Generalizing Polygenic Risk Scores from Europeans to Hispanics/Latinos: Supplementary Material

Kelsey E. Grinde, Qibin Qi, Timothy A. Thornton, Simin Liu, Aaddin H. Shadyab, Kei Hang K. Chan, Alexander P. Reiner, Tamar Sofer

Contents

1 Figures of LD structure in the HCHS/SOL

The following figures display the patterns of LD around (up to 1000 base-pairs away) the SNP rs4628172 on chromosome 7 in each background group in HCHS/SOL. The figures focuses on genotyped, non-monomorphic SNPs.

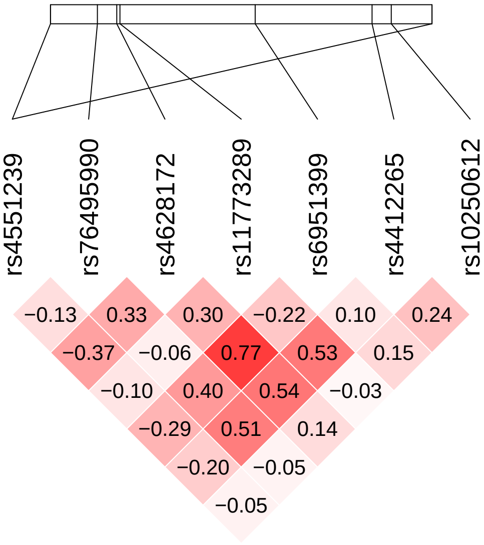

Figure S1: LD between 6 genotyped SNPs 1000bp around rs4628172 in chromosome 7, in the Central American group.

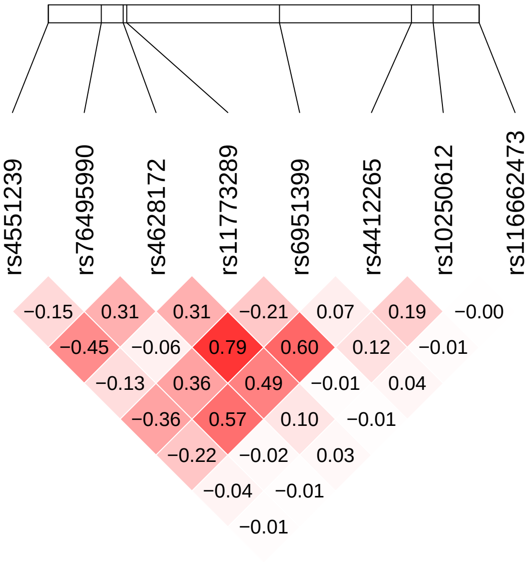

Figure S2: LD between 7 genotyped SNPs 1000bp around rs4628172 in chromosome 7, in the Mexican group.

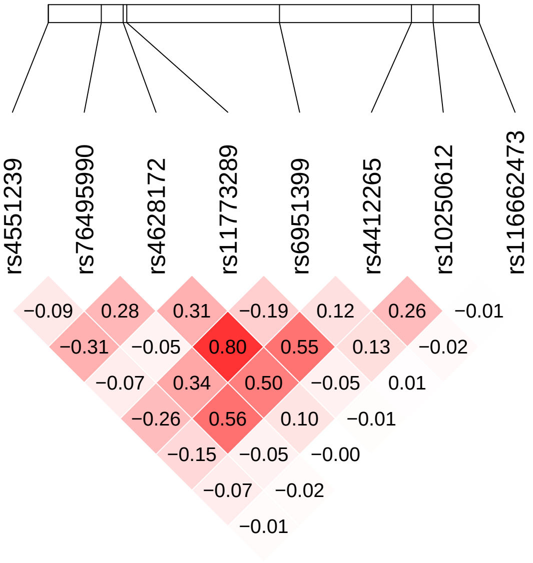

Figure S3: LD between 7 genotyped SNPs 1000bp around rs4628172 in chromosome 7, in the South American group.

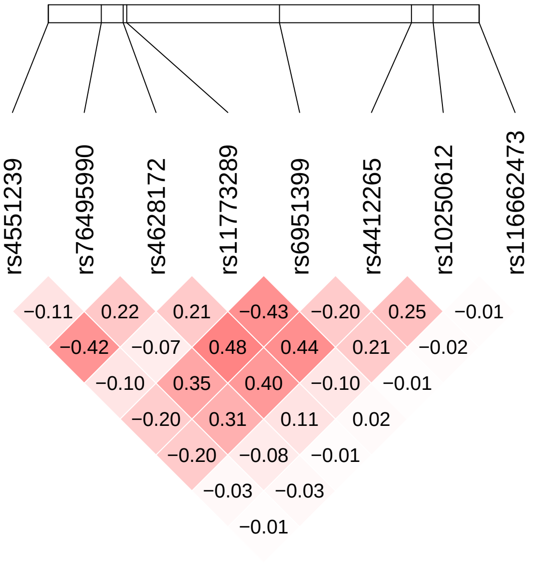

Figure S4: LD between 7 genotyped SNPs 1000bp around rs4628172 in chromosome 7, in the Dominican group.

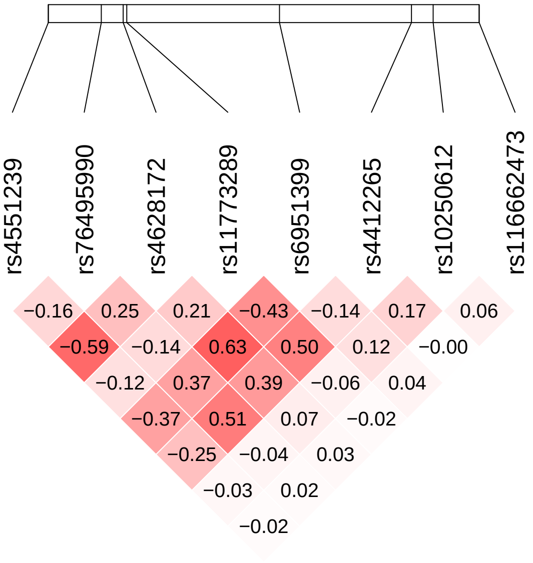

Figure S5: LD between 7 genotyped SNPs 1000bp around rs4628172 in chromosome 7, in the Cuban group.

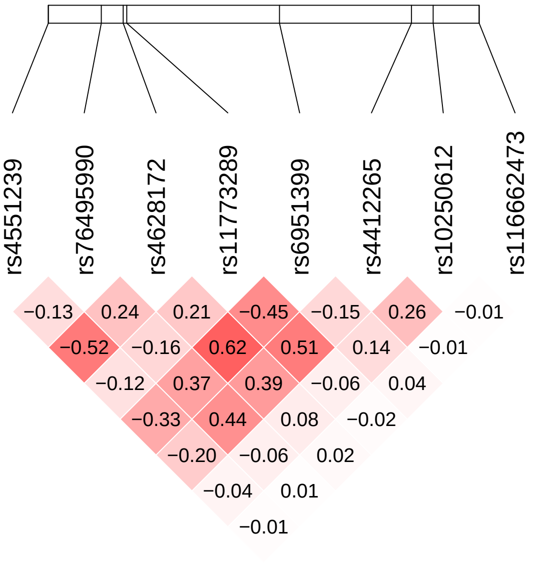

Figure S6: LD between 7 genotyped SNPs 1000bp around rs4628172 in chromosome 7, in the Puerto Rican group.

2 Additional information about GWAS in the HCHS/SOL

2.1 Characteristics of main GWAS

The following table lists the details of each of the GWASs performed in the HCHS/SOL data, which were used as a potential training dataset for PRS construction in WHI SHARe. For each trait, we provide the sample size, covariates included in the model, and additional details relevant to the analysis.

Trait	Pooled/ Stratified	Sample size	Covariates	Transformation	Imputation panel (1000 genome phase)
BMI	Pooled	12705	sex, recruitment center, age, background group, sampling weight	Rank-normalization of residuals	Phase 1
DBP	Stratified	12278	sex, age, age^2, recruitment center, BMI, sampling weight	None (values of HT medication users were adjusted)	Phase 3
Height	Stratified	12652	sex, age, recruitment center, US born indicator, sampling weight	None	Phase 1
HGB	Pooled	12502	sex, recruitment center, age, cigarette use, background group, sampling weight	None	Phase 1
HIP	Pooled	12673	sex, recruitment center, background group, age, age, sampling weight	Rank-normalization of residuals	Phase 1
MAP	Stratified	12278	sex, age, age^2, recruitment center, BMI, sampling weight	None	Phase 3
PLT	Pooled	12491	sex, recruitment center, age, cigarette use, background group, sampling weight	None	Phase 1
PP	Stratified	12277	sex, age, age^2, recruitment center, BMI, sampling weight	None (two values were winsorized to mean + 6 SDs)	Phase 3
SBP	Stratified	12278	sex, age, age^2, recruitment center, BMI, sampling weight	None (values of HT medication users were adjusted)	Phase 3
WBC	Pooled	11809	sex, recruitment center, age, cigarette use, background group, sampling weight	None	Phase 1
WC	Pooled	12679	recruitment center, background group, age, age^2, sex, sampling weight	Rank-normalization of residuals	Phase 1
WHR	Pooled	12672	sex, recruitment center, background group, age, age^2, sampling weight	Rank-normalization of residuals	Phase 1

Table S1: Details about HCHS/SOL GWAS used in the manuscript. BMI: Body Mass Index; DBP: diastolic blood pressure; HGB: Hemoglobin concentration; HIP: Hip circumference; MAP: mean arterial pressure; PLT: platelet count; PP: pulse pressure; SBP: systolic blood pressure; WBC: white blood cell count; WC: waist circumference; WHR: waist-to-hip ratio. All analyses were run via mixed models, with correlation matrices accounting for kinship, household, and block unit sharing, and adjusted for the 5 first principal components of the genetic data in addition to the covariates listed in the table. Stratified analyses were always stratified by background group. When a stratified analysis was performed, the PCs used in each of the background groups were those computed within that group. Sampling weights were always log transformed.

2.2 Best performing combined PRSs in a HCHS/SOL left-out sample and their performance in WHI

We considered an approach that randomly splits the HCHS/SOL dataset in two. One half is the training dataset, used to estimate p-values and effect sizes, and the second half, also called a left-out dataset, is a used for evaluating PRSs performance. We run a GWAS in the first half, and uses it to construct PRSs for multiple p-value thresholds (after pruning), and with either no weights (summing trait-increasing alleles), or with weights being the estimated effect sizes in the training dataset. Then, each of these PRSs is evaluated in the left-out dataset, in a linear mixed model (LMM) regression. The best PRSs is the one with smallest p-value in the left-out dataset. We denote this score SOL_b. Similarly, we constructed PRSs according to the EA GWAS, with clumping based on EA 1000 Genome reference panel. The best EA PRS is again the one with the smallest p-value in the LMM in the left-out dataset. We denote this score EA_b. Finally, we created combinations of SOL_b and EA_b of the form $\alpha\mathrm{EA}_b + (1-\alpha)\mathrm{SOL}_b$, for $\alpha \in \{0, 0.1, \dots, 1\}$ and evaluated these in an LMM. The best combination is the one with the smallest p-value in the left-out HCHS/SOL dataset. The results for each investigated trait are reported in Table S2. For clumping, we used $\rho^2 = 0.2$ as the LD threshold. Note that we used the p-value criterion rather than variance explained because the HCHS/SOL GWAS had heterogeneous residual variances by background groups.

The results from this construction are then used in the WHI PRSs evaluation, and are reported in Table S3. The only PRSs combination that performed well are those for PLT, WBC, and HGB, who were completely based on EA GWAS.

	EA_b characteristics				SOL_b characteristics			
Trait	α	Weights type	p-value threshold	# SNPs	$1-\alpha$	Weights type	p-value threshold	# SNPs
BMI	0	EA	1e-06	262	1	None	0.5	231,883
DBP	0.3	None	1e-04	197	0.7	None	0.001	1,991
Height	0	EA	0.05	54,263	1	SOL1	1e-04	21,584
HGB	1	EA	1e-06	45	0	None	5e-08	1
HIP	0	EA	5e-08	92	1	None	0.5	230,222
MAP	0.6	None	0.05	23,880	0.4	None	0.5	228,810
PLT	1	EA	1e-05	118	0	SOL1	5e-08	3
PP	1	None	5e-08	14	0	SOL1	1e-04	298
SBP	0	EA	0.001	967	1	None	0.05	41,189
WBC	1	EA	1e-07	39	0	SOL1	5e-08	7
WC	0	EA	1e-07	87	1	None	0.05	41,337
WHR	0	EA	0.05	23,480	1	None	0.5	231,573

Table S2: Best performing PRSs in the HCHS/SOL split dataset evaluations. For each trait, the left part of the table describes EA_b, the best performing EA-guided PRSs in the left-out, or second, HCHS/SOL dataset, which was randomly split in half, and the right side, describes SOL_b, the best performing first HCHS/SOL half guided PRSs. The best performing PRSs are described in terms of the weights (none, or effect sizes estimated in EA GWAS (for EA_b), or in HCHS/SOL GWAS in the training dataset, called SOL1), p-value threshold used for SNP selection, and the number of SNPs composing the PRS. We also provide α and $1-\alpha$, the selected weights for the combination $\alpha\text{EA}_b + (1-\alpha)\text{SOL}_b$, selected based on results in the left-out dataset. Grey cells in the table corresponds to instances in which the PRS has selected coefficient 0 (no contribution) in the combined score $\alpha\text{EA}_b + (1-\alpha)\text{SOL}_b$.

Trait	HA variance explained	AA variance explained
BMI	<0.01	0.10
Height	<0.01	<0.01
WC	0.02	0.01
WHR	0.01	<0.01
HIP	0.01	<0.01
HGB	1.13	0.45
WBC	2.49	10.69
PLT	4.34	1.51
SBP	0.10	<0.01
DBP	<0.01	<0.01
MAP	<0.01	0.10
PP	0.01	<0.01

Table S3: Percent variance explained by PRSs constructed in WHI Hispanic American and African American women based on the combinations of EA-based and HCHS/SOL-based scores $\alpha\text{EA}_b + (1-\alpha)\text{SOL}_b$ identified and reported in Table S2.

3 Simulation studies

3.1 Simulating admixed individuals and their haplotypes

Simulating European and African haplotypes using HAPGEN2

We used HAPGEN2 (Su et al., 2011), with data from HapMap3 (The International HapMap Consortium, 2005), to simulate a 1Mbp region at the start of chromosome 3 for a sample of (1) 50,000 CEU individuals, (2) 34,000 CEU individuals, and (3) 34,000 YRI individuals. Phased haplotypes, legend files, and recombination rates for HapMap3 were downloaded from: `https://mathgen.stats.ox.ac.uk/impute/impute_v1.html#Using_IMPUTE_with_the_HapMap_Data`. We ignored the case-control statuses automatically generated by HAPGEN2, and only used simulated controls for later analyses. Sample (1) was used as our large European sample (EA). Samples (2) and (3) were used as the ancestral populations for our admixed samples (ADM_{12}, ADM_5).

Simulating admixture

Our simulations focus on a single 1Mbp genomic region, which is relatively small, so we assume that ancestry is constant over this region (i.e., there was no recombination) because the probability of crossover in a region of this size is very small. We simulated admixture by sampling without replacement two haplotypes for each person from the total set of 136,000 haplotypes (68,000 CEU haplotypes and 68,000 YRI haplotypes), with the probability of sampling a YRI haplotype equal to the admixture proportion (0.2 or 0.4). In the end, each diploid admixed individual had either zero, one, or two CEU haplotypes, and two, one, or zero YRI haplotypes, respectively. This simulation approach is similar to that of Price et al. (2009), but without allowing for recombination. In total, we created four admixed populations, denoted by $\mathrm{ADM}_{n,pr}$ where $n \in \{5, 12\}$ for sample sizes of 5,000 and 12,000 individuals, respectively, and $pr \in \{0.2, 0.4\}$ for admixture proportions of 0.2 and 0.4, respectively.

3.2 Simulating genetic association regions

Simulating phenotypes

We simulated phenotypes (trait values) according to the additive linear model

$$Y_i = \sum_{j \in \mathcal{S}} g_{ij}\alpha_j + \epsilon_i,\ \epsilon_i \sim_{iid} N(0, \sigma_i^2)$$

where the residual variance σ_i^2 differed by population (CEU: $\sigma_i^2 = 472$, ADM$_{n,0.2}$: $\sigma_i^2 = 435$, ADM$_{n,0.4}$: $\sigma_i^2 = 416$), and the effect size was constant ($\alpha_j = 37$). We fixed the effect size in order to focus on the effect of LD on the results, as the effect of having different effect sizes across populations is generally predictable. These residual variance values and effect sizes were selected based on observed values for the pulmonary trait Forced Expiratory Volume in 1 second (FEV1) in the HCHS/SOL. Note that increasing/decreasing residual variance causes decrease/increase in power, a similar effect to decreasing/increasing sample size, and it has minor effect on the results, as long as the basic framework, of much higher power in the CEU population compared to the discovery ADM population, is maintained.

Genetic architecture at the locus

The genetic architecture at the locus is defined by the choice of causal SNPs. In these simulations we focus on the effect of LD, given the availability of only tag SNPs in the GWAS. Therefore, we considered simple scenarios of either 1 or 2 causal SNPs in the locus, where the causal SNP(s) were potentially causal in haplotypes of only one of the ancestries (CEU or YRI). Note that causality is attached to the specific ancestry of the haplotype, rather than to the final admixture status. Therefore, if a SNP was not causal in YRI, it did not have any effect on the phenotype of a person from ADM$_{n,0.2}$ with both haplotypes of YRI ancestry, but it did have an effect on the phenotype of a person from ADM$_{n,0.2}$ with one or two CEU haplotypes. More specifically, we considered the following genetic architectures, described also in Table SS4. For each scenario, we simulated all possible combinations of SNP selections from the simulated haplotypes, based on their polymorphism/monomorphism status in the CEU and YRI ancestries.

1. *Single causal SNP*: A single (the same) causal SNP which is polymorphic in both ancestries

(254 settings).

2. *No causal SNP in YRI*: A single causal SNP which is monomorphic in YRI (7 settings).

3. *Additional causal SNP in YRI*: Two causal SNPs; the first as in scenario (1), the second is polymorphic only in YRI (5,334 settings).

4. *Different causal SNPs*: A single causal SNP in each population, but which SNP is causal differs because the causal SNP in CEU is monomorphic in YRI, and vice versa (147 settings).

		SNP 1		SNP 2	
	Number of combinations	CEU	YRI	CEU	YRI
Scenario 1	254	MAF>0	MAF>0	–	–
Scenario 2	7	MAF>0	MAF$=0$	–	–
Scenario 3	5,334	MAF>0	MAF>0	MAF$=0$	MAF>0
Scenario 4	147	MAF>0	MAF$=0$	MAF$=0$	MAF>0

Table S4: The four simulations scenarios according to the number of causal SNPs and their polymorphism/monomorphism status in the two ancestral populations CEU and YRI. In the two first scenarios, there is one causal SNP in the two ancestries, and it is the same. In Scenarios 3 and 4, there are two potential causal SNPs. Here, MAF>0 denotes that a SNP is polymorphic (and causal) in the given ancestry, and MAF$=0$ denotes that it is monomorphic (and therefore cannot be causal) in the ancestry. The number of combinations is the number of possible SNP 1 and SNP 2 selections in the simulated haplotype that satisfy the polymorphism/monomorphism restrictions.

3.3 Figures comparing distributions of squared-root mean squared prediction error (RMSPE) simulations

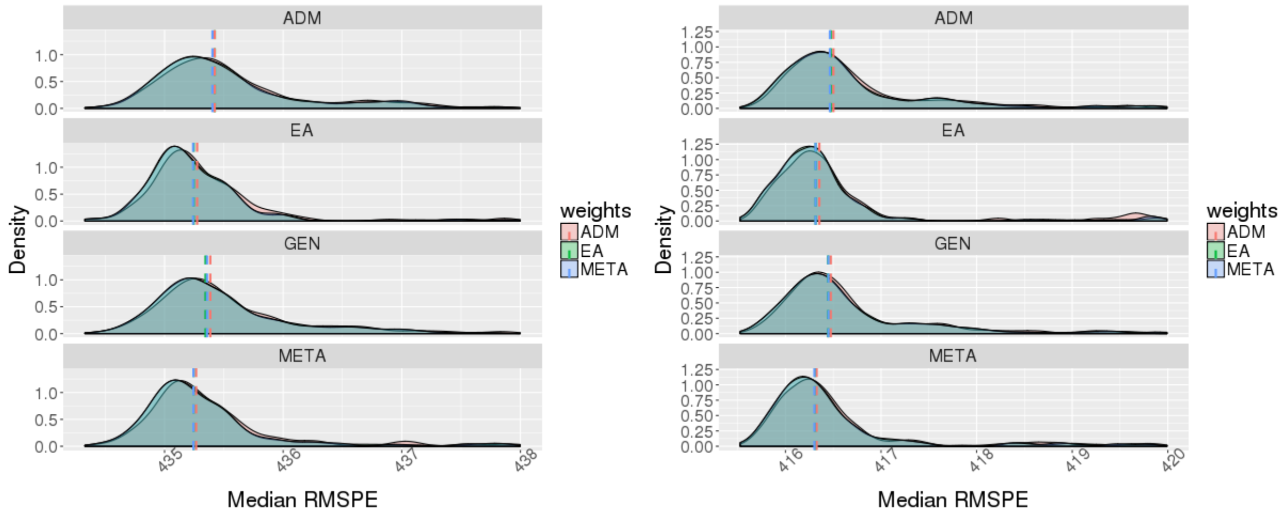

Figure S7: The smoothed distribution of median RMSPEs in simulation scenario 1. Each median was computed across 500 repetitions of the simulations, and the distribution is over all combinations of causal SNPs selection. The left panel corresponds to training datasets were EA and $\text{ADM}_{12,0.2}$ and the test data was $\text{ADM}_{5,0.4}$, and the right panel corresponds to the setting where the training datasets were EA and $\text{ADM}_{12,0.4}$ and the test data was $\text{ADM}_{5,0.2}$. Dashed vertical lines correspond to median of the plotted distribution. In the right panel, the lines corresponding to EA and META weights often overlap.

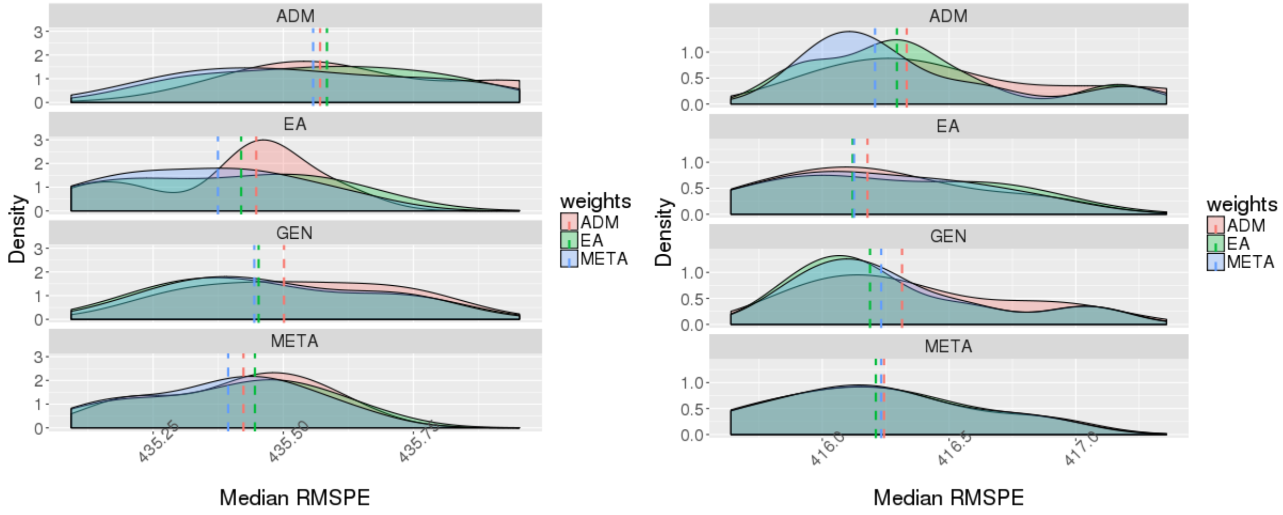

Figure S8: The smoothed distribution of median RMSPEs in simulation scenario 2. Each median was computed across 500 repetitions of the simulations, and the distribution is over all combinations of causal SNPs selection. The left panel corresponds to training datasets were EA and $\text{ADM}_{12,0.2}$ and the test data was $\text{ADM}_{5,0.4}$, and the right panel corresponds to the setting where the training datasets were EA and $\text{ADM}_{12,0.4}$ and the test data was $\text{ADM}_{5,0.2}$. Dashed vertical lines correspond to median of the plotted distribution. In the right panel, the lines corresponding to EA and META weights often overlap.

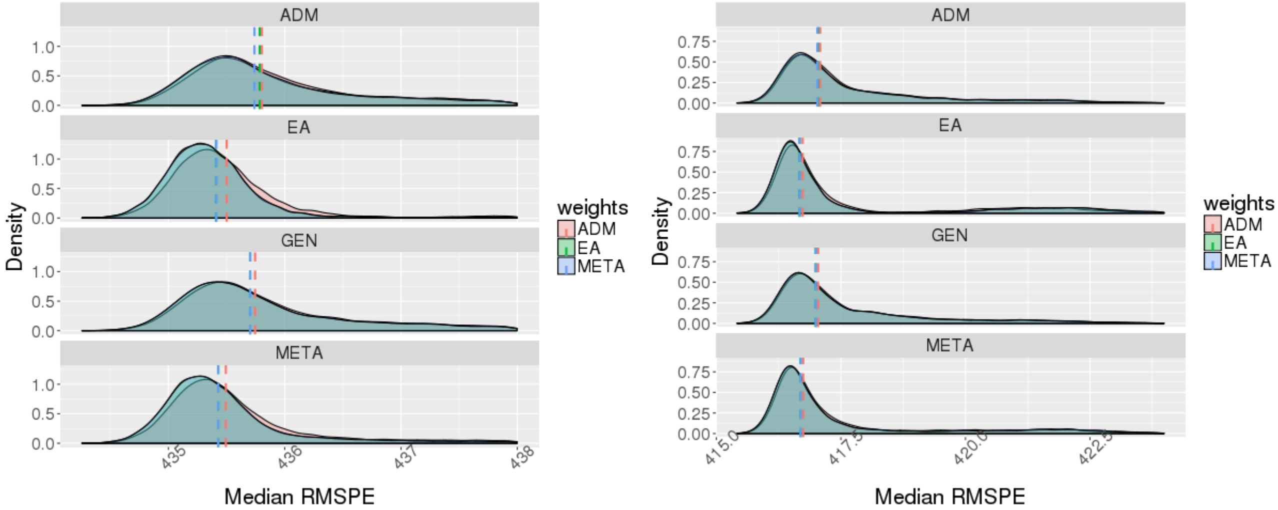

Figure S9: The smoothed distribution of median RMSPEs in simulation scenario 3. Each median was computed across 500 repetitions of the simulations, and the distribution is over all combinations of causal SNPs selection. The left panel corresponds to training datasets were EA and $\text{ADM}_{12,0.2}$ and the test data was $\text{ADM}_{5,0.4}$, and the right panel corresponds to the setting where the training datasets were EA and $\text{ADM}_{12,0.4}$ and the test data was $\text{ADM}_{5,0.2}$.Dashed vertical lines correspond to median of the plotted distribution. In the right panel, the lines corresponding to EA and META weights often overlap.

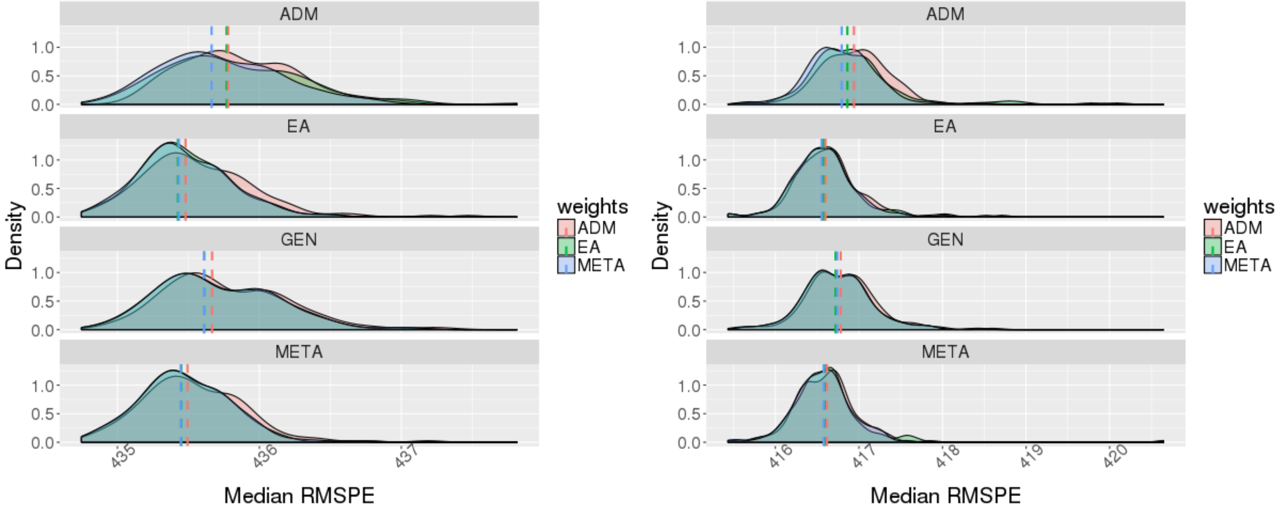

Figure S10: The smoothed distribution of median RMSPEs in simulation scenario 4. Each median was computed across 500 repetitions of the simulations, and the distribution is over all combinations of causal SNPs selection. The left panel corresponds to training datasets were EA and $\text{ADM}_{12,0.2}$ and the test data was $\text{ADM}_{5,0.4}$, and the right panel corresponds to the setting where the training datasets were EA and $\text{ADM}_{12,0.4}$ and the test data was $\text{ADM}_{5,0.2}$.Dashed vertical lines correspond to median of the plotted distribution. In the right panel, the lines corresponding to EA and META weights often overlap.

4 Performance of PRSs in the WHI SHARe African Americans

We study the generalizability of our results to African Americans. African Americans are admixed with European and African ancestry. For the same PRSs constructed based on EA GWAS and the HCHS/SOL and reported in the main manuscript, we provide results form evaluations on WHI African Americans women (n=8,350). Figure S11 demonstrates the highest variance explained obtained by the highest performing EA-based PRS (SNP selection by EA GWAS, and weights by EA GWAS), and the highest variance explained by any of the approaches. Table S5 provides information about the best performing EA-based PRS, and the overall best performing PRS. Performance was measured in variance explained.

	Best performing PRS					Best EA-based performing PRS		
Trait	Selection	Weights	Threshold	# SNPs	Variance explained	Threshold	# SNPs	Variance explained
Height	EA	Meta	1e-04	5,789	4.37	1e-04	5,789	4.33
BMI	EA	Meta	0.5	158,075	2.06	0.01	6,585	1.67
WC	EA	None	0.05	22,122	1.24	0.01	5,951	0.76
HIP	EA	None	0.01	6,329	1.38	0.01	6,329	0.89
WHR	EA	None	0.05	23,304	0.78	0.05	23,304	0.54
PLT	EA	SOL	1e-05	114	1.67	1e-05	114	1.51
WBC	EA	SOL	5e-08	30	12.67	1e-05	69	10.95
HGB	GEN	SOL	0.99 (*r*-value)	41	0.59	1e-06	44	0.45
SBP	META	None	0.05	34210	0.12	1e-05	70	0.01
DBP	META	None	0.5	216,639	0.09	1e-04	192	0.02
MAP	META	EA	0.05	31,755	0.13	0.5	183,970	0.01
PP	EA	SOL	1e-04	179	0.22	0.5	186,907	0.03

Table S5: Characteristics and performance, in terms of variance explained, of the highest performing EA-based PRS and highest performing PRS across all approaches, for all investigated traits, in WHI African Americans. The EA-based PRS selected SNPs based on EA GWAS results, with pruning based on EA populations from 1000 Genomes. The weights used in the PRSs were effect sizes from the EA GWASs. In the best performing GWAS, SNPs were selected based on either EA GWASs, meta-analysis of EA and HCHS/SOL GWASs, or Generalization analysis performed based discovery in EA GWAS and generalization in the HCHS/SOL GWAS. SNP clumping was based on EA populations from 1000 Genomes. Weights were based on EA GWAS, Meta-analysis of EA and HCHS/SOL GWAS (Meta), HCHS/SOL GWAS (SOL), or 'None' - a simple sum of trait-increasing alleles.

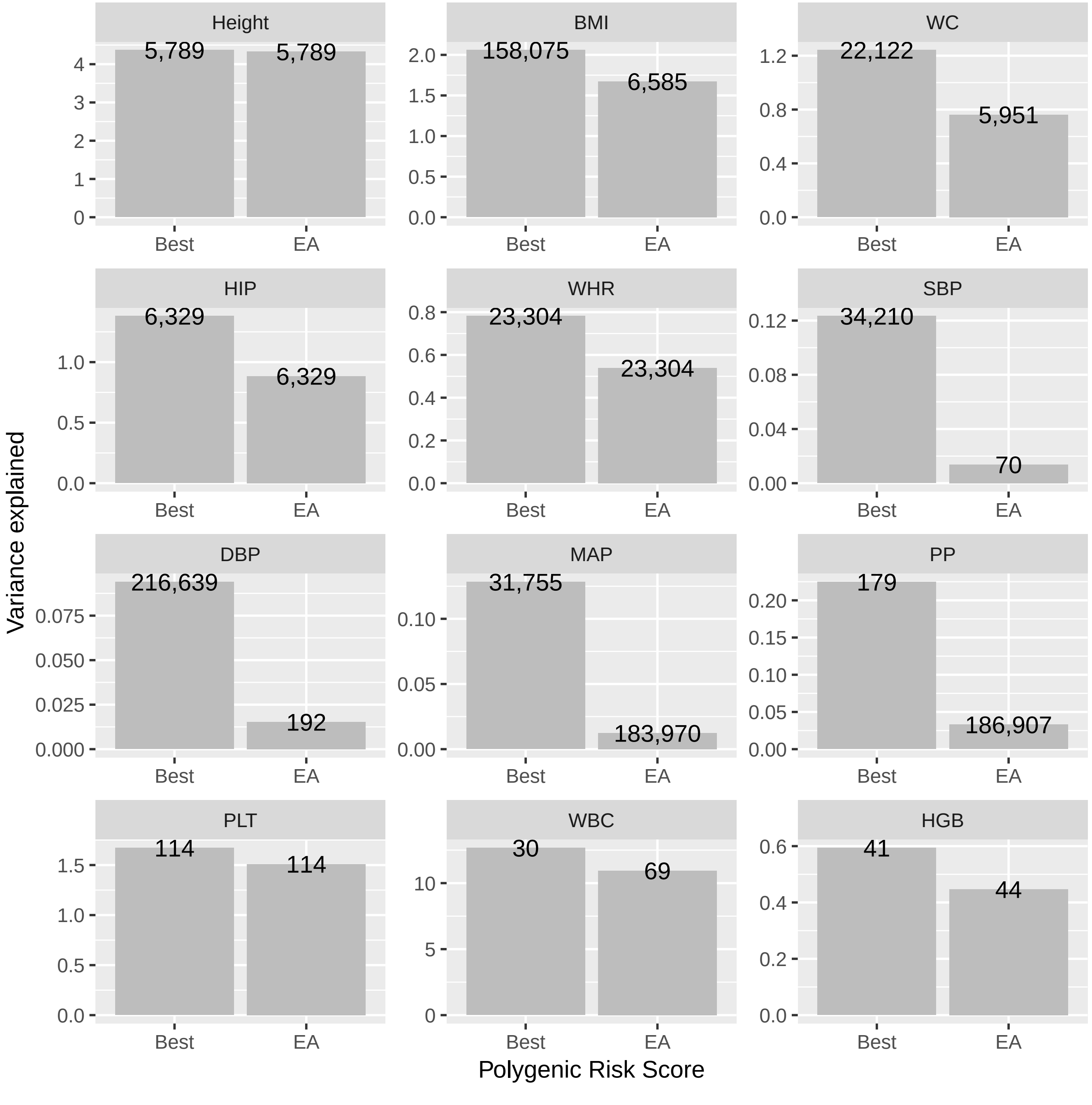

Figure S11: Variance explained by the highest performing EA-based PRS and highest performing PRS across all approaches, for all investigated traits, in WHI African Americans. The numbers on the bars represent the number of SNPs used in the PRS. Table S5 provides more details about the PRSs, including p-value or r-value threshold, weights used, etc.

References

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D. and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS genetics*, **5** e1000519.

Su, Z., Marchini, J. and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, **27** 2304–2305.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, **437** 1299–1320.

5 Comparisons of meta-analysis based PRSs with and without clumping based on an EA reference panel

	Clumping based on EA reference panel		SNPs pruning based on base-pair distance	
Trait	*p*-value threshold	variance explained	*p*-value threshold	variance explained
Height	1e-05	0.27	1e-06	10.09
BMI	0.001	0.14	0.001	4.07
WC	0.001	0.39	1e-04	3.19
HIP	0.001	0.08	1e-04	3.87
WHR	0.001	0.96	0.001	1.31
PLT	1e-04	0.74	1e-05	3.95
WBC	5e-08	1.11	1e-04	3.23
HGB	1e-07	0.03	1e-05	1.33
SBP	1e-05	0.59	1e-05	0.32
DBP	1e-04	0.28	5e-08	0.39
MAP	1e-05	0.83	1e-05	0.74
PP	1e-04	0.47	1e-04	0.35

Table S6: Variance explained in WHI Hispanic/Latina women analysis for each of the investigated traits, in an approach that meta-analyze the EA GWAS results with HCHS/SOL GWAS results, and then constructs PRSs solely based on that (i.e. both SNP selection and weights are based on the meta-analysis results). We compared two approaches: SNP selection based on LD clumping with an EA reference panel from 1000 genome (left part of the table), and SNP selection based on clumping, in which we selected the SNP with the lower *p*-value, then removed all SNPs in a 1Mbp around it, and continued with SNP selection until no more SNPs were left (right part of the table). The *p*-value threshold is the one providing the best results, where considered threshold were: 0.001, 1e-04, 1e-05, 1e-06, 1e-07, 5e-08.

	Clumping based on EA reference panel		SNPs pruning based on base-pair distance	
Trait	p-value threshold	variance explained	p-value threshold	variance explained
Height	1e-07	0.11	1e-06	3.25
BMI	0.001	0.04	1e-04	1.44
WC	0.001	0.01	1e-04	0.71
HIP	1e-05	0.12	0.001	0.90
WHR	1e-04	0.23	1e-07	0.33
PLT	1e-07	0.19	1e-07	1.50
WBC	1e-07	6.35	5e-08	9.97
HGB	1e-07	0.16	1e-05	0.53
SBP	1e-05	0.06	1e-05	0.04
DBP	1e-06	0.02	5e-08	0.01
MAP	1e-05	0.03	1e-06	0.02
PP	1e-04	0.10	1e-07	0.09

Table S7: Variance explained in WHI African American women analysis for each of the investigated traits, in an approach that meta-analyze the EA GWAS results with HCHS/SOL GWAS results, and then constructs PRSs solely based on that (i.e. both SNP selection and weights are based on the meta-analysis results). We compared two approaches: SNP selection based on LD clumping with an EA reference panel from 1000 genome (left part of the table), and SNP selection based on clumping, in which we selected the SNP with the lower p-value, then removed all SNPs in a 1Mbp around it, and continued with SNP selection until no more SNPs were left (right part of the table). The p-value threshold is the one providing the best results, where considered threshold were: 0.001, 1e-04, 1e-05, 1e-06, 1e-07, 5e-08.