Phenomic environment-wide association study (PheEWAS) models complexity in the exposome

Nicole E. Palmiero[1], Tomás González Zarzar[2,3], Kristin Passero[4], Jiayan Zhou[5], Dana C. Crawford[6], and Molly A. Hall[1*]

1. Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
2. Neurology Department, School of Medicine, Pontificia Universidad Católica de Chile, Santiago, Chile
3. Institute for Biological and Medical Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile
4. Virginia Institute of Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, USA.
5. Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA, USA
6. Departments of Population and Quantitative Health Sciences, and Genetics and Genome Sciences, Cleveland Institute for Computational Biology, Case Western Reserve University (CWRU), Cleveland, OH, USA

*** Corresponding author**

Phenome-wide association studies (PheWAS) have successfully identified genomic-based interrelationships between phenotypes but seldom consider environmental exposures. Here, in a phenomic environment-wide association study (PheEWAS), we interrogated relationships between 326 exposures and 55 phenotypes for ~19,000 participants of the National Health and Nutrition Examination Survey (NHANES). Linear regression models adjusted for age, sex, socioeconomic status, BMI, race/ethnicity, and survey year identified and replicated 106 significant exposure–phenotype associations after Bonferroni correction. The top association was for alpha-tocopherol (vitamin E) with triglycerides (Discovery $p = 1.16 \times 10^{-11}$; Replication $p = 8.05 \times 10^{-13}$). The exposure retinol (vitamin A) had the largest number of individual replicating associations (14 phenotypes including total calcium, iron-binding capacity, ferritin, albumin, transferrin saturation, creatinine, gamma-glutamyl transferase, triglycerides, uric acid, alkaline phosphatase, hemoglobin, and blood urea nitrogen). The phenotype with the greatest number of exposure associations was homocysteine (associated with thiamine; alpha- and gamma-tocopherol; dietary fiber, protein, and potassium; riboflavin; cotinine; folate; phosphorus; cadmium; iron intake; supplement count; and niacin). A race/ethnicity-stratified analysis revealed 11 unique population-specific associations. Our findings demonstrate PheEWAS a method to provide new details on the complexity of the exposome at the level of the phenome

INTRODUCTION

The association of genomic variants with multiple phenotypes using phenome-wide association studies (PheWAS) has identified pleiotropic associations and other phenomic relationships using

genetic data from epidemiological studies and clinical trials[1–4]. Applying a phenome-wide approach to genomics research broadens the context of genetic findings and uncovers new information on the development of human diseases[5]. Although exposome research has typically explored phenotypes individually, phenome-wide exposure studies have provided new information on correlations between multiple phenotypes and a single environmental exposure[6,7]. Here, we present a novel approach for investigating the exposome (i.e., the lifetime environmental exposures of an individual): phenomic environment-wide association studies (PheEWAS).

The potential to detect pleiotropy through cross-phenotype associations is a key advantage of PheWAS, as it identifies genetic variants with wide-reaching health impacts. For example, PheWAS have identified genes simultaneously related to lipoprotein metabolism, vitamin E levels, folate levels, and triglycerides[2]. However, exogenous factors such as pollutants, allergens, toxicants, or infectious agents also contribute to human traits and disease[8,9]. Environment-wide association studies (EWAS), which are used to determine the effect of the exposome on human health, have identified exposures associated with type 2 diabetes, body mass index (BMI), blood pressure, and others[10,11]. Identifying exposures with multiple phenotypic effects (i.e., “cross-phenotype exposures”) can reveal critical environmental risk factors with far-reaching impacts on disease outcomes. For example, exposure to heavy metals contributes to multiple diseases with varied phenotypic effects[12,13]. Extending EWAS to the phenome level offers the opportunity to discover new insights into the far-reaching effects of the environment on human health. PheEWAS tests associations between a wide range of exposures and phenotypes, providing information on the complex network of interconnections between the exposome and phenome.

We developed high-throughput methods to identify associations between the exposome and the phenome, with robust quality control (QC) and defined analysis pipelines using our publicly available software, CLeaning to Analysis: Reproducibility-based Interface for Traits and Exposures (CLARITE)[11]. Using data from the National Health and Nutrition Examination Survey (NHANES)[14], we evaluated associations between 326 exposures and 55 phenotypes, using data from more than 18,000 participants. The environmental exposures ranged from physical and chemical measurements to survey responses. The phenotypes included biochemical blood sample measurements. Our PheEWAS reveals interrelationships between the exposome and phenome that are important to consider for human health.

RESULTS

PheEWAS results

We considered 326 exposures and 55 phenotypes that passed our quality control (QC) pipeline (see Methods) for PheEWAS. There were 1,930 statistically significant associations in the discovery dataset with a false discovery rate (FDR) < 0.1. These were considered for replication, and 686 replicated with the same direction of effect in the replication dataset with FDR < 0.1 (Fig. 1). One hundred and six exposure-phenotype associations demonstrated a Bonferroni-adjusted p-value < 0.05 in both discovery and replication datasets (Table 1). The exposures in these 106 PheEWAS associations were categorized as nutrients (30), food component recall (18), heavy metals (14), smoking behavior (14), smoking in the family (9), smoke exposure measured through cotinine (12), alcohol use (5), pharmaceutical (3), and supplement use (1). The phenotypes were categorized as metabolic products (19), blood proteins (19), white blood cells (WBC) (15), red blood cell (RBC)

(14), amino acids (14), enzymes (9), iron measurements (8), and lipids (8). The top exposure-phenotype associations ("d" subscript = discovery and "r" subscript = replication) were alpha-tocopherol (vitamin E) with triglycerides ($P_d = 6.51 \times 10^{-16}$, $\beta_d = 0.43$; $P_r = 4.20 \times 10^{-16}$, $\beta_r = 0.46$), gamma-tocopherol (vitamin E) with triglycerides ($P_d = 2.46 \times 10^{-14}$, $\beta_d = 0.37$; $P_r = 8.54 \times 10^{-16}$, $\beta_r = 0.37$). The 106 Bonferroni-significant associations included 28 unique exposures and 32 unique phenotypes. This shows that many exposures and phenotypes had repeated significant associations. As viewed in the network plot (Fig 2.), the most common significantly associated exposure was retinol, and the most common significantly associated phenotype was homocysteine. The exposure category with the largest number of replicated exposures was nutrients, specifically vitamins (Table 1 & Supplemental Table 1).

<u>Exposures with cross-phenotype associations</u>

To determine which exposures were involved with multiple phenotypes, we evaluated the top 106 exposure-phenotype associations with Bonferroni-significant results in discovery and replication datasets for exposures with cross-phenotype associations (i.e., exposures that demonstrated significant associations with more than one phenotype). We identified 19 exposures with cross-phenotype associations. These exposures belonged to the categories of vitamins (6), smoking (5), dietary supplements (nonmetals and metals) (4), metal pollutants (2), drugs and chemicals (1), and alcohol (1) (Fig. 2). The smoking exposure category had the highest number of associations (34), including cotinine (11) and survey questions: "Does anyone smoke in home?" (9), "Do you now smoke cigarettes?" (6), "Current cigarette smoker?" (4), and "Smoked at least 100 cigarettes in life" (3). The individual exposures that had the most cross-phenotype associations were retinol (14), cotinine (11),

cadmium (10), “Does anyone smoke in home?” (9), and gamma-tocopherol (9) (Table 1, Supplemental Table 1, Fig. 3).

Retinol had the largest number of cross-phenotype associations, which included associations with triglycerides ($P_d = 1.24 \times 10^{-14}$, $\beta_d = 0.31$; $P_r = 4.85 \times 10^{-15}$, $\beta_r = 0.31$), uric acid ($P_d = 9.57 \times 10^{-13}$, $\beta_d = 0.21$; $P_r = 1.11 \times 10^{-13}$, $\beta_r = 0.23$), frozen serum iron ($P_d = 1.70 \times 10^{-12}$, $\beta_d = 0.17$; $P_r = 1.55 \times 10^{-10}$, $\beta_r = 0.25$), blood urea nitrogen ($P_d = 8.04 \times 10^{-12}$, $\beta_d = 0.25$; $P_r = 5.47 \times 10^{-13}$, $\beta_r = 0.24$), ferritin ($P_d = 6.91 \times 10^{-9}$, $\beta_d = 0.14$; $P_r = 8.41 \times 10^{-11}$, $\beta_r = 0.28$), gamma-glutamyl transferase (GGT) ($P_d = 6.91 \times 10^{-9}$, $\beta_d = 0.18$; $P_r = 1.37 \times 10^{-12}$, $\beta_r = 0.19$), hemoglobin ($P_d = 2.48 \times 10^{-8}$, $\beta_d = 0.10$; $P_r = 7.39 \times 10^{-6}$, $\beta_r = 0.09$), and total iron-binding capacity (TIBC) ($P_d = 2.48 \times 10^{-6}$, $\beta_d = 0.18$; $P_r = 6.64 \times 10^{-8}$, $\beta_r = 0.24$).

Cotinine, a biomarker for smoking exposure, had cross-phenotype associations, among which nine had positive beta values. These included WBC phenotypes, such as lymphocyte number ($P_d = 1.24 \times 10^{-11}$, $\beta_d = 0.18$; $P_r = 1.40 \times 10^{-11}$, $\beta_r = 0.16$), segmented neutrophil number ($P_d = 6.65 \times 10^{-8}$, $\beta_d = 0.16$; $P_r = 3.78 \times 10^{-10}$, $\beta_r = 0.17$), and eosinophils ($P_d = 2.47 \times 10^{-6}$, $\beta_d = 0.07$; $P_r = 2.61 \times 10^{-7}$, $\beta_r = 0.08$). Cotinine also had positive associations with RBC measurements, including mean cell volume ($P_d = 1.60 \times 10^{-9}$, $\beta_d = 0.14$; $P_r = 2.43 \times 10^{-9}$, $\beta_r = 0.24$), hemoglobin ($P_d = 1.76 \times 10^{-8}$, $\beta_d = 0.11$; $P_r = 8.04 \times 10^{-12}$, $\beta_r = 0.12$), and hematocrit ($P_d = 2.45 \times 10^{-8}$, $\beta_d = 0.10$; $P_r = 2.38 \times 10^{-11}$, $\beta_r = 0.11$). Cotinine demonstrated two cross-phenotype associations with negative beta values: blood urea nitrogen ($P_d = 3.41 \times 10^{-9}$, $\beta_d = -0.12$; $P_r = 3.12 \times 10^{-11}$, $\beta_r = -0.12$) and bilirubin ($P_d = 1.29 \times 10^{-6}$, $\beta_d = -0.11$; $P_r = 1.37 \times 10^{-6}$, $\beta_r = -0.09$).

The survey question, "Does anyone smoke in the home?" labelled as "Smoking in home" was significantly associated with a number of WBC and RBC traits with positive beta values. The WBC

count phenotypes included lymphocytes ($P_d = 8.71 \times 10^{-9}$, $\beta_d = 0.33$; $P_r = 7.69 \times 10^{-9}$, $\beta_r = 0.32$), monocytes ($P_d = 1.25 \times 10^{-8}$, $\beta_d = 0.28$; $P_r = 2.14 \times 10^{-5}$, $\beta_r = 0.19$), and segmented neutrophils ($P_d = 2.87 \times 10^{-8}$, $\beta_d = 0.34$; $P_r = 4.38 \times 10^{-10}$, $\beta_r = 0.33$). Additionally, they were associated with RBC measurements, including mean cell volume ($P_d = 4.24 \times 10^{-9}$, $\beta_d = 0.31$; $P_r = 5.05 \times 10^{-6}$, $\beta_r = 0.24$) and hematocrit ($P_d = 5.60 \times 10^{-9}$, $\beta_d = 0.21$; $P_r = 6.37 \times 10^{-8}$, $\beta_r = 0.23$)

Cadmium demonstrated 11 cross-phenotype associations. The top cross-phenotype associations were related to blood measurements, including hematocrit ($P_d = 2.22 \times 10^{-9}$, $\beta_d = 0.12$; $P_r = 2.93 \times 10^{-11}$, $\beta_r = 0.13$), mean cell volume ($P_d = 7.48 \times 10^{-9}$, $\beta_d = 0.10$; $P_r = 6.52 \times 10^{-8}$, $\beta_r = 0.10$), and lymphocyte number ($P_d = 2.47 \times 10^{-8}$, $\beta_d = 0.13$; $P_r = 9.74 \times 10^{-10}$, $\beta_r = 0.13$).

Gamma-tocopherol exhibited nine significant cross-phenotype associations, some of which overlapped with alpha-tocopherol, showing positive beta values with lipid-related phenotypes such as triglycerides (alpha-tocopherol: $P_d = 6.51 \times 10^{-16}$, $\beta_d = 0.42$; $P_r = 4.20 \times 10^{-16}$, $\beta_r = 0.46$; gamma-tocopherol: $P_d = 2.46 \times 10^{-14}$, $\beta_d = 0.37$; $P_r = 8.54 \times 10^{-16}$, $\beta_r = 0.37$) and LDL cholesterol (alpha-tocopherol: $P_d = 2.24 \times 10^{-10}$, $\beta_d = 0.33$; $P_r = 1.50 \times 10^{-9}$, $\beta_r = 0.32$; gamma-tocopherol: $P_d = 1.31 \times 10^{-7}$, $\beta_d = 0.23$; $P_r = 4.58 \times 10^{-8}$, $\beta_r = 0.23$). While alpha-tocopherol was associated with a negative beta value with homocysteine ($P_d = 1.04 \times 10^{-7}$, $\beta_d = -0.10$; $P_r = 1.61 \times 10^{-9}$, $\beta_r = -0.11$), the association between gamma-tocopherol and homocysteine showed a positive beta value ($P_d = 5.07 \times 10^{-8}$, $\beta_d = 0.09$; $P_r = 1.16 \times 10^{-7}$, $\beta_r = 0.09$).

<u>Phenotypes with cross-exposure associations</u>

Similar to identifying exposures with cross-phenotype associations, we determined which phenotypes were involved with multiple exposures. As such, we evaluated the top 106 exposure-phenotype

associations with Bonferroni-significant results in discovery and replication datasets for phenotypes with cross-exposure associations (i.e., phenotypes that demonstrated significant associations with more than one exposure). We identified 23 phenotypes with cross-exposure associations, classified as metabolic products (4), RBC measurements (4), WBCs (3), iron measurements (3), lipids (3), blood proteins (3) enzymes (2), and amino acids (1) (Table 1 & Fig. 2). When we considered associations within each phenotype category, phenotypes from the metabolic products category had the most associations (17), which included the traits blood urea nitrogen (7), uric acid (4), creatinine (3), and total bilirubin (3). At the individual level, phenotypes with the most cross-exposure associations were homocysteine (14), blood urea nitrogen (7), lymphocyte number (6), segmented neutrophil number (6), and mean cell volume (6) (Table 1, Supplemental Table 1, Fig. 4).

Homocysteine associated with multiple vitamin exposures with a negative beta coefficient. For instance, homocysteine was associated with folate RBC (vitamin B9) ($P_d = 1.74 \times 10^{-7}$, $\beta_d = -0.15$; $P_r = 1.32 \times 10^{-11}$, $\beta_r = -0.17$), riboflavin (vitamin B2) ($P_d = 7.73 \times 10^{-8}$, $\beta_d = -0.11$; $P_r = 3.84 \times 10^{-8}$, $\beta_r = -0.11$), total supplement count ($P_d = 1.18 \times 10^{-8}$, $\beta_d = -0.11$; $P_r = 2.25 \times 10^{-9}$, $\beta_r = -0.11$), potassium ($P_d = 9.52 \times 10^{-7}$, $\beta_d = -0.07$; $P_r = 3.06 \times 10^{-7}$, $\beta_r = -0.10$), alpha-tocopherol ($P_d = 1.03 \times 10^{-7}$, $\beta_d = -0.09$; $P_r = 1.61 \times 10^{-9}$, $\beta_r = -0.11$) and phosphorus ($P_d = 6.07 \times 10^{-7}$, $\beta_d = -0.09$; $P_r = 6.46 \times 10^{-7}$, $\beta_r = -0.11$). Homocysteine was associated with gamma-tocopherols ($P_d = 5.07 \times 10^{-8}$, $\beta_d = 0.10$; $P_r = 1.16 \times 10^{-7}$, $\beta_r = 0.09$) with positive beta values.

Lymphocyte and segmented neutrophil counts were associated with cotinine and cadmium, as previously described, and were also associated with other smoking-related exposures. Furthermore, lymphocyte levels were associated with survey questions "Current cigarette smoker?" ($P_d = 2.56 \times 10^{-10}$, $\beta_d = 0.48$; $P_r = 9.81 \times 10^{-6}$, $\beta_r = 0.23$), "Do you now smoke cigarettes?" ($P_d = 8.71 \times 10^{-9}$; $P_r = 7.69$

$\times 10^{-9}$), and "Smoked at least 100 cigarettes in life" ($P_d = 1.61 \times 10^{-8}$, $\beta_d = 0.29$; $P_r = 1.89 \times 10^{-7}$, $\beta_r = 0.25$). Similarly, segmented neutrophil count was associated with "Current cigarette smoker?" ($P_d = 1.88 \times 10^{-5}$, $\beta_d = 0.44$; $P_r = 1.48 \times 10^{-7}$, $\beta_r = 0.23$), "Do you now smoke cigarettes?" ($P_d = 1.56 \times 10^{-4}$; $P_r = 3.05 \times 10^{-10}$), and "Smoked at least 100 cigarettes in life" ($P_d = 2.77 \times 10^{-7}$, $\beta_d = 0.22$; $P_r = 4.35 \times 10^{-9}$, $\beta_r = 0.23$).

Finally, mean cell volume demonstrated the largest number of cross-exposure of significant associations with exposures in the smoking category. The top cross-exposure associations for this phenotype related to smoking: "Do you now smoke cigarettes?" ($P_d = 7.44 \times 10^{-7}$; $P_r = 1.53 \times 10^{-8}$) and "Smoked at least 100 cigarettes in life" ($P_d = 3.56 \times 10^{-7}$, $\beta_d = 0.24$; $P_r = 3.97 \times 10^{-12}$, $\beta_r = 0.25$).

<u>Interconnections between the phenome and exposome</u>

To visualize the intricate interplay among the 106 exposure-phenotype associations with Bonferroni-significance across the discovery and replication datasets, we generated a network map that illustrates the complexity of results across the phenome and exposome (Fig. 2). Among the 106 significant associations, 69 exhibited positive beta values, indicating relationships between exposures and phenotypes, whereas 29 demonstrated negative beta values, suggesting inverse associations. The remaining six associations involved categorical predictors, so no beta coefficients were reported because there was no single effect estimate [see Methods]. The resulting network reveals cross-phenotype and cross-exposure links, with some variables acting as highly connected nodes due to their associations with multiple traits. Retinol (Vitamin A) had the highest number of statistically significant cross-phenotype associations, highlighting its extensive links across diverse biological measures.

The visual layout of the network reveals distinct clusters of interconnected variables, where certain phenotypes and exposures act as bridges linking different variables. The size of each node represents its degree centrality, with larger nodes indicating variables that have a higher number of significant connections. Nodes such as retinol (vitamin A), homocysteine, cotinine, and cadmium appear larger, reflecting their roles as key points of interaction within the system. Their positioning sets them as central hubs within the network. Conversely, exposures and phenotypes with fewer connections form more isolated nodes on the periphery. Variables with only one connection, such as niacin (mg) or thiamin (vitamin B1) (mg), appear as smaller nodes, resembling appendages extending from the main network. These periphery nodes, while statistically significant, exhibit more limited influence in terms of cross-exposure or cross-phenotype associations within this study. This structural organization provides insight into how different exposures and biological markers are related, illustrating systemic interdependencies and hierarchical relationships. The varying node sizes and connection densities help to identify key variables, clusters, and association patterns.

PheEWAS results stratified by race-ethnicity

In addition to the multi-ethnic PheEWAS described above, we performed a PheEWAS stratified by self-identified Non-Hispanic Blacks (NHB), Non-Hispanic Whites (NHW), and Mexican Americans (MA) race/ethnicity using the same QC and analytical approaches as our unstratified PheEWAS, as described in the Methods section. For each stratified PheEWAS, in the discovery datasets NHB (1,335), NHW (1,887), and MA (1,887) had FDR < 0.1 and were considered for the replication datasets, which NHB (276), NHW (194), and MA (945) had the same direction of effect with FDR < 0.1. When comparing across the discovery and replication datasets, there were 106 significant

exposure-phenotype results in total that had a p-value < 0.05 (in both) with Bonferroni adjustment for all three groups: NHB (7), NHW (88), and MA (11). Notably, the total number of significant associations in the stratified PheEWAS, equal to that of the unstratified PheEWAS, is purely coincidental, as stratification altered the composition and distribution of associations (Table 2, Fig. 5). Ninety-five of the 106 associations from the unstratified analysis were identified in at least one stratum. The top five exposure-phenotype associations (retinol with blood urea nitrogen, gamma-tocopherol with triglycerides, alpha-tocopherol with triglycerides, retinol with triglycerides, and retinol with uric acid) were significant in all three strata and in the unstratified PheEWAS (Table 2, Fig. 6). Two associations were shared between NHB and NHW (retinol with creatinine and "smoked at least 100 cigarettes in life" with lymphocyte number), four associations were shared between the MA and NHW strata (alpha-tocopherol with LDL cholesterol, retinol with ferritin, retinol with total calcium, and thallium with creatinine) and no associations were shared between the NHB and MA strata (Fig 6, Supplemental Table 2). Eleven associations were unique to the race/ethnicity-specific analysis, two of which were unique to the MA stratum and nine to the NHW stratum (Table 2 & Supplemental Table 2). The two unique associations from the MA stratum were urinary cadmium with urinary creatinine (Bonferroni-adjusted $P_d = 7.83 \times 10^{-4}$, $\beta_d = 1.27$; Bonferroni-adjusted $P_r = 8.77 \times 10^{-5}$, $\beta_r = 0.66$) and platinum with eosinophils (Bonferroni-adjusted $P_d = 3.12 \times 10^{-2}$, $\beta_d = -35.01$; Bonferroni-adjusted $P_r = 3.89 \times 10^{-4}$, $\beta_r = -2.72$). Two examples from the NHW stratum include an association with a positive beta coefficient, between lead in urine and creatinine in urine (Bonferroni-adjusted $P_d = 6.57 \times 10^{-4}$, $\beta_d = 0.72$; Bonferroni-adjusted $P_r = 2.36 \times 10^{-10}$, $\beta_r = 1.14$), and an association with a negative beta coefficient between trans-beta-carotene and homocysteine (Bonferroni-adjusted $P_d = 3.31 \times 10^{-2}$, $\beta_d = -0.10$; Bonferroni-adjusted $P_r = 6.55 \times 10^{-6}$, $\beta_r = -0.12$).

DISCUSSION

PheEWAS assesses the effects of environmental exposures over a wide range of traits, which is important to gain an understanding of the interconnectedness of the environment and the phenome. In this study, we found 106 significant associations between exposures and phenotypes. Multiple exposures were significantly associated with numerous phenotypes, exemplifying the importance of exploring the impact of exposures at the phenome level.

Retinol (vitamin A) had the greatest number of cross-phenotype associations, with connections to 14 phenotypes (Table 1 & Fig. 2). Retinol is an essential, fat-soluble nutrient important in vision, growth, and metabolism[15–18]. It is obtained from plant and animal foods and by supplementation[15]. Although the mechanism is unknown, low levels of vitamin A impair iron metabolism[16], while normal retinol levels have been associated with normal erythropoiesis and hemoglobin concentrations[17]. Thus, vitamin A deficiency is linked to anemia[16]. In this study, our results demonstrated retinol was associated with positive beta coefficients for various iron-related phenotypes, such as ferritin and total iron binding capacity (TIBC), as well as with hemoglobin concentration, reaffirming the role of retinol in iron metabolism and erythropoiesis. Higher retinol levels are also implicated in metabolic syndrome[18]. Because retinol is fat-soluble, it increases with higher body fat. Previous studies found that higher levels of vitamin A, uric acid, and gamma-glutamyl transferase (GGT) are associated with a higher risk of developing metabolic syndrome[18]. Therefore, the positive associations between retinol and uric acid, GGT, and triglyceride levels in our study suggest that retinol is part of a complex network that responds to elevated triglycerides through increased oxidative stress[18]. More specifically, GGT is a liver enzyme that serves as a biomarker for various hepatic and systemic conditions[19]. Elevated GGT levels have been associated with multiple exposures and substances, indicating its role

in diverse metabolic and toxicological pathways where we observed associations with exposures like "drinks(s) per day"[20] and cadmium[21] while being associated with retinol. Although notably, GGT is not specific to lipids: it is similarly elevated after alcohol intake and with exposures to toxic metals including cadmium. The parallel associations we observe therefore reflect a united underlying oxidative-stress pathway that can be triggered by diverse metabolic and toxicological stressors.

Similarly, our results noted that alpha-tocopherol and gamma-tocopherol had cross-phenotype associations with positive beta coefficients related to lipid metabolism, including cholesterol and triglycerides. Vitamin E is a group of eight fat-soluble molecules, consisting of four tocopherols and four tocotrienols [22]. The alpha- and gamma-tocopherols are ingested from food and supplements[23]. Alpha-tocopherol is usually obtained from olives and sunflower oils, and gamma-tocopherol is found in soybeans and corn oil[23]. Previous studies have found associations between alpha-tocopherol and gamma-tocopherol with triglycerides[24,25] (Fig. 3). The fat-soluble tocopherols are transported in plasma by lipoproteins and erythrocytes; therefore, the associations between tocopherols and lipoproteins may reflect their joint transport[26]. We noticed that the phenotype serum homocysteine was associated with a positive beta coefficient for gamma-tocopherol but had a negative beta value for alpha-tocopherol. Homocysteine is reported to cause vascular damage by enhancement of lipoperoxidation, and alpha-tocopherol prevents the damage because of its fat-soluble antioxidant properties[27]. Although our PheEWAS found an association with a positive beta value for gamma-tocopherol and homocysteine, previous studies found inverse associations between alpha- and gamma-tocopherols and homocysteine[28]. Recent studies have also found that while alpha-tocopherol is associated with positive health outcomes, gamma-tocopherol may have a negative effect[28,29]. Although the mechanisms that underlie differences between tocopherol isoforms are unknown, these

differences may result from their cell signaling functions. Multiple genes are differentially regulated by the tocopherol isoforms, including metabolic pathways genes[30]. Our results suggest different roles for alpha- versus gamma-tocopherol.

Cross-exposure associations with negative beta coefficients between homocysteine and the exposures riboflavin (vitamin B2), vitamin B6, and folate (vitamin B9) were also identified. The metabolism of homocysteine involves the demethylation of methionine with the aid of B-vitamins[31]. When homocysteine is broken down by the B-vitamins, it can convert into two substances. One substance is methionine, an essential amino acid for creating proteins, and the other is cysteine, a nonessential amino acid that reduces inflammation by communicating with immune cells[31]. B-vitamin supplements can increase the conversion of homocysteine into methionine and cysteine, thereby decreasing homocysteine in the blood. Dietary recommendations for folate and vitamin B12 take into account their association with homocysteine[32]. Adequate intake of B-vitamins, especially folate, is crucial for maintaining healthy homocysteine levels in the blood[32]. High homocysteine levels are associated with increased risks of cardiovascular diseases, including stroke and ischemic heart disease[30,31]. Research suggests that populations with lower-than-recommended B-vitamin intake, such as those in Sweden and Norway, often have elevated homocysteine levels[32]; thus, the associations of folate and other B-vitamins with homocysteine could result from a vitamin deficiency[31]. However, more research is required to determine whether there is a causal relationship between cardiovascular disease (CVD) and homocysteine and whether vitamin supplements to lower homocysteine in the blood affect CVD[33].

Moreover, the associations between riboflavin (vitamin B2) and uric acid found in this PheEWAS extend this interconnected framework. Riboflavin plays a critical role in oxidative stress modulation,

and its tandem relationship with uric acid—both an antioxidant and pro-oxidant—further emphasizes shared pathways in oxidative metabolism[34,35]. These associations are anchored in the well-established relationship between riboflavin and homocysteine, where riboflavin acts as a cofactor in homocysteine metabolism, promoting its breakdown into methionine and cysteine[34]. When homocysteine builds up, it creates harmful oxidative stress, and uric acid helps neutralize those damaging molecules. Vitamin B2 (riboflavin) is needed to make the enzymes that both break down homocysteine and produce uric acid, linking the two processes. Similarly, folate (vitamin B9), which is crucial for DNA synthesis and erythropoiesis, is also closely linked to homocysteine metabolism, underscoring its role in maintaining cardiovascular and hematological health[36,37].

Other inverse associations with homocysteine were found in this study for several micronutrients, including potassium, which controls nerve function and muscle contraction[38]. Atrial myocytes, which form the muscular walls of the heart, use potassium to regulate their contractile function; decreased potassium can lead to atrial fibrillation[39]. Homocysteine at 50 μmol/L and 500 μmol/L results in a significant decrease in potassium movement in human atrial myocytes[39]. The mechanism by which homocysteine directly affects myocytes is unknown; however, our finding that homocysteine was inversely associated with potassium supports its possible effect on myocytes[39]. Our results are consistent with prior studies showing associations between elevated homocysteine levels and adverse health outcomes, particularly in the context of cardiovascular disease and cardiac arrhythmias, though these specific outcomes were not directly evaluated in our phenotype set. Overall, our results indicate a trend of adverse outcomes associated with increased homocysteine.

Cotinine, a biomarker for tobacco smoking exposure[40], had 12 cross-phenotype associations. Smoking behavior is a risk factor for heart disease, lung cancer, stroke, and chronic obstructive

pulmonary disease (COPD)[41]. Increased cotinine levels and smoking behaviors are associated with inflammation and an increase in WBCs[42,43]. We found associations with a positive beta coefficient between cotinine and smoking-related behaviors with several WBC counts. This supports the hypothesis that smoking behavior can lead to systemic inflammation that increases WBC[44]. This is also supported by our finding of an association with positive beta value between cotinine levels and C-reactive protein, an inflammatory biomarker[45]. In our study, cotinine was also associated with increased RBCs. It has been hypothesized that, due to carbon dioxide being carried by cigarette smoke, increased hematocrit and hemoglobin levels compensate to maintain oxygen transport[45]. Our analysis detected positive directions-of-effect for cotinine levels and other smoking-related behaviors with hemoglobin and hematocrit, indicating an increase in RBCs with smoking exposure[46].

Our analysis showed that serum cadmium levels had 11 cross-phenotype associations, including WBC and RBC phenotypes. Cadmium is a toxic heavy metal that accumulates in plants, particularly tobacco and edible crops grown in cadmium-contaminated soil and water. Therefore, exposure to cadmium is often via smoking and/or food[47]. Exposure to cadmium causes oxidative stress and affects the kidneys, liver, bone, and lungs[12,48]. We observed that cadmium and cotinine shared similar associations, which may result from cadmium exposure via tobacco smoke[13,49,50]. There was a positive direction of effect between cadmium and mean RBC volume. However, WBC counts also increased with cadmium. *Ciarrocca et al.,* found that, in workers exposed to cadmium, there was a positive correlation between cadmium and lymphocyte count. Cadmium induces apoptosis of lymphocytes, thereby modulating the immune system[52]. Our results show similar associations between cotinine and cadmium, which could reflect their co-occurrence in tobacco. However, because cadmium is also an industrial pollutant, it could affect WBC and RBC counts independent of

339 cotinine[53]. As we learn more about the effects of cadmium exposure on multiple phenotypes, we will
340 be able to inform treatment or prevention of diseases related to excess cadmium.

341 When we analyzed each race/ethnicity stratum, we identified many significant associations that
342 were shared with the unstratified analysis. Alpha-tocopherol with triglycerides, retinol with
343 triglycerides, retinol with uric acid, retinol with blood urea nitrogen, and gamma-tocopherol with
344 triglycerides were the top five significant associations for exposures and phenotypes, respectively, in
345 the unstratified and stratified analyses. Race/ethnicity is often adjusted in linear regression models as
346 a covariate which precludes investigation of population-specific associations[54]. To determine if any
347 associations varied by population, we assessed differences between the stratified and unstratified
348 results. Eleven associations were found in the stratified analysis that were not significant in the
349 unstratified dataset, nine of which were identified in the NHW sample. The NHW sample was larger
350 than the NHB and MA samples, which may have driven the higher number of observed associations
351 within that stratum.

352 Two associations were only observed in the MA stratum, an association with a positive direction-
353 of-effect involving cadmium exposure and the creatinine phenotype, and an association with an
354 inverse direction-of-effect between platinum exposure and eosinophil count. The positive direction-of-
355 effect of cadmium and creatinine, implies the potentially toxic effects of cadmium exposure on health.
356 A previous study found that Mexican Americans living in urban and agricultural areas had higher
357 urinary and blood cadmium concentrations compared to other populations[55]. This elevated exposure
358 was linked to increased risk of kidney damage, as reflected in higher serum creatinine levels[55]. These
359 findings demonstrate the need for further study of cadmium exposure in this population.

Only the NHW stratum showed a positive direction-of-effect between urinary lead and urinary creatinine. Lead toxicity is determined by urine lead concentration per gram of creatinine[56]. Lead toxicity produces reactive oxygen species (ROS) causing oxidative stress[5660]. Lead-induced oxidative stress involves the generation and removal of ROS in cells and tissues, which damages membranes, DNA, and proteins[56]. An increased intake of antioxidants can decrease susceptibility to oxidation in those tissues[27]. Supplementation with beta-carotene decreases homocysteine in the blood and has been shown to reduce oxidative stress in chronically lead-exposed workers[57]. Beta-carotene is a carotenoid that protects the body from free radicals; when homocysteine is broken down, beta-carotene increases antioxidant production, reducing oxidative stress[27,57]. The inverse relationship we observed between trans-beta-carotene and homocysteine in the NHW sample supports these findings.

Using PheEWAS, we leveraged data from the NHANES to interrogate the complex relationship between the exposome and phenome. We replicated several associations previously described in the literature and identified novel relationships as well. Limitations to this study should be noted. 1) Like genome-wide association studies, our study provides limited information on the mechanisms linking the exposures and phenotypes; however, these associations can provide novel hypotheses for future studies. 2) The nature of the NHANES provides only cross-sectional snapshots of the exposure and phenotypic associations during adult stages, thereby missing dynamic longitudinal changes in the exposome and their phenotypic effects. In the future, we hope to incorporate longitudinal data and methods in phenomic environment-wide analyses. 3) Because classifying a characteristic or an analyte as a phenotype or exposure can be an arbitrary decision, we were strict in our classifications. We combed the NHANES dataset and dictionary to accurately identify a) phenotypes based on blood and biochemistry lab measurement health outcomes and b) exposures based on their definitions as

exogenous factors. Furthermore, the direction of causality was not explored in this study; in some cases, our defined exposures could be affected by an individual's health status, and in other cases, both exposure and phenotypes could be affected by factors that were not assessed by NHANES. This illustrates the challenges of the high-throughput analysis of exposome and phenome data. 4) Our interpretation of the PheEWAS requires further downstream hypothesis testing. A possible follow-up study could use a more restrictive statistical model that incorporates covariates unique to one of the significant exposure-phenotype associations that we would not necessarily control for in a generic EWAS model. Re-sampling the data using methods like bootstrapping will also improve the interpretation of results.

PheEWAS is a novel approach that connects the exposome with the phenome. As characterization of the exposome becomes more possible with a diverse array of technologies and assessment measures, its complex relationship with the phenome will be further revealed. When applied to a wider range of these exposures, and across multiple timepoints, PheEWAS will enable increased knowledge of the impact of the environment on human health and the ability to identify those most vulnerable to developing adverse health outcomes.

METHODS

National Health and Nutrition Examination Survey (NHANES)

This PheEWAS identified associations between multiple phenotypes and environmental exposures in adults using data from NHANES, which evaluates the health and nutritional status of adults and children in the United States[14]. The survey consists of dietary, socioeconomic, and general health data

obtained through questionnaires and clinical laboratory measurements. These are data collected annually by the Centers for Disease Control and Prevention (CDC) public health policies[14].

We split the NHANES data into discovery and replication cohorts to demonstrate replicability within the dataset. We excluded individuals under age 18, leaving 18,850 individuals, 970 exposure variables, and 56 phenotypes across four survey periods (1999−2000, 2001−2002, 2003−2004, 2005−2006) prior to quality control (QC). The participants in the 1999−2000 and 2001−2002 datasets comprised the discovery dataset (N = 9,063; 52% female, 47% NHW, 19% NHB, 25% MA, 5% other Hispanic, and 3% other ethnicities). The participants in the 2003−2004 and 2005−2006 datasets comprised the replication dataset (N = 9,874; 52% female, 49% NHW, 23% NHB, 21% MA, 3% other Hispanic, 1% other ethnicities) (Table 3). Demographic information was obtained through participant interviews or physical examinations at the mobile exam center[58].

Selection of phenotypes and exposures

The phenotypes were selected from the blood and biochemistry laboratory measurements recorded in the NHANES Dataset Explorer[59]. Exposures were obtained from questionnaires, examinations, and other laboratory measurements, and comprised the following categories: alcohol use, bacterial infection, cotinine, dialkyl, dioxins, food component recall, furans, heavy metals, hydrocarbons, immunizations, nutrients, occupations, polychlorinated biphenyls, pesticides, pharmaceuticals, phthalates, phytoestrogens, sexual behavior, smoking behavior, smoking family, social support, street drugs, supplements, and volatile chemicals. To better summarize the above exposure categories when listing cross-phenotype results, we aggregated them into the following: dietary supplements (nonmetals and metals), metal pollutants, vitamins, alcohol, smoking, and drugs and chemicals. We

utilized phenotype in the biochemistry or blood sample category in the NHANES data dictionary. For blood or biochemistry measurements, the phenotypes were divided into categories for assessing multi-functional effects: WBCs, iron measurements, amino acids, RBC measurements, lipids, metabolic products, enzymes, and blood proteins. All blood or biochemistry markers were deemed a measurable characteristic, as they reflected the physiological state of an individual that could be influenced by environmental factors. For a full list of phenotypes and exposures, see Supplemental Table 3.

Quality control of phenotypes and exposures

Before QC, 970 environmental variables and 56 phenotypes from the NHANES dataset were considered for the discovery dataset. The QC pipelines, protocols, and covariate adjustments for the discovery and replication datasets were identical.

Data filtering and sample/variable-based exclusion

Using CLARITE Software[11], environmental exposures were classified as binary, categorical, or continuous. All phenotypes were continuous. Covariates included for adjustment were age, sex, socioeconomic status (SES), BMI, self-reported race/ethnicity, and data release cycle year. Participants missing a covariate value were excluded, leaving 9,063 samples in the discovery and 9,874 in the replication datasets. Categorical and binary variables that had less than 200 samples in any category were dropped[60]. To determine whether 200 was a sufficient sample size, we simulated over 1,000 observations and found the optimal power of 80% with detection of effect sizes greater than 0.4 (Supplemental Fig. 1 a, b).

Data distribution and transformation of phenotypes

Skewness was calculated and histograms for all phenotypes were visualized for all phenotypes using CLARITE Software (Supplemental Fig. 2), and a log transformation was performed on skewed phenotypes. Phenotypes that had a normal distribution and a skewness value of −0.5 to 0.5 were not transformed. Thirteen discovery phenotypes were normally distributed and were not transformed, whereas 45 phenotypes were log-transformed (Supplemental Fig. 2 a-c). For negatively skewed phenotype p, values of p were adjusted by subtracting them from the constant $[1 + \max(p)]$. The new values were log-transformed, and once again adjusted by subtracting them from the new $[1 + \max(p)]$ to return to their original orientation[61]. Continuous phenotypes for which more than 90% of samples had a value of zero were removed. After QC, 326 exposures (277 continuous, 45 binary, and four categorical variables) and 55 phenotypes remained for analysis. Further, to maintain sufficient sample size, we removed any exposure-phenotype combination with a total sample size or category size (if categorical or binary) < 200, leaving 17,909 out of the 17,930 exposure-phenotype pairs with a total sample size and, if applicable, category size > 200 that was included in the PheEWAS.

Lipid phenotypes

When a participant reported taking a statin (atorvastatin, simvastatin, pravastatin, or fluvastatin), the low-density lipoprotein (LDL)-cholesterol value was divided by 0.7 to estimate the free LDL cholesterol level prior to statin treatment, as recommended by the Global Lipids Genetics Consortium[62,63].

<u>Survey weights</u>

We included survey weights in our analysis to account for the complex survey design of NHANES as recommended[64]. NHANES selected participants from primary sampling units (PSU) and strata while oversampling typically underrepresented groups, such as minorities and the elderly. Survey cycles one and two (1999–2002) were assigned four-year weights as defined by the population estimates used in the 1990 and 2000 census years[64]. In contrast, survey cycles three and four (2003–2006) had two-year weights from which the four-year weights were calculated by dividing their respective two-year weight in half[65].

Phenomic environment-wide association study

Fifty-five phenotypes and 326 exposures that passed QC were included in the PheEWAS (Supplemental Table 1). All PheEWAS models were tested using CLARITE Software. We used a general linear model that was adjusted for sex, self-reported race/ethnicity (Non-Hispanic White, Non-Hispanic Black, and Mexican American), body mass index (BMI), SES, age, and survey year. Complex survey design data was passed into the model including the strata, cluster, and weights for specific variables. A Z-score transformation of continuous or binary exposures and continuous phenotypes was used to standardize the beta coefficients and allowed for a direct comparison of effect sizes across different exposure-phenotype associations[10]. If the predictor was continuous or binary, CLARITE reports the beta coefficient and p-value from the general linear model. For categorical predictors, CLARITE compares the model with and without the categorical predictor using a likelihood ratio test (LRT) to assess whether the inclusion of any (or all the) categories of the predictor improves the model fit. Significant exposure-phenotype associations with a false discovery rate (FDR) p-value < 0.10 in the discovery were considered for replication. Replication criteria for exposure-

phenotype associations were 1) an FDR p-value < 0.10 in the replication dataset and 2) the same direction of effect between subsets.

Three race/ethnicity strata provided by the NHANES questionaries—Non-Hispanic Black, Mexican American, and Non-Hispanic White—had sufficient sample sizes to include in the stratified analysis. Here, we applied the same discovery and replication pipeline as for the unstratified PheEWAS. In the stratified PheEWAS, we accounted for the complex survey design for NHANES similar to the unstratified PheEWAS. The same strata, PSUs, and sampling weights were used. However, stratifying the NHANES data introduced strata with only a single PSU, which can occur in subsets of data[66]. In CLARITE Software, the traditional variance calculation must be adjusted; therefore, we used the 'adjust' option, which was incorporated from the survey package in R. This option provides a conservative estimation of variance by using population residuals rather than stratum residuals, ensuring a more conservative estimate of variance in complex survey designs[66].

Data Availability

More information on the data is available on Dryad (https://doi.org/10.5061/dryad.d5h62)[59]

All code, QC files, and result files can be found on GitHub (https://github.com/HallLab/PheEWAS_paper/tree/main)

1. Pendergrass, S. A. *et al.* Phenome-Wide Association Study (PheWAS) for Detection of Pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet* **9**, (2013).
2. Hall, M. A. *et al.* Detection of Pleiotropy through a Phenome-Wide Association Study (PheWAS) of Epidemiologic Data as Part of the Environmental Architecture for Genes Linked to Environment (EAGLE) Study. *PLoS Genet* **10**, e1004678 (2014).
3. Bush, W. S., Oetjens, M. T. & Crawford, D. C. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nature Reviews Genetics* vol. 17 129–145 Preprint at https://doi.org/10.1038/nrg.2015.36 (2016).
4. Liu, S. & Crawford, D. C. Maturation and application of phenome-wide association studies. *Trends in Genetics* vol. 38 353–363 Preprint at https://doi.org/10.1016/j.tig.2021.12.002 (2022).
5. Denny, J. C., Bastarache, L. & Roden, D. M. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. *Annu Rev Genomics Hum Genet* **17**, 353–373 (2016).
6. Braun, J. M., Kalloo, G., Kingsley, S. L. & Li, N. Using phenome-wide association studies to examine the effect of environmental exposures on human health. *Environment International* vol. 130 104877 Preprint at https://doi.org/10.1016/j.envint.2019.05.071 (2019).
7. González Zarzar, T., Palmiero, N. E., Kim, D., Shen, L. & Hall, M. A. Differential effects of environmental exposures on clinically relevant endophenotypes between sexes. *Scientific Reports 2024 14:1* **14**, 1–15 (2024).
8. Manrai, A. K. *et al.* Informatics and Data Analytics to Support Exposome-Based Discovery for Public Health. *Annu Rev Public Health* **38**, 279–294 (2017).
9. Buck Louis, G. M., Smarr, M. M. & Patel, C. J. The Exposome Research Paradigm: an Opportunity to Understand the Environmental Basis for Human Health and Disease. *Current environmental health reports* vol. 4 89–98 Preprint at https://doi.org/10.1007/s40572-017-0126-3 (2017).
10. Patel, C. J., Bhattacharya, J. & Butte, A. J. An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus. *PLoS One* **5**, e10746 (2010).
11. Lucas, A. M. *et al.* CLARITE Facilitates the Quality Control and Analysis Process for EWAS of Metabolic-Related Traits. *Front Genet* **10**, (2019).

12. Genchi, G., Sinicropi, M. S., Lauria, G., Carocci, A. & Catalano, A. The Effects of Cadmium Toxicity. *Int J Environ Res Public Health* **17**, (2020).

13. Pollack, A. Z. *et al.* Kidney biomarkers associated with blood lead, mercury, and cadmium in premenopausal women: a prospective cohort study. *J Toxicol Environ Health A* **78**, 119 (2015).

14. NHANES - National Health and Nutrition Examination Survey Homepage. https://www.cdc.gov/nchs/nhanes/index.htm.

15. T, O. & R, B. Retinol, Retinoic Acid, and Retinol-Binding Protein 4 are Differentially Associated with Cardiovascular Disease, Type 2 Diabetes, and Obesity: An Overview of Human Studies. *Adv Nutr* **11**, 644–666 (2020).

16. Semba, R. & Bloem, M. The anemia of vitamin A deficiency: epidemiology and pathogenesis. *European Journal of Clinical Nutrition 2002 56:4* **56**, 271–281 (2002).

17. SM, J. *et al.* Serum retinol levels are positively correlated with hemoglobin concentrations, independent of iron homeostasis: a population-based study. *Nutr Res* **33**, 279–285 (2013).

18. KanagasabaiThirumagal, AlkhalaqiKhloud, R., C. & I., A. The Association Between Metabolic Syndrome and Serum Concentrations of Micronutrients, Inflammation, and Oxidative Stress Outside of the Clinical Reference Ranges: A Cross-Sectional Study. *https://home.liebertpub.com/met* **17**, 29–36 (2019).

19. Gamma-Glutamyl Transferase (GGT) Test: What It Is & Results. https://my.clevelandclinic.org/health/diagnostics/22055-gamma-glutamyl-transferase-ggt-test.

20. Nishimura, M. & Teschke, R. Alcohol and gamma-glutamyltransferase. *Klin Wochenschr* **61**, 265–275 (1983).

21. Hong, D., Min, J. Y. & Min, K. B. Association Between Cadmium Exposure and Liver Function in Adults in the United States: A Cross-sectional Study. *Journal of Preventive Medicine and Public Health* **54**, 471–480 (2021).

22. Marriot, B., Birt, D., Stallings, V. & Yates, A. Vitamin E. in *Present Knowledge in Nutrition* 115136 (Elsevier, London, United Kingdom, 2020).

23. Wagner, K.-H., Kamal-Eldin, A. & Elmadfa, I. Gamma-Tocopherol – An Underestimated Vitamin? *Ann Nutr Metab* **48**, 169–188 (2004).

24. Winbauer, A., Pingree, S. & Nuttall, K. Evaluating Serum alpha-Tocopherol (Vitamin E) in Terms of a Lipid Ratio. *Ann Clin Lab Sci* **29**, 185–191 (1999).

25. White, E. *et al.* Correlates of serum α- and γ-tocopherol in the women's health initiative. *Ann Epidemiol* **11**, 136–144 (2001).

26. Bjorneboe, A., Bjorneboe, G. E. A. & Drevon, C. A. Absorption, transport and distribution of vitamin E. *Journal of Nutrition* **120**, 233–242 (1990).

27. Bunout, D. *et al.* Effects of supplementation with folic acid and antioxidant vitamins on homocysteine levels and LDL oxidation in coronary patients. *Nutrition* **16**, 107–110 (2000).

28. Sotomayor, C. G. *et al.* Duality of Tocopherol Isoforms and Novel Associations with Vitamins Involved in One-Carbon Metabolism: Results from an Elderly Sample of the LifeLines Cohort Study. *Nutrients 2020, Vol. 12, Page 580* **12**, 580 (2020).

29. Cooney, R. V., Franke, A. A., Wilkens, L. R., Gill, J. & Kolonel, L. N. Elevated plasma γ-tocopherol and decreased α-tocopherol in men are associated with inflammatory markers and decreased plasma 25-OH vitamin D. *Nutr Cancer* **60**, 21–29 (2008).

30. Azzi, A. *et al.* Regulation of gene expression by α-tocopherol. *Biol Chem* **385**, 585–591 (2004).

31. Homocysteine: Levels, Tests, High Homocysteine Levels. https://my.clevelandclinic.org/health/articles/21527-homocysteine.

32. Dhonukshe-Rutten, R. A. M. *et al.* Dietary intake and status of folate and vitamin B12 and their association with homocysteine and cardiovascular disease in European populations. *Eur J Clin Nutr* **63**, 18–30 (2009).

33. Clarke, R. *et al.* Homocysteine and Risk of Ischemic Heart Disease and Stroke: A Meta-analysis. *JAMA* **288**, 2015–2022 (2002).

34. Riboflavin | Linus Pauling Institute | Oregon State University. https://lpi.oregonstate.edu/mic/vitamins/riboflavin.

35. Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B6, Folate, Vitamin B12, Pantothenic Acid, Biotin, and Choline. *Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B6, Folate, Vitamin B12, Pantothenic Acid, Biotin, and Choline* (1998) doi:10.17226/6015.

36. Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B6, Folate, Vitamin B12, Pantothenic Acid, Biotin, and Choline. *Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B6, Folate, Vitamin B12, Pantothenic Acid, Biotin, and Choline* (1998) doi:10.17226/6015.

37. Folate | Linus Pauling Institute | Oregon State University. https://lpi.oregonstate.edu/mic/vitamins/folate.

38. Lanham-New, S. A., Lambert, H. & Frassetto, L. Potassium. *Advances in Nutrition* **3**, 820 (2012).

39. BZ, C. *et al.* Homocysteine inhibits potassium channels in human atrial myocytes. *Clin Exp Pharmacol Physiol* **34**, 851–855 (2007).

40. Dwoskin, L. P., Teng, L., Buxton, S. T. & Crooks, P. A. (S)-(−)-Cotinine, the Major Brain Metabolite of Nicotine, Stimulates Nicotinic Receptors to Evoke [3H]Dopamine Release from Rat Striatal Slices in a Calcium-Dependent Manner. *Journal of Pharmacology and Experimental Therapeutics* **288**, (1999).

41. Centers for Disease Control and Prevention (CDC). Health Effects | Smoking & Tobacco Use | CDC. https://www.cdc.gov/tobacco/basic_information/health_effects/index.htm (2020).

42. Dinas, P. C. *et al.* Acute effects of second-hand smoke on complete blood count. *http://dx.doi.org/10.1080/09603123.2013.782603* **24**, 56–62 (2013).

43. Abel, G. A. *et al.* Effects of Biochemically Confirmed Smoking Cessation on White Blood Cell Count. *Mayo Clin Proc* **80**, 1022–1028 (2005).

44. Arcavi, L. & Benowitz, N. L. Cigarette Smoking and Infection. *Arch Intern Med* **164**, 2206–2216 (2004).

45. Pedersen, K. M. *et al.* Smoking and Increased White and Red Blood Cells. *Arterioclerosis, Thrombosis, and Vascular Biology* **39**, 965–977 (2019).

46. Malenica, M. *et al.* Effect of Cigarette Smoking on Haematological Parameters in Healthy Population. *Medical Archives* **71**, 132 (2017).

47. Ganguly, K., Levänen, B., Palmberg, L., Åkesson, A. & Lindén, A. Cadmium in tobacco smokers: a neglected link to lung disease? *European Respiratory Review* **27**, (2018).

48. Agency for Toxic Substances and Disease Registry (ATSDR). ToxGuide for Cadmium. *Agency for Toxic Substances and Disease Registry* 2 (2012).

49. S, T. *et al.* Chronic Cadmium Exposure Lead to Inhibition of Serum and Hepatic Alkaline Phosphatase Activity in Wistar Rats. *J Biochem Mol Toxicol* **29**, 587–594 (2015).

50. Chen, Y. *et al.* Blood lead and cadmium levels associated with hematological and hepatic functions in patients from an e-waste-polluted area. (2019) doi:10.1016/j.chemosphere.2018.12.129.

51. Ciarrocca, M. *et al.* Correlation Between Cadmium and Blood Counts in Workers Exposed to Urban Stressor. *http://dx.doi.org/10.1080/19338244.2013.778807* **70**, 70–76 (2015).

52. Wang, Z., Sun, Y., Yao, W., Ba, Q. & Wang, H. Effects of Cadmium Exposure on the Immune System and Immunoregulation. *Front Immunol* **0**, 2892 (2021).

53. Zhang, H., Yan, J., Niu, J., Wang, H. & Li, X. Association between lead and cadmium co-exposure and systemic immune inflammation in residents living near a mining and smelting area in NW China. *Chemosphere* **287**, 132190 (2022).

54. VanderWeele, T. J. & Robinson, W. R. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology* **25**, 473–484 (2014).

55. Weiss, M. C. *et al.* Predictors of Toxic Metal/Metalloid Exposures Among Mexican Americans in Starr County, Texas. *J Racial Ethn Health Disparities* 1–11 (2024) doi:10.1007/S40615-024-02064-3/TABLES/5.

56. Patra, R. C., Rautray, A. K. & Swarup, D. Oxidative Stress in Lead and Cadmium Toxicity and Its Amelioration. *Vet Med Int* **2011**, (2011).

57. Kasperczyk, S. *et al.* The influence of beta-carotene on homocysteine level and oxidative stress in lead-exposed workers - PubMed. *Med Pr.* **3**, 309–316 (2014).

58. Center for Health Statistics, N. *National Health and Nutrition Examination Survey INTERVIEWER PROCEDURES MANUAL*. (2013).

59. Patel, C. J. *et al.* A database of human exposomes and phenomes from the US National Health and Nutrition Examination Survey. *Sci Data* **3**, 160096 (2016).

60. Wilson Vanvoorhis, C. R. & Morgan, B. L. *Understanding Power and Rules of Thumb for Determining Sample Sizes*. *Tutorials in Quantitative Methods for Psychology* vol. 3 http://www.tqmp.org/RegularArticles/vol03-2/p043/p043.pdf (2007).

61. Wuensch, K. L. *Transforming to Reduce Negative Skewness*. http://core.ecu.edu/psyc/wuenschk/StatHelp/NegSkew.pdf (2017).

62. Grundy, S. M. & Feingold, K. R. Guidelines for the Management of High Blood Cholesterol. *Endotext* (2019).

63. Liu, D. J. *et al.* Exome-wide association study of plasma lipids in >300,000 individuals. *Nature Genetics 2017 49:12* **49**, 1758–1766 (2017).

64. CDC. Third National Report on Human Exposure to Environmental Chemicals Department of Health and Human Services Centers for Disease Control and Prevention. (2005).

65. CDC. *NHANES Analytic Guidelines*. http://www.cdc.gov/nchs/data/nhanes/guidelines1.pdf (2004).

66. Lumley, T. *Complex Surveys*. *Biostatistics* (John Wiley & Sons, Inc., Hoboken, NJ, USA, 2010). doi:10.1002/9780470580066.

67. Lucas, A. M. anastasia-lucas/hudson: An R package for creating mirrored Manhattan plots. *2018* https://github.com/anastasia-lucas/hudson#readme.

68. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. Circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).

69. Otasek, D., Morris, J. H., Bouças, J., Pico, A. R. & Demchak, B. Cytoscape Automation: Empowering workflow-based network analysis. *Genome Biol* **20**, (2019).

70. Pendergrass, S. A., Dudek, S. M., Crawford, D. C. & Ritchie, M. D. Visually integrating and exploring high throughput Phenome-Wide Association Study (PheWAS) results using PheWAS-View. *BioData Min* **5**, 1 (2012).

Acknowledgements

This work is supported by the USDA National Institute of Food and Agriculture and Hatch Appropriations under Project #PEN04275 and Accession #1018544

Author contributions

N.E.P., T.G.Z., and K.P. contributed the formal analysis, investigation, methodology, writing and editing the manuscript. J.Z. contributed the formal analysis, methodology, and editing the manuscript. D.C.C. contributed investigation, methodology, and editing the manuscript. M.A.H., contributed the conceptualization, formal analysis, investigation, methodology, writing and editing the manuscript, project administration, and supervision.

Competing interests

The authors declare no competing interests.

Materials and Correspondence

Please contact Molly A. Hall at molly.hall@pennmedicine.upenn.edu for requests of correspondence and materials.

Tables and Figures

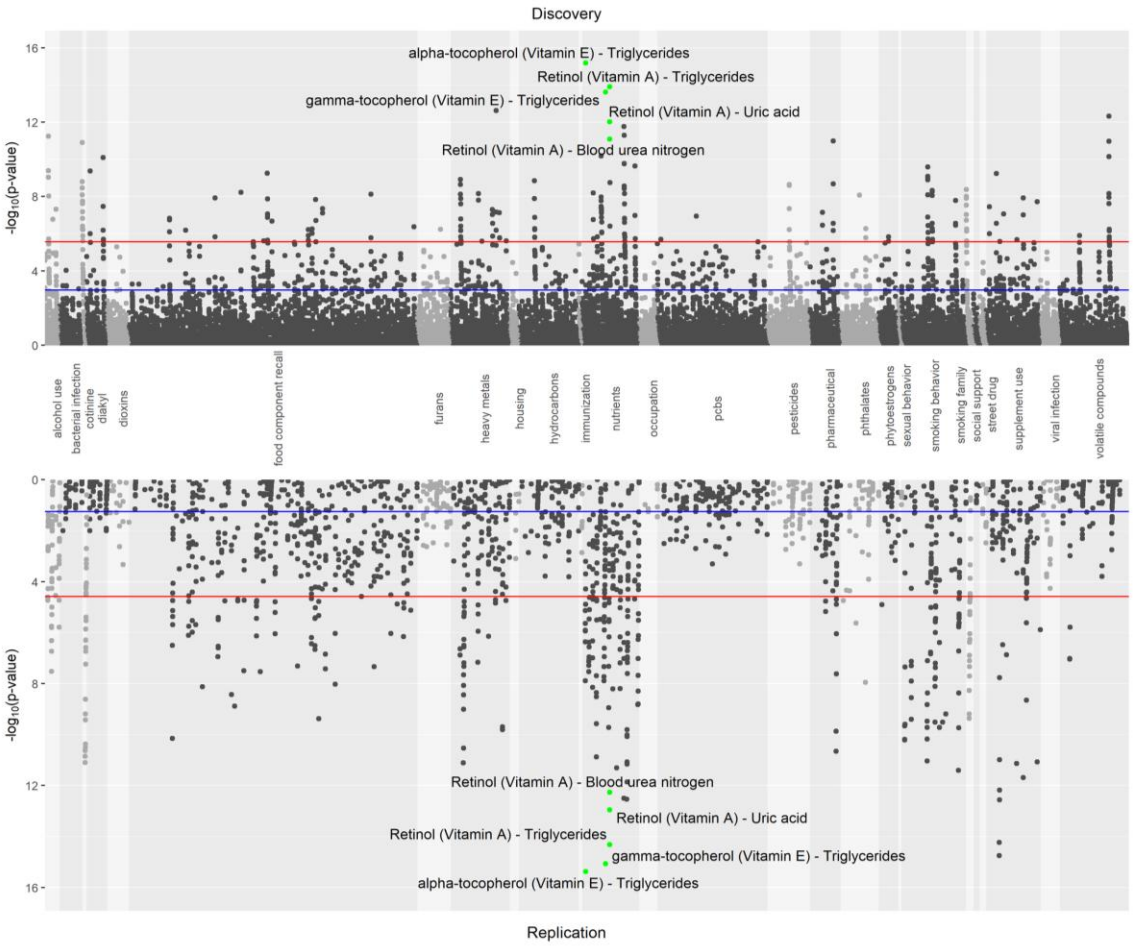

Fig. 1. Phenomic environment-wide association study results. Hudson plot of PheEWAS results. The x-axis shows exposure-phenotype associations, ordered by exposure category, and the y-axis shows the $-\log_{10}$(p-value). The red lines denote a Bonferroni-adjusted p-value of 0.05, and the blue lines denote the threshold for a false discovery rate (FDR) of 0.1. The top replicating results in the nutrients category that were significant after FDR adjustment and Bonferroni correction are labeled. The Hudson plot was created using an R package available at https://github.com/anastasia-lucas/hudson[67].

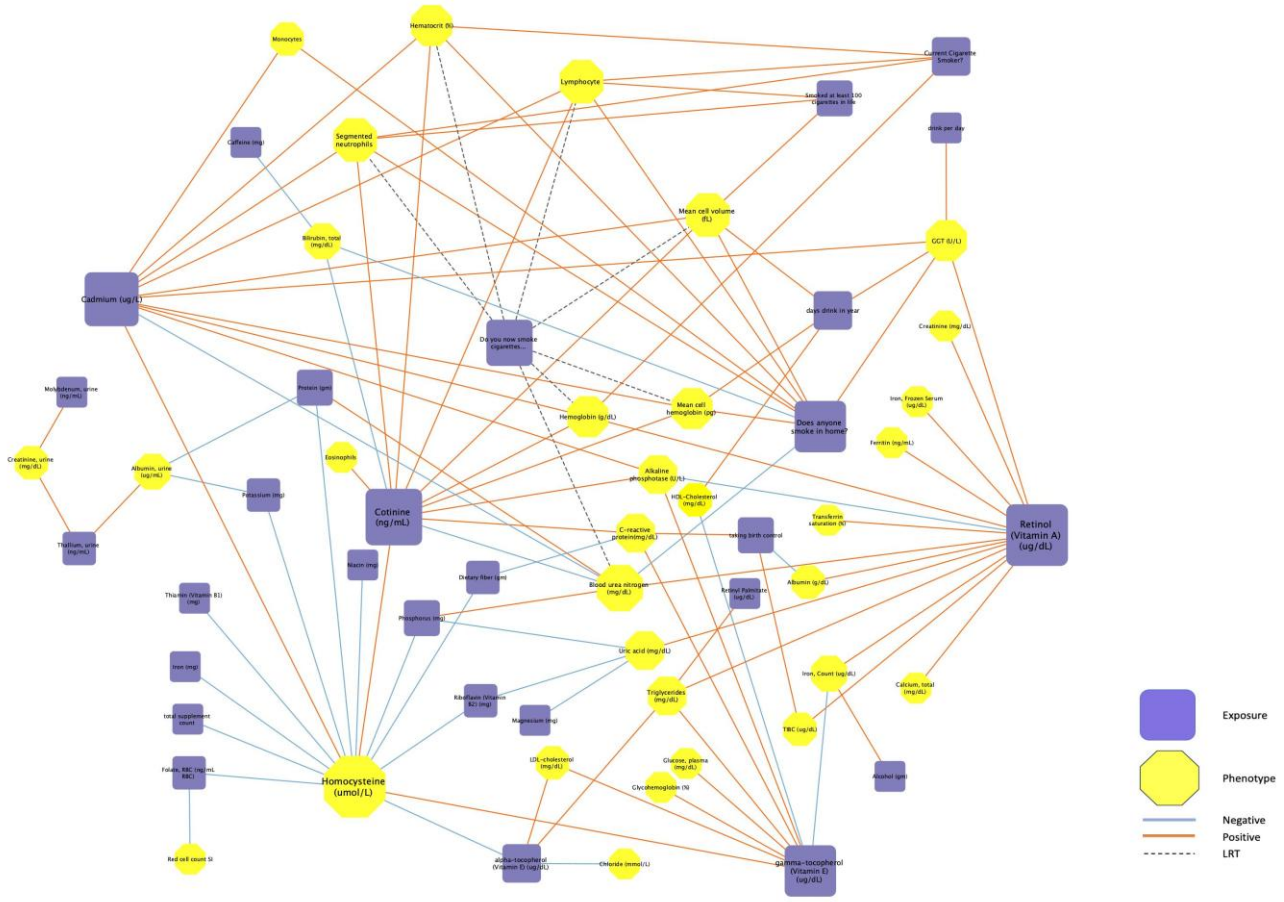

Fig. 2. Network-level exposome-phenome associations. A network plot with exposure nodes (purple squares) and phenotype nodes (yellow hexagons) provides a visual description of the various kinds of associations found[68]. The larger the node the more connections are associated with that node. Line color denotes the type of association. Blue and red connections signify negative and positive GLM coefficients, respectively. The black dashed connections signify a significant LRT because not a single beta value was assigned (See Methods). This plot was generated using Cytoscape software[69].

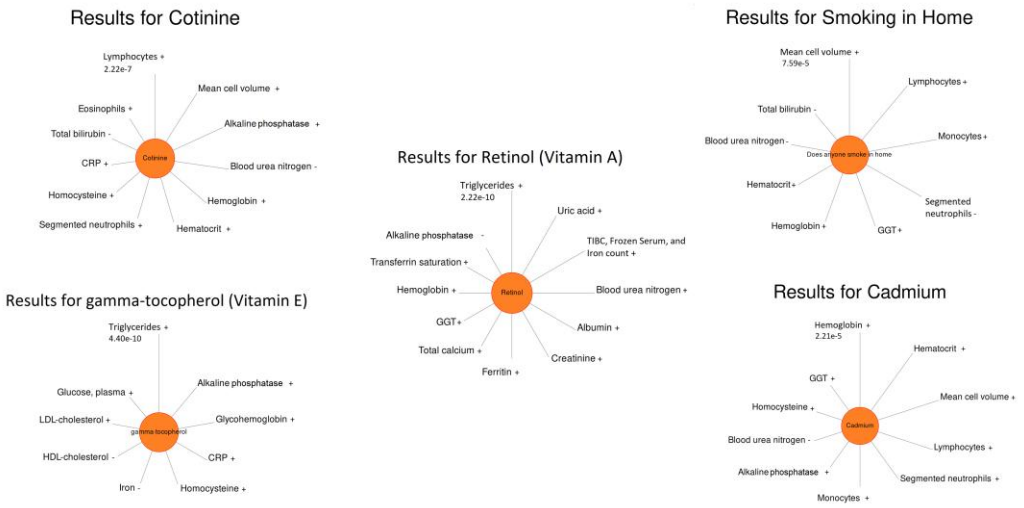

Fig. 3. Exposures with multiple cross-phenotype associations. Sun plot of results for exposures retinol, cotinine, cadmium, *Does anyone smoke in home?* and gamma-tocopherol. Replicated cross-phenotype associations plotted clockwise with the smallest p-value at the top. These exposures had the largest number of cross-phenotypic effects. The length of the line corresponds to the −log(p-value) for each association, and the longest line shows the most significant result and p-value. At the end of the lines, the positive (+) and negative (−) indicate the direction of the effect of each association. The sun plots were generated using PheWAS-View[70]**.**

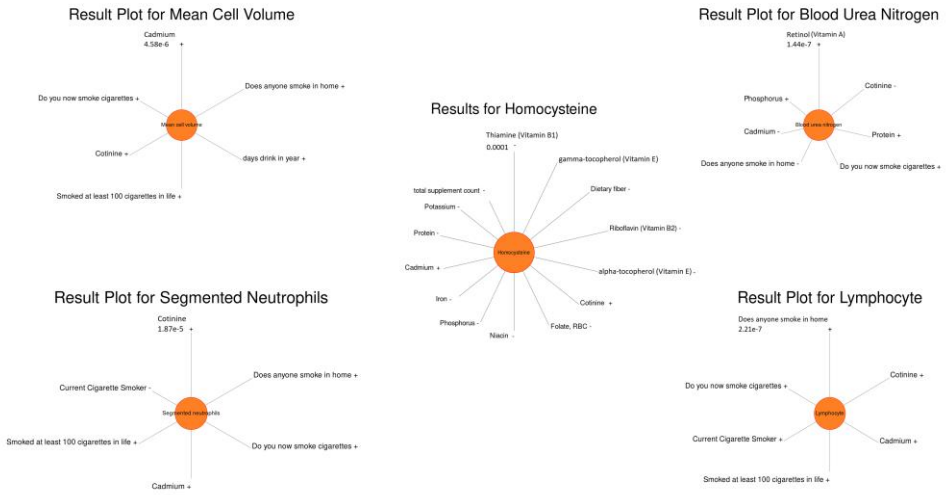

Fig. 4. Phenotypes with multiple cross-exposure associations. Sun plot of results for phenotypes homocysteine, lymphocyte count, segmented neutrophil count, blood urea nitrogen, and mean cell volume. Significant cross-exposure associations are plotted clockwise, with the smallest p-value at the top. These phenotypes exhibited the greatest cross-exposure connections. The length of the line corresponds to the −log(p-value) for each association, and the longest line shows the most significant result and p-value. At the end of the lines, the positive (+) and negative (−) indicate the direction of the effect of each association. The sun plots were generated using PheWAS-View[70].

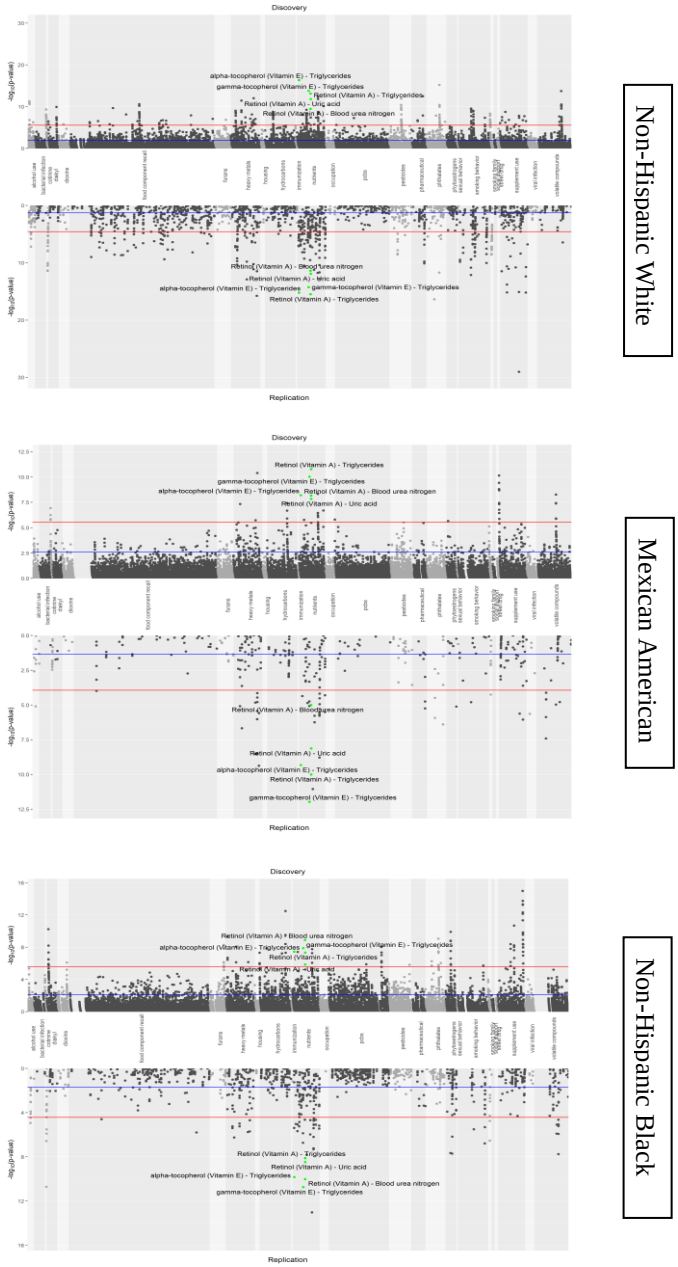

Fig. 5. Stratified PheEWAS results. The three Hudson plots are arranged by race/ethnicity strata: Non-Hispanic White, Mexican American, and Non-Hispanic Black. On the x-axis are exposure-phenotype associations, ordered by exposure category, and on the y-axis is the $-\log_{10}$(p-value). The red lines denote a Bonferroni-adjusted p-value of 0.05, and the blue lines denote the threshold for FDR of 0.1. The top replicating results in the nutrients category that are significant after correction are labeled. Hudson plots were created using an R package available at https://github.com/anastasia-lucas/hudson[67]

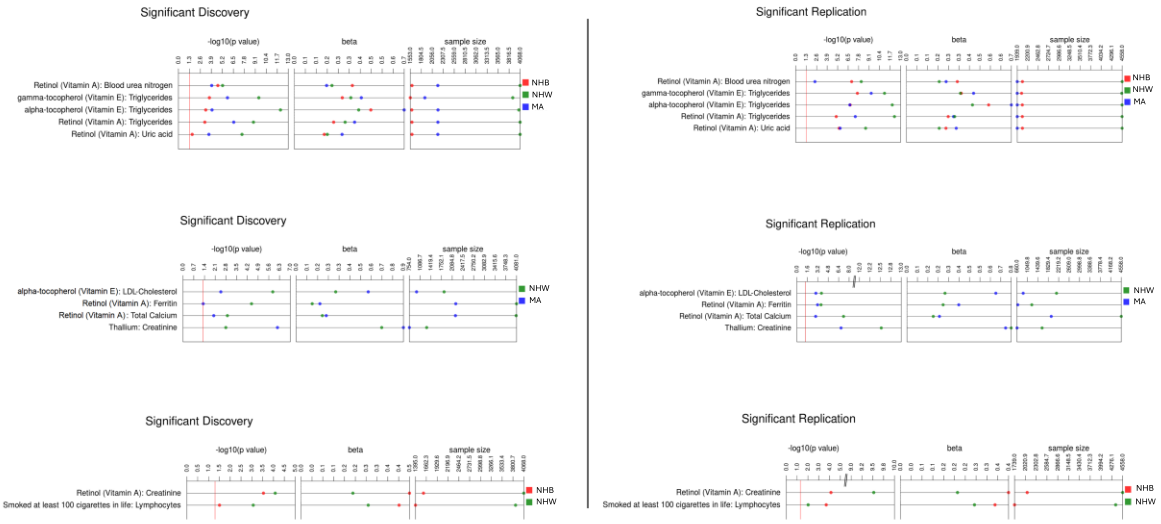

Fig. 6. Significant stratified PheEWAS results. This plot depicts exposure-phenotype associations with Bonferroni-adjusted p-values < 0.05 that overlap for each race-ethnicity stratum. Each color represents a stratified race-ethnicity stratum, red for Non-Hispanic Black (NHB), green for Non-Hispanic White (NHW), and blue for Mexican American (MA), as represented in the NHANES data displaying the same direction of effect. Presented also are the Bonferroni-adjusted p-value for each association and the comparison between all the beta values, indicating the magnitude of effect and varying sample sizes[70].

*Defined abbreviations found in the tables: Gamma-glutamyl transferase (GGT), Red blood cell (RBC), Low-density lipoprotein (LDL), High-density lipoprotein (HDL), Total iron binding capacity (TIBC)

Table 1: The top 106 replicating PheEWAS results.

Exposure	Phenotype	Exposure Category	Phenotype Category	Variable Type	Discovery Bonferroni p-value	Discovery Beta Coefficient	Replication Bonferroni p-value	Replication Beta Coefficient

Retinol (vitamin A) (µg/dL)	Albumin (g/dL)	nutrients	biochemistry	continuous	2.98E−06	1.94E−01	1.87E−07	1.59E−01
Taking birth control	Albumin (g/dL)	pharmaceutical	biochemistry	binary	1.21E−02	−4.31E−01	4.57E−05	−5.95E−01
Potassium (mg)	Albumin, urine (µg/mL)	food component recall	biochemistry	continuous	9.80E−03	−9.22E−02	4.27E−04	−1.07E−01
Protein (gm)	Albumin, urine (µg/mL)	food component recall	biochemistry	continuous	4.82E−02	−8.72E−02	3.25E−02	−7.94E−02
Thallium, urine (ng/mL)	Albumin, urine (µg/mL)	heavy metals	biochemistry	continuous	2.92E−02	2.98E−01	3.88E−07	3.22E−01
Cadmium (µg/L)	Alkaline c(U/L)	heavy metals	biochemistry	continuous	1.20E−02	9.74E−02	1.15E−02	6.27E−02
Cotinine (ng/mL)	Alkaline phosphatase (U/L)	cotinine	biochemistry	continuous	6.10E−05	1.02E−01	9.57E−04	7.97E−02
Gamma-tocopherol (vitamin E) (µg/dL)	Alkaline phosphatase (U/L)	nutrients	biochemistry	continuous	3.02E−04	1.37E−01	2.45E−05	1.07E−01
Retinol (vitamin A) (µg/dL)	Alkaline phosphatase (U/L)	nutrients	biochemistry	continuous	4.40E−03	−8.29E−02	1.81E−04	−9.66E−02
Caffeine (mg)	Bilirubin, total (mg/dL)	food component recall	biochemistry	continuous	1.19E−02	−6.39E−02	3.17E−02	−5.98E−02
Cotinine (ng/mL)	Bilirubin, total (mg/dL)	cotinine	biochemistry	continuous	2.31E−02	−1.14E−01	2.62E−03	−9.13E−02

Does anyone smoke in home?	Bilirubin, total (mg/dL)	smoking family	biochemistry	binary	4.30E−02	−2.26E−01	3.80E−05	−2.36E−01
Cadmium (µg/L)	Blood urea nitrogen (mg/dL)	heavy metals	biochemistry	continuous	2.60E−02	−8.99E−02	1.04E−03	−8.73E−02
Cotinine (ng/mL)	Blood urea nitrogen (mg/dL)	cotinine	biochemistry	continuous	6.12E−05	−1.22E−01	5.99E−08	−1.21E−01
Do you now smoke cigarettes?	Blood urea nitrogen (mg/dL)	smoking behavior	biochemistry	categorical	4.90E−03		1.95E−03	
Does anyone smoke in home?	Blood urea nitrogen (mg/dL)	smoking family	biochemistry	binary	1.33E−02	−2.16E−01	1.25E−06	−2.50E−01
Phosphorus (mg)	Blood urea nitrogen (mg/dL)	food component recall	biochemistry	continuous	4.07E−02	8.54E−02	7.08E−04	1.04E−01
Protein (gm)	Blood urea nitrogen (mg/dL)	food component recall	biochemistry	continuous	2.60E−04	1.33E−01	8.13E−07	1.58E−01
Retinol (vitamin A) (µg/dL)	Blood urea nitrogen (mg/dL)	nutrients	biochemistry	continuous	1.44E−07	2.45E−01	1.05E−09	2.37E−01
Retinol (vitamin A) (µg/dL)	Calcium, total (mg/dL)	nutrients	biochemistry	continuous	6.71E−05	2.24E−01	1.34E−08	2.01E−01
Alpha-tocopherol (vitamin E) (µg/dL)	Chloride (mmol/L)	nutrients	biochemistry	continuous	3.37E−02	−9.52E−02	1.85E−04	−1.31E−01

Cotinine (ng/mL)	C-reactive protein (CRP) (mg/dL)	cotinine	biochemistry	continuous	1.04E−02	9.93E−02	5.64E−03	8.04E−02
Dietary fiber (gm)	C-reactive protein (CRP) (mg/dL)	food component recall	biochemistry	continuous	2.69E−02	−6.82E−02	6.28E−03	−9.90E−02
Gamma-tocopherol (vitamin E) (µg/dL)	C-reactive protein (CRP) (mg/dL)	nutrients	biochemistry	continuous	7.85E−04	9.75E−02	4.73E−04	1.02E−01
Taking birth control	C-reactive protein (CRP) (mg/dL)	pharmaceutical	biochemistry	binary	1.86E−07	7.94E−01	4.20E−08	8.24E−01
Retinol (vitamin A) (µg/dL)	Creatinine (mg/dL)	nutrients	biochemistry	continuous	3.12E−06	2.52E−01	5.55E−10	2.29E−01
Molybdenum, urine (ng/mL)	Creatinine, urine (mg/dL)	heavy metals	biochemistry	continuous	1.67E−03	4.40E−01	2.76E−02	4.51E−01
Thallium, urine (ng/mL)	Creatinine, urine (mg/dL)	heavy metals	biochemistry	continuous	1.34E−03	6.01E−01	2.98E−07	6.07E−01
Days drink in year	Direct HDL-cholesterol (mg/dL)	alcohol use	biochemistry	continuous	1.02E−07	2.28E−01	5.83E−05	1.82E−01
Gamma-tocopherol (vitamin E) (µg/dL)	Direct HDL-cholesterol (mg/dL)	nutrients	biochemistry	continuous	1.38E−03	−1.11E−01	8.03E−03	−1.02E−01
Cotinine (ng/mL)	Eosinophils number	cotinine	blood	continuous	4.42E−02	6.58E−02	5.02E−04	8.35E−02

Retinol (vitamin A) (µg/dL)	Ferritin (ng/mL)	nutrients	biochemistry	continuous	5.00E−05	1.43E−01	1.62E−07	2.81E−01
Cadmium (µg/L)	GGT (U/L)	heavy metals	biochemistry	continuous	4.54E−02	1.07E−01	2.03E−02	8.65E−02
Days drink in year	GGT (U/L)	alcohol use	biochemistry	continuous	1.67E−04	1.77E−01	3.55E−04	1.69E−01
Does anyone smoke in home?	GGT (U/L)	smoking family	biochemistry	binary	1.66E−03	2.71E−01	8.38E−04	2.10E−01
Drinks per day	GGT (U/L)	alcohol use	biochemistry	continuous	8.70E−04	1.46E−01	3.61E−02	1.44E−01
Retinol (vitamin A) (µg/dL)	GGT (U/L)	nutrients	biochemistry	continuous	1.24E−04	1.78E−01	2.63E−09	1.90E−01
Gamma-tocopherol (vitamin E) (µg/dL)	Glucose, plasma (mg/dL)	nutrients	biochemistry	continuous	1.57E−02	1.68E−01	5.03E−03	1.37E−01
Gamma-tocopherol (vitamin E) (µg/dL)	Glycohemoglobin (%)	nutrients	biochemistry	continuous	6.13E−04	1.78E−01	2.43E−03	1.29E−01
Cadmium (µg/L)	Hematocrit (%)	heavy metals	blood	continuous	3.99E−05	1.18E−01	5.62E−08	1.29E−01
Cotinine (ng/mL)	Hematocrit (%)	cotinine	blood	continuous	4.39E−04	1.01E−01	4.56E−08	1.10E−01
Current cigarette smoker?	Hematocrit (%)	smoking behavior	blood	binary	1.48E−05	2.99E−01	3.29E−04	2.07E−01

Do you now smoke cigarettes?	Hematocrit (%)	smoking behavior	blood	categorical	1.00E−04		1.22E−04	
Does anyone smoke in home?	Hematocrit (%)	smoking family	blood	binary	9.98E−03	2.11E−01	1.51E−04	2.28E−01
Cadmium (µg/L)	Hemoglobin (g/dL)	heavy metals	blood	continuous	2.21E−05	1.25E−01	1.50E−08	1.34E−01
Cotinine (ng/mL)	Hemoglobin (g/dL)	cotinine	blood	continuous	3.16E−04	1.10E−01	1.54E−08	1.18E−01
Current cigarette smoker?	Hemoglobin (g/dL)	smoking behavior	blood	binary	2.26E−05	3.13E−01	3.00E−04	2.11E−01
Do you now smoke cigarettes?	Hemoglobin (g/dL)	smoking behavior	blood	categorical	8.35E−05		3.29E−05	
Retinol (vitamin A) (µg/dL)	Hemoglobin (g/dL)	nutrients	blood	continuous	4.44E−04	9.62E−02	1.42E−02	8.71E−02
Alpha-tocopherol (vitamin E) (µg/dL)	Homocysteine (µmol/L)	nutrients	biochemistry	continuous	1.86E−03	−9.80E−02	3.09E−06	−1.14E−01
Cadmium (µg/L)	Homocysteine (µmol/L)	heavy metals	biochemistry	continuous	4.13E−02	8.24E−02	5.21E−04	9.73E−02
Cotinine (ng/mL)	Homocysteine (µmol/L)	cotinine	biochemistry	continuous	2.72E−03	7.86E−02	8.11E−08	1.07E−01
Dietary fiber (gm)	Homocysteine (µmol/L)	food component recall	biochemistry	continuous	2.18E−04	−8.38E−02	5.53E−04	−1.21E−01

Folate, RBC (ng/mL RBC)	Homocysteine (µmol/L)	nutrients	biochemistry	continuous	3.12E−03	−1.54E−01	2.54E−08	−1.73E−01
Gamma-tocopherol (vitamin E) (µg/dL)	Homocysteine (µmol/L)	nutrients	biochemistry	continuous	9.08E−04	1.02E−01	2.23E−04	8.94E−02
Iron (mg)	Homocysteine (µmol/L)	food component recall	biochemistry	continuous	1.05E−04	−1.13E−01	6.14E−05	−1.33E−01
Niacin (mg)	Homocysteine (µmol/L)	food component recall	biochemistry	continuous	3.85E−03	−8.82E−02	1.72E−03	−9.91E−02
Phosphorus (mg)	Homocysteine (µmol/L)	food component recall	biochemistry	continuous	1.09E−02	−8.83E−02	1.24E−03	−1.08E−01
Potassium (mg)	Homocysteine (µmol/L)	food component recall	biochemistry	continuous	1.71E−02	−6.93E−02	5.88E−04	−1.02E−01
Protein (gm)	Homocysteine (µmol/L)	food component recall	biochemistry	continuous	3.43E−03	−8.33E−02	9.07E−03	−1.00E−01
Riboflavin (vitamin B2) (mg)	Homocysteine (µmol/L)	food component recall	biochemistry	continuous	1.38E−03	−1.07E−01	7.37E−05	−1.11E−01
Thiamin (vitamin B1) (mg)	Homocysteine (µmol/L)	food component recall	biochemistry	continuous	1.35E−04	−9.47E−02	9.00E−05	−1.24E−01
Total supplement count	Homocysteine (µmol/L)	supplement use	biochemistry	continuous	2.11E−04	−1.07E−01	4.32E−06	−1.05E−01

Alcohol (gm)	Iron, count (µg/dL)	food component recall	biochemistry	continuous	3.26E−03	8.79E−02	6.09E−04	9.62E−02
Gamma-tocopherol (vitamin E) (µg/dL)	Iron, count (µg/dL)	nutrients	biochemistry	continuous	1.23E−03	−9.01E−02	3.59E−05	−8.92E−02
Retinol (vitamin A) (µg/dL)	Iron, count (µg/dL)	nutrients	biochemistry	continuous	9.10E−08	1.68E−01	1.62E−08	1.91E−01
Retinol (vitamin A) (µg/dL)	Iron, Frozen Serum (µg/dL)	nutrients	blood	continuous	3.04E−08	1.68E−01	2.98E−07	2.53E−01
Alpha-tocopherol (vitamin E) (µg/dL)	LDL-cholesterol (mg/dL)	nutrients	biochemistry	continuous	4.02E−06	3.31E−01	2.88E−06	3.22E−01
Gamma-tocopherol (vitamin E) (µg/dL)	LDL-cholesterol (mg/dL)	nutrients	biochemistry	continuous	2.35E−03	2.27E−01	8.79E−05	2.29E−01
Cadmium (µg/L)	Lymphocyte number	heavy metals	blood	continuous	4.43E−04	1.28E−01	1.87E−06	1.31E−01
Cotinine (ng/mL)	Lymphocyte number	cotinine	blood	continuous	2.22E−07	1.78E−01	2.68E−08	1.57E−01
Current cigarette smoker?	Lymphocyte number	smoking behavior	blood	binary	4.58E−06	4.77E−01	1.88E−02	2.32E−01
Do you now smoke cigarettes?	Lymphocyte number	smoking behavior	blood	categorical	1.56E−04		1.48E−05	

Does anyone smoke in home?	Lymphocyte number	smoking family	blood	binary	1.77E−04	3.28E−01	1.05E−05	3.19E−01
Smoked at least 100 cigarettes in life	Lymphocyte number	smoking behavior	blood	binary	2.88E−04	2.86E−01	3.64E−04	2.47E−01
Cadmium (µg/L)	Mean cell hemoglobin (pg)	heavy metals	blood	continuous	2.30E−04	9.86E−02	4.29E−05	9.52E−02
Cotinine (ng/mL)	Mean cell hemoglobin (pg)	cotinine	blood	continuous	1.46E−04	1.33E−01	1.22E−06	1.18E−01
Days drink in year	Mean cell hemoglobin (pg)	alcohol use	blood	continuous	7.28E−06	1.59E−01	3.08E−03	1.19E−01
Do you now smoke cigarettes?	Mean cell hemoglobin (pg)	smoking behavior	blood	categorical	4.99E−02		8.19E−05	
Does anyone smoke in home?	Mean cell hemoglobin (pg)	smoking family	blood	binary	3.35E−03	2.84E−01	3.36E−02	2.11E−01
Cadmium (µg/L)	Mean cell volume (fL)	heavy metals	blood	continuous	1.34E−04	1.01E−01	1.25E−04	1.00E−01
Cotinine (ng/mL)	Mean cell volume (fL)	cotinine	blood	continuous	2.86E−05	1.39E−01	4.67E−06	1.22E−01
Days drink in year	Mean cell volume (fL)	alcohol use	blood	continuous	1.64E−05	1.64E−01	9.21E−04	1.19E−01
Do you now smoke cigarettes?	Mean cell volume (fL)	smoking behavior	blood	categorical	1.33E−02		2.94E−05	
Does anyone smoke in home?	Mean cell volume (fL)	smoking family	blood	binary	7.59E−05	3.13E−01	9.68E−03	2.44E−01

Smoked at least 100 cigarettes in life	Mean cell volume (fL)	smoking behavior	blood	binary	6.37E−03	2.37E−01	7.62E−09	2.50E−01
Cadmium (µg/L)	Monocyte number	heavy metals	blood	continuous	7.18E−03	1.07E−01	8.18E−03	8.13E−02
Does anyone smoke in home?	Monocyte number	smoking family	blood	binary	2.25E−04	2.81E−01	4.10E−02	1.91E−01
Folate, RBC (ng/mL RBC)	Red cell count SI	nutrients	blood	continuous	3.28E−02	−6.92E−02	8.59E−06	−1.07E−01
Cadmium (µg/L)	Segmented neutrophils number	heavy metals	blood	continuous	2.92E−03	1.19E−01	1.61E−05	1.35E−01
Cotinine (ng/mL)	Segmented neutrophils number	cotinine	blood	continuous	1.19E−03	1.62E−01	7.24E−07	1.65E−01
Current cigarette smoker?	Segmented neutrophils number	smoking behavior	blood	binary	1.88E−05	4.37E−01	2.85E−04	2.25E−01
Do you now smoke cigarettes?	Segmented neutrophils number	smoking behavior	blood	categorical	1.56E−04		5.85E−07	
Does anyone smoke in home?	Segmented neutrophils number	smoking family	blood	binary	5.15E−04	3.35E−01	8.40E−07	3.25E−01
Smoked at least 100 cigarettes in life	Segmented neutrophils number	smoking behavior	blood	binary	4.96E−03	2.21E−01	8.34E−06	2.31E−01

Retinol (vitamin A) (µg/dL)	TIBC (µg/dL)	nutrients	biochemistry	continuous	4.45E−02	1.76E−01	1.27E−04	1.93E−01
Taking birth control	TIBC (µg/dL)	pharmaceutical	biochemistry	binary	3.67E−05	7.00E−01	2.64E−07	7.87E−01
Retinol (vitamin A) (µg/dL)	Transferrin saturation (%)	nutrients	biochemistry	continuous	2.69E−03	9.55E−02	3.79E−04	1.69E−01
Alpha-tocopherol (vitamin E) (µg/dL)	Triglycerides (mg/dL)	nutrients	biochemistry	continuous	1.17E−11	4.25E−01	8.05E−13	4.57E−01
Gamma-tocopherol (vitamin E) (µg/dL)	Triglycerides (mg/dL)	nutrients	biochemistry	continuous	4.40E−10	3.71E−01	1.64E−12	3.69E−01
Retinol (vitamin A) (µg/dL)	Triglycerides (mg/dL)	nutrients	biochemistry	continuous	2.22E−10	3.06E−01	9.30E−12	3.07E−01
Retinyl palmitate (µg/dL)	Triglycerides (mg/dL)	nutrients	biochemistry	continuous	3.24E−05	2.73E−01	1.20E−05	3.27E−01
Magnesium (mg)	Uric acid (mg/dL)	food component recall	biochemistry	continuous	4.98E−02	−6.17E−02	3.50E−02	−5.90E−02
Phosphorus (mg)	Uric acid (mg/dL)	food component recall	biochemistry	continuous	2.24E−02	−7.27E−02	4.14E−02	−6.71E−02
Retinol (vitamin A) (µg/dL)	Uric acid (mg/dL)	nutrients	biochemistry	continuous	1.71E−08	2.08E−01	2.12E−10	2.30E−01

Riboflavin (vitamin B2) (mg)	Uric acid (mg/dL)	food component recall	biochemistry	continuous	8.16E−04	−8.02E−02	2.77E−04	−8.25E−02

Replicated environment-phenotype associations with their discovery and replication Bonferroni-adjusted p-values, categories, data types, and effect sizes.

Table 2: The stratified race/ethnicity PheEWAS results.

Exposure	Phenotype	Group	Significant from Original Dataset	Variable Type	Discovery Bonferroni p-value	Discovery Beta	Replication Bonferroni p-value	Replication Beta
Lead, urine (ng/mL)	Creatinine, urine (mg/dL)	NHW	No	continuous	6.57E−04	7.18E−01	2.36E−10	1.14E+00
Trans-lycopene (µg/dL)	LDL-cholesterol (mg/dL)	NHW	No	continuous	7.50E−04	3.60E−01	3.21E−09	3.54E−01
Mercury, urine (ng/mL)	Creatinine, urine (mg/dL)	NHW	No	continuous	1.35E−03	8.60E−01	1.07E−03	6.41E−01
Lead, urine (ng/mL)	Albumin, urine (µg/mL)	NHW	No	continuous	1.28E−02	3.78E−01	4.35E−03	7.05E−01
Lead, urine (ng/mL)	Albumin, urine (mg/L) SI	NHW	No	continuous	1.29E−02	3.78E−01	2.78E−07	6.75E−01
Lutein and zeaxanthin (µg/dL)	Direct HDL-cholesterol (mg/dL)	NHW	No	continuous	1.80E−02	2.30E−01	5.48E−07	1.87E−01

Does anyone smoke in home?	Hemoglobin (g/dL)	NHW	No	binary	2.81E−02	2.53E−01	1.29E−04	2.30E−01
Thallium, urine (ng/mL)	Albumin, urine (mg/L) SI	NHW	No	continuous	2.93E−02	3.45E−01	3.85E−02	4.13E−01
Trans-beta-carotene (µg/dL)	Homocysteine (µmol/L)	NHW	No	continuous	3.31E−02	−9.89E−02	6.56E−06	−1.19E−01
Cadmium, urine (ng/mL)	Creatinine, urine (mg/dL)	MA	No	continuous	7.83E−04	1.27E+00	8.77E−05	6.67E−01
Platinum, urine (ng/mL)	Eosinophils number	MA	No	continuous	3.12E−02	−3.50E+01	3.89E−04	−2.72E+00
Alpha-tocopherol (µg/dL)	Triglycerides (mg/dL)	NHB	Yes	continuous	8.33E−13	3.99E−01	1.09E−12	4.33E−01
Gamma-tocopherol (vitamin E) (µg/dl)	Triglycerides (mg/dL)	NHB	Yes	continuous	3.07E−10	3.50E−01	9.84E−12	3.54E−01
Retinol (vitamin A) (µg/dL)	Creatinine (mg/dL)	NHB	Yes	continuous	7.86E−09	1.64E−01	2.60E−05	2.24E−01
Retinol (vitamin A) (µg/dL)	Triglycerides (mg/dL)	NHB	Yes	continuous	8.32E−05	2.19E−01	3.01E−10	2.05E−01
Retinol (vitamin A) (µg/dL)	Blood urea nitrogen (mg/dL)	NHB	Yes	continuous	2.91E−04	4.59E−01	7.40E−05	4.15E−01
Retinol (vitamin A) (µg/dL)	Uric acid (mg/dL)	NHB	Yes	continuous	6.04E−04	2.10E−01	3.09E−04	2.34E−01

Smoked at least 100 cigarettes in life	Lymphocyte number	NHB	Yes	binary	8.61E−04	2.84E−01	9.51E−03	2.38E−01
Retinol (vitamin A) (µg/dL)	Albumin (g/dL)	NHW	Yes	continuous	1.32E−09	3.15E−01	5.80E−13	3.16E−01
Taking birth control	C-reactive protein (CRP) (mg/dL)	NHW	Yes	binary	5.81E−09	8.27E−01	2.99E−07	8.63E−01
Retinol (vitamin A) (µg/dL)	Creatinine (mg/dL)	NHW	Yes	continuous	2.99E−08	2.04E−01	2.16E−09	2.15E−01
Days drink in year	Direct HDL-cholesterol (mg/dL)	NHW	Yes	continuous	9.44E−08	2.34E−01	1.21E−04	1.81E−01
Days drink in year	GGT (U/L)	NHW	Yes	continuous	2.31E−07	1.59E−01	2.11E−02	1.10E−01
Retinol (vitamin A) (µg/dL)	Triglycerides (mg/dL)	NHW	Yes	continuous	2.84E−07	3.73E−01	4.09E−08	3.09E−01
Days drink in year	Mean cell volume (fL)	NHW	Yes	continuous	4.36E−07	1.63E−01	1.75E−02	1.06E−01
Gamma-tocopherol (vitamin E) (µg/dl)	Triglycerides (mg/dL)	NHW	Yes	continuous	1.57E−06	4.15E−01	4.41E−10	4.40E−01
Dietary fiber (gm)	Homocysteine (µmol/L)	NHW	Yes	continuous	4.05E−06	−9.53E−02	4.81E−06	−1.51E−01
Current cigarette smoker?	Hemoglobin (g/dL)	NHW	Yes	binary	5.19E−06	3.32E−01	3.59E−04	2.13E−01

Cotinine (ng/mL)	Lymphocyte number	NHW	Yes	continuous	8.90E−06	1.83E−01	8.32E−09	1.56E−01
Current cigarette smoker?	Hematocrit (%)	NHW	Yes	binary	9.40E−06	3.50E−01	2.99E−04	2.19E−01
Current cigarette smoker?	Segmented neutrophils number	NHW	Yes	binary	1.11E−05	4.91E−01	2.19E−03	2.34E−01
Cadmium (µg/L)	Hemoglobin (g/dL)	NHW	Yes	continuous	1.74E−05	1.30E−01	8.94E−08	1.35E−01
Do you now smoke cigarettes?	Segmented neutrophils number	NHW	Yes	categorical	1.84E−05	NA	3.37E−04	NA
Retinol (vitamin A) (µg/dL)	Blood urea nitrogen (mg/dL)	NHW	Yes	continuous	2.15E−05	3.59E−01	1.17E−07	3.34E−01
Do you now smoke cigarettes?	Hematocrit (%)	NHW	Yes	categorical	2.52E−05	NA	8.46E−05	NA
Current cigarette smoker?	Lymphocyte number	NHW	Yes	binary	3.43E−05	5.09E−01	1.57E−02	2.67E−01
Cadmium (µg/L)	Hematocrit (%)	NHW	Yes	continuous	4.04E−05	1.21E−01	1.09E−06	1.29E−01
Retinol (vitamin A) (µg/dL)	TIBC (µg/dL)	NHW	Yes	continuous	4.34E−05	1.86E−01	3.96E−09	2.00E−01
Retinol (vitamin A) (µg/dL)	Calcium, total (mg/dL)	NHW	Yes	continuous	7.40E−05	1.79E−01	2.39E−05	1.43E−01

Cadmium (µg/L)	Mean cell volume (fL)	NHW	Yes	continuous	7.67E−05	1.13E−01	2.82E−05	9.51E−02
Does anyone smoke in home?	Segmented neutrophils number	NHW	Yes	binary	8.83E−05	3.83E−01	4.86E−06	3.33E−01
Alpha-tocopherol (µg/dL)	Triglycerides (mg/dL)	NHW	Yes	continuous	1.06E−04	6.81E−01	1.95E−07	6.88E−01
Retinol (vitamin A) (µg/dL)	Iron, count (µg/dL)	NHW	Yes	continuous	1.16E−04	2.00E−01	4.11E−03	2.60E−01
Do you now smoke cigarettes?	Hemoglobin (g/dL)	NHW	Yes	categorical	1.28E−04	NA	2.70E−06	NA
Iron (mg)	Homocysteine (µmol/L)	NHW	Yes	continuous	1.29E−04	−1.19E−01	2.02E−05	−1.49E−01
Cotinine (ng/mL)	Blood urea nitrogen (mg/dL)	NHW	Yes	continuous	1.43E−04	−1.33E−01	6.41E−08	−1.33E−01
Cotinine (ng/mL)	Hematocrit (%)	NHW	Yes	continuous	2.18E−04	1.15E−01	5.81E−08	1.11E−01
Thiamin (vitamin B1) (mg)	Homocysteine (µmol/L)	NHW	Yes	continuous	2.21E−04	−1.01E−01	1.23E−05	−1.37E−01
Taking birth control	TIBC (µg/dL)	NHW	Yes	binary	2.44E−04	7.69E−01	1.39E−07	8.47E−01
Retinol (vitamin A) (µg/dL)	Iron, Frozen Serum (µg/dL)	NHW	Yes	continuous	2.48E−04	2.96E−01	3.07E−06	3.26E−01
Cadmium (µg/L)	Mean cell hemoglobin (pg)	NHW	Yes	continuous	2.52E−04	1.10E−01	3.28E−05	9.26E−02

Does anyone smoke in home?	Lymphocyte number	NHW	Yes	binary	2.62E−04	3.58E−01	5.26E−05	3.36E−01
Total supplement count	Homocysteine (µmol/L)	NHW	Yes	continuous	2.77E−04	−1.03E−01	7.46E−06	−1.01E−01
Cotinine (ng/mL)	Mean cell hemoglobin (pg)	NHW	Yes	continuous	2.84E−04	1.05E−01	7.83E−03	8.68E−02
Gamma-tocopherol (vitamin E) (µg/dl)	Iron, count (µg/dL)	NHW	Yes	continuous	2.97E−04	−9.27E−02	1.23E−03	−8.57E−02
Gamma-tocopherol (vitamin E) (µg/dl)	Alkaline phosphatase (U/L)	NHW	Yes	continuous	3.39E−04	9.99E−02	1.72E−03	8.20E−02
Cotinine (ng/mL)	Alkaline phosphatase (U/L)	NHW	Yes	continuous	3.74E−04	1.35E−01	2.55E−05	1.13E−01
Cotinine (ng/mL)	Mean cell volume (fL)	NHW	Yes	continuous	5.03E−04	9.27E−02	1.39E−07	1.14E−01
Do you now smoke cigarettes?	Lymphocyte number	NHW	Yes	categorical	5.12E−04	NA	1.11E−05	NA
Cotinine (ng/mL)	Homocysteine (µmol/L)	NHW	Yes	continuous	6.28E−04	1.22E−01	7.26E−09	1.18E−01
Retinyl palmitate (µg/dL)	Triglycerides (mg/dL)	NHW	Yes	continuous	7.38E−04	2.78E−01	9.29E−05	3.14E−01

Retinol (vitamin A) (µg/dL)	Ferritin (ng/mL)	NHW	Yes	continuous	7.54E−04	2.44E−01	9.43E−06	2.74E−01
Gamma-tocopherol (vitamin E) (µg/dl)	Homocysteine (µmol/L)	NHW	Yes	continuous	9.48E−04	1.46E−01	6.04E−04	1.09E−01
Thallium, urine (ng/mL)	Creatinine, urine (mg/dL)	NHW	Yes	continuous	1.61E−03	6.82E−01	3.01E−13	7.54E−01
Does anyone smoke in home?	Hematocrit (%)	NHW	Yes	binary	1.88E−03	3.13E−01	8.34E−03	2.27E−01
Does anyone smoke in home?	Mean cell volume (fL)	NHW	Yes	binary	2.38E−03	2.45E−01	2.99E−05	2.33E−01
Do you now smoke cigarettes?	Blood urea nitrogen (mg/dL)	NHW	Yes	categorical	2.77E−03	NA	1.93E−03	NA
Days drink in year	Mean cell hemoglobin (pg)	NHW	Yes	continuous	2.83E−03	1.56E−01	3.00E−03	1.54E−01
Protein (gm)	Blood urea nitrogen (mg/dL)	NHW	Yes	continuous	3.01E−03	1.62E−01	7.45E−07	1.82E−01
Gamma-tocopherol (vitamin E) (µg/dl)	C-reactive protein (CRP) (mg/dL)	NHW	Yes	continuous	3.33E−03	1.04E−01	5.89E−03	1.01E−01
Alpha-tocopherol (µg/dL)	LDL-cholesterol (mg/dL)	NHW	Yes	continuous	3.41E−03	5.74E−01	1.26E−03	6.42E−01
Potassium (mg)	Albumin, urine (µg/mL)	NHW	Yes	continuous	4.23E−03	−1.04E−01	1.74E−03	−1.19E−01

Cadmium (µg/L)	Lymphocyte number	NHW	Yes	continuous	4.24E−03	1.26E−01	3.44E−07	1.35E−01
Riboflavin (vitamin B2) (mg)	Uric acid (mg/dL)	NHW	Yes	continuous	4.25E−03	−8.27E−02	9.30E−03	−7.94E−02
Cotinine (ng/mL)	Hemoglobin (g/dL)	NHW	Yes	continuous	4.44E−03	1.28E−01	1.98E−05	1.09E−01
Alpha-tocopherol (µg/dL)	Homocysteine (µmol/L)	NHW	Yes	continuous	5.26E−03	−1.07E−01	1.33E−06	−1.14E−01
Caffeine (mg)	Bilirubin, total (mg/dL)	NHW	Yes	continuous	5.27E−03	−6.39E−02	8.76E−05	−7.19E−02
Cadmium (µg/L)	Blood urea nitrogen (mg/dL)	NHW	Yes	continuous	5.38E−03	−1.04E−01	2.28E−03	−9.65E−02
Does anyone smoke in home?	GGT (U/L)	NHW	Yes	binary	7.05E−03	3.03E−01	3.61E−03	2.06E−01
Gamma-tocopherol (vitamin E) (µg/dl)	Direct HDL-cholesterol (mg/dL)	NHW	Yes	continuous	9.62E−03	−1.02E−01	1.17E−02	−1.12E−01
Retinol (vitamin A) (µg/dL)	Uric acid (mg/dL)	NHW	Yes	continuous	9.88E−03	2.39E−01	1.35E−03	2.35E−01
Cadmium (µg/L)	Segmented neutrophils number	NHW	Yes	continuous	1.35E−02	1.30E−01	3.43E−05	1.37E−01
Cotinine (ng/mL)	Segmented neutrophils number	NHW	Yes	continuous	1.40E−02	1.68E−01	1.66E−07	1.63E−01

Does anyone smoke in home?	Blood urea nitrogen (mg/dL)	NHW	Yes	binary	1.51E−02	−2.61E−01	6.75E−06	−2.90E−01
Gamma-tocopherol (vitamin E) (µg/dl)	Glycohemoglobin (%)	NHW	Yes	continuous	2.13E−02	1.64E−01	1.35E−02	1.12E−01
Retinol (vitamin A) (µg/dL)	GGT (U/L)	NHW	Yes	continuous	2.34E−02	1.86E−01	4.18E−06	2.59E−01
Folate, RBC (ng/mL RBC)	Homocysteine (µmol/L)	NHW	Yes	continuous	2.46E−02	−1.56E−01	4.79E−08	−1.71E−01
Riboflavin (vitamin B2) (mg)	Homocysteine (µmol/L)	NHW	Yes	continuous	2.65E−02	−1.08E−01	3.63E−05	−1.30E−01
Thallium, urine (ng/mL)	Albumin, urine (µg/mL)	NHW	Yes	continuous	2.74E−02	3.45E−01	5.94E−09	4.04E−01
Do you now smoke cigarettes?	Mean cell volume (fL)	NHW	Yes	categorical	3.07E−02	NA	1.65E−05	NA
Smoked at least 100 cigarettes in life	Lymphocyte number	NHW	Yes	binary	3.10E−02	4.15E−01	2.04E−04	3.63E−01
Molybdenum, urine (ng/mL)	Creatinine, urine (mg/dL)	NHW	Yes	continuous	3.29E−02	4.33E−01	1.33E−08	5.66E−01
Smoked at least 100 cigarettes in life	Segmented neutrophils number	NHW	Yes	binary	3.64E−02	2.45E−01	5.39E−05	2.32E−01

Smoked at least 100 cigarettes in life	Mean cell volume (fL)	NHW	Yes	binary	3.66E−02	2.27E−01	1.66E−07	2.53E−01
Cotinine (ng/mL)	Eosinophils number	NHW	Yes	continuous	3.86E−02	7.83E−02	5.46E−04	9.52E−02
Do you now smoke cigarettes?	Mean cell hemoglobin (pg)	NHW	Yes	categorical	4.06E−02	NA	1.04E−04	NA
Protein (gm)	Homocysteine (µmol/L)	NHW	Yes	continuous	4.80E−02	−8.58E−02	1.20E−02	−1.23E−01
Retinol (vitamin A) (µg/dL)	Blood urea nitrogen (mg/dL)	MA	Yes	continuous	2.74E−08	1.64E−01	1.37E−06	1.77E−01
Thallium, urine (ng/mL)	Creatinine, urine (mg/dL)	MA	Yes	continuous	6.84E−07	8.55E−01	1.72E−07	7.13E−01
Alpha-tocopherol (µg/dL)	LDL-cholesterol (mg/dL)	MA	Yes	continuous	1.38E−06	3.15E−01	1.77E−04	2.75E−01
Retinol (vitamin A) (µg/dL)	Uric acid (mg/dL)	MA	Yes	continuous	5.95E−06	2.33E−01	7.52E−09	2.15E−01
Retinol (vitamin A) (µg/dL)	Calcium, total (mg/dL)	MA	Yes	continuous	3.38E−05	1.27E−01	2.04E−04	2.61E−01
Gamma-tocopherol (vitamin E) (µg/dl)	Triglycerides (mg/dL)	MA	Yes	continuous	2.16E−04	2.97E−01	2.25E−08	3.60E−01

Alpha-tocopherol (µg/dL)	Triglycerides (mg/dL)	MA	Yes	continuous	5.87E−04	4.74E−01	1.78E−07	5.39E−01
Retinol (vitamin A) (µg/dL)	Ferritin (ng/mL)	MA	Yes	continuous	1.27E−03	2.09E−01	7.15E−08	1.90E−01
Retinol (vitamin A) (µg/dL)	Triglycerides (mg/dL)	MA	Yes	continuous	4.83E−02	1.89E−01	6.89E−04	3.74E−01

The 106 environment-phenotype associations with their discovery and replication Bonferroni-adjusted p-values and effect sizes. They are categorized by race/ethnicity groups in column three and whether the association was replicated from the original dataset.

Table 3: Demographics of the discovery and replication samples.

	Total Sample Size	*Male*	*Female*	*Non-Hispanic White*	*Non-Hispanic Black*	*Mexican American*	*Mean Age*	*Mean BMI*
Discovery (Survey Cycles 1 and 2)	9063	4724	4339	4260	1751	2285	45.4	27.9
Replication (Survey Cycles 3 and 4)	9874	5123	4751	4851	2264	2054	45.8	28.3