

Machine learning models predicting blood pressure phenotypes by combining multiple polygenic risk scores: Supplementary Materials

Supplementary Tables	3
Supplementary Table 1: Participant characteristics of the TOPMed training dataset.	3
Supplementary Table 2: Participant characteristics of the TOPMed testing dataset.....	3
Supplementary Table 3: Number of TOPMed individuals contributed by parent studies.....	4
Supplementary Table 4: Performance of the baseline model using global PRSs and TOPMed data.....	5
Supplementary Table 5: Performance of the genetic and ensemble models using global PRSs and TOPMed data.....	6
Supplementary Table 6: Performance of the genetic and ensemble models using global PRSs in TOPMed data developed with PRS-CSx.	9
Supplementary Table 7: Hyperparameter choice for models' training procedure.	10
Supplementary Table 8: P-value threshold optimizing the PVE across groups in a genetic prediction model using one of UKB-ICBP based PRSs.	11
Supplementary Table 9: Cross-validated training dataset PVEs from genetic models with single PRS based on the UKB-ICBP.....	12
Supplementary Table 10: Computational needs of compared models	15
Supplementary Table 11: Performance of the genetic and ensemble models using local PRSs in TOPMed data.....	16
Supplementary Table 12: Participant characteristics of the MGB Biobank dataset.....	17
Supplementary Table 13: GWAS summary statistics used for SBP and DBP PRS development.	17
Supplementary Figures.....	18
Supplementary Figure 1: Comparison of baseline cross-validated PVE with and without inclusion of genetic PCs.	18
Supplementary Figure 2: Estimated PVE of genetic models fitted using XGBoost and linear models using global PRS.	19
Supplementary Figure 3: Performance of genetic and ensemble models fitted using XGBoost, linear regression and LASSO using local PRSs in TOPMed test dataset.....	20
Supplementary Figure 4: Estimated phenotypic PVE of baseline models in the MGB Biobank data fitted using XGBoost.....	21
Supplementary Figure 5: Estimated PVE of genetic and ensemble models in the MGB Biobank.....	22

Supplementary Note 1: Descriptions of participating TOPMed studies.....	23
Amish.....	23
ARIC.....	24
BioMe.....	25
CARDIA.....	26
CFS.....	27
CHS.....	28
COPDGene.....	30
FHS.....	31
GENOA.....	32
GenSalt.....	33
HCHS/SOL.....	34
JHS.....	35
MESA.....	37
THRV.....	38
WHI.....	39
Supplementary Note 2: Removal of overlap GWAS.....	40
Supplementary Note 3: Sensitivity analysis using PRS-CSx-based global PRS.....	42
Supplementary Note 4: Descriptions of MGB biobank dataset.....	43
Supplementary Note 5: TOPMed and CCDG acknowledgements.....	47
Supplementary References.....	59

Supplementary Tables

Supplementary Table 1: Participant characteristics of the TOPMed training dataset.

Characteristic	White	Black	Hispanic/Latino	Asian	Other/Unknown
N	19,829	10,961	8,018	4,267	481
Gender¹					
Female	12,310 (62%)	6,713 (61%)	4,838 (60%)	2,143 (50%)	198 (41%)
Male	7,519 (38%)	4,248 (39%)	3,180 (40%)	2,124 (50%)	283 (59%)
Age²	59 (48, 68)	54 (47, 63)	53 (43, 62)	47 (39, 55)	59 (50, 67)
SBP²	125 (113, 141)	133 (119, 150)	127 (113, 146)	122 (110, 137)	134 (116, 152)
DBP²	75 (68, 83)	80 (72, 89)	76 (68, 85)	75 (68, 85)	76 (66, 86)
BMI²	26.5 (23.6, 30.0)	29.1 (25.2, 34.0)	29.0 (25.9, 33.0)	23.9 (21.7, 26.2)	27.3 (24.3, 31.5)
Hypertensive¹	11,031 (56%)	7,722 (70%)	4,696 (59%)	2,160 (51%)	323 (67%)

¹n (%)

²Median (IQR)

TOPMed training dataset characteristics combined over the studies broken into race/ethnicity background.

Supplementary Table 2: Participant characteristics of the TOPMed testing dataset.

Characteristic	White	Black	Hispanic/Latino	Asian
N	10,839	3,626	3,886	388
Gender¹				
Female	7,908 (73%)	2,452 (68%)	2,270 (58%)	233 (60%)
Male	2,931 (27%)	1,174 (32%)	1,616 (42%)	155 (40%)
Age²	62 (52, 70)	57 (45, 66)	51 (42, 60)	63 (55, 70)
SBP²	127 (113, 144)	132 (116, 150)	125 (113, 141)	128 (111, 149)
DBP²	75 (68, 83)	79 (71, 88)	76 (68, 84)	76 (68, 84)
BMI²	26.2 (23.4, 30.0)	29.0 (25.0, 33.6)	29.0 (26.0, 33.0)	24.0 (22.0, 27.0)
Hypertensive¹	6,243 (58%)	2,381 (66%)	2,037 (52%)	221 (57%)

¹n (%)

²Median (IQR)

TOPMed testing dataset characteristics combined over the studies broken into race/ethnicity background.

Supplementary Table 3: Number of TOPMed individuals contributed by parent studies.

Study	White	Black	Hispanic / Latino	Asian	Other/Unknown
N	30,668	14,587	11,904	4,655	481
Amish ¹	1,098 (3.6%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
ARIC ¹	5,872 (19%)	1,294 (8.9%)	0 (0%)	0 (0%)	0 (0%)
BioMe ¹	1,675 (5.5%)	1,870 (13%)	3,105 (26%)	103 (2.2%)	481 (100%)
CARDIA ¹	1,647 (5.4%)	1,368 (9.4%)	0 (0%)	0 (0%)	0 (0%)
CFS ¹	282 (0.9%)	379 (2.6%)	0 (0%)	0 (0%)	0 (0%)
CHS ¹	2,686 (8.8%)	654 (4.5%)	31 (0.3%)	0 (0%)	0 (0%)
COPDGene ¹	3,628 (12%)	2,208 (15%)	0 (0%)	0 (0%)	0 (0%)
FHS ¹	3,084 (10%)	0 (0%)	11 (<0.1%)	0 (0%)	0 (0%)
GENOA ¹	0 (0%)	1,041 (7.1%)	0 (0%)	0 (0%)	0 (0%)
GenSalt ¹	0 (0%)	0 (0%)	0 (0%)	1,791 (38%)	0 (0%)
HCHS/SOL ¹	0 (0%)	0 (0%)	7,455 (63%)	0 (0%)	0 (0%)
JHS ¹	0 (0%)	3,281 (2%)	0 (0%)	0 (0%)	0 (0%)
MESA ¹	1,811 (5.9%)	1,079 (7.4%)	1,001 (8.4%)	582 (13%)	0 (0%)
THRV ¹	0 (0%)	0 (0%)	0 (0%)	10986 (43%)	0 (0%)
WHI ¹	8,885 (29%)	1,413 (9.7%)	301 (2.5%)	193 (4.1%)	0 (0%)

¹n (%)

TOPMed testing dataset characteristics broken by study and race/ethnicity background.

Supplementary Table 4: Performance of the baseline model using global PRs and TOPMed data.

Phenotype	Group	Model	Training set (N)	Testing set (N)	Training PVE	Testing PVE
SBP	Overall	XGBoost	43556	18748	31.28%	30.05%
SBP	Black	XGBoost	10914	3674	30.18%	33.35%
SBP	Asian	XGBoost	4281	374	25.04%	19.69%
SBP	White	XGBoost	19853	10823	29.40%	28.93%
SBP	Hispanic/Latino	XGBoost	8027	3877	31.64%	27.66%
DBP	Overall	XGBoost	43472	18768	15.93%	17.35%
DBP	Black	XGBoost	10889	3657	12.95%	22.35%
DBP	Asian	XGBoost	4220	403	17.30%	11.89%
DBP	White	XGBoost	19786	10877	13.18%	13.38%
DBP	Hispanic/Latino	XGBoost	8093	3831	13.26%	17.75%
SBP	Overall	Linear Regression	43556	18748	24.00%	28.07%
SBP	Black	Linear Regression	10914	3674	22.89%	32.03%
SBP	Asian	Linear Regression	4281	374	13.40%	18.52%
SBP	White	Linear Regression	19853	10823	22.97%	27.22%
SBP	Hispanic/Latino	Linear Regression	8027	3877	24.36%	23.98%
DBP	Overall	Linear Regression	43472	18768	11.80%	14.31%
DBP	Black	Linear Regression	10889	3657	9.01%	18.77%
DBP	Asian	Linear Regression	4220	403	10.17%	11.54%
DBP	White	Linear Regression	19786	10877	9.84%	11.40%
DBP	Hispanic/Latino	Linear Regression	8093	3831	8.70%	12.10%

Supplementary Table 5: Performance of the genetic and ensemble models using global PRSs and TOPMed data.

Phenotype	Model	Group	Training set (N)	Testing set (N)	Training PVE Genetic model	Testing PVE Genetic model	Training PVE Ensemble model	Testing PVE Ensemble model	Model complexity level
SBP	XGBoost	Overall	43556	18748	3.10%	4.80%	33.40%	33.40%	Level 1
SBP	XGBoost	Black	10914	3674	0.80%	0.70%	30.80%	33.80%	Level 1
SBP	XGBoost	Asian	4281	374	2.90%	1.60%	27.20%	21.00%	Level 1
SBP	XGBoost	White	19853	10823	6.00%	7.70%	33.60%	34.40%	Level 1
SBP	XGBoost	Hispanic/Latino	8027	3877	0.40%	0.90%	31.90%	28.30%	Level 1
SBP	XGBoost	Overall	43556	18748	3.70%	5.10%	33.80%	33.60%	Level 2
SBP	XGBoost	Black	10914	3674	1.20%	1.10%	31.00%	34.10%	Level 2
SBP	XGBoost	Asian	4281	374	3.60%	2.40%	27.70%	21.60%	Level 2
SBP	XGBoost	White	19853	10823	6.20%	7.60%	33.80%	34.30%	Level 2
SBP	XGBoost	Hispanic/Latino	8027	3877	1.60%	2.00%	32.80%	29.10%	Level 2
SBP	XGBoost	Overall	43556	18748	5.20%	6.00%	34.90%	34.20%	Level 3
SBP	XGBoost	Black	10914	3674	2.60%	2.60%	32.00%	35.10%	Level 3
SBP	XGBoost	Asian	4281	374	5.90%	3.80%	29.50%	22.80%	Level 3
SBP	XGBoost	White	19853	10823	7.10%	7.80%	34.40%	34.50%	Level 3
SBP	XGBoost	Hispanic/Latino	8027	3877	4.70%	4.60%	34.80%	31.00%	Level 3
SBP	Linear Regression	Overall	43556	18748	2.80%	4.40%	33.20%	33.10%	Level 1
SBP	Linear Regression	Black	10914	3674	0.50%	0.80%	30.60%	33.90%	Level 1
SBP	Linear Regression	Asian	4281	374	2.50%	1.40%	26.90%	20.80%	Level 1
SBP	Linear Regression	White	19853	10823	5.50%	7.00%	33.30%	33.90%	Level 1
SBP	Linear Regression	Hispanic/Latino	8027	3877	0.20%	0.90%	31.80%	28.30%	Level 1
SBP	Linear Regression	Overall	43556	18748	3.20%	4.80%	33.50%	33.40%	Level 2
SBP	Linear Regression	Black	10914	3674	0.80%	1.20%	30.70%	34.20%	Level 2
SBP	Linear Regression	Asian	4281	374	3.40%	2.00%	27.60%	21.30%	Level 2
SBP	Linear Regression	White	19853	10823	5.60%	7.10%	33.30%	34.00%	Level 2
SBP	Linear Regression	Hispanic/Latino	8027	3877	1.30%	1.80%	32.50%	28.90%	Level 2

SBP	Linear Regression	Overall	43556	18748	4.30%	6.40%	34.20%	34.50%	Level 3
SBP	Linear Regression	Black	10914	3674	1.70%	3.20%	31.30%	35.50%	Level 3
SBP	Linear Regression	Asian	4281	374	4.90%	3.80%	28.70%	22.70%	Level 3
SBP	Linear Regression	White	19853	10823	5.90%	7.90%	33.60%	34.50%	Level 3
SBP	Linear Regression	Hispanic/Latino	8027	3877	4.20%	5.70%	34.50%	31.70%	Level 3
DBP	XGBoost	Overall	43472	18768	4.20%	4.70%	19.40%	21.20%	Level 1
DBP	XGBoost	Black	10889	3657	1.90%	0.80%	14.60%	23.00%	Level 1
DBP	XGBoost	Asian	4220	403	5.00%	0.90%	21.40%	12.70%	Level 1
DBP	XGBoost	White	19786	10877	6.90%	7.30%	19.10%	19.70%	Level 1
DBP	XGBoost	Hispanic/Latino	8093	3831	1.40%	1.50%	14.50%	19.00%	Level 1
DBP	XGBoost	Overall	43472	18768	4.50%	4.60%	19.70%	21.10%	Level 2
DBP	XGBoost	Black	10889	3657	2.10%	0.30%	14.70%	22.60%	Level 2
DBP	XGBoost	Asian	4220	403	5.60%	2.90%	22.00%	14.40%	Level 2
DBP	XGBoost	White	19786	10877	6.90%	6.70%	19.20%	19.20%	Level 2
DBP	XGBoost	Hispanic/Latino	8093	3831	2.50%	2.70%	15.40%	20.00%	Level 2
DBP	XGBoost	Overall	43472	18768	6.50%	5.60%	21.40%	22.00%	Level 3
DBP	XGBoost	Black	10889	3657	4.60%	1.70%	16.90%	23.60%	Level 3
DBP	XGBoost	Asian	4220	403	7.90%	3.40%	23.90%	14.90%	Level 3
DBP	XGBoost	White	19786	10877	8.00%	7.20%	20.10%	19.60%	Level 3
DBP	XGBoost	Hispanic/Latino	8093	3831	5.70%	5.00%	18.20%	21.90%	Level 3
DBP	Linear Regression	Overall	43472	18768	2.60%	4.40%	18.10%	21.00%	Level 1
DBP	Linear Regression	Black	10889	3657	0.40%	0.90%	13.30%	23.10%	Level 1
DBP	Linear Regression	Asian	4220	403	2.50%	1.80%	19.40%	13.50%	Level 1
DBP	Linear Regression	White	19786	10877	5.70%	6.80%	18.10%	19.30%	Level 1
DBP	Linear Regression	Hispanic/Latino	8093	3831	-0.60%	1.10%	12.80%	18.70%	Level 1
DBP	Linear Regression	Overall	43472	18768	2.90%	4.50%	18.30%	21.10%	Level 2
DBP	Linear Regression	Black	10889	3657	0.90%	1.20%	13.70%	23.30%	Level 2
DBP	Linear Regression	Asian	4220	403	3.00%	2.70%	19.70%	14.30%	Level 2

DBP	Linear Regression	White	19786	10877	5.40%	6.50%	17.80%	19.00%	Level 2
DBP	Linear Regression	Hispanic/Latino	8093	3831	0.60%	2.40%	13.80%	19.70%	Level 2
DBP	Linear Regression	Overall	43472	18768	3.70%	5.20%	19.00%	21.60%	Level 3
DBP	Linear Regression	Black	10889	3657	2.00%	2.20%	14.70%	24.00%	Level 3
DBP	Linear Regression	Asian	4220	403	3.80%	3.40%	20.40%	14.90%	Level 3
DBP	Linear Regression	White	19786	10877	5.40%	6.60%	17.90%	19.10%	Level 3
DBP	Linear Regression	Hispanic/Latino	8093	3831	2.60%	4.30%	15.50%	21.30%	Level 3

Performance results (attained PVEs) from the genetic models (PVEs of predicting residuals from the baseline model) and ensemble models (PVEs of predicting the raw trait) estimated in cross validation on the training dataset, and from the independent test dataset using global PRS in TOPMed dataset.

Supplementary Table 6: Performance of the genetic and ensemble models using global PRSs in TOPMed data developed with PRS-CSx.

Phenotype	Model	Group	Training set (N)	Testing set (N)	Training PVE Genetic model	Testing PVE Genetic model	Training PVE Ensemble model	Testing PVE Ensemble model
SBP	Linear Regression	Overall	43556	18748	2.11%	3.20%	32.75%	32.30%
SBP	Linear Regression	Black	10914	3674	0.54%	1.60%	30.59%	34.50%
SBP	Linear Regression	Asian	4281	374	2.34%	2.80%	26.77%	22.00%
SBP	Linear Regression	White	19853	10823	3.20%	4.00%	31.72%	31.80%
SBP	Linear Regression	Hispanic/Latino	8027	3877	1.84%	2.30%	32.88%	29.30%
SBP	XGBoost	Overall	43556	18748	2.39%	3.20%	32.94%	32.30%
SBP	XGBoost	Black	10914	3674	0.84%	1.60%	30.80%	34.40%
SBP	XGBoost	Asian	4281	374	2.56%	2.80%	26.93%	22.00%
SBP	XGBoost	White	19853	10823	3.48%	4.10%	31.92%	31.80%
SBP	XGBoost	Hispanic/Latino	8027	3877	2.12%	2.20%	33.07%	29.20%
DBP	Linear Regression	Overall	43472	18768	1.93%	2.90%	17.23%	19.30%
DBP	Linear Regression	Black	10889	3657	0.83%	1.40%	13.67%	23.20%
DBP	Linear Regression	Asian	4220	403	2.24%	2.60%	18.52%	14.00%
DBP	Linear Regression	White	19786	10877	2.83%	3.60%	15.29%	16.10%
DBP	Linear Regression	Hispanic/Latino	8093	3831	1.47%	2.30%	14.26%	19.30%
DBP	XGBoost	Overall	43472	18768	2.11%	2.70%	17.54%	19.60%
DBP	XGBoost	Black	10889	3657	1.12%	1.70%	14.47%	23.70%
DBP	XGBoost	Asian	4220	403	2.52%	2.70%	18.53%	14.30%
DBP	XGBoost	White	19786	10877	2.85%	3.20%	15.22%	16.20%
DBP	XGBoost	Hispanic/Latino	8093	3831	1.89%	2.50%	13.91%	19.80%

Performance results (attained PVEs) from the genetic models (PVEs of predicting residuals from the baseline model) and ensemble models (PVEs of predicting the raw trait) estimated in cross validation on the training dataset, and from the independent test dataset using global PRS in TOPMed dataset where PRS were developed using PRS-CSx.

Supplementary Table 7: Hyperparameter choice for models' training procedure.

Phenotype	Model	n estimator	max depth	min child weight	subsample	colsample by tree	lambda	alpha	gamma	eta
DBP	TOPMed baseline model	315	3	100	0.6	1	20	14	32	0.04
SBP	TOPMed baseline model	460	99	20	0.4	0.7	0	49	19	0.01
DBP	TOPMed model 1	90	3	1	0.8	0.9	37	50	45	0.05
SBP	TOPMed model 1	290	1	57	0.3	0.8	48	48	17	0.1
DBP	TOPMed model 2	348	30	74	0.1	0.7	13	37	31	0.01
SBP	TOPMed model 2	192	30	42	0.3	0.5	48	43	24	0.02
DBP	TOPMed model 3	697	2	29	0.9	0.6	31	8	38	0.02
SBP	TOPMed model 3	290	3	16	0.9	0.8	42	36	29	0.04
DBP	MGB Biobank Baseline Model	405	100	9	0.5	0.9	0	44	22	0.01
SBP	MGB Biobank Baseline Model	293	4	40	0.7	0.7	49	35	22	0.04

Hyperparameters and their value choices for models' training procedure after applying Optuna hyperparameter selection.

Supplementary Table 8: P-value threshold optimizing the PVE across groups in a genetic prediction model using one of UKB-ICBP based PRSs.

Group	SBP Linear regression	SBP Non-linear ML	DBP Linear regression	DBP Non-linear ML
Overall	0.001	0.01	0.001	0.01
Black	1e-06	1e-06	1e-06	0.001
Asian	1e-05	1e-05	0.001	0.001
White	0.01	0.01	0.01	0.01
Hispanic/Latino	1e-05	1e-05	1e-07	1e-07

This table provides the optimal p-value threshold from the clump & threshold PRS methodology, as evaluated in the TOPMed-training dataset. The genetic model was fit with a single PRS each time, using the overall group. Cross-validated PVE was computed for the overall and for each self-reported race/ethnicity group. The PRS that optimized the cross-validated PVE for each group is reported. Note that for the overall group, PVEs were higher in the non-linear ML models (fit using XGBoost package). Complete results reporting cross-validated PVE on the training dataset are provided in Supplementary Table 9.

Supplementary Table 9: Cross-validated training dataset PVEs from genetic models with single PRS based on the UKB-ICBP.

phenotype	Model	Group	Training dataset N	UKB-ICBB PRS p-value threshold	Training dataset cross-validated PVE
SBP	Linear Regression	Overall	43556	5E-08	2.18%
SBP	Linear Regression	Black	10914	5E-08	0.72%
SBP	Linear Regression	AsA	4281	5E-08	2.72%
SBP	Linear Regression	White	19853	5E-08	3.37%
SBP	Linear Regression	Hispanic/Latino	8027	5E-08	1.38%
SBP	Linear Regression	Overall	43556	1E-07	2.17%
SBP	Linear Regression	Black	10914	1E-07	0.74%
SBP	Linear Regression	AsA	4281	1E-07	2.65%
SBP	Linear Regression	White	19853	1E-07	3.41%
SBP	Linear Regression	Hispanic/Latino	8027	1E-07	1.28%
SBP	Linear Regression	Overall	43556	1E-06	2.38%
SBP	Linear Regression	Black	10914	1E-06	0.91%
SBP	Linear Regression	AsA	4281	1E-06	2.85%
SBP	Linear Regression	White	19853	1E-06	3.71%
SBP	Linear Regression	Hispanic/Latino	8027	1E-06	1.39%
SBP	Linear Regression	Overall	43556	1E-05	2.68%
SBP	Linear Regression	Black	10914	1E-05	0.83%
SBP	Linear Regression	AsA	4281	1E-05	3.20%
SBP	Linear Regression	White	19853	1E-05	4.12%
SBP	Linear Regression	Hispanic/Latino	8027	1E-05	1.88%
SBP	Linear Regression	Overall	43556	0.0001	2.76%
SBP	Linear Regression	Black	10914	0.0001	0.74%
SBP	Linear Regression	AsA	4281	0.0001	2.91%
SBP	Linear Regression	White	19853	0.0001	4.70%
SBP	Linear Regression	Hispanic/Latino	8027	0.0001	1.40%
SBP	Linear Regression	Overall	43556	0.001	2.89%
SBP	Linear Regression	Black	10914	0.001	0.79%
SBP	Linear Regression	AsA	4281	0.001	2.84%
SBP	Linear Regression	White	19853	0.001	5.32%
SBP	Linear Regression	Hispanic/Latino	8027	0.001	0.76%
SBP	Linear Regression	Overall	43556	0.01	2.76%
SBP	Linear Regression	Black	10914	0.01	0.53%
SBP	Linear Regression	AsA	4281	0.01	2.51%
SBP	Linear Regression	White	19853	0.01	5.54%
SBP	Linear Regression	Hispanic/Latino	8027	0.01	0.25%
SBP	XGBoost	Overall	43556	5E-08	2.29%
SBP	XGBoost	Black	10914	5E-08	0.82%
SBP	XGBoost	AsA	4281	5E-08	2.98%
SBP	XGBoost	White	19853	5E-08	3.47%
SBP	XGBoost	Hispanic/Latino	8027	5E-08	1.39%
SBP	XGBoost	Overall	43556	1E-07	2.33%
SBP	XGBoost	Black	10914	1E-07	0.92%
SBP	XGBoost	AsA	4281	1E-07	2.98%
SBP	XGBoost	White	19853	1E-07	3.57%
SBP	XGBoost	Hispanic/Latino	8027	1E-07	1.27%
SBP	XGBoost	Overall	43556	1E-06	2.51%
SBP	XGBoost	Black	10914	1E-06	1.01%
SBP	XGBoost	AsA	4281	1E-06	3.16%
SBP	XGBoost	White	19853	1E-06	3.82%
SBP	XGBoost	Hispanic/Latino	8027	1E-06	1.44%
SBP	XGBoost	Overall	43556	1E-05	2.81%
SBP	XGBoost	Black	10914	1E-05	1.00%
SBP	XGBoost	AsA	4281	1E-05	3.55%
SBP	XGBoost	White	19853	1E-05	4.23%
SBP	XGBoost	Hispanic/Latino	8027	1E-05	1.85%
SBP	XGBoost	Overall	43556	0.0001	2.90%
SBP	XGBoost	Black	10914	0.0001	0.91%
SBP	XGBoost	AsA	4281	0.0001	3.14%
SBP	XGBoost	White	19853	0.0001	4.81%

SBP	XGBoost	Hispanic/Latino	8027	0.0001	1.46%
SBP	XGBoost	Overall	43556	0.001	3.04%
SBP	XGBoost	Black	10914	0.001	0.91%
SBP	XGBoost	AsA	4281	0.001	3.18%
SBP	XGBoost	White	19853	0.001	5.47%
SBP	XGBoost	Hispanic/Latino	8027	0.001	0.82%
SBP	XGBoost	Overall	43556	0.01	3.08%
SBP	XGBoost	Black	10914	0.01	0.80%
SBP	XGBoost	AsA	4281	0.01	2.78%
SBP	XGBoost	White	19853	0.01	5.98%
SBP	XGBoost	Hispanic/Latino	8027	0.01	0.37%
DBP	Linear Regression	Overall	43472	5E-08	1.99%
DBP	Linear Regression	Black	10889	5E-08	0.93%
DBP	Linear Regression	AsA	4220	5E-08	2.24%
DBP	Linear Regression	White	19786	5E-08	2.83%
DBP	Linear Regression	Hispanic/Latino	8093	5E-08	1.65%
DBP	Linear Regression	Overall	43472	1E-07	2.01%
DBP	Linear Regression	Black	10889	1E-07	0.94%
DBP	Linear Regression	AsA	4220	1E-07	2.27%
DBP	Linear Regression	White	19786	1E-07	2.88%
DBP	Linear Regression	Hispanic/Latino	8093	1E-07	1.67%
DBP	Linear Regression	Overall	43472	1E-06	2.18%
DBP	Linear Regression	Black	10889	1E-06	0.98%
DBP	Linear Regression	AsA	4220	1E-06	2.33%
DBP	Linear Regression	White	19786	1E-06	3.31%
DBP	Linear Regression	Hispanic/Latino	8093	1E-06	1.51%
DBP	Linear Regression	Overall	43472	1E-05	2.33%
DBP	Linear Regression	Black	10889	1E-05	0.97%
DBP	Linear Regression	AsA	4220	1E-05	2.50%
DBP	Linear Regression	White	19786	1E-05	3.82%
DBP	Linear Regression	Hispanic/Latino	8093	1E-05	1.17%
DBP	Linear Regression	Overall	43472	0.0001	2.52%
DBP	Linear Regression	Black	10889	0.0001	0.93%
DBP	Linear Regression	AsA	4220	0.0001	2.53%
DBP	Linear Regression	White	19786	0.0001	4.37%
DBP	Linear Regression	Hispanic/Latino	8093	0.0001	1.05%
DBP	Linear Regression	Overall	43472	0.001	2.77%
DBP	Linear Regression	Black	10889	0.001	0.81%
DBP	Linear Regression	AsA	4220	0.001	2.80%
DBP	Linear Regression	White	19786	0.001	5.07%
DBP	Linear Regression	Hispanic/Latino	8093	0.001	0.89%
DBP	Linear Regression	Overall	43472	0.01	2.57%
DBP	Linear Regression	Black	10889	0.01	0.42%
DBP	Linear Regression	AsA	4220	0.01	2.51%
DBP	Linear Regression	White	19786	0.01	5.68%
DBP	Linear Regression	Hispanic/Latino	8093	0.01	-0.57%
DBP	XGBoost	Overall	43472	5E-08	2.98%
DBP	XGBoost	Black	10889	5E-08	1.73%
DBP	XGBoost	AsA	4220	5E-08	4.21%
DBP	XGBoost	White	19786	5E-08	3.67%
DBP	XGBoost	Hispanic/Latino	8093	5E-08	2.62%
DBP	XGBoost	Overall	43472	1E-07	3.10%
DBP	XGBoost	Black	10889	1E-07	1.78%
DBP	XGBoost	AsA	4220	1E-07	4.28%
DBP	XGBoost	White	19786	1E-07	3.83%
DBP	XGBoost	Hispanic/Latino	8093	1E-07	2.78%
DBP	XGBoost	Overall	43472	1E-06	3.22%
DBP	XGBoost	Black	10889	1E-06	1.80%
DBP	XGBoost	AsA	4220	1E-06	4.46%
DBP	XGBoost	White	19786	1E-06	4.22%
DBP	XGBoost	Hispanic/Latino	8093	1E-06	2.51%
DBP	XGBoost	Overall	43472	1E-05	3.33%
DBP	XGBoost	Black	10889	1E-05	1.79%
DBP	XGBoost	AsA	4220	1E-05	4.57%
DBP	XGBoost	White	19786	1E-05	4.60%
DBP	XGBoost	Hispanic/Latino	8093	1E-05	2.26%
DBP	XGBoost	Overall	43472	0.0001	3.60%

DBP	XGBoost	Black	10889	0.0001	1.83%
DBP	XGBoost	AsA	4220	0.0001	4.74%
DBP	XGBoost	White	19786	0.0001	5.18%
DBP	XGBoost	Hispanic/Latino	8093	0.0001	2.27%
DBP	XGBoost	Overall	43472	0.001	3.92%
DBP	XGBoost	Black	10889	0.001	1.86%
DBP	XGBoost	AsA	4220	0.001	4.93%
DBP	XGBoost	White	19786	0.001	5.92%
DBP	XGBoost	Hispanic/Latino	8093	0.001	2.20%
DBP	XGBoost	Overall	43472	0.01	3.96%
DBP	XGBoost	Black	10889	0.01	1.54%
DBP	XGBoost	AsA	4220	0.01	4.70%
DBP	XGBoost	White	19786	0.01	6.76%
DBP	XGBoost	Hispanic/Latino	8093	0.01	1.24%

Supplementary Table 10: Computational needs of compared models

Benchmarking test	RAM (MiB)	Time (seconds)
Model fitting		
Tunning Model 1 XGB	464	196.2
Tunning Model 2 XGB	761	346
Tunning Model 3 XGB	978.9	646.5
Tunning Model local PRS XGB	1265.1	27951
Tunning Model local PRS LASSO	1734.9	1112.7
Model 1 LR fitting	256.7	< 0.1
Model 2 LR fitting	287.3	< 0.1
Model 3 LR fitting	347.2	0.1
Model local PRS LR fitting	1489.4	2.5
Applying the models for prediction		
Model 1 XGB prediction	281.6	0.1
Model 2 XGB prediction	279	0.3
Model 3 XGB prediction	388.6	0.1
Model 1 LR prediction	242.8	< 0.1
Model 2 LR prediction	279.9	< 0.1
Model 3 LR prediction	337.5	< 0.1
Model local PRS XGB prediction	1058.3	1.2
Model local PRS LR prediction	1231.2	0.2
Model local PRS LASSO prediction	1454.6	0.1

Tunning model 1 XGB refers to the task of fitting Level 1 model using XGBoost software, over the TOPMed training dataset (includes 5-fold cross-validation). Similarly for models 2 and 3. Model 1 XGB refers to the application of the fitted model to obtain prediction over the TOPMed test dataset. These comparisons were performed for SBP prediction models only.

XGB: XGBoost. LR: Linear regression. MiB: mebibyte.

Supplementary Table 11: Performance of the genetic and ensemble models using local PRSs in TOPMed data.

Phenotype	Model	Group	Training set (N)	Testing set (N)	Training PVE Genetic model	Testing PVE Genetic model	Training PVE Ensemble model	Testing PVE Ensemble model
SBP	Lasso	Overall	43556	18748	6.85%	4.52%	36.31%	33.21%
SBP	Lasso	Black	10914	3674	3.76%	1.55%	33.20%	34.39%
SBP	Lasso	Asian	4281	374	5.54%	1.40%	29.48%	20.82%
SBP	Lasso	White	19853	10823	9.09%	5.78%	36.13%	33.03%
SBP	Lasso	Hispanic/Latino	8027	3877	7.04%	4.24%	36.77%	30.72%
SBP	XGBoost	Overall	43556	18748	59.27%	1.61%	72.15%	31.18%
SBP	XGBoost	Black	10914	3674	58.88%	0.68%	71.46%	33.80%
SBP	XGBoost	Asian	4281	374	58.83%	0.28%	69.26%	19.92%
SBP	XGBoost	White	19853	10823	59.45%	2.12%	71.51%	30.43%
SBP	XGBoost	Hispanic/Latino	8027	3877	59.68%	1.62%	72.57%	28.83%
SBP	Linear Regression	Overall	43556	18748	7.81%	4.77%	36.96%	33.39%
SBP	Linear Regression	Black	10914	3674	3.74%	0.89%	33.19%	33.95%
SBP	Linear Regression	Asian	4281	374	6.52%	0.64%	30.21%	20.21%
SBP	Linear Regression	White	19853	10823	10.72%	6.76%	37.27%	33.73%
SBP	Linear Regression	Hispanic/Latino	8027	3877	7.87%	3.41%	37.33%	30.12%
DBP	Lasso	Overall	43472	18768	2.69%	2.04%	18.19%	19.04%
DBP	Lasso	Black	10889	3657	1.37%	0.69%	14.14%	22.89%
DBP	Lasso	Asian	4220	403	1.86%	1.53%	18.84%	13.24%
DBP	Lasso	White	19786	10877	3.05%	1.84%	15.82%	14.97%
DBP	Lasso	Hispanic/Latino	8093	3831	4.06%	3.95%	16.77%	21.00%
DBP	XGBoost	Overall	43472	18768	83.77%	1.93%	86.35%	18.94%
DBP	XGBoost	Black	10889	3657	83.98%	0.69%	86.06%	22.89%
DBP	XGBoost	Asian	4220	403	83.83%	2.00%	86.63%	13.66%
DBP	XGBoost	White	19786	10877	83.34%	1.74%	85.54%	14.89%
DBP	XGBoost	Hispanic/Latino	8093	3831	84.31%	3.79%	86.39%	20.86%
DBP	Linear Regression	Overall	43472	18768	4.75%	2.61%	19.92%	19.51%
DBP	Linear Regression	Black	10889	3657	2.20%	0.10%	14.87%	22.42%
DBP	Linear Regression	Asian	4220	403	3.98%	2.31%	20.59%	13.93%
DBP	Linear Regression	White	19786	10877	5.76%	2.39%	18.17%	15.45%
DBP	Linear Regression	Hispanic/Latino	8093	3831	6.38%	5.66%	18.79%	22.40%

Performance results (attained PVEs) from the genetic models (PVEs of predicting residuals from the baseline model) and ensemble models (PVEs of predicting the raw trait) estimated in cross validation on the training dataset, and from the independent test dataset using local PRS in TOPMed dataset.

Supplementary Table 12: Participant characteristics of the MGB Biobank dataset.

Characteristic	White	Black	Asian	Other/Unknown
N	7,985	412	200	897
Gender¹				
Female	4,975 (62%)	278 (67%)	133 (67%)	628 (70%)
Male	3,010 (38%)	134 (33%)	67 (34%)	269 (30%)
Age²	61 (48, 71)	49 (38, 60)	49 (41, 59)	47 (37, 60)
SBP²	124 (116, 134)	124 (115, 132)	119 (112, 129)	122 (114, 132)
DBP²	74 (68, 80)	73 (68, 80)	73 (68, 79)	74 (69, 80)
BMI²	26.6 (23.5, 30.5)	29.9 (25.8, 34.9)	24.2 (21.5, 26.7)	28.8 (25.4, 33.3)

¹n (%)

²Median (IQR)

MGB Biobank dataset characteristics broken into self-reported race/ethnicity background of the participants.

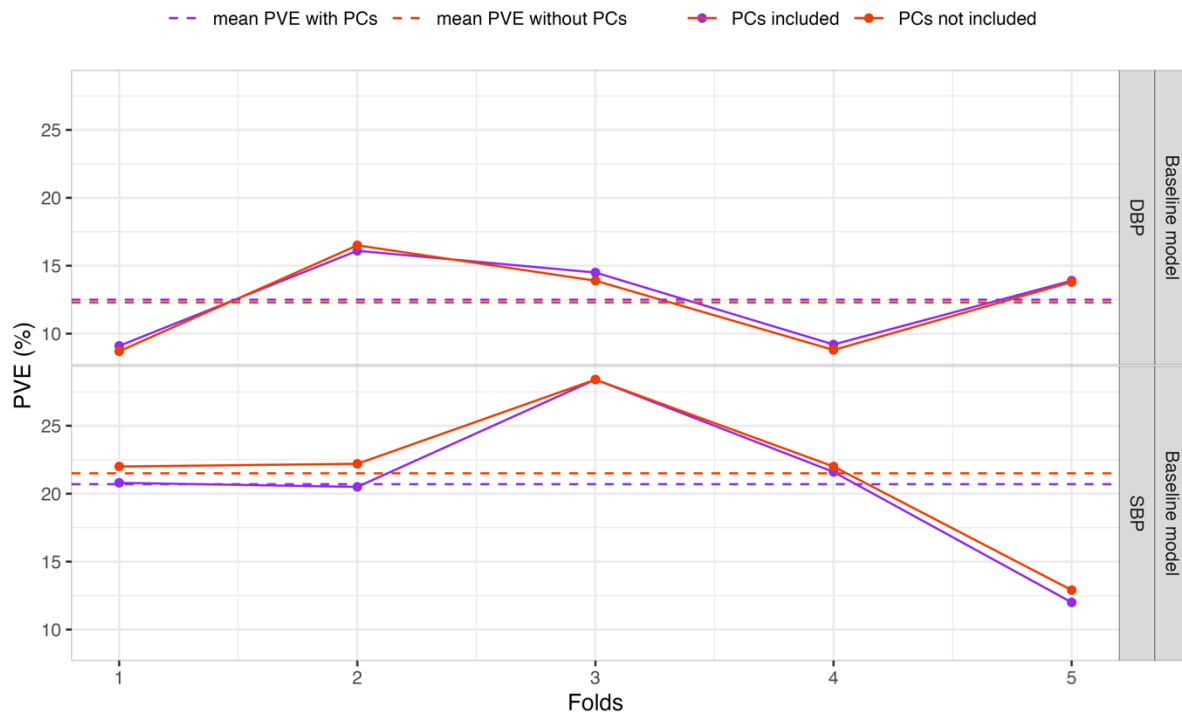
Supplementary Table 13: GWAS summary statistics used for SBP and DBP PRS development.

GWAS name	Reference	Trait	Sample size	Population
BBJ	PMID:29403010(7)	SBP	136,597	Japanese
		DBP	136,615	
UKBB+ICBP	PMID:30224653 (8)	SBP	757,601	European
		DBP	757,601	
MVP	PMID:30578418 (10)	SBP	318,492	Multi-ethnic (69.1% non-Hispanic White, 18.8% non-Hispanic Black, 6.7% Hispanic, 0.77% non-Hispanic Asian and 0.85% non-Hispanic Native American individuals)
		DBP	318,891	

The table provides GWAS source, study population as reported by the manuscript or repository reporting the GWAS, and number of participants used to generate summary statistics. Because some of UKBB+ICBP individuals overlapped with our TOPMed dataset, we performed GWAS using the overlapping TOPMed individuals and applied a numerical procedure (described in Supplementary Note 2 to remove their contribution from the UKBB+ICBP summary statistics.

Supplementary Figures

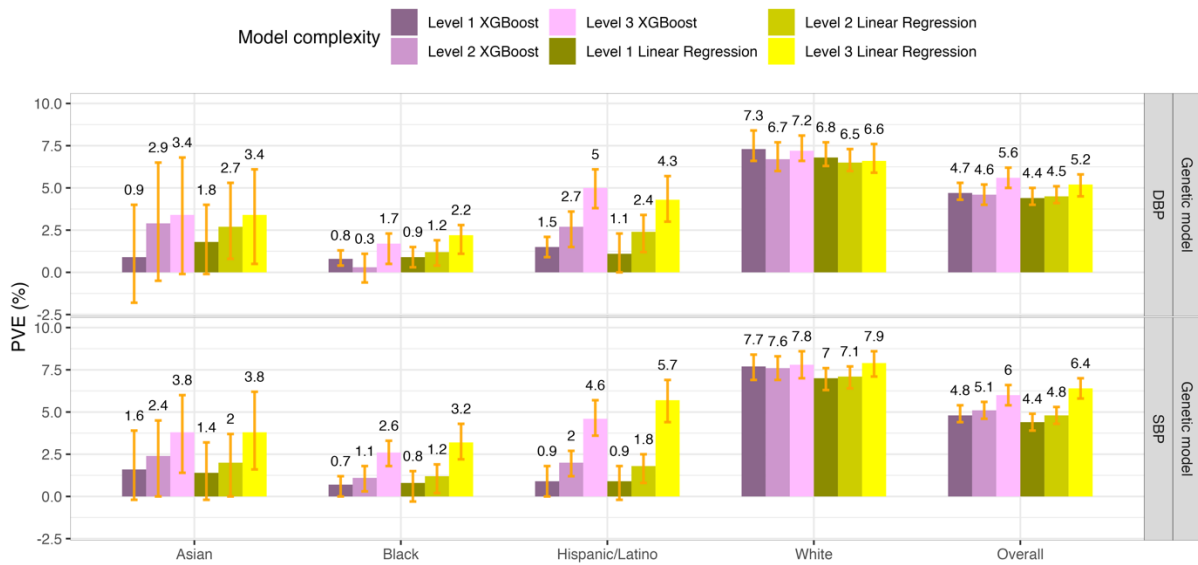
Supplementary Figure 1: Comparison of baseline cross-validated PVE with and without inclusion of genetic PCs.



TOPMed baseline model performance for SBP and DBP prediction, comparing inclusion of genetic PCs to the model performance trained without genetic PCs.

PVE: Percent variance explained. TOPMed: Trans-Omics in Precision Medicine project. SBP: systolic blood pressure. DBP: diastolic blood pressure.

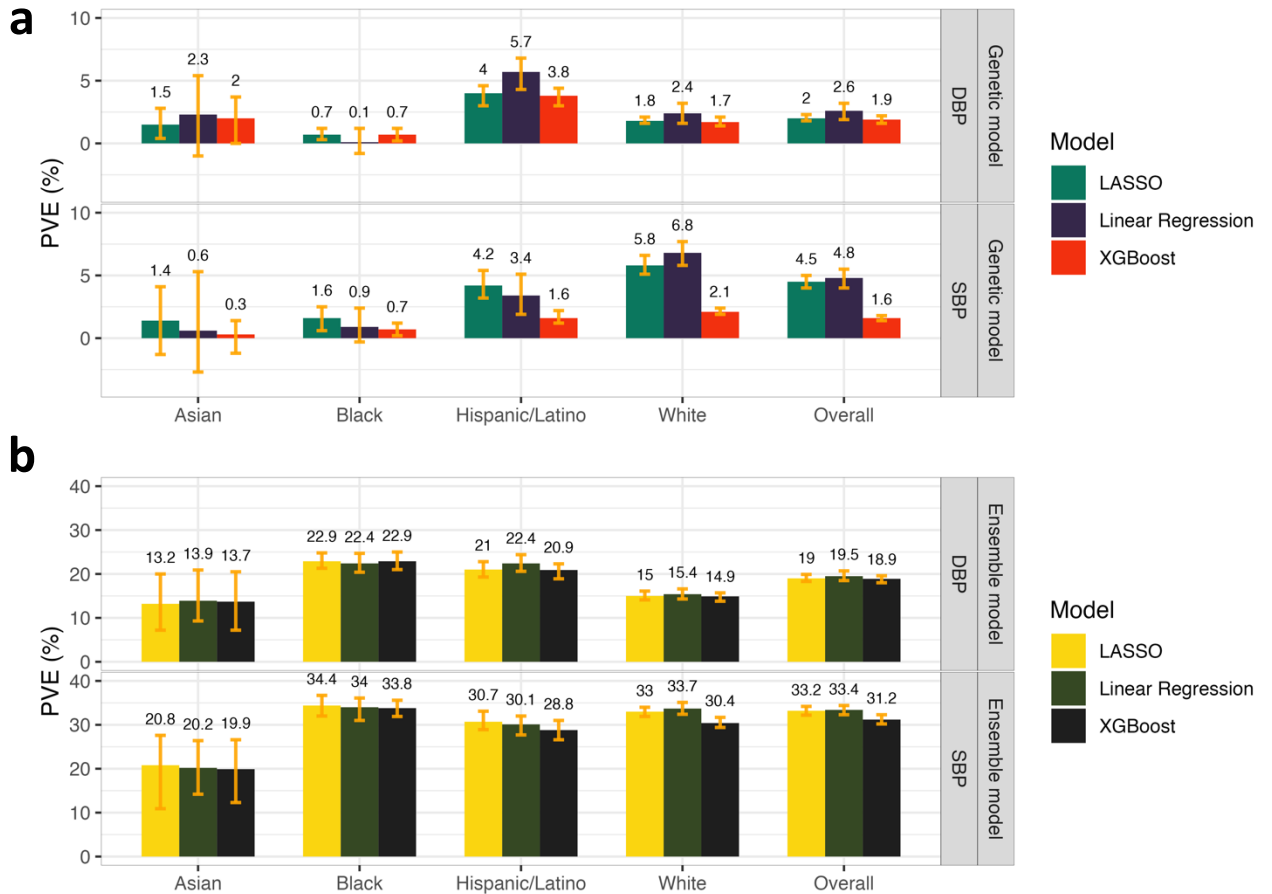
Supplementary Figure 2: Estimated PVE of genetic models fitted using XGBoost and linear models using global PRS.



Estimated PVEs in the TOPMed test dataset for genetic model performance for prediction of SBP and DBP in the overall test dataset and stratified by self-reported race/ethnicity (White N = 10,877, Hispanic/Latino N = 3,831, Black N = 3,657, Asian N = 403 for DBP; White N = 10,823, Hispanic/Latino N = 3,877, Black N = 3,674, Asian N = 374 for SBP).

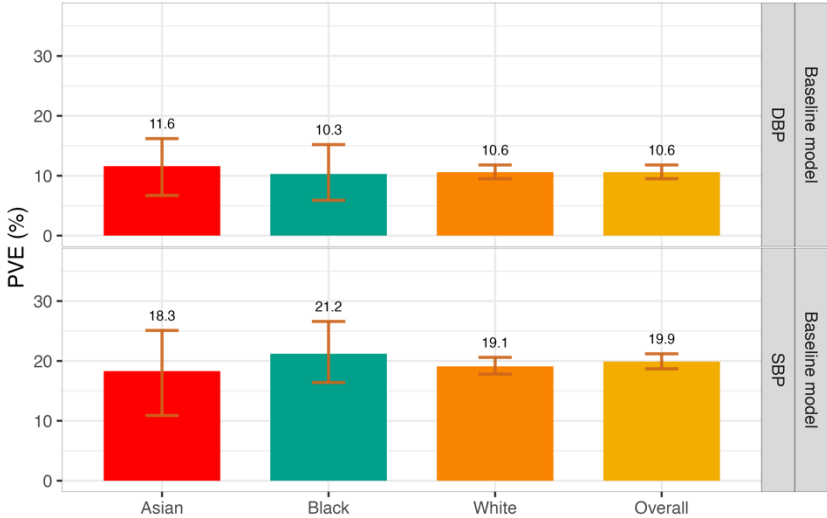
PVE: Percent variance explained. TOPMed: Trans-Omics in Precision Medicine project. SBP: systolic blood pressure. DBP: diastolic blood pressure.

Supplementary Figure 3: Performance of genetic and ensemble models fitted using XGBoost, linear regression and LASSO using local PRSs in TOPMed test dataset.



Panel a: Estimated PVEs in the TOPMed test dataset by genetic models incorporating local PRSs based on the UKB+ICBP GWAS. PVE is reported for predicting residuals from the baseline model, where the baseline model was XGBoost and only used non-genetic covariates. Panel b: Estimated PVE in the TOPMed test dataset for ensemble model at the raw phenotypic level. PVEs are reported for models of SBP and DBP, in the overall test dataset and stratified by self-reported race/ethnicity (White N = 10,877, Hispanic/Latino N = 3,831, Black N = 3,657, Asian N = 403 for DBP; White N = 10,823, Hispanic/Latino N = 3,877, Black N = 3,674, Asian N = 374 for SBP). PVE: Percent variance explained. TOPMed: Trans-Omics in Precision Medicine project. SBP: systolic blood pressure. DBP: diastolic blood pressure. PRS: polygenic risk score.

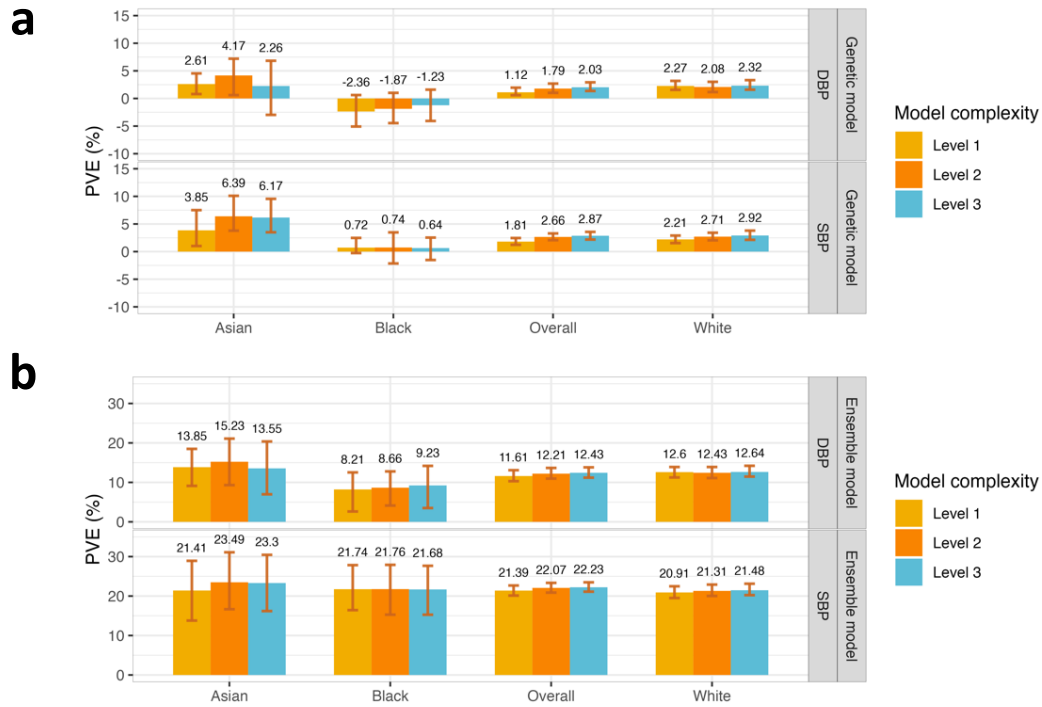
Supplementary Figure 4: Estimated phenotypic PVE of baseline models in the MGB Biobank data fitted using XGBoost.



Estimated PVE in the MGBB test dataset for Baseline models' performance for prediction of SBP and DBP phenotypes in the overall test dataset and stratified by race/ethnicity (White N = 7,985, Black N = 412, Asian N = 200).

PVE: Percent variance explained. TOPMed: Trans-Omics in Precision Medicine project. SBP: systolic blood pressure. DBP: diastolic blood pressure. MGBB: Mass General Brigham Biobank.

Supplementary Figure 5: Estimated PVE of genetic and ensemble models in the MGB Biobank.



Panel a: Estimated PVE in the MGBB test dataset (including only individuals who do not take antihypertensive medication) for XGBoost Genetic models fitted on the TOPMed dataset of three levels of complexity. PVEs are shown for the performance in prediction of the second order of residuals for SBP and DBP phenotypes in the overall test dataset and stratified by race/ethnicity (White N = 7,985, Black N = 412, Asian N = 200). Panel b: Estimated PVE in the MGBB test dataset for Ensemble models' three levels of complexity performance for prediction of the SBP and DBP in the overall test dataset and stratified by race/ethnicity (White N = 7,985, Black N = 412, Asian N = 200). PVE: Percent variance explained. TOPMed: Trans-Omics in Precision Medicine project. SBP: systolic blood pressure. DBP: diastolic blood pressure. MGBB: Mass General Brigham Biobank.

Supplementary Note 1: Descriptions of participating TOPMed studies

Participating TOPMed studies included: Genetics of Cardiometabolic Health in the Amish (Amish; n = 1098), Atherosclerosis Risk in Communities study (ARIC; n = 7166), Mount Sinai BioMe Biobank (BioMe; n = 7234), Coronary Artery Risk Development in Young Adults Study (CARDIA; n = 3015), Cleveland Family Study (CFS; n = 661), Cardiovascular Health Study (CHS; n = 3371), Genetic Epidemiology of COPD (COPDGene; n = 5836), Framingham Heart Study (FHS; n = 3095), Genetic Epidemiology Network of Arteriopathy (GENOA; n = 1041), Genetic Epidemiology Network of Salt Sensitivity (GenSalt, n = 1791), Hispanic Community Health Study/Study of Latinos (HCHS/SOL; n = 7455), Jackson Heart Study (JHS; n = 3281), Multi-Ethnic Study of Atherosclerosis (MESA; n= 4473), Taiwan Study of Hypertension using Rare Variants (THRV; n = 1986) and Women's Health Initiative (WHI; n = 10792). Study specific descriptions are provided below.

Amish

Ethics statement:

All study protocols were approved by the institutional review board at the University of Maryland Baltimore. Informed consent was obtained from each study participant.

Amish acknowledgements:

We gratefully acknowledge our Amish liaisons, research volunteers, field workers, and Amish Research Clinic staff and the extraordinary cooperation and support of the Amish community without which these studies would not have been possible. The Amish studies are supported by

grants and contracts from the NIH, including U01 HL072515, U01 HL84756, U01 HL137181, and P30 DK72488. The TOPMed component of the Amish Research Program was supported by NIH grants R01 HL121007, U01 HL072515, and R01 AG18728.

ARIC

The Atherosclerosis Risk in Communities study (dbGaP accession phs000090) is a population-based prospective cohort study of cardiovascular disease sponsored by the National Heart, Lung, and Blood Institute (NHLBI). ARIC included 15,792 individuals, predominantly European American and African American, aged 45-64 years at baseline (1987-89), chosen by probability sampling from four US communities. Cohort members completed three additional triennial follow-up examinations, a fifth exam in 2011-2013, a sixth exam in 2016-2017, and a seventh exam in 2018-2019, and an eighth exam in 2020. The ARIC study has been described in detail previously [1, 2].

Ethics statement:

The ARIC study has been approved by a single Institutional Review Board (sIRB) at Johns Hopkins School of Medicine and Institutional Review Boards (IRB) at all participating institutions: University of North Carolina at Chapel Hill IRB, Johns Hopkins University School of Public Health IRB, University of Minnesota IRB, Wake Forest University Health Sciences IRB, and University of Mississippi Medical Center IRB. Study participants provided written informed consent at all study visits.

ARIC acknowledgements:

The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health,

Department of Health and Human Services (contract numbers 75N92022D00001, 75N92022D00002, 75N92022D00003, 75N92022D00004, 75N92022D00005). The authors thank the staff and participants of the ARIC study for their important contributions.

WGS for “NHLBI TOPMed: Atherosclerosis Risk in Communities (ARIC)” (phs001211) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201500015C and 3U54HG003273-12S2) and the Broad Institute for MIT and Harvard (3R01HL092577- 06S1). The Genome Sequencing Program (GSP) was funded by the National Human Genome Research Institute (NHGRI), the National Heart, Lung, and Blood Institute (NHLBI), and the National Eye Institute (NEI). The GSP Coordinating Center (U24 HG008956) contributed to cross program scientific initiatives and provided logistical and general study coordination. The Centers for Common Disease Genomics (CCDG) program was supported by NHGRI and NHLBI, and whole genome sequencing was performed at the Baylor College of Medicine Human Genome Sequencing Center (UM1 HG008898).

BioMe

The BioMe Biobank is an ongoing, prospective, hospital- and outpatient- based population research program operated by The Charles Bronfman Institute for Personalized Medicine (IPM) at Mount Sinai. BioMe has enrolled over 50,000 participants between September 2007 and July 2019. BioMe is an Electronic Medical Record (EMR)-linked biobank that integrates research data and clinical care information for consented patients at The Mount Sinai Medical Center, which serves diverse local communities of upper Manhattan with broad health disparities. IPM BioMe populations include 25% of African American ancestry (AA), 36% of Hispanic Latino ancestry (HL), 30% of white European ancestry (EA), and 9% of other ancestry. The BioMe disease burden is reflective of health disparities in the local communities. BioMe operations are fully integrated in clinical care processes, including direct recruitment from clinical sites waiting

areas and phlebotomy stations by dedicated BioMe recruiters independent of clinical care providers, prior to or following a clinician standard of care visit. Recruitment currently occurs at a broad spectrum of over 30 clinical care sites.

Ethics statement:

The BioMe cohort was approved by the Institutional Review Board at the Icahn School of Medicine at Mount Sinai. All BioMe participants provided written, informed consent for genomic data sharing.

BioMe acknowledgements:

The Mount Sinai BioMe Biobank has been supported by The Andrea and Charles Bronfman Philanthropies and in part by Federal funds from the NHLBI and NHGRI (U01HG00638001; U01HG007417; X01HL134588). We thank all participants in the Mount Sinai Biobank. We also thank all our recruiters who have assisted and continue to assist in data collection and management and are grateful for the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

CARDIA

The Coronary Artery Risk Development in Young Adults study (dbGaP accession phs000285) is a prospective multicenter study with 5,115 adults Caucasian and African American participants of the age group 18-30 years at baseline, recruited from four centers at the baseline examination

in 1985-1986 [3]. The recruitment was done from the total community in Birmingham, AL, from selected census tracts in Chicago, IL and Minneapolis, MN; and from the Kaiser Permanente health plan membership in Oakland, CA. Nine examinations have been completed in the years 0, 2, 5, 7, 10, 15, 20, 25 and 30, with high retention rates (91%, 86%, 81%, 79%, 74%, 72%, 72%, and 71%, respectively) and written informed consent was obtained in each visit.

Ethics statement:

All CARDIA participants provided informed consent, and the study was approved by the Institutional Review Boards of the University of Alabama at Birmingham and the University of Texas Health Science Center at Houston.

CARDIA acknowledgements:

The Coronary Artery Risk Development in Young Adults Study (CARDIA) is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with the University of Alabama at Birmingham (HHSN268201800005I & HHSN268201800007I), Northwestern University (HHSN268201800003I), University of Minnesota (HHSN268201800006I), and Kaiser Foundation Research Institute (HHSN268201800004I). CARDIA was also partially supported by the Intramural Research Program of the National Institute on Aging (NIA) and an intra-agency agreement between NIA and NHLBI (AG0005).

CFS

The Cleveland Family Study (CFS) was designed to examine the genetic basis of sleep apnea in 2,534 African-American and European-American individuals from 356 families. Index probands with confirmed sleep apnea were recruited from sleep centers in northern Ohio, supplemented with additional family members and neighborhood control families [4]. Four visits occurred between 1990 and 2006; in the first 3, data were collected in participants' homes while the last

occurred in a clinical research center (2000 - 2006). Measurements included sleep apnea monitoring, blood pressure, anthropometry, spirometry and other related phenotypes. Blood samples (overnight fasting, before bed and following an oral glucose tolerance test), nasal and oral ultrasound, and ECG were also obtained during the 4th exam. Institutional Review Board approval and signed informed consent was obtained for all participants.

Ethics statement:

Cleveland Family Study was approved by the Institutional Review Board (IRB) of Case Western Reserve University and Mass General Brigham (formerly Partners HealthCare). Written informed consent was obtained from all participants.

CHS

The Cardiovascular Health Study (CHS) is a population-based cohort study initiated by the National Heart, Lung and Blood Institute (NHLBI) in 1987 to determine the risk factors for development and progression of cardiovascular disease (CVD) in older adults, with an emphasis on subclinical measures. The study recruited 5,888 adults aged 65 or older at entry in four U.S. communities and conducted extensive annual clinical exams between 1989-1999 along with semi-annual phone calls, events adjudication, and subsequent data analyses and publications. Additional data are collected by studies ancillary to CHS. In June 1990, four Field Centers (Sacramento, CA; Hagerstown, MD; Winston-Salem, NC; Pittsburgh, PA) completed the recruitment of 5201 participants. Between November 1992 and June 1993, an additional 687 adults of primarily African Americans ethnicity were recruited using similar methods.

Blood samples were drawn from all participants at their baseline examination and during follow-up clinic visits and DNA was subsequently extracted from available samples. CHS analyses were limited to participants with available DNA who consented to genetic studies. The baseline examinations consisted of a home interview and a clinic examination that assessed not only traditional risk factors but also measures of subclinical disease, including carotid

ultrasound, echocardiography, electrocardiography, and pulmonary function. Between enrollment and 1998-99, participants were seen in the clinic annually, and contacted by phone at 6-month intervals to collect information about hospitalizations and potential cardiovascular events. Major exam components were repeated during annual follow-up examinations through 1999. Cranial MRI scans, retinal photography, and tests of endothelial function were added as new components. Standard protocols for the identification and adjudication of events were implemented during follow-up. The adjudicated events are CHD, angina, heart failure (HF), stroke, transient ischemic attack (TIA), claudication and mortality. Adjudication of cause of death continues using a streamlined protocol; adjudication of other events ended in June 2015. Deep venous thrombosis and pulmonary embolism events from baseline through 2001 were adjudicated in an ancillary study: the Longitudinal Investigation of Thromboembolism Etiology (LITE). Since 1999, participants have been contacted every 6 months by phone, primarily to ascertain health status and for events follow-up. The study was initially approved by institutional review boards at the Field Centers (Wake Forest, University of California – Davis, Johns Hopkins University, University of Pittsburgh), the Core Laboratory (University of Vermont) and at the Coordinating Center (University of Washington). The University of Washington now handles CHS Data Repository approvals.

Ethics statement:

All CHS participants provided informed consent, and the study was approved by the Institutional Review Board [or ethics review committee] of University Washington.

CHS acknowledgements:

Cardiovascular Health Study: This research was supported by contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, 75N92021D00006, and grants U01HL080295, U01HL130114, and HL105756 from the National Heart, Lung, and Blood Institute

(NHLBI), with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

COPDGene

COPDGene [5] is a cohort study for respiratory disease research, recruiting more than 10,000 subjects between the ages of 45 and 80 who had at least 10 pack-years of smoking during January 2008 - June 2011 at 21 clinical centers. Participants were characterized using spirometry, six-minute walk, inspiratory and expiratory chest CT scans, respiratory symptoms, medical history, medication history and 36-Item short form health survey. In the current analysis, we only used COPDGene control participants (meaning, individuals without COPD).

Ethics statement:

All COPDGene participants provided written informed consent, and the study was approved by the Institutional Review Boards of the participating clinical centers.

COPDGene acknowledgements:

The COPDGene project described was supported by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health. The COPDGene project is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Novartis,

Pfizer, Siemens and Sunovion. A full listing of COPDGene investigators can be found at:
<http://www.copdgene.org/directory>

FHS

The Framingham Heart Study (dbGaP accession phs000007) began in 1948 with the recruitment of an original cohort of 5,209 men and women (mean age 44 years; 55 percent women). In 1971 a second generation of study participants was enrolled; this cohort (mean age 37 years; 52% women) consisted of 5,124 children and spouses of children of the original cohort. A third-generation cohort of 4,095 children of offspring cohort participants (mean age 40 years; 53 percent women) was enrolled in 2002-2005 and are seen every 4 to 8 years. Details of study designs for the three cohorts are summarized elsewhere [6-8]. At each clinic visit, a medical history was obtained, and participants underwent a physical examination. Only study participants consented for genetic and non-genetic data are included. FHS has been approved by the Boston University IRB

Ethics statement:

The Framingham Heart Study was approved by the Institutional Review Board of the Boston University Medical Center. All study participants provided written informed consent.

FHS acknowledgements:

The Framingham Heart Study (FHS) acknowledges the support of contracts NO1-HC-25195, HHSN268201500001I and 75N92019D00031 from the National Heart, Lung and Blood Institute and grant supplement R01 HL092577-06S1 for this research. We also acknowledge the dedication of the FHS study participants without whom this research would not be possible. Dr. Vasani is supported in part by the Evans Medical Foundation and the Jay and Louis Coffman Endowment from the Department of Medicine, Boston University School of Medicine.

GENOA

The Genetic Epidemiology Network of Arteriopathy (GENOA) study (dbGaP accession phs000379), a part of the Family Blood Pressure Program (FBPP Investigators, 2002), consists of hypertensive sibships that were recruited for linkage and association studies in order to identify genes that influence blood pressure and its target organ damage (Daniels, 2004). In the initial phase of the GENOA study (Phase I: 1996-2001), all members of sibships containing ≥ 2 individuals with essential hypertension clinically diagnosed before age 60 were invited to participate, including both hypertensive and normotensive siblings. In the second phase of the GENOA study (Phase II: 2000-2004), 1,239 non-Hispanic white and 1,482 African American participants were successfully re-recruited to measure potential target organ damage due to hypertension.

Ethics statement:

Written informed consent was obtained from all subjects and approval was granted by participating institutional review boards (University of Michigan, University of Mississippi Medical Center, and Mayo Clinic).

GENOA acknowledgements:

Support for the Genetic Epidemiology Network of Arteriopathy (GENOA) was provided by the National Heart, Lung and Blood Institute (U01 HL054457, U01 HL054464, U01 HL054481, R01 HL119443, and R01 HL087660) of the National Institutes of Health. DNA extraction for “NHLBI TOPMed: Genetic Epidemiology Network of Arteriopathy” (phs001345) was performed at the Mayo Clinic Genotyping Core, and WGS was performed at the DNA Sequencing and Gene Analysis Center at the University of Washington (3R01HL055673-18S1) and the Broad Institute (HHSN268201500014C). We would like to thank the GENOA participants.

GenSalt

The GenSalt study (dbGaP accession phs000784) is a unique NHLBI- sponsored family feeding-study designed to examine the interaction between genes and dietary sodium and potassium intake on BP. A detailed description of the GenSalt study design and participants has been reported previously [9]. Briefly, 3,142 participants from 633 Han families from rural, north China were ascertained through a proband with untreated pre-hypertension or stage-1 hypertension identified from a population-based BP screening. A total of 1,906 GenSalt probands and their siblings, spouses, and offspring were eligible. Among them, 1,818 took part in the TOPMed WGS program and had BP and covariable data available for the current analysis. Three morning BP measurements were obtained according to a standard protocol during each of the 3-days of baseline observation. All BP readings were measured by trained and certified observers using a random-zero sphygmomanometer using a standard protocol [10]. BP was measured with the participant in the sitting position after 5 minutes of rest. In addition, participants were advised to avoid alcohol, cigarette smoking, coffee/tea, and exercise for at least 30 minutes prior to their BP measurements. Systolic and diastolic BP measures were taken in triplicate during each day of the three-day baseline observation. After throwing out the first measure, the subsequent two measures obtained on the first day of baseline observation were averaged and used in this analysis.

Ethics statement:

All subjects provided informed consent and the GenSalt study was approved by the Institutional Review Board (IRB) of all participating institutes in the US and China.

GenSalt acknowledgements:

The Genetic Epidemiology Network of Salt-Sensitivity (GenSalt) was supported by research grants (U01HL072507, R01HL087263, and R01HL090682) from the National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD.

HCHS/SOL

The Hispanic Community Health Study/Study of Latinos (dbGaP accession phs000810) is a community-based longitudinal cohort study of 16,415 self-identified Hispanic/Latino persons aged 18–74 years and selected from households in predefined census-block groups across four US field centers (in Chicago, Miami, the Bronx, and San Diego). The census-block groups were chosen to provide diversity among cohort participants with regard to socioeconomic status and national origin or background [11, 12]. The HCHS/SOL cohort includes participants who self-identified as having a Hispanic/Latino background; the largest groups are Central American (n = 1,730), Cuban (n = 2,348), Dominican (n = 1,460), Mexican (n = 6,471), Puerto Rican (n = 2,728), and South American (n = 1,068). The HCHS/SOL baseline clinical examination occurred between 2008 and 2011 and included comprehensive biological, behavioral, and sociodemographic assessments. Visit 2 took place between 2014 and 2017, which re-examined 11,623 participants from the baseline sample. Visit 3 has started in 2020 and will last 4 years, ending on January 2024. In addition to clinic visit, participants are contacted annually to assess clinical outcomes. The study was approved by the Institutional Review Boards at each participating institution and written informed consent was obtained from all participants.

Ethics statement:

This study was approved by the institutional review boards (IRBs) at each field center, where all participants gave written informed consent, and by the Non-Biomedical IRB at the University of North Carolina at Chapel Hill, to the HCHS/SOL Data Coordinating Center. All IRBs approving the study are: Non-Biomedical IRB at the University of North Carolina at Chapel Hill. Chapel Hill, NC; Einstein IRB at the Albert Einstein College of Medicine of Yeshiva University. Bronx, NY; IRB at Office for the Protection of Research Subjects (OPRS), University of Illinois at Chicago. Chicago,

IL; Human Subject Research Office, University of Miami. Miami, FL; Institutional Review Board of San Diego State University. San Diego, CA.

HCHS/SOL acknowledgements:

The Hispanic Community Health Study/Study of Latinos is a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (HHSN268201300001I / N01-HC-65233), University of Miami (HHSN268201300004I / N01-HC- 65234), Albert Einstein College of Medicine (HHSN268201300002I / N01-HC-65235), University of Illinois at Chicago – HHSN268201300003I / N01- HC-65236 Northwestern Univ), and San Diego State University (HHSN268201300005I / N01-HC-65237). The following Institutes/Centers/Offices have contributed to the HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases, National Institute of Neurological Disorders and Stroke, NIH Institution-Office of Dietary Supplements.

JHS

The Jackson Heart Study (dbGaP accession phs000286) is a longitudinal investigation of genetic and environmental risk factors associated with the disproportionate burden of cardiovascular disease in African Americans [13, 14]. At baseline, the JHS recruited 5306 African American residents of the Jackson, Mississippi Metropolitan Statistical Area, which included approximately 6.6% of all African American adults aged 35-84 residing in the area. Participants were recruited via random sampling (17% of participants), volunteers (30%), prior participants in the Atherosclerosis Risk in Communities (ARIC) study (31%), and secondary family members (22%). Among these participants, approximately 3400 gave consent that allows genetic research. JHS participants received three back-to-back clinical examinations (Exam 1, 2000-2004; Exam 2, 2005-2008; and Exam 3, 2009-2013), and a fourth clinical examination started in

2020. Participants are also contacted annually by telephone to update personal and health information including vital status, interim medical events, hospitalizations, functional status and sociocultural information.

Ethics statement:

The Institutional Review Boards at Jackson State University, Tougaloo College, and the University of Mississippi Medical Center approved the study, and all participants provided written informed consent.

JHS acknowledgements:

The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute for Minority Health and Health Disparities (NIMHD). The authors also wish to thank the staffs and participants of the JHS.

Genome sequencing (dbGap accession phs000964) was performed at the Northwest Genomics Center (HHSN268201100037C). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (R01HL117626; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL120393; U01HL120393; contract HHSN268201800001I).

MESA

The Multi-Ethnic Study of Atherosclerosis (dbGaP accession phs000209) is a study of the characteristics of subclinical cardiovascular disease (disease detected non-invasively before it has produced clinical signs and symptoms) and the risk factors that predict progression to clinically overt cardiovascular disease or progression of the subclinical disease [15]. MESA consisted of a diverse, population-based sample of an initial 6,814 asymptomatic men and women aged 45-84. 38 percent of the recruited participants were white, 28 percent African American, 22 percent Hispanic, and 12 percent Asian, predominantly of Chinese descent. Participants were recruited from six field centers across the United States: Wake Forest University, Columbia University, Johns Hopkins University, University of Minnesota, Northwestern University and University of California - Los Angeles. Participants are being followed for identification and characterization of cardiovascular disease events, including acute myocardial infarction and other forms of coronary heart disease (CHD), stroke, and congestive heart failure; for cardiovascular disease interventions; and for mortality. The first examination took place over two years, from July 2000 - July 2002. It was followed by five examination periods that were 17-20 months in length. Participants have been contacted every 9 to 12 months throughout the study to assess clinical morbidity and mortality.

Ethics statements:

All MESA participants provided written informed consent, and the study was approved by the Institutional Review Boards at The Lundquist Institute (formerly Los Angeles BioMedical Research Institute) at Harbor-UCLA Medical Center, University of Washington, Wake Forest School of Medicine, Northwestern University, University of Minnesota, Columbia University, and Johns Hopkins University.

MESA acknowledgements:

MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420. MESA Family is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support is provided by grants and contracts R01HL071051, R01HL071205, R01HL071250, R01HL071251, R01HL071258, R01HL071259, and by the National Center for Research Resources, Grant UL1RR033176. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

THRv

The THRv-TOPMed study comprises 2,353 Taiwan Chinese participants in three cohorts: The SAPPHIRe (Stanford-Asian Pacific Program in Hypertension and Insulin Resistance) Family cohort of approximately 300 hypertensive sibships (N=1,271) and two hospital-based cohorts, the TSGH (Tri-Service General Hospital) cohort (N=160) and the TCVGH (Taichung Veterans General Hospital) cohort (N=922) that provide population-based controls (unrelated hypertensive or non-hypertensive) matched to SAPPHIRe samples. All three cohorts are based in Taiwan. The 1,271 SAPPHIRe subjects were previously recruited as part of the SAPPHIRe Network of the NHLBI-sponsored Family Blood Pressure Program (FBPP). The SAPPHIRe families were recruited to have two or more hypertensive sibs, with some families having one normotensive/hypotensive sib. The two Hospital-based cohorts (TSGH and TCVGH) both recruited unrelated subjects at the SAPPHIRe field centers/hospitals in Taiwan, that matched with the SAPPHIRe subjects for age, sex, and BMI category. Several metabolic variables associated with blood pressure and insulin resistance were measured in the first 5-year SAPPHIRe funding from the NHLBI (1995-2000). Additional phenotyping through return visits

and regular follow ups occurred between 2001 and 2008 which included echocardiographic and multi-detector row CT imaging procedures.

Ethics statements:

All THRV participants provided informed consent, and the study was approved by the Institutional Review Board at The Lundquist Institute (formerly Los Angeles BioMedical Research Institute, or LA BioMed) at Harbor-UCLA Medical Center, and at Washington University in St. Louis.

THRV acknowledgments:

The Rare Variants for Hypertension in Taiwan Chinese (THRV) is supported by the National Heart, Lung, and Blood Institute (NHLBI) grant (R01HL111249) and its participation in TOPMed is supported by an NHLBI supplement (R01HL111249-04S1). THRV is a collaborative study between Washington University in St. Louis, LA BioMed at Harbor UCLA, University of Texas in Houston, Taichung Veterans General Hospital, Taipei Veterans General Hospital, Tri-Service General Hospital, National Health Research Institutes, National Taiwan University, and Baylor University. THRV is based (substantially) on the parent SAPPHIRe study, along with additional population-based and hospital-based cohorts. SAPPHIRe was supported by NHLBI grants (U01HL54527, U01HL54498) and Taiwan funds, and the other cohorts were supported by Taiwan funds.

WHI

The Women's Health Initiative (WHI) cohort. The WHI is a prospective national health study focused on identifying optimal strategies for preventing chronic diseases that are the major causes of death and disability in postmenopausal women. The WHI initially recruited 161,808

women between 1993 and 1997 with the goal of including a socio-demographically diverse population with diversity background groups proportionate to the total minority population of US women aged 50-79 years. The WHI consists of two major parts: a set of randomized Clinical Trials and an Observational Study. The WHI Clinical Trials (CT; N=68,132) includes three overlapping components, each a randomized controlled comparison: the Hormone Therapy Trials (HT), Dietary Modification Trial, and Calcium and Vitamin D Trial. A parallel prospective observational study (OS; N = 93,676) examined biomarkers and risk factors associated with various chronic diseases. While the HT trials ended in the mid-2000s, active follow-up of the WHI-CT and WHI-OS cohorts has continued for over 25 years, with the accumulation of large numbers of diverse clinical outcomes, risk factor measurements, medication use, and many other types of data.

Ethics statement:

All WHI participants provided informed consent and the study was approved by the Institutional Review Board (IRB) of the Fred Hutchinson Cancer Research Center.

WHI acknowledgements:

The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts 75N92021D00001, 75N92021D00002, 75N92021D00003, 75N92021D00004, 75N92021D00005.

Supplementary Note 2: Removal of overlap GWAS

Inverse-variance fixed effects meta-analysis (usually performed by large GWAS meta-analysis efforts) when combining two studies:

Let $\widehat{\beta}_1, \widehat{\beta}_2$ be the effect estimates from study 1 and study 2. Let $\widehat{v}_1, \widehat{v}_2$ be their estimated variances. Let $w_1 = \frac{1}{\widehat{v}_1}, w_2 = \frac{1}{\widehat{v}_2}$, Then:

$$\widehat{\beta} = \frac{w_1 \widehat{\beta}_1 + w_2 \widehat{\beta}_2}{w_1 + w_2}$$

And

$$\widehat{v} = var(\widehat{\beta}) = \frac{w_1^2 \widehat{v}_1 + w_2^2 \widehat{v}_2}{(w_1 + w_2)^2} = \frac{w_1 + w_2}{(w_1 + w_2)^2} = \frac{1}{w_1 + w_2}.$$

The same formula straightforwardly extends for an arbitrary number of studies. Further, it can be easily shown that the meta-analytic estimator over m studies will be the same regardless of the order of the meta-analysis, meaning, that one can first meta-analyze statistics from studies 1 and 2, next from studies 3 and 4, and finally meta-analyze the results of these two meta-analyses, and receive the same results as from a meta-analysis of the four studies at the same time.

Therefore, if we have meta-analysis results from m studies, and we can obtain (or compute) the meta-analysis results from a subset of m_1 studies, we can “back compute” the results from the $m - m_1$ studies. We also assume that the meta-analysis results from m_1 studies are the same as a pooled-summary statistics GWAS results for the same m_1 studies (even though likely the results are not *precisely* the same due to different modelling).

Suppose that we had $\widehat{\beta}, \widehat{v}$ (the meta-analytic estimator of all studies), and $\widehat{\beta}_1, \widehat{v}_1$ estimates from m_1 studies, a subset of these studies, together. We want to compute the estimates $\widehat{\beta}_2, \widehat{v}_2$

because these are independent of our m_1 studies and using them will prevent overfitting when computing polygenic risk models.

Clearly:

$$\widehat{v}_2 = \frac{1}{w_2} = \frac{\widehat{v}}{1 - w_1 \widehat{v}} = \frac{\widehat{v}}{1 - \widehat{v}/\widehat{v}_1}$$

And:

$$\widehat{\beta}_2 = \frac{\widehat{\beta}(w_1 + w_2) - w_1 \widehat{\beta}_1}{w_2}.$$

These can be used to compute p-values of the variants based on the non-overlapping sample of m_2 studies from the original meta-analysis.

Using these summary statistics, we selected SNPs and their weights to calculate Polygenic Risk Scores (PRS) and included them in the Machine Learning (ML) models as features to predict the phenotypes.

Supplementary Note 3: Sensitivity analysis using PRS-CSx-based global PRS

For comparison, we developed global SBP and DBP PRSs, using PRS-CSx [16]. We used the same GWAS summary statistics used for the main analysis, and coupled BBJ with the UKB East Asian LD reference panel, UKB+ICBP (summary statistics after “removing” the estimated effects of overlapping TOPMed White participants) with the UKB European reference panel, and MVP with the UKB African reference panel. Reference panels are implemented in the PRS-CSx

software. We required that SNPs have $MAF \geq 0.01$ in the source GWAS, and used the overlapping SNPs between the GWAS and the reference panel, as well as other default parameters and the “auto” option, as in previous publication [17] (though note that the GWAS summary statistics are different than the ones used in Kurniansyah et al., 2023). The application of PRS-CSx resulted in three ancestry-specific PRSs for each BP trait. For each BP trait, we constructed the three PRSs in the TOPMed dataset using PRSice but without any clumping and thresholding. The three PRSs were then scaled to have mean 0 and variance, and summed. We next trained the genetic model component of the ensemble using conventional linear regression and non-linear ML (implemented using XGBoost), and compared its performance to that of the main analysis model. Supplementary Table 6 reports the complete performance results as attained PVEs from the genetic models (PVEs of predicting residuals from the baseline model) and ensemble models (PVEs of predicting the raw trait) estimated in cross validation on the training dataset, and from the independent test dataset. Overall, performance of prediction models using a single PRS developed by PRS-CSx showed had lower PVE in comparison to models trained using PRSs developed using PRSice2.

Supplementary Note 4: Descriptions of MGB biobank dataset

Samples, genomic data, and health information were obtained from the Mass General Brigham (MGB) Biobank, a biorepository of consented patient samples at Mass General Brigham.

DNA samples

DNA samples are processed from whole blood that was collected as a dedicated research draw or as a clinical discard. Dedicated research samples are aimed to be processed within four hours of collection. Clinical discards are processed 24+ hours after collection. Whole blood is spun to buffy coat with a centrifuge and the buffy coat is stored in a freezer up to several months. The buffy coat is then extracted to DNA. The DNA is then placed in an ultralow freezer (-80oC). Each DNA aliquot contains a minimum of 2 ug of DNA. The concentration varies.

Genotyping

Samples have been genotyped using three versions of the biobank SNP array offered by Illumina that is designed to capture the diversity of genetic backgrounds across the globe. The first batch of data was generated on the Multi-Ethnic Genotyping Array (MEGA) array, the first release of this SNP array. The second, third, and fourth batches were generated on the Expanded Multi-Ethnic Genotyping Array (MEGA Ex) array. All remaining data were generated on the Multi-Ethnic Global (MEG) BeadChip.

Imputation

Prior to performing imputation, files were converted to VCF format, separated by chromosomes. When multiple probes measured the same genotypes, they were checked for concordance and were set to a missing value if the genotypes did not match. Files were

uploaded to the Michigan Imputation Server, and Genotypes were imputed using TOPMed reference panel. Genomic coordinates are provided in GRCh38.

Quality control

We performed quality control using PLINK (v2.0). We filtered SNPs with low-quality imputation ($r < 0.5$), with missing call rates > 0.1 , HWE p-value less than 1×10^{-6} and MAF $< 1\%$. We computed principal component (PC) using PLINK: we pruned the genotype data using a window size of 1000 variants, sliding across the genome with a step size of 250 variants at a time, filtering out any SNPs with LD $R^2 > 0.1$. We used unrelated individuals (3rd degree, identified using PLINK) to compute the loadings for the first 10 PCs.

PRS construction

We constructed PRS using PRSice2, using the same SNPs as those based on clumping performed on the TOPMed dataset, and otherwise the same methodology.

Hypertension status based on Curated Disease Populations

We used the hypertension outcome from the “curated disease populations” provided by the MGB Biobank team. These phenotypes were developed by the Biobank Portal team using both structured and unstructured electronic medical record (EMR) data and clinical, computational and statistical methods. Natural Language Processing (NLP) was used to extract data from narrative text. Chart reviews by disease experts helped identify features and variables

associated with particular phenotypes and were also used to validate results of the algorithms. The process produced robust phenotype algorithms that were evaluated using metrics such as sensitivity, the proportion of true positives correctly identified as such, and positive predictive value (PPV), the proportion of individuals classified as cases by the algorithm [16]. The high throughput phenotyping algorithm is as follows:

1. Create an initial phenotype definition using ICD-9 diagnosis codes.
2. Broaden the definition by determining the most up-to-date features (comorbidities, symptoms, medications) that create a more accurate profile of the phenotype when combined with ICD-9 codes. Features are extracted from online medical literature and knowledge bases via an Automated Feature Extraction Protocol (AFEP).
3. Narrow and refine the definition by determining the features that occur most often in the Biobank data. Extract, code, and rank features contained in clinical narratives with Natural Language Processing (NLP).
4. Create a gold-standard patient set for training the method. Query coded EMR data for the set of patients having at least one ICD-9 code for the phenotype. Apply a statistical sampling algorithm to select a random subset of those patients for full chart review. A clinical expert performs a full chart review to classify the patients as positive or negative for the phenotype.
5. Train a statistical model that incorporates all features in the definition to predict the presence or absence of the phenotype against the gold-standard patient set.
6. Apply the trained model to the entire Biobank Population.

The prevalence of hypertension in the MGB Biobank in general at the time of phenotype development (not restricted to the set of individuals used in our analysis) was 42% and the AUC computed based on the 4 steps above was 0.912.

Ethics statement

All Biobank subjects have provided their consent to join the MGB Biobank, which includes agreeing to provide a blood sample linked to the electronic medical record. Subjects also agree to be recontacted by the Partners Biobank staff as needed.

Acknowledgements

We thank Mass General Brigham Biobank for providing samples, genomic data, and health information data.

Supplementary Note 5: TOPMed and CCDG acknowledgements

Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Genome sequencing for “NHLBI TOPMed: Genetics of Cardiometabolic Health in the Old Order Amish Study” (phs000956) were performed at the Broad Institute of MIT and Harvard (HHSN268201500014C).

Genome sequencing for “NHLBI TOPMed: Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study” (phs000974.v4.p3) was performed at the Broad Institute Genomics Platform (3R01HL092577-06S1, 3U54HG003067-12S2). Genome sequencing for the “NHLBI TOPMed: Genetic Epidemiology Network of Arteriopathy (GENOA)” (phs001345.v2.p1) was performed at the Broad Institute Genomics Platform (HHSN268201500014C) and the Northwest Genomics Center (3R01HL055673-18S1). Genome sequencing for “NHLBI TOPMed: The Jackson Heart Study” (phs000964.v1.p1) was performed at the Northwest Genomics Center (HHSN268201100037C). Genome sequencing for the “NHLBI TOPMed: The Atherosclerosis Risk in Communities Study” (phs001211.v3.p2) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201500015C and 3U54HG003273-12S2) and the Broad Institute for MIT and Harvard (3R01HL092577-06S1). Genome sequencing for “NHLBI TOPMed: Coronary Artery Risk Development in Young Adults Study” (phs001612.v1.p1) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201600033I). Genome sequencing for “NHLBI TOPMed: Cleveland Family Study” (phs000954.v3.p2) was performed at the Northwest Genomics Center (3R01HL098433-05S1, HHSN268201600032I). Genome sequencing for “NHLBI TOPMed: Genetic Epidemiology of COPD (COPDGene) in the TOPMed Program” (phs000951) was performed at the University of Washington Northwest Genomics Center (3R01 HL089856-08S1) and the Broad Institute of MIT and Harvard (HHSN268201500014C). Genomics sequencing for “NHLBI TOPMed: Cardiovascular Health Study” (phs001368.v2.p1) was performed at the Baylor College of Medicine Human Genome Sequencing Center (3U54HG003273-12S2, HHSN268201500015C, HHSN268201600033I). Genome sequencing for “NHLBI TOPMed: Hispanic Community Health

Study/Study of Latinos” (phs001395.v1.p1) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201600033I). Genome sequencing for “NHLBI TOPMed: Women’s Health Initiative (WHI)” (phs001237.v2.p1) was performed at the Broad Institute of MIT and Harvard (HHSN268201500014C). Genome sequencing for “NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis” (phs001416.v2.p1) was performed at Broad Institute Genomics Platform (HHSN268201500014C, 3U54HG003067-13S1). Genome sequencing for “NHLBI TOPMed: Genetic Epidemiology Network of Salt Sensitivity (GenSalt)” (phs001217.v3.p1) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201500015C). Genome sequencing for “NHLBI TOPMed: Rare Variants for Hypertension in Taiwan Chinese (THRv)” (phs001387.v3.p1) was performed at the Baylor College of Medicine Human Genome Sequencing Center (3R01HL111249-04S1, HHSN26820150015C). Genome sequencing for “NHLBI TOPMed: Mount Sinai BioMe Biobank (BioMe)” (phs001644.v3.p2) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201600033I) and at McDonnell Genome Institute (3UM1HG008853-01S2, HHSN268201600037I). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The Genome Sequencing Program (GSP) was funded by the National Human

Genome Research Institute (NHGRI), the National Heart, Lung, and Blood Institute (NHLBI), and the National Eye Institute (NEI). The GSP Coordinating Center (U24 HG008956) contributed to cross-program scientific initiatives and provided logistical and general study coordination. The Centers for Common Disease Genomics (CCDG) program was supported by NHGRI and NHLBI, and whole genome sequencing was performed at the Baylor College of Medicine Human Genome Sequencing Center (UM1 HG008898).

Supplementary Note 6: TOPMed consortium members

Namiko Abe¹, Gonçalo Abecasis², Francois Aguet³, Christine Albert⁴, Laura Almasy⁵, Alvaro Alonso⁶, Seth Ament⁷, Peter Anderson⁸, Pramod Anugu⁹, Deborah Applebaum-Bowden¹⁰, Kristin Ardlie³, Dan Arking¹¹, Donna K Arnett¹², Allison Ashley-Koch¹³, Stella Aslibekyan¹⁴, Tim Assimes¹⁵, Paul Auer¹⁶, Dimitrios Avramopoulos¹¹, Najib Ayas¹⁷, Adithya Balasubramanian¹⁸, John Barnard¹⁹, Kathleen Barnes²⁰, R. Graham Barr²¹, Emily Barron-Casella¹¹, Lucas Barwick²², Terri Beaty¹¹, Gerald Beck²³, Diane Becker²⁴, Lewis Becker¹¹, Rebecca Beer²⁵, Amber Beitelshes⁷, Emelia Benjamin²⁶, Takis Benos²⁷, Marcos Bezerra²⁸, Larry Bielak², Joshua Bis²⁹, Thomas Blackwell², John Blangero³⁰, Eric Boerwinkle³¹, Donald W. Bowden³², Russell Bowler³³, Jennifer Brody⁸, Ulrich Broeckel³⁴, Jai Broome⁸, Deborah Brown³⁵, Karen Bunting¹, Esteban Burchard³⁶, Carlos Bustamante³⁷, Erin Buth³⁸, Brian Cade³⁹, Jonathan Cardwell⁴⁰, Vincent Carey⁴¹, Julie Carrier⁴², April Carson⁴³, Cara Carty⁴⁴, Richard Casaburi⁴⁵, Juan P Casas Romero⁴⁶, James Casella¹¹, Peter Castaldi⁴⁷, Mark Chaffin³, Christy Chang⁷, Yi-Cheng Chang⁴⁸, Daniel Chasman⁴⁹, Sameer Chavan⁴⁰, Bo-Juen Chen¹, Wei-Min Chen⁵⁰, Yii-Der Ida Chen⁵¹, Michael Cho⁴¹, Seung Hoan Choi³, Lee-Ming Chuang⁵², Mina Chung⁵³, Ren-Hua Chung⁵⁴, Clary Clish⁵⁵, Suzy Comhair⁵⁶, Matthew Conomos³⁸, Elaine Cornell⁵⁷, Adolfo Correa⁵⁸, Carolyn Crandall⁴⁵, James Crapo⁵⁹, L. Adrienne Cupples⁶⁰, Joanne Curran⁶¹, Jeffrey Curtis⁶², Brian Custer⁶³, Coleen Damcott⁷, Dawood Darbar⁶⁴, Sean David⁶⁵, Colleen Davis⁸, Michelle Daya⁴⁰, Mariza de

Andrade⁶⁶, Lisa de las Fuentes⁶⁷, Paul de Vries⁶⁸, Michael DeBaun⁶⁹, Ranjan Deka⁷⁰, Dawn DeMeo⁴¹, Scott Devine⁷, Huyen Dinh¹⁸, Harsha Doddapaneni¹⁸, Qing Duan⁷¹, Shannon Dugan-Perez¹⁸, Ravi Duggirala⁷², Jon Peter Durda⁵⁷, Susan K. Dutcher⁷³, Charles Eaton⁷⁴, Lynette Ekunwe⁹, Adel El Boueiz⁷⁵, Patrick Ellinor⁷⁶, Leslie Emery⁸, Serpil Erzurum¹⁹, Charles Farber⁵⁰, Jesse Farek¹⁸, Tasha Fingerlin⁷⁷, Matthew Flickinger², Myriam Fornage³¹, Nora Franceschini⁷⁸, Chris Frazar⁸, Mao Fu⁷, Stephanie M. Fullerton⁸, Lucinda Fulton⁷⁹, Stacey Gabriel³, Weiniu Gan²⁵, Shanshan Gao⁴⁰, Yan Gao⁹, Margery Gass⁸⁰, Heather Geiger⁸¹, Bruce Gelb⁸², Mark Geraci²⁷, Soren Germer¹, Robert Gerszten⁸³, Auyon Ghosh⁴¹, Richard Gibbs¹⁸, Chris Gignoux¹⁵, Mark Gladwin²⁷, David Glahn⁸⁴, Stephanie Gogarten⁸, Da-Wei Gong⁷, Harald Goring⁸⁵, Sharon Graw⁸⁶, Kathryn J. Gray⁸⁷, Daniel Grine⁴⁰, Colin Gross², C. Charles Gu⁷⁹, Yue Guan⁷, Xiuqing Guo⁵¹, Namrata Gupta³, David M. Haas⁸⁸, Jeff Haessler⁸⁰, Michael Hall⁸⁹, Yi Han¹⁸, Patrick Hanly⁹⁰, Daniel Harris⁹¹, Nicola L. Hawley⁹², Jiang He⁹³, Ben Heavner³⁸, Susan Heckbert⁹⁴, Ryan Hernandez³⁶, David Herrington⁹⁵, Craig Hersh⁹⁶, Bertha Hidalgo¹⁴, James Hixson³¹, Brian Hobbs⁴¹, John Hokanson⁴⁰, Elliott Hong⁷, Karin Hoth⁹⁷, Chao (Agnes) Hsiung⁹⁸, Jianhong Hu¹⁸, Yi-Jen Hung⁹⁹, Haley Huston¹⁰⁰, Chii Min Hwu¹⁰¹, Marguerite Ryan Irvin¹⁴, Rebecca Jackson¹⁰², Deepti Jain⁸, Cashell Jaquish¹⁰³, Jill Johnsen¹⁰⁴, Andrew Johnson²⁵, Craig Johnson⁸, Rich Johnston⁶, Kimberly Jones¹¹, Hyun Min Kang¹⁰⁵, Robert Kaplan¹⁰⁶, Sharon Kardia², Shannon Kelly³⁶, Eimear Kenny⁸², Michael Kessler⁷, Alyna Khan⁸, Ziad Khan¹⁸, Wonji Kim¹⁰⁷, John Kimoff¹⁰⁸, Greg Kinney¹⁰⁹, Barbara Konkle¹¹⁰, Charles Kooperberg⁸⁰, Holly Kramer¹¹¹, Christoph Lange¹¹², Ethan Lange⁴⁰, Leslie Lange¹¹³, Cathy Laurie⁸, Cecelia Laurie⁸, Meryl LeBoff⁴¹, Jiwon Lee⁴¹, Sandra Lee¹⁸, Wen-Jane Lee¹⁰¹, Jonathon LeFaive², David Levine⁸, Dan Levy²⁵, Joshua Lewis⁷, Xiaohui Li⁵¹, Yun Li⁷¹, Henry Lin⁵¹, Honghuang Lin¹¹⁴, Xihong Lin¹¹⁵, Simin Liu¹¹⁶, Yongmei Liu¹¹⁷, Yu Liu¹¹⁸, Ruth J.F. Loos¹¹⁹, Steven Lubitz⁷⁶, Kathryn Lunetta¹¹⁴, James Luo²⁵, Ulysses Magalang¹²⁰, Michael Mahaney⁶¹, Barry Make¹¹, Ani Manichaikul⁵⁰, Alisa Manning¹²¹, JoAnn Manson⁴¹, Lisa Martin¹²², Melissa Marton⁸¹, Susan Mathai⁴⁰, Rasika Mathias¹¹, Susanne May³⁸, Patrick McArdle⁷, Merry-Lynn McDonald¹²³, Sean McFarland¹⁰⁷, Stephen McGarvey¹²⁴, Daniel McGoldrick¹²⁵, Caitlin McHugh³⁸, Becky McNeil¹²⁶, Hao Mei⁹, James Meigs¹²⁷, Vipin Menon¹⁸, Luisa Mestroni⁸⁶, Ginger Metcalf¹⁸, Deborah A Meyers¹²⁸, Emmanuel Mignot¹²⁹, Julie Mikulla²⁵, Nancy Min⁹, Mollie Minear¹³⁰, Ryan L Minster²⁷, Braxton D. Mitchell⁷, Matt Moll⁴⁷, Zeineen

Momin¹⁸, May E. Montasser⁷, Courtney Montgomery¹³¹, Donna Muzny¹⁸, Josyf C Mychaleckyj⁵⁰, Girish Nadkarni⁸², Rakhi Naik¹¹, Take Naseri¹³², Pradeep Natarajan³, Sergei Nekhai¹³³, Sarah C. Nelson³⁸, Bonnie Neltner⁴⁰, Caitlin Nessner¹⁸, Deborah Nickerson¹³⁴, Osuji Nkechinyere¹⁸, Kari North⁷¹, Jeff O'Connell¹³⁵, Tim O'Connor⁷, Heather Ochs-Balcom¹³⁶, Geoffrey Okwuonu¹⁸, Allan Pack¹³⁷, David T. Paik¹³⁸, Nicholette Palmer¹³⁹, James Pankow¹⁴⁰, George Papanicolaou²⁵, Cora Parker¹⁴¹, Gina Peloso¹⁴², Juan Manuel Peralta⁷², Marco Perez¹⁵, James Perry⁷, Ulrike Peters¹⁴³, Patricia Peyser², Lawrence S Phillips⁶, Jacob Pleiness², Toni Pollin⁷, Wendy Post¹⁴⁴, Julia Powers Becker¹⁴⁵, Meher Preethi Boorgula⁴⁰, Michael Preuss⁸², Bruce Psaty⁸, Pankaj Qasba²⁵, Dandi Qiao⁴¹, Zhaohui Qin⁶, Nicholas Rafaels¹⁴⁶, Laura Raffield¹⁴⁷, Mahitha Rajendran¹⁸, Vasam S. Ramachandran¹¹⁴, D.C. Rao⁷⁹, Laura Rasmussen-Torvik¹⁴⁸, Aakrosh Ratan⁵⁰, Susan Redline⁴⁷, Robert Reed⁷, Catherine Reeves¹⁴⁹, Elizabeth Regan⁵⁹, Alex Reiner¹⁵⁰, Muagututiã€a Sefuiva Reupena¹⁵¹, Ken Rice⁸, Stephen Rich⁵⁰, Rebecca Robillard¹⁵², Nicolas Robine⁸¹, Dan Roden¹⁵³, Carolina Roselli³, Jerome Rotter¹⁵⁴, Ingo Ruczinski¹¹, Alexi Runnels⁸¹, Pamela Russell⁴⁰, Sarah Ruuska¹⁰⁰, Kathleen Ryan⁷, Ester Cerdeira Sabino¹⁵⁵, Danish Saleheen²¹, Shabnam Salimi¹⁵⁶, Sejal Salvi¹⁸, Steven Salzberg¹¹, Kevin Sandow¹⁵⁷, Vijay G. Sankaran¹⁵⁸, Jireh Santibanez¹⁸, Karen Schwander⁷⁹, David Schwartz⁴⁰, Frank Sciurba²⁷, Christine Seidman¹⁵⁹, Jonathan Seidman¹⁶⁰, Frédéric Sériès¹⁶¹, Vivien Sheehan¹⁶², Stephanie L. Sherman¹⁶³, Amol Shetty⁷, Aniket Shetty⁴⁰, Wayne Hui- Heng Sheu¹⁰¹, M. Benjamin Shoemaker¹⁶⁴, Brian Silver¹⁶⁵, Edwin Silverman⁴¹, Robert Skomro¹⁶⁶, Albert Vernon Smith¹⁶⁷, Jennifer Smith², Josh Smith⁸, Nicholas Smith⁹⁴, Tanja Smith¹, Sylvia Smoller¹⁰⁶, Beverly Snively¹⁶⁸, Michael Snyder¹⁵, Tamar Sofer⁴¹, Nona Sotoodehnia⁸, Adrienne M. Stilp⁸, Garrett Storm¹⁶⁹, Elizabeth Streeten⁷, Jessica Lasky Su¹⁷⁰, Yun Ju Sung⁷⁹, Jody Sylvia⁴¹, Adam Szpiro⁸, Daniel Taliun², Hua Tang¹⁷¹, Margaret Taub¹¹, Kent D. Taylor¹⁷², Matthew Taylor⁸⁶, Simeon Taylor⁷, Marilyn Telen¹³, Timothy A. Thornton⁸, Machiko Threlkeld¹⁷³, Lesley Tinker¹⁷⁴, David Tirschwell⁸, Sarah Tishkoff¹⁷⁵, Hemant Tiwari¹⁷⁶, Catherine Tong¹⁷⁷, Russell Tracy¹⁷⁸, Michael Tsai¹⁴⁰, Dhananjay Vaidya¹¹, David Van Den Berg¹⁷⁹, Peter VandeHaar², Scott Vrieze¹⁴⁰, Tarik Walker⁴⁰, Robert Wallace⁹⁷, Avram Walts⁴⁰, Fei Fei Wang⁸, Heming Wang¹⁸⁰, Jiongming Wang¹⁶⁷, Karol Watson⁴⁵, Jennifer Watt¹⁸, Daniel E. Weeks²⁷, Joshua Weinstock¹⁰⁵, Bruce Weir⁸, Scott T Weiss¹⁸¹, Lu-Chen Weng⁷⁶, Jennifer Wessel¹⁸², Cristen Willer⁶², Kayleen Williams³⁸, L. Keoki Williams¹⁸³, Carla Wilson⁴¹, James Wilson¹⁸⁴, Lara

Winterkorn⁸¹, Quenna Wong⁸, Joseph Wu¹³⁸, Huichun Xu⁷, Lisa Yanek¹¹, Ivana Yang⁴⁰, Ketian Yu², Seyedeh Maryam Zekavat³, Yingze Zhang¹⁸⁵, Snow Xueyan Zhao⁵⁹, Wei Zhao¹⁸⁶, Xiaofeng Zhu¹⁸⁷, Michael Zody¹, Sebastian Zoellner²

1 - New York Genome Center, New York, New York; 2 - University of Michigan, Ann Arbor, Michigan; 3 - Broad Institute, Cambridge, Massachusetts; 4 - Cedars Sinai, Boston, Massachusetts; 5 - Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, Pennsylvania; 6 - Emory University, Atlanta, Georgia; 7 - University of Maryland, Baltimore, Maryland; 8 - University of Washington, Seattle, Washington; 9 - University of Mississippi, Jackson, Mississippi; 10 - National Institutes of Health, Bethesda, Maryland; 11 - Johns Hopkins University, Baltimore, Maryland; 12 - University of Kentucky, Lexington, Kentucky; 13 - Duke University, Durham, North Carolina; 14 - University of Alabama, Birmingham, Alabama; 15 - Stanford University, Stanford, California; 16 - Medical College of Wisconsin, Milwaukee, Wisconsin; 17 - Medicine, Providence Health Care, Vancouver; 18 - Baylor College of Medicine Human Genome Sequencing Center, Houston, Texas; 19 - Cleveland Clinic, Cleveland, Ohio; 20 - Tempus, University of Colorado Anschutz Medical Campus, Aurora, Colorado; 21 - Columbia University, New York, New York; 22 - LTRC, The Emmes Corporation, Rockville, Maryland; 23 - Quantitative Health Sciences, Cleveland Clinic, Cleveland, Ohio; 24 - Medicine, Johns Hopkins University, Baltimore, Maryland; 25 - National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland; 26 - Boston University School of Medicine, Boston University, Massachusetts General Hospital, Boston, Massachusetts; 27 - University of Pittsburgh, Pittsburgh, Pennsylvania; 28 - Fundação de Hematologia e Hemoterapia de Pernambuco - Hemope, Recife; 29 - Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, Washington; 30 - Human Genetics, University of Texas Rio Grande Valley School of Medicine, Brownsville, Texas; 31 - University of Texas Health at Houston, Houston, Texas; 32 - Department of Biochemistry, Wake Forest Baptist Health, Winston-Salem, North Carolina; 33 - National Jewish Health, National Jewish Health, Denver, Colorado; 34 - Pediatrics, Medical College of Wisconsin, Milwaukee, Wisconsin; 35 - Pediatrics, University of Texas Health at Houston, Houston, Texas; 36 - University of California, San

Francisco, San Francisco, California; 37 – Biomedical Data Science, Stanford University, Stanford, California; 38 - Biostatistics, University of Washington, Seattle, Washington; 39 - Brigham and Women's Hospital, Brigham & Women's Hospital, Boston, Massachusetts; 40 - University of Colorado at Denver, Denver, Colorado; 41 - Brigham & Women's Hospital, Boston, Massachusetts; 42 - University of Montreal; 43 - Medicine, University of Mississippi, Jackson, Mississippi; 44 - Washington State University, Pullman, Washington; 45 - University of California, Los Angeles, Los Angeles, California; 46 - Brigham & Women's Hospital; 47 - Medicine, Brigham & Women's Hospital, Boston, Massachusetts; 48 - National Taiwan University, Taipei; 49 - Division of Preventive Medicine, Brigham & Women's Hospital, Boston, Massachusetts; 50 - University of Virginia, Charlottesville, Virginia; 51 - Lundquist Institute, Torrance, California; 52 - National Taiwan University Hospital, National Taiwan University, Taipei; 53 - Cleveland Clinic, Cleveland Clinic, Cleveland, Ohio; 54 - National Health Research Institute Taiwan, Miaoli County; 55 - Metabolomics Platform, Broad Institute, Cambridge, Massachusetts; 56 - Immunity and Immunology, Cleveland Clinic, Cleveland, Ohio; 57 - University of Vermont, Burlington, Vermont; 58 - Population Health Science, University of Mississippi, Jackson, Mississippi; 59 - National Jewish Health, Denver, Colorado; 60 - Biostatistics, Boston University, Boston, Massachusetts; 61 - University of Texas Rio Grande Valley School of Medicine, Brownsville, Texas; 62 - Internal Medicine, University of Michigan, Ann Arbor, Michigan; 63 - Vitalant Research Institute, San Francisco, California; 64 - University of Illinois at Chicago, Chicago, Illinois; 65 - University of Chicago, Chicago, Illinois; 66 - Health Quantitative Sciences Research, Mayo Clinic, Rochester, Minnesota; 67 - Department of Medicine, Cardiovascular Division, Washington University in St Louis, St. Louis, Missouri; 68 - Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, University of Texas Health at Houston, Houston, Texas; 69 - Vanderbilt University, Nashville, Tennessee; 70 - University of Cincinnati, Cincinnati, Ohio; 71 - University of North Carolina, Chapel Hill, North Carolina; 72 - University of Texas Rio Grande Valley School of Medicine, Edinburg, Texas; 73 - Genetics, Washington University in St Louis, St Louis, Missouri; 74 - Brown University, Providence, Rhode Island; 75 - Channing Division of Network Medicine, Harvard University, Cambridge, Massachusetts; 76 - Massachusetts General Hospital, Boston,

Massachusetts; 77 - Center for Genes, Environment and Health, National Jewish Health, Denver, Colorado; 78 - Epidemiology, University of North Carolina, Chapel Hill, North Carolina; 79 - Washington University in St Louis, St Louis, Missouri; 80 - Fred Hutchinson Cancer Research Center, Seattle, Washington; 81 - New York Genome Center, New York City, New York; 82 - Icahn School of Medicine at Mount Sinai, New York, New York; 83 - Beth Israel Deaconess Medical Center, Boston, Massachusetts; 84 - Department of Psychiatry, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts; 85 - University of Texas Rio Grande Valley School of Medicine, San Antonio, Texas; 86 - University of Colorado Anschutz Medical Campus, Aurora, Colorado; 87 - Obstetrics and Gynecology, Mass General Brigham, Boston, Massachusetts; 88 - OB/GYN, Indiana University, Indianapolis, Indiana; 89 - Cardiology, University of Mississippi, Jackson, Mississippi; 90 - Medicine, University of Calgary, Calgary; 91 - Genetics, University of Maryland, Philadelphia, Pennsylvania; 92 - Department of Chronic Disease Epidemiology, Yale University, New Haven, Connecticut; 93 - Tulane University, New Orleans, Louisiana; 94 - Epidemiology, University of Washington, Seattle, Washington; 95 - Wake Forest Baptist Health, Winston-Salem, North Carolina; 96 - Channing Division of Network Medicine, Brigham & Women's Hospital, Boston, Massachusetts; 97 - University of Iowa, Iowa City, Iowa; 98 - Institute of Population Health Sciences, NHRI, National Health Research Institute Taiwan, Miaoli County; 99 - Tri-Service General Hospital National Defense Medical Center; 100 - Blood Works Northwest, Seattle, Washington; 101 - Taichung Veterans General Hospital Taiwan, Taichung City; 102 - Internal Medicine, Division of Endocrinology, Diabetes and Metabolism, Oklahoma State University Medical Center, Columbus, Ohio; 103 - NHLBI, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland; 104 - Research Institute, Blood Works Northwest, Seattle, Washington; 105 - Biostatistics, University of Michigan, Ann Arbor, Michigan; 106 - Albert Einstein College of Medicine, New York, New York; 107 - Harvard University, Cambridge, Massachusetts; 108 - McGill University, Montré@al; 109 - Epidemiology, University of Colorado at Denver, Aurora, Colorado; 110 - Medicine, Blood Works Northwest, Seattle, Washington; 111 - Public Health Sciences, Loyola University, Maywood, Illinois; 112 - Biostats, Harvard School of Public Health, Boston, Massachusetts; 113 - Medicine, University of Colorado at Denver, Aurora, Colorado; 114 - Boston University, Boston,

Massachusetts; 115 - Harvard School of Public Health, Boston, Massachusetts; 116 - Epidemiology and Medicine, Brown University, Providence, Rhode Island; 117 - Cardiology, Duke University, Durham, North Carolina; 118 - Cardiovascular Institute, Stanford University, Stanford, California; 119 - The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York; 120 - Division of Pulmonary, Critical Care and Sleep Medicine, Ohio State University, Columbus, Ohio; 121 - Broad Institute, Harvard University, Massachusetts General Hospital; 122 - cardiology, George Washington University, Washington, District of Columbia; 123 - University of Alabama at Birmingham, University of Alabama, Birmingham, Alabama; 124 - Epidemiology, Brown University, Providence, Rhode Island; 125 - Genome Sciences, University of Washington, Seattle, Washington; 126 - RTI International; 127 - Medicine, Massachusetts General Hospital, Boston, Massachusetts; 128 - University of Arizona, Tucson, Arizona; 129 - Center For Sleep Sciences and Medicine, Stanford University, Palo Alto, California; 130 - National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland; 131 - Genes and Human Disease, Oklahoma Medical Research Foundation, Oklahoma City, Oklahoma; 132 - Ministry of Health, Government of Samoa, Apia; 133 - Howard University, Washington, District of Columbia; 134 - Department of Genome Sciences, University of Washington, Seattle, Washington; 135 - University of Maryland, Baltimore, Maryland; 136 - University at Buffalo, Buffalo, New York; 137 - Division of Sleep Medicine/Department of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania; 138 - Stanford Cardiovascular Institute, Stanford University, Stanford, California; 139 - Biochemistry, Wake Forest Baptist Health, Winston-Salem, North Carolina; 140 - University of Minnesota, Minneapolis, Minnesota; 141 - Biostatistics and Epidemiology Division, RTI International, Research Triangle Park, North Carolina; 142 - Department of Biostatistics, Boston University, Boston, Massachusetts; 143 - Fred Hutch and UW, Fred Hutchinson Cancer Research Center, Seattle, Washington; 144 - Cardiology/Medicine, Johns Hopkins University, Baltimore, Maryland; 145 - Medicine, University of Colorado at Denver, Denver, Colorado; 146 - CCPM, University of Colorado at Denver, Denver, Colorado; 147 - Genetics, University of North Carolina, Chapel Hill, North Carolina; 148 - Northwestern University, Chicago, Illinois; 149 - New York Genome Center, New

York Genome Center, New York City, New York; 150 - Fred Hutchinson Cancer Research Center, University of Washington, Seattle, Washington; 151 - Lutia I Puava Ae Mapu I Fagalele, Apia; 152 - Sleep Research Unit, University of Ottawa Institute for Mental Health Research, University of Ottawa, Ottawa; 153 - Medicine, Pharmacology, Biomedical Informatics, Vanderbilt University, Nashville, Tennessee; 154 - Pediatrics, Lundquist Institute, Torrance, California; 155 - Faculdade de Medicina, Universidade de Sao Paulo, Sao Paulo; 156 - Pathology, University of Maryland, Seattle, Washington; 157 - TGPS, Lundquist Institute, Torrance, California; 158 - Division of Hematology/Oncology, Harvard University, Boston, Massachusetts; 159 - Genetics, Harvard Medical School, Boston, Massachusetts; 160 - Harvard Medical School, Boston, Massachusetts; 161 - Université Laval, Quebec City; 162 - Pediatrics, Emory University, Atlanta, Georgia; 163 - Human Genetics, Emory University, Atlanta, Georgia; 164 - Medicine/Cardiology, Vanderbilt University, Nashville, Tennessee; 165 - UMass Memorial Medical Center, Worcester, Massachusetts; 166 - University of Saskatchewan, Saskatoon; 167 - University of Michigan; 168 - Biostatistical Sciences, Wake Forest Baptist Health, Winston-Salem, North Carolina; 169 - Genomic Cardiology, University of Colorado at Denver, Aurora, Colorado; 170 - Channing Department of Medicine, Brigham & Women's Hospital, Boston, Massachusetts; 171 - Genetics, Stanford University, Stanford, California; 172 - Institute for Translational Genomics and Populations Sciences, Lundquist Institute, Torrance, California; 173 - University of Washington, Department of Genome Sciences, University of Washington, Seattle, Washington; 174 - Cancer Prevention Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington; 175 - Genetics, University of Pennsylvania, Philadelphia, Pennsylvania; 176 - Biostatistics, University of Alabama, Birmingham, Alabama; 177 - Department of Biostatistics, University of Washington, Seattle, Washington; 178 - Pathology & Laboratory Medicine, University of Vermont, Burlington, Vermont; 179 - USC Methylation Characterization Center, University of Southern California, University of Southern California, California; 180 - Brigham & Women's Hospital, Mass General Brigham, Boston, Massachusetts; 181 - Channing Division of Network Medicine, Department of Medicine, Brigham & Women's Hospital, Boston, Massachusetts; 182 - Epidemiology, Indiana University, Indianapolis, Indiana; 183 - Henry Ford Health System, Detroit, Michigan; 184 -

Cardiology, Beth Israel Deaconess Medical Center, Cambridge, Massachusetts; 185 - Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania; 186 - Department of Epidemiology, University of Michigan, Ann Arbor, Michigan; 187 - Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio

Supplementary References

1. *The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators.* Am J Epidemiol, 1989. **129**(4): p. 687-702.
2. Wright, J.D., et al., *The ARIC (Atherosclerosis Risk In Communities) Study: JACC Focus Seminar 3/8.* J Am Coll Cardiol, 2021. **77**(23): p. 2939-2959.
3. Friedman, G.D., et al., *CARDIA: study design, recruitment, and some characteristics of the examined subjects.* J Clin Epidemiol, 1988. **41**(11): p. 1105-16.
4. Redline, S., et al., *The familial aggregation of obstructive sleep apnea.* Am J Respir Crit Care Med, 1995. **151**(3 Pt 1): p. 682-7.
5. Regan, E.A., et al., *Genetic epidemiology of COPD (COPDGene) study design.* COPD, 2010. **7**(1): p. 32-43.
6. Dawber, T.R., W.B. Kannel, and L.P. Lyell, *An approach to longitudinal studies in a community: the Framingham Study.* Ann N Y Acad Sci, 1963. **107**: p. 539-56.
7. Splansky, G.L., et al., *The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination.* Am J Epidemiol, 2007. **165**(11): p. 1328-35.
8. Kannel, W.B., et al., *An investigation of coronary heart disease in families. The Framingham offspring study.* Am J Epidemiol, 1979. **110**(3): p. 281-90.
9. GenSalt Collaborative Research, G., *GenSalt: rationale, design, methods and baseline characteristics of study participants.* J Hum Hypertens, 2007. **21**(8): p. 639-46.
10. Perloff, D., et al., *Human blood pressure determination by sphygmomanometry.* Circulation, 1993. **88**(5 Pt 1): p. 2460-70.
11. Lavange, L.M., et al., *Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos.* Ann Epidemiol, 2010. **20**(8): p. 642-9.
12. Sorlie, P.D., et al., *Design and implementation of the Hispanic Community Health Study/Study of Latinos.* Ann Epidemiol, 2010. **20**(8): p. 629-41.
13. Wyatt, S.B., et al., *A community-driven model of research participation: the Jackson Heart Study Participant Recruitment and Retention Study.* Ethn Dis, 2003. **13**(4): p. 438-55.
14. Taylor, H.A., Jr., et al., *Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study.* Ethn Dis, 2005. **15**(4 Suppl 6): p. S6-4-17.
15. Bild, D.E., et al., *Multi-Ethnic Study of Atherosclerosis: objectives and design.* Am J Epidemiol, 2002. **156**(9): p. 871-81.
16. Yu, S., et al., *Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources.* J Am Med Inform Assoc, 2015. **22**(5): p. 993-1000.