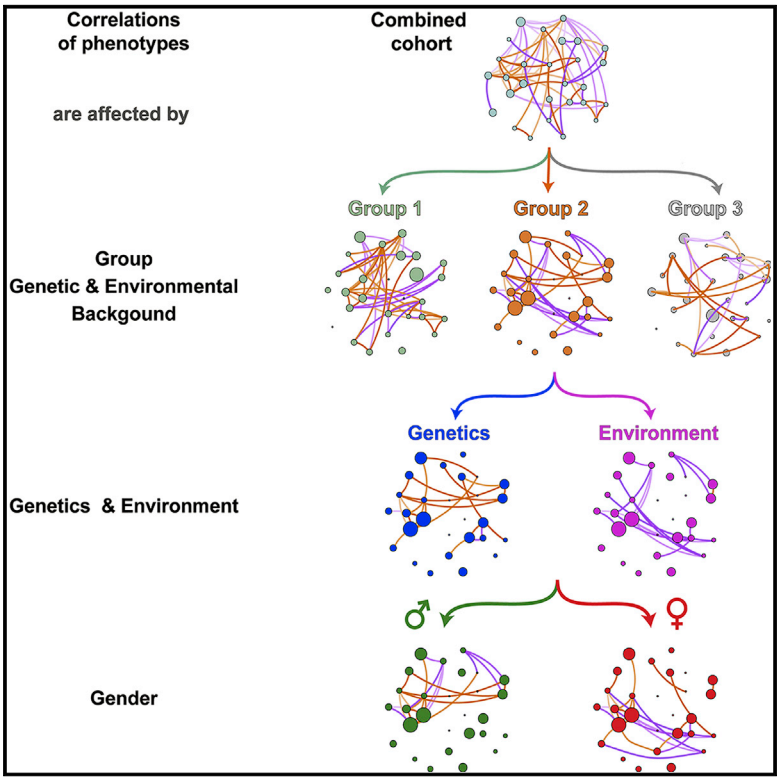
Cell Reports
Medicine

Article

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

Graphical abstract



Highlights

- An estimator of genetic correlation generalizes the Pearson correlation estimator
- Fractional genetic correlation quantifies genetic fraction of a phenotypic correlation
- Genetic correlations between phenotypes vary by genetic and environmental determinants

Authors

Michael Elgart, Matthew O. Goodman, Carmen Isasi, ..., Susan Redline, The Trans-Omics for Precision Medicine (TOPMed) Consortium, Tamar Sofer

Correspondence

melgart@bwh.harvard.edu (M.E.), tsofer@bwh.harvard.edu (T.S.)

In brief

Elgart et al. develop a closed-form estimator for genetic correlation between phenotypes called HEc, generalizing the Pearson estimator. Fractional correlation estimates that quantify contributions of genetics and environment to the overall phenotypic correlation demonstrate that both genetics and environment contribute to phenotypic correlation differences between groups of individuals.

Elgart et al., 2022, Cell Reports Medicine 3, 100844
December 20, 2022
https://doi.org/10.1016/j.xcrm.2022.100844

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

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

Michael Elgart,[1,2,*] Matthew O. Goodman,[1,2] Carmen Isasi,[3] Han Chen,[4,5] Alanna C. Morrison,[4] Paul S. de Vries,[4] Huichun Xu,[6] Ani W. Manichaikul,[7] Xiuqing Guo,[8] Nora Franceschini,[9] Bruce M. Psaty,[10] Stephen S. Rich,[11] Jerome I. Rotter,[8] Donald M. Lloyd-Jones,[12] Myriam Fornage,[4,13] Adolfo Correa,[14] Nancy L. Heard-Costa,[15,16] Ramachandran S. Vasan,[15,17] Ryan Hernandez,[18] Robert C. Kaplan,[3,19] Susan Redline,[1,2] The Trans-Omics for Precision Medicine (TOPMed) Consortium, and Tamar Sofer[1,2,20,21,*]

[1]Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, MA, USA
[2]Department of Medicine, Harvard Medical School, Boston, MA, USA
[3]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA
[4]Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA
[5]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA
[6]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA
[7]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA
[8]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA
[9]Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA
[10]Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology, and Health Services, University of Washington, Seattle, WA, USA
[11]Center for Public Health Genomics, University of Virginia School of Medicine, Charlottesville, VA, USA
[12]Department of Preventive Medicine, Northwestern University, Chicago, IL, USA
[13]Brown Foundation Institute of Molecular Medicine, McGovern Medical School, University of Texas Health Science Center at Houston, Houston, TX, USA
[14]Department of Population Health Science, University of Mississippi Medical Center, Jackson, MS, USA
[15]Boston University and National Heart Lung and Blood Institute's Framingham Heart Study, Framingham, MA, USA
[16]Department of Neurology, Boston University School of Medicine, Boston, MA, USA
[17]Preventive Medicine & Epidemiology, and Cardiovascular Medicine, Medicine, Boston University School of Medicine, and Epidemiology, Boston University School of Public Health, Boston, MA, USA
[18]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, USA
[19]Fred Hutchinson Cancer Research Center, Division of Public Health Sciences, Seattle, WA, USA
[20]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[21]Lead contact
*Correspondence: melgart@bwh.harvard.edu (M.E.), tsofer@bwh.harvard.edu (T.S.)
https://doi.org/10.1016/j.xcrm.2022.100844

SUMMARY

We develop a closed-form Haseman-Elston estimator for genetic and environmental correlation coefficients between complex phenotypes, which we term HEc, that is as precise as GCTA yet ~20× faster. We estimate genetic and environmental correlations between over 7,000 phenotype pairs in subgroups from the Trans-Omics in Precision Medicine (TOPMed) program. We demonstrate substantial differences in both heritabilities and genetic correlations for multiple phenotypes and phenotype pairs between individuals of self-reported Black, Hispanic/Latino, and White backgrounds. We similarly observe differences in many of the genetic and environmental correlations between genders. To estimate the contribution of genetics to the observed phenotypic correlation, we introduce "fractional genetic correlation" as the fraction of phenotypic correlation explained by genetics. Finally, we quantify the enrichment of correlations between phenotypic domains, each of which is comprised of multiple phenotypes. Altogether, we demonstrate that the observed correlations between complex human phenotypes depend on the genetic background of the individuals, their gender, and their environment.

INTRODUCTION

Both genetics and environment determine human phenotypes and the correlations between them.[1,2] Correlations can arise due to multiple forms of causal relationships including common genetic and environmental determinants, (bi)directional causal associations and others. The correlations between phenotypes can reveal genetic architecture, help uncover gene functions

and disease mechanisms, improve diagnosis, and aid in therapeutic interventions.[3] Given appropriate data, phenotypic correlations can be decomposed into genetic and environmental components by estimating corresponding measures, such as genetic correlation.[4]

Several studies have leveraged the data generated by large studies with genotyped individuals and multiple measured phenotypes (e.g., BioBank Japan[5] and UK Biobank[6]), to estimate genetic correlations between various phenotypes.[7–10] Dozens of pairwise correlations between phenotypes were estimated and reported, mostly based on studies with participants of European or East Asian ancestries[9,10] and of mixed genders. Genetic ancestry, as well as biological sex and sociocultural roles of gender, all contribute to differences in phenotypic distributions between different groups of people due to both underlying genetics and different environmental exposures.[11–13] The genetic background manifests in allele frequencies, effect sizes, and, more generally, genetic architectures.[14,15] Sociocultural measures, captured in part by self-reported race/ethnicity, are related to behavioral, environmental, and psychosocial exposures such as smoking, alcohol, nutrition, physical activity, and stress,[16,17] modifying the effect of genetic variants. In this work, we refer to the collective characteristics of a studied race and ethnicity-based group as a "background" to reflect the fact that no single genetic or social measure can be used to define the group, although its individuals may self-identify using a specific, pre-defined label, whether by choice or by the set of options presented to them. Notably, these background groups are enriched with patterns of both genetics and sociodemographic (and environmental and cultural) similarity, and all factors may ultimately affect the expression of genetic effects. However, we cannot, using existing data, separate these different influences. While a handful of studies reported differences in genetic correlation across different background groups,[18–20] these are limited in the number of studied phenotypes and backgrounds. Similarly, while biological sex-specific heritabilities of complex phenotypes have been reported previously,[21–23] gender-specific correlations between phenotypes have not been comprehensively studied. Genetic correlations between phenotypes can be leveraged for prediction of health conditions[24] and differences between these correlations across groups may affect the application of prediction models. An environmental correlation between two phenotypes may suggest the possibility of a lifestyle change to alleviate the burden of disease related to these phenotypes.

The two main computational approaches that are used to estimate genetic correlation are the genetic restricted maximum likelihood analysis (GREML),[25–27] which requires individual-level genotypes and is computationally challenging when analyzing datasets with thousands of individuals; and the linkage disequilibrium score regression (LDSC),[28] which uses genome-wide association study (GWAS) summary statistics. LDSC requires reliable GWASs, and can be inaccurate when there is genetic heterogeneity between the target sample and reference linkage disequilibrium (LD) panel,[29,30] and thus cannot be used reliably for admixed or multi-ancestry analyses. Thus, both GREML and LDSC approaches are limited when analyzing datasets that include tens of thousands of genetically diverse individuals.

Here, we estimated the heritabilities, as well as genetic and environmental correlations between phenotypes in joint as well as background- and gender-stratified analysis. First, we derived a closed-form solution for the estimation of genetic and environmental correlation coefficients within the Haseman-Elston regression framework, which we termed HEc (Haseman-Elston closed form). Second, we applied the algorithm to study heritabilities and genetic correlations for 28 phenotypes in the Trans-Omics in Precision Medicine (TOPMed) program[31,32] dataset, with a large representation of individuals of White, Black, and Hispanic/Latino backgrounds. We then focused on Hispanics/Latinos from the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) cohort,[33] and utilized available data on shared household (representing shared environmental exposure) and compared gender-specific genetic and environmental correlations across a larger panel of 61 phenotypes. Finally, we performed domain-level enrichment analysis to identify genetic and environmental correlations between phenotypic domains.

RESULTS

HEc is a fast and precise estimator for genetic correlation

We developed an estimator, HEc, for genetic correlation within the Haseman-Elston regression framework (see STAR Methods for detailed derivation). The HEc estimator generalizes the Pearson correlation coefficient, and it has the closed-form formula:

$$\widehat{\rho}_k = \frac{\widehat{\epsilon}_1^T \boldsymbol{S_k} \widehat{\epsilon}_2}{\sqrt{(\widehat{\epsilon}_1^T \boldsymbol{S_k} \widehat{\epsilon}_1)(\widehat{\epsilon}_2^T \boldsymbol{S_k} \widehat{\epsilon}_2)}}$$

where $\widehat{\epsilon}_1$, $\widehat{\epsilon}_2$ are the residuals of traits y_1, y_2 after regression on covariates, and $\boldsymbol{S_k}$ is a matrix that depends on the genetic relationship between individuals in the dataset and potentially other matrices used to model correlation between individuals (e.g., matrices modeling shared environmental influences). The reader may see that it generalizes the standard Pearson estimator since replacing the matrix $\boldsymbol{S_k}$ by the identity matrix and by assuming that the residuals are those obtained by regressing the traits on intercepts (i.e., subtracting their mean), reduces the estimate to the Pearson correlation coefficient.

We evaluated the performance of HEc in simulations across multiple parameters and background groups and compared it with the GREML algorithm[4] implemented in the GCTA (Genome-wide Complex Trait Analysis) software[32] (Figures 1A and S2–S5).

Figure S1 visualizes TOPMed and HCHS/SOL individuals in the space generated by the first two genetic principal components (PCs). One can see that the White, Black, and Hispanic/Latino background groups are only partly separated in PC space, and individuals from Hispanic/Latino background, who are admixed with high proportions of European, Amerindian, and, to lower extent, African genetic ancestries, are represented throughout the PC space. In Figure S1B that focuses on HCHS/SOL, we also highlight finer self-reported Hispanic/Latino backgrounds, including Mexican, Puerto Rican, etc., which show some clustering, indicating some genetic structure within Hispanic/Latino individuals (as is known). This figure demonstrates that

CellPress
OPEN ACCESS

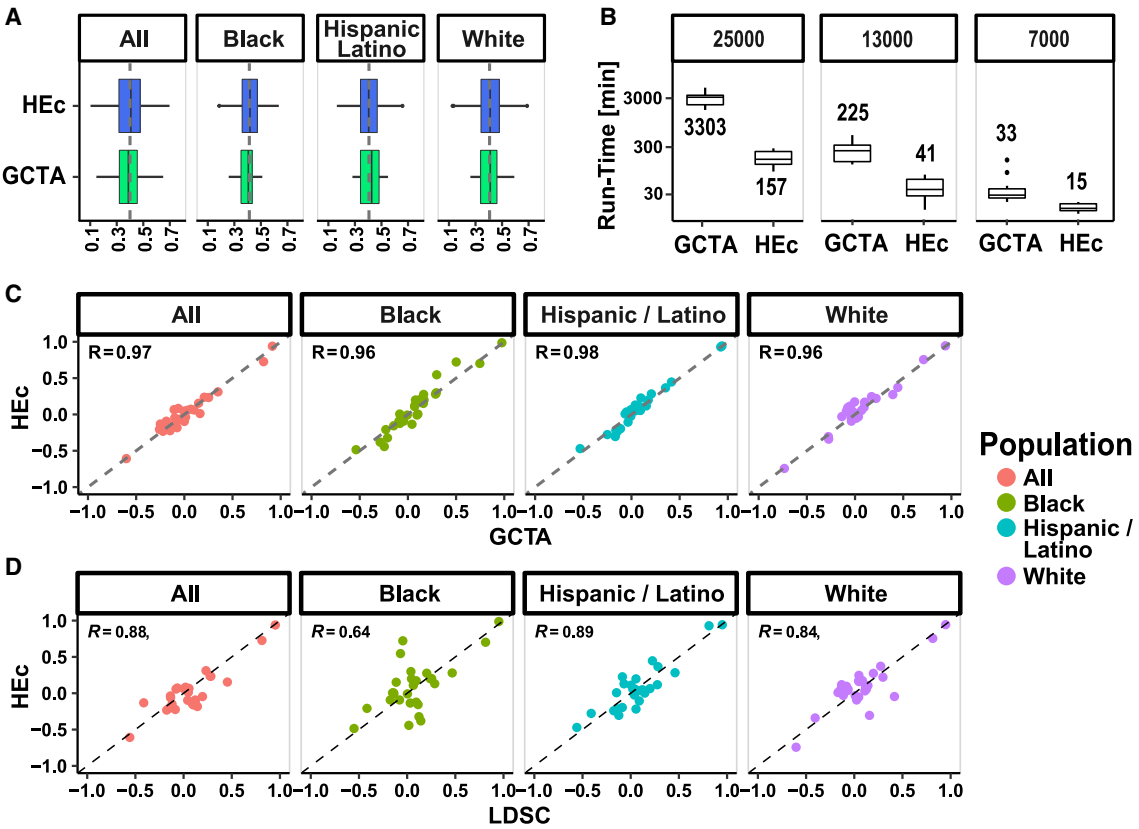


Figure 1. Closed-form HEc is as accurate as GCTA on diverse and admixed populations while being up to 20× faster

Comparison of accuracy and computation speed between our HEc method, GCTA-GREML, and an LD-score-based method (LDSC) based on simulations and the TOPMed dataset. Relatedness data from three self-reported TOPMed backgrounds as well as a combined group of the same size consisting of equal number of all three (n = 7,706) were used to simulate phenotype pairs with known genetic correlation coefficients (ρ_k).

(A) Boxplots of distributions of estimated ρ_k from HEc (blue color) and GCTA-GREML (green) when we set ρ_k = 0.4. See Figure S1 and S2 for multiple simulation values.

(B) Comparison of runtime between HEc and GCTA-GREML for samples comprised of increasing number of individuals (note the log scale).

(C) Comparison of ρ_k estimates between HEc and GCTA-GREML for all pairs of eight phenotypes selected from the diverse TOPMed cohort either joint or stratified by self-reported background.

(D) Comparison of ρ_k estimates between HEc and LDSC for all pairs of eight phenotypes from the diverse TOPMed cohort either joint or stratified by self-reported background. See supplemental information and STAR Methods for description of the methods and GWAS used.

individuals within the background groups are on average more similar to each other in their genetic patterns compared with individuals in other background groups. However, the separation in the figure is far from perfect, as background groups capture self-identification, which does not precisely correlate with genetic patterns. Self-identified background also captures shared culture and systematic environmental exposures, which modify genetic effects, and which motivated us to use these grouping rather than ones based on genetic ancestry.

To simulate realistic data, we used kinship matrices from participants of the different self-reported backgrounds (Black, Hispanic/Latino, White) and a joint group of same size comprised of people from each group (n = 7,706, which is the also the size of the smallest subgroup from the TOPMed dataset). We evaluated the time required for each algorithm as a function of the number of individuals (Figure 1B). In all cases, HEc matched or exceeded the accuracy of GCTA while improving the speed up to 20-fold and beyond (Figure 1B).

HEc outperforms existing approaches on complex real-world data

We selected eight phenotypes from the TOPMed cohort (height, BMI, diastolic blood pressure, systolic blood pressure, total cholesterol, HDL cholesterol, LDL cholesterol, triglycerides) to compare the performance of HEc with existing algorithms (GCTA and LDSC; Figures 1C and 1D). We estimated the heritabilities of the eight phenotypes using GCTA and LDSC and estimated genetic correlations between all phenotype pairs using HEc and GCTA. Estimates were computed based on the combined TOPMed cohort (Figure 1, "All") as well as on the different background groups. While we see excellent agreement of HEc and GCTA estimates (Figure 1C; correlation above 0.96 for all groups), there is lower agreement between HEc and LDSC (Figure 1D; same for GCTA versus LDSC). This is especially evident for the self-reported Black group where the correlation is only 0.64 (Figure 1D). We also performed comparisons to a method called cov-LDSC, which was reported to account for population

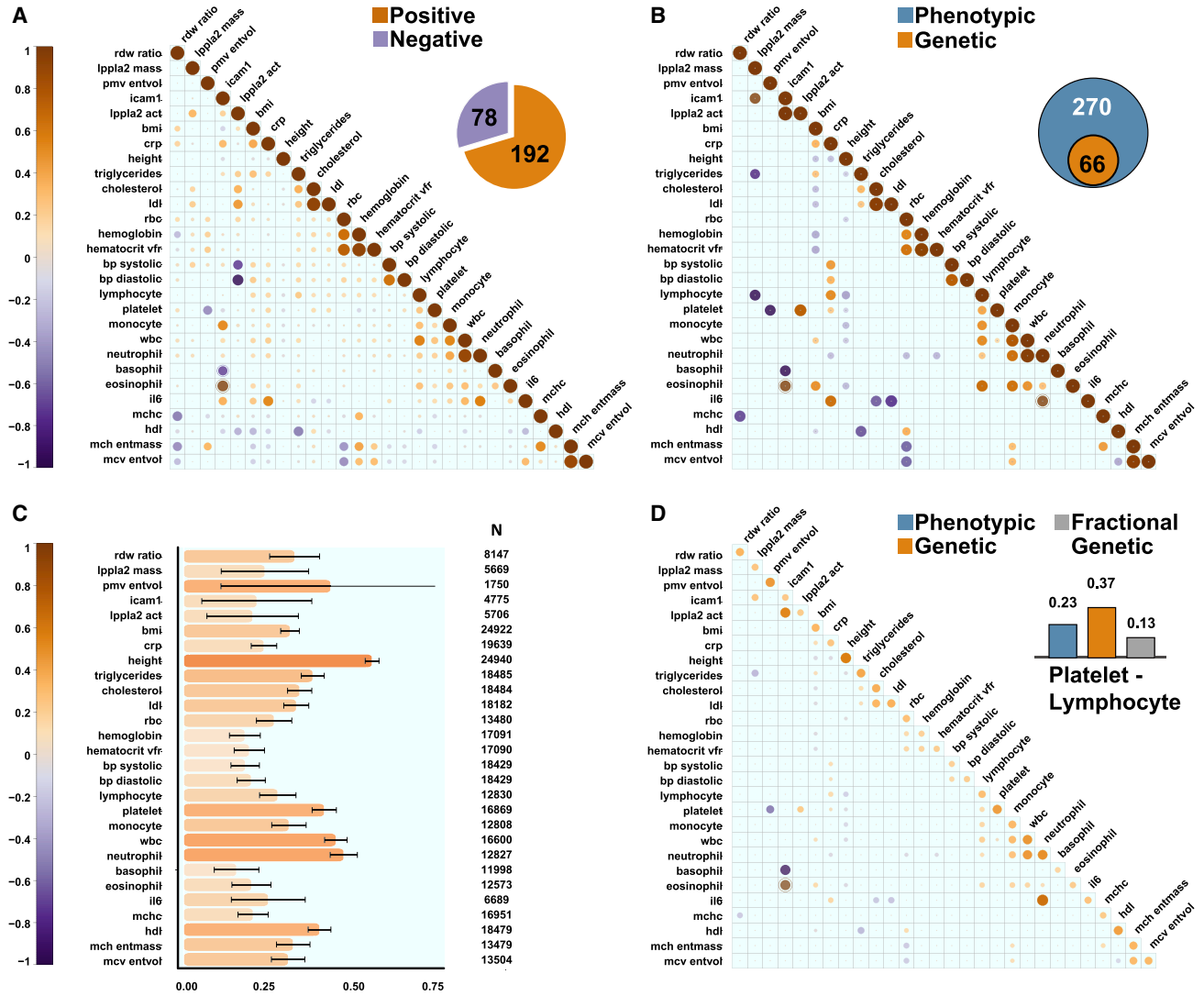


Figure 2. Genetic basis of observed phenotypic correlations between phenotypes in the combined TOPMed dataset

Correlation matrices where each column and row represent one of the 28 phenotypes in the TOPMed dataset (n = 33,959), and the intersection is the estimated correlation magnitude. Size and color of circle indicates the correlation strength: dark orange, positive; dark purple, negative correlation.

(A) Phenotypic correlations between the phenotypes. Inset: number of positive and negative correlated phenotype pairs with $p < 0.05$.

(B) Estimated genetic correlations (ρ_k) (shown only for phenotype pairs with $p < 0.05$ between the phenotypes). Inset: number of phenotype pairs with both phenotypic and genetic correlations with $p < 0.05$ in this dataset.

(C) Estimated heritabilities for the studied phenotypes.

(D) Fractional genetic correlations (ρ_{fk}) between the phenotypes (shown only for phenotype pairs with $p < 0.05$). Inset: example of phenotypic and genetic correlation where the absolute value of the genetic correlation is larger than that of the phenotypic correlation which complicates interpretability.

structure in admixed populations when estimating heritability[34] (Figure S5). Although the method indeed resulted in a lesser bias for heritability (Figure S5D), when extending the method for estimating genetic correlation there was little improvement (Figure S5B).

A compendium of genetic correlations and heritabilities in the multi-ethnic TOPMed dataset

We next expanded the number of analyzed phenotypes and calculated phenotypic and genetic correlations between 28 harmonized phenotypes in 33,959 TOPMed individuals. Results are provided in Figure 2 and Data S1. There were 378 phenotype pairs in the dataset. Of these, 270 (~71%) had phenotypic correlations R with $p < 0.05$ (Figure 2A), and 66 (18%) had genetic correlations ($p < 0.05$; Figure 2B, and inset).

Our results in the multi-ethnic dataset agree well with previous reports. For example, we estimated a genetic correlation of $\widehat{\rho}_k = 0.23$ (95% CI 0.07 $-$ 0.4) between body mass index (BMI) and triglycerides, while Bulik-Sullivan et al. estimated it

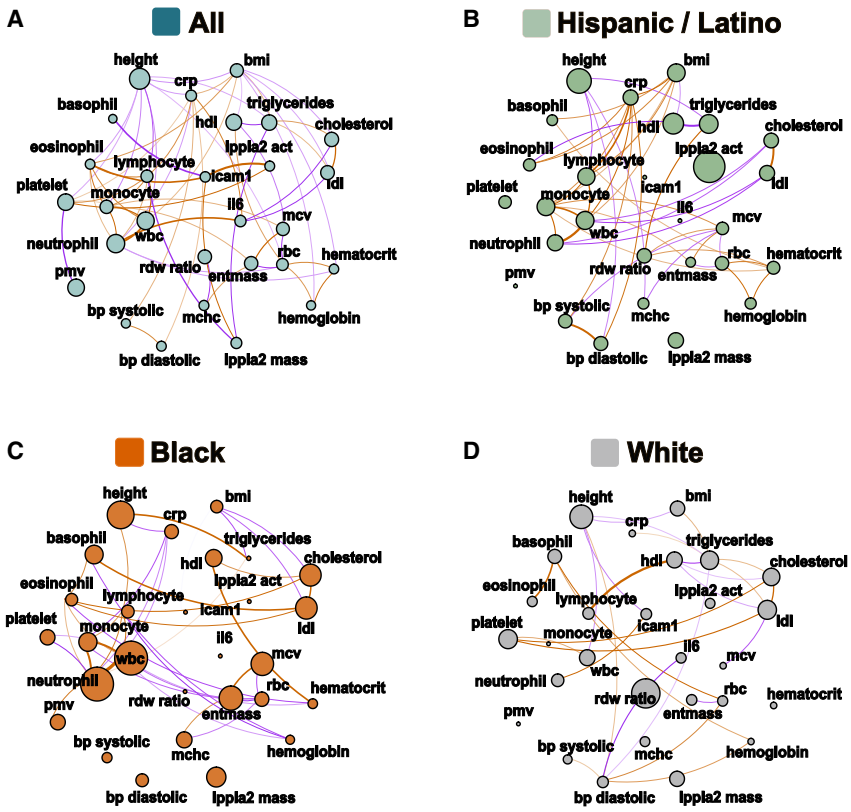


Figure 3. Many genetic correlations and heritabilities vary by self-reported background

Fractional genetic correlation plots where each phenotype is represented by a node and the fractional genetic 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 magnitude of the fractional genetic correlation and the color represents direction: orange represents positive and purple negative correlation.

(A) Fractional genetic correlations (ρ_{fk}) between the 28 phenotypes in the combined TOPMed dataset ($p < 0.05$, n = 33,959).

(B–D) Fractional genetic correlations (ρ_{fk}) between the 28 phenotypes in subsets of TOPMed individuals grouped by self-reported race/ethnic background: Hispanic/Latino (B) (marine, n = 8,762), Black (C) (orange, n = 8,054), and White (D) (gray, n = 17,143).

at 0.26[10] and Zhang et al. at 0.2[30] in European populations. We also identified potentially clinically relevant genetic correlations, including between white blood cell types and blood pressure, that are consistent with literature implicating a biological association between inflammation and hypertension[35,36] (Table S2 in Data S3).

We also estimated heritabilities for all the phenotypes in our combined dataset (Figure 2C; Table S3). The most highly heritable phenotypes are height ($\widehat{h^2} = 0.56$, SE = 0.02) as well as multiple blood cell measurements, such as neutrophil counts (Figure 2C; $\widehat{h^2} = 0.4$, SE = 0.04) and total white blood cell (WBC) counts (Figure 2C; $\widehat{h^2} = 0.45$, SE = 0.03) similar to previous estimates.[37,38]

Fractional genetic correlation quantifies the contribution of genetics to phenotypic correlations

We identified multiple instances where the genetic correlation coefficient is larger than the phenotypic correlation (Figure 2D, inset). For example, for lymphocytes and platelets the estimated phenotypic correlation is $\widehat{R} = 0.23$ but the estimated genetic correlation is larger: $\widehat{\rho_k} = 0.37\ (95\%\ \text{CI}\ 0.2 - 0.6)$. This emphasizes that the genetic correlation coefficient is not directly related to the phenotypic correlation.[39,40] To address this, we introduce the concept of "fractional genetic correlation" (ρ_{fk}), which we define as the fraction of the observed phenotypic correlation R explained by genetics in the decomposition $\widehat{R} = \widehat{\rho}_{fk} + \widehat{\rho}_{f\epsilon}$ (see STAR Methods Equations 34–36), where $\widehat{\rho}_{f\epsilon}$ is the estimated fractional residual correlation. In the example of lymphocytes and platelets, the estimated fractional genetic correlation is $\widehat{\rho_{fk}} = 0.13$, and we conclude that 56% of the observed phenotypic correlation is due to genetics ($\widehat{\rho_{fk}} = 0.13$ out of $\widehat{R} = 0.23$).

We estimated ρ_{fk} for all the pairs of phenotypes in our dataset (Figure 2D; Table S4 in Data S3). Of phenotype pairs with genetic correlation with $p < 0.05$, 58% had substantial $\widehat{\rho_{fk}}$, defined as $|\widehat{\rho_{fk}}| > 0.1$, corresponding to 19% of phenotypic correlations with $p < 0.05$. For most of the phenotype pairs, $\widehat{\rho_{fk}}$ is much lower than $\widehat{\rho_k}$ and is lower than the estimated R (by construction).

Genetic correlations and heritabilities differ between self-reported background groups

Out dataset includes 8,054 Black, 17,143 White, and 8,762 Hispanic/Latino participants (Table S2). We estimated heritability, phenotypic, and genetic correlations within these groups and compared them with the estimates of the combined group (Figures 3 and S7; Data S1).

While some phenotypes, such as HDL and eosinophil counts, have similar heritabilities across self-reported background groups (Figures 3 and S4), many other heritabilities vary by group. For example, CRP is similarly heritable in the Black and Hispanic/Latino backgrounds ($\widehat{h_2} = 0.34$, SE = 0.13 and 0.38, SE = 0.07, respectively), but much less so in White individuals ($\widehat{h_2} = 0.09$, SE = 0.1). In contrast, neutrophil counts are very heritable in Black individuals ($\widehat{h_2} > 0.99$, SE = 0.06), but are less so in White individuals ($\widehat{h_2} = 0.29$, SE = 0.21) and Hispanic/Latino individuals ($\widehat{h_2} = 0.41$, SE = 0.08).

Similarly, multiple phenotype pairs, such as systolic and diastolic blood pressure are genetically correlated across backgrounds (Figures 3 and S6). However, many other $\widehat{\rho}_k$ differ by background (Figure 3). For example, lymphocyte counts and height have estimated genetic correlation $\widehat{\rho_k} = 0.68$ (95% CI 0.1 – 0.99) in individuals from a Black background,

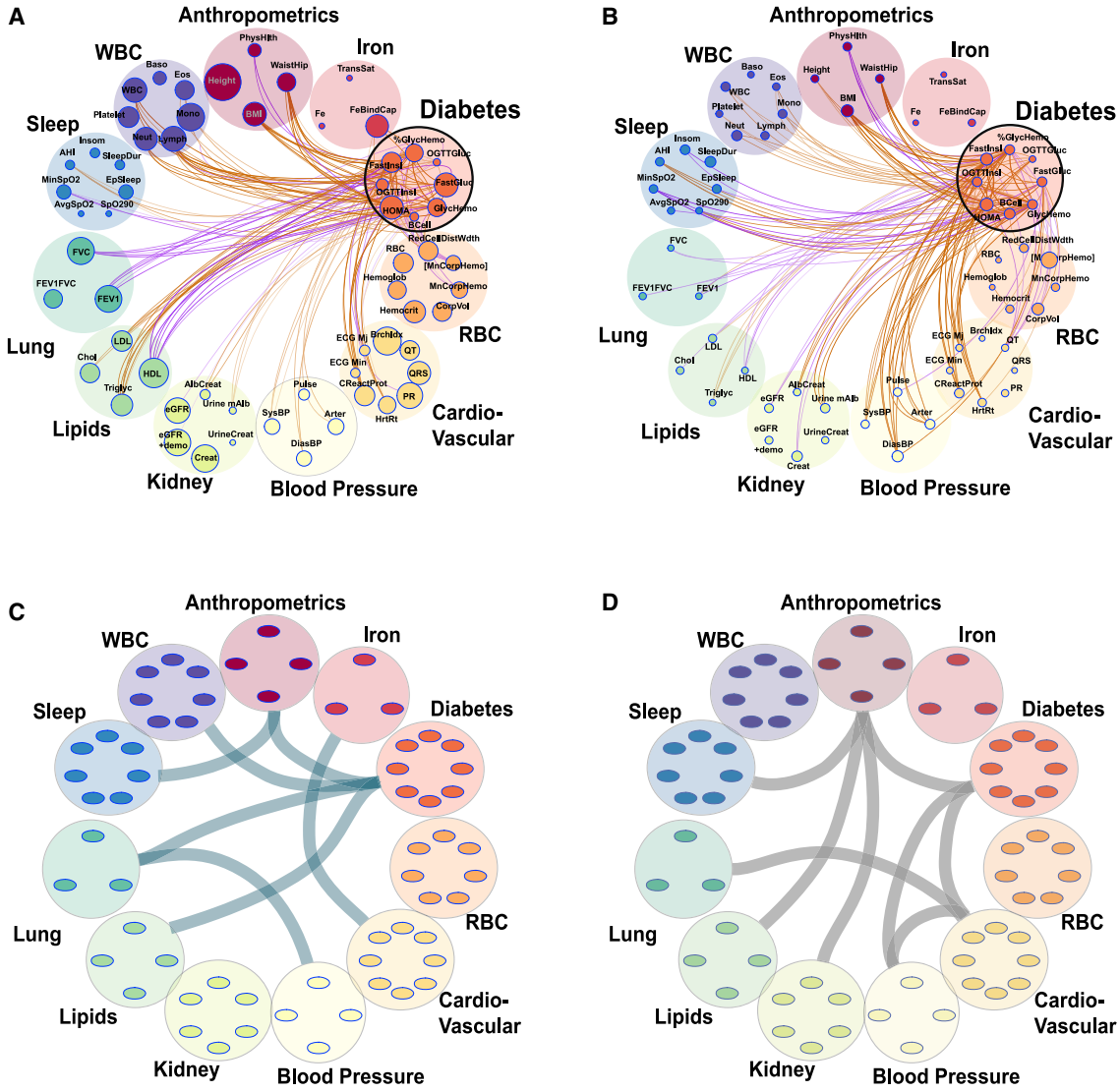


Figure 4. Genetics and shared household factors contribute to associations between phenotypes

(A and B) Correlation plots between phenotypes in the HCHS/SOL dataset (n = 12,565). 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 labeled 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 focused look on the diabetes phenotype domain showing fractional genetic (A) and fractional household (B) correlations with all other phenotypes (see Figure S7 for full graph).

(C and D) Same as (A and B) but connections represent enriched correlations between phenotypic domains. (C) Represents genetic correlations (teal) and (D) household correlations (gray).

but an opposite estimate $\widehat{\rho_k} = -0.34$ (95% CI -0.5 to -0.17) of genetic correlation in individuals from an Hispanic/Latino background ($p = 0.02$ for correlation coefficients difference), and genetic correlation with $p > 0.05$ in individuals from a White background $\widehat{\rho_k} = -0.22$ (95% CI -0.99 -0.7) (Figures 3 and S6; Data S1). Overall, out of the 378 examined phenotype pairs, in 211 (55%) we detect a difference in genetic correlation values ($p < 0.05$) for at least two of the background groups in the multi-ethnic TOPMed dataset.

Genetics and shared household differentially affect phenotypic correlations in Hispanic/Latino individuals

We next studied genetic correlations among a larger panel of 61 phenotypes in n = 12,565 self-reported Hispanic/Latino individuals from the HCHS/SOL.[33,41] The phenotypes represent 11 phenotypic domains: diabetes, cardiovascular disease, blood pressure, kidney function, lipids, lung function, sleep, anthropometrics, iron, RBC (red blood cell), and WBC (Figures 4 and S8; Table S1). HCHS/SOL also has information about household

sharing between participants, allowing for estimation of both genetic and environmental correlations between phenotypes.

We estimated ρ_k and ρ_{fk} in conjunction with the corresponding household correlation measures ρ_h and ρ_{fh} for all the 1,830 pairs of phenotypes (Figures 4 and S8). Out of the 1,830 phenotype pairs, 1,499 (or ~81%) have phenotypic correlations with $p < 0.05$ (Data S2). Of these, 427 (~28%) also have genetic correlations with $p < 0.05$ (Figure 3A) and 380 (~25%) have household correlations with $p < 0.05$. An interesting contrast between the genetic and household correlations can be seen for multiple phenotype pairs. For example, the diabetes domain (Figures 4A and 4B) has many household correlations with $p < 0.05$ with blood pressure domain phenotypes (but less so for genetic correlations) and has many genetic correlations with $p < 0.05$ with lung and lipid domain phenotypes (but not household correlations).

Domain enrichment analysis highlights associations between phenotypic domains

The genetic correlations are distributed non-uniformly with regard to the phenotypic domains. Domain enrichment analysis, in which we measured over-abundance of correlations with $p < 0.05$ between phenotypes within domain-pairs, showed a strong enrichment of the number of intra-group correlations for all the 11 phenotype domains (Data S2). Figures 4C and 4D visualize the estimated between-domain enrichment ($p < 0.05$). While some observed correlations, such as the ones between anthropometrics and diabetes or sleep, are driven both by genetics and shared household, many other domain-level correlations due to shared household do not mirror the genetic ones (Figure 4D). For example, shared household affects the correlations between diabetes and the blood-pressure and cardiovascular domains (but not genetic); however, the correlations between diabetes and lung and lipids domains are driven by genetics.

Heritabilities and genetic correlations differ across gender groups

The HCHS/SOL dataset of 12,565 participants has 5,175 males and 7,390 females (both self-identified genders and identified by sex chromosome checks using genetic data). Thus, we estimated heritabilities and genetic correlations stratified by gender. We identify a large number of differences between genders (Figures 5 and S9) in both genetic and environmental correlations. For example, the phenotype pair DiasBP (diastolic blood pressure) and FEV1FVC (forced expiratory volume to forced vital capacity ratio) has a negative estimated $\widehat{\rho_k} = -0.77(95\%\ \text{CI}\ -0.99\ -\ -0.18)$ in males but a positive genetic correlation $\widehat{\rho_k} = 0.43(95\%\ \text{CI}\ 0.13 - 0.84)$ in females. Similarly, the household correlation for lymphocyte count and height is $\widehat{\rho_h} = 0.68(95\%\ \text{CI}\ 0.02 - 0.99)$ in males but $\widehat{\rho_h} = -0.39(95\%\ \text{CI}\ -1\ \text{to}\ -0.03)$ in females (Figures 5 and S9; Data S2).

Overall, out of 363 phenotype pairs with either male or female genetic correlation with $p < 0.05$, there were 128 phenotype pairs (35%) in which the correlations were detected only in one gender groups (Figure S9C). Similarly, out of 349 phenotype pairs with either male or female household correlation with $p < 0.05$, there were 214 phenotype pairs (61%) in which the correlations were detected only in one gender group (Figure S9F).

These differences were also apparent at the domain level (Figures 5C and 5D). For example, while the correlations between blood pressure and diabetes domains are predominantly environmental in both gender groups (Figures 5C and 5D), the correlations between anthropometrics and sleep domains are enriched for genetic correlations only in the male group (Figure 5C and 5D). Figure S10 and Table S5 further provide similar results from gender-stratified genetic correlation analysis in the TOPMed White background group.

DISCUSSION

We sought to study how observed correlations between complex human phenotypes can vary by socially constructed groups, and their characterization using both genetics and shared environment. To achieve that, we developed and implemented a computationally efficient framework HEc to estimate genetic and environmental correlations between phenotypes. We validated our method in simulations guided by data from multiple TOPMed background groups. HEc showed similar accuracy to GCTA while being up to 20 times faster. The GCTA speed dropped significantly with increased number of people, and it took up to 55 h to calculate a single genetic correlation for the combined TOPMed dataset (~30k people), while it took ~2.5 h for HEc.

We also compared HEc with GCTA and LDSC genetic correlation estimates using real data on a number of phenotypes in different background groups. While HEc and GCTA results were highly concordant, HEc and LDSC results differed. This is expected, as LDSC uses summary statistics from GWAS and relies on a reference panel for computing LD, assuming that the LD matches that of the population used for GWAS. Here, we implemented LDSC on summary statistics from the pan-UKBB GWAS, a population of mostly European genetic ancestry, and our target TOPMed background groups for LD. Thus, differences in LDSC-estimated genetic correlations across backgrounds are only due to differences in LD, and differences between HEc (and GCTA) and LDSC for the same TOPMed group are due to mismatch in the underlying genetic associations with the phenotypes. Notably, there are currently no available high-powered GWAS for either the Black or Hispanic/Latino groups that could be used by LDSC in lieu of the pan-UKBB GWAS summary statistics for any of the phenotypes analyzed here. We also adapted and studied a more recent algorithm (cov-LDSC[34]) that was developed to estimate heritability in admixed populations using summary statistics. While it did improve estimated heritability across backgrounds, its estimated genetic correlations were similar to those from LSDC (Figure S3). Finally, while LDSC was indeed very fast when calculating genetic correlations (which takes only several minutes) after preparation of the LD reference, it is important to note that it required a very long preprocessing: over 5 days for calculating a group-specific LD panel for each background group. In this case, methods that use individual-level data, such as GCTA and HEc, are advantageous, especially given that preparation of the LD panel needs to be adapted for both the summary statistics and the genetic data used by matching on available genotypes and on effect alleles.

We next employed HEc to systematically interrogate factors that may affect the observed correlation between complex

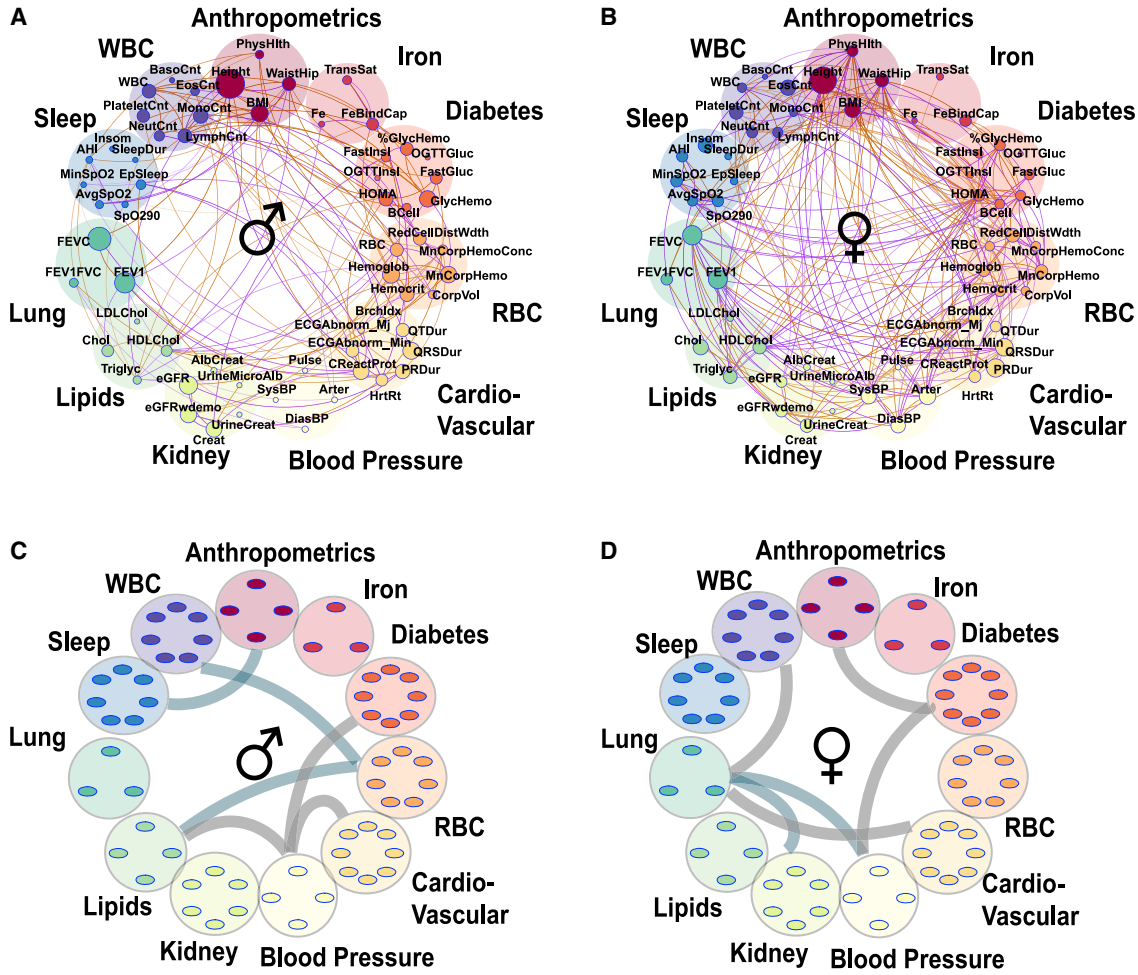


Figure 5. Gender differences in genetic correlations and heritabilities

(A and B) 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 fractional genetic correlations (ρ_{fk}) between the 61 phenotypes in the extended HCHS/SOL dataset ($p < 0.05$) in males (A) (n = 5,175) and females (B) (n = 7,390).

(C and D) Enriched ($p < 0.05$) correlations between the phenotypic domains. Teal represents genetic correlations and gray represents household correlation in males (C) and females (D).

human phenotypes, including shared genetics and shared environment. We assessed differences by self-reported background, capturing a combination of genetic patterns and sociocultural and environmental exposures that may modify genetic effects; and gender, capturing both biological sex effects and downstream sociocultural-related modifications; all affecting the underlying determinants of phenotypic distributions. We identified differences in some of the correlations and heritabilities across groups, agreeing with some previous reports of differences in phenotypic distributions and disease prevalence across race/ethnicities[11–13] (Figure S7). Overall, we identified 26 phenotype pairs that had different genetic correlations ($p < 0.05$) between the Hispanic/Latino background and other backgrounds, 41 such phenotype pairs for the White background and 39 phenotype pairs with genetic correlation $p < 0.05$ only in the Black background group (Figure S6A).

We estimated high heritabilities of some of the blood phenotypes specifically in the Black background group. The heritabilities of neutrophil counts ($\widehat{h_2} > 0.99$, SE = 0.07), WBC counts ($\widehat{h_2} > 0.99$, SE = 0.05), MCV_{EntVol} ($\widehat{h_2} = 0.64, \text{SE} = 0.14$); and monocytes ($\widehat{h_2} = 0.5 \pm 0.16$) in the Black background were much higher than in the White background, with neutrophil ($\widehat{h_2} = 0.3$, SE = 0.21), WBC ($\widehat{h_2} = 0.4$, SE = 0.1), MCV_{EntVol} ($\widehat{h_2} = 0.1$, SE = 0.25 and Monocytes ($\widehat{h_2} = 0$, SE = 0.28). Similarly high heritabilities were previously reported for Black individuals based on 236 African American pedigrees from the GeneSTAR study,[42] and are usually attributed to the Duffy antigen receptor for chemokine gene, which accounts for ~20% of total variation in the blood measures.[43,44] The differences in the distribution of the Duffy antigens in population were first reported in 1954, when it was found that the overwhelming

majority of people of African descent had the erythrocyte phenotype Fy(a-b-) which is rare in European genetic ancestries, and therefore in individuals of White background, who have predominately European ancestry. This region was shown to confer protection against malaria while inducing benign neutropenia. This genotype likely has high influence on estimated heritabilities and genetic correlations related to blood counts.

Traditionally in the GWAS era, such analyses have been performed for groups with clearly defined genetic ancestry, typically European. An important difference now is that we have been using whole-genome sequencing (WGS) data, with joint estimation of PCs and kinship matrix. Thus, our analysis does not suffer from limitations from focusing on sets of SNPs with differing LD patterns in different genetic ancestries and with different imputation qualities, which may affect analyses using genotyping array and imputed data. In principle, it is therefore appropriate to attempt to use these data to characterize the population in aggregate. However, we still see differences between estimates in the aggregated analysis and in the background-stratified analyses. These differences are likely driven by gene-environment interactions, where individuals from different backgrounds are exposed differently to modifiers of genetic effects. Such modifiers may include lifestyle factors such as smoking, sleep, nutrition,[45,46] and although less studied, structural determinants such as the built environment and access to health care. Another potential reason for differences between backgrounds are gene-gene interactions leading to different genetic effects in haplotypes of different genetic ancestries.[47] Although group differences reduce our ability to interpret the estimates in the combined group, we think that it was important to demonstrate the places where such differences are observed, as these are likely areas where there are stronger environmental effects on genetic effects and therefore policy or lifestyle interventions may be more useful to improve health.

We used a few measures of correlation throughout. The phenotypic correlation is the correlation between phenotypes without further modeling of contribution of specific factors. The genetic and household correlations measure the similarity between the contribution of genetic factors and household environment, respectively, to the phenotypes. Although genetic correlation could, by some statistical models, be traced to additive effects of a set of genetic variants, the household correlation was not developed under the same modeling assumption. Yet, they are estimated in the same manner, as different parameters corresponding to different matrices, once defined based on a statistical model that relies on measures of similarity of genetics (genetic relationship) and of household environment (household sharing) between individuals. Both of types of correlations are not restricted by the phenotypic correlation, where the phenotypic correlation may be very low while the genetic (or household) correlation can be large. This motivated the concept of "fractional genetic correlation coefficient" ρ_{fk} that we introduced, defined as the fraction of the observed phenotypic correlation R explained by genetics: $R = \widehat{\rho}_{fk} + \widehat{\rho}_{f\epsilon}$. The fractional genetic correlation ρ_{fk} is the genetic correlation normalized by the two traits' heritabilities (Equations 34–36) and is algorithm-agnostic, i.e., it does not depend on which algorithm is used to estimate heritabilities and genetic correlations. It addresses the limitations of the genetic correlations where (1) it is sometimes higher than the phenotypic correlation, and (2) it can have high estimate when the estimated heritabilities are low. The fractional genetic correlation allows for identification of phenotype pairs where genetics is a large contributor to the overall observed correlations. Fractional genetic correlations are typically lower than genetic correlations and are more highly correlated with the observed phenotypic correlations (e.g., correlations of 0.92 between the phenotype and fractional genetic correlations estimated in the TOPMed White background compared with correlation of 0.67 between the estimated phenotypic and genetic correlations; similarly 0.84 versus 0.52 between estimated phenotype and fractional versus genetic correlation in the Black background group). We believe that this measure is useful as it is interpretable with respect to its relationship to phenotypic correlation, and we report it for all correlations estimated in this work.

We next assess the contribution of genetics and environment to the correlations between phenotypes using the rich data collected in the HCHS/SOL cohort, including measurements from a wide range of phenotypes along with genetics and information on shared households. We estimated genetic and environmental (due to shared household) correlations between 61 phenotypes from diabetes, cardiovascular, blood pressure, kidney, lipids, lung, sleep, anthropometrics, iron, RBC, and WBC domains. While shared household does not capture all of the contribution of the environment to the correlations between phenotypes, it does contribute substantially to 22% of all the 1,830 phenotype pairs ($p < 0.05$, Figure 4; Figure S8). Moreover, the contribution of the shared household to phenotypic correlations varies by phenotype pairs. In some cases, estimated genetic and household correlation are in opposite directions. For example, for albumin-creatinine ratio and PR duration (an echocardiogram measure of heart rate) $\widehat{\rho_k} = 0.31\,(95\%\ \text{CI}\ 0.12 - 0.67)$ while $\widehat{\rho_H} = -0.65\,(95\%\ \text{CI}\ (-0.99\ \text{to}\ -0.17)$; and for major ECG abnormalities and BMI $\widehat{\rho_k} = -0.33(95\%\ \text{CI} -0.88\ \text{to} -0.07)$ and $\widehat{\rho_H} = 0.75(95\%\ \text{CI}\ 0.04 - 1)$).

We also performed domain-level enrichment analysis. We defined domains as sets of phenotypes that capture similar underlying "latent" phenotypes, with the limitations that groups of phenotypes assigned for the same domain may still capture complex underlying biology, i.e., are not measures of exactly the same latent phenotype (e.g., insomnia and mean oxygen saturation during sleep, while correlated in individuals with obstructive sleep apnea, may capture different pathophysiological disorders). While more study is needed, domain analysis should be less sensitive to individual variation in any particular phenotype. The results of this analysis highlight the domains and their mode of association (genetic and/or environmental) and, in the case of the correlations that are driven by the shared household, present a way to increase or disrupt the correlation via lifestyle changes. Interestingly, it seems that some domains have multiple phenotypes that are correlated with phenotypes in another domain predominantly via genetics or shared household. For example, the interactions between diabetes and blood pressure and cardiovascular domains are strongly influenced by shared household and therefore are a possible target for lifestyle

interventions. On the other hand, the diabetes domain has an abundance of genetically correlated phenotypes with the lung and anthropometrics domains.

Finally, we stratified participants from the Hispanic/Latino background by gender and analyzed genetic and household correlations. We found multiple differences between the genders such that there were 35% phenotype pairs with genetic correlations and 61% phenotype pairs with household correlations with $p < 0.05$ only in one gender group but not the other. For example, eosinophil counts versus PhysHealth (Aggregate Physical Health Scale) had a high $\widehat{\rho_k} = 0.98$ (95% CI. 0.19–1) in males but $\widehat{\rho_k} = -0.46 (95\%\ \text{CI}\ -0.99 \text{ to } -0.09)$ in females. Similarly, the estimated household correlation between lymphocyte counts and height was $\widehat{\rho_h} = 0.68$ (95% CI, 0.02–1) in males but was reversed and equal to $\widehat{\rho_k} = -0.39$ (95% CI, -1 to -0.03). Multiple correlations between phenotypic domains were also gender specific. Other recent studies considered gender differences in genetic determinants of phenotypes, focusing on UK Biobank participants of European ancestries.[48,49] Both studies computed genetic correlations between male and female genetic effects for the same phenotype, which we will denote here by ρ_{gender} to differentiate it from ρ_k, and found that often this genetic correlation is different than 1, indicating differences in genetic architecture between the gender groups. Zhu et al. further showed that gender differences are often due to "amplification effects," where genetic associations in one gender group, e.g., females, on average, are the same as the effects in males, multiplied by a constant, suggesting different regulations, for example, due to hormone levels, in males and females. In contrast, our analysis focused on gender differences in genetic correlation ρ_k between pairs of phenotypes. While mathematical modeling is needed to study whether the amplification model is consistent with downstream large differences in ρ_k between a pair of phenotypes in males compared with females, it is possible that differences in the systematic regulation of sets of genes between males and females will lead to such observed differences.

In interpreting such gender-stratified genetic and household correlations, we note that gender is a social construct that is related to sex, and drivers of some of the estimated quantities depend on complex interactions between biological determinants of sex and the gendered environment. Thus, both genetic and environmental correlations may differ between genders due to sociocultural differences between them (differences in environmental exposures may lead to differences in genetic effects via gene-environment interactions). Here, we were not able to assess specific sociocultural contributions related to gender roles to the estimated correlations, but we think that observed differences in household correlations are largely driven by them.

A specific strength of our study is the use of high-quality phenotypic and genotyping data from the diverse multi-ethnic TOPMed program. Furthermore, all participating studies are population-based cohort studies, reducing the likelihood of selection and other biases, which may arise in studies following selected populations, such as case-control studies. Other strengths are the investigation of a large panel of phenotypes, evaluation of both genetic and environmental correlations, stratification by both self-reported background and gender, and the domain-level enrichment analysis.

Nevertheless, this study also has several limitations. For example, the use of self-reported background rather than strata defined by genetic ancestry is also somewhat imprecise, as self-reported background may change over time.[50] Still, we chose to proceed with these groupings; first, to reflect on currently used groups in medical research, and second, because individuals of Hispanic/Latino background are highly admixed and there is no natural grouping that is based on genetic ancestry. We also note that socially constructed background groups may be meaningful due to differences in exposures across groups, which may translate to differences in the expression of genetic effects via gene-environment interactions. As the field of genetic medicine grapples with the use of genetic ancestry and social definition of race/ethnicity,[51] both potentially leading to the wrong and harmful reification of the biological basis of race,[52] it would be important to re-consider models for genetic correlation analyses. Finally, while our results are consistent with other studies[14,15] that demonstrated that genetic effects on specific traits vary by age, gender, and other environmental exposures, additional data and mathematical models are needed to untangle how specific factors influence measures of genetic and environmental correlations.

In summary, in this work we establish that multiple factors, including genetics, gender, sociocultural environment, and household environment, all shape the correlations between complex human phenotypes. We demonstrate how stratification by groups that encapsulate these factors, such as background groups capturing both genetic patterns and social race/ethnicity constructs, and gender groups capturing biological sex effects and gendered environment, uncovers differences in heritabilities and genetic correlations between them. We report thousands of genetic and environmental correlations between phenotypes. This work should lay the foundation for additional research in identifying personalized treatment and intervention strategies in understudied populations. Future work includes the application of approaches from the graph analysis field, such as Gaussian graphical models[53–55] to discover directionality and causality, utilizing genetically correlated phenotypes to improve polygenic risk prediction models,[56] and studying genetic correlations by categories of genetic variants to capture the contributions of rare variants.

LIMITATIONS OF THE STUDY

One limitation of this study is the somewhat limited sample sizes. Larger sample sizes, especially within stratified analyses, would enable stronger inferences. Another limitation is imperfect definitions of phenotypic domains, which may not accurately capture underlying pathophysiology. Next, although the use of population-based studies reduces the likelihood of bias in estimation of heritabilities and genetic correlations, additional biases could remain as all studies employed some preferential sampling, e.g., of specific age ranges, geographic regions, etc., and therefore none of the studies, separately or combined, accurately represents a random sample from a specific US population. Lastly, we note that because the p values

were derived via a bootstrap procedure (which is quite slow), and there were thousands of estimated genetic correlations in this paper, we are limited by the resulting p values with regard to FDR correction (which we provide for all calculated correlation values in Data S1–S4).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - The Trans-Omics for Precision Medicine (TOPMed) program
  - The Hispanic community health study/study of latinos
- METHOD DETAILS
  - Statistical model when genetic relatedness in the only modeled source of correlation
  - An estimator of genetic correlation between two phenotypes
  - Extension to multiple correlation matrices and generalization
  - Fractional genetic correlation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Simulation studies
  - Heritability and genetic/environment correlation estimation via HEc and GCTA-GREML
  - Heritability and genetic correlation estimation via LD-based methods
  - Domain-level enrichment analysis
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2022.100844.

ACKNOWLEDGMENTS

M.E., T.S., and S.R. were supported by National Heart, Lung, and Blood Institute grants R21HL145425 to T.S. and R35HL135818 to S.R. Acknowledgments for the TOPMed, CCDG, and parent cohorts participating in this study are provided in the supplemental information.

AUTHOR CONTRIBUTIONS

M.E. and T.S. conceived the study. M.E., M.O.G., and T.S. developed statistical methods for correlation analysis. M.E. performed analysis and prepared the figures and tables. M.E. and T.S. drafted the manuscript. M.O.G., C.I., H.C., P.S.d.V., H.X., A.W.M., X.G., N.F., B.M.P., S.S.R., J.I.R., D.M.L.-J., M.F., A.C., N.L.H.-C., R.S.V., R.H., R.C.K., A.C.M., and S.R. critically reviewed and approved the manuscript. C.I., H.X., B.M.P., S.S.R., J.I.R., D.M.L.-J., M.F., A.C., N.L.H.-C., R.S.V., R.C.K., A.C.M., and S.R. were involved in collection of data for studies that were used in this manuscript.

DECLARATION OF INTERESTS

B.M.P. serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. R.H. serves as a consultant for Invitae and is on the Scientific Advisory Board for Variant Bio. These roles are unrelated to the work in this paper.

Received: November 10, 2021
Revised: July 11, 2022
Accepted: November 9, 2022
Published: December 12, 2022

REFERENCES

1. Pickrell, J.K., Berisa, T., Liu, J.Z., Ségurel, L., Tung, J.Y., and Hinds, D.A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. Nat. Genet. *48*, 709–717.
2. Shi, H., Mancuso, N., Spendlove, S., and Pasaniuc, B. (2017). Local genetic correlation gives insights into the shared genetic architecture of complex traits. Am. J. Hum. Genet. *101*, 737–751.
3. van Rheenen, W., Peyrot, W.J., Schork, A.J., Lee, S.H., and Wray, N.R. (2019). Genetic correlations of polygenic disease traits: from theory to practice. Nat. Rev. Genet. *20*, 567–581.
4. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M., and Wray, N.R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics *28*, 2540–2542.
5. Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., et al. (2017). Overview of the BioBank Japan project: study design and profile. J. Epidemiol. *27*, S2–S8.
6. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
7. Zheng, J., Erzurumluoglu, A.M., Elsworth, B.L., Kemp, J.P., Howe, L., Haycock, P.C., Hemani, G., Tansey, K., Laurin, C., et al.; Early Genetics and Lifecourse Epidemiology EAGLE Eczema Consortium (2017). LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics *33*, 272–279.
8. Sakaue, S., Kanai, M., Karjalainen, J., Akiyama, M., Kurki, M., Matoba, N., Takahashi, A., Hirata, M., Kubo, M., Matsuda, K., et al. (2020). Trans-biobank analysis with 676, 000 individuals elucidates the association of polygenic risk scores of complex traits with human lifespan. Nat. Med. *26*, 542–548.
9. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. *50*, 390–400.
10. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.R., ReproGen Consortium; Psychiatric Genomics Consortium; Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3; and Duncan, L., et al. (2015). An atlas of genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.
11. Chen, M.H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., et al. (2020). Trans-ethnic and ancestry-specific blood-cell genetics in 746, 667 individuals from 5 global populations. Cell *182*, 1198–1213.e14.
12. Mogil, L.S., Andaleon, A., Badalamenti, A., Dickinson, S.P., Guo, X., Rotter, J.I., Johnson, W.C., Im, H.K., Liu, Y., and Wheeler, H.E. (2018). Genetic architecture of gene expression traits across diverse populations. PLoS Genet. *14*, e1007586.

13. Burt, V.L., Whelton, P., Roccella, E.J., Brown, C., Cutler, J.A., Higgins, M., Horan, M.J., and Labarthe, D. (1995). Prevalence of hypertension in the US adult population: results from the third national health and nutrition examination survey, 1988-1991. Hypertension *25*, 305–313.

14. Ge, T., Chen, C.Y., Neale, B.M., Sabuncu, M.R., and Smoller, J.W. (2017). Phenome-wide heritability analysis of the UK Biobank. PLoS Genet. *13*, e1006711.

15. Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J.K., and Przeworski, M. (2020). Variable prediction accuracy of polygenic scores within an ancestry group. Elife *9*, e48376. https://doi.org/10.7554/ELIFE.48376.

16. Rao, D.C., Sung, Y.J., Winkler, T.W., Schwander, K., Borecki, I., Adrienne Cupples, L., James Gauderman, W., Rice, K., Munroe, P.B., and Psaty, B.M. (2017). Multiancestry study of gene-lifestyle interactions for cardiovascular traits in 610 475 individuals from 124 cohorts: design and rationale. Circ Cardiovasc Genet. *10*, e001649. https://doi.org/10.1161/CIRCGENETICS.116.001649.

17. Walters, R.K., Polimanti, R., Johnson, E.C., McClintick, J.N., Adams, M.J., Adkins, A.E., Aliev, F., Bacanu, S.A., Batzler, A., Bertelsen, S., et al. (2018). Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. Nat. Neurosci. *21*, 1656–1669.

18. Brown, B.C., Asian Genetic Epidemiology Network Type 2 Diabetes Consortium; Ye, C.J., Price, A.L., and Zaitlen, N. (2016). Transethnic genetic-correlation estimates from summary statistics. Am. J. Hum. Genet. *99*, 76–88.

19. Galinsky, K.J., Reshef, Y.A., Finucane, H.K., Loh, P.R., Zaitlen, N., Patterson, N.J., Brown, B.C., and Price, A.L. (2019). Estimating cross-population genetic correlations of causal effect sizes. Genet. Epidemiol. *43*, 180–188.

20. Wientjes, Y.C.J., Bijma, P., Vandenplas, J., and Calus, M.P.L. (2017). Multi-population genomic relationships for estimating current genetic variances within and genetic correlations between populations. Genetics *207*, 503–515.

21. Schousboe, K., Willemsen, G., Kyvik, K.O., Mortensen, J., Boomsma, D.I., Cornes, B.K., Davis, C.J., Fagnani, C., Hjelmborg, J., Kaprio, J., et al. (2003). Sex differences in heritability of BMI: a comparative study of results from twin studies in eight countries. Twin Res. *6*, 409–421.

22. Weiss, L.A., Pan, L., Abney, M., and Ober, C. (2006). The sex-specific genetic architecture of quantitative traits in humans. Nat. Genet. *38*, 218–222.

23. Ober, C., Loisel, D.A., and Gilad, Y. (2008). Sex-specific genetic architecture of human disease. Nat. Rev. Genet. *9*, 911–922.

24. Maier, R.M., Zhu, Z., Lee, S.H., Trzaskowski, M., Ruderfer, D.M., Stahl, E.A., Ripke, S., Wray, N.R., Yang, J., Visscher, P.M., and Robinson, M.R. (2018). Improving genetic prediction by leveraging genetic correlations among human diseases and traits. Nat. Commun. *9*, 989–1017.

25. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.

26. Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2012). Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. *91*, 1011–1021.

27. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. *88*, 294–305.

28. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium; Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.

29. Ni, G., Moser, G., Schizophrenia Working Group of the Psychiatric Genomics Consortium; Wray, N.R., and Lee, S.H. (2018). Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood. Am. J. Hum. Genet. *102*, 1185–1194.

30. Zhang, Y., Cheng, Y., Jiang, W., Ye, Y., Lu, Q., and Zhao, H. (2021). Comparison of methods for estimating genetic correlation between complex traits using GWAS summary statistics. Briefings Bioinf. *22*, bbaa442.

31. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., Taliun, S.A.G., Corvelo, A., Gogarten, S.M., Kang, H.M., et al. (2021). Sequencing of 53, 831 diverse genomes from the NHLBI TOPMed Program. Nature *590*, 290–299.

32. Stilp, A.M., Emery, L.S., Broome, J.G., Buth, E.J., Khan, A.T., Laurie, C.A., Wang, F.F., Wong, Q., Chen, D., D'Augustine, C.M., et al. (2021). A system for phenotype harmonization in the NHLBI trans-omics for precision medicine (TOPMed) program. Am. J. Epidemiol. *190*, 1977–1992. https://doi.org/10.1093/aje/kwab115.

33. Sorlie, P.D., Avilés-Santa, L.M., Wassertheil-Smoller, S., Kaplan, R.C., Daviglus, M.L., Giachello, A.L., Schneiderman, N., Raij, L., Talavera, G., Allison, M., et al. (2010). Design and implementation of the hispanic community health study/study of Latinos. Ann. Epidemiol. *20*, 629–641.

34. Luo, Y., Li, X., Wang, X., Gazal, S., Mercader, J.M., 23 and Me Research Team; SIGMA Type 2 Diabetes Consortium; Neale, B.M., Florez, J.C., Auton, A., et al. (2021). Estimating heritability and its enrichment in tissue-specific gene sets in admixed populations. Hum. Mol. Genet. *30*, 1521–1534.

35. Siedlinski, M., Jozefczuk, E., Xu, X., Teumer, A., Evangelou, E., Schnabel, R.B., Welsh, P., Maffia, P., Erdmann, J., Tomaszewski, M., et al. (2020). White blood cells and blood pressure: a mendelian randomization study. Circulation *141*, 1307–1317.

36. Schillaci, G., Pirro, M., Pucci, G., Ronti, T., Vaudo, G., Mannarino, M.R., Porcellati, C., and Mannarino, E. (2007). Prognostic value of elevated white blood cell count in hypertension. Am. J. Hypertens. *20*, 364–369.

37. Wainschtein, P., Jain, D.P., Yengo, L., Cupples, L.A., Shadyab, A.H., McKnight, B., Shoemaker, B.M., Mitchell, B.D., Psaty, B.M., Kooperberg, C., and Liu, C.T. (2019). Recovery of trait heritability from whole genome sequence data. Preprint at bioRxiv. https://doi.org/10.1101/588020.

38. Reiner, A.P., Lettre, G., Nalls, M.A., Ganesh, S.K., Mathias, R., Austin, M.A., Dean, E., Arepalli, S., Britton, A., Chen, Z., et al. (2011). Genome-Wide association study of white blood cell count in 16, 388 african americans: the continental Origins and Genetic Epidemiology network (COGENT). PLoS Genet. *7*, e1002108.

39. Sodini, S.M., Kemper, K.E., Wray, N.R., and Trzaskowski, M. (2018). Comparison of genotypic and phenotypic correlations: cheverud's conjecture in humans. Genetics *209*, 941–948.

40. Searle, S.R. (1961). Phenotypic, genetic and environmental correlations. Biometrics *17*, 474.

41. LaVange, L.M., Kalsbeek, W.D., Sorlie, P.D., Avilés-Santa, L.M., Kaplan, R.C., Barnhart, J., Liu, K., Giachello, A., Lee, D.J., Ryan, J., et al. (2010). Sample design and cohort selection in the hispanic community health study/study of Latinos. Ann. Epidemiol. *20*, 642–649.

42. Reiner, A.P., Lettre, G., Nalls, M.A., Ganesh, S.K., Mathias, R., Austin, M.A., Dean, E., Arepalli, S., Britton, A., Chen, Z., and Couper, D. (2011). Genome-wide association study of white blood cell count in 16,388 African Americans: the continental origins and genetic epidemiology network (COGENT). PLoS Genet. *7*, e1002108. https://doi.org/10.1371/JOURNAL.PGEN.1002108.

43. Nalls, M.A., Couper, D.J., Tanaka, T., van Rooij, F.J.A., Chen, M.H., Smith, A.V., Toniolo, D., Zakai, N.A., Yang, Q., Greinacher, A., et al. (2011). Multiple loci are associated with white blood cell phenotypes. PLoS Genet. *7*, e1002113.

44. Reich, D., Nalls, M.A., Kao, W.H.L., Akylbekova, E.L., Tandon, A., Patterson, N., Mullikin, J., Hsueh, W.C., Cheng, C.Y., Coresh, J., et al. (2009). Reduced neutrophil count in people of african descent is due to a regulatory variant in the duffy antigen receptor for chemokines gene. PLoS Genet. *5*, e1000360.

45. Wang, H., Noordam, R., Cade, B.E., Schwander, K., Winkler, T.W., Lee, J., Sung, Y.J., Bentley, A.R., Manning, A.K., Aschard, H., et al. (2021). Multi-ancestry genome-wide gene–sleep interactions identify novel loci for blood pressure. Mol. Psychiatr. *26*, 6293–6304.

46. Laville, V., Majarian, T., Sung, Y.J., Schwander, K., Feitosa, M.F., Chasman, D.I., Bentley, A.R., Rotimi, C.N., Cupples, L.A., de Vries, P.S., et al. (2022). Gene-lifestyle interactions in the genomics of human complex traits. Eur. J. Hum. Genet. *30*, 730–739.

47. Patel, R.A., Musharoff, S.A., Spence, J.P., Pimentel, H., Tcheandjieu, C., Mostafavi, H., Sinnott-Armstrong, N., Clarke, S.L., Smith, C.J., et al.; V.A. Million Veteran Program (2022). Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. Am. J. Hum. Genet. *109*, 1286–1297.

48. Bernabeu, E., Canela-Xandri, O., Rawlik, K., Talenti, A., Prendergast, J., and Tenesa, A. (2021). Sex differences in genetic architecture in the UK Biobank. Nat. Genet. *53*, 1283–1289.

49. Zhu, C., Ming, M.J., Cole, J.M., Kirkpatrick, M., and Harpak, A. (2022). Amplification is the primary mode of gene-by-sex interaction in complex human traits. Preprint at bioRxiv. https://doi.org/10.1101/2022.05.06.490973.

50. Hitlin, S., Scott Brown, J., and Elder, G.H. (2006). Racial self-categorization in adolescence: multiracial development and social pathways. Child Dev. *77*, 1298–1308.

51. Hernandez, L.M., and Blazer, D.G.; Institute of Medicine (US) Committee on Assessing Interactions Among Social B and GF in H (2006). Sex/Gender, Race/Ethnicity, and Health (National Academies Press).

52. Lewis, A.C.F., Molina, S.J., Appelbaum, P.S., Dauda, B., Di Rienzo, A., Fuentes, A., Fullerton, S.M., Garrison, N.A., Ghosh, N., Hammonds, E.M., et al. (2022). Getting genetic ancestry right for science and society. Science *376*, 250–252.

53. Talluri, R., and Shete, S. (2014). Gaussian graphical models for phenotypes using pedigree data and exploratory analysis using networks with genetic and nongenetic factors based on Genetic Analysis Workshop 18 data. In BMC Proc (BioMed Central Ltd.), p. S99.

54. Zhao, H., and Duan, Z.H. (2019). Cancer genetic network inference using Gaussian graphical models. Bioinf. Biol. Insights *13*, 1177932219839402. https://doi.org/10.1177/1177932219839402.

55. Glymour, C., Zhang, K., and Spirtes, P. (2019). Review of causal discovery methods based on graphical models. Front. Genet. *10*, 524.

56. Torkamani, A., Wineinger, N.E., and Topol, E.J. (2018). The personal and clinical utility of polygenic risk scores. Nat. Rev. Genet. *19*, 581–590.

57. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76–82.

58. Csardi, G., and Nepusz, T. (2006). The Igraph Software Package for Complex Network Research (InterJournal Complex Sy), p. 1695.

59. Epskamp, S., Cramer, A.O.J., Waldorp, L.J., Schmittmann, V.D., and Borsboom, D. (2012). Qgraph: network visualizations of relationships in psychometric data. J. Stat. Software *48*, 1–18.

60. Wei, T., and Simko, V. (2021). R Package "Corrplot": Visualization of a Correlation Matrix (corrplot). https://cran.r-project.org/web/packages/corrplot/index.html.

61. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.

62. Conomos, M.P., Miller, M.B., and Thornton, T.A. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet. Epidemiol. *39*, 276–293.

63. Conomos, M.P., Reiner, A.P., Weir, B.S., and Thornton, T.A. (2016). Model-free estimation of recent genetic relatedness. Am. J. Hum. Genet. *98*, 127–148.

64. 1000 Genomes Project Consortium; Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1, 092 human genomes. Nature *491*, 56–65.

65. Conomos, M.P., Laurie, C.A., Stilp, A.M., Gogarten, S.M., McHugh, C.P., Nelson, S.C., Sofer, T., Fernández-Rhodes, L., Justice, A.E., Graff, M., et al. (2016). Genetic diversity and association studies in US hispanic/latino populations: applications in the hispanic community health study/study of Latinos. Am. J. Hum. Genet. *98*, 165–184.

66. Sofer, T. (2017). Confidence intervals for heritability via Haseman-Elston regression. Stat. Appl. Genet. Mol. Biol. *16*, 259–273.

67. Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika *10*, 507.

68. Bishara, A.J., and Hittner, J.B. (2017). Confidence intervals for correlations when data are not normal. Behav. Res. Methods *49*, 294–309.

69. Kunsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. Ann. Stat. *17*, 1217–1241.

70. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag, New York).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Summary statistics, PAN-UK BioBank	Pan-UKB team. https://pan.ukbb.broadinstitute.org. 2020	https://pan.ukbb.broadinstitute.org/
TOPMed data, by study	Taliun et al.[31]	Amish: phs000956, ARIC: phs001211, CARDIA: phs001612, CHS: phs001368, FHS: phs000974, HCHS_SOL: phs001395, JHS: phs000964, MESA: phs001416
Software and algorithms		
All original computer code	This paper	https://github.com/tamartsi/HE_Genetic_Correlation
Principal Components, kinship matrices, unrelated individuals	This paper	TOPMed DCC pipeline https://github.com/UW-GAC/analysis_pipeline
GCTA	Yang et al.[57]	https://yanglab.westlake.edu.cn/software/gcta/#Overview
R package: igraph	Csardi et al.[58]	https://igraph.org/r/
R package: qgraph	Epskamp et al.[59]	https://cran.r-project.org/web/packages/qgraph/index.html
R package: corrplot	Friendly[60]	https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html
ldsc package	Bulik-Sullivan[28]	https://github.com/bulik/ldsc
cov-LDSC	Luo[34]	https://github.com/immunogenomics/cov-ldsc

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to the lead contact, Dr. Tamar Sofer (tsofer@bwh.harvard.edu).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

- The data is publicly available upon request from the TOPMed consortium (https://topmed.nhlbi.nih.gov/), and all original code used in this work is freely available at https://github.com/tamartsi/HE_Genetic_Correlation.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The Trans-Omics for Precision Medicine (TOPMed) program

We used harmonized phenotype data from eight cohort studies participating in TOPMed (9) Freeze 8 [https://www.nhlbiwgs.org/topmed-whole-genome-sequencing-methods-freeze-8], which included 33,959 genotyped individuals from the Amish Study (n = 1,105), JHS (n = 2,807), FHS (n = 3,658), HCHS/SOL (n = 7,693), ARIC (n = 7,479), CHS (n = 3,482), MESA (n = 4,665), and CARDIA (n = 3,070) with available self-reported race/ethnic identification, which we refer to as "background". Descriptions of each of these studies are provided in the Supplemental information. This dataset included 8,054 Black participants, 17,143 White participants and 8,762 participants of Hispanic/Latino background. All participants provided informed consent and the study was approved by IRBs in each of the participating institutions. For TOPMed WGS data acquisition and QC report see ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?id=phd006969.1. The phenotype harmonization was performed by TOPMed Data Coordinating Center (DCC) as

described in.[32] The phenotype names and description, exclusion criteria and transformations are described in Table S1. Phenotypes' characteristics across backgrounds are reported in Table S2 (Data S3). All analyses were adjusted for age, gender, study, and reported race/ethnicity as well as 11 first principal components (PCs) to adjust for population structure. The PCs, kinship matrices, and unrelated individual pools were computed by TOPMed DCC via a robust pipeline [https://github.com/UW-GAC/analysis_pipeline] via a combination of KING,[61] PC-AiR,[62] and PC-Relate.[63]

The Hispanic community health study/study of latinos

The HCHS/SOL is a community-based cohort study of Hispanic individuals from four field centers across the US[33,41] with almost 13,000 genotyped participants. A two-stage sampling scheme for participant selection was employed, with sampled community block units followed by households. Correlation matrices to model environmental variance due to households and community block units were generated so that the i, j entry of a given matrix was set to 1 if the i and j individuals live in the same household (or community block unit), and 0 otherwise. This study was approved by the institutional review boards at each field center, where all participants gave written informed consent. Genotyping and quality control for HCHS/SOL have been described in detail elsewhere.[62] In brief, DNA extracted from blood was genotyped on the HCHS Custom 15,041,502 array (Illumina Omni2.5M + custom content). Genotyping and downstream quality-control procedures yielded 2 232 944 genetic variants for genotyped HCHS/SOL participants. These were used for genotype imputation with the 1000 Genomes Project phase 3 reference panel[64] as previously described.[65] Variants with at least two copies of the minor allele and present in any of the four 1000 Genomes continental panels were imputed yielding about 50 million imputed variants prior to quality filtering. A subset of 7,693 individuals from HCHS/SOL participated in TOPMed and was used in the first part of this study (with the whole dataset used in the latter parts), with additional phenotypes that were not harmonized across other TOPMed studies. HCHS/SOL phenotype names and description are described in Table S3. The number of participants with non-missing information per phenotype as well as means and standard deviations per phenotype per gender are reported in Table S4 (Data S3). We estimated genetic correlations between phenotypes from the following domains: anthropometric, blood pressure, lipids, blood cell counts, and inflammation markers. All analyses were adjusted for age, gender, sampling weights and 11 first principal components to account for population structure.

METHOD DETAILS

Statistical model when genetic relatedness in the only modeled source of correlation

Consider the linear model

$$y_i = x_i^T\beta + g_{i,1}\alpha_1 + \ldots + g_{i,d}\alpha_d + \epsilon_i^0,\ i = 1, \ldots, n \qquad \text{Model 1}$$

in which the quantitative outcome y_i is modeled by a regression on covariates x_i and the additive effects of d genetic variants $g_1, \ldots, g_d$; and $\epsilon_i^0 \sim \mathcal{N}(0, \sigma_e^2)$ are normally distributed errors across n participants. Assuming that the genetic variants are independent random variables, each centered and scaled to have mean 0 and variance 1, the mean and variance of y_i are

$$E[y_i] = x_i^T\beta \qquad \text{(Equation 1)}$$

$$var(y_i) = var(g_1) \times \alpha_1^2 + \ldots + var(g_d) \times \alpha_d^2 + \sigma_e^2 =$$

$$= \sum_{j=1}^{d} \alpha_j^2 + \sigma_e^2 = \sigma_k^2 + \sigma_e^2 \qquad \text{(Equation 2)}$$

Here, σ_k^2, σ_e^2 are the genetic and error variance components. Accordingly, narrow-sense heritability, which is the proportion of trait variance that is due to additive genetic factors is:

$$\hat{h}^2 = \frac{\sigma_k^2}{\sigma_k^2 + \sigma_e^2} \qquad \text{(Equation 3)}$$

To model genetic correlation, we extend Model 1 into a two-trait model. For person i:

$$y_{i,1} = x_i^T\beta_1 + g_{i,1}\alpha_{1,1} + \ldots + g_{i,d}\alpha_{1,d} + \epsilon_{i,1}^0 \qquad \text{(Equation 4)}$$

$$y_{i,2} = x_i^T\beta_2 + g_{i,1}\alpha_{2,1} + \ldots + g_{i,d}\alpha_{2,d} + \epsilon_{i,2}^0 \qquad \text{(Equation 5)}$$

with error terms $\epsilon_{i,1}^0, \epsilon_{i,2}^0$ satisfying $\epsilon_{i,l}^0 \sim \mathcal{N}(0, \sigma_{\epsilon,l}^2)$, $cor(\epsilon_{i,1}^0, \epsilon_{i,2}^0) = \rho_\epsilon$, $\epsilon_{i,1}^0 \perp \epsilon_{j,2}^0$ for $k \in \{1,2\}$, $i = 1,..,n$.

Thus, the errors of the same person may be correlated for the two traits, but for different traits the error of person i is independent of the error of person *j*. Consider the covariance between the two traits, again while making the simplifying assumption of independence between genetic variants:

$$cov(y_{i,1},\ y_{i,2}) \ = \ cov\left(x_i^T\beta_1 \ + g_{i,1}\alpha_{1,1} \ + \ \ldots \ + \ g_{i,d}\alpha_{1,d} \ + \ \epsilon_{i,1}^0,\ x_i^T\beta_2 \ + g_{i,1}\alpha_{2,1} \ + \ \ldots \ + \ g_{i,d}\alpha_{2,d} \ + \ \epsilon_{i,2}^0\right)$$

$$= cov\left(g_{i,1}\alpha_{1,1} \ + \ \ldots \ + \ g_{i,d}\alpha_{1,d} \ + \ \epsilon_{i,1}^0,\ g_{i,1}\alpha_{2,1} \ + \ \ldots \ + \ g_{i,d}\alpha_{2,d} \ + \ \epsilon_{i,2}^0\right)$$

$$= \sum_{k\ =\ 1}^{d}\alpha_{1,k}\alpha_{2,k} + \ \sigma_1\sigma_2\rho_e \ = \ \sigma_{g_1}\sigma_{g_2}\rho_k + \ \sigma_{\epsilon,1}\sigma_{\epsilon,2}\rho_e \qquad \text{(Equation 6)}$$

where ρ_k is the correlation between the genetic effect of the k-th variant on the two outcomes, and ρ_e is the correlation between the residual errors of the two outcomes. Note that the transition to using ρ_k at the final step treats the vectors of causal genetic effects $\alpha_1 = (\alpha_{1,1}\ldots\alpha_{1,d})^T, \alpha_2 = (\alpha_{2,1}\ldots\alpha_{2,d})^T$ as random variables with mean 0, i.e., $E[\alpha_{i,j}] \ = \ 0$, $var(\alpha_{i,j}) \ = \ \sigma_{g,j}^2$, $i \in \{1,2\}$, $j \ = \ 1, \ldots, d$.

Noting that for individuals i, l, $cov(g_{i,j},\ g_{l,j}) \ = \ k_{i,l}$, the probability of the two individuals sharing the same allele identically-by-decent,[61,63] an equivalent formulation supposes that the genetic effects can be modeled via K, the $n \times n$ kinship matrix, tabulating the measure of genetic relationship between the i and j participants in its i,j entry. Consider the vector form of the model for the l outcome with correlated errors of trait l = 1,2:

$$y_l = X\beta_l + \ \epsilon_l, \quad l = 1,2 \qquad \text{Model 2}$$

$$cov(\epsilon_l) \ = \ \sigma_{\epsilon,l}^2\boldsymbol{I}_{n\times n} + \sigma_{k,l}^2\boldsymbol{K}$$

$$cov(\epsilon_1, \epsilon_2) \ = \ \sigma_{\epsilon,1}\sigma_{\epsilon,2}\rho_\epsilon\boldsymbol{I}_{n\times n} + \sigma_{k,1}\sigma_{k,2}\rho_k\boldsymbol{K}$$

Now the genetic correlation can be estimated using mixed model with two traits. However, this is computationally demanding, especially for very large datasets. Recently,[66] discussed the Haseman-Elston regression for variance components estimation, and demonstrated that the genetic variance components estimator corresponding to the kinship matrix, if it is independent of all other correlation matrices in a model with potentially multiple sources of correlation (which holds here, because we only have a single correlation matrix, the kinship matrix) are given by $\widehat{\sigma}_{\epsilon,l}^2 \ = \ \frac{\widehat{\epsilon}_l^T\boldsymbol{K}^-\widehat{\epsilon}_l}{tr(\boldsymbol{K}^-\boldsymbol{K}^-)}$, where $\boldsymbol{K}^-$ is the kinship matrix with all diagonal values set to zero.

An estimator of genetic correlation between two phenotypes

We extend the Haseman-Elston approach for modeling the genetic correlations between two phenotypes. For the errors of persons *i* and *j*, and phenotypes 1 and 2, under *Model 2* we get:

$$E[\epsilon_{1,j}, \epsilon_{2,i}] \ = \ \sigma_{\epsilon,1}\sigma_{\epsilon,2}\rho_\epsilon + \sigma_{k,1}\sigma_{k,2}\rho_k \qquad \text{(Equation 7)}$$

$$E[\epsilon_{1,j}, \epsilon_{2,j}] \ = \ \sigma_{k,1}\sigma_{k,2}\rho_k k_{i,j} \qquad \text{(Equation 8)}$$

Suppose for now that $\sigma_{k,1} = \left|\sqrt{\sigma_{k,1}^2}\right|$ and $\sigma_{k,2} = \left|\sqrt{\sigma_{k,2}^2}\right|$ are known. For estimating the genetic correlation between the two phenotypes, we take all pairs $\widehat{\epsilon}_{1,i}\widehat{\epsilon}_{2,j}$ of residuals (estimating the error terms) after regression on mean-model covariates for $i \neq j$, and regress them against the "covariate" $\sigma_{k,1}\sigma_{k,2}k_{i,j}$. From properties of linear regression, we get:

$$\widehat{\rho}_k \ = \ \frac{\sum_{i\ =\ 1}^n\sigma_{k,1}\sigma_{k,2}k_{i,j}\widehat{\epsilon}_{1,i}\widehat{\epsilon}_{2,j}}{\sum_{i\ =\ 1}^n\sigma_{k,1}^2\sigma_{k,2}^2k_{i,j}^2} = \frac{\sum_{i\ =\ 1}^n k_{i,j}\widehat{\epsilon}_{1,i}\widehat{\epsilon}_{2,j}}{\sigma_{k,1}\sigma_{k,2}\sum_{i\ =\ 1}^n k_{i,j}^2} = \frac{\widehat{\epsilon}_1^T\boldsymbol{K}^-\widehat{\epsilon}_2}{\sigma_{k,1}\sigma_{k,2}\ tr(\boldsymbol{K}^-\boldsymbol{K}^-)} \qquad \text{(Equation 9)}$$

Now we can plug-in the estimators of $\sigma_{k,1}, \sigma_{k,2}$ to get:

$$\widehat{\rho}_k \ = \ \frac{\widehat{\epsilon}_1^T\boldsymbol{K}^-\widehat{\epsilon}_2}{tr\left(\boldsymbol{K}^-\boldsymbol{K}^-\right)\sqrt{\left(\frac{\widehat{\epsilon}_1^T\boldsymbol{K}^-\widehat{\epsilon}_1}{tr(\boldsymbol{K}^-\boldsymbol{K}^-)}\right)\left(\frac{\widehat{\epsilon}_2^T\boldsymbol{K}^-\widehat{\epsilon}_2}{tr(\boldsymbol{K}^-\boldsymbol{K}^-)}\right)}} = \frac{\widehat{\epsilon}_1^T\boldsymbol{K}^-\widehat{\epsilon}_2}{\sqrt{(\widehat{\epsilon}_1^T\boldsymbol{K}^-\widehat{\epsilon}_1)(\widehat{\epsilon}_2^T\boldsymbol{K}^-\widehat{\epsilon}_2)}} \qquad \text{(Equation 10)}$$

This estimator resembles that of the Pearson correlation parameter between the variables $\widehat{\epsilon}_1$ and $\widehat{\epsilon}_2$, as can be seen if one replaces the matrix $\boldsymbol{K}^{-}$ by the identify matrix $\boldsymbol{I}$. Interestingly, this estimator does not involve the unknown variance parameters. It does include the estimated kinship parameters, which are treated as fixed.

Extension to multiple correlation matrices and generalization

We can use multiple relatedness matrices $(A, \ldots, K)$ with elements $(a_{i,j}..k_{i,j})$ indicating the measure of relatedness between the *i-th* and *j-th* participants in its *i,j* entry to model the variance. We can then estimate the variance components obtained from expressions of the form

$$E[\epsilon_i, \epsilon_j] = \sigma^2_{\epsilon} \times 1_{(i=j)} + \sigma^2_a a_{i,j} + \ldots + \sigma^2_k k_{i,j} \quad \text{(Equation 11)}$$

(corresponding to *Model 2* above) via a residual regression, i.e. by taking the vector all pairs of residuals $\widehat{\epsilon}_{1,i}\widehat{\epsilon}_{2,j}$ for all i,j. The HE design matrix, now re-defined (compared to[66]) to include rows corresponding to $\widehat{\epsilon}_{1,j}\widehat{\epsilon}_{2,j}$ with $i = j$, is given by:

$$X_{\sigma} = \begin{pmatrix} 1 & a_{1,1} & \ldots & k_{1,1} \\ 0 & a_{1,2} & \ldots & k_{1,2} \\ \vdots & \vdots & \ldots & \vdots \\ 0 & a_{1,n} & \ldots & k_{1,n} \\ 0 & a_{2,1} & \ldots & k_{2,1} \\ 1 & a_{2,2} & \ldots & k_{2,2} \\ \vdots & \vdots & \ldots & \vdots \\ 0 & a_{2,n} & \ldots & k_{2,n} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & a_{n,n} & \ldots & k_{n,n} \end{pmatrix} \quad \text{(Equation 12)}$$

Similarly, the design matrix for estimating genetic correlation, obtained from expression of the form

$$E[\epsilon_{1,j}, \epsilon_{2,j}] = \sigma_{\epsilon,1}\sigma_{\epsilon,2}\rho_{\epsilon}1_{i=j} + \sigma_{a,1}\sigma_{a,2}\rho_a a_{i,j} + \ldots + \sigma_{k,1}\sigma_{k,2}\rho_k k_{i,j} \quad \text{(Equation 13)}$$

Can be written (if the variance parameters were known) as:

$$X_{\rho} = \begin{pmatrix} \sigma_{\epsilon,1}\sigma_{\epsilon,2} & \sigma_{a,1}\sigma_{a,2}a_{1,1} & \ldots & \sigma_{k,1}\sigma_{k,2}k_{1,1} \\ 0 & \sigma_{a,1}\sigma_{a,2}a_{1,2} & \ldots & \sigma_{k,1}\sigma_{k,2}k_{1,2} \\ \vdots & \vdots & \ldots & \vdots \\ 0 & \sigma_{a,1}\sigma_{a,2}a_{1,n} & \ldots & \sigma_{k,1}\sigma_{k,2}k_{1,n} \\ 0 & \sigma_{a,1}\sigma_{a,2}a_{2,1} & \ldots & \sigma_{k,1}\sigma_{k,2}k_{2,1} \\ \sigma_{\epsilon,1}\sigma_{\epsilon,2} & \sigma_{a,1}\sigma_{a,2}a_{2,2} & \ldots & \sigma_{k,1}\sigma_{k,2}k_{2,2} \\ \vdots & \vdots & \ldots & \vdots \\ 0 & \sigma_{a,1}\sigma_{a,2}a_{2,n} & \ldots & \sigma_{k,1}\sigma_{k,2}k_{2,n} \\ \vdots & \vdots & \ldots & \vdots \\ \sigma_{\epsilon,1}\sigma_{\epsilon,2} & \sigma_{a,1}\sigma_{a,2}a_{n,n} & \ldots & \sigma_{k,1}\sigma_{k,2}k_{n,n} \end{pmatrix} \quad \text{(Equation 14)}$$

Noting that:

$$X_{\rho} = X_{\sigma}\begin{pmatrix} \sigma_{\epsilon,1}\sigma_{\epsilon,2} & 0 & \ldots & 0 \\ 0 & \sigma_{b,1}\sigma_{b,2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_{k,1}\sigma_{k,2} \end{pmatrix} = X_{\sigma}D_{\sigma} \quad \text{(Equation 15)}$$

We get that:

$$X^T_{\rho}X_{\rho} = D_{\sigma}X^T_{\sigma}X_{\sigma}D_{\sigma} \quad \text{(Equation 16)}$$

$$\left(X^T_{\rho}X_{\rho}\right)^{-1} = D^{-1}_{\sigma}\left(X^T_{\sigma}X_{\sigma}\right)^{-1}D^{-1}_{\sigma} \quad \text{(Equation 17)}$$

We also note the outcome matrices for estimating variance components and correlation parameters are:

$$Y_{\sigma,1} = \begin{pmatrix} \epsilon_{1,1}\epsilon_{1,1} \\ \epsilon_{1,1}\epsilon_{1,2} \\ \vdots \\ \epsilon_{1,1}\epsilon_{1,n} \\ \epsilon_{1,2}\epsilon_{1,1} \\ \epsilon_{1,2}\epsilon_{1,2} \\ \vdots \\ \epsilon_{1,2}\epsilon_{1,n} \\ \vdots \\ \epsilon_{1,n}\epsilon_{1,n} \end{pmatrix}, Y_{\sigma,2} = \begin{pmatrix} \epsilon_{2,1}\epsilon_{2,1} \\ \epsilon_{2,1}\epsilon_{2,2} \\ \vdots \\ \epsilon_{2,1}\epsilon_{2,n} \\ \epsilon_{2,2}\epsilon_{2,1} \\ \epsilon_{2,2}\epsilon_{2,2} \\ \vdots \\ \epsilon_{2,2}\epsilon_{2,n} \\ \vdots \\ \epsilon_{2,n}\epsilon_{2,n} \end{pmatrix}, Y_{\rho} = \begin{pmatrix} \epsilon_{1,1}\epsilon_{2,1} \\ \epsilon_{1,1}\epsilon_{2,2} \\ \vdots \\ \epsilon_{1,1}\epsilon_{2,n} \\ \epsilon_{1,2}\epsilon_{2,1} \\ \epsilon_{1,2}\epsilon_{2,2} \\ \vdots \\ \epsilon_{1,2}\epsilon_{2,n} \\ \vdots \\ \epsilon_{1,n}\epsilon_{2,n} \end{pmatrix} \quad \text{(Equation 18)}$$

Therefore:

$$\begin{pmatrix} \sigma^2_{\epsilon,1} \\ \sigma^2_{a,1} \\ \vdots \\ \sigma^2_{k,1} \end{pmatrix} = \left(X_\sigma^T X_\sigma\right)^{-1} X_\sigma^T Y_{\sigma,1}, \begin{pmatrix} \sigma^2_{\epsilon,2} \\ \sigma^2_{a,2} \\ \vdots \\ \sigma^2_{k,2} \end{pmatrix} = \left(X_\sigma^T X_\sigma\right)^{-1} X_\sigma^T Y_{\sigma,2} \quad \text{(Equation 19)}$$

and:

$$\begin{pmatrix} \sigma^2_{\epsilon,1} \\ \sigma^2_{a,1} \\ \vdots \\ \sigma^2_{k,1} \end{pmatrix} = \left(X_\rho^T X_\rho\right)^{-1} X_\rho^T Y_\rho \quad \text{(Equation 20)}$$

$$= D_\sigma^{-1}\left(X_\sigma^T X_\sigma\right)^{-1} D_\sigma^{-1} D_\sigma X_\sigma^T Y_\rho \quad \text{(Equation 21)}$$

$$= D_\sigma^{-1}\left(X_\sigma^T X_\sigma\right)^{-1} X_\sigma^T Y_\rho \quad \text{(Equation 22)}$$

Because the l_{th} entry of D_σ is $\sigma_{1,l}\sigma_{2,l}$ we have then for ρ_l:

$$\rho_l = \frac{\left[\left(X_\sigma^T X_\sigma\right)^{-1} X_\sigma^T Y_\rho\right]_l}{\sigma_{1,l}\sigma_{2,l}} \quad \text{(Equation 23)}$$

$$= \frac{\left[\left(X_\sigma^T X_\sigma\right)^{-1} X_\sigma^T Y_\rho\right]_l}{\sqrt{\left[\left(X_\sigma^T X_\sigma\right)^{-1} X_\sigma^T Y_{\sigma,1}\right]_l \left[\left(X_\sigma^T X_\sigma\right)^{-1} X_\sigma^T Y_{\sigma,2}\right]_l}} \quad \text{(Equation 24)}$$

To prove that this is a generalized Pearson correlation, we only need to show that $[(X_\sigma^T X_\sigma)^{-1} X_\sigma^T Y_\rho]_l$ is a bilinear form, with the matrix completely defined by the l_{th} row of $(X_\sigma^T X_\sigma)^{-1} X_\sigma^T$. This is simple to see, because the entries of $X_\sigma^T Y_\rho$ are:

$$X_\sigma^T Y_\rho = \begin{pmatrix} \epsilon_1^T \boldsymbol{I} \epsilon_2 \\ \epsilon_1^T \boldsymbol{A} \epsilon_2 \\ \vdots \\ \epsilon_1^T \boldsymbol{K} \epsilon_2 \end{pmatrix} \quad \text{(Equation 25)}$$

Similarly:

$$X_\sigma^T Y_{\sigma,1} = \begin{pmatrix} \epsilon_1^T \boldsymbol{I} \epsilon_1 \\ \epsilon_1^T \boldsymbol{A} \epsilon_1 \\ \vdots \\ \epsilon_1^T \boldsymbol{K} \epsilon_1 \end{pmatrix}, X_\sigma^T Y_{\sigma,2} = \begin{pmatrix} \epsilon_2^T \boldsymbol{I} \epsilon_2 \\ \epsilon_2^T \boldsymbol{A} \epsilon_2 \\ \vdots \\ \epsilon_2^T \boldsymbol{K} \epsilon_2 \end{pmatrix} \quad \text{(Equation 26)}$$

and the l_{th} row of $(X_\sigma^T X_\sigma)^{-1}$ determines the weights in the following expression:

$$\left[(X_\sigma^T X_\sigma)^{-1} X_\sigma^T Y_\rho \right]_l = \epsilon_1^T [\omega_{\epsilon,l} \boldsymbol{I} + \omega_{a,l} \boldsymbol{A} + \ldots + \omega_{k,l} \boldsymbol{K}] \epsilon_1 \quad \text{(Equation 27)}$$

Thus for

$$\boldsymbol{S_l} = \omega_{\epsilon,l} \boldsymbol{I} + \omega_{a,l} \boldsymbol{A} + \ldots + \omega_{k,l} \boldsymbol{K} \quad \text{(Equation 28)}$$

We get

$$\widehat{\rho}_l = \frac{\widehat{\epsilon}_1^T \boldsymbol{S_l} \widehat{\epsilon}_2}{\sqrt{(\widehat{\epsilon}_1^T \boldsymbol{S_l} \widehat{\epsilon}_1)(\widehat{\epsilon}_2^T \boldsymbol{S_l} \widehat{\epsilon}_2)}} \quad \text{(Equation 29)}$$

Deriving confidence intervals for estimated correlation coefficients; We propose two methods to compute confidence intervals for the estimated correlation coefficients. First, using the Fisher's transformation, which was developed to estimate confidence intervals for the standard Pearson correlation coefficient, and second, using block bootstrap.

Confidence intervals using the Fisher's transform

Fisher's transformation converts the distribution of the correlation coefficients to a normal one and thus allows us to calculate confidence intervals (and corresponding p values) for the correlation coefficient using the values of the correlation coefficient and the sample size.[67,68] Since we show that calculating genetic correlation is equal to calculating standard correlation for adjusted phenotypes (Equation 29), the Fisher method is equally applicable for genetic correlation coefficient with the modification of plugging-in ''effective sample size'' to account for the modeled correlation structure between the two traits. Specifically, Fisher's z-transformation of a correlation coefficient ρ is defined as:

$$z = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right) = arctanh(\rho) \quad \text{(Equation 30)}$$

If the two variables for which the correlation is measured have a bivariate normal distribution and are independent and identically distributed, then z is approximately normally distributed with mean μ and SE σ given by:

$$\mu = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right) \quad \text{(Equation 31)}$$

$$\sigma = \frac{1}{\sqrt{N_{eff} - 3}} \quad \text{(Equation 32)}$$

and N_{eff} being the effective sample size of our sample equal to:

$$N_{eff} = \sqrt{tr\left(\boldsymbol{S_l} \boldsymbol{S_l^T} \right)} \quad \text{(Equation 33)}$$

where $\boldsymbol{S_l}$ is the weighted matrix described in Equation 28. The coverage of this approach was verified in simulations via comparisons to the block Bootstrap method described below (see Figure S1).

Confidence intervals using block bootstrap

Multiple participants in our datasets are genetically related (and thus correlated), which violates the assumption of the standard bootstrap method. We thus performed the block bootstrap procedure to derive the confidence intervals and p values as described in.[69] Briefly, related individuals were grouped into blocks (via third degree kinship) and the sampling procedure was at the level of blocks. We applied the Fisher's transformation on each correlation value estimated in the bootstrap. Standard deviations and consequently confidence intervals and p values were calculated based on the Fisher's transformed values. Finally, to obtain confidence intervals on the original genetic correlation scale we applied the inverse transformation of the Fisher's transform (given by $f(x) = (e^{2x} - 1)/(e^{2x} + 1)$) to the endpoints of the confidence intervals obtained on the Fisher's transformation scale. In addition, we used the quantiles method to derive 95% confidence interval from the bootstrap results and report whether the corresponding p value is < 0.05 based on the null value being in the confidence interval.

Testing for difference of two correlation coefficients

To compare two correlation coefficients, we transform both correlations using the Fisher-z transformation (as described in Equation 30). Once the correlations have been converted into z values, the normal distribution is used to conduct the test of $Z_1 - Z_2$.

Fractional genetic correlation

For simplicity, focus on the single correlation matrix settings (the derivation here naturally extends to multiple sources of correlation). Recall *Model 2* - the phenotypic correlation coefficient $\boldsymbol{R}$ is equal to:

$$R = \frac{cov(\epsilon_1, \epsilon_2)}{\sqrt{cov(\epsilon_1)cov(\epsilon_2)}} = \left(\frac{\sigma_{\epsilon,1}\sigma_{\epsilon,2}}{\sqrt{\sigma^2_{\epsilon,1} + \sigma^2_{k,1}}\sqrt{\sigma^2_{\epsilon,2} + \sigma^2_{k,2}}} \times \rho_\epsilon \right) + \left(\frac{\sigma_{k,1}\sigma_{k,2}}{\sqrt{\sigma^2_{\epsilon,1} + \sigma^2_{k,1}}\sqrt{\sigma^2_{\epsilon,2} + \sigma^2_{k,2}}} \times \rho_k \right) \quad \text{(Equation 34)}$$

We then define Fractional Genetic Correlation (ρ_{fk}) as

$$\rho_{fk} = \left(\frac{\sigma_{k,1}\sigma_{k,2}}{\sqrt{\sigma^2_{\epsilon,1} + \sigma^2_{k,1}}\sqrt{\sigma^2_{\epsilon,2} + \sigma^2_{k,2}}} \times \rho_k \right) \quad \text{(Equation 35)}$$

and Fractional Residual Correlation as

$$\rho_{f\epsilon} = \frac{\sigma_{\epsilon,1}\sigma_{\epsilon,2}}{\sqrt{\sigma^2_{\epsilon,1} + \sigma^2_{k,1}}\sqrt{\sigma^2_{\epsilon,2} + \sigma^2_{k,2}}} \times \rho_\epsilon$$

.

This is a natural decomposition of phenotypic correlation into two components:

$$R = \rho_{fk} + \rho_{f\epsilon} \quad \text{(Equation 36)}$$

Unlike standard genetic correlation, ρ_{fk} is the genetic correlation adjusted (fractional) for both traits' heritabilities and variances and crucially it represents a fraction of the phenotypic correlation that is due to genetics. As expected, the fractional correlation terms are never larger than the phenotypic correlation R, because they sum to R.

QUANTIFICATION AND STATISTICAL ANALYSIS

Simulation studies

We studied the accuracy of the proposed method for estimating genetic correlations and for calculating confidence intervals in simulations. We used correlation matrices from the HCHS/SOL representing kinship and shared household to generate realistic correlation structures. In all simulations, data were generated by first sampling two uncorrelated error vectors (ϵ_1, ϵ_2) from a standard normal distribution. We next simulated the covariance structure according to our model:

$$cov(\epsilon_l) = \sigma^2_{\epsilon,l}\boldsymbol{I}_{n\times n} + \sigma^2_{k,l}\boldsymbol{K} + \sigma^2_{h,l}\boldsymbol{H} cov(\epsilon_1, \epsilon_2) = \sigma_{\epsilon,1}\sigma_{\epsilon,2}\rho_\epsilon\boldsymbol{I}_{n\times n} + \sigma_{k,1}\sigma_{k,2}\rho_k\boldsymbol{K} + \sigma_{h,1}\sigma_{h,2}\rho_h\boldsymbol{H}$$

The matrix K **represents** kinship, and H represents shared household. All simulations were performed 1,000 times with different sample sizes (1000, 4000, and 7706, the latter is sample size of HCHS/SOL individuals in TOPMed freeze 8 which is the smallest subgroup in this study) and values of ρ_k and ρ_h ranging from 0 to 1 in increments of 0.1. The variance components reported here were set to typical values for phenotypes from our dataset and equal to $\sigma = (\sigma_{k,1}\sigma_{k,2}, \sigma_{h,1}, \sigma_{h,2}) = (0.6, 0.7, 0.4, 0.3)$ (corresponding to HDL, height, fasting glucose levels and eosinophil counts accordingly). The confidence intervals and p values were calculated from the block bootstrap method using the Fisher's transformation and the percentile method (Figure S1).

We repeated all the simulations for a single kinship matrix on different self-reported background groups using kinship matrices from TOPMed participants from the available background groups (Black, Hispanic/Latino, and White). We also simulated a joint group in which we sampled an equal number of TOPMed individuals from each of the background groups, and used the corresponding kinship matrix. The number of participants was kept at 7706, and we compared the performance of HEc method (1000 repeats per group per ρ_k) to GCTA-GREML (with 40 repeats due to the slow computation speeds).

Heritability and genetic/environment correlation estimation via HEc and GCTA-GREML

The relatedness between individuals is modeled via a kinship (K) matrix, and an additional household matrix (H) for modeling environmental effects (available only for the HCHS/SOL cohort). Each phenotype was regressed on age, gender, sampling weights and 11 first principal components (and race/ethnicity and study for TOPMed combined cohorts) and the residuals were rank-normalized. We estimated the correlation coefficients corresponding to the relatedness matrices for all trait pairs by plugging in the normalized residuals to Equation 29. This was implemented via R scripts provided in GitHub repository [https://github.com/tamartsi/HE_Genetic_Correlation]. The genetic and environment variance components as well as the corresponding heritabilities were

calculated via the GCTA software.[57] Following sensitivity analysis for the presence of related individuals in the cohorts (Figure S5) we removed all individuals related at third degree or more for the calculation of heritabilities, however as we did not see any substantial effects of relatives on the estimated genetic correlations (Figure S2B), the relatives were kept in for genetic and environmental correlation coefficients estimation. We provide the confidence intervals for both the Fisher Method and the two Bootstrap-Based approaches in the supplementary data files. Visualizations were performed via the R packages igraph,[58] qgraph,[59] ggplot2[70] and corrplot[60] followed by Adobe Illustrator. The figures are based on uncorrected p values. FDR-corrected p values are provided in the Supplementary Data files.

Heritability and genetic correlation estimation via LD-based methods

Summary statistics for 8 selected phenotypes (Table S5) were downloaded from PAN-UK BioBank (https://pan.ukbb.broadinstitute.org/). The GWAS underwent further cleanup and quality control as recommended by the ldsc package (https://github.com/bulik/ldsc).[10,28] LD scores were computed for each of the TOPMed self-reported background groups as well as for the joint dataset as recommended by the ldsc software. Alternatively, covariate-adjusted LD scores were calculated via cov-LDSC (https://github.com/immunogenomics/cov-ldsc)[34] using 11 principal components (calculated as described in section 2.4). Finally, genetic correlations and heritabilities were estimated using the computed LD scores.

Domain-level enrichment analysis

We calculated the enrichment of inter-domain correlation via a permutation approach. Specifically, for 1000 repeats, we generated random connections between nodes in our correlation graph such that each node will receive a same number of connections as in the real dataset as well as keeping the overall number of connections identical. We then calculated the distribution of number of connections between each pair of domains and used it to obtain a domain enrichment p value as follows:

$$\text{p-value}\left(\text{domain 1},\ \text{domain 2}\right) = \frac{1}{1000}\sum_{i=1}^{1000} 1\left(N_i^c > N^c\right)$$

where N^c is the number of connections between domains 1 and 2, and N_i^c, $i = 1, \ldots, 1000$ is the number of connections between domains 1 and 2 in the i th permutation. We considered two domains to be enriched if their enrichment p value was <0.05.

ADDITIONAL RESOURCES

No additional resources were used.

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

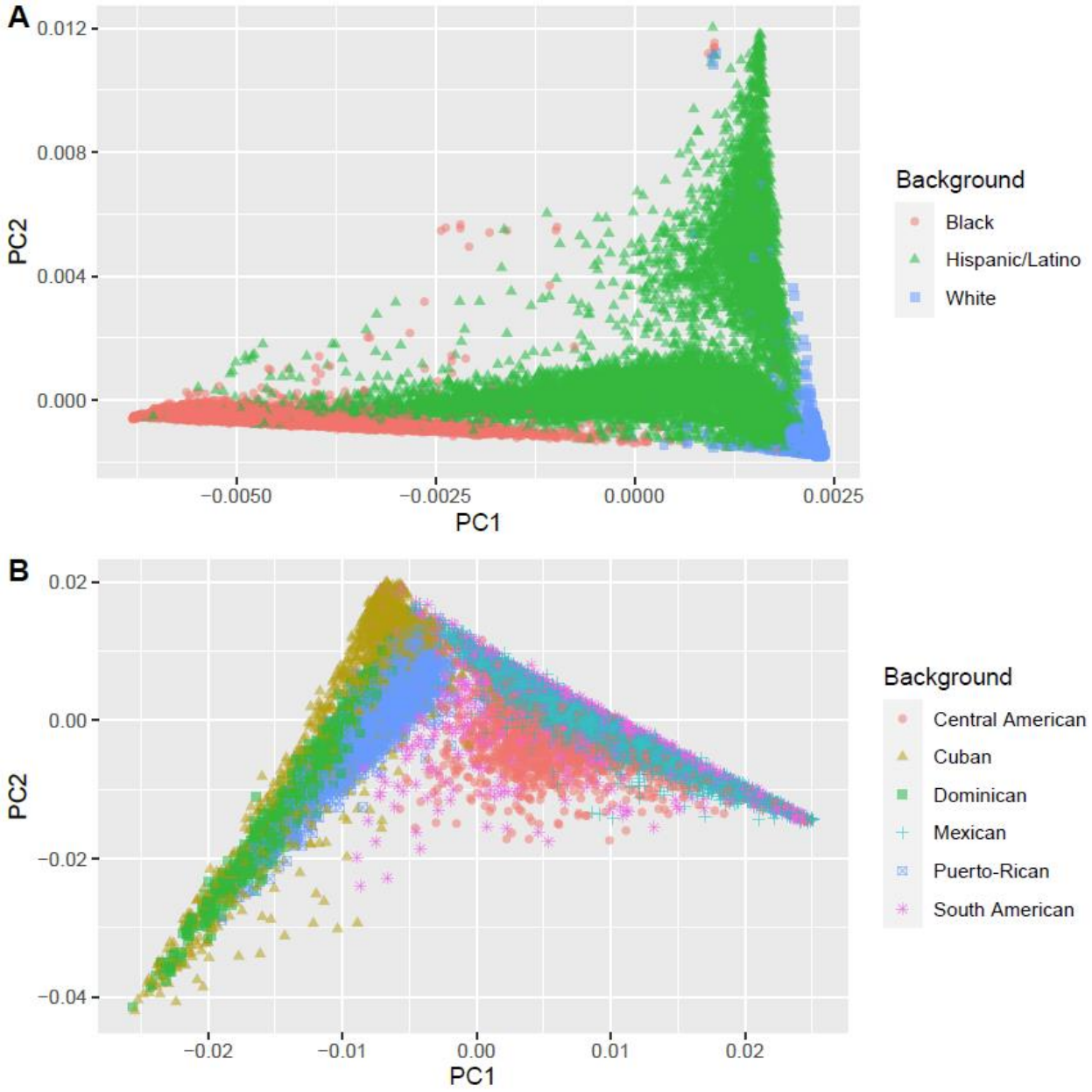


Figure S1 (related to Figures 2-4). Populations backgrounds as captured by PCA analysis

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)**.

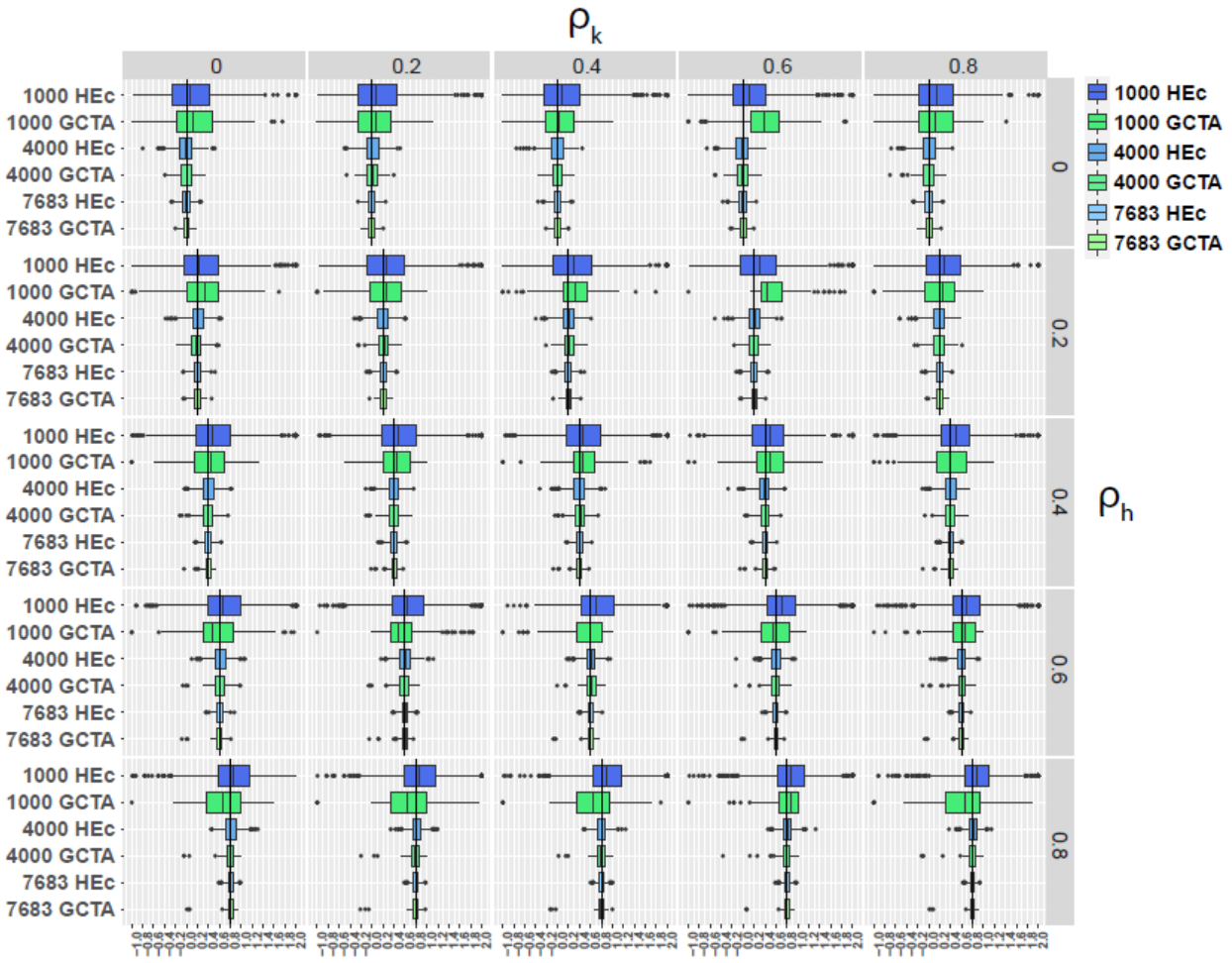


Figure S2 (related to Figure 1). Comparison of simulations compare HEc, the ground truth and GCTA-GREML
Two relatedness matrices were used to simulate phenotypes with known correlation coefficients (ρ_h, ρ_k). Each phenotype was simulated 1000 times in 1000, 4000 and 7683 people. Shown here are the boxplots 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].

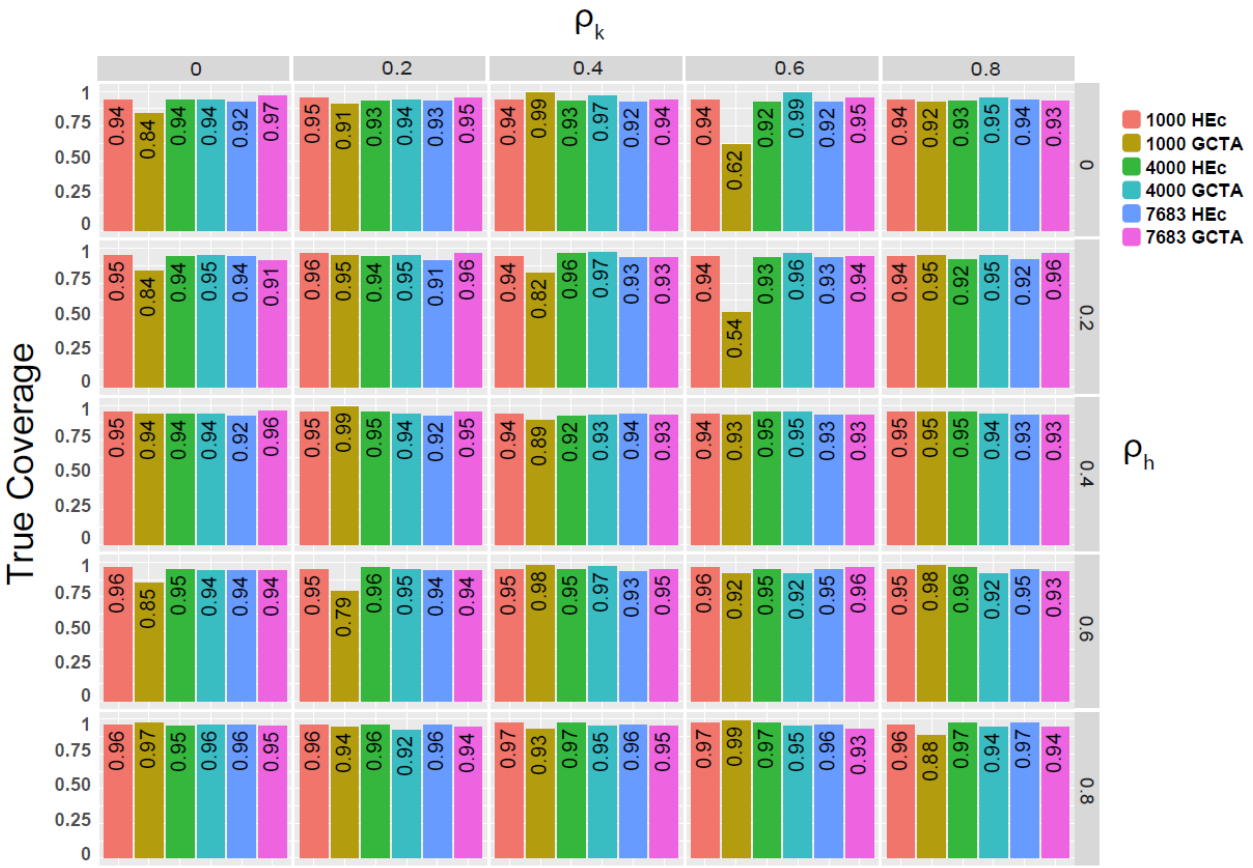


Figure S3 (related to Figure 1). Comparison of confidence interval coverages in simulations of HEc and GCTA-GREML approach

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).

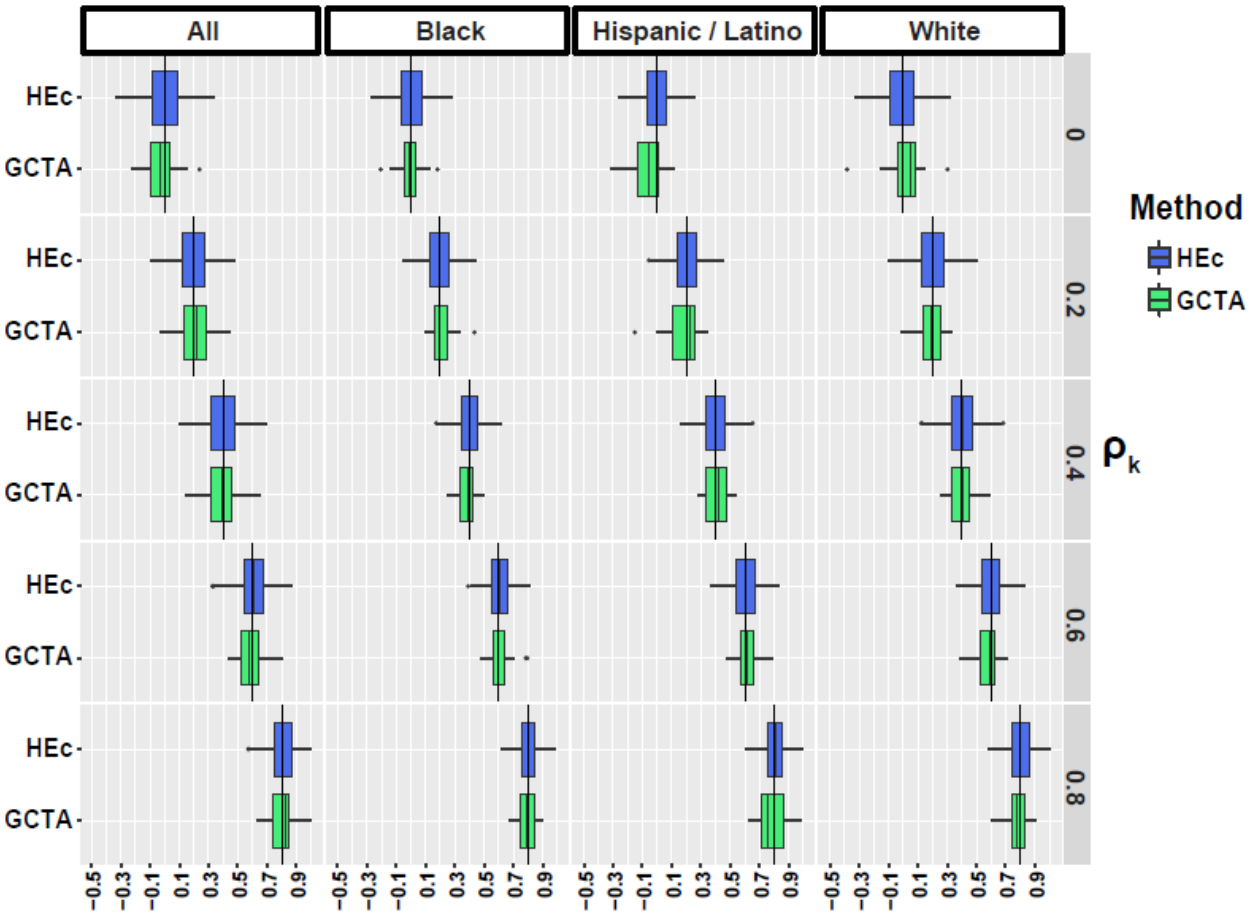


Figure S4 (related to Figure 1). Comparison HEc and GCTA-GREML applied on different populations in simulations

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.

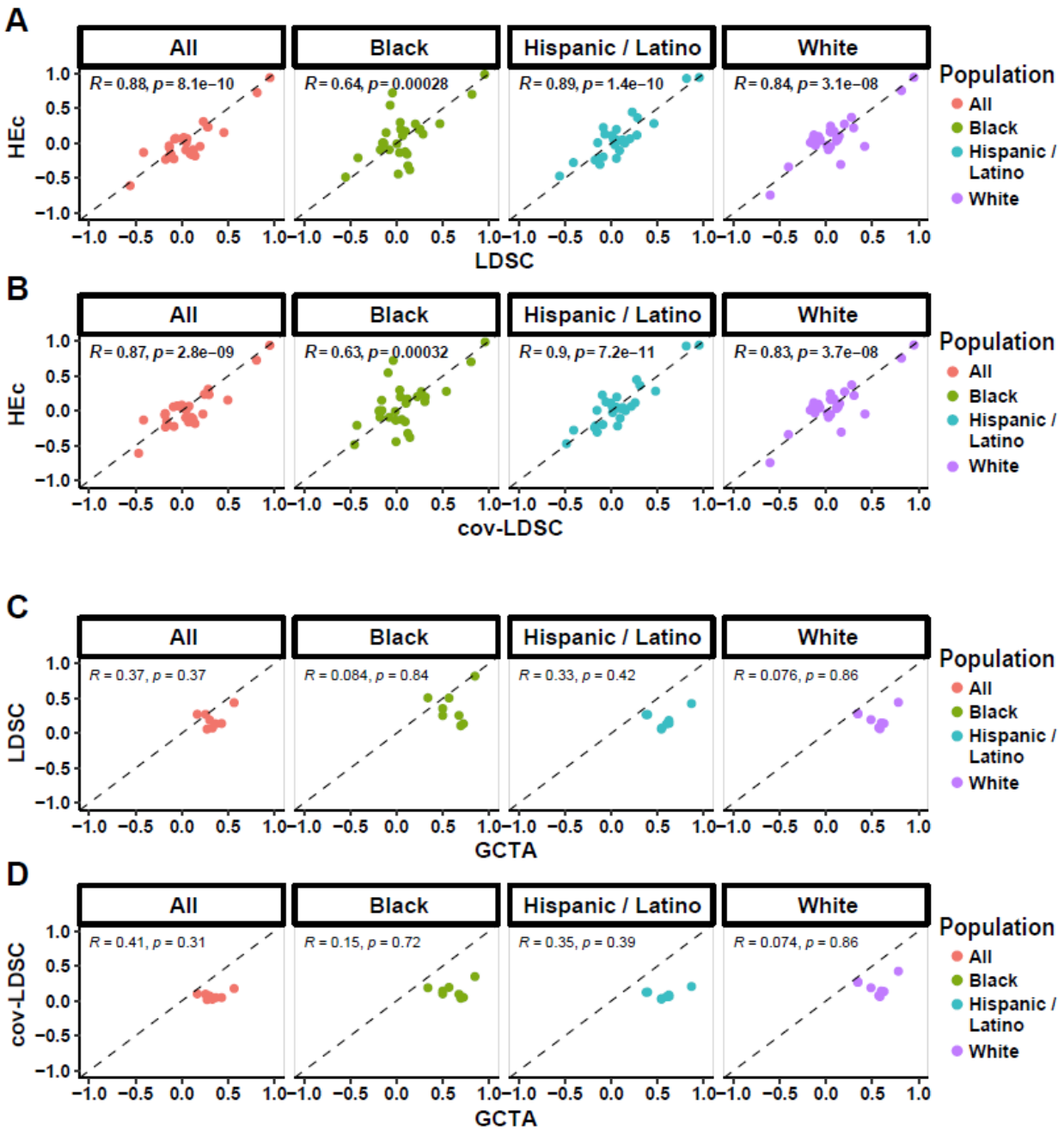


Figure S5 (related to Figure 1). Comparison of results of HEc, GCTA-GREML and two LD-score based methods for computation of phenotype heritabilities and genetic correlations from the TOPMed dataset stratified by populations

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.

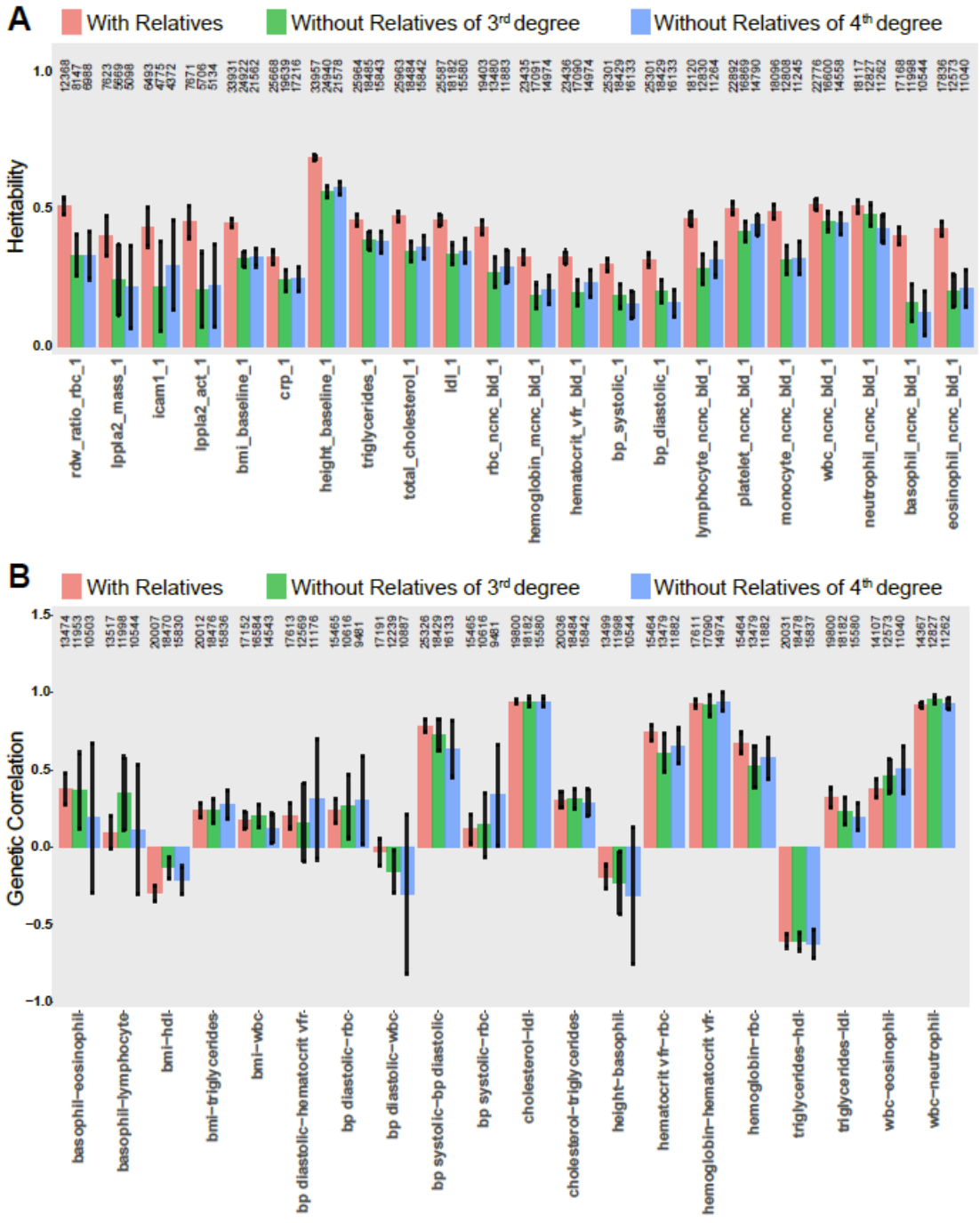


Figure S6 (related to Figure 1). Sensitivity analysis studying the effect of presence of relatives in the data on heritability and genetic correlation estimates

(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).

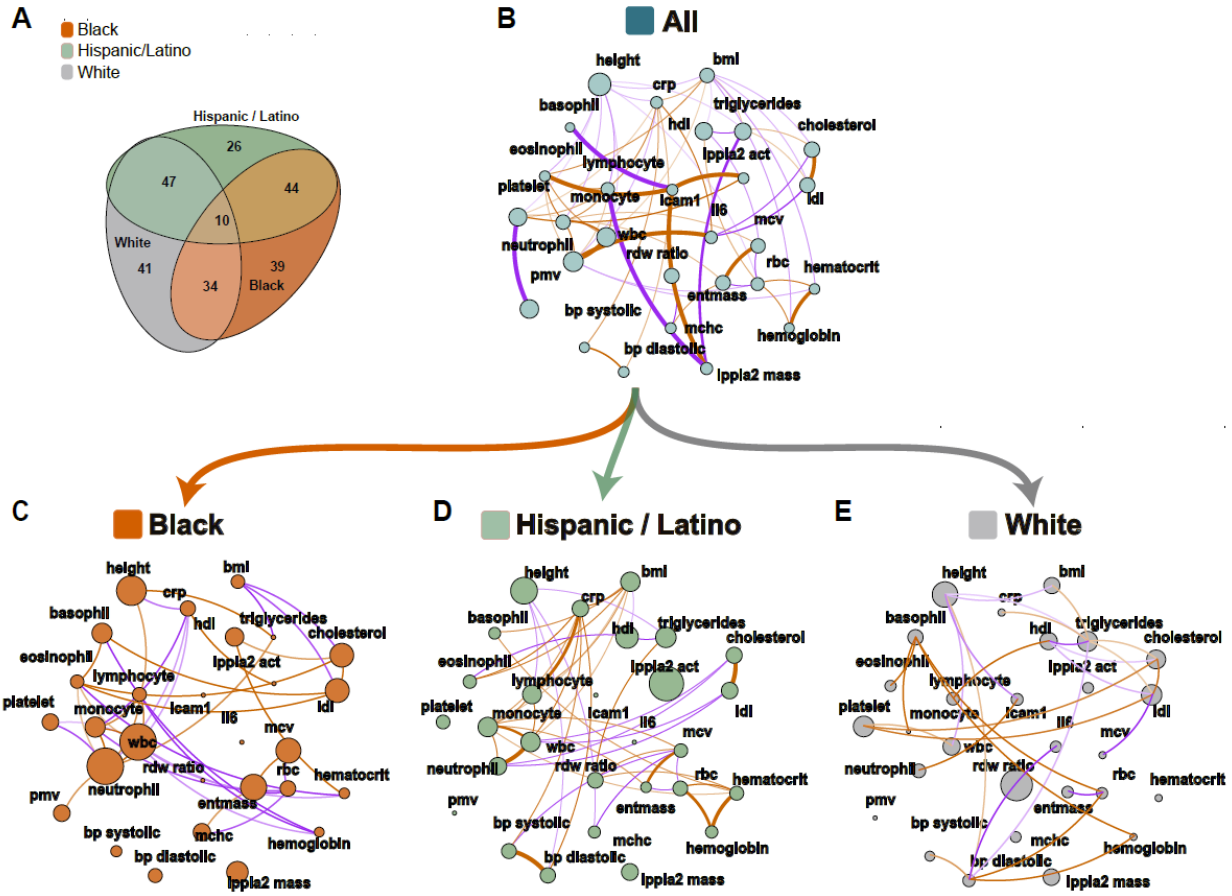


Figure S7 (related to Figure 3). Certain genetic correlations are background-specific.

(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.

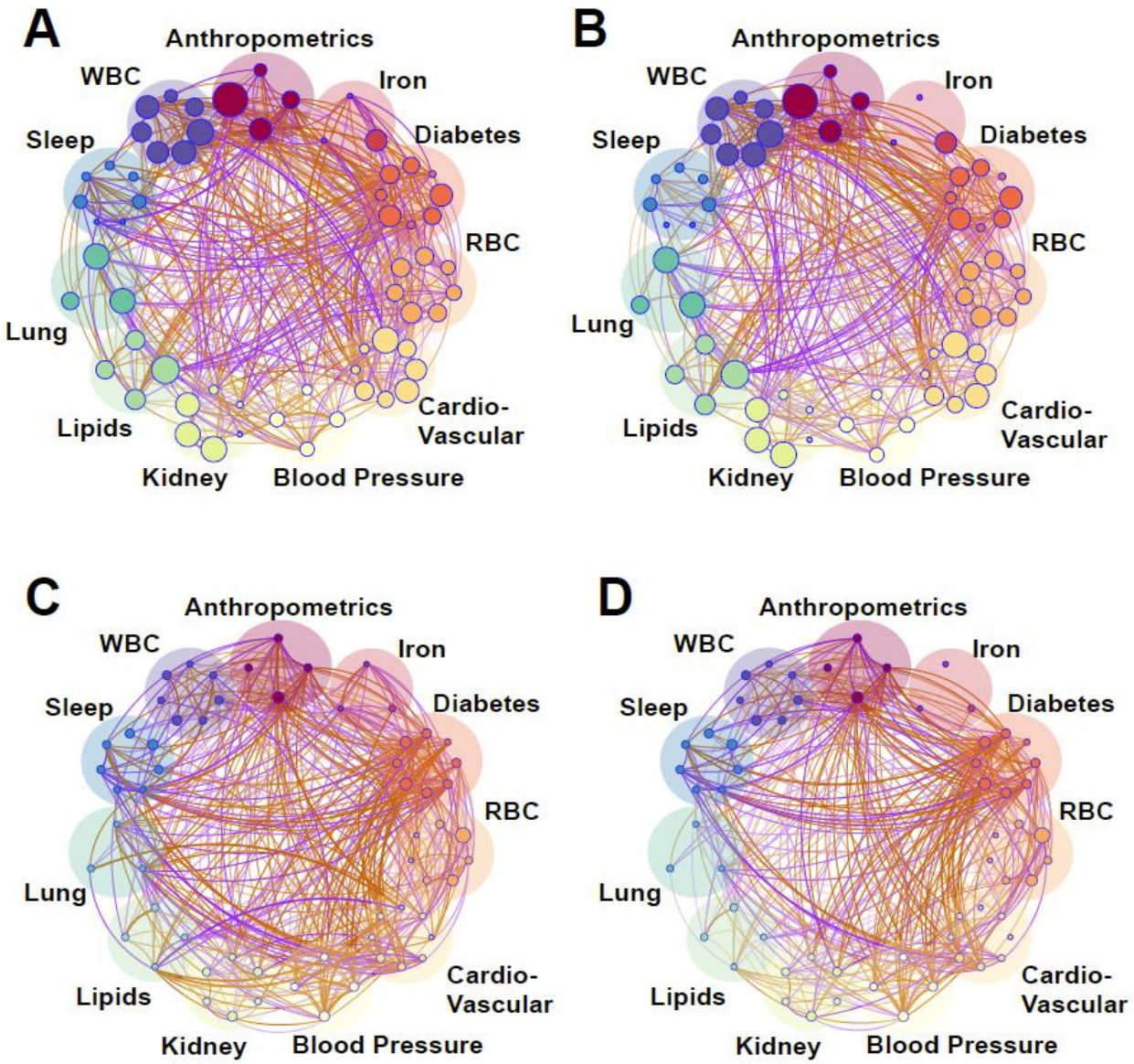


Figure S8 (related to Figure 4). Genetic and environmental correlations and heritabilities of 61 phenotypes in self-reported Hispanics/Latinos.

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).

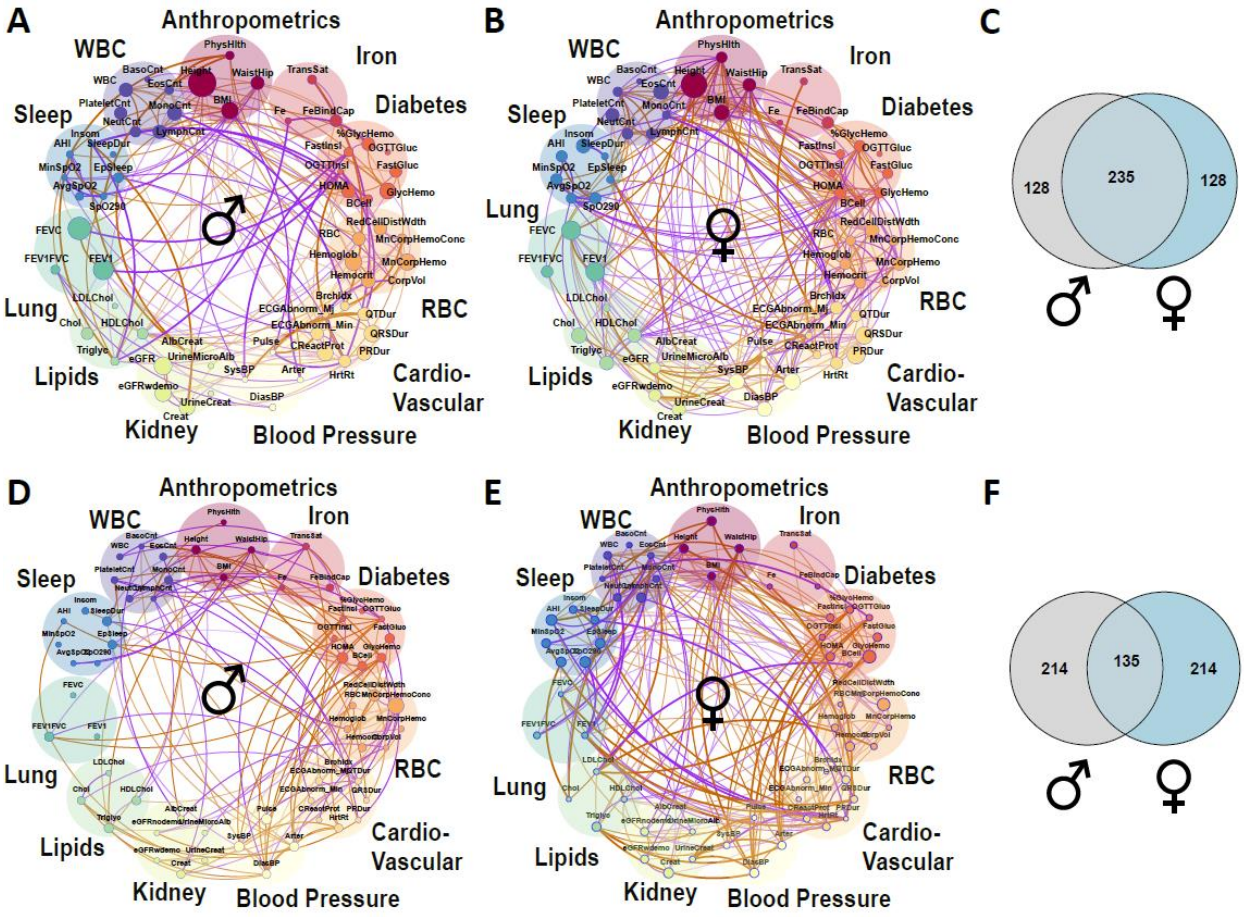


Figure S9 (related to Figure 5). Genetic and environmental correlations and heritabilities differ by gender in Hispanics/Latinos.

(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)** .

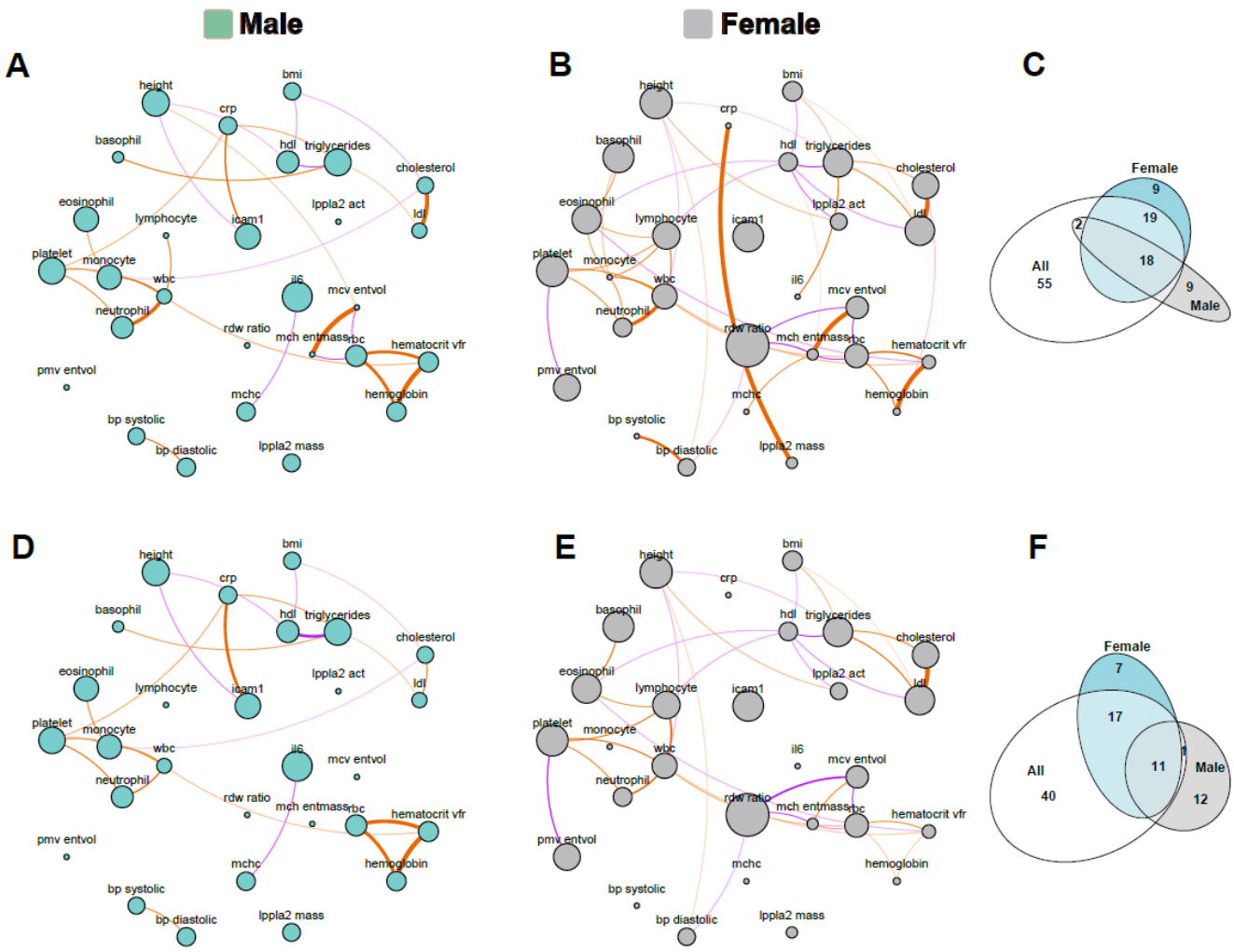


Figure S10 (related to Figure 5). Genetic correlations in individuals of White background stratified by gender.
(A-D) Correlation plots where each phenotype is represented by a node and the correlations are represented by connections (edges) between nodes. The size of the node is proportional to the 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