Integrated exposomic analysis of lipid phenotypes: Leveraging GE.db in environment by environment interaction studies

ANDRE LUIS GARAO RICO and NICOLE PALMIERO
Department of Genetics, University of Pennsylvania, 3700 Hamilton Walk
Philadelphia, PA 19104, USA
Email: andreluis.rico@pennmedicine.upenn.edu, nicole.palmiero@pennmedicine.upenn.edu

MARYLYN D. RITCHIE
Department of Genetics, University of Pennsylvania, 3700 Hamilton Walk
Philadelphia, PA 19104, USA
Email: marylyn@pennmedicine.upenn.edu

MOLLY A. HALL
Department of Genetics, University of Pennsylvania, 3700 Hamilton Walk
Philadelphia, PA 19104, USA
Email: molly.hall@pennmedicine.upenn.edu

Gene-environment interaction (GxE) studies provide insights into the interplay between genetics and the environment but often overlook multiple environmental factors' synergistic effects. This study encompasses the use of environment by environment interaction (ExE) studies to explore interactions among environmental factors affecting lipid phenotypes (e.g., HDL, LDL, and total cholesterol, and triglycerides), which are crucial for disease risk assessment. We developed a novel curated knowledge base, GE.db, integrating genomic and exposomic interactions. In this study, we filtered NHANES exposure variables (available 1999-2018) to identify significant ExE using GE.db. From 101,316 participants and 77 exposures, we identified 263 statistically significant interactions (FDR $p < 0.1$) in discovery and replication datasets, with 21 interactions significant for HDL-C (Bonferroni $p < 0.05$). Notable interactions included docosapentaenoic acid (22:5n-3) (DPA) - arachidic acid (20:0), stearic acid (18:0) - arachidic acid (20:0), and blood 2,5-dimethyfuran - blood benzene associated with HDL-C levels. These findings underscore GE.db's role in enhancing -omics research efficiency and highlight the complex impact of environmental exposures on lipid metabolism, informing future health strategies.

Keywords: Knowledge-Based Filtering; Interaction Analysis; Exposome; Lipid Metabolism

1. Introduction

Understanding the intricate interplay between genetics and the environment is pivotal in unraveling the complexities of human traits and diseases. While gene-environment interaction (GxE) studies have provided valuable insights into how genetic variants interact with environmental factors, they often overlook the synergistic effects of multiple environmental variables[1,2]. This limitation

necessitates the need for utilizing environment by environment interaction (ExE) studies, which explore how different environmental factors interact with each other to influence phenotypic outcomes. The outcomes of interest used in this study are lipid traits, including high-density lipoprotein-cholesterol (HDL-C), low-density lipoprotein-cholesterol (LDL-C), total cholesterol, and triglycerides, all of which are important risk factors for a multitude of diseases[3–5]. It is well established that lipid traits are influenced by a variety of factors, including genetic inheritance, environmental and occupational exposures, medication use, ethnicity, and sex[6,7]. In this study, we define environmental exposure as any physical, chemical, or biological agent that someone is exposed to and has potential to cause a wide range of health effects. The dietary exposures in this study refer to the intake of nutrients that can either benefit, harm, or have no effect on one's health.

Due to the scale of risk variables available in contemporary cohort and biobank datasets, many researchers perform variable selection (or filtering) prior to statistical or computational modeling. The shift towards knowledge-based filtering in these studies has been shown to be an effective alternative to main effect filtering (whereby variables are filtered based on having a statistically significant independent effect), especially for variables that only exhibit an effect in the context of another variable. The incorporation of prior biological knowledge to prioritize genetic variants that are more likely to interact with one another has revealed numerous GxG for complex diseases[8-11]. However, these studies have been restricted to knowledge about genes and have not included knowledge of the biological relationship between exposures. Thus, we propose that ExE coupled with knowledge-based filtering represents a promising approach to further elucidate the complexities of ExE in health and disease. This paper introduces the Gene x Exposome database (GE.db) module of the Integrative Genome-Exposome Method (IGEM) system[12], a knowledge base of genomic and exposomic interactions derived from various public databases [see Methods]. The development of GE.db aims to leverage prior knowledge to filter high-volume research datasets, retaining only variables with known biological relationships. This approach significantly reduces the number of variables for analysis, conserves computational resources and processing time, and minimizes type I errors following multiple testing corrections.

To demonstrate the utility of GE.db, we conducted an ExE analysis with lipid traits using the National Health and Nutritional Examination Survey (NHANES)[13] data from 1999-2018. By focusing on an exposome-wide interaction approach and utilizing GE.db, this research can provide important insights for the prevention and management of lipid-based health risk factors. Additionally, this study highlights the potential of GE.db to enhance the efficiency and accuracy of -omics research by providing a knowledge base resource for filtering datasets based on known interactions, thereby facilitating more focused and reliable statistical and computational analyses.

2. Methods

2.1 NHANES Dataset

The National Health and Nutrition Examination Survey (NHANES) is an ongoing initiative conducted by the Centers for Disease Control and Prevention (CDC) aimed at evaluating the health and nutritional status of the U.S. population[14]. Its primary objectives include identifying risk factors

for prevalent diseases and informing the development of public health policies. Data collection encompasses a wide range of participant information including demographics, dietary recalls, health examinations, toxin exposures, and laboratory measurements, all obtained through structured interviews and physical examinations conducted either at participants' homes or mobile testing centers.

Datasets were extracted from the NHANES website[15], covering the cycles from 1999 to 2018. Specifically, the focus was on testing the exposomic variables only for this study. These datasets were integrated into a comprehensive table, where each row corresponds to a participant and each column represents a specific NHANES variable. This cumulative dataset consists of 101,316 participants and 11,274 variables spanning multiple domains, including demographic, dietary, health, examination, laboratory, questionnaire, socioeconomic, and occupational categories including all phenotype, exposure, and covariate information sourced from the NHANES database. From this comprehensive data, we were able to select the specified lipid phenotypes and exposures relevant to our study. It is noteworthy that NHANES fields are not consistently maintained across cycles; fields may be modified or discontinued over time, posing challenges for longitudinal analyses[16].

2.2 GE.db

The GE.db module is an integral component of the IGEM system[12], designed as a comprehensive knowledge base of genomic and exposomic interactions. This module aggregates data from various public databases, providing a curated repository of interactions that can be leveraged to filter high-volume research datasets effectively. The primary purpose of GE.db is to utilize prior knowledge of gene-exposure and exposure-exposure interactions to filter datasets, thereby retaining only variables with known biological relationships. The aim of strategic filtering is to significantly reduce the number of variables requiring analysis so as to conserve computational resources, reduce processing time, and minimize the occurrence of type I errors after multiple testing corrections.

2.2.1 Data Sources

GE.db derives its data from multiple reputable public databases that are frequently updated and maintained. As a foundational step in developing the exposure terms, IGEM incorporates an integration system of environmental and genetic data as it uses a rigorous process of standardizing and mapping terms. To facilitate this task, we use MeSH (Medical Subject Headings)[17] from the National Center for Biotechnology Information (NCBI), a widely recognized database of biomedical descriptors, as part of the word pre-processing procedure. The main function of MeSH in IGEM is to serve as a reference dictionary to standardize and consolidate different forms of terms that appear in various data sources. For instance, in the context of chemical exposures, the same chemical compound might be referred to in different ways, either by its chemical formula (e.g., "C6H12O6" for glucose), its full name (e.g., "glucose"), or a numeric code or identifier. The word pre-processing procedure in IGEM uses MeSH to identify all these variations and then assigns a unique and consolidated identifier to each term. This unified identifier ensures that IGEM recognizes all these

forms as the same concept, providing consistency across the data and facilitating the integration of external sources. Moreover, this mapping allows IGEM to link data from multiple external databases, ensuring that the same terms can be identified in different contexts, such as environmental exposures or clinical records, regardless of how they were originally represented. The final product is a standardized and unified knowledge base that simplifies the analysis of interactions of the environmental terms, improving both the efficiency and accuracy of scientific discoveries.

For this analysis, the following databases were considered as they provide relevant environmental information: Human Metabolome Database (HMDB)[18], a detailed resource containing information on small molecule metabolites found in the human body, crucial for understanding metabolic interactions and pathways; Comparative Toxicogenomics Database (CTD)[19], which integrates information on chemical-gene/protein interactions, chemical-disease, and gene-disease relationships, facilitating insights into the molecular mechanisms of environmental diseases; and Kyoto Encyclopedia of Genes and Genomes (KEGG)[20], which provides comprehensive data on gene functions, biological pathways, diseases, drugs, and chemical substances, supporting the integration of genomic and metabolic information. This methodology allowed the identification and recording of interactions where multiple exposure factors were found in the same record.

At the time of analysis, the GE.db contained 1,057,827 terms grouped into categories such as anatomy, chemicals, diseases, chromosomes, genes, metabolites, pathways, and SNPs, along with 15,667,807 interactions among these terms. The GE.db module is designed with a flexible architecture that allows for the seamless integration of new data sources. It includes several key components: Term Table, which contains key terms and concepts essential for the analysis, organized into groups and categories for efficient retrieval; Interaction Table, which stores documented interactions between various genomic and exposomic variables, providing a robust foundation for filtering datasets; and Mapping Algorithms, which utilize advanced algorithms to match external data terms to internal GE.db terms, ensuring consistency and reliability in the filtering process. To maintain the GE.db, the IGEM system employs version control routines and layers of data ingestion and data transformation to fetch data from their sources and transform them into term links (Figure 1). The GE.filter is another component of IGEM that enables various operations on the GE.db knowledge base, including term matching, interaction identification, and data reduction. The IGEM system, along with its modules GE.db and GE.filter, is deployed in a Python environment on an institutional linux computing cluster. The database utilized is SQLite, which currently has a size of 2.7 GB. For a more detailed explanation on the workflow and filtration parameters used within each command involving Ge.db and GE.filter, please refer to our user guide located on Github[21].

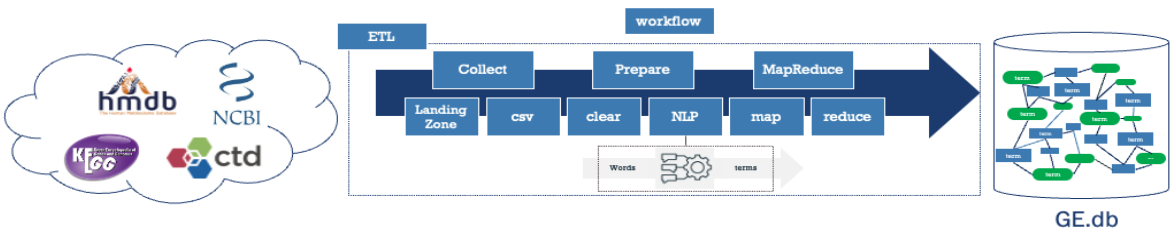


Figure 1. Visualization of GE.db workflow from database to interaction term identification.

2.3Phenotypes and Confounder Variables

Within the NHANES dataset, specific variables were identified as phenotypes and confounders for this analysis. The selected phenotypes included are listed in Table 1. For HDL-C, NHANES altered the calculation method for this indicator over different cycles. NHANES encountered method-related bias for calculating the HDL-C values for 1999-2000, 2001-2002, and 2005-2006; the bias for 2003-2004 was acceptable (<4%) and required no correction[22]. The adjustments implemented improved consistency across various years and methodologies, ensuring that the differences observed in HDL-C levels more accurately reflected true variations rather than being impacted by measurement bias. Consequently, these three fields were maintained separately, creating three distinct datasets. The selected confounders included Gender (RIAGENDR), Age (RIDAGEYR), BMI (BMXBMI), Race/Ethnicity (RIDRETH1), and Survey Cycle (SDDSRVYR).

2.4 Adjusting for Cholesterol Medications

To account for the influence of cholesterol-lowering medications on lipid measurements, we adjusted the LDL-C and Total Cholesterol (TC) values for participants who reported using statins (Figures S-1,2). This adjustment is crucial for accurately assessing lipid levels and their associations with various exposures, as statins significantly alter cholesterol levels. We utilized the NHANES dataset RXQ_RX to identify participants who reported using at least one of the following statin components: ATORVASTATIN CALCIUM, SIMVASTATIN, PRAVASTATIN SODIUM, and FLUVASTATIN SODIUM.

For these participants, we adjusted the LDL and TC values as follows: LDL-cholesterol (LBDLDL) values were divided by 0.7 to account for the reduction effect of statins, and Total Cholesterol (LBXTC) values were divided by 0.8 to adjust for statin usage[23]. By incorporating these adjustments, we enhanced the precision of our lipid measurements, ensuring that our analysis of exposure-lipid interactions was both accurate and reliable.

2.5 NHANES Exposure Filtering for the Interaction Models

To align the NHANES variables with GE.db, all NHANES variable descriptions (excluding lipid phenotypes and confounders) were processed through the GE.filter function. GE.filter utilizes an internal NLP (Natural Language Processing) engine to identify corresponding GE.db terms based on textual descriptions. This process identified 3,619 NHANES variables related to 534 GE.db terms.

A subsequent review of these related NHANES variables identified 1,136 exposure factors, corresponding to 217 unique terms. These 217 terms were then used as filter parameters for another GE.filter function run, which searched the GE.db knowledge base for all interactions among these terms, resulting in the identification of 382,613 putative Exposure x Exposure interactions. We performed this step prior to quality controlling the exposure variables, ensuring that only exposures present in the NHANES data were included for curation of the interactions to be tested.

2.6 Quality Control (QC)

Quality Control is a critical step to ensure the integrity, reliability, and validity of the dataset used in the analysis. The IGEM system includes specialized functions that accelerate and assist in the

application of QC procedures to -omics data analyses. The following procedures were applied to the NHANES dataset after filtering and modifications from previous steps.

For continuous data type QC, all variables with more than 90% missing values were removed. The distribution of phenotypes was calculated using the skewness (3(mean-median)/standard deviation.) and all phenotypes were log-transformed to normalize the distribution (Figure S-3).

Participants were then separated into discovery and replication groups for the six cohorts of phenotypes, resulting in twelve datasets. For each dataset, a minimum of 200 participants for categorical and binary exposures was maintained. Only variables present in both discovery and replication datasets for each phenotype were retained to ensure consistency and reliability (Table 1).

Table 1. Overview of lipid phenotypes sorted by survey cycle, including sample sizes, exposures, and interactions that passed quality control.

Phenotype	NHANES Cycles	NHANES ID	NHANES Description	N Discovery	N Replication	Exposures	Interactions
HDL-C	1999 – 2002	LBDHDL	HDL-cholesterol, mg/dL	4,572	4,949	96	2,073
HDL-C	2003 – 2004	LBXHDD	Direct HDL-Cholesterol, mg/dL	3,425	1,469	219	11,093
HDL-C	2005 – 2018	LBDHDD	HDL-Cholesterol, mg/dL	21,442	16,000	231	11,721
LDL-C	1999 – 2018	LBDLDL	LDL-cholesterol, mg/dL	11,453	12,695	181	6,934
Total Cholesterol	1999 – 2018	LBXTC	Total Cholesterol, mg/dL	24,836	27,023	193	7,873
Triglycerides	1999 – 2018	LBXSTR	Triglycerides, mg/dL	19,305	26,916	177	6,446

2.7 Statistical Analysis Models (Discovery and Replication)

The IGEM system, inheriting functionalities from the CLARITE system[24], performs interaction analyses by calculating the p-value of the Likelihood Ratio Test (LRT) between two models. In the full and reduced model $Y_{phenotype}$ is the outcome variable, β_0 is the intercept, $\beta_1 term1$ and $\beta_2 term2$ are the coefficients for the individual predictors, and $\beta_{n+1} cov_n$ are the coefficients for the covariates with n adding on to the number of covariates used in the model. Exclusive to the full model, $\beta_3(term1 \times term2)$ is the interaction term between term 1 and term 2.

Full Model:

$$Y_{phenotype} = \beta_0 + \beta_1 term1 + \beta_2 term2 + \beta_3(term1 \times term2) + \beta_4 cov_1 + \ldots + \beta_{n+1} cov_n \quad (1)$$

Reduced Model:

$$Y_{phenotype} = \beta_0 + \beta_1 term1 + \beta_2 term2 + \beta_3 cov_1 + \ldots + \beta_{n+1} cov_n \quad (2)$$

The LRT is utilized to compare the fit of the two models, with the full model including the interaction term ($\beta_3(term1 \times term2)$) and the reduced model excluding it. The analysis involves fitting the full model to the data to obtain the log-likelihood (L_{full}) and fitting the reduced model to obtain the log-likelihood ($L_{restricted}$). The LRT statistic represented as D with -2 used as a scaling factor that makes the likelihood ratio test statistic approximately follow a chi-squared distribution under the null hypothesis is calculated as:

$$D = -2(L_{restricted} - L_{full}) \quad (3)$$

The difference in degrees of freedom between the two models is 1, since the full model has one additional parameter ($\beta_3(term1 \times term2)$). The p-value is derived from the probability (P) that a random variable following a chi-squared distribution ($\chi 2$) with 1 degree of freedom takes a value greater than or equal to the observed test statistic (D):

$$p\text{-}value = P(\chi 2 \geq D \mid df = 1) \quad (4)$$

The LRT p-values were calculated for each interaction identified in the discovery dataset for each phenotype.

However, in some cases, the p-value of the LRT cannot be calculated. The following messages inform the user of the reasons:

- Too few complete observations (min_n filter: N < 200)
- The number of complete observations is insufficient to perform the analysis, as the minimum required is 200
- Both models are equivalent in terms of fit: the two models are equivalent in terms of fit, with no significant difference between them
- No Overlap (min_n filter: 0 < 200): there is insufficient data overlap to perform the analysis, as the minimum required is 200.

Following the interaction model analysis, the IGEM function was applied to adjust the p-values for multiple testing using both Bonferroni correction and False Discovery Rate (FDR) adjustment. From the discovery analysis, interactions with an FDR-adjusted p-value < 0.1 were filtered. These significant interactions that met the FDR adjustment threshold were then tested in the replication dataset. The same interaction analysis was conducted in the replication cohort, applying identical model specifications and LRT. The replication criteria also required that interactions exhibit consistent directional effects between the discovery and replication interaction betas, with all significant interactions retaining a Bonferroni-adjusted p-value < 0.05, across both datasets. This rigorous approach ensures that the identified interactions are robust and not due to random chance.

3. Results

In this study, we examined the interactions between various exposure variables and lipid phenotypes using the NHANES dataset. We performed a comprehensive analysis to identify significant exposure-exposure interactions (ExE) that are associated with lipid levels. Below are the key findings from our discovery and replication datasets. Of all the 26,107 interactions tested that included exposures that passed QC, a total of 263 interactions were statistically significant in the

discovery dataset with an FDR p < 0.1 (Table 2). A total of 61 interactions were found to be significant in both discovery and replication when allowing for an FDR p < 0.1 (assorted by lipid phenotype) and 21 interactions associated with the HDL-cholesterol trait was significant with a Bonferroni corrected p < 0.05 (Figure 2). Additionally, these interactions demonstrated consistent directions of effect across both discovery and replication datasets (Table S-1).

Table 2. Frequency table of all the interactions tested for every lipid phenotype.

Phenotype	Discovery Interactions	Replication Interactions	FDR p < 0.1 in both	Bonferroni p < 0.05 in both
HDL-C [1999-2002]	1,116	4	1	1
HDL-C [2003-2004]	5,459	93	2	0
HDL-C [2005-2018]	6,584	141	58	20
LDL-C	4,339	9	0	0
Total Cholesterol	4,764	10	0	0
Triglycerides	3,845	6	0	0
Total	26,107	263	61	21

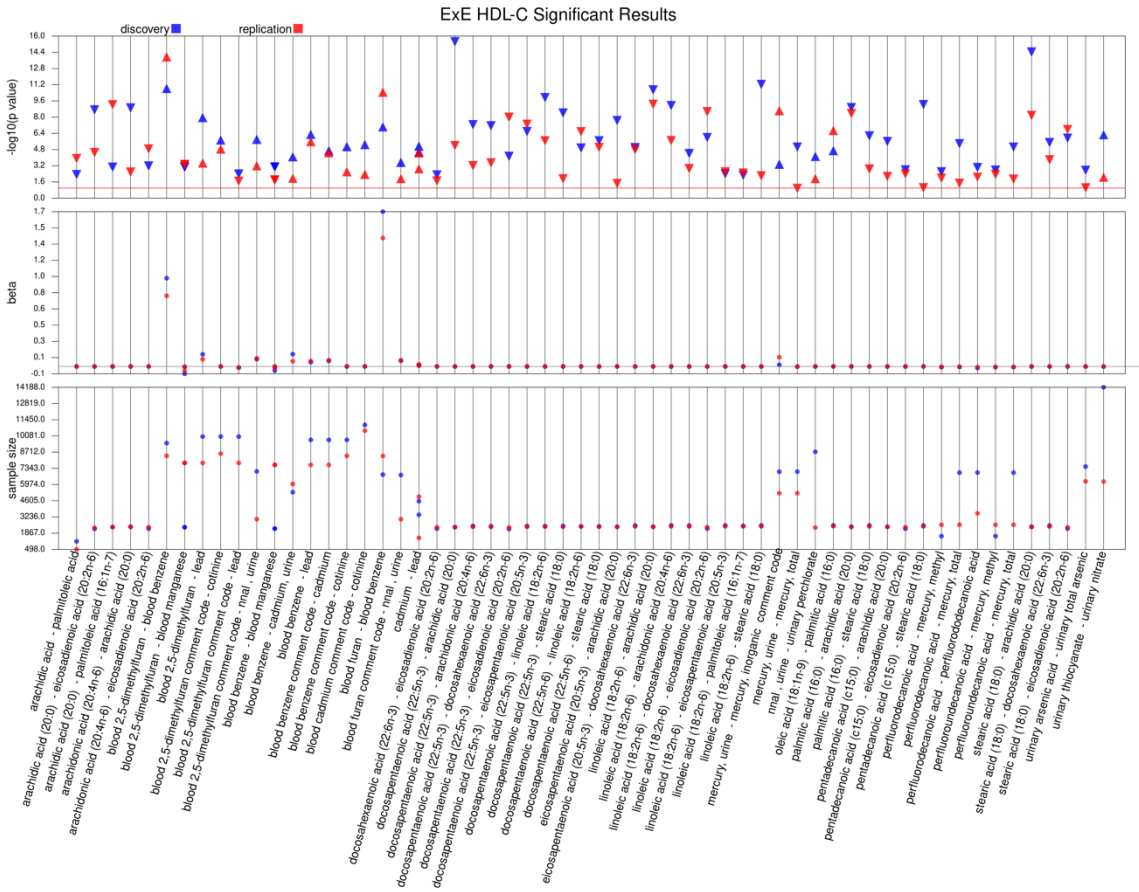
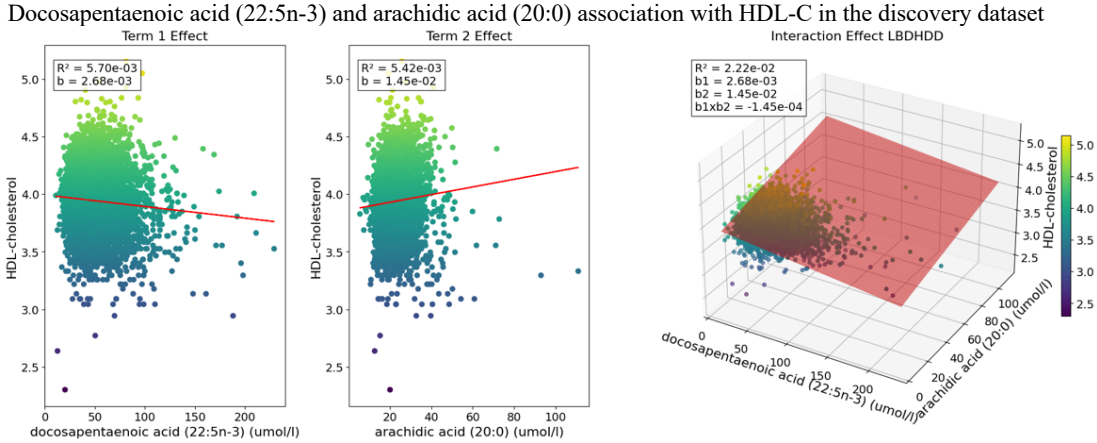


Figure 2. The sixty-one significant results starting from the top showcasing all the interactions with FDR LRT p-value < 0.1(denoted by the redline) and the twenty-one significant results with a Bonferroni adjusted LRT p-value < 0.05 (direction of effect pointing down is negative and up is positive), the interaction beta for both exposures, and the sample sizes. PheWAS-View was the software used to generate this plot[25].

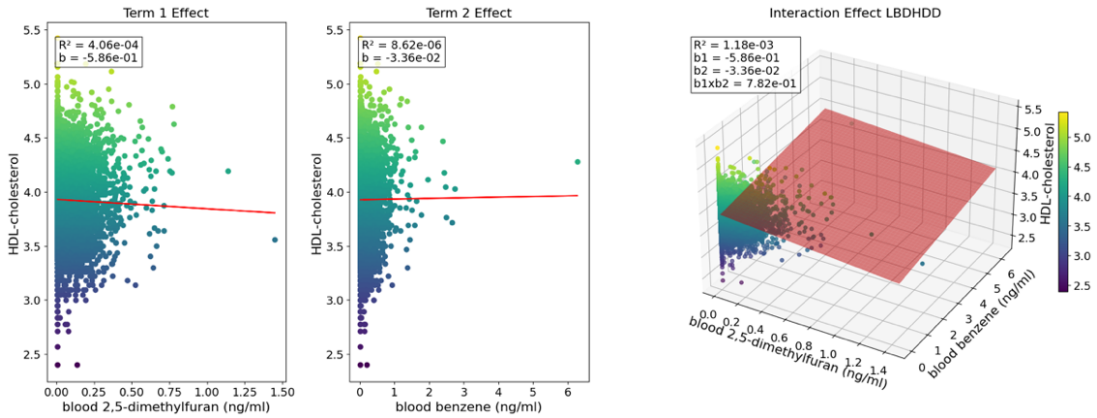
3.1 Significant Interactions

The top three results with the lowest LRT p-values associated with HDL-cholesterol include: 1) Docosapentaenoic acid (22:5n-3) (DPA) - arachidic acid (20:0) (Discovery: Bonferroni adjusted LRT p-value = 8.43×10^{-13}, $\beta = -1.4\times10^{-4}$; Replication: Bonferroni adjusted LRT p-value = 3.25×10^{-4}, $\beta = -1.2\times10^{-4}$) (Figure 3A). 2) Blood 2,5-dimethyfuran - blood benzene (Discovery: Bonferroni adjusted LRT p-value = 2.75×10^{-7}, $\beta = 0.97$; Replication: Bonferroni adjusted LRT p-value = 4.48×10^{-12}, $\beta = 0.78$) (Figure 3B). 3) Stearic acid (18:0) - arachidic acid (20:0) (Discovery: Bonferroni adjusted LRT p-value = 8.88×10^{-12}, $\beta = -7.79\times10^{-6}$; Replication: Bonferroni adjusted LRT p-value = 3.47×10^{-7}, $\beta = -1.26\times10^{-5}$) (Figure 3C).

(A) Docosapentaenoic acid (22:5n-3) and arachidic acid (20:0) association with HDL-C in the discovery dataset



(B) Blood 2,5 dimethylfuran (ng/ml) and blood benzene (ng/ml) association with HDL-C in the replicate dataset



(C) Stearic acid (18:0) (umol/l) and arachidic acid (20:0) (umol/l) association with HDL-C in the discovery dataset

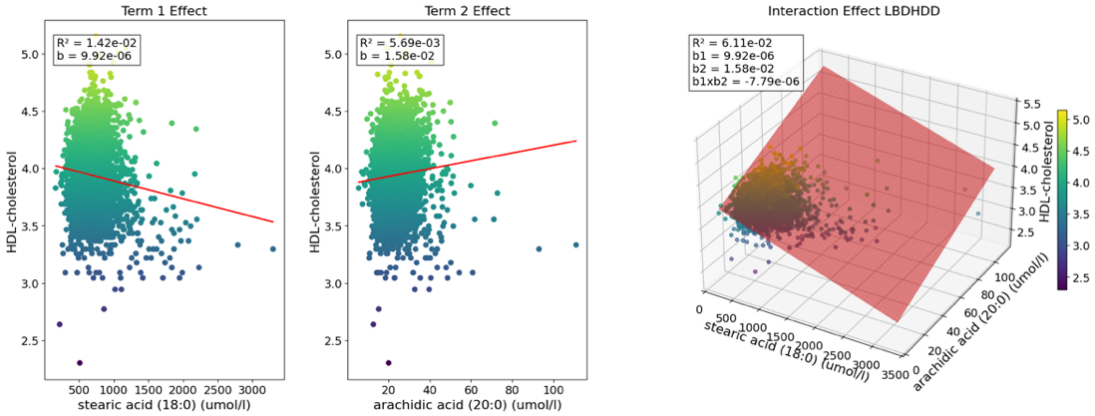


Figure 3A-C. The top three results plots observing the individual main effect correlation line, and the 3D plot showing the interaction correlation along the square plane.

4. Discussion

In this study, we leveraged the comprehensive exposomic knowledge base provided by the GE.db module of IGEM to investigate exposure-exposure interactions (ExE) associated with lipid phenotypes. By utilizing data from the NHANES dataset spanning 1999 to 2018, we identified several significant interactions between various exposures and lipid levels. The replication of these findings across independent datasets underscores the robustness of our approach and highlights the potential of GE.db in facilitating large-scale -omics research.

4.1 Clinical and Public Health Implications

Our analysis revealed several key interactions, notably DPA and stearic acid with arachidic acid associated with HDL-C. DPA is a known essential omega-3 fatty acid, and stearic and arachidic acid are saturated fatty acids[26-28]. These results suggest that specific combinations of environmental exposures may have synergistic effects on lipid metabolism, though most research only touches on their individual effects on lipid profiles. For instance, omega-3 fatty acids, such as DPA, are generally linked with increased HDL cholesterol levels[29], while high consumption of saturated fatty acids like arachidic acid may unfavorably affect lipid profiles, potentially leading to elevated LDL-C levels[27]. Our findings indicate a negative impact on HDL-C when arachidic acid interacts with fatty acids typically associated with positive HDL-C effects, suggesting that arachidic acid could potentially diminish the benefits of HDL-C promoting fatty acids. Other research suggests that stearic acid may have a neutral or even beneficial effect on cholesterol levels, possibly not adversely affecting HDL-C on its own[30]. However, as seen in our results, when combined with arachidic acid, this interaction could overall have a negative impact, counteracting any neutral or positive effects on HDL-C.

Additionally, the interaction between blood 2,5-dimethylfuran and blood benzene highlights the potential combined impact of exposure to volatile organic compounds (VOCs) on HDL-C levels. Benzene has been observed to increase LDL-C levels which would naturally displace or plateau

HDL-C levels procuring a negative effect[31,33]. Measures of 2,5-dimethyfuran though, have limited research indicating influence on lipids, but may pose health risks similar to other VOCs. These risks can include respiratory irritation, and potential systemic effects that could indirectly affect lipid metabolism and cardiovascular health[34–36]. Conversely, our results demonstrate a positive interaction effect on HDL-C with benzene and 2,5-dimethylfuran. Therefore, further study of this interaction is warranted, especially considering the known detrimental impact of VOCs on public health. In summary, all these findings have important implications for public health, as they point to the need for considering multiple concurrent exposures in dietary and environmental risk assessments. Public health strategies could be developed to mitigate the combined effects of specific dietary and environmental exposures on lipid metabolism.

4.1.1 Significant Interaction Effects Sizes

As stated previously for HDL-C, the bias adjustment was accounted for whether the survey cycle year had been corrected or not as they were all approved to use for statistical analysis. The HDL-C variable still had to be labeled and categorized differently to identify which ones were corrected vs. not corrected for transparency. Given that the LBDHDD variable spanned the largest survey cycle from 2005-2018 of the three, showed an increased sample size disparity by about 17,000 participants when comparing the other two survey cycles which had around 4,000 participants each. Thus, presuming that even if the effect size remains similar across those survey cycles, a larger sample size in one cycle can lead to a significant p-value, while a smaller sample size in another cycle could result in a non-significant p-value for the same effect size.

Regarding the effect sizes of the three significant interactions mentioned, we believe the positive beta for DPA can coexist with a slight negative trend due to the small effect size and interaction with arachidic acid (Figure 3A). The combined effect of DPA and arachidic acid as described by the interaction term, may influence the overall outcome more than the individual effect of DPA alone. In the dataset, the interaction between the two terms might reduce or counterbalance DPA's small positive effect on HDL-C. Figure 3B depicts another story where the two weaker effect sizes of blood 2,5-dimethylfuran and blood benzene alone hold less weight than compared to the larger effect size of when both blood 2,5-dimethylfuran and blood benzene increase together. Their combined effect led to an overall increase in HDL-C despite their individual negative contributions. Lastly, the negative interaction effect size for stearic acid and arachidic acid is very small in association with HDL-C, and largely driven by the positive influence of arachidic acid (Figure 3C). The overall interaction appears to slightly counteract the combined positive effects of both terms but not enough to reverse the trend significantly. Thus, a large amount of the variation is most likely not fully explained in this model and further testing is required.

4.2 Methodological Strengths, Limitations, and Future Directions

A major strength of this study is the use of the GE.db knowledge base, which allowed us to filter high-volume research datasets effectively, focusing only on variables with known biological relationships. This approach significantly reduced the computational burden and enhanced the reliability of our findings by minimizing type I errors through multiple testing corrections. By

employing multiple IGEM modules, we streamlined quality control (QC) processes, which involved variable categorization, data cleaning, and adjustment for confounders like statin use. This approach improved the integrity and accuracy of our analysis, making it user-friendly for whomever uses this tool, and ensuring alignment to bioinformatics practices. The split of data into discovery and replication datasets based on NHANES cycles further increased the validity of our results, as significant interactions identified in the discovery phase were consistently replicated. Another consideration to note is the main-effect interaction model when incorporated without the use of knowledge-driven filters, is typically performed to determine the isolated impact of each variable (in this case, each exposure factor) on phenotypes. However, the goal of this study was not to identify individual main effects but to examine how the combination of multiple exposures influences lipid phenotypes. While the standard main-effect model is valuable in other contexts like simple-trait analysis or in situations where the effects of multiple variables are purely additive, our primary focus was to highlight IGEM's strengths, particularly in capturing interactions based on the pre-existing knowledge within GE.db. GE.db was specifically designed to filter highly relevant variables based on known relationships between exposures. Using this filtering approach allows the analysis to focus on variables with biological context, avoiding the processing of many irrelevant exposures or statistical noise that could arise when including non-interactive main effects.

Despite the robustness of our findings, several limitations warrant consideration. First, since GE.db relies on public databases such as HMDB, CTD, and KEGG, the quality, completeness, and update frequency of these external databases can directly affect the accuracy and relevance of the information in GE.db. Any gaps, errors, or outdated information in these sources could introduce bias or limitations in the results. Regular updates are imperative to ensure the data remains current, but the complexity of fetching and processing new data might slow down the user's analysis pipeline. Furthermore, the interactions stored in GE.db are curated from specific sources, and their generalizability to other populations, environmental contexts, or less-studied interactions may be limited. Results may not always be applicable outside the scope of the databases from which they were derived.

In the context of the NHANES dataset, the observational nature of the data limits the ability to infer causal relationships between exposures and lipid levels. Interaction effects, as we have noted, may have opposite signs of effect when compared to the main effect betas, which complicates the interpretability of the results. Other datasets with repeated measures of QC and analysis as we have specified with the NHANES data, can help with cross checking all the betas, refining the elucidation of significant interactions. Inclusion of more datasets that host the same kinds of environmental exposures such as the UK Biobank[37] and All of Us Research Program[38], will also help address the possibility of false negatives as some interactions may not have been flagged as significant given our designated thresholds used for the NHANES dataset. Future studies could also incorporate longitudinal data and more sophisticated causal inference methods to address this limitation.

Moreover, while our analysis accounted for several covariates, there may be other unmeasured factors that could influence the observed interactions. Further research should aim to include a broader range of potential confounders and explore the underlying biological mechanisms driving these interactions. Another limitation is the reliance on self-reported data for certain exposures,

which may introduce reporting biases. The integration of more objective measures of exposure, such as well-established biomarkers, could enhance the reliability of future analyses.

4.3 Conclusion

In conclusion, this study demonstrates the utility of the GE.db module in identifying significant ExE influencing lipid traits. The consistent replication of key interactions across independent variables highlights the robustness of our approach and its potential to uncover future novel insights into the complex interplay between environmental exposures and lipid metabolism. These findings pave the way for future research aimed at understanding and mitigating the multifactorial nature of dyslipidemias, ultimately contributing to improved public health outcomes.

This project was supported by the the National Institute of Child Health and Human Development under award number U2C OD023375-06 and the National Heart Lung, and Blood Institute under awards HL169458 and HL168841. This work was additionally supported by the USDA National Institute of Food and Agriculture and Hatch Appropriations under Project #PEN04275 and Accession #1018544.

Code for GE.db, GE.db filter, and quality control steps used in this study are made available here: https://github.com/HallLab/pbs_igem/tree/main. The IGEM package and user guide are available here: https://github.com/HallLab/IGEM.
Supplemental table and figures S-1, S-2, and S-3 are available at https://ritchielab.org/publications/supplementary-data/psb-2025/igem.

References

1. Virolainen, S. J., VonHandorf, A., Viel, K. C. M. F., Weirauch, M. T. & Kottyan, L. C. Gene-environment interactions and their impact on human health. *Genes Immun.* **24**, 1–11 (2023).
2. Ottman, R. Gene-environment interaction: definitions and study designs. *Prev. Med.* **25**, 764–770 (1996).
3. Castelli, W. P. Lipids, risk factors and ischaemic heart disease. *Atherosclerosis* **124 Suppl**, S1–9 (1996).
4. Emerging Risk Factors Collaboration *et al.* Major lipids, apolipoproteins, and risk of vascular disease. *JAMA* **302**, 1993–2000 (2009).
5. Dayimu, A. *et al.* Trajectories of Lipids Profile and Incident Cardiovascular Disease Risk: A Longitudinal Cohort Study. *J. Am. Heart Assoc.* **8**, e013479 (2019).
6. Amin, K. A., Homeida, A. M., El Mazoudy, R. H., Hashim, K. S. & Garelnabi, M. Dietary Lipids in Health and Disease. *J. Lipids* **2019**, 5729498 (2019).

7. Hornburg, D. *et al.* Dynamic lipidome alterations associated with human health, disease and ageing. *Nat Metab* **5**, 1578–1594 (2023).
8. Ritchie, M. D. *et al.* Incorporation of Biological Knowledge Into the Study of Gene-Environment Interactions. *Am. J. Epidemiol.* **186**, 771–777 (2017).
9. Pendergrass, S. A. *et al.* Genomic analyses with biofilter 2.0: knowledge driven filtering, annotation, and model development. *BioData Min.* **6**, 25 (2013).
10. Kim, D. *et al.* Biofilter as a functional annotation pipeline for common and rare copy number burden. *Pac. Symp. Biocomput.* **21**, 357–368 (2016).
11. Hall, M. A. *et al.* Biology-driven gene-gene interaction analysis of age-related cataract in the eMERGE Network: Biology-driven tool to identify genetic interactions. *Genet. Epidemiol.* **39**, 376–384 (2015).
12. Term — IGEM 0.1.0 documentation. https://igem.readthedocs.io/en/latest/ge/md/term.html.
13. National health and nutrition examination survey. https://www.cdc.gov/nchs/nhanes/index.htm (2024).
14. About the national health and nutrition examination survey. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm (2024).
15. CDC. Centers for disease control and prevention. https://www.cdc.gov/ (2024).
16. Nguyen, V. K. *et al.* Harmonized US National Health and Nutrition Examination Survey 1988-2018 for high throughput exposome-health discovery. *medRxiv* (2023) doi:10.1101/2023.02.06.23284573.
17. Sievert, M., Patrick, T. & Reid, J. Need a bloody nose be a nosebleed? or, lexical variants cause surprising results. *Bull. Med. Libr. Assoc.* **89**, 68–71 (2001).
18. Wishart, D. S. *et al.* HMDB: the Human Metabolome Database. *Nucleic Acids Res.* **35**, D521–6 (2007).
19. Davis, A. P. *et al.* Comparative Toxicogenomics Database (CTD): update 2023. *Nucleic Acids Res.* **51**, D1257–D1262 (2023).
20. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
21. *User Guider: IGEM v.0.1.4*. (Github, 2023).

22. NHANES 2005-2006: Cholesterol - HDL Data Documentation, Codebook, and Frequencies. https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/HDL_D.htm#LBDHDD.
23. Liu, D. J. *et al.* Exome-wide association study of plasma lipids in >300,000 individuals. *Nat. Genet.* **49**, 1758–1766 (2017).
24. Lucas, A. M. *et al.* CLARITE Facilitates the Quality Control and Analysis Process for EWAS of Metabolic-Related Traits. *Front. Genet.* **10**, 1240 (2019).
25. Wolfe, D., Dudek, S., Ritchie, M. D. & Pendergrass, S. A. Visualizing genomic information across chromosomes with PhenoGram. *BioData Min.* **6**, 18 (2013).
26. Human Metabolome Database: Showing metabocard for Stearic acid (HMDB0000827). https://hmdb.ca/metabolites/HMDB0000827.
27. Human Metabolome Database: Showing metabocard for Docosapentaenoic acid (22n-3) (HMDB0006528). https://hmdb.ca/metabolites/HMDB0006528.
28. Human Metabolome Database: Showing metabocard for Arachidic acid (HMDB0002212). https://hmdb.ca/metabolites/HMDB0002212.
29. Peña-de-la-Sancha, P. *et al.* Eicosapentaenoic and Docosahexaenoic Acid Supplementation Increases HDL Content in n-3 Fatty Acids and Improves Endothelial Function in Hypertriglyceridemic Patients. *Int. J. Mol. Sci.* **24**, (2023).
30. Siri-Tarino, P. W., Sun, Q., Hu, F. B. & Krauss, R. M. Saturated fatty acids and risk of coronary heart disease: modulation by replacement nutrients. *Curr. Atheroscler. Rep.* **12**, 384–390 (2010).
31. Grundy, S. M. Influence of stearic acid on cholesterol metabolism relative to other long-chain fatty acids. *Am. J. Clin. Nutr.* **60**, 986S–990S (1994).
32. Tualeka, N. A. R., Martiana, N. T., Wibrata, A. & Rahmawati, P. Effect of food consumption contain glutahione anti-oxidant towards LDL cholesterol concentrations on benzene-exposed-workers at the romokalisari shoe industry, Surabaya. *Indian J. Forensic Med. Toxicol.* (2019) doi:10.5958/0973-9130.2019.00333.5.

33. Ye, L. *et al.* Moderate body lipid accumulation in mice attenuated benzene-induced hematotoxicity via acceleration of benzene metabolism and clearance. *Environ. Int.* **178**, 108113 (2023).
34. Fu, X. *et al.* Airborne 2,5-dimethylfuran as a marker to indicate exposure to indoor tobacco and biomass burning smoke. *Atmos. Environ.* **259**, 118509 (2021).
35. Jing, L., Chen, T., Yang, Z. & Dong, W. Association of the blood levels of specific volatile organic compounds with nonfatal cardio-cerebrovascular events in US adults. *BMC Public Health* **24**, 616 (2024).
36. Chen, X. *et al.* Association of Smoking with Metabolic Volatile Organic Compounds in Exhaled Breath. *Int. J. Mol. Sci.* **18**, (2017).
37. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
38. All of Us Research Program Investigators *et al.* The 'All of Us' Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).