Examining the Effect of Social Determinants of Health on Human Trait Heritability

Oona Shigeno Risse-Adams[1,2*], Jaclyn L. Liquori[1], Nasa Sinnott-Armstrong[2,3,4*†], Shaila A. Musharoff[1*†]

[1]Department of Computational Biology, Cornell University, Ithaca, NY, USA.
[2]Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA, USA.
[3]Department of Genome Sciences, University of Washington, Seattle, WA, USA.
[4]Brotman Baty Institute, University of Washington, Seattle, WA, USA.

*Corresponding author(s). E-mail(s): or93@cornell.edu; nasa@fredhutch.org; sam442@cornell.edu;
†These authors contributed equally to this work.

Abstract

Social determinants of health (SDOH) data are often excluded from genetic models of human traits. We use individual-level data from the All of Us Research Program to assess whether SDOH survey variables alter heritability estimates. Incorporating SDOH summaries decreases heritability significantly for 4 of 18 anthropometric and metabolic traits in individuals whose self-reported race is "White" (n=67,545). There are no such significant changes in individuals whose self-reported race is "Black or African American" (n=6,538), likely reflecting reduced statistical power. Incorporating genetic principal components consistently lowers heritability estimates in both groups, whether or not SDOH summaries are included. These findings demonstrate that survey-derived SDOH summaries can change heritability estimates and should be considered along with genetic summaries.

Keywords: heritability, social determinants of health, human genetics

1 Introduction

Understanding the environmental contributions to human traits remains a critical challenge in statistical genetics. While genome-wide studies routinely adjust for genetic structure, trait-relevant environmental exposures are typically unmeasured or oversimplified. Social determinants of health (SDOH)—factors such as housing, financial insecurity, education, and perceived stress—shape both disease risk and access to health care. SDOH factors are increasingly captured in structured survey form in biobank-scale cohorts, including the All of Us Research Program [1]. However, SDOH data are rarely explicitly incorporated in heritability and genetic association models [2, 3], which could result in SDOH-stratified phenotypic residuals. While the All of Us SDOH data has been previously characterized [4] and connected to health outcomes [5], to our knowledge, it has not been incorporated in heritability models.

Here, we investigate how high-dimensional environmental structure captured by the All of Us SDOH survey influences heritability estimates of 18 anthropometric and metabolic traits. To summarize the SDOH survey responses, we apply one of three strategies: (i) dimension reduction via principal component analysis (PCA) [6], (ii) dimension reduction via multiple correspondence analysis (MCA), and (iii) an environmental relatedness matrix (ERM), from which we computed an individual-level environmental score (ERM row sum) by weighting and summing one-hot encoded survey responses that have been corrected for question complexity and missingness. Each SDOH summarization strategy yields participant-level covariates that can be used in genetic models.

2 Results

We analyzed data from individuals who had available SDOH survey and genotyping array data and whose self-identified race is either "White" (n=67,545) or "Black or African American" (n=6,583; here, "Black/African American" or "Black/AA"). Collinearity and structure analyses were performed on this full sample, while heritability estimation analyses were restricted to the subsets of individuals who additionally had measurements available for the 18 traits of interest.

Collinearity

We first assessed collinearity between the three survey-derived social determinants of health (SDOH) summaries (ERM row sum, MCA, PCA) and genetic principal components (PCs) (Fig. 1). For the ERM row-sum, the maximum absolute Pearson correlation with any PC was 0.313 and all variance inflation factors (VIFs) remained below 1.2. For the top six MCA axes, the maximum correlation with any PC was 0.301 (VIFs < 1.4), and for the top twenty PCA axes it was also 0.301 (VIFs $<$ 1.5). Thus each SDOH summary can safely be modeled alongside genetic PCs without problematic multicollinearity.

However, when analyzing self-identified White and Black/African American groups separately, elevated VIFs (3.99) prompted further investigation. Pearson correlation plots of SDOH and genetic PCs within each subgroup revealed modest correlations

between genetic components that were not evident in the combined sample (Figs. S4 and S5). These correlations may reflect the expected loss of orthogonality when principal components computed on a full sample are applied to subsets rather than meaningful structure. As such, caution is warranted when interpreting within-group relationships, as restricting analyses to specific demographic groups may amplify spurious associations among otherwise orthogonal PCs.

Trait heritability

We next used Haseman–Elston regression to obtain heritability estimates via cross product (HE-CP) for the following six covariate models:

Model 0: Base, which is age + sex
Model 1: Base + Genetic PCs 1:3
Model 2: Base + (ERM Row sum or SDOH PCs 1:3 or MC Axes 1:6)
Model 3: Base + (SDOH PCs 1:20 or MC Axes 1:40)
Model 4: Base + (ERM Row sum or SDOH PCs 1:3 or MC Axes 1:6) + Genetic PCs 1:3
Model 5: Base + (SDOH PCs 1:20 or MC Axes 1:40) + Genetic PCs 1:3

Unlike the SDOH MCA and PCA summaries, which can be partitioned into different axis subsets for inclusion across four models (Models 2-5), the SDOH ERM row sum is a single statistic and only appears in two models (Models 2 and 4).

In White individuals (Fig. 2), adding genetic PCs 1:3 (Models 1, 4, and 5) consistently reduced heritability relative to the Base (Model 0) and SDOH-only models (Models 2 and 3). These reductions are significant in 11 of 18 traits: BMI, HDL cholesterol, carbon dioxide, creatinine, diastolic blood pressure, glucose, hip circumference, heart rate, sodium, urea nitrogen, and waist circumference. This implies that genetic PCs capture relevant population structure.

In White individuals, adding SDOH MCA or PCA covariates to the Base model (Models 2 and 3) also reduced the heritability estimate of 4 of 18 traits - BMI, hip circumference, waist circumference, and HDL cholesterol - while the ERM row sum had no significant effect on any trait. Specifically, including SDOH MCA axes 1:6 (Model 2) significantly reduced heritability estimates relative to the Base model (Model 0) for BMI (by 0.233 from 0.526 to 0.293), hip circumference (by 0.208 from 0.633 to 0.425), and waist circumference (by 0.231 from 0.686 to 0.455). Including the MCA axes 7:40 (Model 3) further reduced the heritability of these traits, though not significantly. While SDOH PCs 1:3 (Model 2) had little effect on trait heritability, adding SDOH PCs 4:20 (Model 3) significantly reduced heritability relative to both the Base model (Model 0) and SDOH PCs 1:3 alone (Model 2) for BMI (by 0.275 from 0.526 to 0.251), HDL cholesterol (by 0.216 from 0.475 to 0.259), hip circumference (by 0.250 from 0.633 to 0.383), and waist circumference (by 0.261 from 0.686 to 0.425).

In Black/African American individuals, (Fig. 3), including only SDOH covariates (Models 2 and 3) did not significantly change heritability relative to the Base model, and adding genetic PCs 1:3 (Models 1, 4, and 5) reduced heritability point estimates, though not significantly so, likely reflecting reduced statistical power. These analyses demonstrate that survey-derived SDOH data summarized with PCA and MCA can influence heritability estimates, sometimes significantly so. Our results indicate that

genetic PCs should be included with SDOH summaries in heritability models to reduce potential confounding.

Structure analysis

We assessed the extent to which SDOH and genetic population structure are associated by examining bidirectional relationships between SDOH PCs and genetic PCs in the combined sample of Black/African American and White individuals, as well as within each group separately.

To assess how social determinants explain genetic structure, we regressed each of genetic PCs 1:10 on SDOH PCs 1:20 in the combined sample of individuals. A linear combination of SDOH PCs explained non-trivial variation in genetic PC1 (multiple linear regression $R^2 = 0.184$), likely reflecting alignment between global ancestry and social groupings (Table S8). For genetic PCs 2:10, the variance explained by SDOH PCs was considerably lower (max R^2 for any genetic PC is 0.035), suggesting limited influence beyond the primary ancestry axis, genetic PC 1.

Next, to assess how genetic structure explains social determinant patterns, we regressed each of SDOH PCs 1:20 on genetic PCs 1:10. Genetic PCs explained modest amounts of variation in SDOH structure. While all 20 SDOH PCs showed statistically significant associations with genetic PCs, most R^2 values remained under 0.029, with the strongest associations observed for SDOH PCs 2 and 5 ($R^2 = 0.041$ and 0.062, respectively; Table S9). These components (SDOH PCs 2 and 5) likely capture SDOH topics such as healthcare discrimination, social support, neighborhood disorder, religious attendance, religious spirituality, and perceived stress (Fig. S6).

When we repeated these analyses separately in Black/African American and White individuals, we observed similar trends as in the combined sample, but the associations were generally weaker. In both Black/African American and White subgroups, genetic PCs explained little of the variation in SDOH PCs ($R^2 < 0.015$), though statistical significance was observed more frequently in the White group (Table S11, S13). Similarly, SDOH PCs explained only a small fraction of variance in genetic PCs, with R^2 values ranging from 0.001 to 0.036, again with more significant associations observed in the White group (Table S12). In the Black/African American group, associations were weaker and more diffuse (Table S10). It is important to note that the sample size between these two groups differ substantially with the size for the Black/African American group being an order of magnitude smaller than the White group (Table S2). This may contribute to reduced power and greater estimate variability in any stratified analyses.

These findings suggest that while SDOH and genetic PCs are statistically associated—particularly where social and ancestral structures align—the practical overlap is limited. Most of the variation in each domain is not captured by the other. This underscores the importance of treating social and genetic factors as distinct sources of variation in analyses of complex traits and reinforces the need to measure environmental exposures directly, rather than relying on race as a proxy. Further, this apparent distinction between domains supports the use of both sets of PCs in downstream analyses.

3 Methods

All of Us Study

The All of Us Research Program is a major longitudinal study designed to enhance biomedical research and improve health outcomes by enrolling at least one million diverse participants across the USA. This initiative aims to address historical gaps in research by including groups that have been underrepresented in biomedical studies. The program recently released clinical-grade genome sequences from 245,388 (now 312,924) participants, with a notable emphasis on diversity—77% of these participants come from historically under-represented groups, and 46% are from racial and ethnic minorities. The dataset includes over one billion genetic variants and has facilitated the identification of genetic associations with various diseases. Data is accessible to researchers via the All of Us Researcher Workbench, which allows for broad scientific use while ensuring participant privacy. The program collects a wide range of health data, including electronic health records, physical measurements, and survey responses, in addition to genetic information. This extensive dataset is expected to advance the field of genomic medicine by providing valuable insights into the genetic basis of health and disease across diverse populations.

The Social Determinants of Health (SDOH) survey, developed by the All of Us Research Program, aims to gather data on social factors influencing health. The survey was designed to be valid, reliable, and relevant to precision medicine, capturing dimensions of social determinants such as income, education, and housing quality. Out of 397,732 eligible participants, 117,783 (29.6%) completed the optional survey by June 30, 2022.

Of these, only 94,030 have array data available. Among those who completed the SDOH survey, 78.4% were from underrepresented populations in biomedical research. The survey was available in English and Spanish, and was designed to be completed in about 10-15 minutes. Participants who responded were generally more likely to be White and female.

Quality Control

Genotyping was performed using a custom Illumina array built on the design framework of the Global Screening Array. The genome build used is HG38. Please refer to the All of Us Research Program technical documentation and data release notes for further details.

Individuals who self-identified as either White or Black/African American, were included if they had genetic array data available, completed the SDOH survey, and had a sex of assigned male or female at birth.

Variant filtering was conducted using PLINK 2.0 [7] on the All of Us version 7 controlled-tier array data, restricted to autosomal SNPs. This reassigned variant identifiers and removed all duplicate variants. To mitigate the effects of linkage disequilibrium, we pruned markers using a sliding window approach (200 variants per window, with a step size of 50 variants) and an r^2 threshold of 0.2. Markers with a minor allele frequency below 0.01 were excluded, as were variants with a genotype

missingness rate exceeding 1%, markers deviating from Hardy-Weinberg equilibrium ($p < 1x10^{-6}$), and individuals with more than 2% missing genotype calls. Relatedness up to the third degree was then assessed via KING [8] (kinship coefficient ≥ 0.0625), and one member of each related pair was removed. This resulted in 7,348 Black/African American (5,355 Females, 1,993 Males) and 71,124 White (44,777 Females, 26,347 Males) individuals.

Trait Processing

We obtained biomarkers from the All of Us Research Program and performed data cleaning and deduplication procedures prior to downstream analyses. We restricted the data to 18 clinically relevant anthropometric and metabolic traits: alanine aminotransferase (AAT), alkaline phosphatase (AlkPhos), aspartate aminotransferase (AspAminTrans), body mass index (BMI), carbon dioxide (CO2), chloride (Cl), cholesterol in HDL (CholHDL), creatinine (Creat), diastolic blood pressure (DBP), glucose, heart rate (HR), hematocrit, hip circumference (HipCirc), protein, sodium, systolic blood pressure (SBP), urea nitrogen (UreaNitro), and waist circumference (WC). Sample sizes varied by trait and race due to differential availability of measurements. To mitigate the influence of extreme values, we calculated the IQR for each biomarker independently, and observations falling outside these bounds were truncated to the corresponding threshold values. For individuals with multiple measurements, we retained the most recent measurement. Systolic blood pressure was adjusted for antihypertensive use based on binary indicator variable.

Estimating Variance Contribution by Genome-wide Complex Trait Analysis (GCTA)

Genome-wide Complex Trait Analysis (GCTA [9]) was used to compute heritability estimates with Haseman-Elston (HE) regression applied to the analyzed traits. HE-CP estimates were taken as point estimates and 95% confidence intervals (CIs) were constructed as the point estimate plus or minus 1.96 times the HE standard error based on a jackknife over individuals. Non-overlapping confidence intervals were interpreted as evidence of significant changes in heritability estimates.

Summarizations of the SDOH Survey

Survey Preparation

Survey items were organized into themes (i.e. conceptual domains grouping related questions, such as "perceived stress," "social support," "housing instability," etc.) based on the SDOH Task Force's consensus-based conceptual frameworks. Missing responses coded as "PMI: Skip", "PMI: Don't Know", "PMI: Not Sure", "PMI: Other" or "PMI: Prefer Not To Answer" were excluded. At the question-level, missingness was very low (<5%) for nearly all themed items(Table S7), with only two un-themed "orphan" questions exhibiting high skip rates. Theme-level missingness was similar across race groups (2–8%) (Table S5), and at the individual level the median per-participant missingness was $\approx 3\%$, with $< 1\%$ of participants missing $> 20\%$ of items.

PCA & MCA

Each categorical SDOH survey question was first one-hot encoded into a set of binary dummy variables. Principal component analysis (PCA) and multiple correspondence analysis (MCA) were then applied to these indicators in order to extract a lower dimensional summary of survey variation. The leading SDOH components—that is, the PCs from PCA (Fig. S2) and the dimensions from MCA (Fig. S3), capture the major axes of social-determinant heterogeneity in our cohort and were included as linear covariates in downstream heritability analyses. SDOH PCs and MCA axes were calculated from all individuals with SDOH survey data. In addition, genetic PCs were calculated on all unrelated individuals with array data (Fig. S1).

Environmental Relatedness Matrix (ERM)

In addition to the summarizations above, we construct an environmental relatedness matrix (ERM) from the same one-hot encoded data. In this approach, every individual's binary response vector forms a row of the data matrix

$$X \in \{0,1\}^{n \times p},$$

where n is the number of individuals and p is the total number of response-level columns.

To ensure that each question contributes equally regardless of how many response levels it has, we form a diagonal weight matrix

$$W = \operatorname{diag}(w_1, \ldots, w_p),$$

with

$$w_\ell = \frac{1}{\left|\{\text{levels of question corresponding to column } \ell\}\right|}.$$

We then define the weighted design matrix

$$X_w = X\,W,$$

so that

$$(X_w)_{i\ell} = w_\ell\, X_{i\ell}.$$

Next, to account for missing responses, we build a binary mask matrix

$$M \in \{0,1\}^{n \times p},$$

by setting

$$M_{i\ell} = \begin{cases} 1, & \text{if individual } i \text{ answered level column } \ell, \\ 0, & \text{otherwise,} \end{cases}$$

and its weighted version

$$M_w = M\,W.$$

Rather than ever forming the full $n \times n$ matrix, each row-sum

$$s_i = \sum_{j=1}^{n} \mathrm{ERM}_{ij}$$

can be computed in $O(n\,p)$ time by first computing the two vectors

$$c = X_w^\top \mathbf{1}_n, \qquad d = M^\top \mathbf{1}_n,$$

and then

$$u = X_w\, c, \qquad v = M_w\, d.$$

Each individual's ERM row sum is then

$$s_i = \frac{u_i}{v_i}, \quad i = 1, \ldots, n,$$

with any indeterminate 0/0 set to zero (treating pairs with no overlapping responses as unrelated). The resulting row-sum s_i quantifies the total environmental relatedness of individual i to the cohort; before inclusion as a covariate in downstream regression models, we standardize these row-sums to have mean zero and unit variance. All matrix operations are carried out on the sparse weighted matrices X_w and M_w. Although $\mathrm{ERM}_{ii} = 1$ holds by definition, we never explicitly compute or store diagonal entries, as they are not required for the ERM row sum calculation.

4 Discussion

Our results show that SDOH structure as summarized by dimension-reduced axes (via MCA and PCA) reduces heritability estimates using HE regression across multiple traits. However, a pairwise similarity measure (the ERM row sum) does not. In White individuals, including SDOH covariates significantly decreased heritability estimates for 4 of 18 traits (BMI, hip circumference, waist circumference, and HDL cholesterol), but not for the other traits nor in Black/African American individuals, possibly due to their smaller sample size. Including genetic PCs consistently reduced heritability estimates in both groups, indicating the presence of population structure and underscoring the necessity of including genetic PCs when SDOH summaries are included.

We note that sample size differences and our methodological usage of HE regression and self-identified race have important potential consequences. The size imbalance between racial groups is a major limitation of this study: the Black/African American cohort (n=6,538) is an order of magnitude smaller than the White cohort (n=67,545). The absence of significant SDOH covariate effects in the Black/African American cohort likely reflects insufficient power rather than a true absence of SDOH influence on heritability. In addition, between-group comparisons are not meaningful because true population differences are indistinguishable from induced statistical artifacts.

Methodologically, our choice of HE regression, while computationally efficient, may yield less precise heritability estimates compared to GREML, particularly in smaller

subgroups. This compounds the sample size issue as more precise estimation methods might detect SDOH effects that HE regression cannot capture. While we stratified by self-identified race, we acknowledge that race functions as a social construct that imperfectly captures both genetic ancestry and social experiences. Our stratified analyses risks conflating genetic background with socially mediated SDOH exposures, and our findings should not be interpreted as evidence of biological differences between racial groups. To address these limitations, we are conducting additional analyses using alternative heritability estimation methods, such as GREML, and complementary analytical approaches using balanced sample sizes. We are also investigating which factors most substantially influence our findings, with particular focus on traits showing the greatest sensitivity to the inclusion of SDOH covariates.

These results highlight the influence of both genetic and social structures on trait heritability and demonstrate how accounting for social factors can alter trait-specific heritability estimates.

5 Data availability

Individual-level genotype and survey data used in this study are available to registered researchers through the All of Us Researcher Workbench (https://workbench.researchallofus.org/) following Data and Research Center (DRC) access procedures. Genomic data (v7 array release) and survey responses, including the Social Determinants of Health (SDOH) survey, are part of the All of Us [10] controlled tier dataset. Other information is available upon request.

6 Code availability

All analyses were performed using previously published or developed tools, as indicated in the Methods.

Acknowledgments. We thank the All of Us Research Program and its participants. We thank Dylan Ong for helpful comments that improved the manuscript. This research was supported by NHGRI CEGS RM1: RM1HG010461 (NSA) and NIH NCI U54CA267738 (SAM) and The Cornell University Graduate School Deans Excellence Fellowship (OSRA).

Declarations. The authors declare no competing interests.

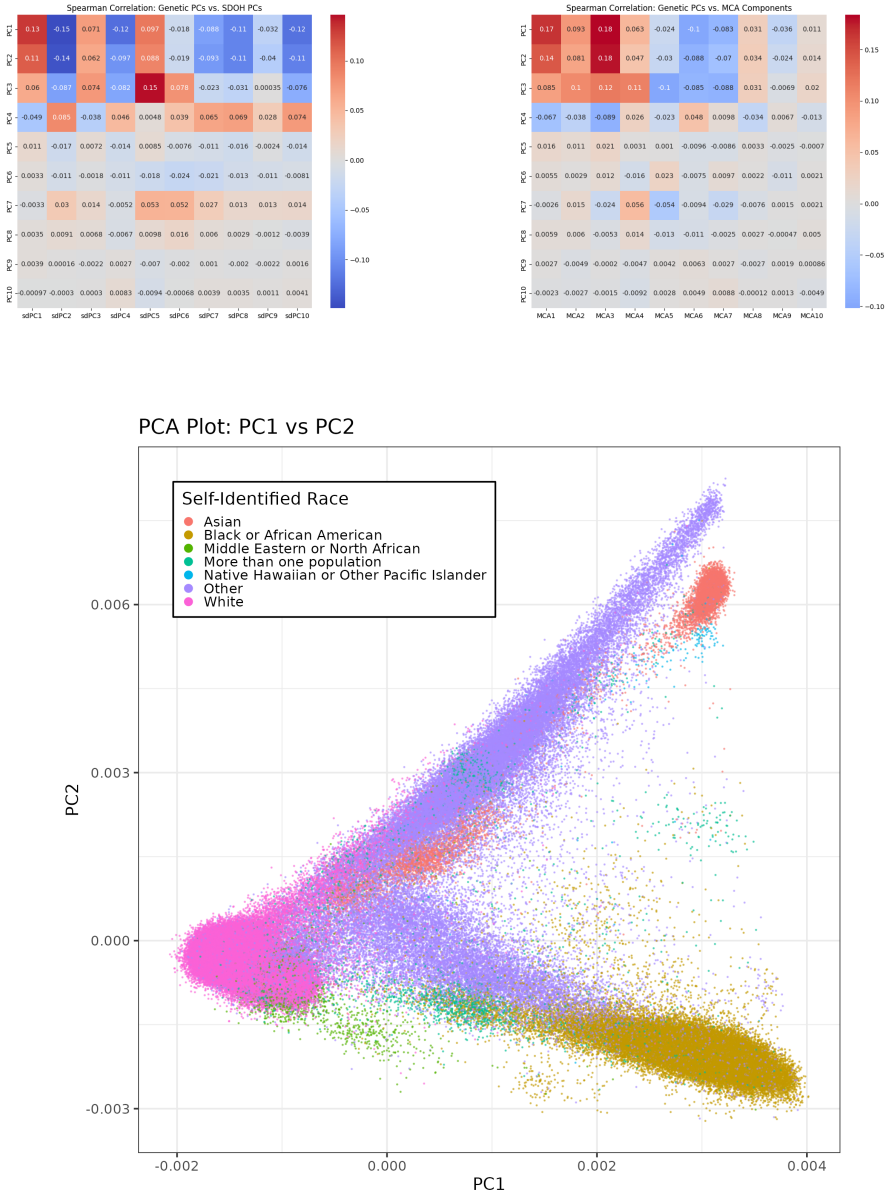

Fig. 1: Relationship between genetic structure and SDOH summaries. Panels (a) and (b) display heatmaps showing Spearman correlations between the top 10 genetic PCs and the first 10 SDOH axes derived from PCA and MCA respectively. Panel (c) shows a PCA plot of genetic data with the x- and y-axes representing genetic PCs 1 and 2 which explain 36.8% and 21.5% of the total variance, respectively. Each point represents an individual, colored by self-identified race.

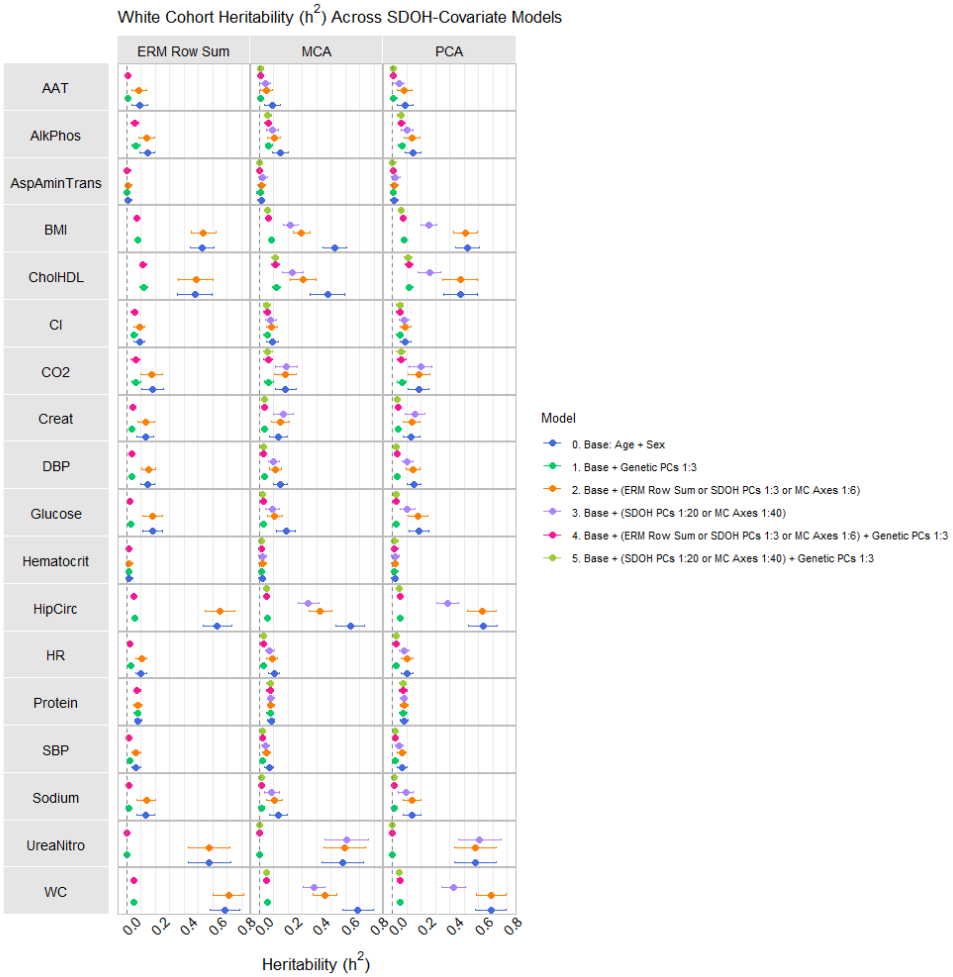

Fig. 2: White individuals: Haseman–Elston (HE) regression heritability estimates for 18 traits (gray rows) utilizing one of the following SDOH summarization strategies (gray columns): ERM row sum, which was run with one of four possible models (Models 0, 1, 2, and 4); MCA, run with one of six models (Models 0-5); or PCA, run with one of six models (Models 0-5). Points are HE-CP estimates, intervals are 95% CIs based on HE jackknife standard errors. Traits: alanine aminotransferase (AAT), alkaline phosphatase (AlkPhos), aspartate aminotransferase (AspAminTrans), body mass index (BMI), carbon dioxide (CO2), chloride (Cl), cholesterol in HDL (CholHDL), creatinine (Creat), diastolic blood pressure (DBP), glucose, heart rate (HR), hematocrit, hip circumference (HipCirc), protein, sodium, systolic blood pressure (SBP), urea nitrogen (UreaNitro), and waist circumference (WC).

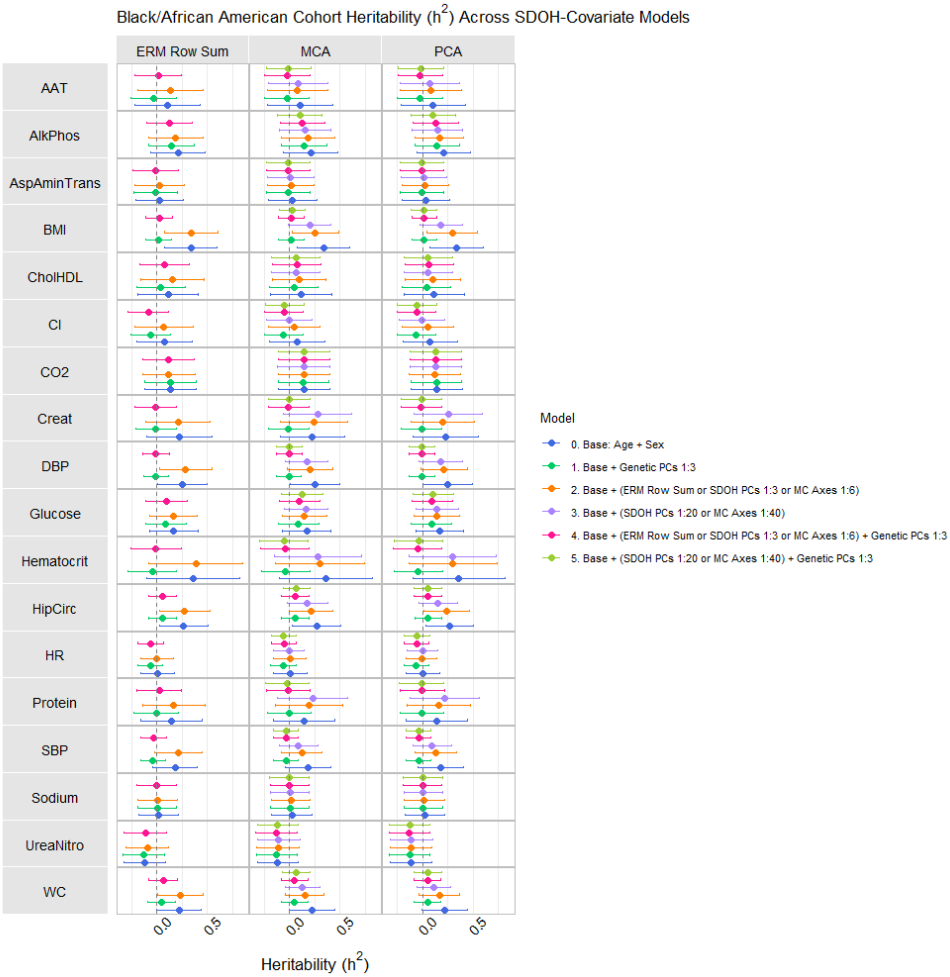

Fig. 3: Black/African American individuals: Haseman–Elston (HE) regression heritability estimates for 18 traits (gray rows) utilizing one of the following SDOH summarization strategies (gray columns): ERM row sum, which was run with one of four possible models (Models 0, 1, 2, and 4); MCA, run with one of six models (Models 0-5); or PCA, run with one of six models (Models 0-5). Points are HE-CP estimates, intervals are 95% CIs based on HE jackknife standard errors. Traits: alanine aminotransferase (AAT), alkaline phosphatase (AlkPhos), aspartate aminotransferase (AspAminTrans), body mass index (BMI), carbon dioxide (CO2), chloride (Cl), cholesterol in HDL (CholHDL), creatinine (Creat), diastolic blood pressure (DBP), glucose, heart rate (HR), hematocrit, hip circumference (HipCirc), protein, sodium, systolic blood pressure (SBP), urea nitrogen (UreaNitro), and waist circumference (WC).

References

[1] The "All of Us" Research Program | New England Journal of Medicine.

[2] Hailan Liu and Guo-Bo Chen. A novel genomic prediction method combining randomized Haseman-Elston regression with a modified algorithm for Proven and Young for large genomic data. *The Crop Journal*, 10(2):550–554, April 2022. ISSN 2214-5141. doi: 10.1016/j.cj.2021.09.001.

[3] Ali Pazokitoroudi, Zhengtong Liu, Andrew Dahl, Noah Zaitlen, Saharon Rosset, and Sriram Sankararaman. A scalable and robust variance components method reveals insights into the architecture of gene-environment interactions underlying complex traits. *The American Journal of Human Genetics*, 111(7): 1462–1480, July 2024. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2024.05.015. Publisher: Elsevier.

[4] Suresh K. Bhavnani, Weibin Zhang, Daniel Bao, Mukaila Raji, Veronica Ajewole, Rodney Hunter, Yong-Fang Kuo, Susanne Schmidt, Monique R. Pappadis, Elise Smith, Alex Bokov, Timothy Reistetter, Shyam Visweswaran, and Brian Downer. Subtyping Social Determinants of Health in All of Us: Network Analysis and Visualization Approach. *medRxiv*, page 2023.01.27.23285125, August 2023. doi: 10.1101/2023.01.27.23285125.

[5] Sonali Gupta, Vincent Lam, I. King Jordan, and Leonardo Mariño-Ramírez. A composite socioeconomic deprivation index from All of Us survey data: associations with health outcomes and disparities. *medRxiv*, page 2024.10.04.24314904, October 2024. doi: 10.1101/2024.10.04.24314904.

[6] Cole Brokamp, Andrew F. Beck, Neera K. Goyal, Patrick Ryan, James M. Greenberg, and Eric S. Hall. Material community deprivation and hospital utilization during the first year of life: an urban population–based cohort study. *Annals of Epidemiology*, 30:37–43, February 2019. ISSN 1047-2797. doi: 10.1016/j.annepidem.2018.11.008.

[7] Second-generation PLINK: rising to the challenge of larger and richer datasets | GigaScience | Oxford Academic.

[8] Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, November 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq559.

[9] GCTA: A Tool for Genome-wide Complex Trait Analysis - PMC.

[10] Alexander G. Bick, Ginger A. Metcalf, Kelsey R. Mayo, Lee Lichtenstein, Shimon Rura, Robert J. Carroll, Anjene Musick, Jodell E. Linder, I. King Jordan, Shashwat Deepali Nagar, Shivam Sharma, Robert Meller, Melissa Basford, Eric

Boerwinkle, Mine S. Cicek, Kimberly F. Doheny, Evan E. Eichler, Stacey Gabriel, Richard A. Gibbs, David Glazer, Paul A. Harris, Gail P. Jarvik, Anthony Philippakis, Heidi L. Rehm, Dan M. Roden, Stephen N. Thibodeau, Scott Topper, Ashley L. Blegen, Samantha J. Wirkus, Victoria A. Wagner, Jeffrey G. Meyer, Mine S. Cicek, Donna M. Muzny, Eric Venner, Michelle Z. Mawhinney, Sean M. L. Griffith, Elvin Hsu, Hua Ling, Marcia K. Adams, Kimberly Walker, Jianhong Hu, Harsha Doddapaneni, Christie L. Kovar, Mullai Murugan, Shannon Dugan, Ziad Khan, Eric Boerwinkle, Niall J. Lennon, Christina Austin-Tse, Eric Banks, Michael Gatzen, Namrata Gupta, Emma Henricks, Katie Larsson, Sheli McDonough, Steven M. Harrison, Christopher Kachulis, Matthew S. Lebo, Cynthia L. Neben, Marcie Steeves, Alicia Y. Zhou, Joshua D. Smith, Christian D. Frazar, Colleen P. Davis, Karynne E. Patterson, Marsha M. Wheeler, Sean McGee, Christina M. Lockwood, Brian H. Shirts, Colin C. Pritchard, Mitzi L. Murray, Valeria Vasta, Dru Leistritz, Matthew A. Richardson, Jillian G. Buchan, Aparna Radhakrishnan, Niklas Krumm, Brenna W. Ehmen, Sophie Schwartz, M. Morgan T. Aster, Kristian Cibulskis, Andrea Haessly, Rebecca Asch, Aurora Cremer, Kylee Degatano, Akum Shergill, Laura D. Gauthier, Samuel K. Lee, Aaron Hatcher, George B. Grant, Genevieve R. Brandt, Miguel Covarrubias, Eric Banks, Ashley Able, Ashley E. Green, Robert J. Carroll, Jennifer Zhang, Henry R. Condon, Yuanyuan Wang, Moira K. Dillon, C. H. Albach, Wail Baalawi, Seung Hoan Choi, Xin Wang, Elisabeth A. Rosenthal, Andrea H. Ramirez, Sokny Lim, Siddhartha Nambiar, Bradley Ozenberger, Anastasia L. Wise, Chris Lunt, Geoffrey S. Ginsburg, Joshua C. Denny, The All of Us Research Program Genomics Investigators, Manuscript Writing Group, All of Us Research Program Genomics Principal Investigators, Mayo Biobank, Genome Center: Baylor-Hopkins Clinical Genome Center, Color Genome Center: Broad, and Mass General Brigham Laboratory for Molecular Medicine, Genome Center: University of Washington, Data and Research Center, All of Us Research Demonstration Project Teams, and NIH All of Us Research Program Staff. Genomic data in the All of Us Research Program. *Nature*, 627(8003):340–346, March 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06957-x. Publisher: Nature Publishing Group.

Supplemental Material

Table S1: Counts by Age Group

	age_group	count
1	18-29	3514
2	30-39	7746
3	40-49	8834
4	50-59	11571
5	60-69	18711
6	70-79	21679
7	80-89	6680
8	90+	615
9	NA	8666

Table S2: Counts by Self-Identified Race

Category	Count
Black or African American	6583
Asian	2323
Middle Eastern or North African	383
More than one population	1522
Native Hawaiian or Other Pacific Islander	44
White	67545

Table S3: Missingness by "Theme" and Age

Theme	Age Group	Total Responses	Missing	% Missing
discrimination_healthcare	<30	29407	570	1.9
discrimination_healthcare	30–50	130004	2184	1.7
discrimination_healthcare	50–70	241822	6200	2.6
discrimination_healthcare	>70	214879	8129	3.8
everyday_discrimination	<30	37809	583	1.5
everyday_discrimination	30–50	167148	2408	1.4
everyday_discrimination	50–70	310914	7221	2.3
everyday_discrimination	>70	276273	10419	3.8
food_insecurity	<30	8402	110	1.3
food_insecurity	30–50	37144	295	0.8
food_insecurity	50–70	69092	564	0.8
food_insecurity	>70	61394	653	1.1
housing_insecurity	<30	4201	80	1.9
housing_insecurity	30–50	18572	313	1.7
housing_insecurity	50–70	34546	1045	3.0
housing_insecurity	>70	30697	1203	3.9
housing_instability	<30	4201	93	2.2
housing_instability	30–50	18572	386	2.1
housing_instability	50–70	34546	1356	3.9
housing_instability	>70	30697	2759	9.0
loneliness	<30	33608	432	1.3
loneliness	30–50	148576	1674	1.1
loneliness	50–70	276368	5187	1.9
loneliness	>70	245576	6914	2.8
neighborhood_cohesion	<30	16804	144	0.9
neighborhood_cohesion	30–50	74288	725	1.0
neighborhood_cohesion	50–70	138184	2415	1.7
neighborhood_cohesion	>70	122788	3943	3.2
neighborhood_disorder	<30	54613	635	1.2
neighborhood_disorder	30–50	241436	3797	1.6
neighborhood_disorder	50–70	449098	12680	2.8
neighborhood_disorder	>70	399061	17831	4.5
neighborhood_environment	<30	33608	2004	6.0
neighborhood_environment	30–50	148576	8286	5.6
neighborhood_environment	50–70	276368	19252	7.0
neighborhood_environment	>70	245576	21933	8.9
perceived_stress	<30	42010	888	2.1
perceived_stress	30–50	185720	3684	2.0
perceived_stress	50–70	345460	12895	3.7
perceived_stress	>70	306970	19652	6.4
religious_attendance	<30	4201	308	7.3
religious_attendance	30–50	18572	1472	7.9
religious_attendance	50–70	34546	4234	12.3
religious_attendance	>70	30697	4702	15.3
religious_spirituality	<30	25206	428	1.7
religious_spirituality	30–50	111432	1406	1.3
religious_spirituality	50–70	207276	2933	1.4
religious_spirituality	>70	184182	3751	2.0
social_support	<30	33608	431	1.3
social_support	30–50	148576	1612	1.1
social_support	50–70	276368	5549	2.0
social_support	>70	245576	8353	3.4

Table S5: Missingness by Theme and Race. Rows containing less than 20 individuals removed as per All of Us data usage agreement.

Theme	Race	Total Responses	Missing	% Missing
discrimination_healthcare	Race_Black_or_African_American	46081	2079	4.5
discrimination_healthcare	Race_Other	67312	2909	4.3
discrimination_healthcare	Race_Middle_Eastern_or_North_African	2681	88	3.3
discrimination_healthcare	Race_White	472815	11513	2.4
discrimination_healthcare	Race_Asian	16261	299	1.8
discrimination_healthcare	Race_More_than_one_population	10654	187	1.8
everyday_discrimination	Race_Black_or_African_American	59247	2621	4.4
everyday_discrimination	Race_Other	86544	3538	4.1
everyday_discrimination	Race_White	607905	13792	2.3
everyday_discrimination	Race_Asian	20907	383	1.8
everyday_discrimination	Race_More_than_one_population	13698	223	1.6
everyday_discrimination	Race_Middle_Eastern_or_North_African	3447	55	1.6
food_insecurity	Race_Black_or_African_American	13166	274	2.1
food_insecurity	Race_Other	19232	339	1.8
food_insecurity	Race_Asian	4646	52	1.1
food_insecurity	Race_White	135090	927	0.7
housing_insecurity	Race_Black_or_African_American	6583	387	5.9
housing_insecurity	Race_Other	9616	518	5.4
housing_insecurity	Race_Asian	2323	60	2.6
housing_insecurity	Race_White	67545	1625	2.4
housing_insecurity	Race_More_than_one_population	1522	29	1.9
housing_instability	Race_White	67545	3668	5.4
housing_instability	Race_Other	9616	506	5.3
housing_instability	Race_Black_or_African_American	6583	259	3.9
housing_instability	Race_Asian	2323	90	3.9
housing_instability	Race_More_than_one_population	1522	48	3.2
loneliness	Race_Black_or_African_American	52664	1865	3.5
loneliness	Race_Other	76928	2698	3.5
loneliness	Race_Middle_Eastern_or_North_African	3064	53	1.7
loneliness	Race_White	540360	9128	1.7
loneliness	Race_Asian	18584	292	1.6
loneliness	Race_More_than_one_population	12176	162	1.3
neighborhood_cohesion	Race_Black_or_African_American	26332	833	3.2
neighborhood_cohesion	Race_Other	38464	1155	3.0
neighborhood_cohesion	Race_White	270180	5063	1.9
neighborhood_cohesion	Race_Middle_Eastern_or_North_African	1532	22	1.4
neighborhood_cohesion	Race_More_than_one_population	6088	65	1.1
neighborhood_cohesion	Race_Asian	9292	81	0.9
neighborhood_disorder	Race_Other	125008	5532	4.4
neighborhood_disorder	Race_Black_or_African_American	85579	3776	4.4
neighborhood_disorder	Race_White	878085	24622	2.8
neighborhood_disorder	Race_Middle_Eastern_or_North_African	4979	131	2.6
neighborhood_disorder	Race_Asian	30199	542	1.8
neighborhood_disorder	Race_More_than_one_population	19786	319	1.6
neighborhood_environment	Race_Black_or_African_American	52664	4794	9.1
neighborhood_environment	Race_Other	76928	6401	8.3
neighborhood_environment	Race_Middle_Eastern_or_North_African	3064	220	7.2
neighborhood_environment	Race_White	540360	38320	7.1
neighborhood_environment	Race_More_than_one_population	12176	726	6.0
neighborhood_environment	Race_Asian	18584	982	5.3
perceived_stress	Race_Black_or_African_American	65830	3945	6.0
perceived_stress	Race_Other	96160	5316	5.5
perceived_stress	Race_Middle_Eastern_or_North_African	3830	173	4.5
perceived_stress	Race_White	675450	26568	3.9
perceived_stress	Race_Asian	23230	728	3.1
perceived_stress	Race_More_than_one_population	15220	369	2.4
religious_attendance	Race_Black_or_African_American	6583	942	14.3
religious_attendance	Race_White	67545	8405	12.4
religious_attendance	Race_Middle_Eastern_or_North_African	383	46	12.0
religious_attendance	Race_Other	9616	1039	10.8
religious_attendance	Race_More_than_one_population	1522	114	7.5
religious_attendance	Race_Asian	2323	165	7.1
religious_spirituality	Race_Black_or_African_American	39498	1127	2.9
religious_spirituality	Race_Other	57696	1502	2.6
religious_spirituality	Race_Middle_Eastern_or_North_African	2298	58	2.5
religious_spirituality	Race_White	405270	5522	1.4
religious_spirituality	Race_More_than_one_population	9132	120	1.3
religious_spirituality	Race_Asian	13938	180	1.3
social_support	Race_Black_or_African_American	52664	1987	3.8
social_support	Race_Other	76928	2744	3.6
social_support	Race_Middle_Eastern_or_North_African	3064	94	3.1
social_support	Race_White	540360	10638	2.0
social_support	Race_Asian	18584	333	1.8
social_support	Race_More_than_one_population	12176	144	1.2

Table S7: Overall Missingness by Theme

Theme	Total Responses	Missing	% Missing
religious_attendance	88016	10716	12.2
neighborhood_environment	704128	51475	7.3
housing_instability	88016	4594	5.2
perceived_stress	880160	37119	4.2
neighborhood_disorder	1144208	34943	3.1
housing_insecurity	88016	2641	3.0
discrimination_healthcare	616112	17083	2.8
everyday_discrimination	792144	20631	2.6
social_support	704128	15945	2.3
neighborhood_cohesion	352064	7227	2.1
loneliness	704128	14207	2.0
religious_spirituality	528096	8518	1.6
food_insecurity	176032	1622	0.9

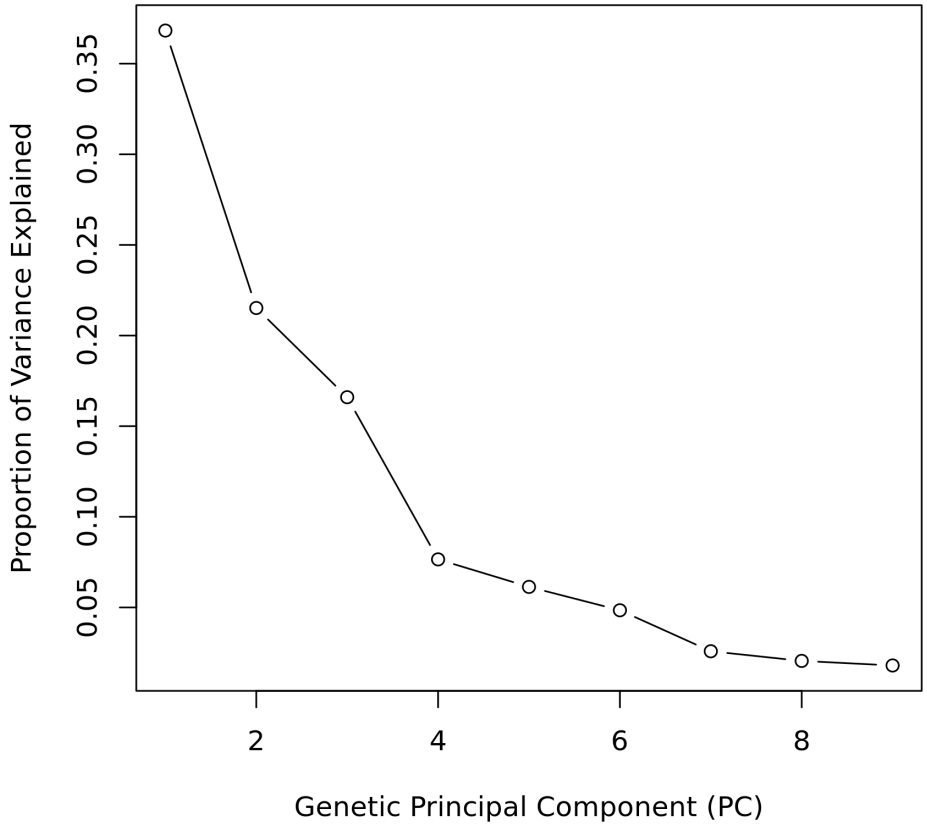

Fig. S1: Genetic PCA scree plot

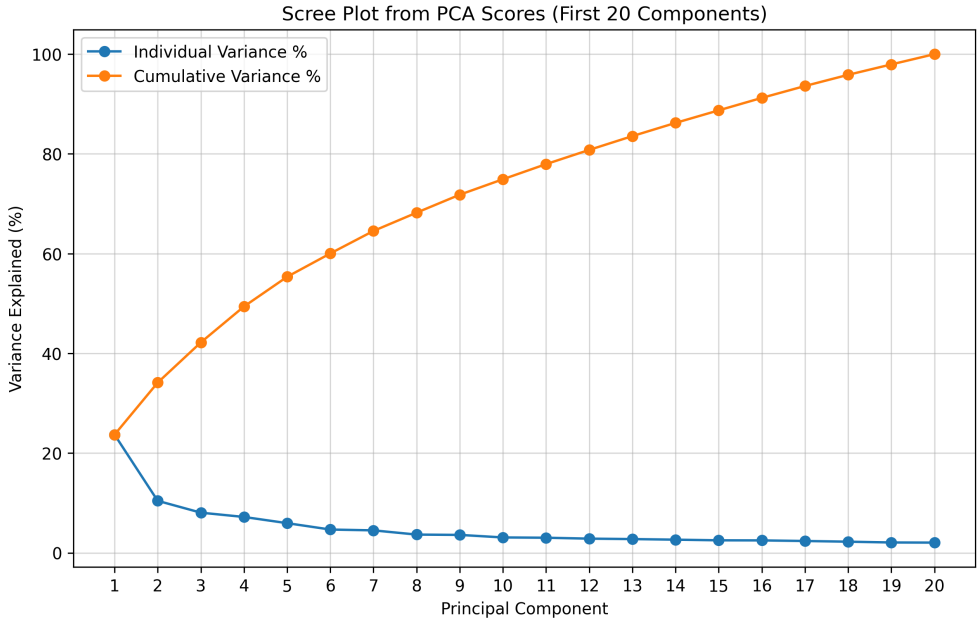

Fig. S2: SDOH PCA scree plot

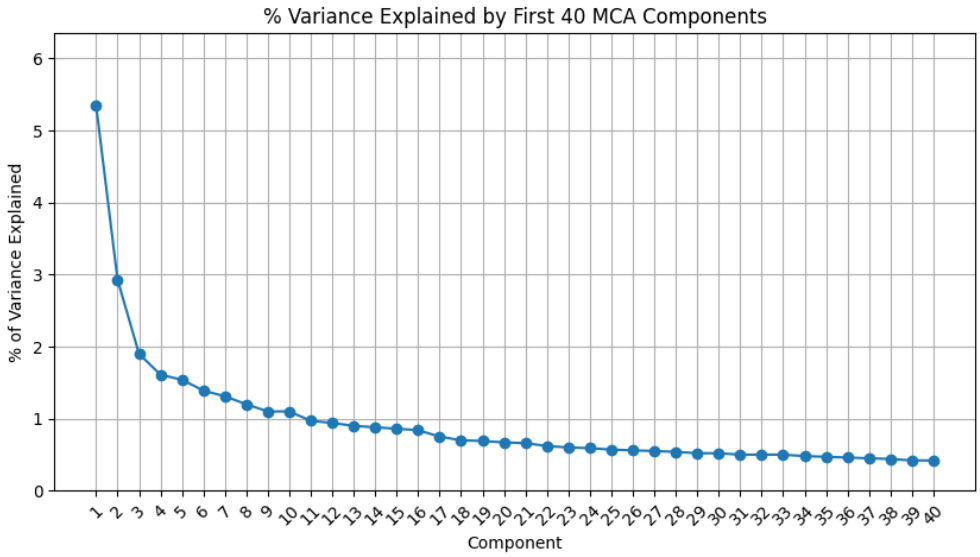

Fig. S3: SDOH MCA scree plot

Table S8: Black/AA and White combined Genetic Prediction

GEN_PC	R_squared	F_stat	F_pvalue
GENPC1	0.184	992.652	0.000
GENPC2	0.026	118.039	0.000
GENPC3	0.035	159.044	0.000
GENPC4	0.031	140.070	0.000
GENPC5	0.010	43.227	0.000
GENPC6	0.015	65.924	0.000
GENPC7	0.004	19.109	0.000
GENPC8	0.006	25.444	0.000
GENPC9	0.006	26.041	0.000
GENPC10	0.003	12.362	0.000

Table S9: Black/AA and White combined SDOH Prediction

SDOH_PC	R_squared	F_stat	F_pvalue
SDOHPC1	0.026	236.895	0.000
SDOHPC2	0.041	375.185	0.000
SDOHPC3	0.015	136.914	0.000
SDOHPC4	0.029	258.842	0.000
SDOHPC5	0.062	584.787	0.000
SDOHPC6	0.019	165.936	0.000
SDOHPC7	0.015	131.894	0.000
SDOHPC8	0.021	191.462	0.000
SDOHPC9	0.004	32.169	0.000
SDOHPC10	0.029	259.003	0.000
SDOHPC11	0.003	27.209	0.000
SDOHPC12	0.014	125.033	0.000
SDOHPC13	0.010	86.146	0.000
SDOHPC14	0.006	52.523	0.000
SDOHPC15	0.006	49.701	0.000
SDOHPC16	0.004	33.414	0.000
SDOHPC17	0.005	41.492	0.000
SDOHPC18	0.002	21.156	0.000
SDOHPC19	0.012	111.350	0.000
SDOHPC20	0.002	16.871	0.000

Table S10: Black/AA Genetic Prediction

GEN_PC	R_squared	F_stat	F_pvalue
GENPC1	0.031	10.519	0.000
GENPC2	0.013	4.409	0.000
GENPC3	0.008	2.476	0.000
GENPC4	0.002	0.703	0.827
GENPC5	0.004	1.139	0.300
GENPC6	0.009	2.811	0.000
GENPC7	0.007	2.236	0.001
GENPC8	0.003	0.827	0.682
GENPC9	0.002	0.775	0.747
GENPC10	0.004	1.306	0.162

Table S11: Black/AA SDOH Prediction

SDOH_PC	R_squared	F_stat	F_pvalue
SDOHPC1	0.003	2.080	0.023
SDOHPC2	0.005	3.428	0.000
SDOHPC3	0.008	5.358	0.000
SDOHPC4	0.003	1.925	0.037
SDOHPC5	0.004	2.859	0.001
SDOHPC6	0.005	3.385	0.000
SDOHPC7	0.001	0.769	0.659
SDOHPC8	0.004	2.522	0.005
SDOHPC9	0.002	1.558	0.113
SDOHPC10	0.003	1.765	0.061
SDOHPC11	0.002	1.186	0.295
SDOHPC12	0.004	2.574	0.004
SDOHPC13	0.002	1.166	0.308
SDOHPC14	0.004	2.486	0.006
SDOHPC15	0.002	1.167	0.308
SDOHPC16	0.004	2.286	0.011
SDOHPC17	0.002	1.345	0.200
SDOHPC18	0.001	0.795	0.633
SDOHPC19	0.002	1.114	0.347
SDOHPC20	0.002	1.289	0.230

Table S12: White Genetic Prediction

GEN_PC	R_squared	F_stat	F_pvalue
GENPC1	0.021	69.574	0.000
GENPC2	0.009	28.679	0.000
GENPC3	0.036	123.472	0.000
GENPC4	0.005	16.352	0.000
GENPC5	0.021	70.146	0.000
GENPC6	0.008	28.356	0.000
GENPC7	0.006	20.239	0.000
GENPC8	0.003	10.209	0.000
GENPC9	0.008	27.795	0.000
GENPC10	0.001	4.864	0.000

Table S13: White SDOH Prediction

SDOH_PC	R_squared	F_stat	F_pvalue
SDOHPC1	0.002	13.339	0.000
SDOHPC2	0.005	31.238	0.000
SDOHPC3	0.001	9.555	0.000
SDOHPC4	0.003	21.540	0.000
SDOHPC5	0.011	72.907	0.000
SDOHPC6	0.013	89.254	0.000
SDOHPC7	0.003	18.598	0.000
SDOHPC8	0.004	28.064	0.000
SDOHPC9	0.003	17.211	0.000
SDOHPC10	0.002	13.380	0.000
SDOHPC11	0.001	7.952	0.000
SDOHPC12	0.006	37.549	0.000
SDOHPC13	0.001	9.072	0.000
SDOHPC14	0.002	13.690	0.000
SDOHPC15	0.001	3.805	0.000
SDOHPC16	0.000	2.429	0.007
SDOHPC17	0.001	9.486	0.000
SDOHPC18	0.002	13.773	0.000
SDOHPC19	0.001	9.650	0.000
SDOHPC20	0.001	8.890	0.000

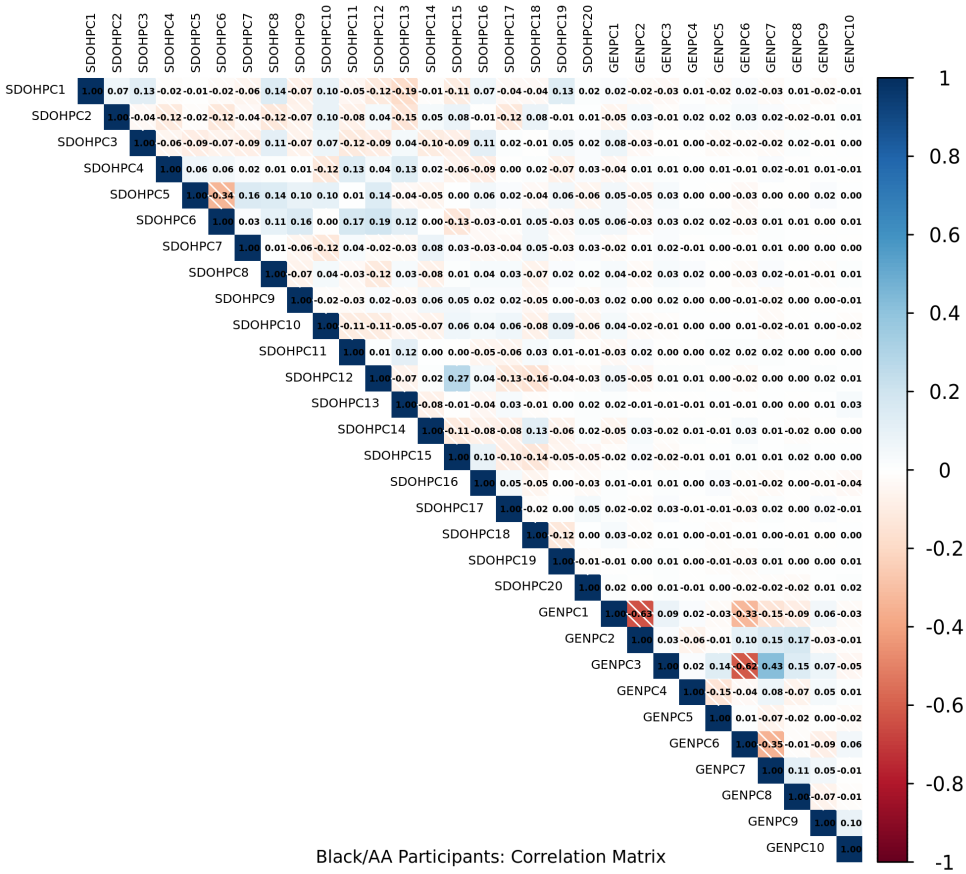

Fig. S4: PC Pearson correlation matrix for Black/African American individuals

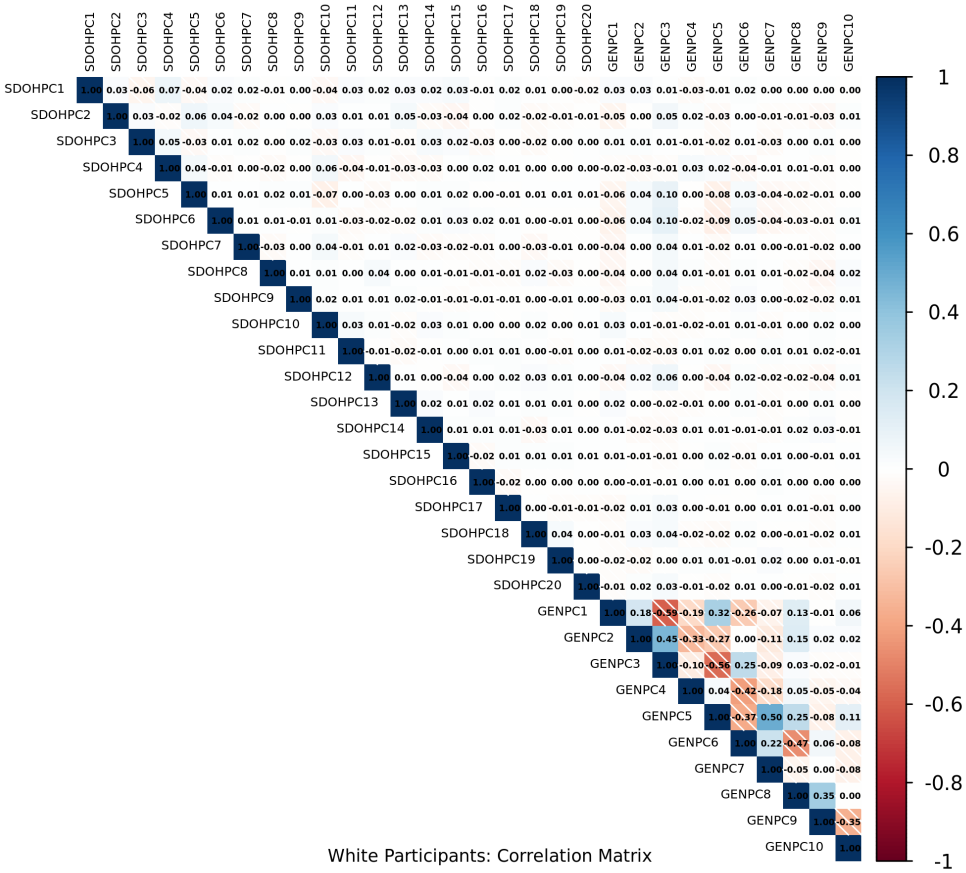

Fig. S5: PC Pearson correlation matrix for White individuals

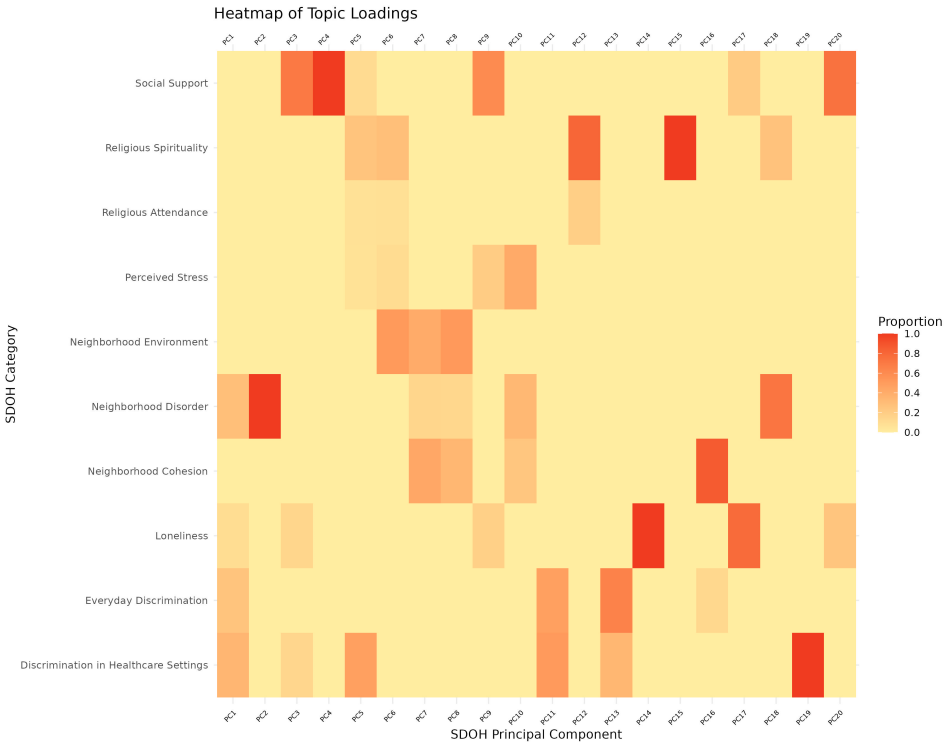

Fig. S6: Heatmap of SDOH topic loadings for SDOH PCs

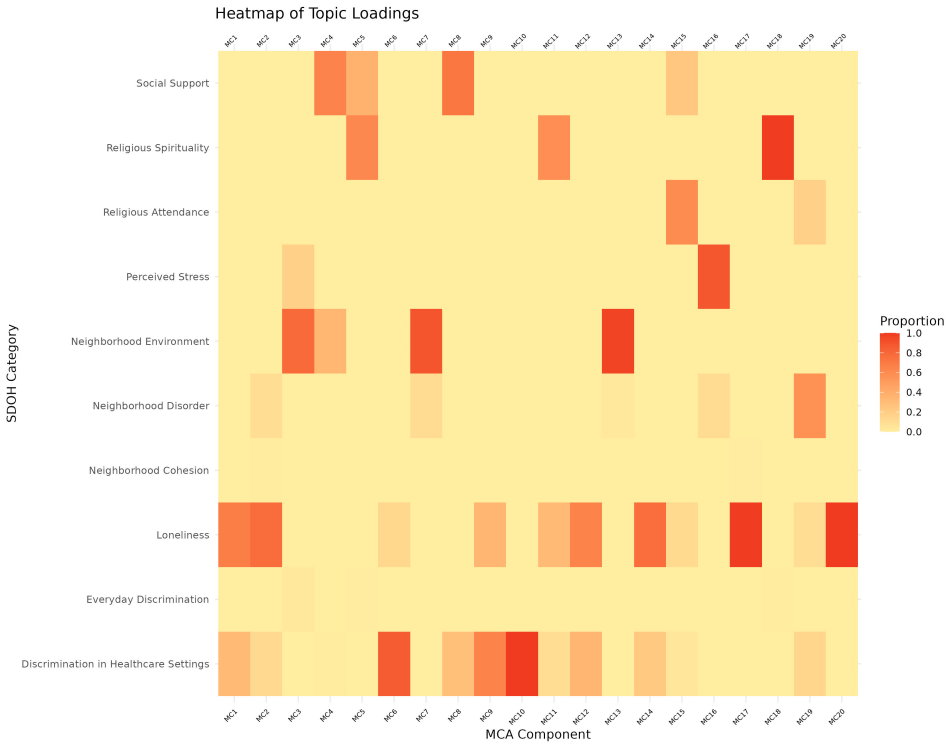

Fig. S7: Heatmap of SDOH topic loadings for SDOH MCA axes.