistics. *Genome Biol.* **22**, 1–19 (2021).
40. Brown, B. C., Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic Genetic-Correlation Estimates from Summary Statistics. *Am. J. Hum. Genet.* **99**, 76–88 (2016).
41. Shi, H. *et al.* Population-specific causal disease effect sizes in functionally important regions impacted by selection. *Nat. Commun.* **12**, 1–15 (2021).
42. van der Laan, M. J., Polley, E. C. & Hubbard, A. E. Super learner. *Stat. Appl. Genet. Mol. Biol.* **6**, (2007).
43. Polley, E. & van der Laan, M. J. Super Learner In Prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series* (2010).
44. van der Laan, M. J. & Rose, S. *Targeted learning: causal inference for observational and experimental data.* vol. 4 (Springer New York , 2011).
45. Ledell, E., Petersen, M. & van der Laan, M. J. Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. *Electron. J. Stat.* **9**, 1583–1607 (2015).
46. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 267–288 (1996).
47. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1 (2010).
48. Ripley, B. D. *Pattern recognition and neural networks*. (Cambridge university press, 2007).
49. Ruan, Y. *et al.* Improving Polygenic Prediction in Ancestrally Diverse Populations. *medRxiv* (2021).
50. Consortium, T. I. H. 3. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52 (2010).
51. Bien, S. A. *et al.* Strategies for Enriching Variant Coverage in Candidate Disease Loci on a Multiethnic Genotyping Array. *PLoS One* **11**, 167758 (2016).
52. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
53. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
54. Márquez-Luna, C. *et al.* Incorporating functional priors improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *Nat. Commun.* **12**, (2021).
55. Ge, T., Chen, C. Y., Neale, B. M., Sabuncu, M. R. & Smoller, J. W. Phenome-wide heritability analysis of the UK Biobank. *PLoS Genet.* **13**, (2017).
56. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).

57. Cai, M. *et al.* A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. *Am. J. Hum. Genet.* **108**, 632–655 (2021).
58. Weissbrod, O. *et al.* Leveraging fine-mapping and non-European training data to improve trans-ethnic polygenic risk scores. *medRxiv* 2021.01.19.21249483 (2021) doi:10.1101/2021.01.19.21249483.
59. Zhang, J. *et al.* Large Bi-Ethnic Study of Plasma Proteome Leads to Comprehensive Mapping of cis-pQTL and Models for Proteome-wide Association Studies. *bioRxiv* (2021).
60. Pritchard, J. K. & Przeworski, M. Linkage Disequilibrium in Humans: Models and Data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
61. Polley, E., LeDell, E., Kennedy, C. & van der Laan, M. J. SuperLearner: Super Learner Prediction. *R package version 2.0-26* (2019).
62. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304–2305 (2011).
63. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
64. Foucher, Y. *et al.* RISCA: Causal Inference and Prediction in Cohort-Based Analyses. *R package version 0.9* (2020).

Online Methods

We assume there are $l = 1, \dots L$ populations with $l = 1$ indexing the EUR population. We assume that for each population, summary-statistics data from underlying GWAS are available in the form of $(\hat{\beta}_{kl}, s_{kl}, p_{kl})$ for $k = 1,2, \dots, K_L$ SNPs, where $\hat{\beta}, s$ and p denote estimates of effect-size, standard errors and p-values associated with individual SNPs, respectively. We further assume that additional datasets are available for each target population of interest, which could be split into tuning and validation sets. Our proposed CT-SLEB method contains three steps: 1. Two-dimensional clumping and thresholding (CT); 2. EB procedure; 3. Super-learning algorithm, which will be described in detail in the following three subsections.

CT. In this step, we extend the traditional CT to a two-dimensional setting so that PRS for a target population can be built using approximately independent SNPs that show significant association in at least one of the two populations (majority population and the target population). The CT method has two components, the Clumping step and the Thresholding step. In the two-dimensional setting where the lead SNPs might be informed by GWAS of either the EUR or target population, it is unclear what reference sample is the most suited for LD clumping. After initial exploration of alternative approaches through simulation studies, we find the most informative approach is to split the SNPs into two sets depending on which population they show stronger signals and then perform LD clumping for each set separately based on the reference sample for the respective population. For the thresholding step, we select SNPs based on two distinct thresholds for their respective p-values in the two populations. As the optimal threshold for p-value selection is known to depend on sample size for underlying GWAS[23,36,37], and sample sizes for GWAS across EUR and minority populations are highly differential, we anticipate (and confirm through simulation studies) that a two-dimensional approach for threshold selection is more optimal than using a single p-value threshold across both populations. Following, we describe details of the CT step:

1. The clumping r^2-cutoff and base size of the clumping window size w_b vary across (0.01, 0.05, 0.1, 0.2, 0.5, 0.8) and (50kb, 100kb), respectively. The

clumping window size w_s is defined as w_b/r^2 because LD is inversely proportional to the genetic distance between variants[25,60].

2. Select all SNPs with smaller p-values in EUR ($p_{k1} < p_{k2}$), and then, clump based on p_{k1} using LD estimates from the EUR reference samples with selected r^2 and w_s.
3. Select all variants with smaller p-values in the target population ($p_{k2} < p_{k1}$) and the population-specific SNPs, and then, clump based on p_{k2} using LD estimates from the reference samples of the target population with the same r^2-cutoff and w_b.
4. Combine the post-clumping variants from the second and third steps as the candidate variants set.
5. Define two different p-value cutoffs (p_{t1}, p_{t2}) for the EUR and the target population. A variant is selected if $p_{k1} < p_{t1}$ or $p_{k2} < p_{t2}$. We allow p_{t1} and p_{t2} to vary in the set: $(5 \times 10^{-8}, 5 \times 10^{-7}, 5 \times 10^{-6}, \dots, 5 \times 10^{-1}, 1.0)$. With the cross combination of p_{t1} and p_{t2}, a total of 81 different p-value cutoffs are applied.
6. With the cross combination of p_{t1}, p_{t2}, r^2 and w_b, a total of 972 PRSs are evaluated on the tuning dataset using estimated regression coefficients ($\hat{\beta}_{k2}$) from GWAS of the target population.

EB to calibrate regression coefficients. In the CT step above, we use $\hat{\beta}_{k2}$ from the target population to calculate PRS. However, $\hat{\beta}_{k2}$ can be noisy when the GWAS sample size of the target population is small. Meanwhile, given the high genetic correlation across different ancestries[40,41], effect sizes from other populations can be used to calibrate the regression coefficients for the PRS. Although we only use p-values from GWAS for the EUR and the target population for selecting SNPs in the CT step, the EB step takes advantage of existing GWAS from multiple populations. Suppose $\hat{\boldsymbol{u}}_k = (\hat{u}_{k1}, \dots, \hat{u}_{kL}) = (\hat{\beta}_{k1}\sqrt{2f_{k1}(1-f_{k1})}, \dots, \hat{\beta}_{kL}\sqrt{2f_{kl}(1-f_{kL})})$, is the vector of the standardized effect-size for the kth SNP in L different populations, with $\hat{\boldsymbol{s}}^*_{\boldsymbol{k}} = (s^*_{k1}, \dots, s^*_{kL}) = \left(s_{k1}\sqrt{2f_{k1}(1-f_{k1})}, \dots, \hat{s}_{kL}\sqrt{2f_{kL}(1-f_{kL})}\right)$ being the vector of the

corresponding standard errors of $\widehat{\boldsymbol{u}}_k$. We assume that $\widehat{\boldsymbol{u}}_k|u_k \sim N(\boldsymbol{u}_k, \boldsymbol{\Sigma_k})$, where $\boldsymbol{\Sigma_k} = diag\{(\widehat{\boldsymbol{s}}_{\boldsymbol{k}}^*)^{\boldsymbol{2}}\}$ given that the GWAS for different populations are independent. Additionally, we assume that the prior distribution of the mean of $\widehat{\boldsymbol{u}}_k$ is $\boldsymbol{u}_k \sim N(\boldsymbol{0}, \boldsymbol{\Sigma_0})$. By integrating the conditional and prior distribution, we can obtain the marginal distribution of $\widehat{\boldsymbol{u}}_k$ as $N(\boldsymbol{0}, \boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_k})$. Suppose the SNP set selected from the CT step has K^* variants overlapped across all the populations. We estimate the prior covariance matrix $\boldsymbol{\Sigma_0}$ using the K^*overlapped variants shared across all populations as:

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{0}} = \frac{1}{K^* - 1}\sum_{k=1}^{K^*} \widehat{\boldsymbol{u}}_k^T \widehat{\boldsymbol{u}}_k - \boldsymbol{\Sigma_k}.$$

We note that we ignore any potential correlation across selected SNPs in this step, but the estimate is still expected to be consistent for $\boldsymbol{\Sigma_0}$ which represents marginal variance-covariance matrices for effect sizes associated with an individual SNP across populations. Applying the Bayes formula, the posterior distribution of $\boldsymbol{u}_k$ becomes

$$\boldsymbol{u}_k|\widehat{\boldsymbol{u}}_k \sim N\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{0}}(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{0}} + \boldsymbol{\Sigma_k})^{-\boldsymbol{1}}\widehat{\boldsymbol{u}}_k, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{0}}(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{0}} + \boldsymbol{\Sigma_k})^{-\boldsymbol{1}}\boldsymbol{\Sigma_k}\right).$$

The EB coefficients for the kth SNP are defined as:

$$\widehat{\boldsymbol{\beta}}_k^{EB} = \mathbf{F_k}\,\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{0}}\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{0}} + \boldsymbol{\Sigma_k}\right)^{-\boldsymbol{1}}\widehat{\boldsymbol{u}}_k,$$

where $\mathbf{F_k} = \mathrm{diag}\left\{\frac{1}{\sqrt{2f_{kl}(1-f_{kl})}}\right\}_{\mathrm{L\times L}}$ is the scaling matrix to scale effect sizes from the standardized scale back to the original scale. Preliminary simulation studies indicate that EB step of effect-size calibration leads to distinct improvement in PRS performance (compared to using effect-size estimates from the target population) irrespective of all other steps.

To save computational time, we estimate $\widehat{\boldsymbol{\Sigma}}_0$ in the above step only once based on the SNP set that gives the best PRS in the CT step across all different p-value thresholds, r^2-cutoff, and window sizes. We then apply the same $\widehat{\boldsymbol{\Sigma}}_0$ to derive the EB-calibrated effect sizes for SNPs included in all different PRSs corresponding to cross combination of p_{t1}, p_{t2}, r^2-cutoff and w_b. In all analyses, we compute the 1944 PRSs using EB-calibrated effect sizes of the target population and EUR (972 PRSs using the posterior coefficients for each population). When more than two ancestries are involved, we use

data from all populations to derive the EB estimates of effect-sizes for SNPs for each population. However, to save computational time at the super-learning step, we derive the final PRS for a target population by only incorporating the initial PRSs derived for the larger EUR population and those for the specific target population. All 1944 PRSs are used as input for the super-learning step to predict the outcome for the target population. Because many PRSs are highly correlated with each other, we filter out the highly redundant one with pairwise correlations higher than 0.98.

Super learning. We combine all PRSs generated from the above steps into an input dataset and train them on the tuning dataset to predict the outcome Y. The super-learning algorithm generates an optimally weighted combination from a set of distinct prediction algorithms[42–45] (**Supplementary Note**). The set of prediction algorithms can be self-designed or chosen from classical prediction algorithms e.g., Lasso[46], ridge regression[47], neural networks[48], etc. We use three different prediction algorithms implemented in the SuperLearner package[61] to generate the super learning estimate: Lasso[46], ridge regression[47] and neural networks[48]. For binary traits, since the ridge regression algorithm is not supported by the SuperLearner package now, we only use Lasso and neural networks in the data analysis. To use AUC as the objective function, we use the flag “method = method.AUC” in the SupearLearner package.

Simulation. Large-scale multi-ancestry genotype data are generated using HAPGEN2 (version 2.1.2)[62] mimicking the LD of EUR, AFR, Americas (AMR), East Asia (EAS) and South Asia (SAS). The 1000 Genomes Project (Phase 3)[52] is used as the reference panel which include 503 EUR, 661 AFR, 347 AMR, 504 EAS and 489 SAS subjects. Biallelic SNPs with MAF more than 0.01 in any of the populations are kept in the reference panel, resulting in ~8.6 million SNPs for EUR, ~14.8 million SNPs for AFR, ~9.8 million SNPs for AMR, ~7.6 million SNPs for EAS, and ~9.0 million SNPs for SAS. The genotype data are generated with a total of ~19.2 million SNPs. Different populations have population-specific SNPs and shared SNPs with other populations. The proportion of population-specific SNPs range from 2.92% for AMR to 43.84% for

AFR (**Supplementary Figure 16**). We simulate a total of 120,000 independent subjects for each of the population.

For generating trait values, we select causal SNPs randomly across the whole genome with the causal SNP proportion being set to 0.01, 0.001, or 5×10^{-4}. We consider two alternative models for generating heritability distribution within each population: (A) Constant common-SNP heritability. (B) Constant per-SNP heritability that implies the total heritability is proportional to the number of common SNPs. We also consider three different models for negative selection pattern: strong, mild and no negative selection.

We denote by u_{kl} the standardized effect-size for kth causal SNP for the lth population. Under strong negative selection and constant heritability model, the standardized effect-sizes are drawn from a multivariate normal distribution of the form:

$$u_{kl} \sim N\left(0, \frac{h^2}{C_l}\right), cov\left(u_{kl_1}, u_{kl_2}\right) = \frac{\rho h^2}{\sqrt{C_{l_1} C_{l_2}}},$$

where C_l is the number of causal SNPs with MAF > 0.01 in the lth population, the heritability h^2 associated with common SNPs for each population is set to 0.4, and the genetic correlation ρ is set to 0.8. We then generate the phenotype using linear model of the form $Y_{il} = \sum_{k=1}^{C_l} \frac{G_{ikl}}{\sqrt{2(f_{kl}(1-f_{kl})}} u_{kl} + \epsilon_{il}$ for the ith subject in the lth population, where $f_{kl}$is the effect allele frequency for the kth causal SNP in lth population. The error terms are generated as $\epsilon_{il} \sim N(0, 1 - h^2)$. We also consider mild negative selection ($u_{kl}^2 \propto [f_{kl}(1 - f_{kl})]^{0.75}$) and no negative selection ($u_{kl}^2 \propto [f_{kl}(1 - f_{kl})]$) scenarios (see **Supplemental Notes** for details). Finally, we simulate data under an assumption of total heritability of all ~19 million SNPs being 0.4 across all populations, but the common SNP heritability varying proportionately to their number within each the populations. The model assumes per SNP heritability to be the same across all populations and thus leads to the common SNP heritability value of 0.32, 0.21, 0.16, 0.19 and 0.17 for AFR, AMR, EAS, EUR and SAS, respectively. The genetic correlation is set to 0.8 or 0.6.

We set the training sample sizes for each target population to 15,000, 45,000, 80,000, or 100,000. We generate GWAS summary statistics for each population based on the training samples using PLINK version 1.90 with the command "--linear". We fixed the sample sizes for the EUR population at 100,000. We further simulate the tuning and validation dataset of size 10,000 for each target population. The final prediction R^2 is reported as the average of ten independent simulation replicates for each simulation setting. For evaluating CT-SLEB that incorporates data across all five ancestries, we assume the training sample size for the each of the other non-EUR populations to be the same as that of the target population.

Existing PRS methods. The CT method selects clumped SNPs with different p-value thresholds and picks a single optimal PRS based on its performance on the tuning dataset. We implement CT using PLINK version 1.90[63] with the clumping step command "--clump --clump-r2 0.1 --clump-kb 500". We estimate LD based on 3,000 randomly selected unrelated subjects from the training dataset for each population. We set the candidate p-value thresholds to be $(5 \times 10^{-8}, 1 \times 10^{-7}, 5 \times 10^{-7}, 1 \times 10^{-6}, \ldots, 5 \times 10^{-1}, 1.0)$ and for computing PRS, we use the PLINK command "--score no-sum no-mean-imputation". The optimal p-value threshold is determined based prediction R^2 (variation explained by the corresponding PRS) on the tuning dataset.

The LDpred2 method infers SNP effect sizes by a shrinkage estimator that combines GWAS summary statistics with a prior on effect sizes while leveraging LD information from an external reference panel. LDpred2 is implemented using the R package "bigsnpr"[31]. The tuning parameters included are: (1) the proportion of causal SNPs, with candidate values set to a sequence of length 17 that are evenly spaced on a logarithmic scale from 10^{-4} to 1; (2) per-SNP heritability, with candidate values set to 0.7, 1, or 1.4 times the total heritability estimated by LD score regression divided by the number of causal SNPs; (3) "sparse" option, which is set to "yes" or "no" (the "sparse" option sets some weak effects to zero). The method selects tuning parameters based on the performance on the tuning dataset.

The EUR PRS based on CT or LDpred2 are built based on training dataset from the EUR population and estimates tuning parameters based on tuning sample for the EUR population. When the EUR PRSs are evaluated in the target population, we exclude the SNPs that do not exist in the target population.

The weighted-PRS linearly combines the CT PRS generated from the EUR and from the target population. The weights for EUR PRS and for the target population PRS in the linear combination are estimated using the tuning dataset from the target population through a linear regression. We implement the weighted-PRS using R version 4.0.0.

The PRS-CSx method estimates population-specific SNP effect sizes based on a Bayesian framework using continuous shrinkage priors to jointly model the GWAS summary statistics from multiple populations. Besides the Bayesian modeling step, PRS-CSx further conducts a step similar to weighted-PRS, which is to linearly combine the PRS based on the posterior effect-sizes obtained from the EUR and the target population with the weights in the linear combination being estimated based on the tuning dataset of the target population. We implemented PRS-CSx following the guidance provided in https://github.com/getian107/PRScsx. We set the hyperparameters a and b in the gamma-gamma prior to their default values of 1 and 0.5, respectively. Further, the parameter ϕ is varied over the default set of values 10^{-6}, 10^{-4}, 10^{-2}, and 1. The optimal ϕ is determined based on the performance on the tuning dataset.

Runtimes and memory usage. The computation time and memory usage of CT-SLEB (two ancestries), CT-SLEB (five ancestries), and PRS-CSx are compared based on their performance on chromosome 22 and assuming AFR is the target population. All analyses are performed using a single core with Intel E5-26840v4 CPU. The reported performance is averaged over 100 replicates. The training dataset includes GWAS summary statistics for AFR (N_{GWAS}=15,000) and for EUR (N_{GWAS}=100,000) population. The tuning dataset and validation dataset each contains 10,000 subjects. For five

ancestries analyses, the training GWAS sample sizes for AMR, EAS and SAS are all set to 15,000.

23andMe Data analysis. The individuals included in our analyses are part of the 23andMe participant cohort. All these individuals included have provided informed consent and answered surveys online according to our human subject protocol reviewed and approved by Ethical & Independent Review Services, a private institutional review board (http://www.eandireview.com). Detailed information about genotyping, quality control, imputation, removing related individuals, and ancestry determination is provided in **Supplementary Note**. Participants were included in the analysis based on consent status as checked when data analyses were initiated.

The analyses include five ancestries: African American, East Asian, European, Latino, and South Asian. Meanwhile, the analyses include two continuous and five binary traits: 1. Heart metabolic disease burden 2. Height 3. Any CVD 4. Depression 5. Migraine Diagnosis 6. Morning Person 7. SBMN. We randomly split the data for each population into training, tuning, and validation datasets with the proportion of 70%, 20%, and 10% (**Supplementary Table 1**). We perform GWAS for the seven traits using the training dataset for each population, adjusting for PC 1-5, sex, and age using standard quality control procedures (**Supplementary Note**). SNPs with MAF > 0.01 in at least one of the five populations are kept in the analyses. We further restrict analyses to SNPs that are on HM3 + MEGA chips with ~2.8 million SNPs (**Supplementary Table 3**). We use LDSC version 1.01[53] to estimate the heritability using the GWAS summary statistics of European populations for the seven traits. We estimate the LD score using the 503 unrelated samples of EUR ancestry from 1000 Genomes Project. We restrict heritability analyses to EUR populations since some non-EUR populations don't have sufficient sample size to get stable estimate from LD-score regression.

We apply seven methods to compare PRS prediction performance, CT, LDPred2, best EUR PRS based on CT and LDpred2, weighted-PRS, PRS-CSx, CT-SLEB using data from EUR and the target population, and CT-SLEB using all five populations. Since

individual-level data is not available in the training step, we use the reference data from the 1000 Genomes Project (Phase 3) to estimate the LD for each population. Specifically, we use AFR and AMR from the 1000 Genomes Project as the reference for the AA and Latino population in 23andMe, respectively. All PRS prediction performances are reported based on the independent validation dataset that is independent of the training and tuning datasets. To calculate the adjusted R^2 for continuous traits, we first regress the traits on covariates and then evaluate performance for the PRS to predict residualized trait values. To calculate the adjusted AUC for binary traits, we used the roc.binary function in the R package RISCA version 0.9[64].

Figure 1: CT-SLEB Workflow. The method contains three major steps: 1. Two-dimensional clumping and thresholding method for selecting SNPs (**Figure 1a**) ; 2. Empirical-Bayes procedure for utilizing correlation in effect sizes of genetic variants across populations (**Figure 1b**); 3. Super-learning model for combing the PRSs derived from the first two steps under different tuning parameters (**Figure 1c**). The GWAS summary-statistics data are obtained from the training data. The tuning dataset is used to train the super learning model. The final prediction performance is evaluated based on an independent validation dataset.

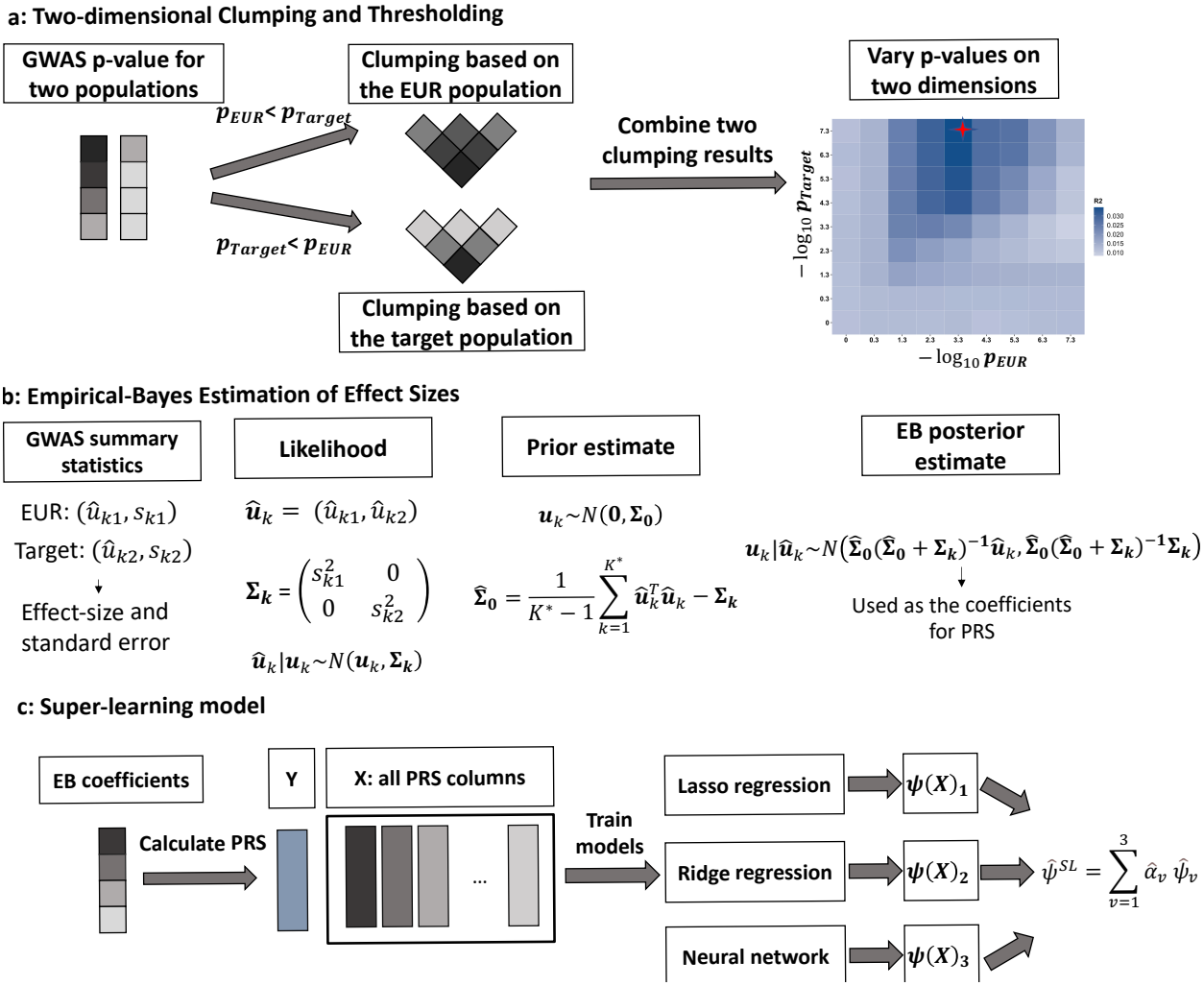

Figure 2: Simulation results showing performances of different methods for generating PRS in the multi-ancestry setting. The training sample size for each of the four non-EUR populations is 15,000 (**Figure 2a**) or 80,000 (**Figure 2b**). The training sample size for the EUR population is fixed at 100,000. The sample size for the tuning dataset of each population is fixed at 10,000. Prediction R^2 values are reported based on an independent validation dataset with 10,000 subjects for each population. Common SNP heritability is assumed to be 0.4 across all populations, and effect-size correlation is assumed to be 0.8 across all pairs of populations. The causal SNPs proportion (degree of polygenicity) is varied across 0.01, 0.001, 5×10^{-4} ($N_{causal} = 192K, 19.2K, 9.6K$), and effect sizes for causal variants are assumed to be related to allele frequency under a strong negative selection model. All data are generated based on ~19 million common SNPs across the five populations, but analyses are restricted to ~2.8 million SNPs that are used on Hapmap3 + Multi-Ethnic Genotyping Arrays chip (PRS-CSx analysis is further restricted to ~1.3 million HM3 SNPs).

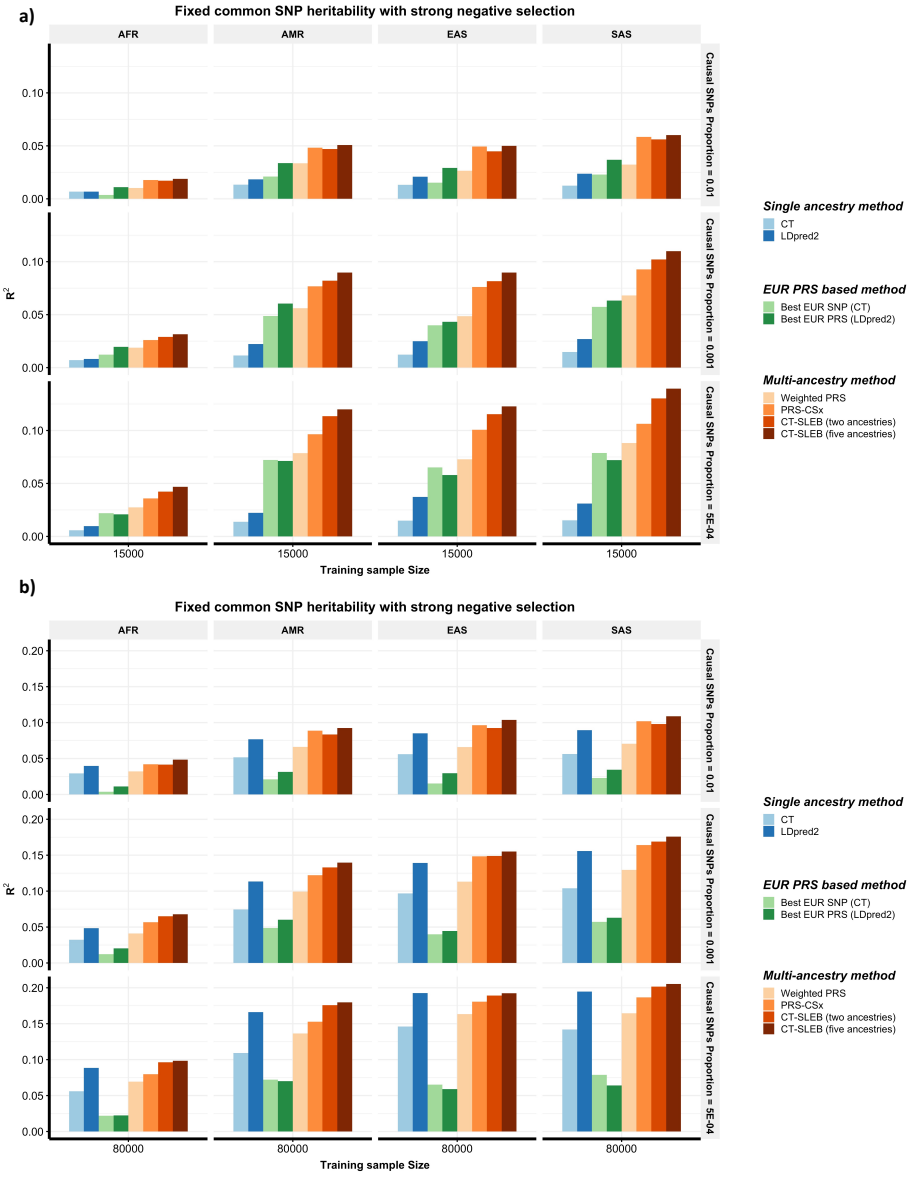

Figure 3: Prediction performance of CT-SLEB PRS across different ancestries relative to single ancestry EUR PRS in the EUR population. The training sample size for each of the four non-EUR populations is 15,000, 45000, 80,000, or 100,000. The training sample size for the EUR population is fixed at 100,000, and PRS performance is assessed using single ancestry CT or LDpred2, whichever performs the best in each setting. Two different models for genetic architectures are considered where either the common SNP heritability is fixed (at 0.4) (**Figure 3a and 3b**) or per-SNP heritability is fixed (Figure 3c and 3d) across the five populations (**Figure 3c and 3d**). The effect-size correlation is assumed to be 0.8 across all pairs of populations. The effect sizes for causal variants are assumed to be related to allele frequency under a strong negative selection model.

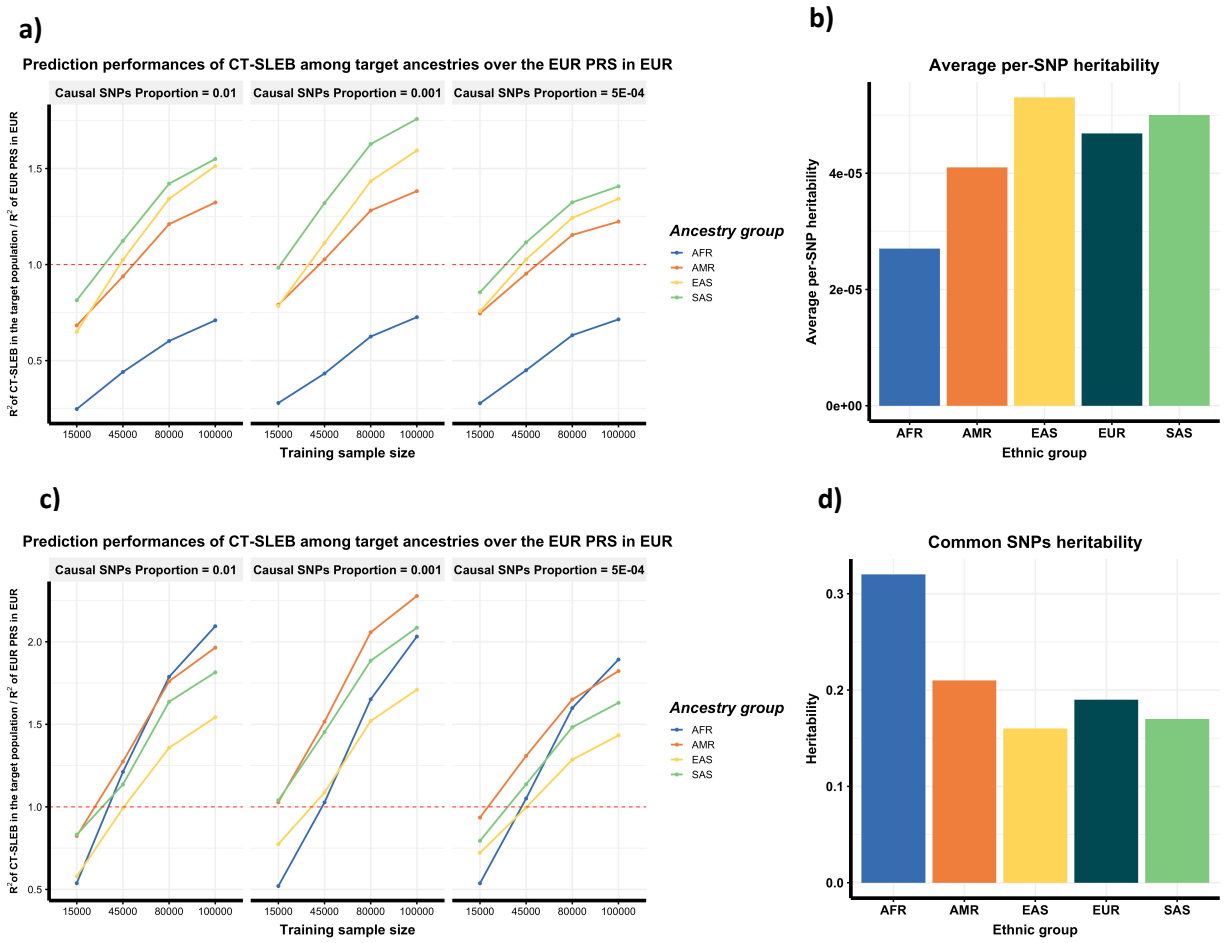

Figure 4: Prediction performance of CT-SLEB PRS under different SNP density. Analysis of each simulated data based on ~19 million SNPs are restricted to three different SNP sets Hapmap3 (~1.3 million SNPs), Hapmap3 + Multi-Ethnic Genotyping Arrays (~2.8 million SNPs), 1000 Genomes Project (~19 million SNPs). The training sample size for each of the four non-EUR populations is 15,000 (**Figure 4a**) or 80,000 (**Figure 4b**). The training sample size for the EUR population is fixed at 100,000. Prediction R^2 values are reported based on independent validation dataset with 10,000 subjects for each population. Common SNP heritability is assumed to be 0.4 across all populations and effect-size correlation is assumed to be 0.8 across all pairs of populations. The causal SNPs proportion are varied across 0.01, 0.001, 5×10^{-4} ($N_{causal} = 192K, 19.2K, 9.6K$) and effect sizes for causal variants are assumed to be related to allele frequency under a strong negative selection model.

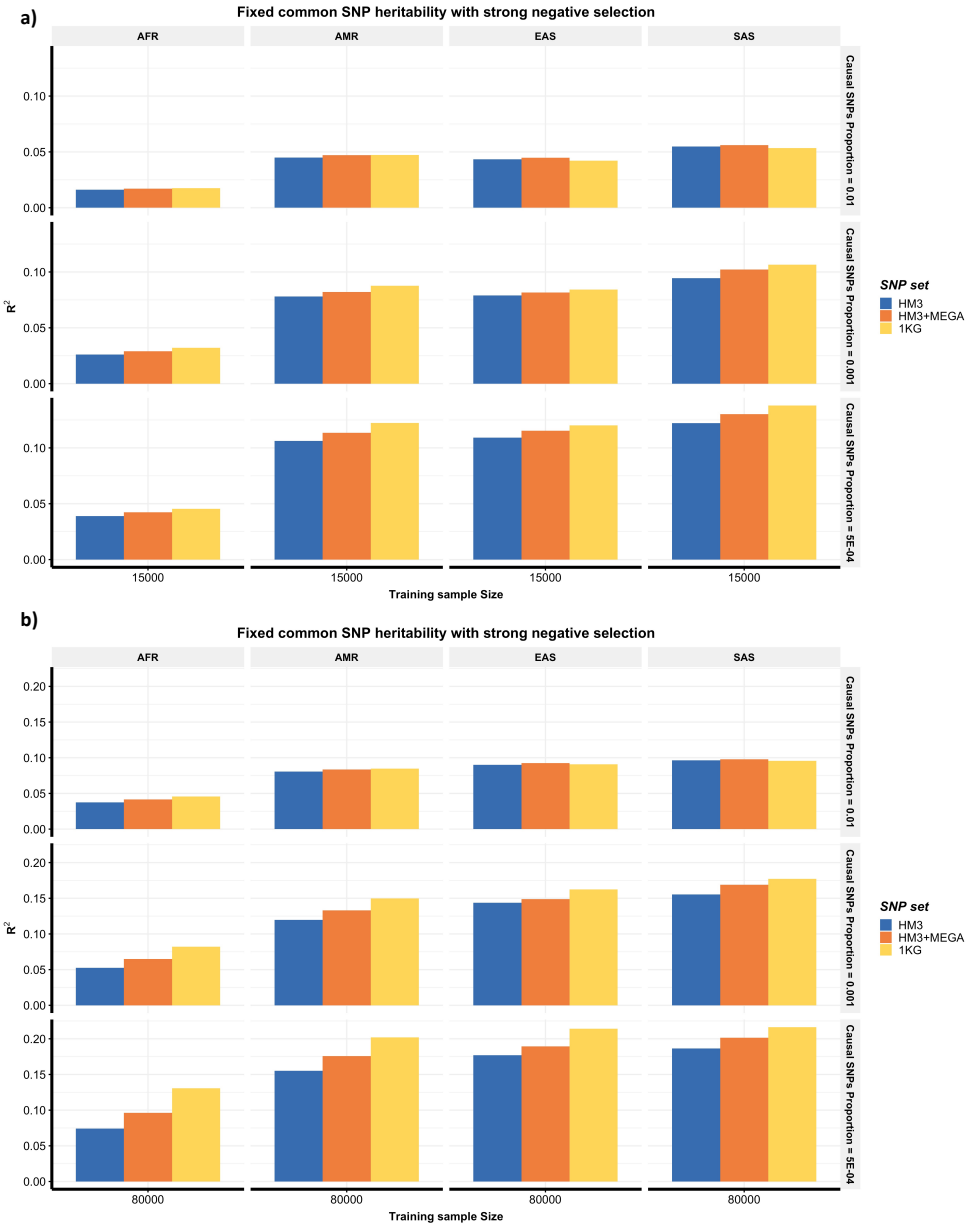

Figure 5: Prediction accuracy of PRS for heart metabolic disease burden and height in 23andMe, Inc. datasets. The total sample size for heart metabolic disease burden and height is, respectively, 2.46 million and 2.93 million for European, 131K and 141K for African American, 375K and 509K for Latino, 110K and 121K for East Asian, and 29K and 32K for South Asian. The dataset is randomly split into 70%, 20%, 10% for training, tuning and validation dataset, respectively. The prediction R^2 values are reported based on the performance of the PRS in the validation dataset. The red dashed line represents the prediction performance of EUR PRS generated using single ancestry method (best of CT or LDpred2) in the EUR population. Analyses are restricted to ~2.8 million SNPs that are included in Hapmap3, or the Multi-Ethnic Genotyping Arrays chips array or both. However, PRS-CSx is further restricted to ~1.3 million Hapmap3 SNPs as has been implemented in the provided software.

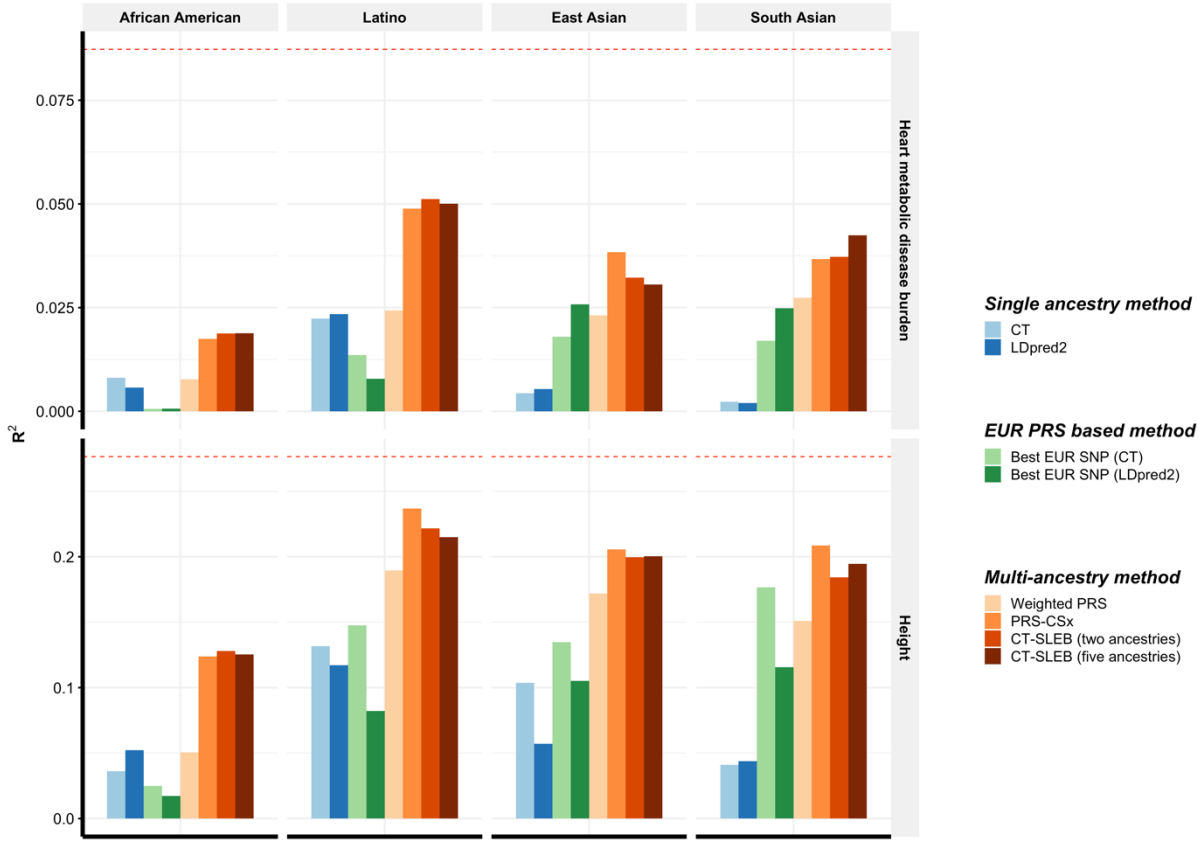

Figure 6: Prediction accuracy of five binary traits in 23andMe, Inc datasets. We used the data from five populations: EUR (averaged N ≈ 2.37 million), African American (averaged N ≈ 109K), Latino (averaged N ≈ 401K), EAS (averaged N≈ 86K), SAS (averaged N≈ 24K). The datasets are randomly split into 70%, 20%, 10% for training, tuning and validation dataset, respectively. The AUC values are reported based on the validation dataset. The red dashed line represents the prediction performance of EUR PRS generated using single ancestry method (best of CT or LDpred2) in the EUR population. Analyses are restricted to ~2.8 million SNPs that are included in Hapmap3, or the Multi-Ethnic Genotyping Arrays chips array or both. However, PRS-CSx is further restricted to ~1.3 million Hapmap3 SNPs as has been implemented in the provided software.