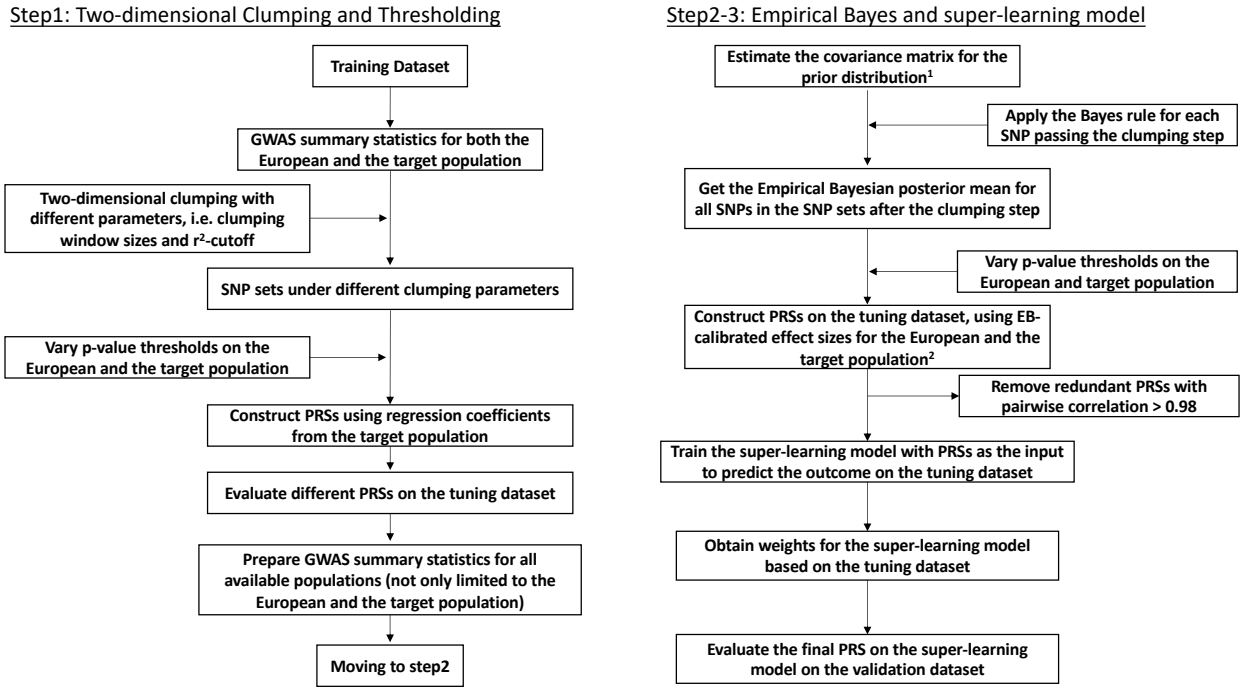
Supplementary figure 1: CT-SLEB detailed flowchart. The method contains three major steps: 1. Two-dimensional clumping and thresholding; 2. Empirical-Bayes procedure for utilizing genetic correlations of effect sizes across populations; 3. Super-learning model for combining PRSs under different tuning parameters. The tuning dataset is used to train the super learning model. The final prediction performance is evaluated based on an independent validation dataset. For continuous traits, the prediction is evaluated using R^2 obtained from the linear regression between outcome and PRS after adjusting for covariates (**Methods**). For binary traits, the prediction is evaluated using area under the ROC curve (AUC).

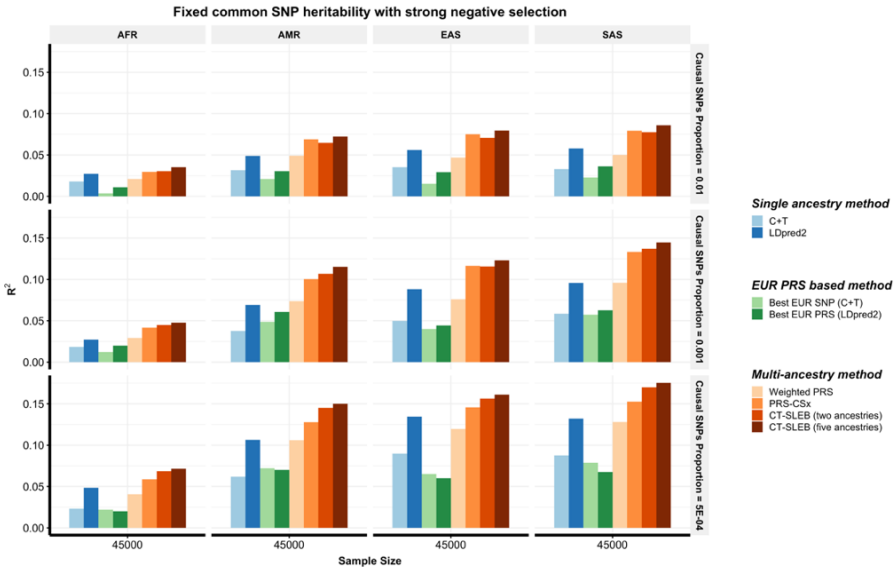


[1] The prior distribution is only estimated once based on the SNP set that gives the best PRS in the CT step across all different p-value thresholds, r^2-cutoff and window sizes

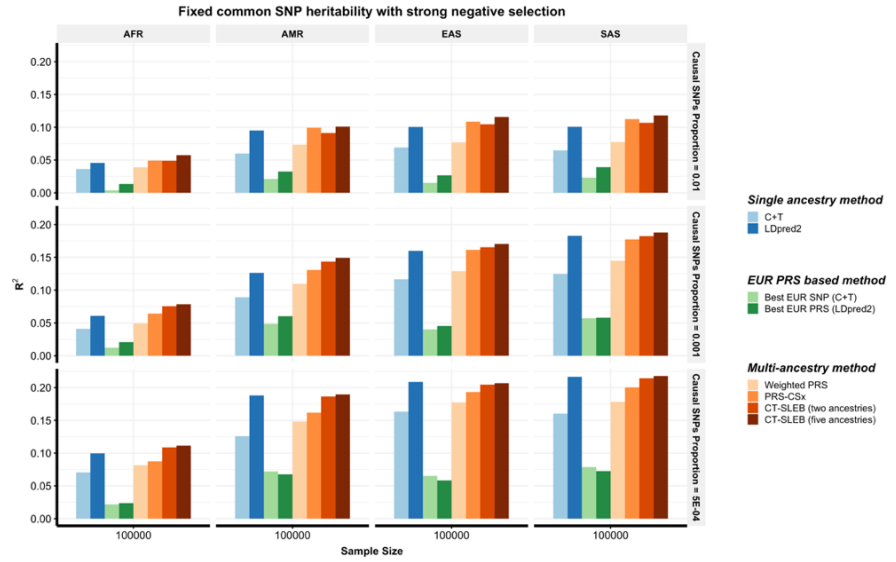
[2] When more than two ancestries are involved, we use data from all populations to derive the EB estimates of effect-sizes for SNPs for each population. However, to save computational time at the super-learning step, we derive the final PRS for a target population by only incorporating the initial PRSs derived for the larger EUR population and those for the specific target population.

Supplementary figure 2: Simulation results showing performances of different methods for generating PRS in the multi-ancestry setting under a strong negative selection model. The training sample size for each of the four non-EUR populations is 45,000 (**Sup. Fig. 2a**) or 100,000 (**Sup. fig. 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 is 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 are varied across 0.01, 0.001, 5×10^{-4}. 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 (HM3) + Multi-Ethnic Genotyping Arrays (MEGA) chip (PRS-CSx analysis is further restricted to ~1.3 million HM3 SNPs).

a)

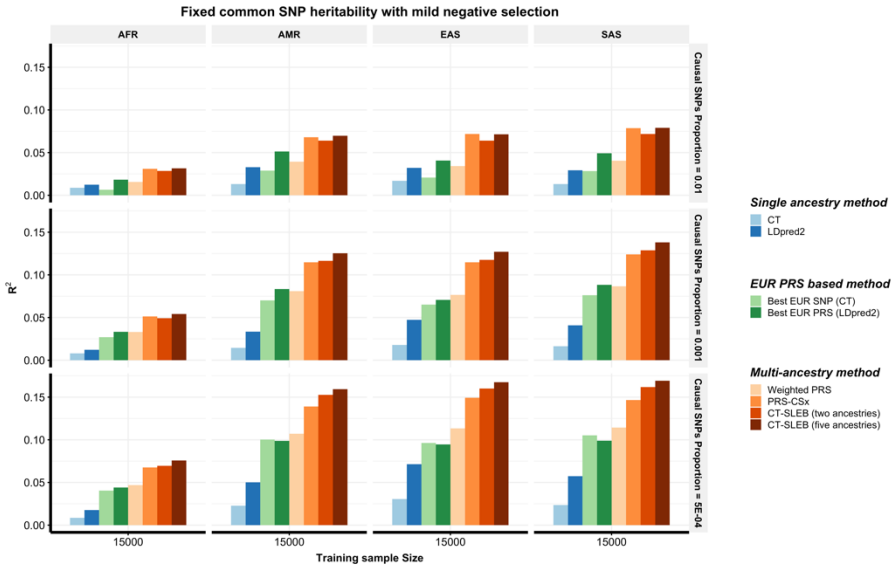


b)

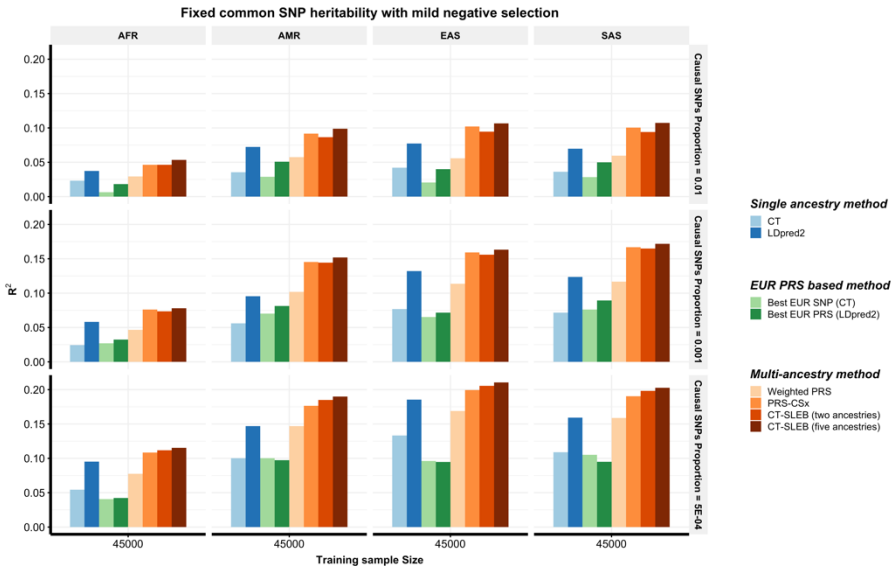


Supplementary figure 3: Simulation results showing performances of different methods for generating PRS in the multi-ancestry setting under a mild negative selection model. The training sample size for each of the four non-EUR populations is a) 15,000 b) 45,000 c) 80,000 d) 100,000. The training sample size for the EUR population is fixed at 100,000. The sample size for the tuning dataset for each population is fixed at 10,000. Prediction R^2 is 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 are varied across 0.01, 0.001, 5×10^{-4} 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 HM3 + MEGA chip (PRS-CSx analysis is further restricted to ~1.3 million HM3 SNPs).

a)

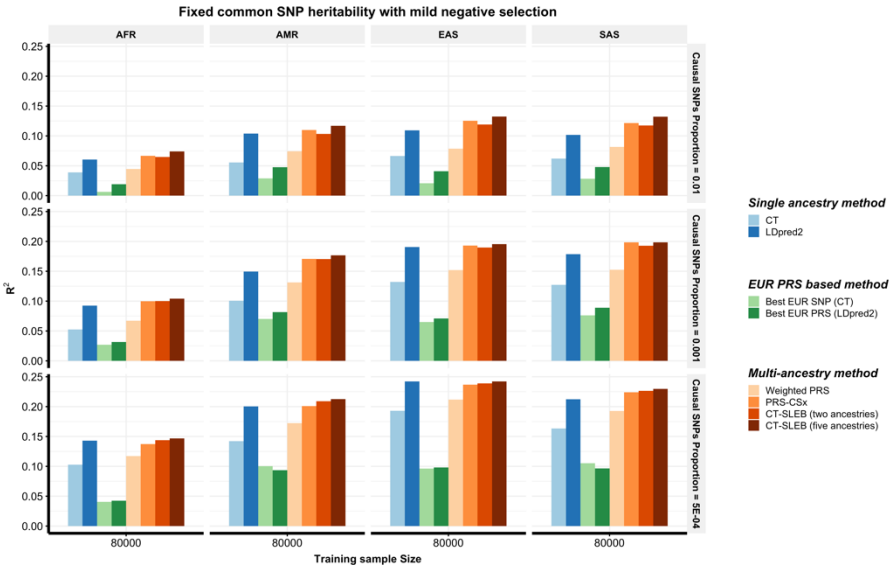


b)

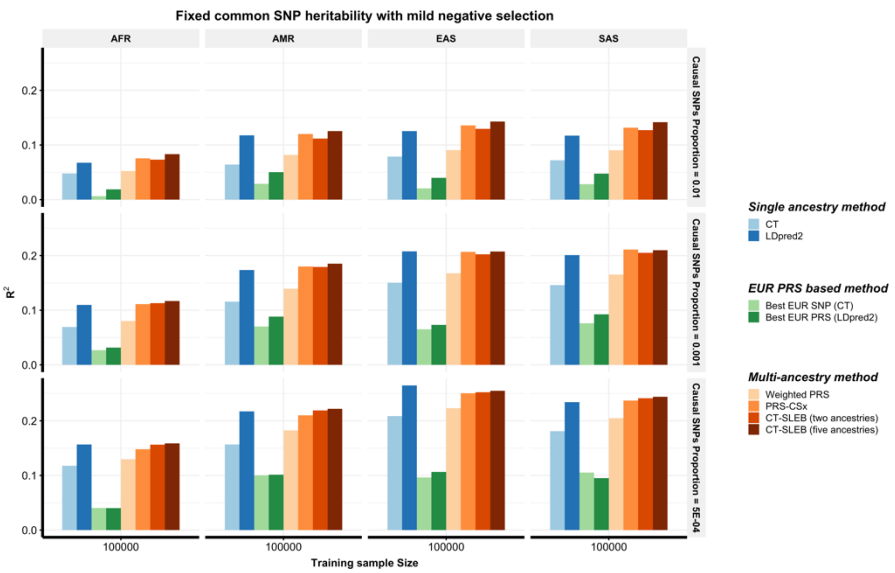


Supplementary figure 3 continued: Simulation results showing performances of different methods for generating PRS in the multi-ancestry setting under a mild negative selection model. The training sample size for each of the four non-EUR populations is a) 15,000 b) 45,000 c) 80,000 d) 100,000. The training sample size for the EUR population is fixed at 100,000. The sample size for the tuning dataset for each population is fixed at 10,000. Prediction R^2 is 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 are varied across 0.01, 0.001, 5×10^{-4}. 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 HM3 + MEGA chip (PRS-CSx analysis is further restricted to ~1.3 million HM3 SNPs).

c)

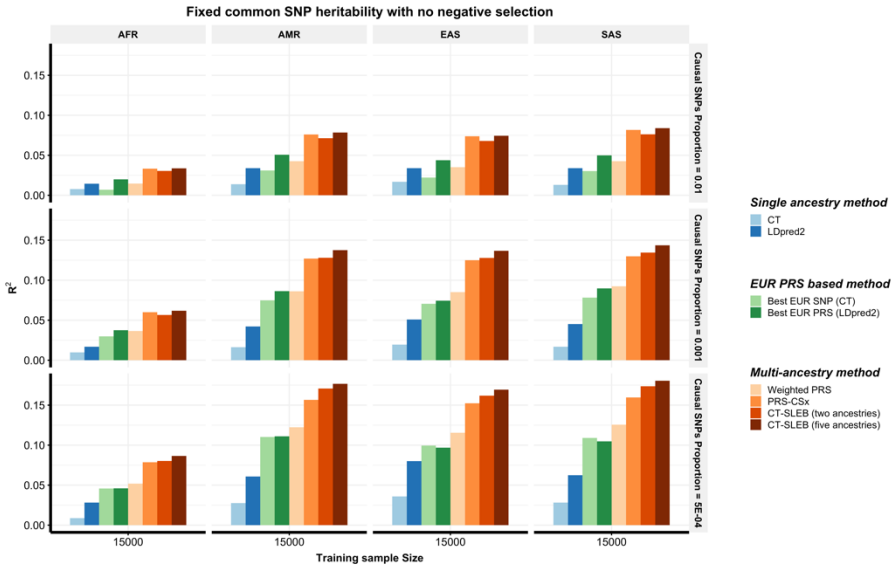


d)

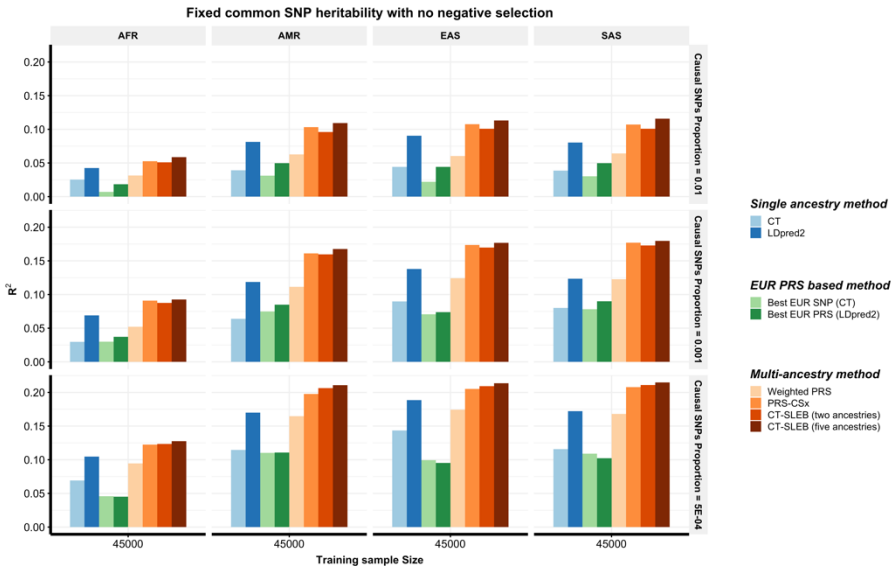


Supplementary figure 4: Simulation results showing performances of different methods for generating PRS in the multi-ancestry setting under a no negative selection model. The training sample size for each of the four non-EUR populations is a) 15,000 b) 45,000 c) 80,000 d) 100,000. The training sample size for the EUR population is fixed at 100,000 The sample size for the tuning dataset for each population is fixed at 10,000. Prediction R^2 is 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 are varied across 0.01, 0.001, 5×10^{-4}. 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 HM3 + MEGA chip (PRS-CSx analysis is further restricted to ~1.3 million HM3 SNPs).

a)

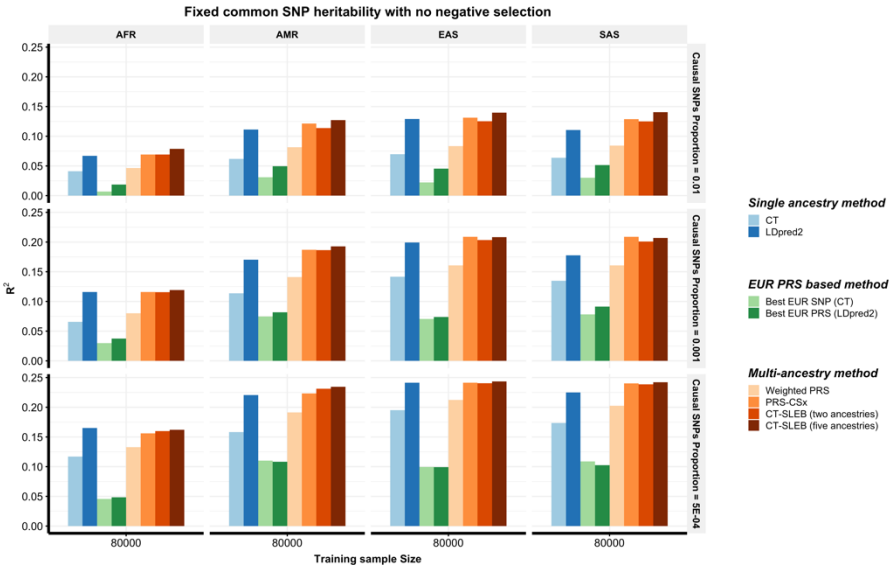


b)

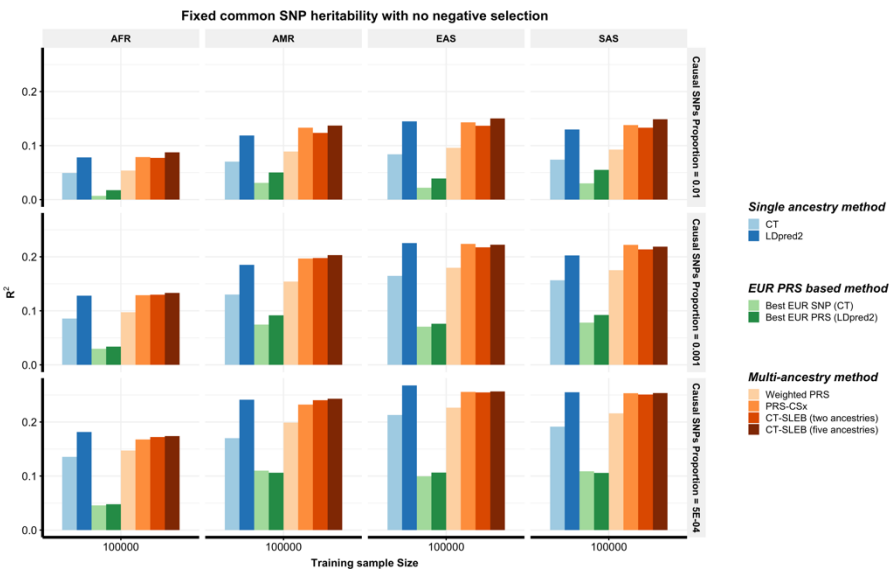


Supplementary figure 4 continued: Simulation results showing performances of different methods for generating PRS in the multi-ancestry setting under a no negative selection model. The training sample size for each of the four non-EUR populations is a) 15,000 b) 45,000 c) 80,000 d) 100,000. The training sample size for the EUR population is fixed at 100,000. The sample size for the tuning dataset for each population is fixed at 10,000. Prediction R^2 is 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 are varied across 0.01, 0.001, 5×10^{-4}. 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 HM3 + MEGA chip (PRS-CSx analysis is further restricted to ~1.3 million HM3 SNPs).

c)

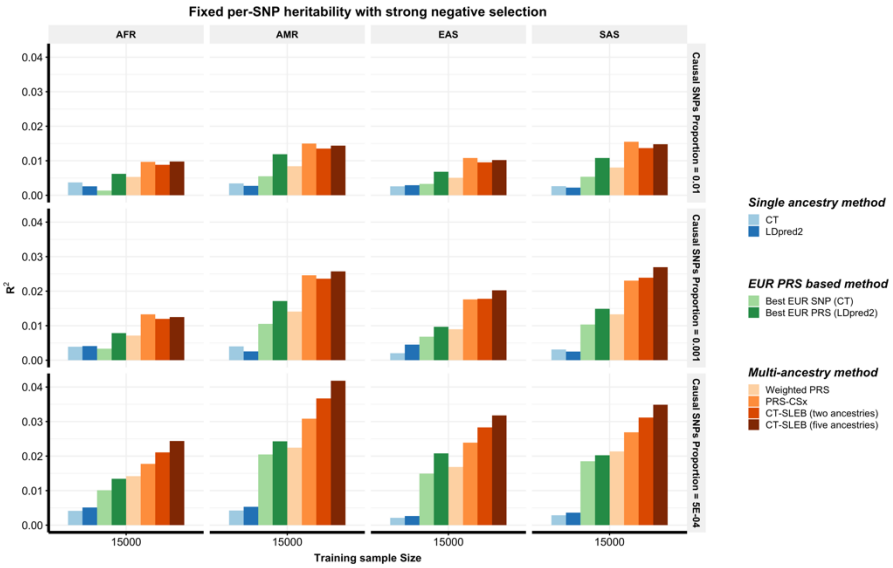


d)

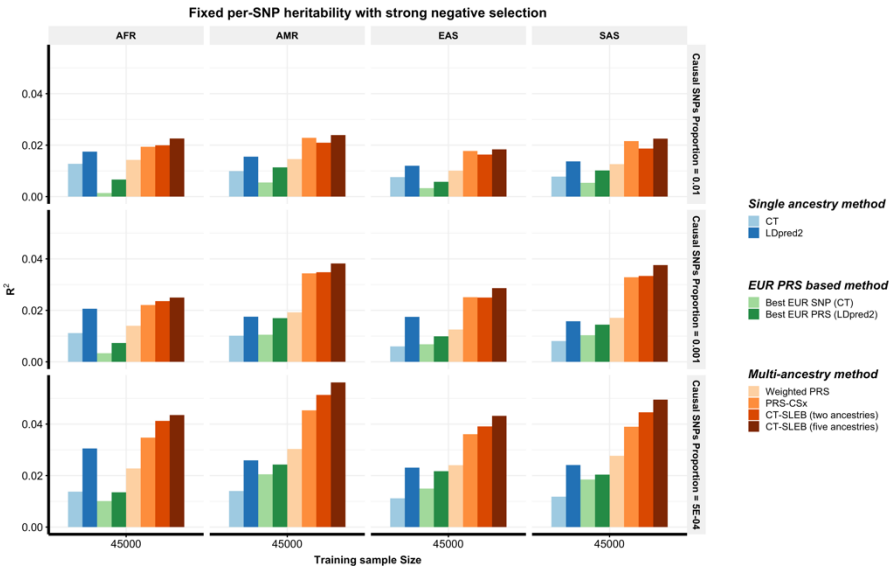


Supplementary figure 5: Simulation results showing different performances of different methods for generating PRS in the multi-ancestry setting under a strong negative selection model with per-SNP heritability fixed across all populations. The training sample size for each of the four non-EUR populations is a) 15,000 b) 45,000 c) 80,000 d) 100,000. The training sample size for the EUR population is fixed at 100,000. The sample size for the tuning dataset for each population is fixed at 10,000. Prediction R^2 is reported based on an independent validation dataset with 10,000 subjects for each population. The 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}. 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 HM3 + MEGA chip (PRS-CSx analysis is further restricted to ~1.3 million HM3 SNPs).

a)

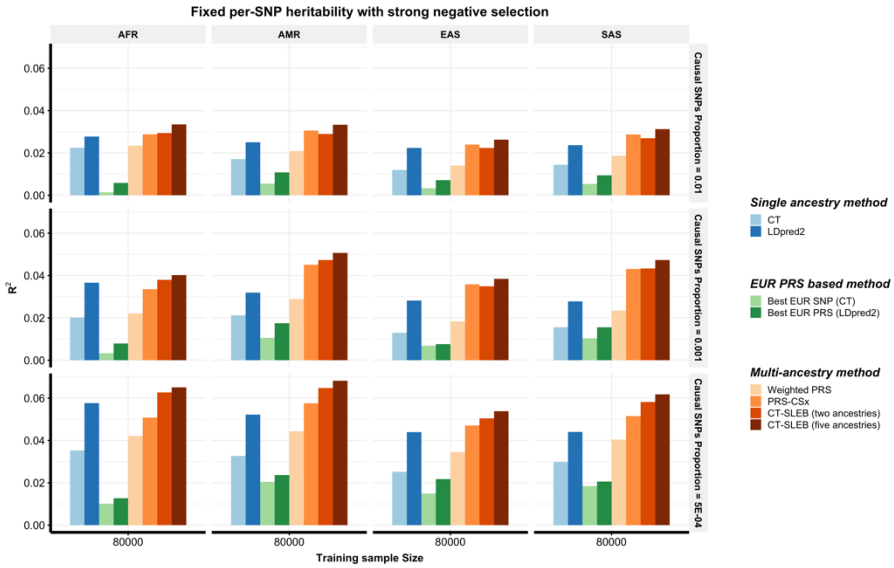


b)

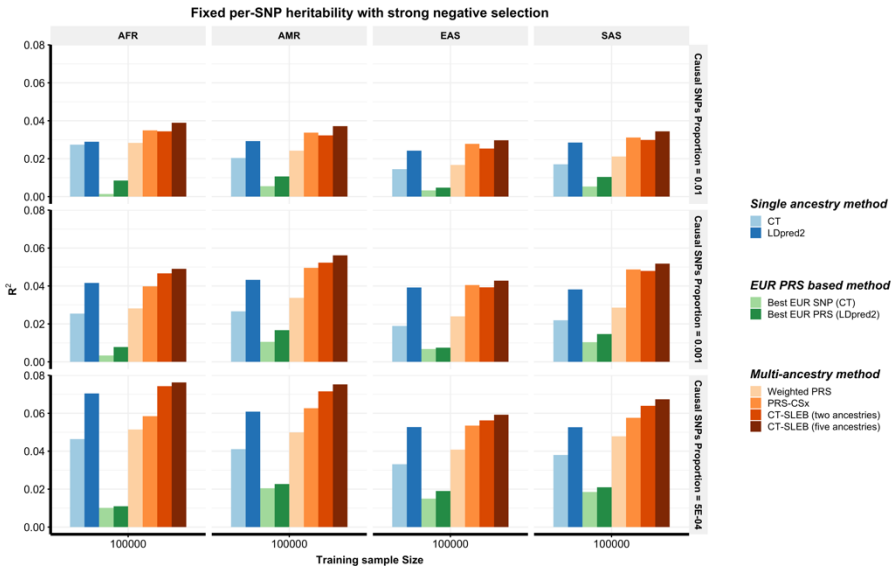


Supplementary figure 5 continued: Simulation results showing performances of different methods for generating PRS in the multi-ancestry setting under a strong negative selection model with per-SNP heritability fixed across all populations. The training sample size for each of the four non-EUR populations is a) 15,000 b) 45,000 c) 80,000 d) 100,000. The training sample size for the EUR population is fixed at 100,000. The sample size for the tuning dataset for each population is fixed at 10,000. Prediction R^2 is reported based on an independent validation dataset with 10,000 subjects for each population. The 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}. 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 HM3 + MEGA chip (PRS-CSx analysis is further restricted to ~1.3 million HM3 SNPs).

c)

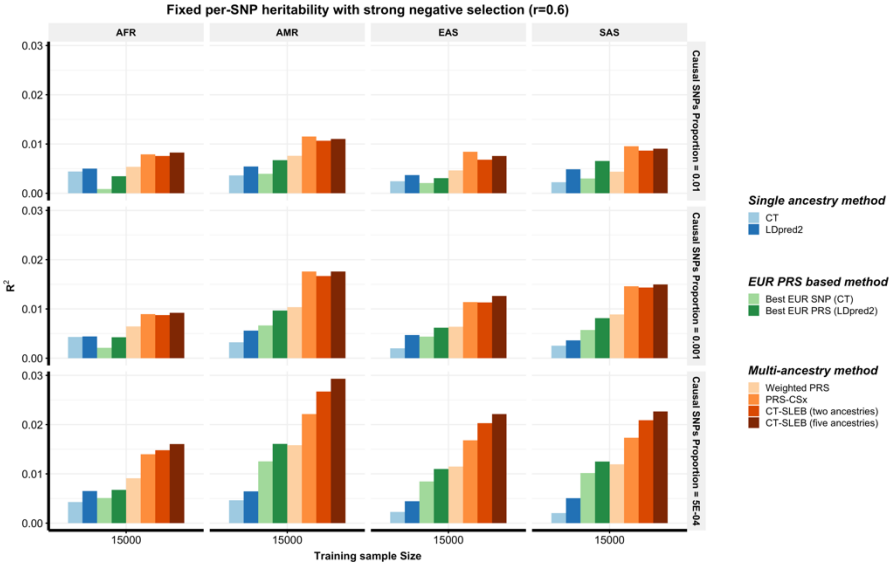


d)

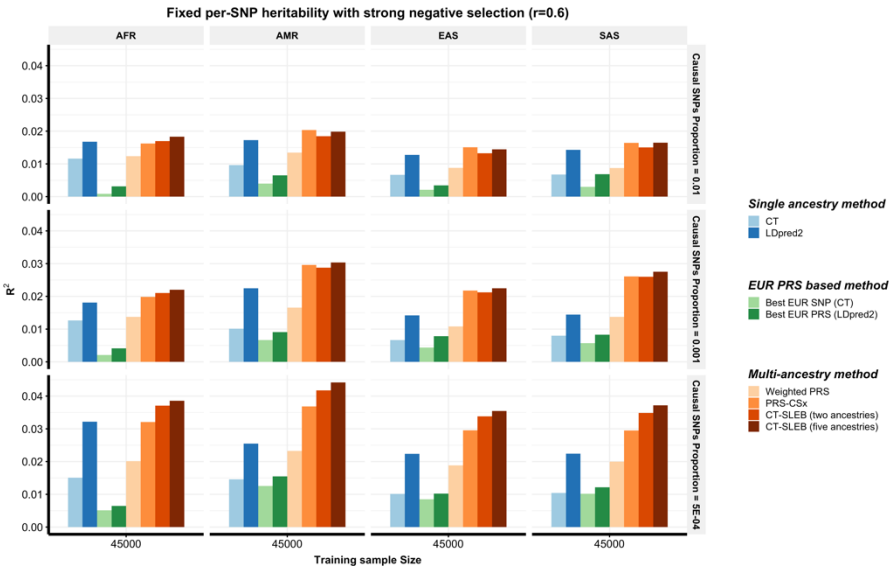


Supplementary figure 6: Simulation results showing performances of different methods for generating PRS in the multi-ancestry setting under a strong negative selection model with per-SNP heritability fixed across all populations (genetic correlation = 0.6). The training sample size for each of the four non-EUR populations is a) 15,000 b) 45,000 c) 80,000 d) 100,000. The training sample size for the EUR population is fixed at 100,000. The sample size for the tuning dataset for each population is fixed at 10,000. Prediction R^2 is reported based on an independent validation dataset with 10,000 subjects for each population. The effect-size correlation is assumed to be 0.6 across all pairs of populations. The causal SNPs proportion are varied across 0.01, 0.001, 5×10^{-4}. 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 HM3 + MEGA chip (PRS-CSx analysis is further restricted to ~1.3 million HM3 SNPs).

a)

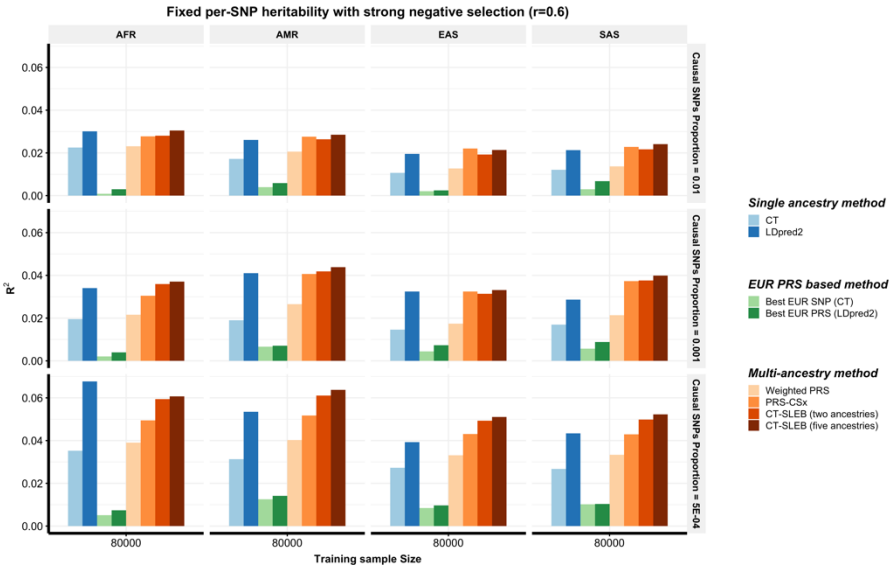


b)

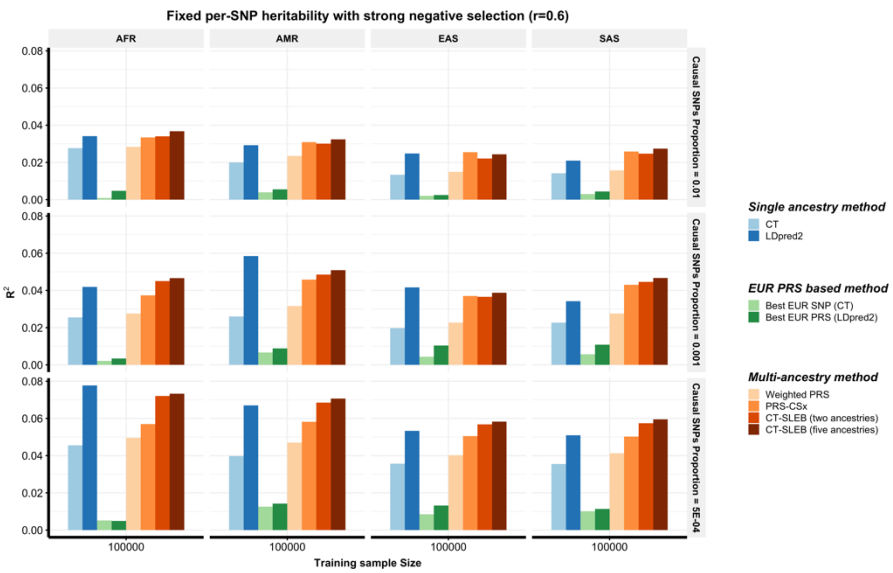


Supplementary figure 6 continued: Simulation results showing performances of different methods for generating PRS in the multi-ancestry setting under a strong negative selection model with per-SNP heritability fixed across all populations (genetic correlation = 0.6). The training sample size for each of the four non-EUR populations is a) 15,000 b) 45,000 c) 80,000 d) 100,000. The training sample size for the EUR population is fixed at 100,000. The sample size for the tuning dataset size for each population is fixed at 10,000. Prediction R^2 is reported based on an independent validation dataset with 10,000 subjects for each population. The effect-size correlation is assumed to be 0.6 across all pairs of populations. The causal SNPs proportion are varied across 0.01, 0.001, 5×10^{-4} . 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 HM3 + MEGA chip (PRS-CSx analysis is further restricted to ~1.3 million HM3 SNPs).

c)

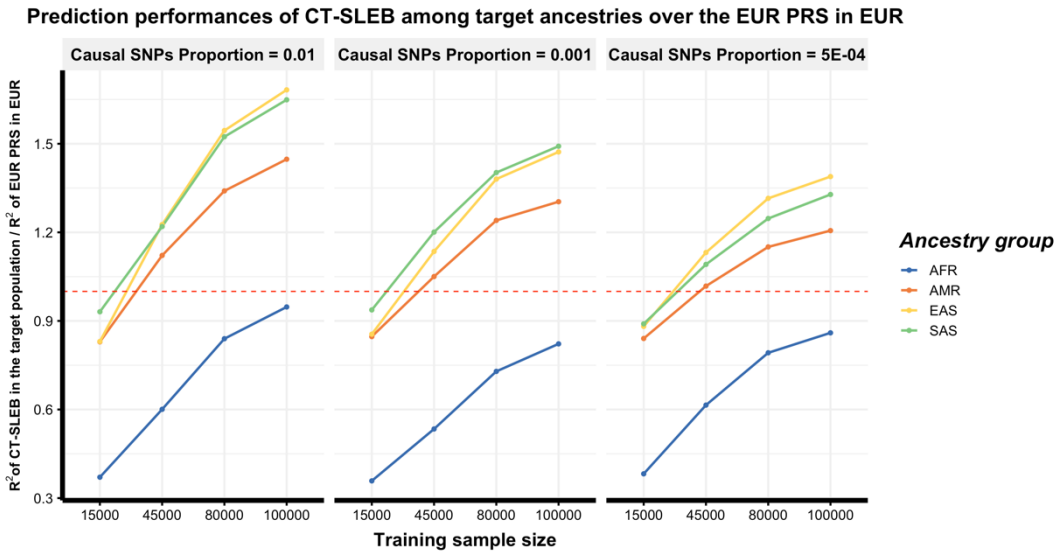


d)

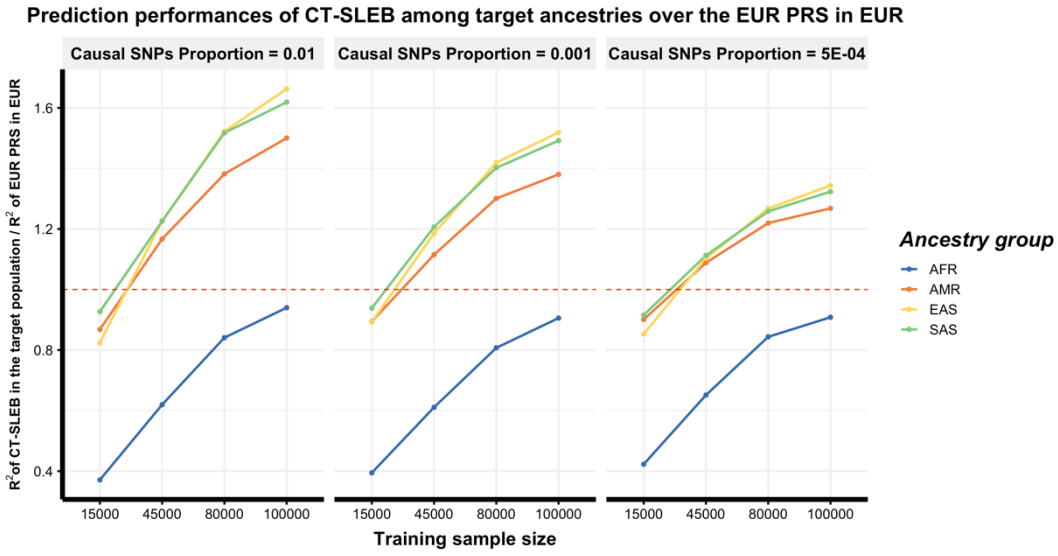


Supplementary figure 7: Prediction performance of CT-SLEB PRS across different ancestries relative to single ancestry EUR PRS in the EUR population. The training sample size for the EUR population is fixed at $100{,}000$, and PRS performance is assessed using single ancestry CT or LDpred2, whichever performs the best in each setting. Three different models for genetic architectures are considered: the common SNP heritability is fixed (at 0.4) with mild negative selection (Sup. Fig. 7a) and with no negative selection (Sup. Fig. 7b) and fixed per-SNP heritability with strong negative selection (Sup. Fig. 7c). The effect-size correlation is assumed to be 0.8 across all pairs of populations for settings in Sup. Fig. 7a and 7b. The effect-size correlation is assumed to be 0.6 for the setting in Sup. Fig. 7c.

a)

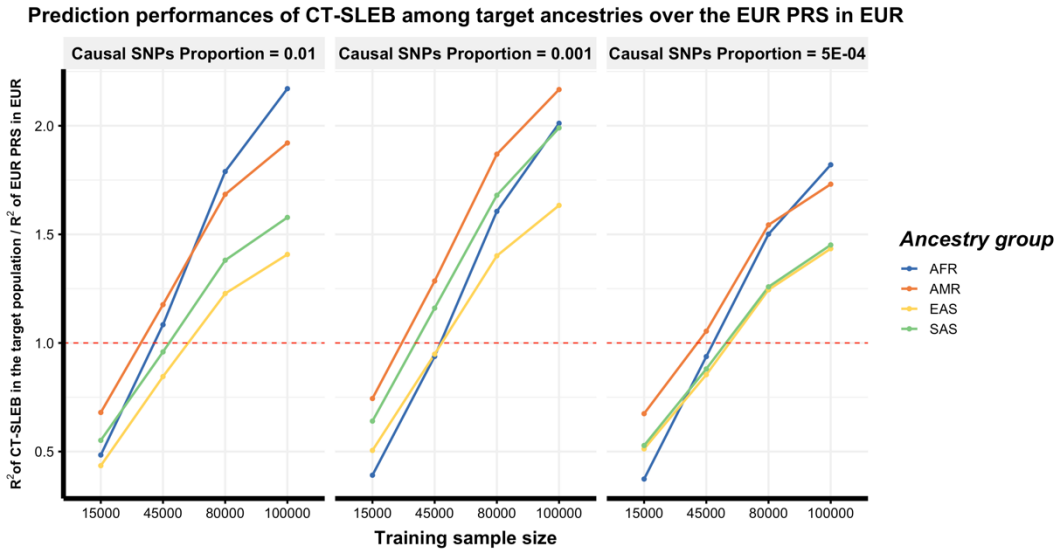


b)



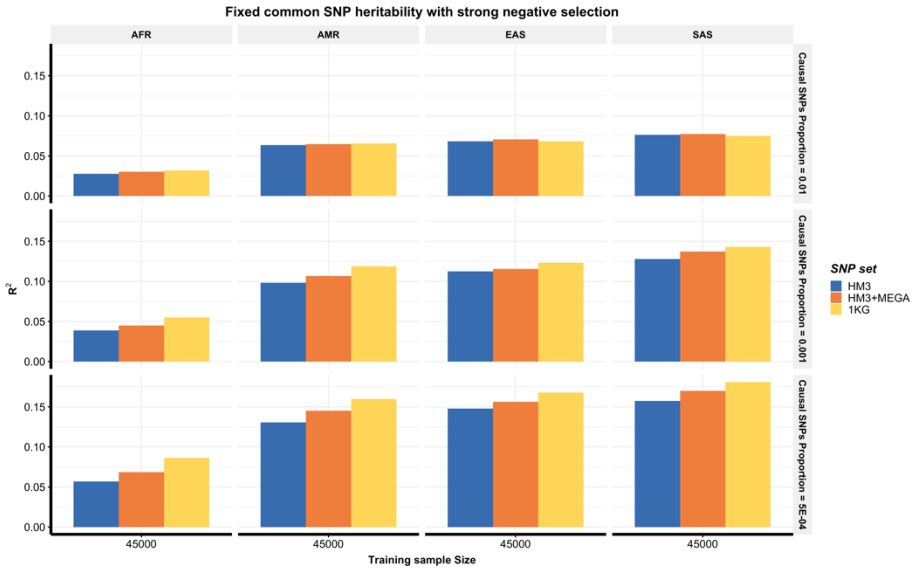
Supplementary figure 7 continued: Prediction performance of CT-SLEB PRS across different ancestries relative to single ancestry EUR PRS in the EUR population. The training sample size for the EUR population is fixed at 100,000, and PRS performance is assessed using single ancestry CT or LDpred2, whichever performs the best in each setting. Three different models for genetic architectures are considered: the common SNP heritability is fixed (at 0.4) with mild negative selection (Sup. Fig. 7a) and with no negative selection (Sup. Fig. 7b) and fixed per-SNP heritability with strong negative selection (Sup. Fig. 7c). The effect-size correlation is assumed to be 0.8 across all pairs of populations for settings in Sup. Fig. 7a and 7b. The effect-size correlation is assumed to be 0.6 for the setting in Sup. Fig. 7c.

c)

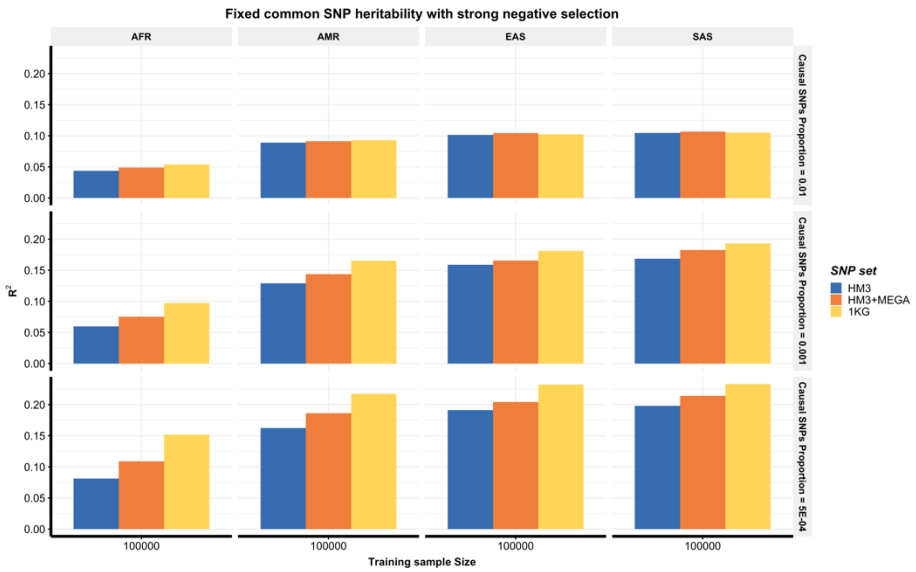


Supplementary figure 8: Prediction performance of CT-SLEB PRS under different SNP density. Analysis of each simulated data based on ~19 million SNPs are restricted to three different SNP sets Hapmap3 (~1.3 million SNPs), Hapmap3 + Multi-Ethnic Genotyping Arrays (~2.8 million SNPs), 1000 Genomes Project (~19 million SNPs). The training sample sizes for each of the four non-EUR populations is 45,000 (Sup. Fig. 8a) or 100,000 (Sup. Fig. 8b).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.

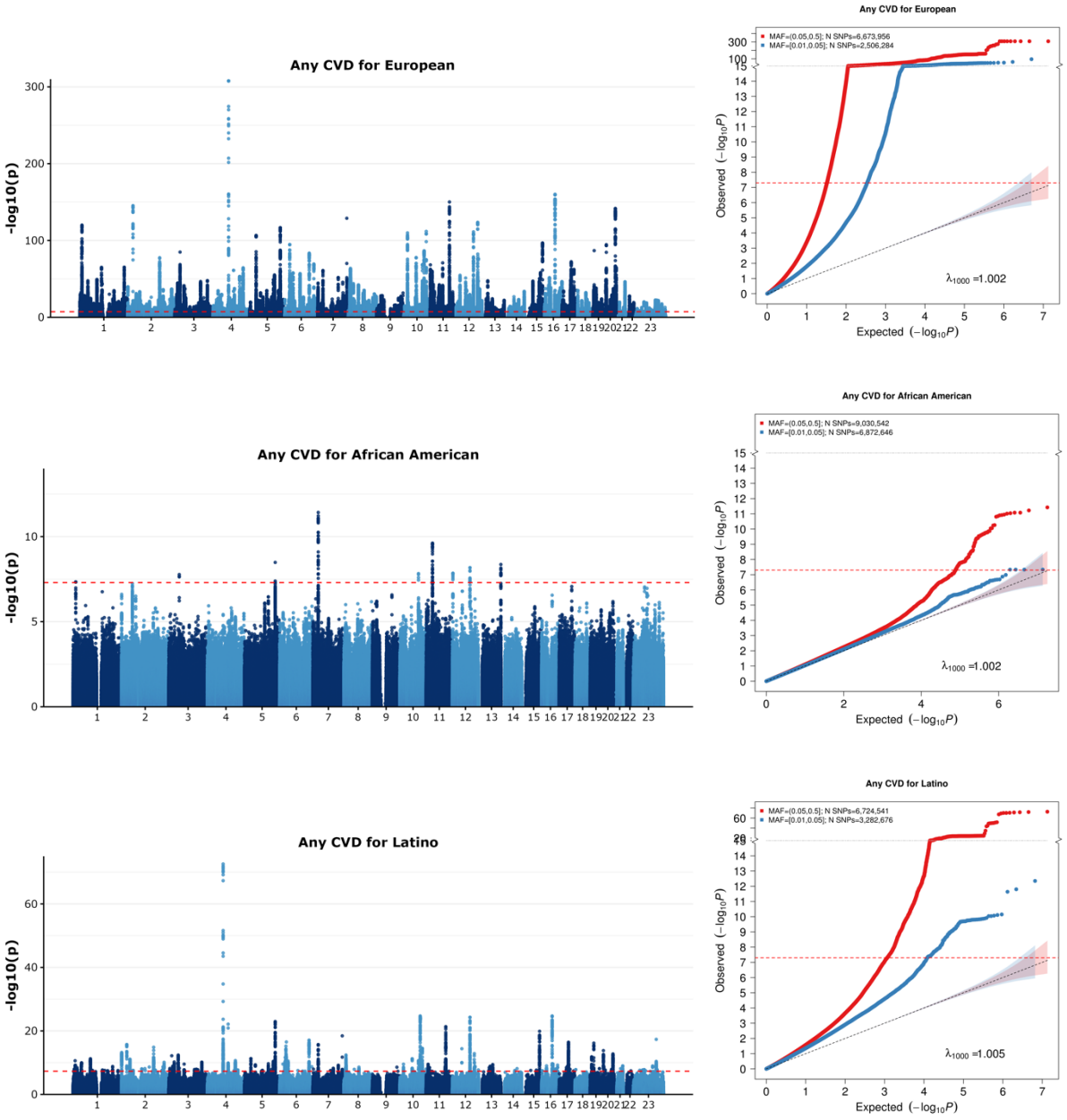
a)



b)

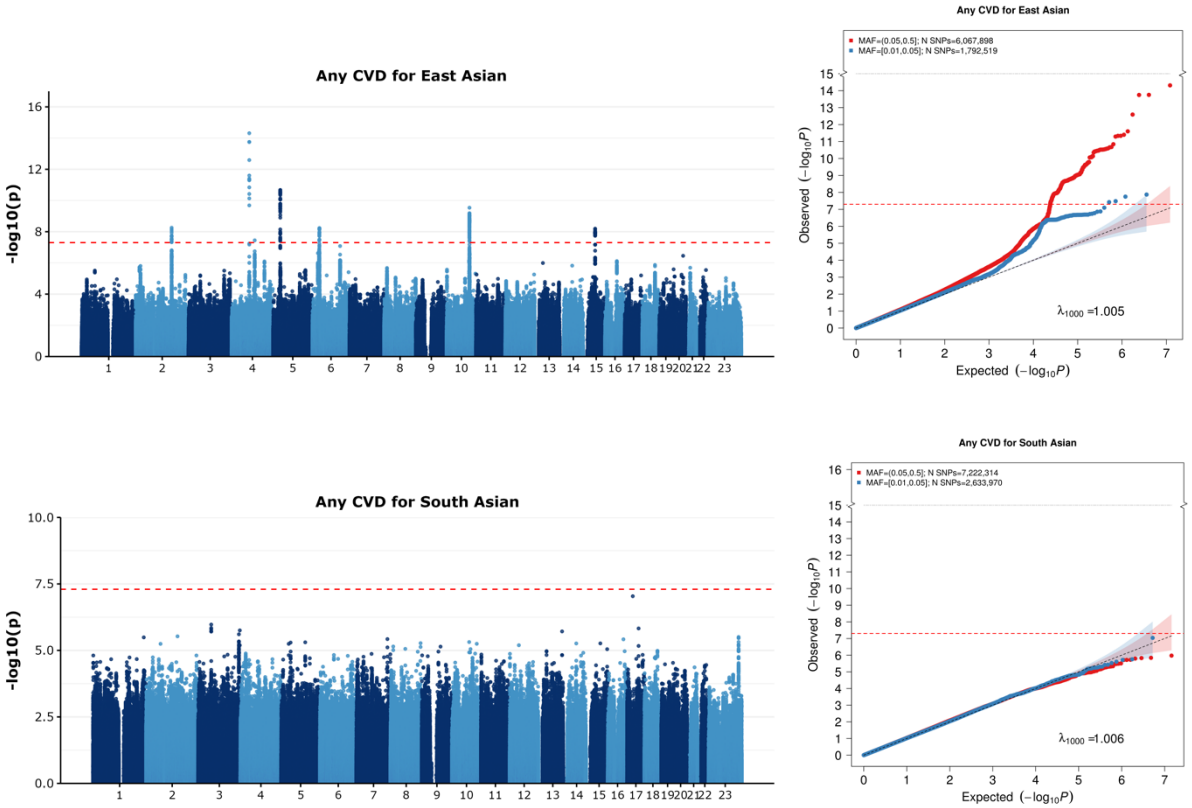


Supplementary figure 9: Manhattan plot and QQ plot[1] based on the 23andMe, Inc. GWAS summary statistics for any cardiovascular disease (any CVD) in five populations: European, African American, Latino, East Asian, South Asian.



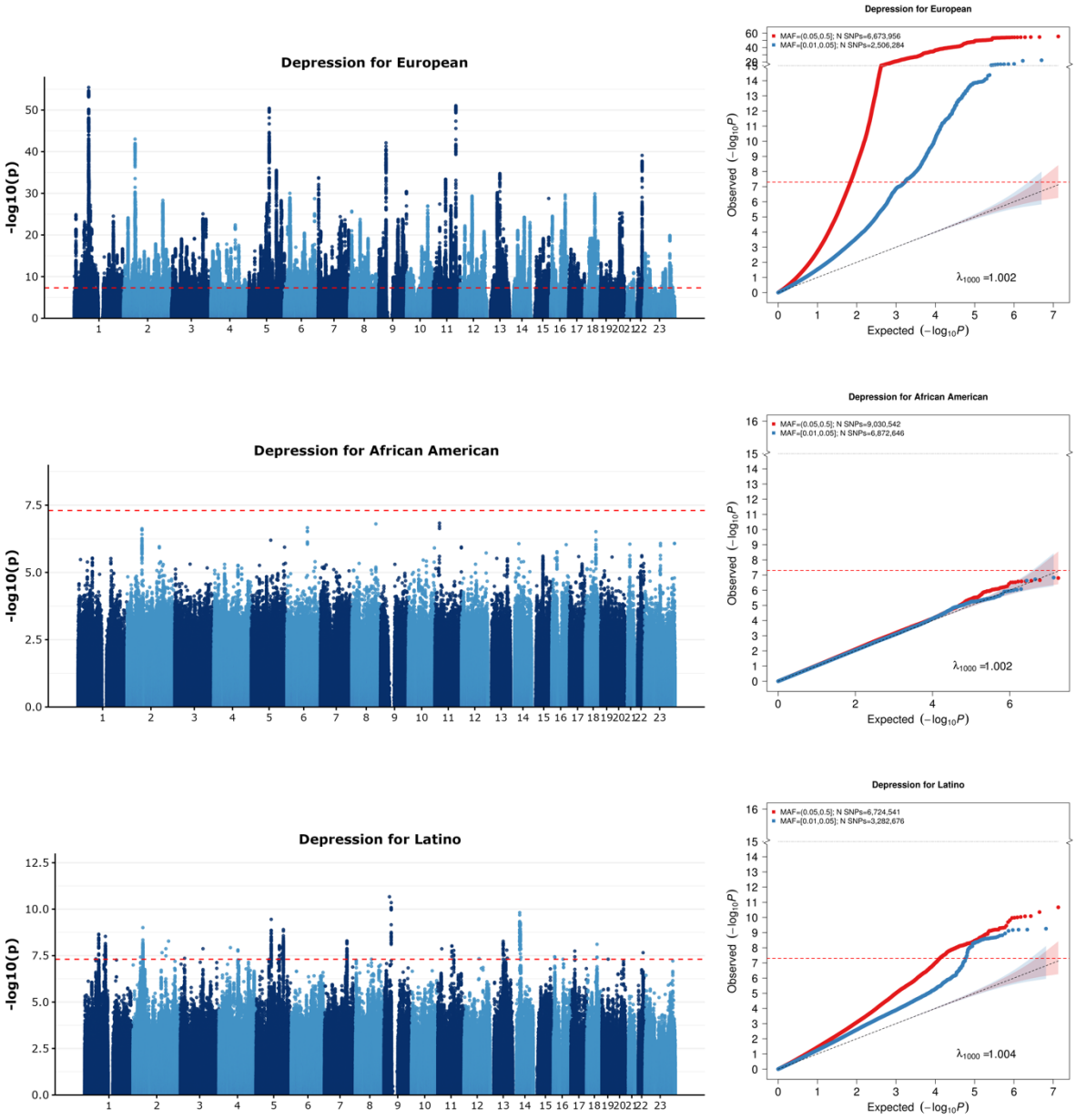
[1] For continuous traits, λ_{1000} scales the genomic inflation factor λ to a study with 1000 subjects using $\lambda_{1000} = 1 + 1000 * (\lambda - 1)/N$, where N is the total sample size. For binary traits, λ_{1000} scales λ to a study with 1000 cases and 1000 controls using $\lambda_{1000} = 1 +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

Supplementary figure 9 continued: Manhattan plot and QQ plot[1] based on the 23andMe, Inc. GWAS summary statistics for any cardiovascular disease (any CVD) in five populations: European, African American, Latino, East Asian, South Asian.



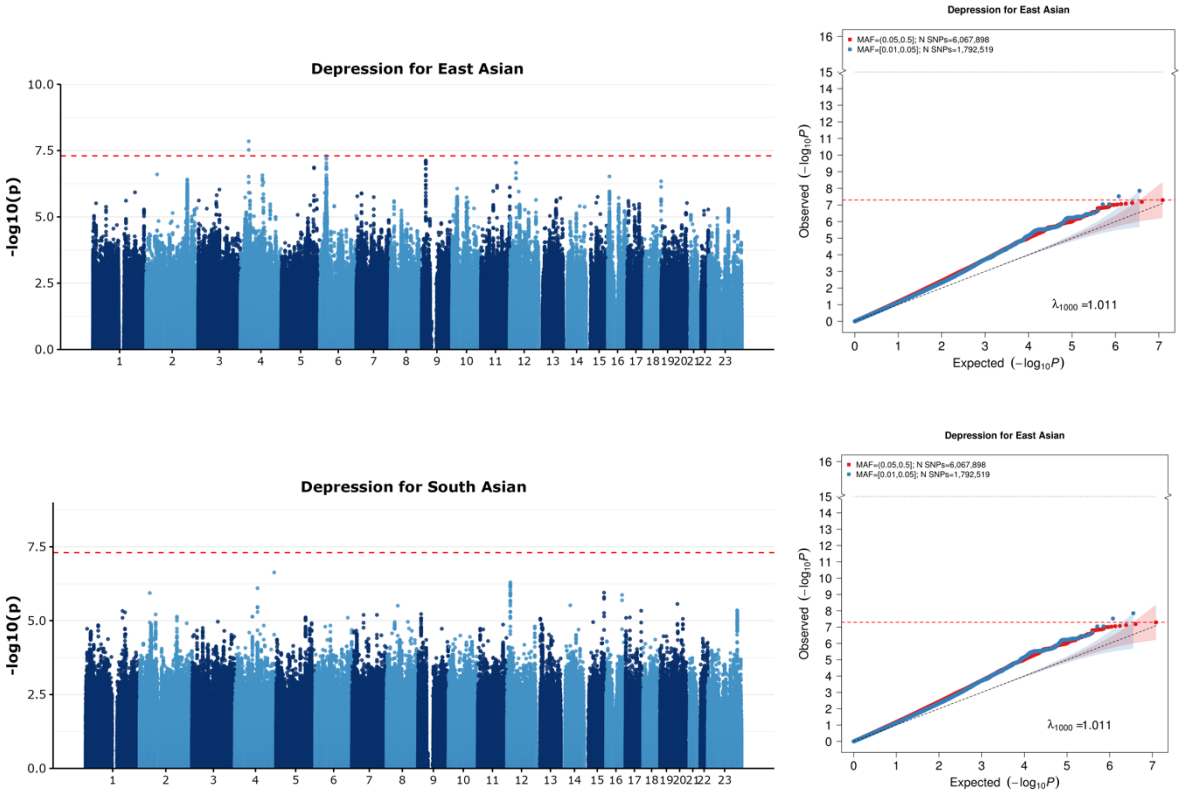
[1] For continuous traits, λ_{1000} scales the genomic inflation factor λ to a study with 1000 subjects using $\lambda_{1000} = 1 + 1000 * (\lambda - 1)/N$, where N is the total sample size. For binary traits, λ_{1000} scales λ to a study with 1000 cases and 1000 controls using $\lambda_{1000} = 1 +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

Supplementary figure 10: Manhattan plot and QQ plot[1] based on the 23andMe, Inc. GWAS summary statistics for depression in five populations: European, African American, Latino, East Asian, South Asian.



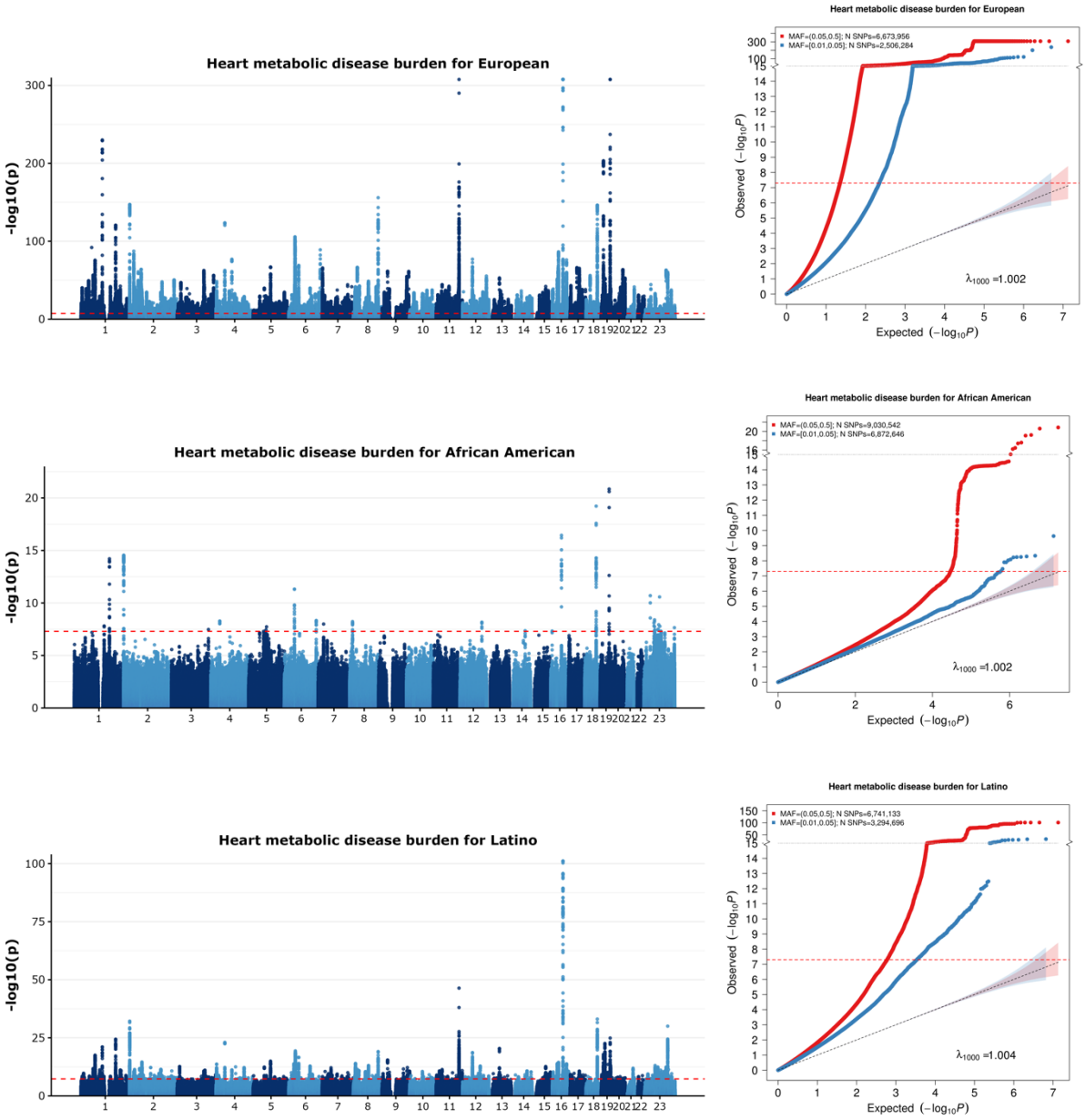
[1] For continuous traits, λ_{1000} scales the genomic inflation factor λ to a study with 1000 subjects using $\lambda_{1000} = 1 + 1000 * (\lambda - 1)/N$, where N is the total sample size. For binary traits, λ_{1000} scales λ to a study with 1000 cases and 1000 controls using $\lambda_{1000} = 1 +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

Supplementary figure 10 continued: Manhattan plot and QQ plot[1] based on the 23andMe, Inc. GWAS summary statistics for depression in five populations: European, African American, Latino, East Asian, South Asian.



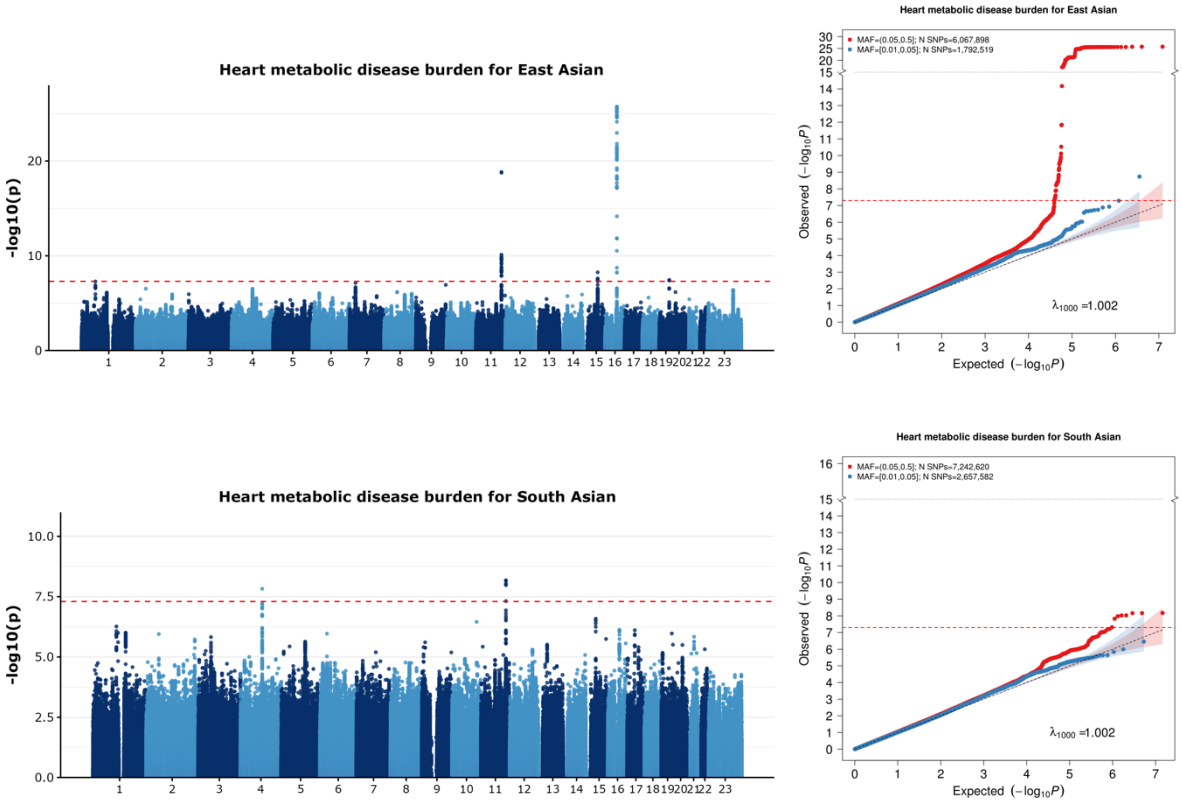
[1] For continuous traits, λ_{1000} scales the genomic inflation factor λ to a study with 1000 subjects using $\lambda_{1000} = 1 + 1000 * (\lambda - 1)/N$, where N is the total sample size. For binary traits, λ_{1000} scales λ to a study with 1000 cases and 1000 controls using $\lambda_{1000} = 1 +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

Supplementary figure 11: Manhattan plot and QQ plot[1] based on the 23andMe, Inc. GWAS summary statistics for heart metabolic disease burden in five populations: European, African American, Latino, East Asian, South Asian.



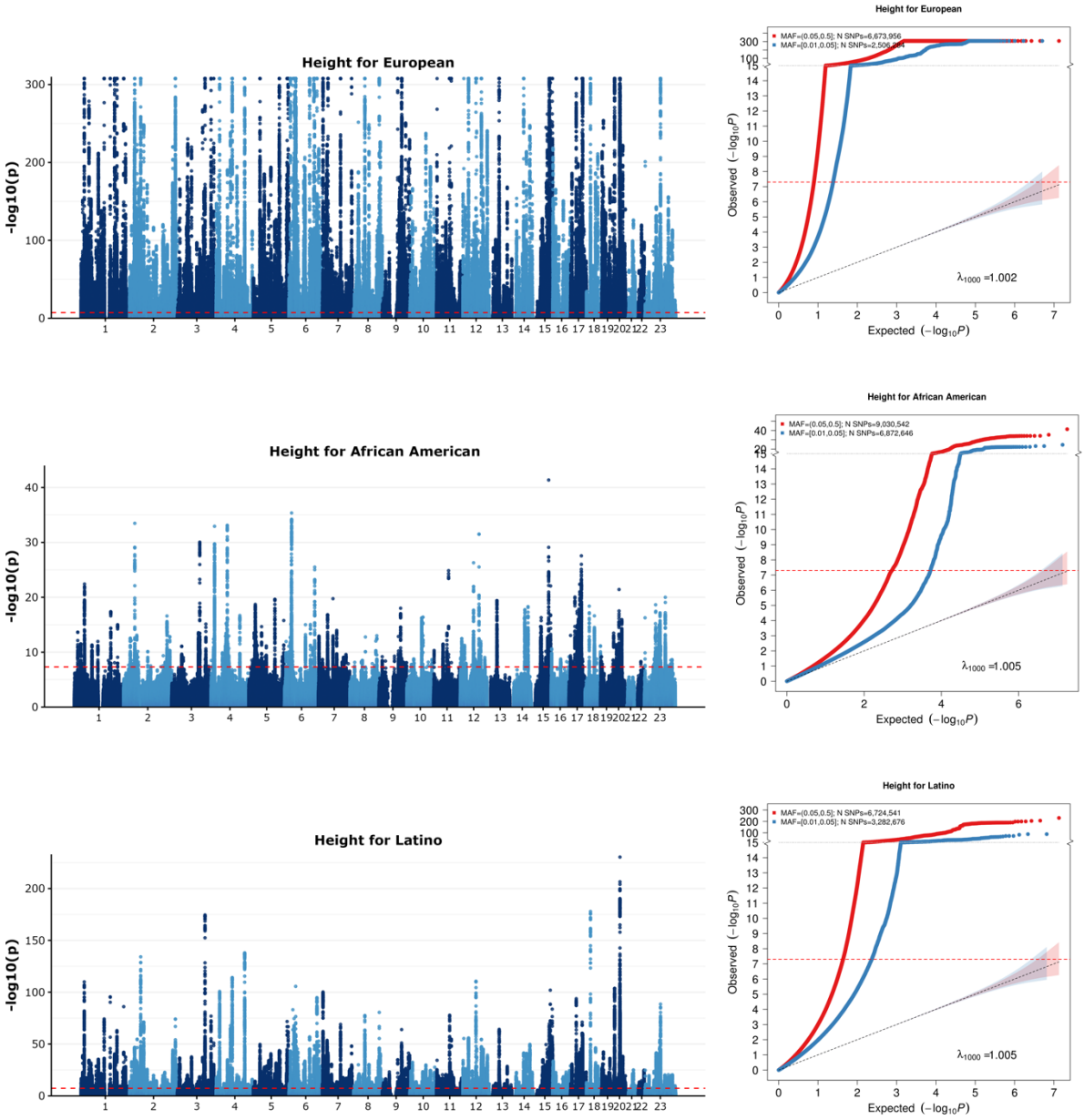
[1] For continuous traits, λ_{1000} scales the genomic inflation factor λ to a study with 1000 subjects using $\lambda_{1000} = 1 + 1000 * (\lambda - 1)/N$, where N is the total sample size. For binary traits, λ_{1000} scales λ to a study with 1000 cases and 1000 controls using $\lambda_{1000} = 1 +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

Supplementary figure 11 continued: Manhattan plot and QQ plot[1] based on the 23andMe, Inc. GWAS summary statistics for depression in five populations: European, African American, Latino, East Asian, South Asian.



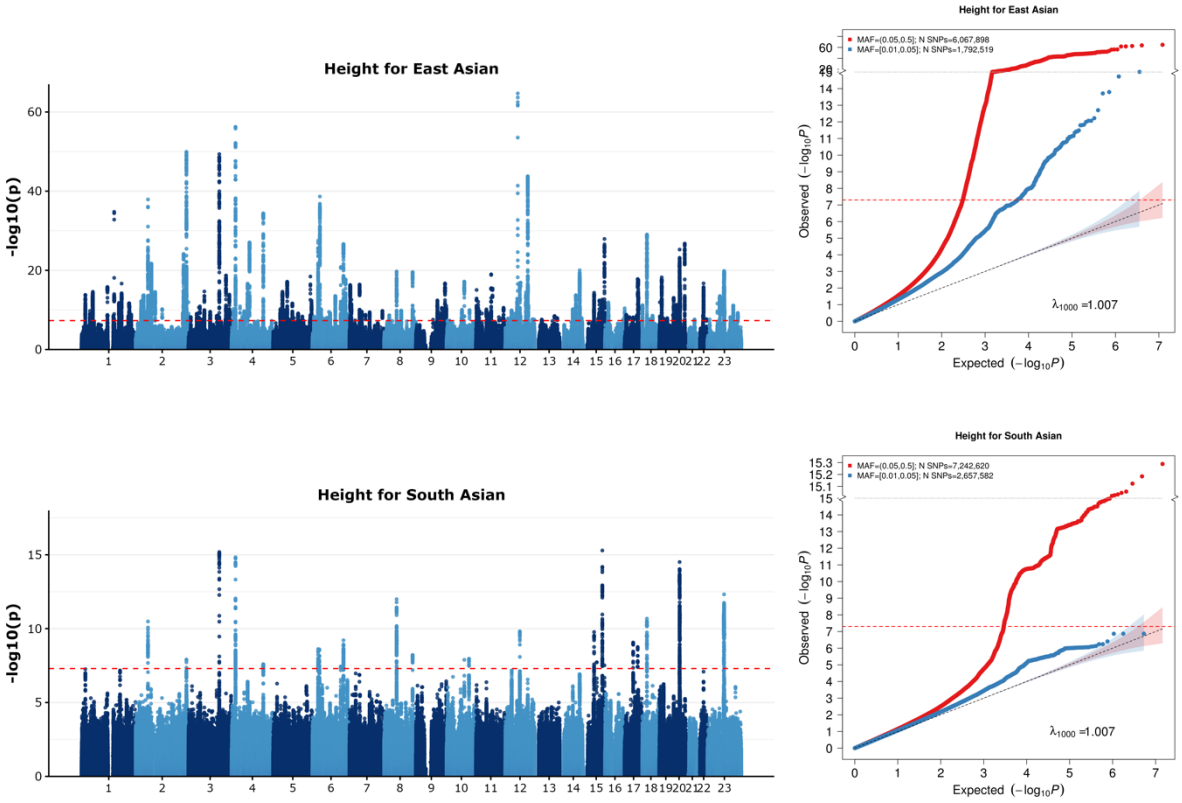
[1] For continuous traits, λ_{1000} scales the genomic inflation factor λ to a study with 1000 subjects using $\lambda_{1000} = 1 + 1000 * (\lambda - 1)/N$, where N is the total sample size. For binary traits, λ_{1000} scales λ to a study with 1000 cases and 1000 controls using $\lambda_{1000} = 1 +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

Supplementary figure 12: Manhattan plot and QQ plot[1] based on the 23andMe, Inc. GWAS summary statistics for height in five populations: European, African American, Latino, East Asian, South Asian.



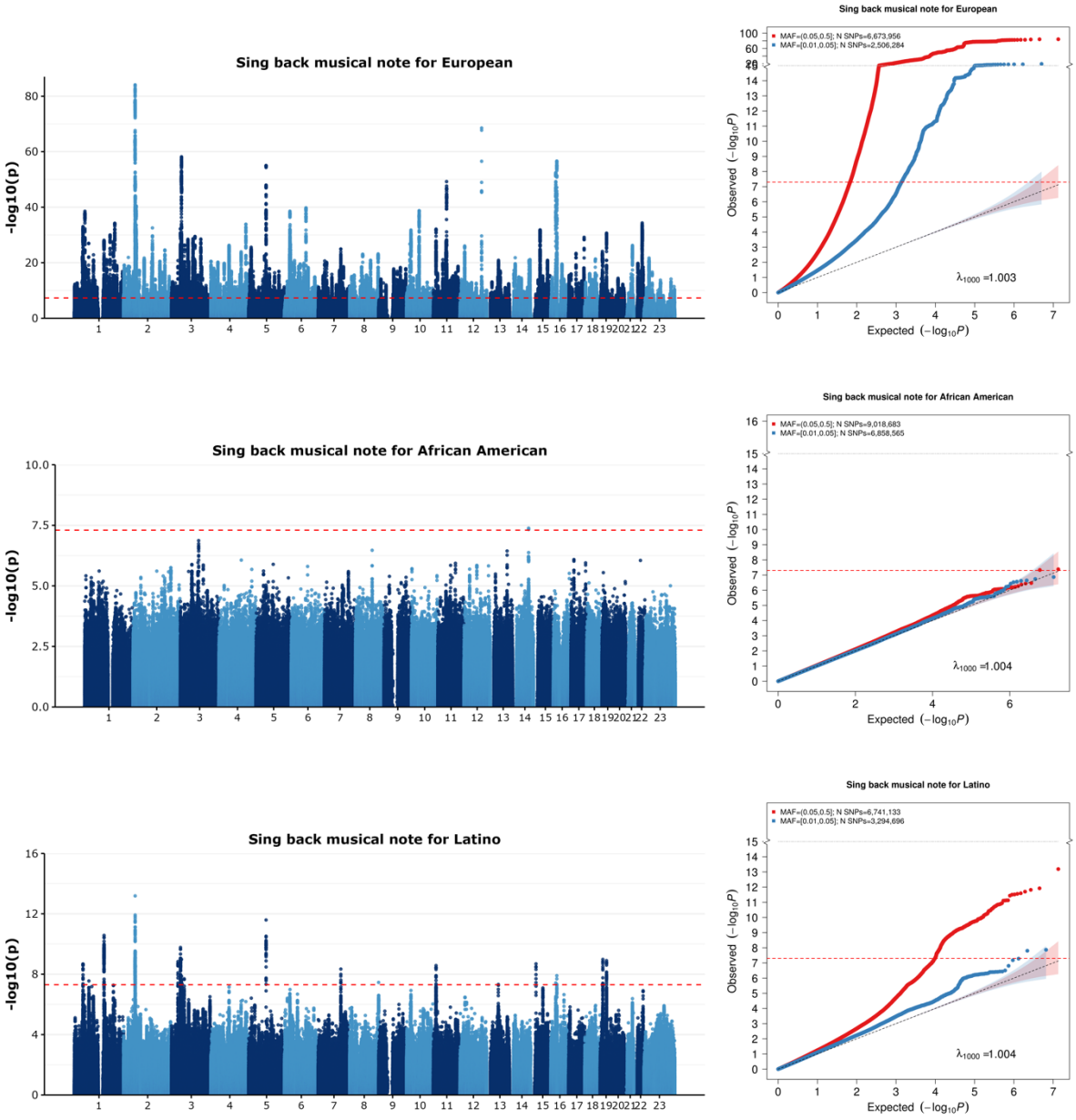
[1] For continuous traits, λ_{1000} scales the genomic inflation factor λ to a study with 1000 subjects using $\lambda_{1000} = 1 + 1000 * (\lambda - 1)/N$, where N is the total sample size. For binary traits, λ_{1000} scales λ to a study with 1000 cases and 1000 controls using $\lambda_{1000} = 1 +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

Supplementary figure 12 continued: Manhattan plot and QQ plot[1] based on the 23andMe, Inc. GWAS summary statistics for height in five populations: European, African American, Latino, East Asian, South Asian.



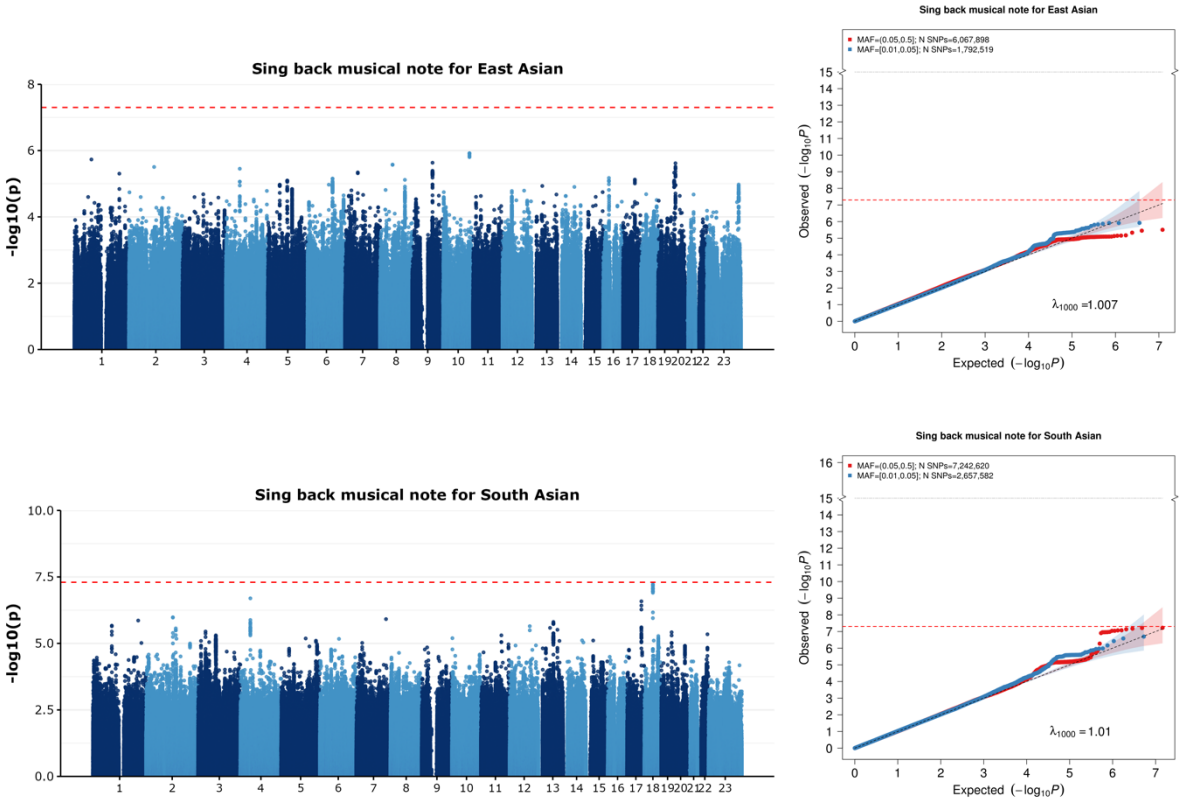
[1] For continuous traits, λ_{1000} scales the genomic inflation factor λ to a study with 1000 subjects using $\lambda_{1000} = 1 + 1000 * (\lambda - 1)/N$, where N is the total sample size. For binary traits, λ_{1000} scales λ to a study with 1000 cases and 1000 controls using $\lambda_{1000} = 1 +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

Supplementary figure 13: Manhattan plot and QQ plot[1] based on the 23andMe, Inc. GWAS summary statistics for sing back musical note (SBMN) in five populations: European, African American, Latino, East Asian, South Asian.



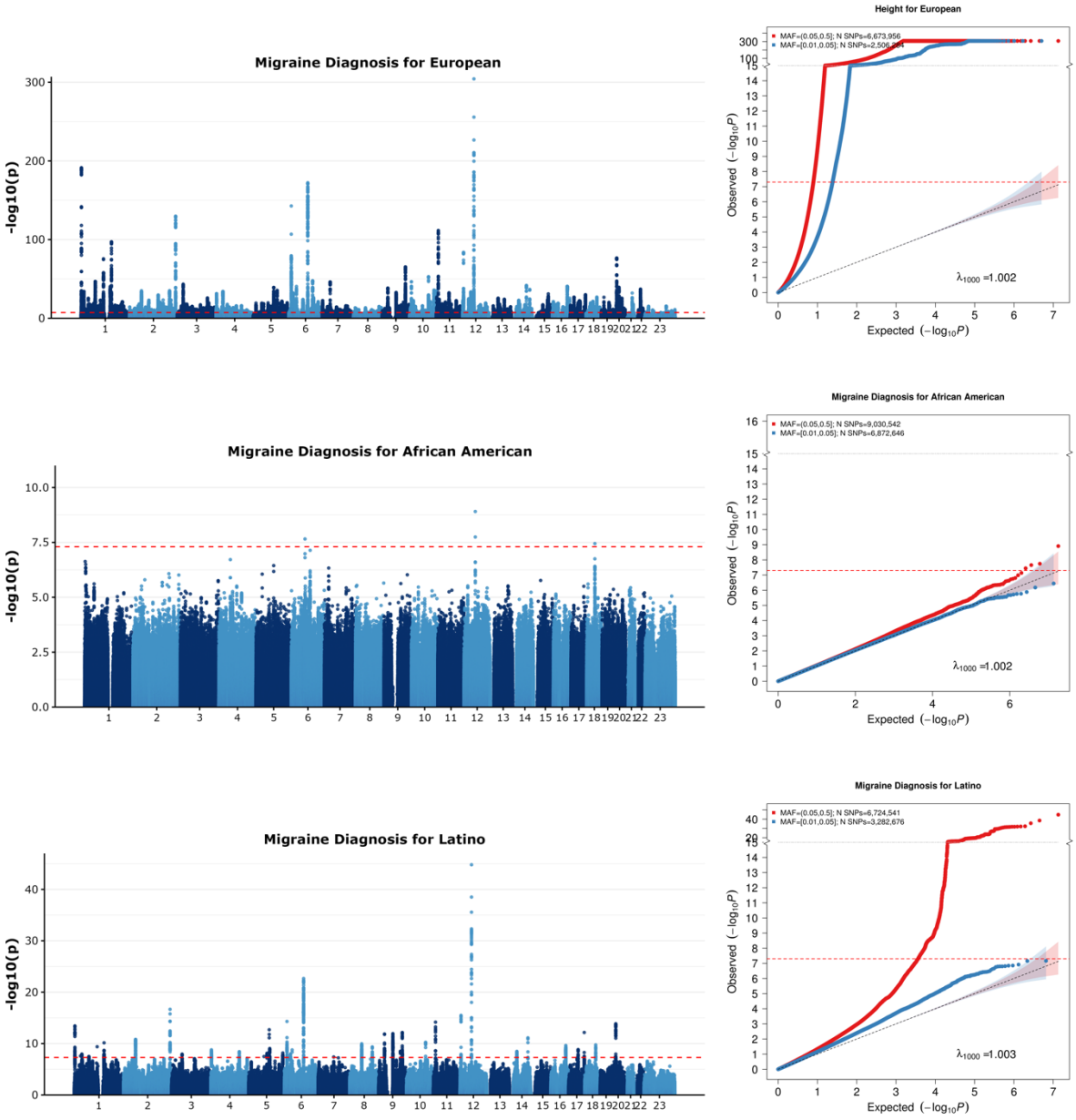
[1] For continuous traits, λ_{1000} scales the genomic inflation factor λ to a study with 1000 subjects using $\lambda_{1000} = 1 + 1000 * (\lambda - 1)/N$, where N is the total sample size. For binary traits, λ_{1000} scales λ to a study with 1000 cases and 1000 controls using $\lambda_{1000} = 1 +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

Supplementary figure 13 continued: Manhattan plot and QQ plot[1] based on the 23andMe, Inc. GWAS summary statistics for sing back musical note (SBMN) in five populations: European, African American, Latino, East Asian, South Asian.



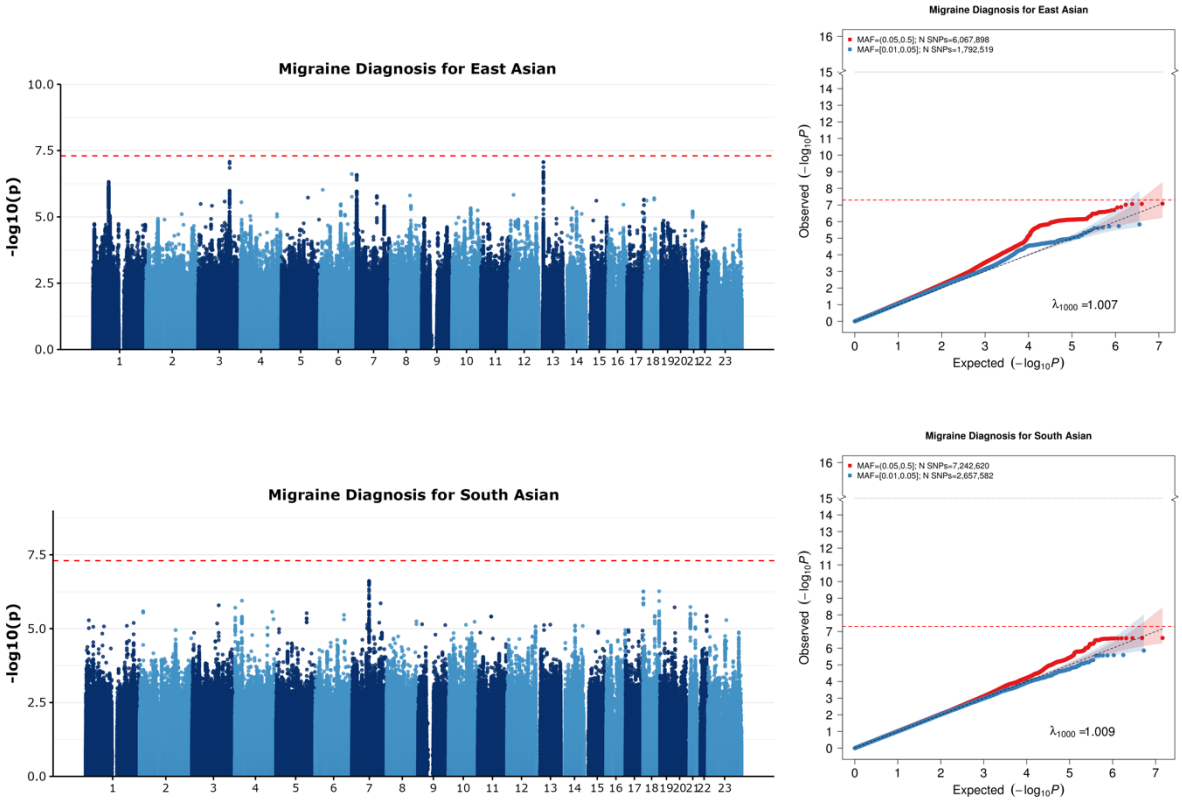
[1] For continuous traits, λ_{1000} scales the genomic inflation factor λ to a study with 1000 subjects using $\lambda_{1000} = 1 + 1000 * (\lambda - 1)/N$, where N is the total sample size. For binary traits, λ_{1000} scales λ to a study with 1000 cases and 1000 controls using $\lambda_{1000} = 1 +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

Supplementary figure 14: Manhattan plot and QQ plot[1] based on the 23andMe, Inc. GWAS summary statistics for migraine diagnosis in five populations: European, African American, Latino, East Asian, South Asian.



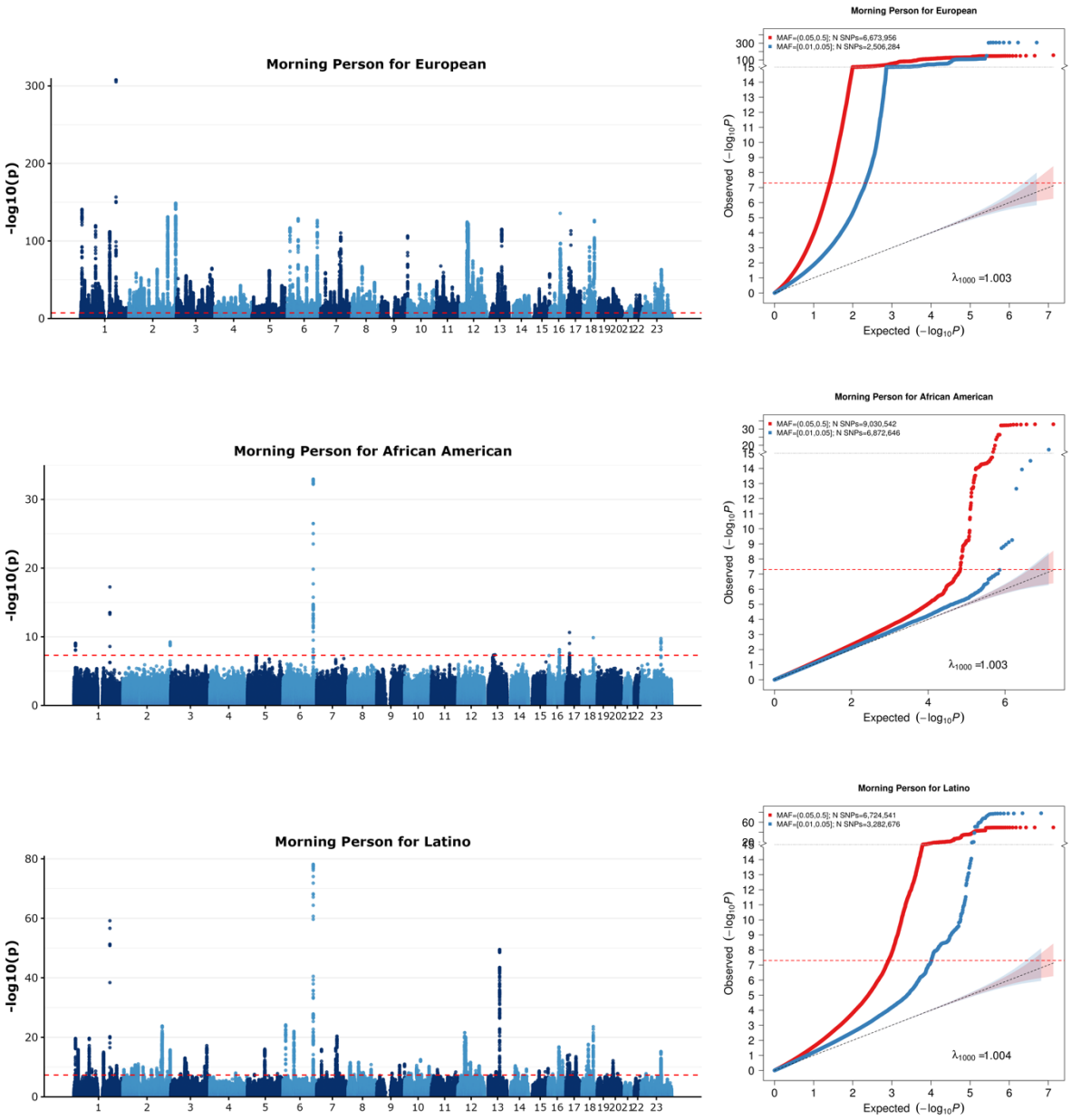
[1] For continuous traits, λ_{1000} scales the genomic inflation factor λ to a study with 1000 subjects using $\lambda_{1000} = 1 + 1000 * (\lambda - 1)/N$, where N is the total sample size. For binary traits, λ_{1000} scales λ to a study with 1000 cases and 1000 controls using $\lambda_{1000} = 1 +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

Supplementary figure 14 continued: Manhattan plot and QQ plot[1] based on the 23andMe, Inc. GWAS summary statistics for migraine diagnosis in five populations: European, African American, Latino, East Asian, South Asian.



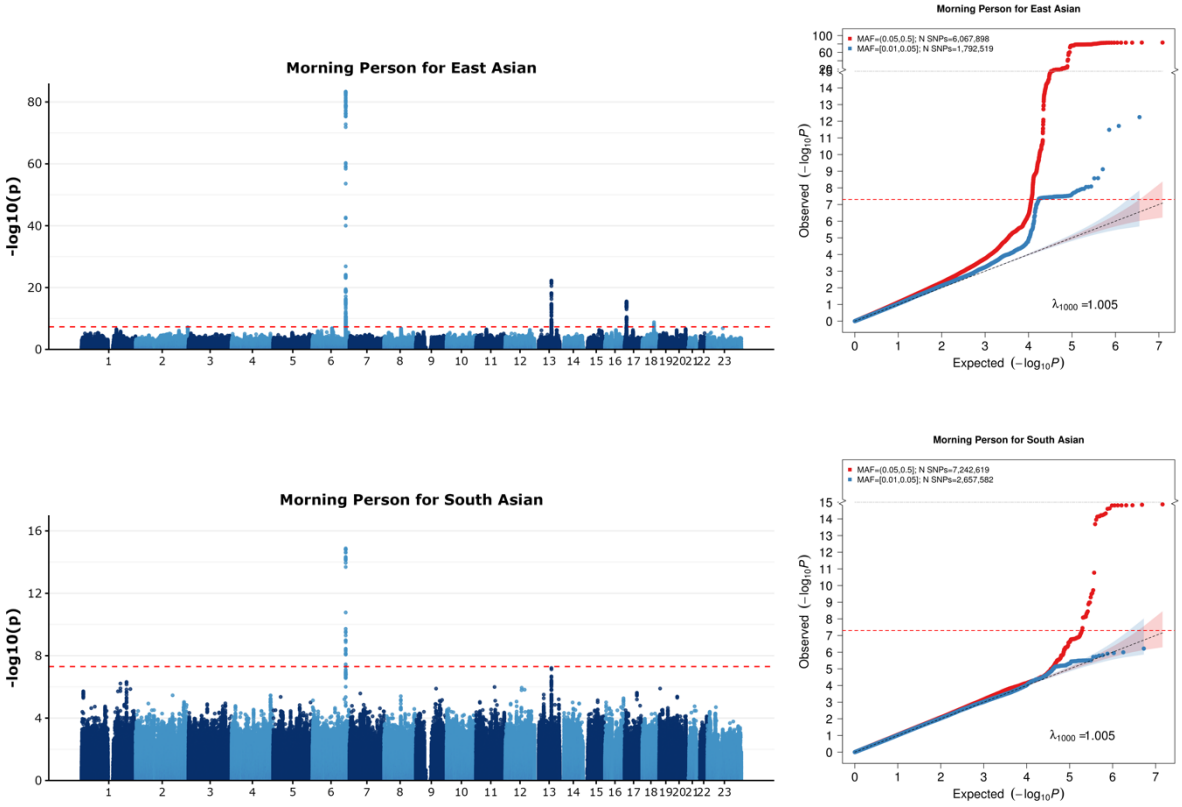
[1] For continuous traits, λ_{1000} scales the genomic inflation factor λ to a study with 1000 subjects using $\lambda_{1000} = 1 + 1000 * (\lambda - 1)/N$, where N is the total sample size. For binary traits, λ_{1000} scales λ to a study with 1000 cases and 1000 controls using $\lambda_{1000} = 1 +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

Supplementary figure 15: Manhattan plot and QQ plot[1] based on the 23andMe, Inc. GWAS summary statistics for morning person in five populations: European, African American, Latino, East Asian, South Asian.



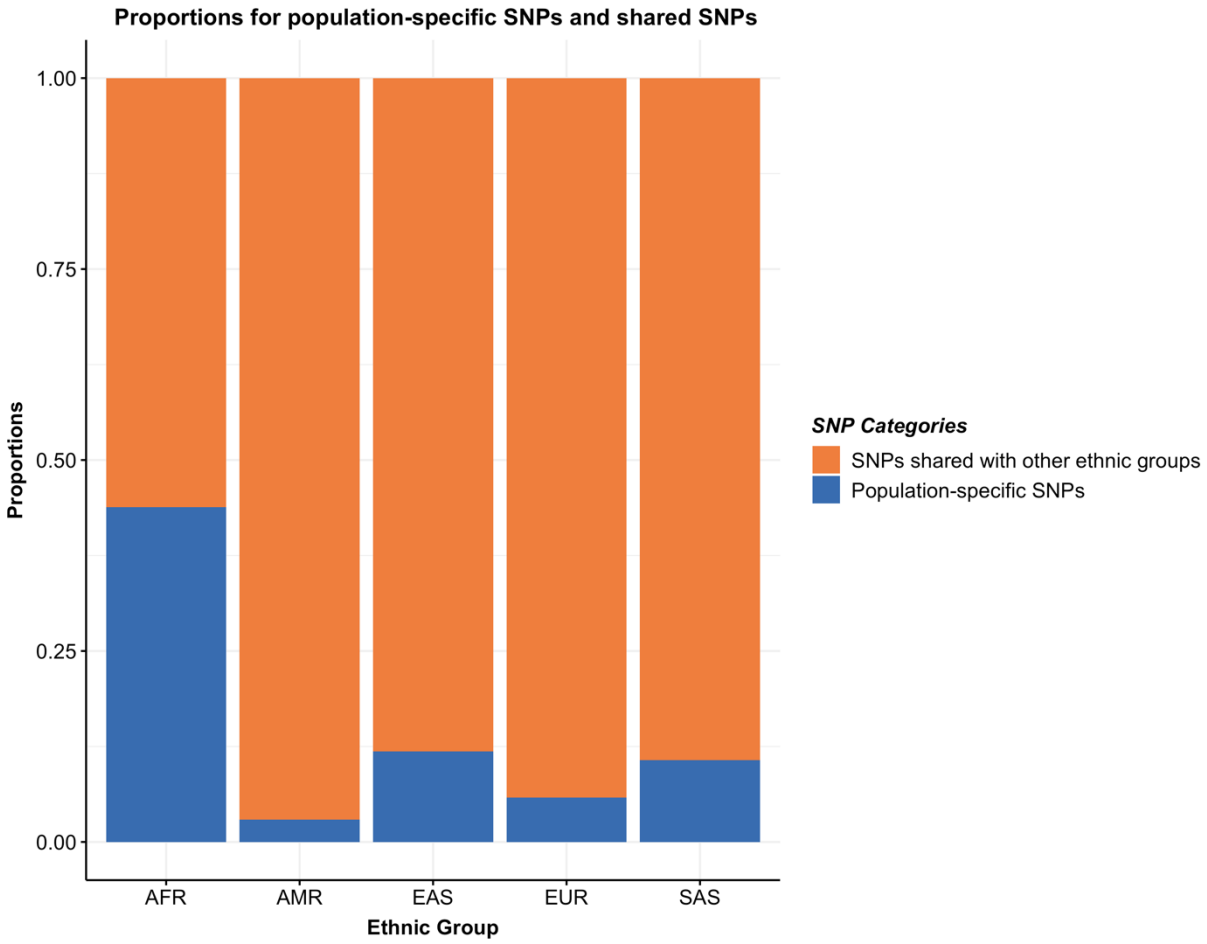
[1] For continuous traits, λ_{1000} scales the genomic inflation factor λ to a study with 1000 subjects using $\lambda_{1000} = 1 + 1000 * (\lambda - 1)/N$, where N is the total sample size. For binary traits, λ_{1000} scales λ to a study with 1000 cases and 1000 controls using $\lambda_{1000} = 1 +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

Supplementary figure 15 continued: Manhattan plot and QQ plot[1] based on the 23andMe, Inc. GWAS summary statistics for morning person in five populations: European, African American, Latino, East Asian, South Asian.



[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 $\lambda_{1000} = 1 + 1000 * (\lambda - 1) * (\frac{1}{N_{case}} + \frac{1}{N_{control}})$

Supplementary figure 16: Number of SNPs across five populations in 1000 Genomes Project (Phase 3). SNPs are selected with MAF > 0.01 in at least one of five populations: African (AFR), American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS).



Supplementary Note

Super learning

Let $X = (PRS_1, PRS_2, \dots, PRS_M)$ be the input dataset. Suppose the predicted outcome are $\psi(X)_v$ for $v = 1,2, \dots V$ different prediction algorithms. For continuous traits, the objective function for each of the prediction algorithm is defined to minimize $\{Y - \psi(X)_v\}^2$. We propose a weighted combination of the V different predictors as $F(X) = \sum_{v=1}^{V} \alpha_v \, \hat{\psi}(X)_v$ with $\sum_{v=1}^{V} \alpha_v = 1$ and $\alpha_v \geq 0$ for any v. Then the weights α are determined using the tuning dataset with:

$$\hat{\alpha} = \arg\min_{\alpha} \{Y - F(X)\}^2,$$

For binary traits, the objective function for each of the algorithm is to maximize the AUC. The weights α are determined using the tuning dataset by maximizing the AUC while using $F(X)$ to predict Y. With the optimal weights of the V different algorithm, the super learner estimator is:

$$\hat{\psi}^{SL} = \sum_{v=1}^{V} \hat{\alpha}_v \, \hat{\psi}_v.$$

In practice, we use three different prediction algorithms implemented in the SuperLearner package[9] to generate the super learning estimate: Lasso[10], ridge regression[11] and neural networks[12]. The tuning parameters for these algorithms are put as the defaults as SuperLearner package. Other common prediction algorithms, such as random forest[13] and XGBoost[14], are also available in the R package, and can be added for constructing PRSs by the user. For binary traits, since the ridge regression algorithm is not supported by SuperLearner package now, we only use Lasso and neural networks in real data analysis. To use AUC as the objective function, we use the flag “method = method.AUC” in the SupearLearner package.

Mild and no negative selection simulation generation

Suppose u_{kl} is the standardized effect size for kth causal SNP for the lth population. We consider generating the effect size of causal variants are related to allele frequency under 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. To generate the required effect-size, we first draw from a multivariate normal distribution:

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

where v_{kl} is on a temporary generating scale, and h^2 is set to 0.4. Then, the temporary effect size is defined $u_{kl}^* = v_{kl}\{\sqrt{2f_{kl}(1-f_{kl})}\}^{\alpha}$, where $\alpha = 0.75$ is for mild negative selection and $\alpha = 1$ was for no negative selection. To control the common SNPs heritability as 0.4, the standardized effect size is generated by scaling the temporary effect-size as $u_{kl} = u_{kl}^*\{\sqrt{\frac{0.4}{\sum_{l=1}^{C_l}(u_{kl}^*)^2}}\}$. The genetic correlation ρ is defined on the generating scale. For both mild and no negative selection setting, ρ is set to 0.8.

Genotyping of 23andMe data

DNA extraction and genotyping were performed on saliva samples by National Genetics Institute (NGI), a CLIA licensed clinical laboratory and a subsidiary of Laboratory Corporation of America. Samples were genotyped on one of five genotyping platforms. The v1 and v2 platforms were variants of the Illumina HumanHap550+ BeadChip, including about 25,000 custom SNPs selected by 23andMe, with a total of about 560,000 SNPs. The v3 platform was based on the Illumina OmniExpress+ BeadChip, with custom content to improve the overlap with our v2 array, with a total of about 950,000 SNPs. The v4 platform was a fully customized array, including a lower redundancy subset of v2 and v3 SNPs with additional coverage of lower-frequency coding variation, and about 570,000 SNPs. The v5 platform, in current use, is an Illumina Infinium Global Screening Array (~640,000 SNPs) supplemented with ~50,000 SNPs of custom content. This array was specifically designed to better capture global

genetic diversity and to help standardize the platform for genetic research. Samples that failed to reach 98.5% call rate were re-analyzed. Individuals whose analyses failed repeatedly were re-contacted by 23andMe customer service to provide additional samples.

Ancestry determination of 23andMe data

For our standard GWAS, we restrict participants to a set of individuals who have a specified ancestry determined through an analysis of local ancestry[1]. Briefly, our algorithm first partitions phased genomic data into short windows of about 300 SNPs. Within each window, we use a support vector machine (SVM) to classify individual haplotypes into one of 31 reference populations (https://www.23andme.com/ancestry-composition-guide/). The SVM classifications are then fed into a hidden Markov model (HMM) that accounts for switch errors and incorrect assignments, and gives probabilities for each reference population in each window. Finally, we use simulated admixed individuals to recalibrate the HMM probabilities so that the reported assignments are consistent with the simulated admixture proportions. The reference population data is derived from public datasets (the Human Genome Diversity Project, HapMap, and 1000 Genomes), as well as 23andMe customers who have reported having four grandparents from the same country.

Ancestries are defined as follow:

Ancestry	Classification criteria
European	European + Middle Eastern > 0.97, European > 0.90
East Asian	East Asian + Southeast Asian > 0.97

South Asian	South Asian > 0.97
Middle Eastern (& North African)	Middle Eastern + European > 0.97, Middle Eastern > 0.90
African American + Latinos	European + African + East Asian + Native American + Middle Eastern > 0.90, African + Native American > 0.01

African Americans and Latinos are admixed with broadly varying contributions from Europe, Africa and the Americas. Therefore, no single threshold of genome-wide ancestry will be able to effectively discriminate African Americans and Latinos. However, the distributions of the length of segments of European, African and American ancestry are very different between African Americans and Latinos, because of distinct admixture timing between the three ancestral populations in the two ethnic groups. Therefore, we trained a logistic classifier that takes one customer's length histogram of segments of African, European and American ancestry, and predict whether the customer is likely African American or Latino.

Selecting unrelated individuals within 23andMe data

A maximal set of unrelated individuals was chosen for each analysis using a segmental identity-by-descent (IBD) estimation algorithm[2]. Individuals were defined as related if they shared more than 700 cM IBD, including regions where the two individuals share either one or both genomic segments IBD. This level of relatedness (roughly 20% of the genome) corresponds approximately to the minimal expected sharing between first cousins in an outbred population. When selecting individuals for case/control phenotype analyses, the selection process is designed to maximize case sample size by

preferentially retaining cases over controls. Specifically, if both an individual case and an individual control are found to be related, then the case is retained in the analysis.

Imputation of 23andMe data

Imputation panels created by combining multiple smaller panels have been shown to give better imputation performance than the individual constituent panels alone[3]. To that end, we combined the May 2015 release of the 1000 Genomes Phase 3 haplotypes[4] with the UK10K imputation reference panel[5] to create a single unified imputation reference panel. To do this, multiallelic sites with N alternate alleles were split into N separate biallelic sites. We then removed any site whose minor allele appeared in only one sample. For each chromosome, we used Minimac3[6] to impute the reference panels against each other, reporting the best-guess genotype at each site. This gave us calls for all samples over a single unified set of variants. We then joined these together to get, for each chromosome, a single file with phased calls at every site for 6,285 samples. Throughout, we treated structural variants and small indels in the same way as SNPs.

In preparation for imputation we split each chromosome of the reference panel into chunks of no more than 300,000 variants, with overlaps of 10,000 variants on each side. We used a single batch of 10,000 individuals to estimate Minimac3 imputation model parameters for each chunk.

To generate phased participant data for the v1 to v4 platforms, we used an internally-developed tool, Finch, which implements the Beagle graph-based haplotype phasing algorithm[7], modified to separate the haplotype graph construction and phasing steps. Finch extends the Beagle model to accommodate genotyping error and recombination, in order to handle cases where there are no consistent paths through the haplotype graph for the individual being phased. We constructed haplotype graphs for all participants from a representative sample of genotyped individuals, and then performed out-of-sample phasing of all genotyped individuals against the appropriate graph. For the X chromosome, we built separate haplotype graphs for the non-pseudoautosomal region and each pseudoautosomal region, and these regions were phased separately. For the 23andMe participants genotyped on the v5 array, we used a similar approach, but using a new phasing algorithm, Eagle2[8]. We imputed phased participant data against the merged reference panel using Minimac3, treating males as homozygous pseudo-diploids for the non-pseudoautosomal region.

We compute association test results for the genotyped and the imputed SNPs. For case control phenotypes, we compute association by logistic regression assuming additive allelic effects. For tests using imputed data, we use the imputed dosages rather than best-guess genotypes. As standard, we include covariates for age, gender, the top five principal components to account for residual population structure, and indicators for genotype platforms to account for genotype batch effects. The association test *P* value we report is computed using a likelihood ratio test, which in our experience is better behaved than a Wald test on the regression coefficient. For quantitative traits,

association tests are performed by linear regression. Results for the X chromosome are computed similarly, with male genotypes coded as if they were homozygous diploid for the observed allele.

Principal calculation using 23andMe data

A principal component analysis was performed independently for each ancestry, using ~65,000 high quality genotyped variants present in all five genotyping platforms. It was computed on a subset of participants randomly sampled across all the genotyping platforms (137K, 102K, 1000K, 360K and 32K participants were used for African-American, East-Asian, European, Latino, and South-Asian, respectively). PC scores for participants not included in the analysis were obtained by projection, combining the eigenvectors of the analysis and the SNP weights.

Quality control of 23andMe data

The vast majority of SNPs are only imputed and not genotyped, and therefore they only have imputed GWAS results. A small proportion of SNPs (often rare) have only genotyped GWAS results. Finally, the majority of genotyped SNPs are also imputed. When choosing between imputed and genotyped GWAS results for these SNPs, if they both passed quality control (QC), we report the imputed result.

Variant QC is applied independently to genotyped and imputed GWAS results, and we flag the SNPs failing QC. For QC of genotyped GWAS results, we flagged SNPs that were only genotyped on our “v1” and/or “v2” platforms due to small sample size, and

SNPs on chrM or chrY because many of these are not currently called reliably. Using trio data, we flagged SNPs that failed a test for parent-offspring transmission; specifically, we regressed the child's allele count against the mean parental allele count and flagged SNPs with fitted $\beta<0.6$ and $P<10^{-20}$ for a test of $\beta<1$. We flagged SNPs with a Hardy-Weinberg $P<10^{-20}$, or a call rate of <90%. We also tested genotyped SNPs for genotype date effects, and flagged SNPs with $P<10^{-50}$ by ANOVA of SNP genotypes against a factor dividing genotyping date into 20 roughly equal-sized buckets. We flagged SNPs with large sex effect (ANOVA of SNP genotypes, $r^2>0.1$). Finally, we flag SNPs with probes matching multiple genomic positions in the reference genome ('self chain').

For imputed GWAS results, we flagged SNPs with rsq<0.3, as well as SNPs that had strong evidence of a platform batch effect. The batch effect test is an F test from an ANOVA of the SNP dosages against a factor representing v4 or v5 platform; we flagged results with $P<10^{-50}$. Prior to GWAS, we identified, for each SNP, the largest subset of the data passing these criteria, based on their original genotyping platform -- either v2+v3+v4+v5, v4+v5, v4, or v5 only -- and computed association test results for whatever was the largest passing set. As a result, there are no imputed results for SNPs that fail these filters.

Across all results, we flag SNPs that have an available sample size of less than 20% of the total GWAS sample size. We also flag logistic regression results that did not converge due to complete separation, identified by abs (effect) >10 or stderr>10 on the

log odds scale. We also flag linear regression results for SNPs with MAF < 0.1% because tests of low frequency variants can be sensitive to violations of the regression assumption of normally distributed residuals.

References

1. Durand, E. Y., Do, C. B., Mountain, J. L. & Macpherson, J. M. Ancestry composition: a novel, efficient pipeline for ancestry deconvolution. *biorxiv* 10512 (2014).

2. Henn, B. M. *et al.* Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PloS One* **7**, (2012).

3. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, (2015).

4. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

5. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–89 (2015).

6. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

7. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).

8. Loh, P. R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).

9. Polley, E., LeDell, E., Kennedy, C. & van der Laan, M. SuperLearner: Super Learner Prediction. (2019).

10. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 267–288 (1996).

11. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1 (2010).

12. Venables, W. N. & Ripley, B. D. Modern applied statistics with S fourth edition. World. (2002).

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

14. Chen, T. *et al.* xgboost: Extreme Gradient Boosting. *R package version 0.6-4* 1–4 (2015).