**Relationship of Omega-6 PUFAs and DNA Methylation in the** **Multi-Ethnic Study of Atherosclerosis**

**Presenter: Xizhi Xu**

**Plan B Advisor: Weihua Guan**

**Date: 06/11/2024**

**Abstract**

Inflammation is a key response to infection and injury and is part of the normal, innate immune response. Biologically, fatty acids are involved in the inflammatory process on multiple levels. In our study, we would like to understand the epigenetic background of fatty acids.

We carried out epigenome-wide association studies (EWASs) using data collected in the Multi-Ethnic Study of Atherosclerosis (MESA) study at 4 clinical centers with N = 1264 participants. Epigenetic data were measured by Illumina HumanMethylation450 BeadChip including probes for ~485K CpG sites. Our study was a cross-sectional analysis and we were interested in five Omega-6 polyunsaturated fatty acids (PUFAs), including linoleic acid (LA), eicosadienoic acid (EDA), gamma-linolenic acid (GLA), dihommo-gamma-linolenic acid (DGLA), and arachidonic acid (AA). We applied linear regressions to examine associations between blood DNA methylation (DNAm) and each fatty acid, adjusting for Gender, Site, Age, Race, Cigarette, BMI, Glucose. To identify and remove unobserved confounders, Surrogate Variable Analysis (SVA) was applied to DNA methylation data and significant SVs were also included as covariates. The Bonferroni correction (p = 0.05 / 484817 = 1.03e-07) was used to determine the significant threshold. Additionally, we performed differentially methylated region (DMR) analysis to group CpG sites into regions and compare them. The motivation was to understand the methylation change in a broader region. As results of association results, the top CpG signal, cg19610905, appeared epigenome-wide significant in its associations with LA, AA, EDA and DGLA, and this CpG site is located in the Fatty Acid Desaturase 2 (FADS2) gene region. Functionally, the FADS2 gene encodes an enzyme involved in the desaturation process of fatty acids, regulating unsaturation of fatty acids through the introduction of double bonds between defined carbons of the fatty acyl chain. Additional significant CpG sites were identified, for example, in the PFKP, PLXNB3 gene regions. The findings of DMR were different than linear associations, which discovered MEST gene for fatty acids LA, AA, EDA and DGLA. The genes related to GLA were PCDHGA and RP3-400N23.6. Both linear association and DMR results were reasonable, depending on whether viewing it from an individual CpG perspective or a regional perspective.

Our study has the potential to revolutionize dietary recommendations, enhance disease prevention strategies, and lead to novel therapeutic approaches that leverage the power of epigenetic regulation.

**Introduction**

**Background and existing literature:** My topic is about the relationship of phospholipid fatty acids, DNA methylation, and gene expression of inflammatory loci in the Multi-Ethnic Study of Atherosclerosis (MESA). Inflammation is a key response to infection and injury and is part of the normal, innate immune response. Uncontrolled or inappropriate inflammation causes excessive damage to tissues and contributes to a wide range of conditions. Fatty acids are involved in the inflammatory process on multiple levels. Previous studies have proved systematic inflammation was significantly reduced by omega-3 polyunsaturated fatty acids (PUFAs) through the reduction of interleukin-6 (IL-6) and TNF-α concentrations. Conversely, increased intake of omega-6 PUFAs, however, has no influence on TNF-α, IL-6 production. Therefore, we suspect that fatty acids may influence the inflammatory process up to individual or type of fatty acid.

**My topic and its significance:** The existing evidence has shown that fatty acids impact cytokine signaling through global DNA methylation because of roles of individuals in one-carbon metabolism. (DNA methylation is an epigenetic mechanism where a methyl group is added to the cytosine or adenine DNA nucleotides. It plays a critical role in regulating gene expression, genomic stability, and development. In mammals, DNA methylation typically occurs at CpG sites, where a cytosine nucleotide is followed by a guanine nucleotide.) These epigenetic changes possibly make a difference to the effects of PUFAs observed on gene expression and inflammation.

The overall goal of my research is to examine the relationship between phospholipid omega-6 polyunsaturated fatty acids (PUFAs) and DNA methylation patterns of the TNF encoding gene at multiple cytosine-guanine dinucleotide (CpG) sites. Our expected outcome is to identify which PUFAs are significantly associated with DNA methylation. More generally, from public health perspective, the significance of the study is to potentially revolutionize dietary recommendations, enhance disease prevention strategies, and lead to novel therapeutic approaches that leverage the power of epigenetic regulation.

**Research Hypothesis**

Our main research question is whether Omega-6 PUFAs influence methylation patterns of individual several CpG sites of the TNF gene in differing ways depending on individual fatty acid. The secondary research question is whether Omega-6 PUFAs influence methylation patterns of other genes in similar manner to those in aim 1.

**Methods**

**Data collection:** The current research was mostly based on the analysis of purified monocyte samples of 1264 randomly selected participants from 4 MESA clinical centers (Baltimore, MD; Forsyth County, NC; New York, NY; and St. Paul, MN). The study protocol was approved by the Institutional Review Board at each site. All participants signed informed consent.

**Quality control and processing of microarray data:** Data preprocessing and quality control analyses were conducted in R studio using Bioconductor packages. The Illumina HumanMethylation450 BeadChip included probes for 485K CpG sites. Based on the quality control elimination criteria, where detected methylation levels <90% of MESA samples using a detection P value cutoff of 0.05 or overlap with a repetitive element or region, 484817 CpG sites passed the criteria and remained as the final methylation data.

**Missing data**: There were 38 missing observations regarding different covariates in phenotype data, approximately 3% of total sample size. For simplicity and interpretation, we determined to use the complete case study and drop these observations. So the actual sample size would be 1226 for the real analysis.

**SVA:** (Why do we use?) Large scale gene expression studies aim to measure the variables of interest, treatments or phenotypes. For the primary variables, we usually explicitly model them of the expression study. However, in real study, there are also some “unmeasured factors”. The unmeasured factors are sources of expression variation that cannot explained by included variables. The major sources of expression variation are environmental, demographic, technical and genetic factors. As a simple example, consider a human expression study where disease status on a particular cell type is the primary variable. Suppose except for expression change being associated with disease status, the BMI of individuals can also potentially impact the expression. Therefore, if BMI is not included in the model while estimating differential expression with respect to disease status, there would be extra variation because of the effect of BMI, leading to decreased statistical power and ambiguous associations with disease and primary variables.

Therefore, the SVA is used to identify and estimate the surrogate variables for unknown sources of variation in high-dimensional data, like gene expression data. The algorithm is decomposed into two parts: calculate the residual matrix and apply the principal component analysis (PCA).

Suppose the generalized statistical model:

Calculate the residuals and repeat this step to form a **.** The goal is to remove the effects of measured variables of interest. Then, compute the covariance matrix **M** and derive the eigenvalues and eigenvectors. The eigenvalues represent the amount of variance explained by each principal component. Finally, sort the top k eigenvectors and choose them to be our surrogate variables.

Briefly, the intuition of SVA is pretty similar to PCA. They both play crucial roles in dimension reduction of the high-throughput data. The key difference is that PCA applies to whole data matrix, while SVA uses at residual matrix. Also, PCA aims to find the principal factors that explain the most variation of data, and SVA aims to identify and adjust for hidden sources of variation.

**Model / Association analysis:**

We ran the linear regression to test the association between DNA methylation of Cpg sites and five Omega-6 fatty acids: linoleic acid (LA), eicosadienoic acid (EDA), gamma-linolenic acid (GLA); dihommo-gamma-linolenic acid (DGLA), arachidonic acid (AA), adjusting for the covariates: Gender, Site, Age, Race, Cigarette use, BMI, Glucose. Additionally, the surrogate variables are added to ensure accounting for unknown sources of variation.

**DMR Analysis:**

In addition to finding out the individual CpG site effect, we also performed differentially methylated region (DMR) analysis. DMR is a method used to look for regions if multiple CpG sites show a consistent pattern of differential methylation. The tool we used was DMRcate package (Tim Peters. *DMRcate*. 2024). It aims to capture more biologically meaningful changes across a broader genomic context. Instead of purely based on the statistical significance of individual associations, we zoomed out our analysis to see a larger picture. On the other hand, we were curious to check if this result was consistent as before.

In the model specification, we account for one fatty acid each time at design matrix, including all the covariates. The annotation package for Illumina's 450k methylation arrays was *IlluminaHumanMethylation450kanno.ilmn12.hg19.*

**Results**

The demographics of the study population (n = 1226) are shown in *Table 1: Descriptive Table* below. We have 7 covariates in total. For the categorical variables (eg. Gender, Race), the proportion and percentage of each level is presented. For the continuous variables (eg. BMI, Glucose), the mean and standard deviation are presented.

A screenshot of a computer screen

Description automatically generated

As the result of association analysis, we created 5 Manhattan plots in *qqman* package to visualize the significant CpG sites (*Intro to the qqman package,* 2023). The x-axis is the chromosome number: from chromosome 1 to chromosome 22. The chromosome X and chromosome Y are recoded as Chr 23 and Chr 24, respectively, as the Manhattan function requires all the numeric inputs. And the y-axis represents the p-values for the CpG. The red horizontal line is the genome-wide significance line, the CpG sites above this line are considered as significant. The Bonferroni correction is applied to determine the significance threshold ( = 0.05 / N, N = 484817). In an overview, we discovered there were approximately 3 or 4 significant CpG sites on average for each fatty acid. Here, we would mostly focus on the top signal.

A collage of graphs and charts

Description automatically generated

A graph of a number of molecules

Description automatically generated with medium confidence

*Figure 1: Manhattan plots of five fatty acids (FAs)*

**LA:** In LA, we found the top signal, cg19610905, with p-value = 3.865e-28. The reference gene region is FADS2, standing for Fatty Acid Desaturase 2, which is a protein coding gene. (Detailed gene information and functions will be described in **Discussion.)** Then we could get the conclusion: the gene, LA, significantly impacts the DNA methylation site cg19610905, cg01400685, cg22837486. These CpG are all in the gene region FADS. As for the other two significant CpG that passed the threshold, we marked them and added them to *Table 2: Supplemental table of significant signals*.

**AA:** Similarly in AA, we also found the top signal cg19610905 with p-value = 9.38e-30. In addition to the top signal CpG, cg22837486 and cg01400685 also behaved significance here. There are also several new significant CpG appearing in AA: cg11250194, cg26786045 and cg05237114. Here, we could draw similar conclusions: the fatty acid AA significantly impacts the DNA methylation sites cg19610905, cg22837486, cg01400685, cg26786045, cg11250194, cg05237114.The reference gene of the first two CpG are WFDS2 which we talked about before. The cg26786045 reference gene region is RAB3IL1. RAB3IL1 encodes a guanine nucleotide exchange factor for the ras-related protein Rab3A. The encoded protein binds Rab3a and the inositol hexakisphosphate kinase InsP6K1. Alternative splicing results in multiple transcript variants.

**DGLA/ EDA:** Similarly, in DGLA and EDA, the significant CpG sites all appeared before, so we will not cover every details again.

**GLA:** However, GLA has pretty different results from the others. It is worth noting the significant CpG here are all new and spread to chromosome 6, 23 (which is chromosome X), 17, 3, 10. For this fatty acid, there is additional consideration: if the significant CpG has the SNPs nearby and minor allele frequency > 0.05, those SNPs may confound the results, which means we need to be careful to conclude the association. The reason we didn’t mention earlier in other fatty acids is because previously all the significant CpG signals were all in the same chromosome region. Those CpG sites formed a significant area, like a “cluster”. Thus, even if one or two CpG has the nearby SNPs, the remaining CpG in the cluster can still support our conclusion. Then, following the same process, the top signal here is cg27069691 in chromosome 10, with p-value 1.01e-10. However, there is another SNP, rs72778561, nearby and minor allele frequency > 0.05. In this case, we cannot arbitrarily draw the conclusion, and further investigation is required. For the remaining 4 significant signals, they don’t have nearby confounding SNPs and we are safe to conclude that these CpG signals are significantly influenced by fatty acid GLA.

The *Table 2: Supplemental table of significant signals* recorded all the significant CpG sites in the five fatty acids, along with the p-values, gene names. Below is the preview. The link of the complete table is [here](file:///C:\Users\13915\OneDrive\Desktop\Plan%20B\Table2.html).

A screenshot of a computer

Description automatically generated

*Table 2: Supplemental table of significant signals*

*Figure 3: QQ-plots* are attached below to verify the assumptions. The red line means the expected line. Obviously, we do not see the significant deviation in the QQ-plots, suggesting that we do not have uncontrolled confounding issues. Each plot has only several points out of the red line, which is consistent with the results in Manhattan plots. Therefore, our results are quite reasonable.

A group of graphs with numbers

Description automatically generated with medium confidence

A graph with a red line

Description automatically generated

**DMR Results**

The findings of DMR analysis were quite different from the linear association results. In fatty acids LA, AA, EDA and DGLA, we found one differential methylated region at chromosome 7, with slightly different range for each fatty acid (LA and EDA: 130132413-130132790, AA: 130132259-130132790, DGLA: 130132453-130132790). The gene found in this region is MEST, which is an imprinted gene, primarily expressed by the paternal allele. It plays a role in early development, particularly in mesodermal tissues.

However, in GLA, we found 2 different regions: chr 5: 140812236-140812558 and chr 22: 31688639-31688645, with genes PCDHGA and RP3-400N23.6, respectively. PCDHGA1 is one of many gamma-protocadherins, each playing a role in cell identity and adhesion processes. It could hypothetically influence cell adhesion mechanisms or inflammatory signaling, both key factors in atherosclerosis development. RP3-400N23.6 is a less-characterized gene, sometimes referred to as a long non-coding RNA (lncRNA). Although specific functions of RP3-400N23.6 are not well-defined, some lncRNAs modulate inflammatory pathways, endothelial cell function, or vascular smooth muscle cell behavior—key components in the development of atherosclerosis.

**Discussion**

**Summary / Findings:** Summing up the previous findings, we found some CpG sites were significantly associated with multiple fatty acids. It’s quite reasonable because of the biological mechanisms: metabolic pathways. For example, the cg19610905 reference gene region is FADS2. The protein encoded by this gene is a member of the fatty acid desaturase (FADS) gene family. FADS2 encodes an enzyme involved in the desaturation process of fatty acids, specifically in the conversion of linoleic acid to gamma-linolenic acid and arachidonic acid. Since LA, AA, and EDA are all substrates or products in the same metabolic pathway, they can influence the activity and regulation of the FADS2 gene. The gene functions include regulating unsaturation of fatty acids through the introduction of double bonds between defined carbons of the fatty acyl chain. ([GeneCards](https://www.genecards.org/cgi-bin/carddisp.pl?gene=FADS2)).

We may also notice for the first four fatty acids, the significant CpG are all in chromosome 11. One possible reason is gene clustering. Chromosome 11 contains a cluster of genes involved in lipid metabolism, including the fatty acid desaturase (FADS) gene cluster (FADS1, FADS2, and FADS3). These genes are important for the desaturation and elongation of fatty acids, making them the crucial roles in the metabolism of LA, AA, DGLA and EDA. The other reason is probably because of the shared regulatory elements. Chromosome 11 may have shared enhancers, promoters, or other regulatory elements that respond to fatty acid levels. These regulatory elements could control multiple genes within the same region.

GLA is the exception among those fatty acids. Unlike others, the significant CpG of GLA spread to five different chromosomes, which are not in FADS gene region. For example, the second significant CpG, cg00938641, with the reference gene PLXNB3, coding a member of the plexin family. It plays a role as a receptor for semaphorin 5A, and functions in axon guidance, invasive growth and cell migration. Alternatively spliced transcript variants encoding different isoforms have been found for this gene. ([GeneCards](https://www.genecards.org/cgi-bin/carddisp.pl?gene=PLXNB3))

**DMR interpretation:** While we got different outcomes for linear association and DMR analysis, both results were reasonable. Previously, we ran the linear regression to individual CpG sites. In other words, we analyzed every CpG site independently and calculated the p-values. Under this situation, we found several CpG sites that had the most significant p-values at chromosome 11, which reference gene was FADS2.

However, in the DMR analysis, we focused on the regions that grouped nearby CpG sites with similar methylation patterns. It meant under this region, each CpG site may not extreme significant itself, but when they combined and acted together, potentially reflecting a more general or coordinated epigenetic regulation.

**Limitation / Future work:** While most of our results were reliable, there were still some limitations: Some of the methylation measurements on the array may be confounded by proximity to SNPs, and cross-hybridisation to other areas of the genome. (cross-hybridisation means in genome study, some probes not only match and detect the targeting DNA sequence, but also match the similar and repeated DNA sequence, which potentially interrupts the DNA signal and leads to inaccurate results).

Another limitation was DNA methylation data was measured at a later visit as compared to the FA measurement, which could lead to bias. To validate our results, the replication study of a new FA and DNA methylation data was recommended.

In conclusion, we identified the relationship between PUFAs and DNA methylation patterns at multiple CpG sites. Furthermore, we found the related gene, realizing the gene functions and the potential biological applications. Future investigation is necessary to remove the confounding nearby SNPs and do the replication studies.

**References**

Lindsay M.Reynolds, Ma Wan, Jingzhong Ding, et al. *DNA Methylation of the Aryl Hydrocarbon Receptor Repressor Associations With Cigarette Smoking and Subclinical Atherosclerosis.* (2015)**AHA Journals**.

<https://www.ahajournals.org/doi/10.1161/CIRCGENETICS.115.001097?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%20%200pubmed>

Tim Peters. *DMRcate*. (2024) **Bioconductor**.

<https://bioconductor.org/packages/release/bioc/html/DMRcate.html>

Kasper Daniel Hansen. *IlluminaHumanMethylation450kanno.ilmn12.hg19* (2024)

<https://bioconductor.org/packages/release/data/annotation/html/IlluminaHumanMethylation450kanno.ilmn12.hg19.html>

**GeneCards.**

<https://www.genecards.org/cgi-bin/carddisp.pl?gene=FADS2>

Stephen D. Turner. *Intro to the qqman package.* (2023). **CRAN.R-Project.**

<https://cran.r-project.org/web/packages/qqman/vignettes/qqman.html>