Cell Reports Medicine, Volume *3*

Supplemental information

Correlations between complex human phenotypes vary by genetic background, gender, and environment

Michael Elgart, Matthew O. Goodman, Carmen Isasi, Han Chen, Alanna C. Morrison, Paul S. de Vries, Huichun Xu, Ani W. Manichaikul, Xiuqing Guo, Nora Franceschini, Bruce M. Psaty, Stephen S. Rich, Jerome I. Rotter, Donald M. Lloyd-Jones, Myriam Fornage, Adolfo Correa, Nancy L. Heard-Costa, Ramachandran S. Vasan, Ryan Hernandez, Robert C. Kaplan, Susan Redline, The Trans-Omics for Precision Medicine (TOPMed) Consortium, and Tamar Sofer

Correlations between complex human phenotypes vary by genetic background, gender, and environment

Elgart et al.

Supplementary Tables

Code		Description
annotated_sex_1		Biological sex
Race/ethnicity	*hispanic_or_latino_1*	Indicator of reported Hispanic or Latino ethnicity; only used samples where this agreed with "race_us_1"
	race_us_1	Reported race of participant according to the United States administrative definition of race; only used samples where this agreed with "hispanic_or_latino_1"
age_at_height_baseline_1		Age
height_baseline_1		Body height

bmi_baseline_1	Body mass index
antihypertensive_meds_1	Indicator for use of antihypertensive medication at the time of blood pressure measurement
bp_systolic_1	Resting systolic blood pressure from the upper arm in a clinical setting; only used samples if blood pressure lowering medications were not used (antihypertensive_meds_1)
bp_diastolic_1	Resting diastolic blood pressure from the upper arm in a clinical setting; only used samples if blood pressure lowering medications were not used (antihypertensive_meds_1)
lipid_lowering_medication_1	Indicates whether participant was taking any lipid-lowering medication at blood draw to measure lipids phenotypes
total_cholesterol_1	Blood mass concentration of total cholesterol; only used samples where no lipid medication was used (lipid_lowering_medication_1)
triglycerides_1	Blood mass concentration of triglycerides; only used samples where no lipid medication was used (lipid_lowering_medication_1)
hdl_1	Blood mass concentration of high-density lipoprotein cholesterol; only used samples where no lipid medication was used (lipid_lowering_medication_1)
ldl_1	Blood mass concentration of low-density lipoprotein cholesterol; only used samples where no lipid medication was used (lipid_lowering_medication_1)
hemoglobin_mcnc_bld_1	Measurement of mass per volume, or mass concentration (mcnc), of hemoglobin in the blood (bld)
hematocrit_vfr_bld_1	Measurement of hematocrit, the fraction of volume (vfr) of blood (bld) that is composed of red blood cells
rbc_ncnc_bld_1	Count by volume, or number concentration (ncnc), of red blood cells in the blood (bld)
wbc_ncnc_bld_1	Count by volume, or number concentration (ncnc), of white blood cells in the blood (bld)
basophil_ncnc_bld_1	Count by volume, or number concentration (ncnc), of basophils in the blood (bld)
eosinophil_ncnc_bld_1	Count by volume, or number concentration (ncnc), of eosinophils in the blood (bld)
neutrophil_ncnc_bld_1	Count by volume, or number concentration (ncnc), of neutrophils in the blood (bld)
lymphocyte_ncnc_bld_1	Count by volume, or number concentration (ncnc), of lymphocytes in the

	blood (bld)
monocyte_ncnc_bld_1	Count by volume, or number concentration (ncnc), of monocytes in the blood (bld)
platelet_ncnc_bld_1	Count by volume, or number concentration (ncnc), of platelets in the blood (bld)
mch_entmass_rbc_1	Measurement of the average mass (entmass) of hemoglobin per red blood cell(rbc), known as mean corpuscular hemoglobin (MCH)
mchc_mcnc_rbc_1	Measurement of the mass concentration (mcnc) of hemoglobin in a given volume of packed red blood cells (rbc), known as mean corpuscular hemoglobin concentration (MCHC)
mcv_entvol_rbc_1	Measurement of the average volume (entvol) of red blood cells (rbc), known as mean corpuscular volume (MCV)
pmv_entvol_bld_1	Measurement of the mean volume (entvol) of platelets in the blood (bld), known as mean platelet volume (MPV or PMV)
rdw_ratio_rbc_1	Measurement of the ratio of variation in width to the mean width of the red blood cell (rbc) volume distribution curve taken at +/- 1 CV, known as red cell distribution width (RDW)
cd40_1	Cluster of differentiation 40 ligand (CD40) concentration in blood.
crp_1	C-reactive protein (CRP) concentration in blood
eselectin_1	E-selectin concentration in blood.
icam1_1	Intercellular adhesion molecule 1 (ICAM1) concentration in blood
il1_beta_1	Interleukin 1 beta (IL1b) concentration in blood
il6_1	Interleukin 6 (IL6) concentration in blood
il10_1	Interleukin 10 (IL10) concentration in blood
il18_1	Interleukin 18 (IL18) concentration in blood
isoprostane_8_epi_pgf2a_1	Isoprostane 8-epi-prostaglandin F2 alpha (8-epi-PGF2a) concentration in urine
lppla2_act_1	Activity of lipoprotein-associated phospholipase A2 (LP-PLA2), also known as platelet-activating factor acetylhydrolase, measured in blood
lppla2_mass_1	Mass of lipoprotein-associated phospholipase A2 (LP-PLA2), also known as platelet-activating factor acetylhydrolase, measured in blood

mcp1_1	Monocyte chemoattractant protein-1 (MCP1), also known as C-C motif chemokine ligand 2, concentration in blood
mmp9_1	Matrix metalloproteinase 9 (MMP9) concentration in blood
mpo_1	Myeloperoxidase (MPO) concentration in blood
opg_1	Osteoprotegerin (OPG) concentration in blood
pselectin_1	P-selectin concentration in blood.
tnfa_1	Tumor necrosis factor alpha (TNFa) concentration in blood
tnfa_r1_1	Tumor necrosis factor alpha receptor 1 (TNFa-R1) concentration in blood
tnfr2_1	Tumor necrosis factor receptor 2 (TNFR2) concentration in blood

Table S1 (related to Figures 2,3). Phenotypes harmonized across participating studies by the TOPMed DCC used in this study

Code	**Description**
AHI	Apnea/Hypopnea Events (3% desat)
MinSpO2	Minimum oxyhemoglobin saturation during sleep
AvgSpO2	Mean oxygen saturation during sleep
SpO290	Percent sleep time with oxygen saturation less than 90%
Height	Height
BMI	Body Mass Index (kg/m^2)
WaistHip	Waist to Hip Ratio
FEV1FVC	FEV1 to FVC Ratio (%)
FEV1	Forced Expiratory Volume
FEVC	Forced Vital Capacity
PhysHlth	Aggregate Physical Health Scale
BrchIdx	Overall Ankle Brachial Index (occ. failure incl.)
FastInsl	Insulin, fasting (converted to mU/L)
OGTTInsl	Insulin, post OGTT (converted to mU/L)
eGFRnodemo	eGFR based on serum cystatin C w/o demographics
eGFRwdemo	eGFR based on serum cystatin C, serum creatinine, gender, age and race
HOMA	HOMA-IR index of Insulin Resistance
BCell	HOMA-BCELL index of Insulin Resistance
GlycHemo	Glycosylated Hemoglobin in SI units (mmol/mol)
ECGAbnorm_Mj	Major ECG Abnormalities
ECGAbnorm_Min	Minor ECG Abnormalities
EpSleep	Epworth Sleepiness Scale

SleepDur	Self-reported sleep duration (difference in bed and wake times) (hours)
Insom	Women's Health Initiative Insomnia Rating Scale
SysBP	Systolic Blood Pressure
DiasBP	Diastolic Blood Pressure
Arter	Mean arterial pressure
Pulse	Pulse pressure
WBC	White Blood Count (x10e9)
RBC	Red Blood Count (x10e12)
Hemoglob	Hemoglobin (g/dL)
Hemocrit	% Hematocrit
CorpVol	Mean Corpuscular Volume (fl)
MnCorpHemo	Mean Corpuscular Hemoglobin (pg)
MnCorpHemoConc	Mean Corpuscular Hemoglobin Concentration (g/dL)
RedCellDistWdth	% Red Cell Distribution Width
PlateletCnt	Platelet Count (x10e9)
NeutCnt	Neutrophil Count (x10e9)
LymphCnt	Lymphocyte Count (x10e9)
MonoCnt	Monocyte Count (x10e9)
EosCnt	Eosinophil Count (x10e9)
BasoCnt	Basophil Count (x10e9)
Chol	Total cholesterol (mg/dL)
Triglyc	Triglycerides (mg/dL)
HDLChol	HDL-cholesterol (mg/dL)
LDLChol	LDL-cholesterol (mg/dL)
FastGluc	Glucose, fasting (mg/dL)
OGTTGluc	Glucose, post OGTT (mg/dL)
GlycoHemo	% Glycosylated Hemoglobin
Creat	Creatinine (mg/dL)
UrineCreat	Urine creatinine, random (mg/dL)
UrineMicroAlb	Urine microalbumin, random (mg/dL)
AlbCreat	Albumin/creatinine ratio (mg/g)
Fe	Iron (ug/dL)
FeBindCap	Total Iron Binding Capacity (TIBC) (ug/dL)
TransSat	% Transferrin saturation
CReactProt	High-sensitivity C-Reactive Protein (mg/L)
HrtRt	Heart Rate
PRDur	PR duration
QRSDur	QRS duration
QTDur	QT duration
Sex	Sex

Table S3 (related to Figure 4). Code and descriptions of HCHS/SOL phenotypes used in this study

Code	*Phenotype*	*N*	*Females*	*Males*	*AFR*	*AMR*	*CSA*	*EAS*	*EUR*	*MID*
30690	Cholesterol	420607	227266	193341	6212	938	8422	2572	400963	1500
30760	HDL cholesterol	385023	206578	178445	5754	854	7688	2342	367021	1364
30780	LDL direct	419831	226901	192930	6200	938	8404	2568	400223	1498
30870	Triglycerides	420271	227138	193133	6211	937	8415	2570	400639	1499
1160	Sleep duration	429528	232661	196854	6382	959	NA	2631	418009	1547
21001	Body mass index (BMI)	439590	237771	201805	6545	971	8646	2693	419163	1572
50	Standing height	438478	237363	201115	6556	972	8657	2697	419596	NA
4079	Diastolic blood pressure	416959	225161	191786	6551	959	8641	2600	396667	1541
4080	Systolic blood pressure	416955	225159	191784	6551	959	8641	2600	396663	1541

Table S5 (related to Figure 1). Pan-UKBB summary statistics for phenotypes used in LDSC estimations of heritability and genetic correlations. Description of all the GWAS used in this work. All data were downloaded from the Pan-UKBB website (https://pan.ukbb.broadinstitute.org/). Phenotype names along with internal codes are provided for reproducibility. Each phenotype is further broken down by ancestry groups (AFR – African ancestry, AMR – Admixed American ancestry, CSA - Central/South Asian ancestry, EAS - East Asian ancestry, EUR – European ancestry, MID - Middle Eastern ancestry)

Supplementary Figures

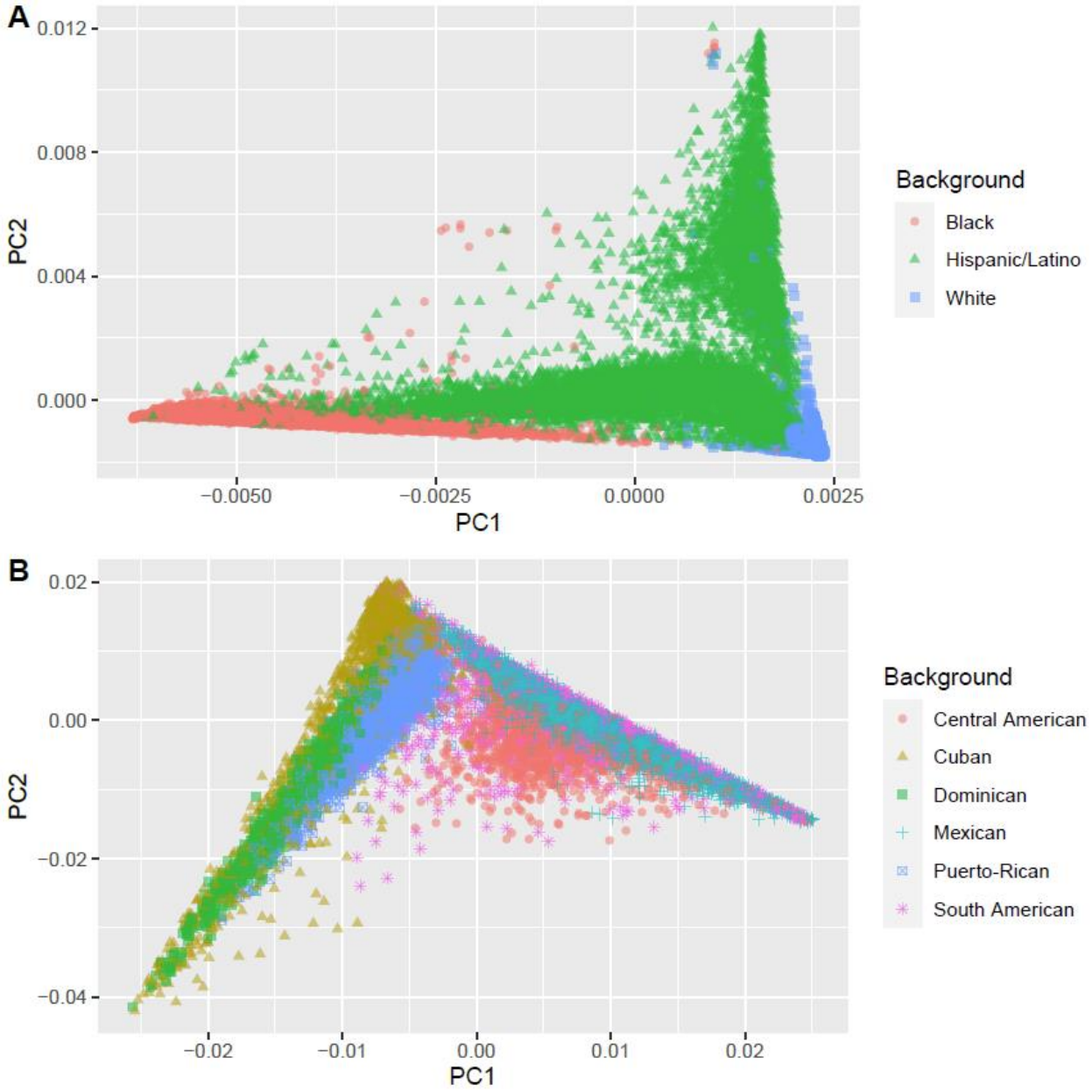

Figure S1 (related to Figures 2-4). Populations backgrounds as captured by PCA analysis

Shown here are PCA plots for first and second principal components for the two datasets used in this study. The TOPMed dataset **(A)** and the HCHS/SOL dataset **(B)**.

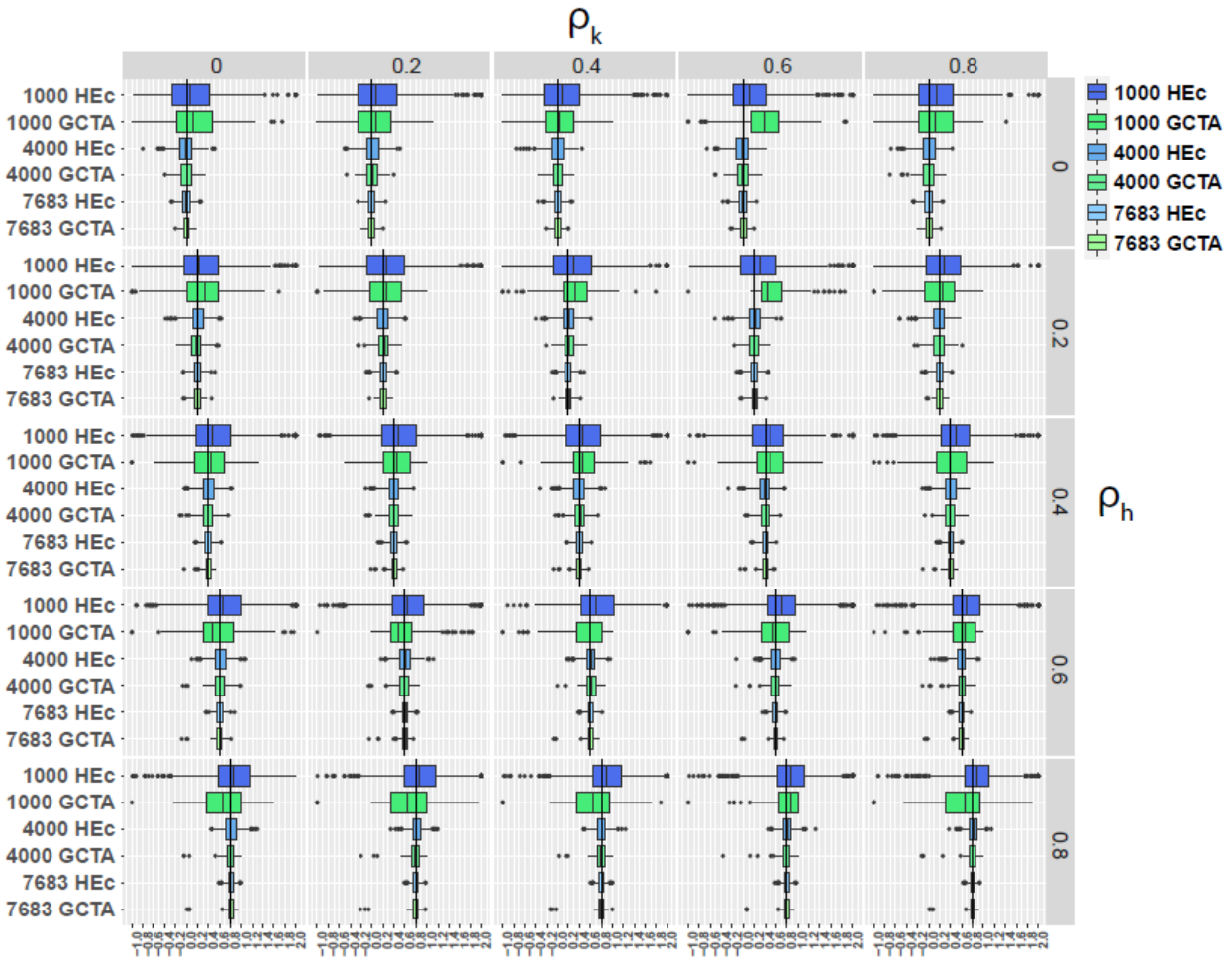

Figure S2 (related to Figure 1). Comparison of simulations compare HEc, the ground truth and GCTA-GREML
Two relatedness matrices were used to simulate phenotypes with known correlation coefficients (ρ_h, ρ_k). Each phenotype was simulated 1000 times in 1000, 4000 and 7683 people. Shown here are the boxplots of distributions of estimated ρ_k for both our closed-form HE approach (HEc, blue colors) as well as the gold-standard REML method GCTA (green colors) [12].

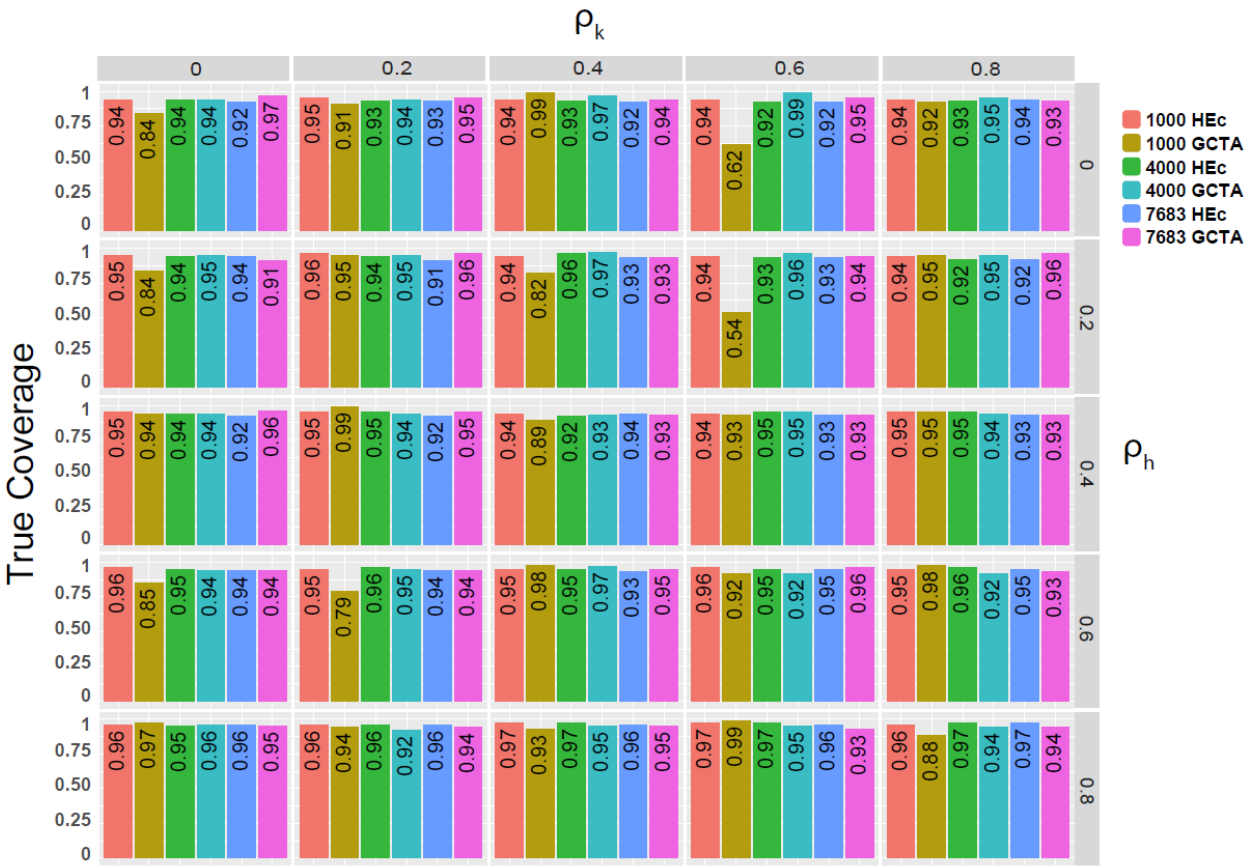

Figure S3 (related to Figure 1). Comparison of confidence interval coverages in simulations of HEc and GCTA-GREML approach

Two relatedness matrices were used to simulate phenotypes with known correlation coefficients (ρ_h, ρ_k). Each phenotype was simulated 1000 times in 1000, 4000 and 7683 people. Next 95% Confidence Intervals were calculated via HEc using the quantile methos, as well as GCTA [12]. For each such combinations we display the true coverage (i.e. the fraction of the simulated cases where true value of ρ_k was within the reported confidence interval).

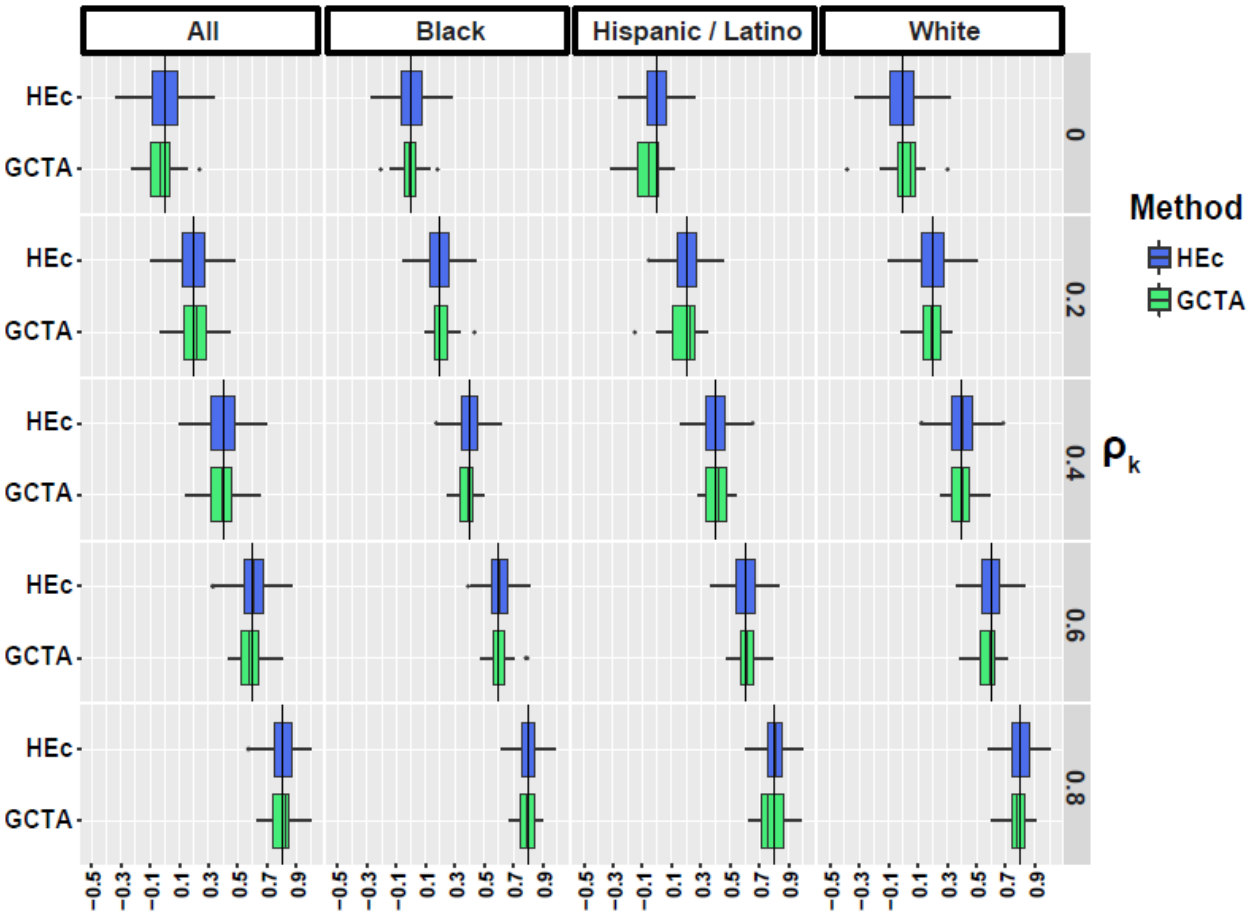

Figure S4 (related to Figure 1). Comparison HEc and GCTA-GREML applied on different populations in simulations

Relatedness data from three TOPMed populations (as well as a joint population consisting of equal number of all three) were used to simulate phenotypes pairs with known genetic correlation coefficients (ρ_k). Each phenotype was simulated 1000 times for 7706 people. Shown here are the boxplots of distributions of estimated ρ_k for HEc (blue colors) as well as GCTA-GREML.

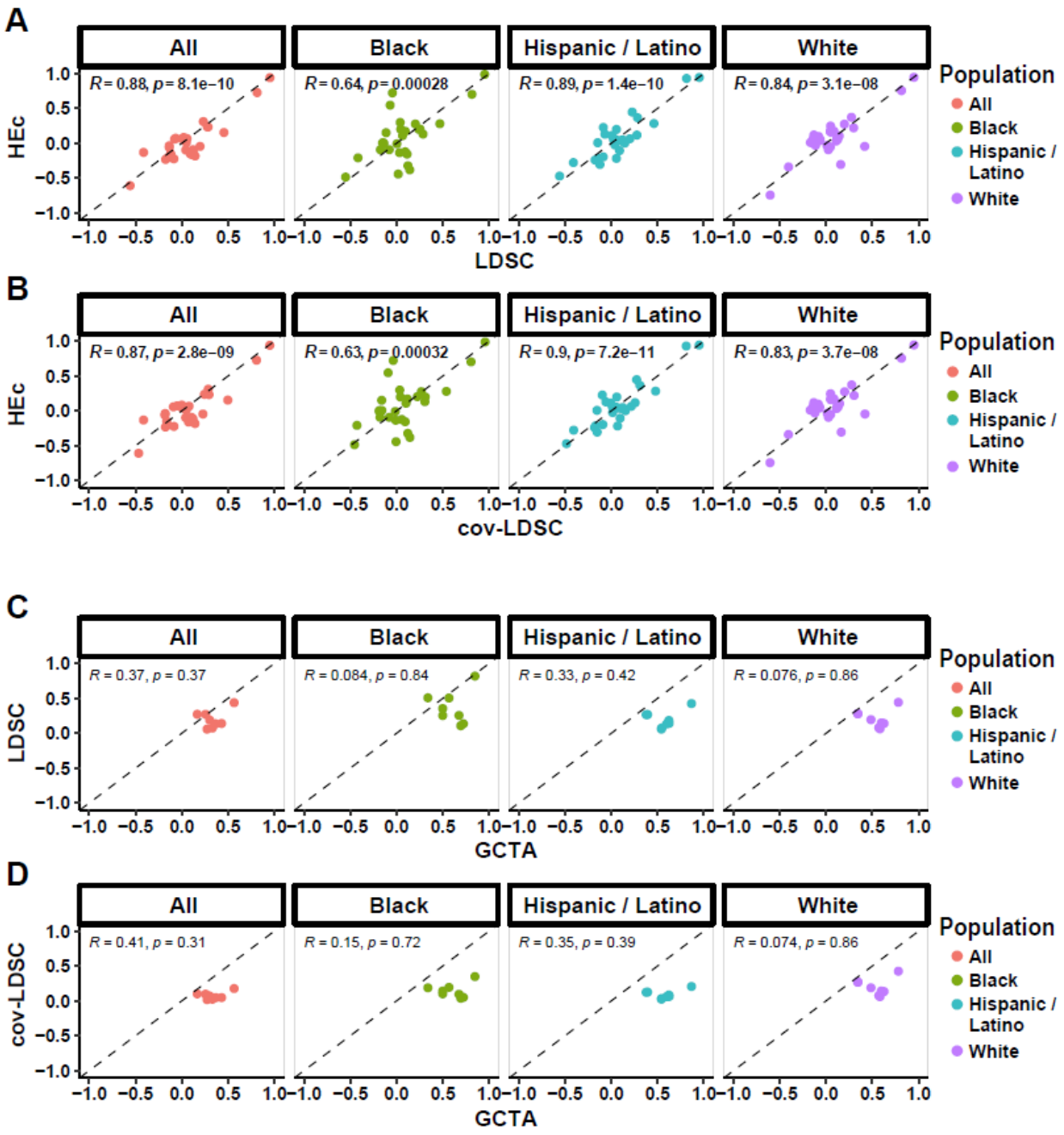

Figure S5 (related to Figure 1). Comparison of results of HEc, GCTA-GREML and two LD-score based methods for computation of phenotype heritabilities and genetic correlations from the TOPMed dataset stratified by populations

Comparison of results between our HEc method, GREML and two LD-score based methods for computation of genetic correlations and genetic heritabilities in the TOPMed dataset either joint or stratified by populations **(A)** Comparison of ρ_k (genetic correlation coefficient) estimations between HEc and LDSC for 112 pairs of 8 phenotypes selected from the diverse TOPMed cohort **(B)** Comparison of $\boldsymbol{\rho_k}$ estimations for same cohorts between HE and cov-LDSC which is a novel methods which may improve LDSC performance in mismatched populations (see Materials and Methods for description of the GWASes used) **(C)** Comparison of heritability estimates between GCTA and LDSC for the 8 phenotypes either in joint dataset or stratified by self-reported ancestry. **(D)** same as (C) but using cov-LDSC.

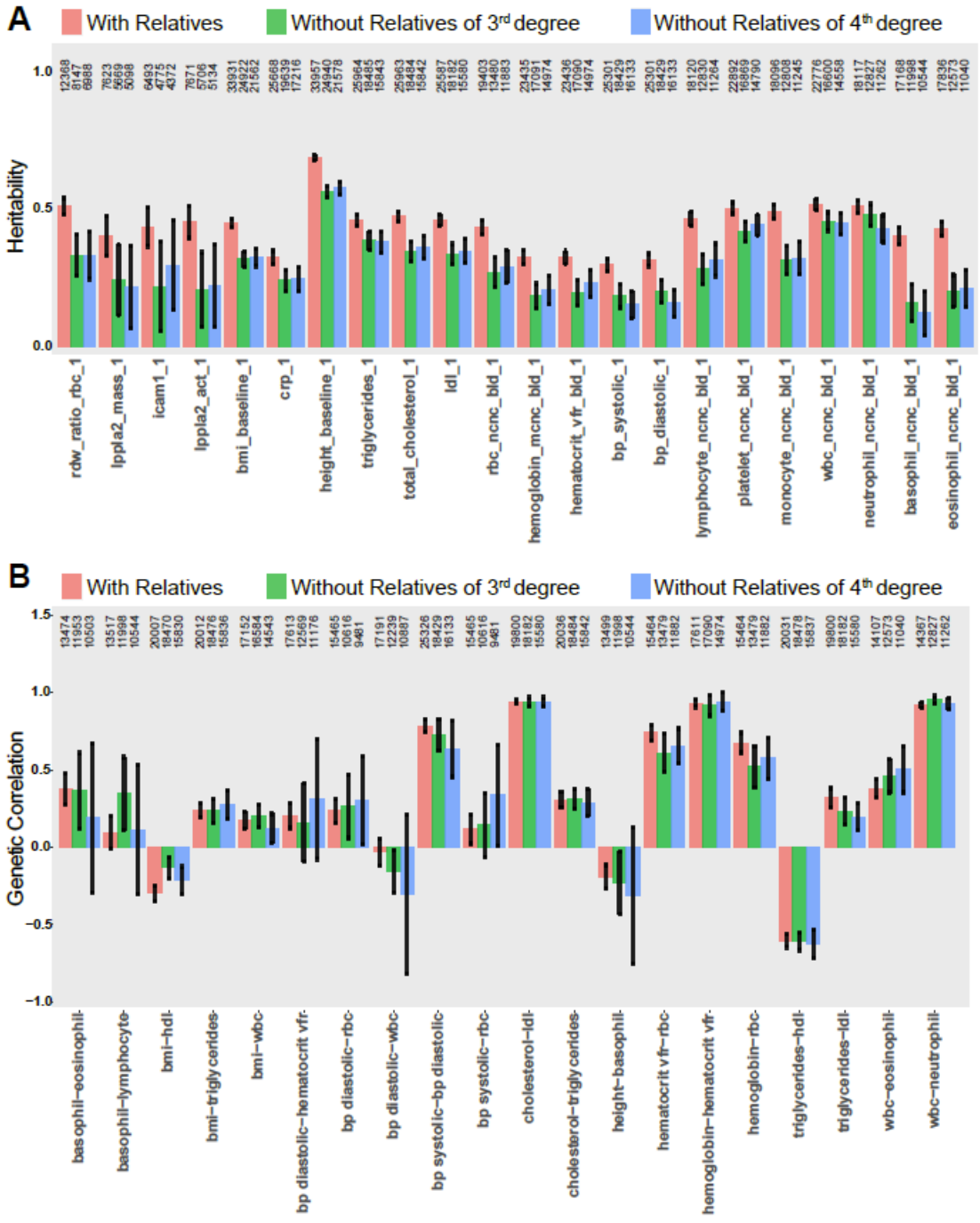

Figure S6 (related to Figure 1). Sensitivity analysis studying the effect of presence of relatives in the data on heritability and genetic correlation estimates

(A, B) Analysis of the TOPMed cohort with the relatives present (pink), with relatives of 3rd degree and more distant removed (green), and with relatives of 4th degree and more distant removed (blue) for selected examples of phenotypes with regard to **(A)** Heritability and **(B)** Genetic correlations (ρ_k).

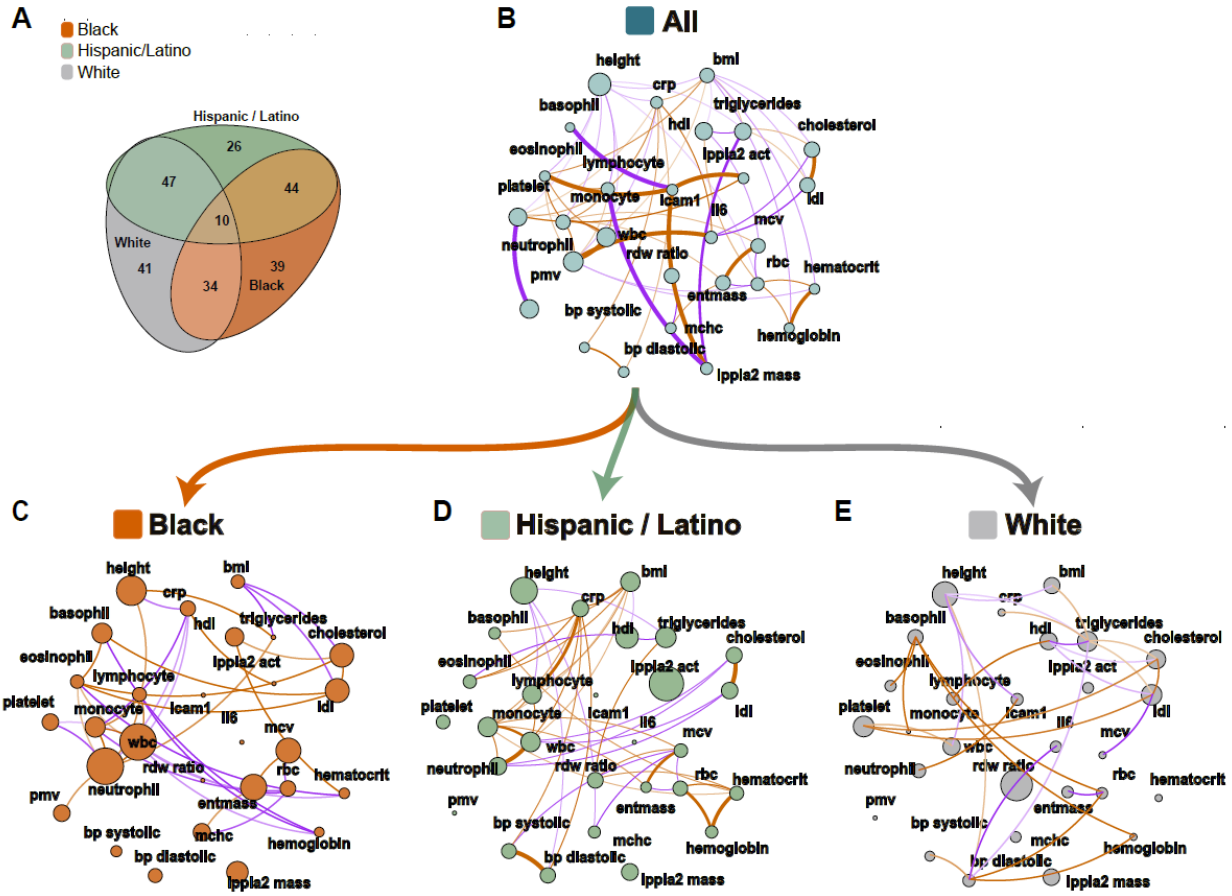

Figure S7 (related to Figure 3). Certain genetic correlations are background-specific.

(B-E) Correlation plots where each phenotype is represented by a node and the correlations are represented by connections (edges) between nodes. The size of the node is proportional to the phenotype heritability. The thickness of the edge is proportional to the strength of correlation and the color represents magnitude: orange represents positive and purple negative correlation. **(B)** Genetic correlations (ρ_k) between the 28 phenotypes in the combined TOPMed dataset (p-value < 0.05) **(C, D, E)** Genetic correlations (ρ_k) between the 28 phenotypes in the background-specific subsets of the TOPMed dataset - Black (**C**, orange), Hispanic/Latino (**D**, marine) and White (**E**, grey). **(A)** Venn diagram depicting the overlap in significant phenotype-pairs between Black (orange), Hispanic/Latino (marine) and White (grey) for genetic correlation.

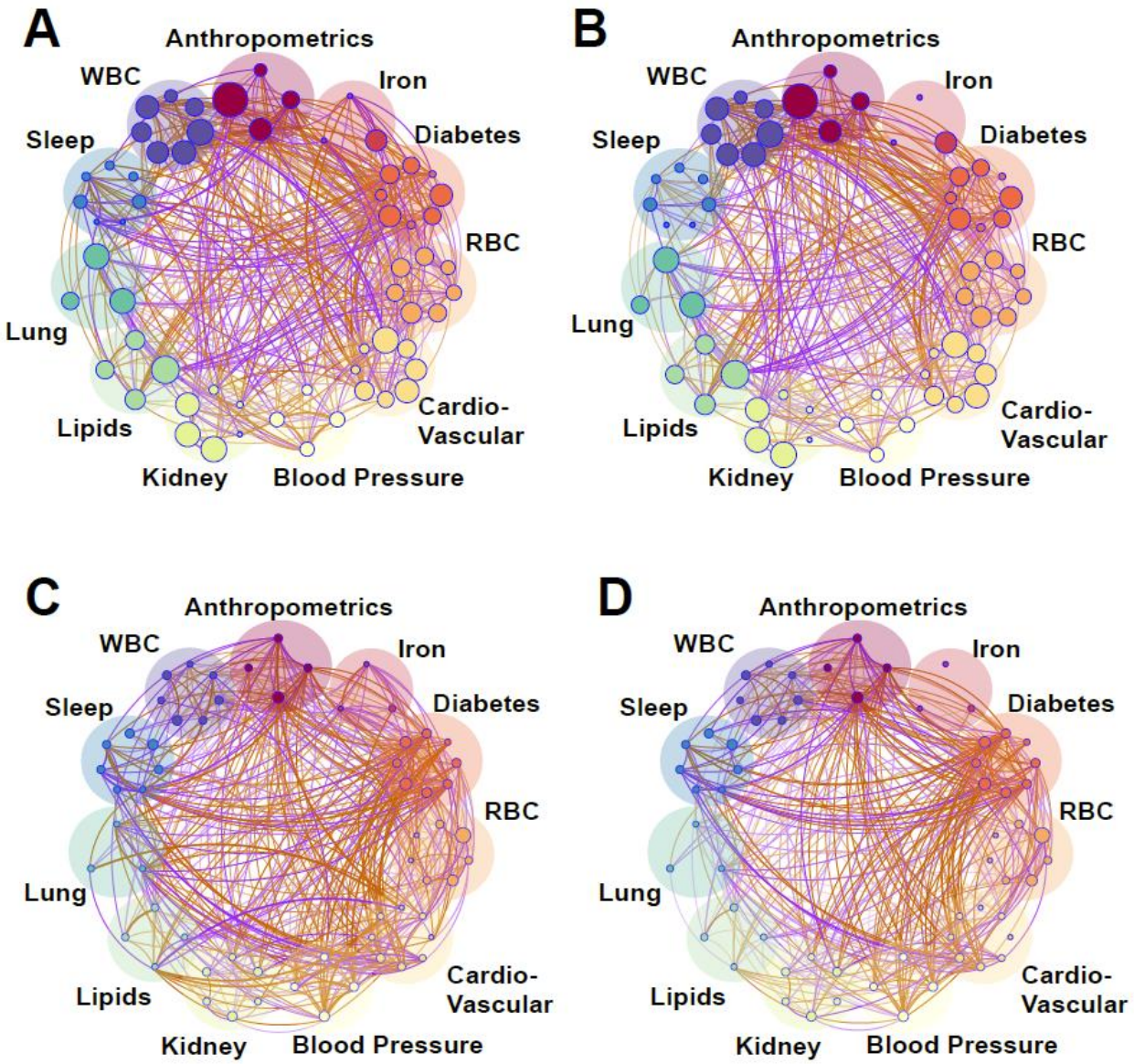

Figure S8 (related to Figure 4). Genetic and environmental correlations and heritabilities of 61 phenotypes in self-reported Hispanics/Latinos.

Correlation plots between the 61 phenotypes in the TOPMed HCHS/SOL dataset. Each phenotype is represented by a node (colored small circles) with the size of the circle proportional to the phenotype heritability. The correlations are represented by connections (edges) between nodes (phenotypes). The nodes are grouped into phenotypic domains (colored semi-transparent circles labelled Anthropometrics, Iron, etc.). The thickness of the edge is proportional to the strength of correlation and the color represents magnitude: orange represents positive and purple negative correlation. **(A)** Genetic correlations (ρ_k) between the 61 phenotypes (p-value < 0.05) **(B)** Fractional genetic correlations (ρ_{Nk}) between the 61 phenotypes (p-value < 0.05) **(C)** Household correlations (analog of genetic correlation but for household data; see Materials and Methods) between the 61 phenotypes (p-value < 0.05) **(D)** Fractional household correlations (ρ_{Nh}) between the 61 phenotypes (p-value < 0.05).

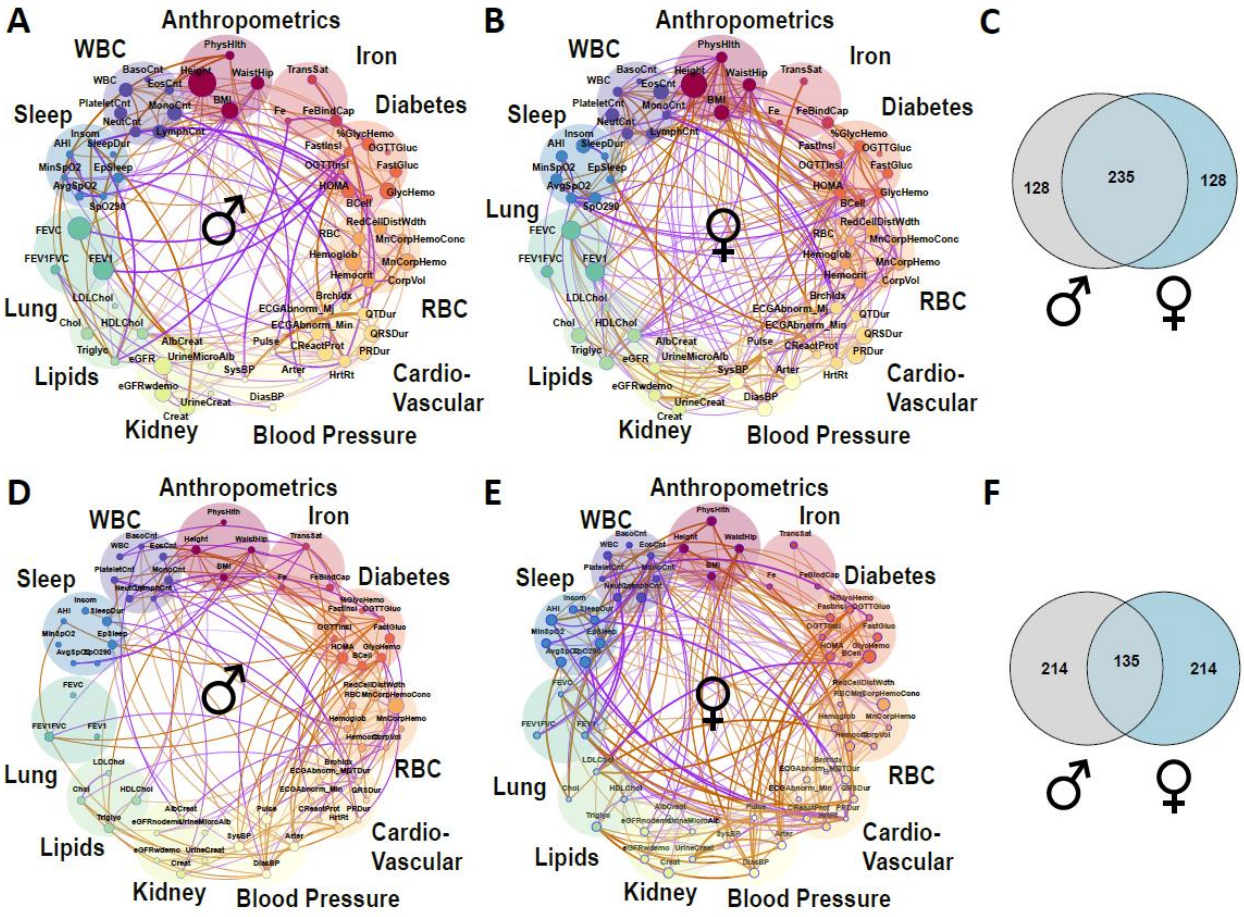

Figure S9 (related to Figure 5). Genetic and environmental correlations and heritabilities differ by gender in Hispanics/Latinos.

(A,B-D,E) Correlation plots where each phenotype is represented by a node and the correlations are represented by connections (edges) between nodes. The size of the node is proportional to the phenotype heritability. The thickness of the edge is proportional to the strength of correlation and the color represents magnitude: orange represents positive and purple negative correlation. Shown are genetic correlations (ρ_k) between the 61 phenotypes in the extended HCHS/SOL dataset (p-value < 0.05). Correlations and heritabilities as measured in males **(A, D)** and females **(B, E).** The top panels represent genetic correlations **(A, B)** and the bottom panels **(D, E)** represent the household correlations. (**C, F**) Venn diagrams depicting the overlap in significant phenotype-pairs between Males and Females for genetic correlation **(C)**, and household correlation **(F)** .

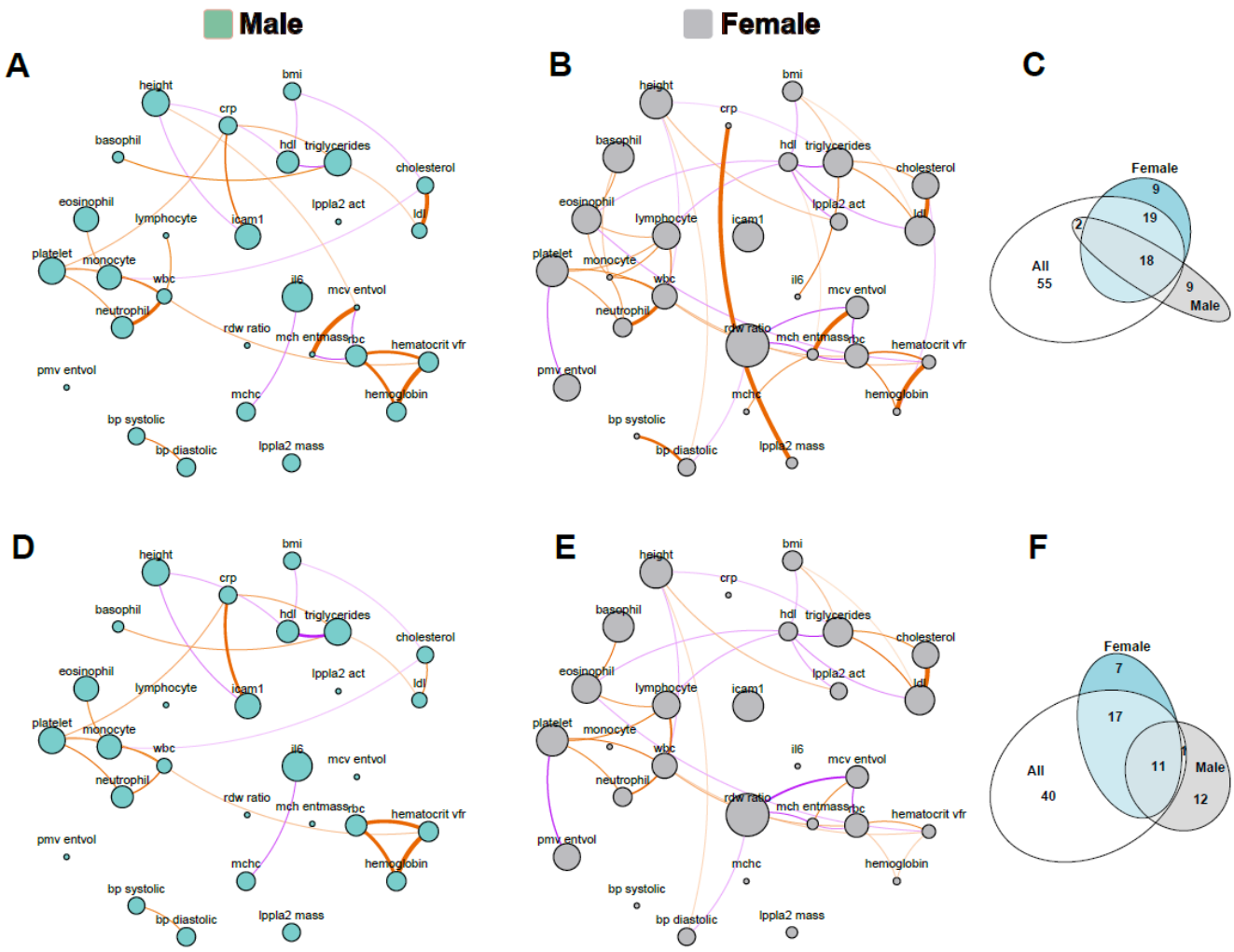

Figure S10 (related to Figure 5). Genetic correlations in individuals of White background stratified by gender.
(A-D) Correlation plots where each phenotype is represented by a node and the correlations are represented by connections (edges) between nodes. The size of the node is proportional to the phenotype heritability. The thickness of the edge is proportional to the strength of correlation and the color represents magnitude: orange represents positive and purple negative correlation. Shown are genetic correlations (ρ_k) **(A, D)** between the 18 phenotypes in the TOPMed dataset (p-value < 0.05; $|\rho_k| > 0.05$); and fractional genetic correlations (ρ_{fk}) **(A, D)** between the 18 phenotypes in the TOPMed dataset (p-value < 0.05; $|\rho_{fk}| > 0.05$). The dataset was stratified by males **(A, D)** and females **(B, E).** Venn diagrams depicting the overlap in significant phenotype-pairs **(C,F)** between males, females and a combined dataset (All) for genetic correlation **(C)**, and fractional household correlation **(F)**.

References

1. LaVange LM, Kalsbeek WD, Sorlie PD, et al (2010) Sample Design and Cohort Selection in the Hispanic Community Health Study/Study of Latinos. Ann Epidemiol 20:642–649
2. Sorlie PD, Avilés-Santa LM, Wassertheil-Smoller S, et al (2010) Design and Implementation of the Hispanic Community Health Study/Study of Latinos. Ann Epidemiol 20:629–641
3. Kannel WB, Feinleib M, Mcnamara PM, Garrison RJ, Castelli WP (1979) An investigation of coronary heart disease in families: The framingham offspring study. Am J Epidemiol 110:281–290
4. Splansky GL, Corey D, Yang Q, et al (2007) The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: Design, recruitment, and initial examination. Am J Epidemiol 165:1328–1335
5. Dawber TR, Kannel WB, Lyell LP (1963) AN APPROACH TO LONGITUDINAL STUDIES IN A COMMUNITY: THE FRAMINGHAM STUDY. Ann N Y Acad Sci 107:539–556
6. Wright JD, Folsom AR, Coresh J, et al (2021) The ARIC (Atherosclerosis Risk In Communities) Study: JACC Focus Seminar 3/8. J Am Coll Cardiol 77:2939–2959
7. Bild DE, Bluemke DA, Burke GL, et al (2002) Multi-Ethnic Study of Atherosclerosis: Objectives and design. Am J Epidemiol 156:871–881
8. Friedman GD, Cutter GR, Donahue RP, Hughes GH, Hulley SB, Jacobs DR, Liu K, Savage PJ (1988) Cardia: study design, recruitment, and some characteristics of the examined subjects. J Clin Epidemiol 41:1105–1116
9. Taylor HA, Wilson JG, Jones DW, Sarpong DF, Srinivasan A, Garrison RJ, Nelson C, Wyatt SB TOWARD RESOLUTION OF CARDIOVASCULAR HEALTH DISPARITIES IN AFRICAN AMERICANS: DESIGN AND METHODS OF THE JACKSON HEART STUDY.
10. Wyatt SB, Diekelmann N, Henderson F, Andrew ME, Billingsley G, Felder SH, Fuqua S, Jackson PB (2003) A community-driven model of research participation: The Jackson Hearth Study participant recruitment and retention study. Ethn Dis 13:438–455
11. Redline S, Tishler P V., Tosteson TD, Williamson J, Kump K, Browner I, Ferrette V, Krejci P (1995) The Familial Aggregation of Obstructive Sleep Apnea. Am J Respir Crit Care Med 151:682–687
12. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: A tool for genome-wide complex trait analysis. Am J Hum Genet 88:76–82