Leveraging Multi-Omics data to Understand Genetic and Biological Mechanisms underlying Fatty Acid Metabolism and Coronary Artery Disease

Chaojie Yang
Guangdong, China

Bachelor of Science, Biostatistics, Southern Medical University, 2015
Master of Science, Biostatistics, Georgetown University, 2016

A Dissertation presented to the Graduate Faculty of the University of Virginia in Candidacy for the Degree of Doctor of Philosophy

Department of Biochemistry and Molecular Genetics

University of Virginia
December 2023

Dr. Ani Manichaikul (Mentor)
Dr. Stefan Bekiranov (Committee head)
Dr. Clint Miller (Member)
Dr. Coleen McNamara (Member)

Table of Contents

Abstract

Polyunsaturated fatty acids (PUFAs) and their metabolites play critical roles in various aspects of human physiology and health and coronary artery disease (CAD) is a leading cause of mortality worldwide. In my dissertation work, I leveraged multi-omics data to understand genetic and biological mechanisms underlying Fatty Acid metabolism and coronary artery disease using multiple statistical approaches. In the first chapter, I described the different types of molecular ‘omics data, the available analytical techniques, and the backgrounds of PUFAs and CAD. In the second chapter, I estimated the global proportional of Amerind ancestry in 1102 Hispanic Americans from the Multi-Ethnic Study of Atherosclerosis (MESA), and demonstrated strong negative associations between Amerind genetic ancestry and PUFA levels. In the third chapter, I performed a meta-analysis of Genome-wide association study (GWAS) of PUFAs in Hispanic Americans and African Americans to identify multiple novel signals spanning a > 9 Mb region on chromosome 11 (57.5 Mb ~ 67.1 Mb) and demonstrated that multiple associations are unique to Hispanic Americans. In the fourth chapter, I applied colocalization analysis and correlation network analysis to prioritize seven potential causal genes of coronary artery diseases and subclinical atherosclerosis using MESA multi-omics data. In the last chapter, I summarized my research work and discussed the future direction of leveraging the molecular ‘omics data to provide a comprehensive insight of the biological mechanism of the complex human diseases.

Dedication

I would like to extend my heartfelt gratitude to my PhD mentor, Dr. Ani Manichaikul, for her unwavering dedication, profound expertise, and tireless commitment to nurturing my academic and professional development. Her patience, never-ending support and unwavering encouragement have proven invaluable to my journey.

To committee member, Dr. Stefan Bekiranov, Dr. Coleen McNamara and Dr. Clint Miller, I am profoundly appreciative of the invaluable contributions you have made in shaping the depth and excellence of this work. I also would like to express my sincere gratitude for your valuable suggestion regarding my career path.

To the current and previous group member at Dr. Ani Manichaikul's lab, Xiaowei Hu, Catherine Debban, Yanlin Ma, Chansuk Kang, Jenny Nguyen, I deeply appreciate the comments and the suggestion you provided for my research work. I also cherish the moments of our collaboration and discussions.

To my parents, Jiquan Yang and Jinhui Li, my brother, Chaoqun Yang, thank you for providing me with love, understanding, and the freedom to pursue my dreams.

To my girlfriend, Jiamin Yao, thanks for your company, unconditional support, and kindness during my entire PhD journey. Your presence has been a source of strength and comfort.

List of Figures

Chapter 1

Introduction

Sections of this chapter are adapted from:
Yang C, Hallmark B, Chai JC, O'Connor TD, Reynolds LM, Wood AC, Seeds M, Chen YI, Steffen LM, Tsai MY, Kaplan RC, Daviglus ML, Mandarino LJ, Fretts AM, Lemaitre RN, Coletta DK, Blomquist SA, Johnstone LM, Tontsch C, Qi Q, Ruczinski I, Rich SS, Mathias RA, Chilton FH, Manichaikul A. Impact of Amerind ancestry and FADS genetic variation on omega-3 deficiency and cardiometabolic traits in Hispanic populations. Commun Biol. 2021 Jul 28;4(1):918. doi: 10.1038/s42003-021-02431-4

Yang C, Veenstra J, Bartz TM, Pahl MC, Hallmark B, Chen YI, Westra J, Steffen LM, Brown CD, Siscovick D, Tsai MY, Wood AC, Rich SS, Smith CE, O'Connor TD, Mozaffarian D, Grant SFA, Chilton FH, Tintle NL, Lemaitre RN, Manichaikul A. Genome-wide association studies and fine-mapping identify genomic loci for n-3 and n-6 polyunsaturated fatty acids in Hispanic American and African American cohorts. Commun Biol. 2023 Aug 16;6(1):852. doi: 10.1038/s42003-023-05219-w.

Chilton FH, Manichaikul A, **Yang C**, O'Connor TD, Johnstone LM, Blomquist S, Schembre SM, Sergeant S, Zec M, Tsai MY, Rich SS, Bridgewater SJ, Mathias RA, Hallmark B. Interpreting Clinical Trials With Omega-3 Supplements in the Context of Ancestry and FADS Genetic Variation. Front Nutr. 2022 Feb 8;8:808054. doi: 10.3389/fnut.2021.808054.

1.1 Human genetics

1.1.1 Background of human genetics

Human genetics is the study of how human traits are influenced by genetic factors. The origins of the field of human genetics can be historically traced to the year 1949, marked by two seminal discoveries of paramount significance.[1] Firstly, sickle cell anemia was classified as an autosomal recessive phenotype by James V. Neel.[2] Secondly, this disorder was described as a “molecular” disease by Linus Pauling.[3] Notably, it was also in 1949 that the American Journal of Human Genetics was founded.[1]

There are enormous advancements in the evolution of modern human genetics. For example, the first advancement is widely attributed to the field of cytogenetics, including substantial progress in cell culture techniques and preparation of mitotic chromosomes suitable for examination under light microscopy. These breakthroughs demonstrated that a multitude of human disorders can be attributed to specific aberrations in the numerical or structural integrity of chromosomes.[4,5] Secondly, the advancements in high-throughput sequencing technologies, commonly referred to as next-generation sequencing (NGS), have facilitated the generation of extensive and diverse genomic data (as discussed in **Section 1.1.5**). These data types include DNA Sequencing Data (Whole-Genome Sequencing and Exome Sequencing), Transcriptome data, Proteome data, DNA Methylation Data and Functional Genomics Data (ChIP-Seq Data and ATAC-seq Data). These data play a pivotal role in offering valuable insights into the structure and function of different levels of molecular targets.[6–8] Thirdly, the advancement of analytical approaches applied to molecular omics data, including Genome-wide association study (GWAS, as discussed in **Section 1.1.4**), statistical fine-mapping (as discussed in **Section 1.6.1.1**), Quantitative Trait Loci (QTL, as discussed in **Section 1.6.1.2**) mapping and integrative analysis (as discussed in **Section 1.6.1.3** and **Section 1.6.1.3**), have substantially enhanced our capacity to interpret the genetic underpinnings of traits and phenotypes by enabling better interpretation of their genetic basis.

Table 1.1 demonstrates the major advances in human genetics from 1949 to 2020.[1] Collectively, these advancements have provided critical insights into the biological mechanisms and pathways of human diseases, underscoring the crucial importance of human genetics in the field of medicine. For example, human genetics enables early disease diagnosis, often before symptoms manifest and further allows for timely interventions and appropriate treatments, especially in cases of hereditary conditions and certain cancers. [9–11]

	New concepts and approaches in human genetics	**Main author(s)**
1949	First "molecular" disease. Sickle cell anemia genetics	L Pauling, JV Neel
	Term "Human Genetics". First textbook Human Genetics	C. Stern
1950	First Medical Genetics Unit at Montreal Children´s Hospital	C Scriver, FC Fraser
1952	First human enzyme defect	C Cori & G Cori
	First autosomal linkage group in man (Lutheran/Secretor)	J Mohr
1953	Dietary therapy in phenylketonuria	H Bickel
1954	Leukocyte drumsticks	Davidson & Smith
	Book on *Counseling in Medical Genetics*	Sheldon Reed
1955	Buccal smear X-Chromatin analysis	Moore, Barr, Marberger
1956	Concept of genetic heterogeneity	CF Fraser; H. Harris
	Heritable Disorders of Connective Tissue	VA McKusick
1957	Triplets encode the 20 amino acids	S Brenner
1958	HLA antigen genetic system	J Dausset; PI Terasaki
	Somatic cell genetics	G Ponetcorvo
1959	First chromosomal aberrations (Trisomy 21 / XXY / XO)	J Lejeune; PA Jacobs; CE Ford
	Pharmacogenetics as a new concept	AG Motulsky; F Vogel; ES Vesell

1960	New autosomal trisomy 18 and trisomy 13.	JH Edwards; K Patau
	Philadelphia chromosome	PC Nowell & D Hungerford
⋮	⋮	⋮
2004	Evolutiionary Medicine	PD Glucksman & MA Hanson
2005	X-inactivation profile / X-chromosome sequenced	I Carrel & HF WIllard; MT Ross et al
	DNA sequence of the human X-cromosome	MT Ross et al
2006	Human genome chromosome by chromosome	Nature Suppl. 1 June 2016
2008	Epigenome Project / Fanconi anemia protein complex	D Schindler & H Hoehn
2009	Principles of Evolutionary Medicine	P Gluckman, A Beadle, M Hanson
2011	Chromothripsis in oncogenesis	PJ Stephens
2012	Topological domains in chromatin structure	JR Dixon; RE Thurmann
2013	Cancer genomics	B Vogelstein
2015	Cancer Genome Atlas	National Cancer Institute (NIH)
2017	Autologous transgenic skin gene therapy	M De Luca, T Hirsch
2018	GTEx-Genotype-Tissue Expression Project/ Cancer immunotherapy	NIH Common Fund
2019	Mutated cell clones in normal tissues	K Yizhak et al (Science 364(6444)p970
	Diabetes type 2 associated with more than 400 gene variants	M Roden & GI Shulman
2020	Structural variation in 17,795 human genomes	HJ Abel et al (Nature 583:83-89)
	Embryonic neurodevelopment disrupted in Chorea Huntington	M Barnat et al

Table 1.1: The major advances relevant to Human Genetics from 1949 to 2020. Table adapted from Passarge E et al. 2021.

1.1.2 Heritability

Heritability is the most common measurement used to study the genetic contribution to quantitative traits and disease outcomes.[12,13] Heritability estimates range from zero to one. A heritability estimate of zero implies that the majority of the variability observed in a trait among individuals is primarily attributable to environmental factors, with minimal influence stemming from genetic distinctions, while estimate of one indicates that nearly all of the variability observed in a trait is driven by genetic differences among individuals. One of the approaches of qualifying heritability is to use the liability threshold model to explain how a large number of environmental and genetic factors result in a disease. The liability threshold model uses a continuous latent liability, capturing genetic and environmental factors that influence disease risk, to describe the disease status. This model assumes that the disease occurs when the subject's liability exceeds a certain threshold, which means the disease prevalence can be represented by the probability of liability exceeds the threshold. Moreover, this model assumes that the genetic and environmental risks are independent and normally distributed. Thus, in this model, the liability is normally distributed with mean zero and reflects the sum of the variance of genetic and environmental risks. Based on the liability threshold model, heritability is defined as the ratio of genetic variation to phenotypic variation.[14–17]

1.1.3 Genetic ancestry and population structure

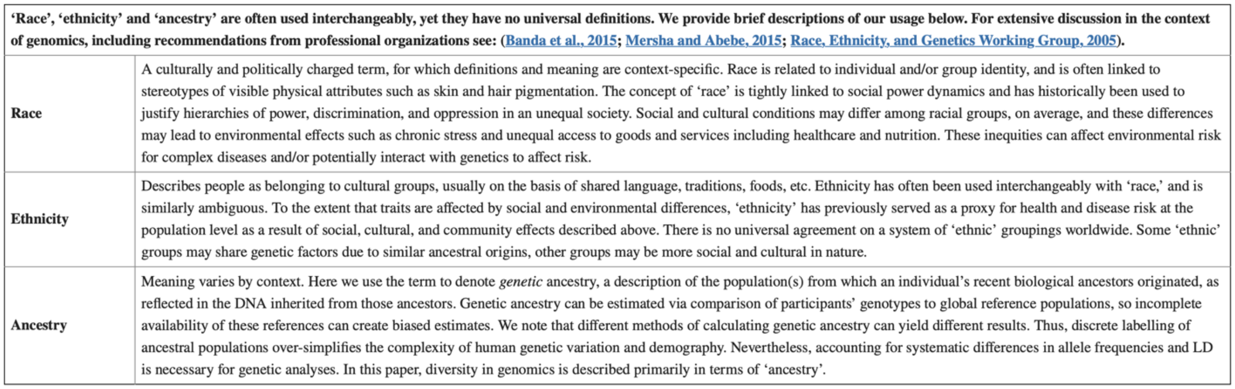
Genetic ancestry is one of the most critical topics in human genetics and refers to an individual's ancestral origins as quantified by genetics. Individuals who share similar ancestral origins exhibit common genomic signatures.[18] Compared to the traditional self-reported race/ethnicity, genetic ancestry is increasingly utilized in modern human genetics due to its precise characterization of individuals' biological ancestry.[19,20] **Box 1.1** demonstrates the concepts and differences among race, ethnicity and genetic ancestry. Genetic ancestry studies have conclusively shown disparities between estimation of an individual's genetic ancestry and the self-reported race or ethnicity information provided by the same individuals.[21] For example, self-identified African Americans can exhibit significant variation in their levels of African and European ancestry.[22–25]

Interestingly, Hispanic Americans, constituting the largest ancestral minority population in the United States, exhibit a complex genetic structure and genetic admixture, with large variation in the contributions from Amerind, European, and African ancestry. Manichaikul A, et al. conducted a population structure analysis for Hispanic Americans in the Multi-Ethnic Study of Atherosclerosis (MESA), with MESA Hispanic Americans representing six major countries/regions of origin: Central America, Cuba, the Dominican Republic, Mexico, Puerto Rico, and South America. This study revealed that the global proportion of European ancestry in MESA Hispanic Americans ranged from 37% in Central Americans to 73% among Cubans. The global proportion of African ancestry was 43% in Dominicans, while the global proportion of Amerind ancestry was 45 and 48% in Central Americans and Mexicans, respectively.[26]

Genetic ancestry can be estimated based on an individual's genetic information, for example, the individual's genotypes as compared to genotypes from global references of human genetic variation (including the Human Genetic Diversity project [HGDP][27] and the 1000 Genomes project[28]). Principal component analysis (PCA)[29] and model-based cluster analysis (ADMIXTURE[30] , STRUCTURE[31]) are employed routinely for the investigation of population structures. The key feature of PCA is to reduce the dimensionality of genetic data to identify the patterns of genetic variation among the individuals. Unlike PCA, model-based cluster analysis aims to identify a specific number of clusters and assign the individual to these clusters probabilistically.

Diverse ancestral groups comprise a spectrum of distinct disease prevalence profiles, highlighting the complexity of health disparities across ancestry groups. For example, African Americans exhibit a lower level of coronary artery calcium (CAC) but greater level of carotid intima-media thickness (cIMT) compared to European Americans.[32] On the other hand, Hispanic Americans tend to show reduced CAC compared to African Americans and lower levels of cIMT compared to African Americans.[32] In addition, Hispanic Americans exhibit notably elevated prevalence of both diabetes and nonalcoholic fatty liver disease (NAFLD) in comparison to other ancestral groups within the United States.[33,34]

Incorporating genetic ancestry into human genetic studies plays an critical role in advancing human health. Firstly, genetic ancestry enhances precision medicine, which could provide improvements in efficacy and accuracy of medical treatments and interventions for individuals from diverse ancestral groups. Secondly, genetic ancestry offers a comprehensive perspective on health disparities among various ancestral groups. Thirdly, genetic ancestry can enable a better understanding of historical migration patterns and admixture events.[18,20,35–39]



'Race', 'ethnicity' and 'ancestry' are often used interchangeably, yet they have no universal definitions. We provide brief descriptions of our usage below. For extensive discussion in the context of genomics, including recommendations from professional organizations see: (Banda et al., 2015; Mersha and Abebe, 2015; Race, Ethnicity, and Genetics Working Group, 2005).	
Race	A culturally and politically charged term, for which definitions and meaning are context-specific. Race is related to individual and/or group identity, and is often linked to stereotypes of visible physical attributes such as skin and hair pigmentation. The concept of 'race' is tightly linked to social power dynamics and has historically been used to justify hierarchies of power, discrimination, and oppression in an unequal society. Social and cultural conditions may differ among racial groups, on average, and these differences may lead to environmental effects such as chronic stress and unequal access to goods and services including healthcare and nutrition. These inequities can affect environmental risk for complex diseases and/or potentially interact with genetics to affect risk.
Ethnicity	Describes people as belonging to cultural groups, usually on the basis of shared language, traditions, foods, etc. Ethnicity has often been used interchangeably with 'race,' and is similarly ambiguous. To the extent that traits are affected by social and environmental differences, 'ethnicity' has previously served as a proxy for health and disease risk at the population level as a result of social, cultural, and community effects described above. There is no universal agreement on a system of 'ethnic' groupings worldwide. Some 'ethnic' groups may share genetic factors due to similar ancestral origins, other groups may be more social and cultural in nature.
Ancestry	Meaning varies by context. Here we use the term to denote *genetic* ancestry, a description of the population(s) from which an individual's recent biological ancestors originated, as reflected in the DNA inherited from those ancestors. Genetic ancestry can be estimated via comparison of participants' genotypes to global reference populations, so incomplete availability of these references can create biased estimates. We note that different methods of calculating genetic ancestry can yield different results. Thus, discrete labelling of ancestral populations over-simplifies the complexity of human genetic variation and demography. Nevertheless, accounting for systematic differences in allele frequencies and LD is necessary for genetic analyses. In this paper, diversity in genomics is described primarily in terms of 'ancestry'.

Box1.1: the concepts and differences among 'race', 'ethnicity' and 'genetic ancestry', Box adapted from Peterson RE O et al. 2019.

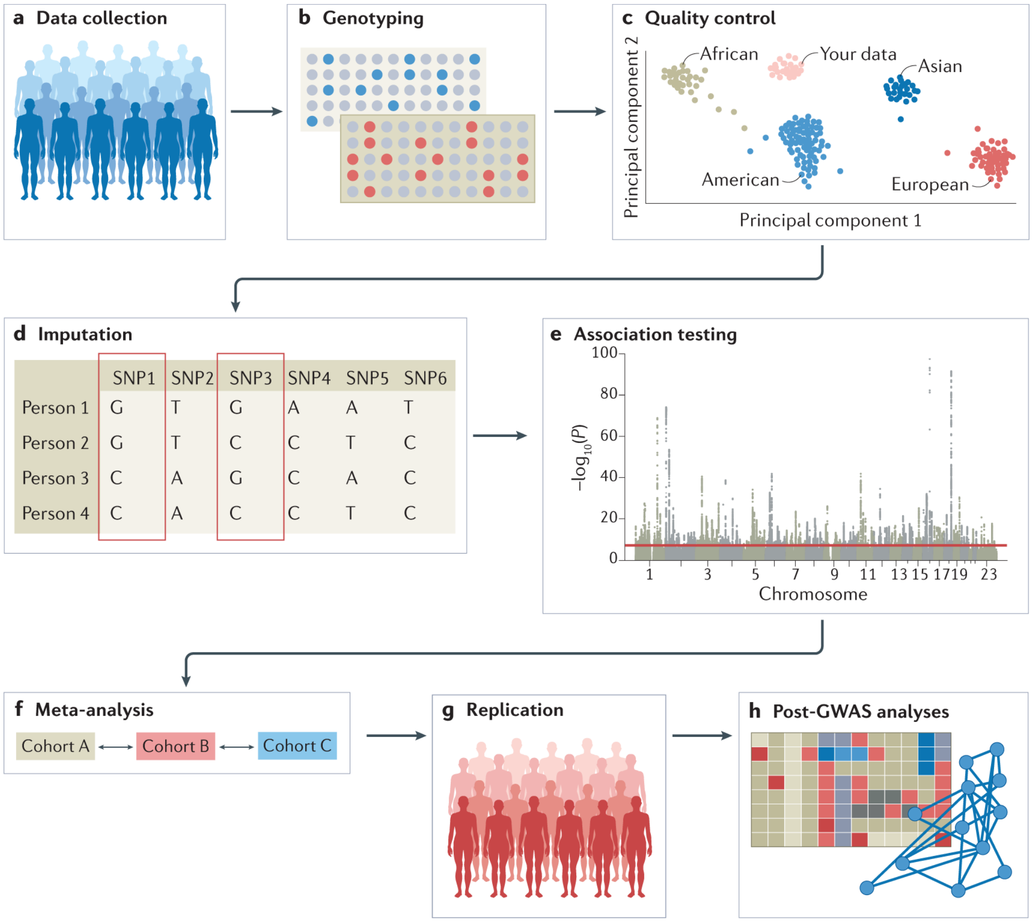
1.1.4 Genome-Wide Association Studies

Genome-wide association studies (GWAS) have become an established approach for testing comprehensively the association between genetic variants or single nucleotide polymorphisms (SNPs) and particular traits. GWAS have been applied to study a variety of human diseases, including breast cancer, coronary artery disease and type 2 diabetes.[40,41] The NHGRI-EBI GWAS Catalog (www.ebi.ac.uk/gwas) is a curated collection of all human genome-wide association studies, produced by a collaboration between EMBL-EBI and NHGRI. The NHGRI-EBI GWAS Catalog is a Findable, Accessible, Interoperable, and Reusable (FAIR) knowledgebase, which contains ~400,000 SNP-trait associations across > 5,000 human traits from > 45,000 individual GWAS.[42,43]

The steps of conducting GWAS involve: (1) Selection of study cohorts, phenotypes and covariates measurements -- notably, large sample sizes are required for GWAS to have sufficient statistical power; (2) Obtaining individual-level genotype for GWAS using microarray-based genotyping and next-generation sequencing -- in most cases, imputation is required for the variants that have not been assayed directly using a reference panel such as the 1000 Genomes Project or TOPMed; and (3). Association testing between the phenotypes and genotypes using appropriate models, such as linear or logistic regression models. In addition, covariates can be adjusted in the models, for example, age, sex and genetic ancestry. Multiple comparison correction is applied routinely after the association testing, to take into account the vast number of statistical associations examine by GWAS analysis. **Figure 1.1** demonstrates the workflow of conducting GWAS.[40]

GWAS results exhibit a wide spectrum of applications within the field of genetics research, prominently exemplified by their utility in disease risk prediction. Notably, the development of Polygenic Risk Scores (PRS), derived from GWAS summary statistics, has emerged as a valuable tool for prediction of disease risk, which can enable the identification of individuals at high risk of disease for clinical interventions and preventions. In addition, GWAS results provide a foundation for comprehensive insights for understanding genetic architecture of traits and estimating their heritability, representing the proportion of trait variance attributable to genetic factors within the population. These applications can guide pharmaceutical research and the development of targeted therapies, ultimately improving treatment efficacy and reducing side effects.[44–47]

However, GWAS approaches alone fall short in identification of causal variants due to linkage disequilibrium (LD) which complicates identification of the disease-causing molecular targets. In addition, the majority of lead associated variants from GWAS lie within non-coding regions, making it difficult to predict their functions and interpret the biological significance.[44] Follow-up analyses including statistical fine-mapping and integrative analysis using molecular ‘omics data are necessary for the identification of the causal variants and downstream target genes.



	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Person 1	G	T	G	A	A	T
Person 2	G	T	C	C	T	C
Person 3	C	A	G	C	A	C
Person 4	C	A	C	C	T	C

Figure 1.1: Overview of steps for conducting GWAS. Figure adapted from Uffelmann, E et al. 2021.

1.1.5 Molecular ‘omics data

Advancements in high-throughput sequencing technology offer an opportunity to obtain the molecular phenotypes within a tissue or cell and the types of molecular ‘omics data include genomics, transcriptomics, proteomics, metabolomics and epigenomics. The genome is the complete sequence of DNA and genomics mainly focuses on identifying genetic features associated with disease, including identification of genomic variants by genome-wide association study. The transcriptome represents the entire set of RNA transcripts derived from DNA, including messenger RNA (mRNA), non-coding RNA and more. Transcriptomics consists of both qualitative aspects, involving the identification of which transcripts are present, the discovery of novel splice sites, and the detection of RNA editing sites, as well as quantitative aspects, involving the measurement of the abundance of each transcript expressed within a biological sample. The proteome represents the complete set of proteins expressed by a cell or tissue and proteomics is employed for the quantification of peptide abundance, the investigation of protein modification and interaction.[48–50] **Figure 1.2** demonstrates multiple molecular ‘omics data types. Multiple analytic approaches can be conducted for the investigation of molecular omics data, such as integrative analysis (see **Section 1.1.6.3 and 1.1.6.4**), correlation network analysis, machine learning and deep learning. Additionally, studying multi-omics data offers a more comprehensive view of the downstream effects of genetic variation compared to studies of a single omics type. For example, genetic effects on plasma protein abundance are often, but not exclusively, driven by regulation of mRNA and pQTL associations without corresponding eQTL evidences may reflect genetic effects on processes other than transcription, including protein degradation, binding, secretion, or clearance from circulation.[48,51]

Integration of different types of molecular omics data can prioritize a list of candidate disease-altering biomarkers and molecular targets, which can help us come to an improved understanding of the biological mechanisms and pathways of human disease and offer guidance in the diagnosis and prognosis of diseases, such as stroke, cardiovascular diseases, diabetes and others.[51,52] For example, in cancer research, G protein-coupled receptor clusters have been identified as breast cancer-associated biomarkers by studying the proteomics data, providing useful insight in guiding breast cancer therapy.[53,54] **Figure 1.3** demonstrates the application of multi-omics in multiple research areas.

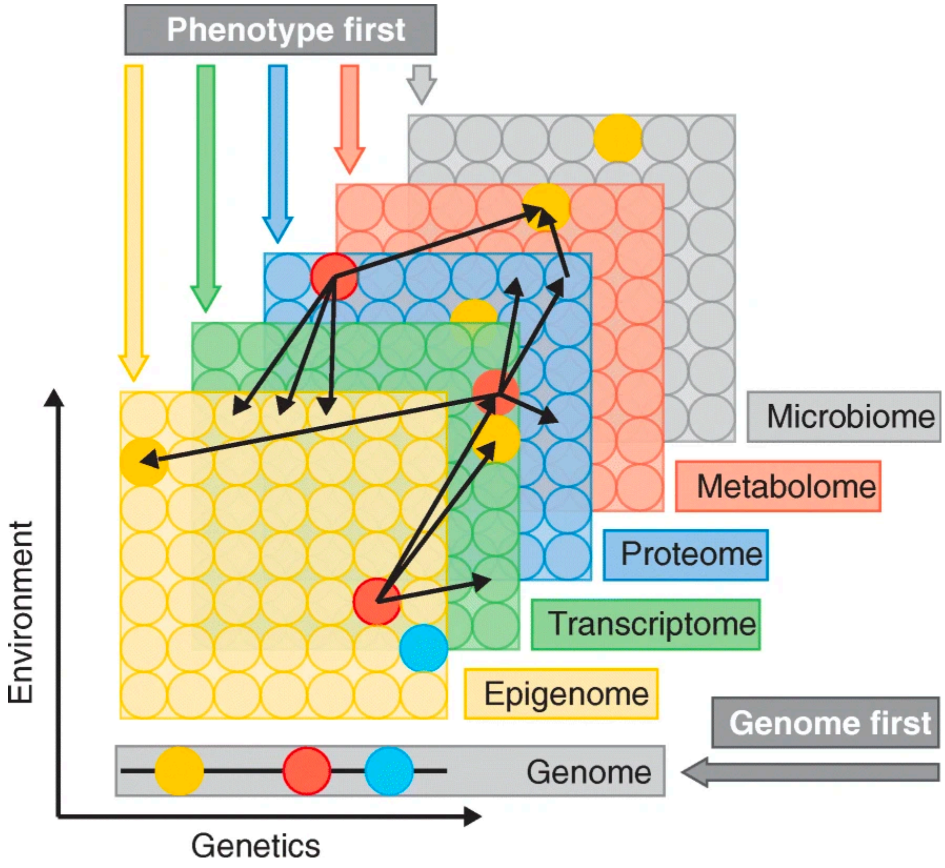


Figure 1.2: multiple molecular 'omics data types. Figure adapted from Hasin, Y et al. 2017.

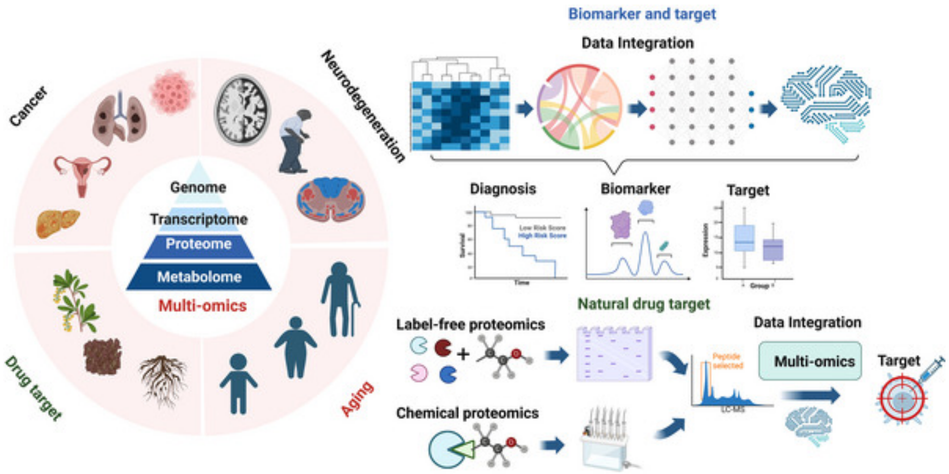


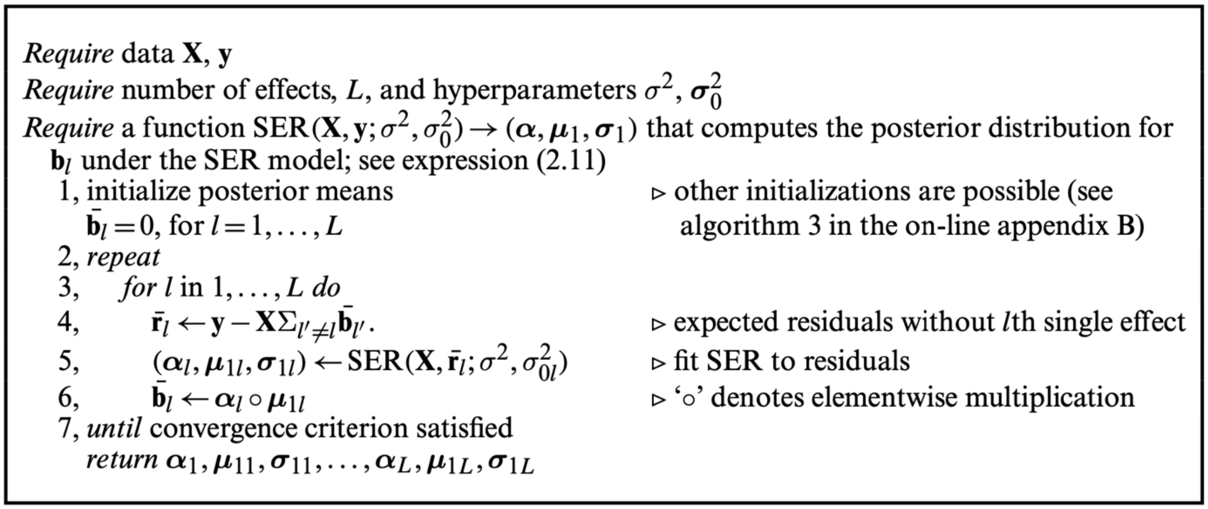
Figure 1.3: The application of multi-omics in disease, aging, and natural drug target identification. Figure adapted from Chen C, et al. 2020.

1.1.6 Follow-up analysis of GWAS leveraging molecular ‘omics data

1.1.6.1 Identification of causal variants using statistical fine-mapping

Genome-wide association studies (GWAS) have emerged as a powerful tool for uncovering a multitude of genetic associations with complex diseases and the primary goal of fine-mapping is to discern the genetic variants that causally affect the examined traits. However, determining the causal variants is challenging as the genetic variants within a locus can be highly correlated due to linkage disequilibrium among neighboring SNPs.[55–57]

One of the popular approaches in statistical fine-mapping is ‘sum of single effects' model (SuSiE), which introduces multiple single-effect vectors and constructs the overall effect vector as the sum of these single effects. A key feature of this model is that it can generate "Credible Sets" (CSs), which are each designed to have high probability to contain a signal with non-zero effect, while at the same time being as small as possible. The fitting procedure of this model is Iterative Bayesian Step-wise Selection (IBSS), which is a simple and intuitive procedure.[58] **Box 1.2** shows the IBSS algorithm. Compared with currently available methods, for example, DAP-G, FINEMAP and CAVIAR, SuSiE offers several significant advantages in terms of computational efficiency and its capacity to yield credible sets capturing the causal variants. For example, SuSiE operates at approximately four times the speed of DAP-G, 30 times the speed of FINEMAP, and a staggering 4000 times the speed of CAVIAR. Additionally, the credible sets generated by SuSiE consistently outperform those produced by DAP-G, exhibiting higher power, smaller size, and greater purity.[58]



```
Require data $\mathbf{X}, \mathbf{y}$
Require number of effects, $L$, and hyperparameters $\sigma^2$, $\boldsymbol{\sigma}_0^2$
Require a function $\mathrm{SER}(\mathbf{X}, \mathbf{y}; \sigma^2, \sigma_0^2) \rightarrow (\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1)$ that computes the posterior distribution for
  $\mathbf{b}_l$ under the SER model; see expression (2.11)
  initialize posterior means                            ▹ other initializations are possible (see
    $\bar{\mathbf{b}}_l = 0$, for $l = 1, \dots, L$              algorithm 3 in the on-line appendix B)
  repeat
      for $l$ in $1, \dots, L$ do
          $\bar{\mathbf{r}}_l \leftarrow \mathbf{y} - \mathbf{X}\Sigma_{l' \neq l}\bar{\mathbf{b}}_{l'}$.       ▹ expected residuals without $l$th single effect
          $(\boldsymbol{\alpha}_l, \boldsymbol{\mu}_{1l}, \boldsymbol{\sigma}_{1l}) \leftarrow \mathrm{SER}(\mathbf{X}, \bar{\mathbf{r}}_l; \sigma^2, \sigma_{0l}^2)$    ▹ fit SER to residuals
          $\bar{\mathbf{b}}_l \leftarrow \boldsymbol{\alpha}_l \circ \boldsymbol{\mu}_{1l}$                  ▹ '○' denotes elementwise multiplication
  until convergence criterion satisfied
    return $\boldsymbol{\alpha}_1, \boldsymbol{\mu}_{11}, \boldsymbol{\sigma}_{11}, \dots, \boldsymbol{\alpha}_L, \boldsymbol{\mu}_{1L}, \boldsymbol{\sigma}_{1L}$
```

Box 1.2: IBSS algorithm. Box adapted from Wang et al. 2020.

1.1.6.2 Quantitative Trait Loci (QTL) mapping of high-throughput molecular 'omics traits

Most common disease-associated genetic variants identified by GWAS are located within non-coding genomic regions, suggesting they may affect gene regulation.[59] This observation has promoted studies of the relationship between regulatory variants and potential downstream molecular targets (QTLs). Advancements in high-throughput sequencing technology have provided data on various molecular phenotypes (gene expression, protein abundance, metabolites, and DNA methylation). The goal of quantitative trait loci (QTL) mapping is to examine the association between genetic variants and downstream molecular phenotypes (e.g., gene expression [eQTL], protein abundance [pQTL] and DNA methylation [mQTL]). QTL mapping is of paramount importance because they yield invaluable resources with diverse downstream applications. These resources empower researchers to identify and prioritize therapeutic targets for complex diseases, facilitating the development of more effective treatments. Additionally, they enable the implementation of precision medicine strategies, allowing healthcare professionals to tailor medical interventions to individual genetic profiles, thereby enhancing patient outcomes.[60–62]

MatrixeQTL and FastQTL are commonly used approaches to perform the QTL mapping. MatrixeQTL is employed to assess the associations between SNPs and gene expression by implementing either additive linear model or ANOVA model with various covariates adjustment. The computation time of Matrix eQTL is relatively faster due to its specific processing and its use of large matrix operations.[63] FastQTL further incorporates permutation strategies, including a direct permutation scheme, an adaptive permutation scheme and a beta approximation to address the issues of multiple testing of correlated SNPs.[64] QTL mapping is computationally intensive due to the complexity of analyzing large scale of datasets containing genetic information for numerous traits and individuals. It involves the exploration of numerous genetic variants across the genome for the identification of associations with specific traits or diseases. This process requires extensive computational power and time, especially as datasets continue to grow and complexity. Recently, TensorQTL was developed to conduct QTL mapping, which is a tool designed to implement on GPU, which is 100 times faster than CPU.

1.1.6.3 Bayesian colocalization analysis

Integrating molecular QTLs with GWAS represents an important step for interpreting the biological and clinical relevance of the GWAS results.[65,66] Bayesian colocalization analysis has proven a rigorous and efficient computational approach for identification of downstream molecular targets underlying GWAS loci.[67] Colocalization analysis quantifies the probability that a single variant is causally linked to both disease and molecular traits, with the following hypotheses: H0. neither GWAS nor molecular QTL has a genetic association in the region; H1. only GWAS has a genetic association in the region; H2. only molecular QTL has a genetic association in the region; H3. both GWAS and molecular QTL are associated, but with different causal variants; H4. both GWAS and molecular QTL are associated and share a single causal variant. **Figure 1.4** shows the example of one configuration under different hypotheses. A distinguishing feature of this approach is the minimal input data requirement, which includes p-values of the trait-associated SNPs and their minor allele frequencies (MAFs), or effect size of SNPs and the corresponding standard error. Compared to the original implementation of colocalization analysis, one recent methodological improvement involves incorporation of statistical fine-mapping (SuSiE), which can account for multiple causal variants underlying the GWAS and molecular QTL signals in a region.[68]

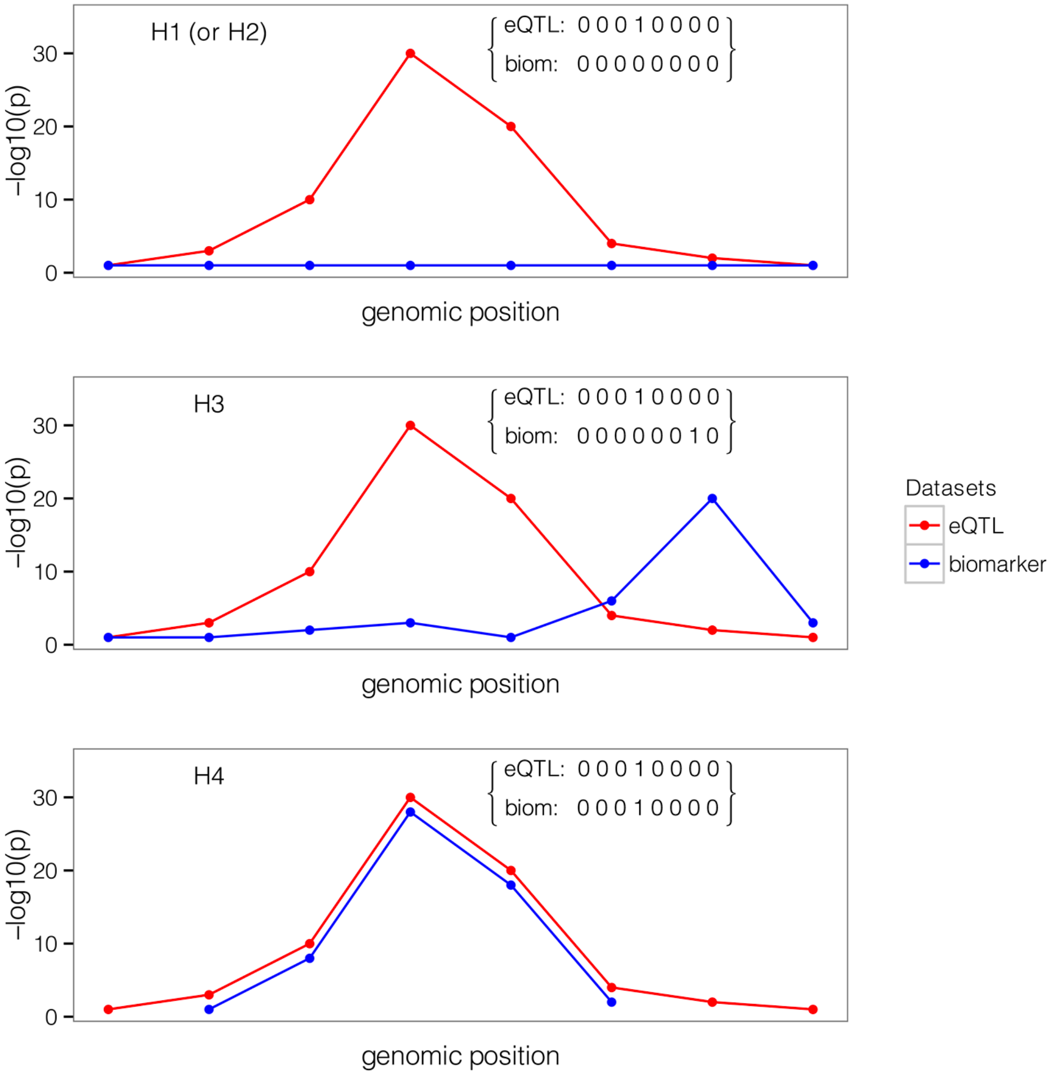
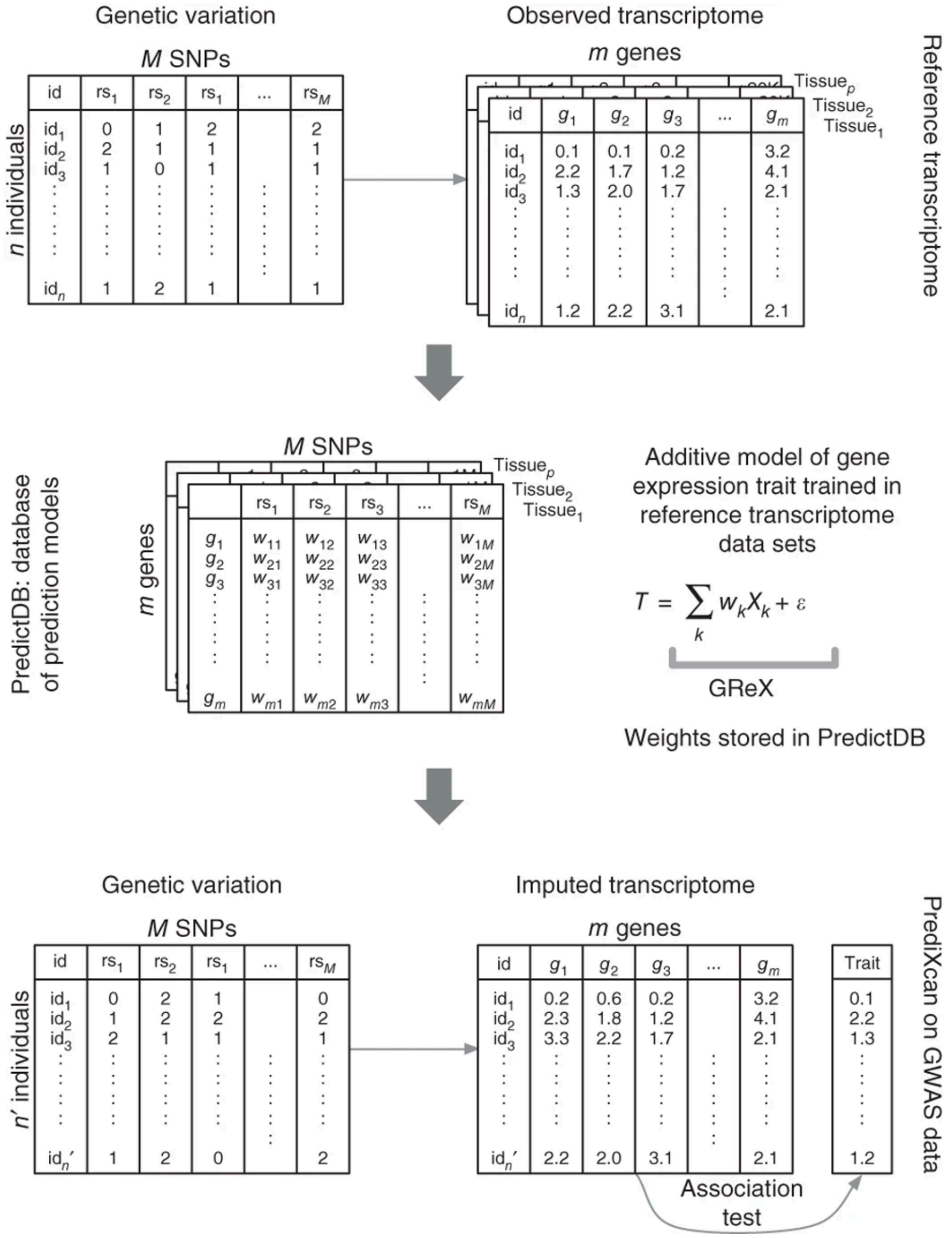


Figure 1.4: Example of one configuration under different hypotheses.. Figure adapted from Giambartolomei C et al. 2014.

1.1.6.4 Transcriptome-wide association studies or PrediXcan

The transcriptome-wide association study (TWAS) or PrediXcan framework represents a gene-based association method designed to identify trait-associated genes by predicting genetically regulated gene expression and quantifying the effect of predicted gene expression on the phenotype on interest. An individual's gene expression level can be dissected into distinct components. These components encompass (1) a genetically regulated expression (GReX) component, which signifies the portion of gene expression influenced by genetic variants, (2) a component subject to alteration by the phenotypes/traits under investigation, (3) a component that accounts for variations in gene expression attributable to environmental influences and other pertinent factors. PrediXcan examines the mediating role of gene expression on the phenotype of interest by quantifying the association between GReX and the specific phenotype of interest.[69,70] **Figure 1.5** demonstrates the workflow of PrediXcan.

The prediction model is a fundamental component in PrediXcan because it enables the estimation of gene expression levels in tissues that might not be directly measured in a particular study. By using this model, researchers can predict the gene expression levels for various tissues based on genetic information. This prediction is crucial because it allows for the assessment of how genetic variants influence gene expression and, consequently, impact complex traits or diseases. Prediction models for gene expression were developed for incorporation in the PrediXcan framework using large-scale transcriptome study data sets, including DGN[71], GEUVADIS[72] and GTEx[73]. The advantages of PrediXcan include a smaller multiple-testing burden compared to single-variant tests, independence from actual transcriptome data and straightforward construction of informative priors and groupings of functional units. In addition, PrediXcan provides effect estimates and direction of effect, which can enhance interpretation of the phenotype-gene expression relationship and inform downstream therapeutic development.



id	rs$_1$	rs$_2$	rs$_1$	...	rs$_M$
id$_1$	0	1	2		2
id$_2$	2	1	1		1
id$_3$	1	0	1		1
⋮	⋮	⋮	⋮	⋮	⋮
id$_n$	1	2	1		1

id	g_1	g_2	g_3	...	g_m
id$_1$	0.1	0.1	0.2		3.2
id$_2$	2.2	1.7	1.2		4.1
id$_3$	1.3	2.0	1.7		2.1
⋮	⋮	⋮	⋮	⋮	⋮
id$_n$	1.2	2.2	3.1		2.1

	rs$_1$	rs$_2$	rs$_3$	...	rs$_M$
g_1	w_{11}	w_{12}	w_{13}		w_{1M}
g_2	w_{21}	w_{22}	w_{23}		w_{2M}
g_3	w_{31}	w_{32}	w_{33}		w_{3M}
⋮	⋮	⋮	⋮	⋮	⋮
g_m	w_{m1}	w_{m2}	w_{m3}		w_{mM}

$$\underbrace{T = \sum_k w_k X_k + \varepsilon}_{\text{GReX}}$$

id	rs$_1$	rs$_2$	rs$_1$	...	rs$_M$
id$_1$	0	2	1		0
id$_2$	1	2	2		2
id$_3$	2	1	1		1
⋮	⋮	⋮	⋮	⋮	⋮
id$_{n}$′	1	2	0		2

id	g_1	g_2	g_3	...	g_m
id$_1$	0.2	0.6	0.2		3.2
id$_2$	2.3	1.8	1.2		4.1
id$_3$	3.3	2.2	1.7		2.1
⋮	⋮	⋮	⋮	⋮	⋮
id$_{n}$′	2.2	2.0	3.1		2.1

Trait
0.1
2.2
1.3
⋮
1.2

Figure 1.5: The workflow of PrediXcan. Figure adapted from Gamazon E et al. 2015.

1.2 Fatty Acid Metabolism

1.2.1 Background of Polyunsaturated Fatty Acids

Polyunsaturated fatty acids (PUFAs) are critical structural components of cell membranes. PUFAs, along with the signaling metabolites originating from them, hold pivotal roles in processes associated with inflammation and thrombosis. These processes are implicated in the pathogenesis of various medical conditions, encompassing cardiovascular disease (CVD), Alzheimer's disease (AD), type 2 diabetes, autoimmune disorders, cancer, hypersensitivity conditions, skin and gastrointestinal disorders, and infectious diseases such as COVID-19.[74,75]

Polyunsaturated fatty acids (PUFAs) are characterized by the position of their first double bond counted from the methyl terminal, often referred to as omega (ω) or n−FAs. These PUFAs are typically categorized into two primary families, known as n-3 and n-6. Within the n-3 PUFA family, the most prevalent members include alpha-linolenic acid (ALA), eicosapentaenoic acid (EPA), docosapentaenoic acid (DPA), and docosahexaenoic acid (DHA). Conversely, the primary n-6 PUFAs encompass linoleic acid (LA), gamma-linolenic acid (GLA), dihomo-γ-linolenic acid (DGLA), and arachidonic acid (AA). ALA is synthesized in plants and once produced and subsequently consumed by humans, ALA can be converted to EPA, DPA, and DHA. and their metabolic products have different and often opposing effects. Indeed, n-3 and n-6 PUFAs and their corresponding metabolic products exert diverse and frequently contradictory effects, as supported by numerous studies. Metabolites derived from the n-6 PUFA arachidonic acid (ARA) primarily operate at localized sites to stimulate inflammatory responses. Conversely, n-3 LC-PUFAs, including eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA), along with their metabolites, possess anti-inflammatory properties and actively promote the resolution of inflammation.[76,77] In addition to their effects on inflammation, circulating levels of n-3 LC-PUFAs, including EPA and DHA, are inversely associated with fasting and postprandial serum TG concentrations, largely through attenuation of hepatic very-low-density lipoprotein (VLDL)-TG production.[78]

1.2.2 Previous GWAS of Polyunsaturated Fatty Acids

Genome-wide association studies (GWAS) of n-3 and n-6 PUFAs were performed by the CHARGE consortium in European ancestry (EUR) participants.[79,80]

The CHARGE GWAS of n-3 PUFAs in 8,866 European Americans across population-based cohorts identified that SNPs in/near *FADS* cluster are showing the significant association with higher levels of ALA ($p = 3 \times 10^{-64}$) and lower levels of EPA ($p = 5 \times 10^{-58}$) and DPA ($p = 4 \times 10^{-154}$) as well as SNPs in/near *ELOVL2* strongly associated with higher EPA ($p = 2 \times 10^{-12}$) and DPA ($p = 1 \times 10\text{-}^{43}$) and lower DHA ($p = 1 \times 10^{-15}$). Overall, the most significantly associated SNPs on chromosome 11 explained 3.8%, 2.0%, 8.6% of total variation in ALA, EPA and DPA, respectively and the most significantly associated SNPs on chromosome 6 explained 0.4%, 2.8%, 0.7% of total variation in EPA, DPA and DHA, respectively. This GWAS results of n-3 PUFAs suggested that the genetic variation in *FADS* cluster affect the conversion of ALA to EPA and DPA, while the genetic variation in *ELOVL2* decrease the conversion of EPA and DPA to DHA, which provided a comprehensive insight of genetic variation in shaping the circulating levels of n-3 PUFAs. **Figure 1.6** demonstrates the major metabolic pathway connecting n-3 polyunsaturated fatty acids and presents a summary of genome-wide associations overlapping that pathway.

For n-6 PUFAs, a CHARGE GWAS was performed in five prospective studies comprising 8,631 European Americans. In addition to the confirmation of the genetic association of *FADS* cluster in LA and AA, which explained a large proportion of total variation in n-6 PUFA (8.7% -11.1% for DGLA), this study identified that multiple novel significant SNPs in/near *NRBF2* on chromosome 10 associated with LA (rs10740118, p-value = 8.1×10^{-9}) and in/near *NTAN1* on chromosome 6 associated with LA, GLA, DGLA and AA (rs16966952, p-value = 1.2×10^{-15}, 5.0×10^{-11}, 7.6×10^{-65}, and 2.4×10^{-10}). Overall, *NRBF2* variant rs10740118 explained 0.2-0.7% of total variation in LA and *NTAN1* rs16966952 explained 0.1-0.6% to 2.0-4.5% of total variation in AA and DGLA. **Figure 1.6** demonstrates the n-6 polyunsaturated fatty acid metabolic pathway and presents a summary of genome-wide associations overlapping that pathway.[81]

Collectively, these findings of n-3 and n-6 PUFAs in European Americans provide a valuable foundation for guiding future investigations into the genetic and metabolic pathways that potentially affect the levels of n-3 and n-6 PUFAs. However, one of the limitations of these GWAS of n-3 and n-6 PUFAs is the paucity in non-European ancestry cohorts, for example, Hispanic American and African American. Due to the potential difference of genetic associations with ancestry, it is important to perform the GWAS of n-3 and n-6 PUFAs across diverse ancestry groups.[82]

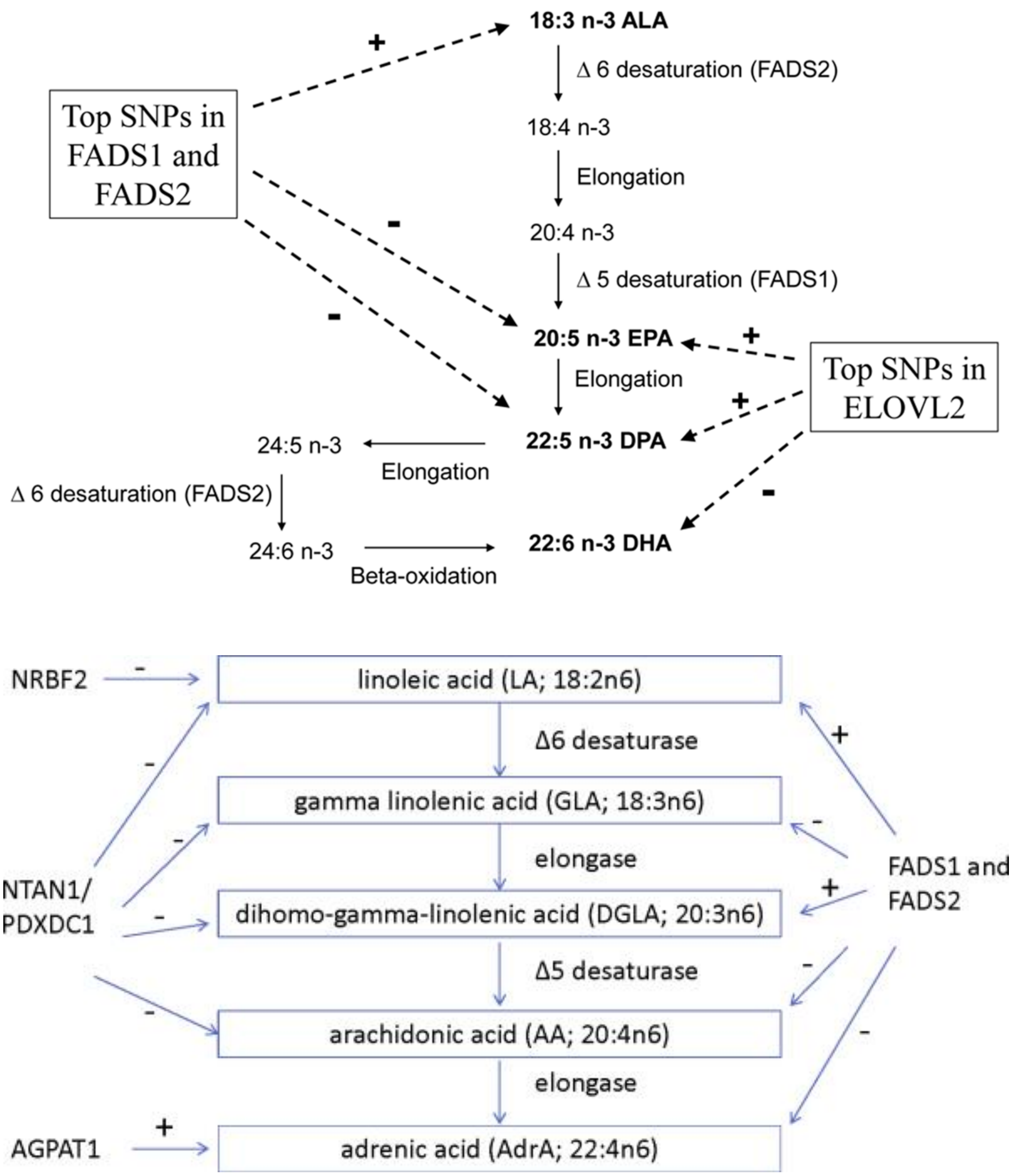


Figure 1.6: n-3 and n-6 polyunsaturated fatty acid metabolic pathway and summary of genome-wide associations in pathway. Figure adapted from Lemaitre RN et al. 2011 and Guan W et al. 2014

.

1.3 Coronary Artery Disease (CAD)

1.3.1 Background of Coronary Artery Disease (CAD)

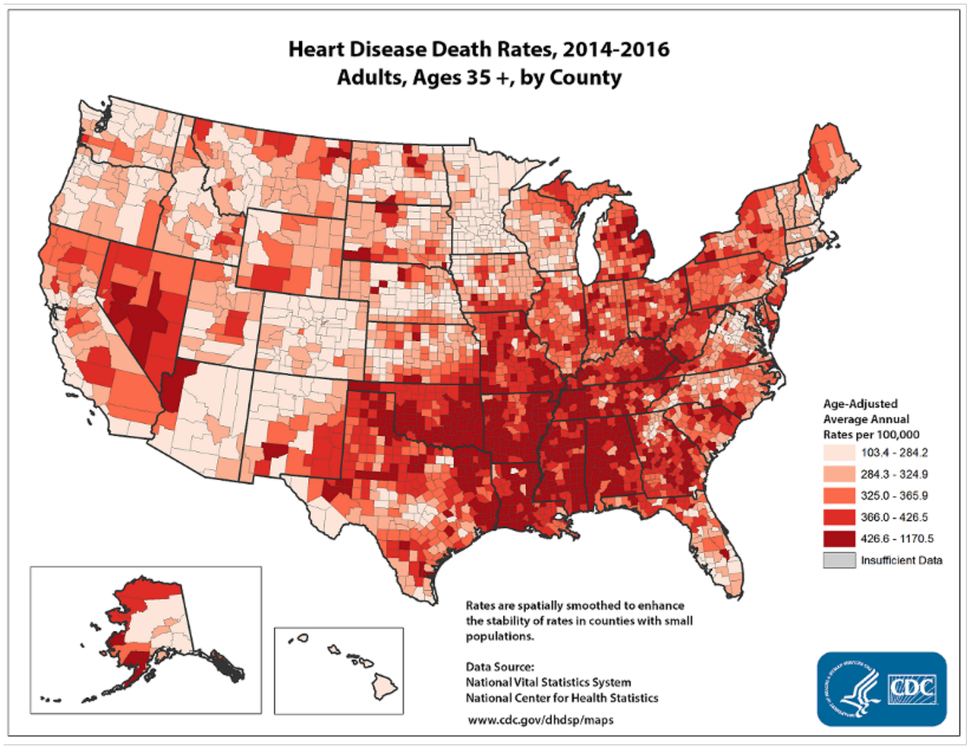
Coronary artery disease (CAD) is a leading cause of death and disability worldwide. About 18.2 million Americans adults suffer from coronary artery disease and half of all deaths correlated with cardiovascular disease arise from coronary artery disease.[83,84] **Figure 1.7** shows the heart disease death rate. CAD represents an archetypal common complex disease with both genetic and environmental determinants. It is usually caused by atherosclerosis, which is the buildup of plaque inside the arteries. Atherosclerotic plaque is composed of a complex mixture, such as, cholesterol, fatty substances, waste products, calcium deposits, and the clot-making substance fibrin. Over time, as plaque accumulates along the inner walls of arteries, it induces a narrowing and stiffening of these vessels. The presence of plaque can lead to significant arterial obstruction and damage, thereby obstructing the flow of blood to the heart muscle. Without an adequate blood supply, the heart becomes starved of oxygen and the vital nutrients it needs to work properly. [85–89] The most common symptoms of CAD include angina, chest pain, shortness of breath and back breath.

Risk factors for CAD can be categorized into non-modifiable risk factors, which include male gender, family history of heart disease, advanced age and race, and modifiable factors, which include cigarette smoking, high blood pressure, overweight and unhealthy diet. Primary prevention and treatment strategies are designed to reduce the modifiable risk factors of CAD, including smoking cessation, promoting healthy dietary habits, encouraging regular physical activity, monitoring and controlling high blood pressure and managing cholesterol levels. [90,91]

Coronary artery disease, characterized by the narrowing of the arteries in the heart, is a prevalent factor contributing to heart failure.[92] Several epidemiological studies suggest that coronary artery disease account for 23-73% of the heart failure in the patients evaluated.[93] In addition, a population-based study of incidence and aetiology of heart failure shows that CAD is usually the major cause on patients with a reduced left ventricular ejection fraction.[94] Moreover, a GWAS meta-analysis of heart failure estimated that the genetic correlation between heart failure and CAD is 0.67, which suggests they are sharing the genetic aetiology.[95]

1.3.2 Previous GWAS of Coronary Artery Disease (CAD)

GWAS for CAD have yielded numerous significant and promising findings, tracing their origins back to 2007 when the risk locus on chromosome 9p21 risk locus was identified.[96–98] Starting form 2011, multiple large GWAS consortia including Coronary ARtery DIsease Genome-wide Replication and Meta-analysis (CARDIoGRAM) Consortium,[99] the Coronary Artery Disease (C4D) Genetics Consortium[100] and UK Biobank (UKBB) performed their individual GWAS of CAD with a large sample size and identified multiple loci associating with CAD at a genome-wide level of significance. In 2017, Pim van der Harst *et al*. conducted a meta-analysis GWAS of CAD using CARDIoGRAMplusC4D and UK Biobank resourcese comprising 34,541 CAD cases and 261,984 controls of UK Biobank resource and 88,192 cases and 162,544 controls from CARDIoGRAMplusC4D. They identified 64 novel genetic loci of CAD.[101] In 2022, Tcheandjieu, C. *et al.* performed a large scale GWAS of CAD in genetically diverse population using Million Veteran Program (MVP), UK Biobank, CARDIoGRAMplusC4D and Biobank Japan. This large-scale multi-ethnic study includes European, African American, Hispanics and Asian and they identified 95 novel genetic loci of CAD. This study provides the significance of including diverse population in the genetic study to understand genetic architecture of CAD.[102] Collectively, these GWAS of CAD findings offers valuable and comprehensive insights into the genetic basis of CAD, further enhance our understanding of the potential biological mechanism and pathway of CAD.[103] **Figure 1.8** demonstrates milestones in cardiovascular genome research from 2007 to 2017 and beyond. However, the molecular consequences of the GWAS variants of CAD and their relevance to subclinical atherosclerosis have not been explored comprehensively in human cohorts.



Heart Disease Death Rates, 2014-2016
Adults, Ages 35 +, by County

Rates are spatially smoothed to enhance the stability of rates in counties with small populations.

Data Source:
National Vital Statistics System
National Center for Health Statistics

www.cdc.gov/dhdsp/maps

Figure 1.7: Heart disease death rate. 2014-2016. Figure adapted from Centers for Disease Control and Prevention.

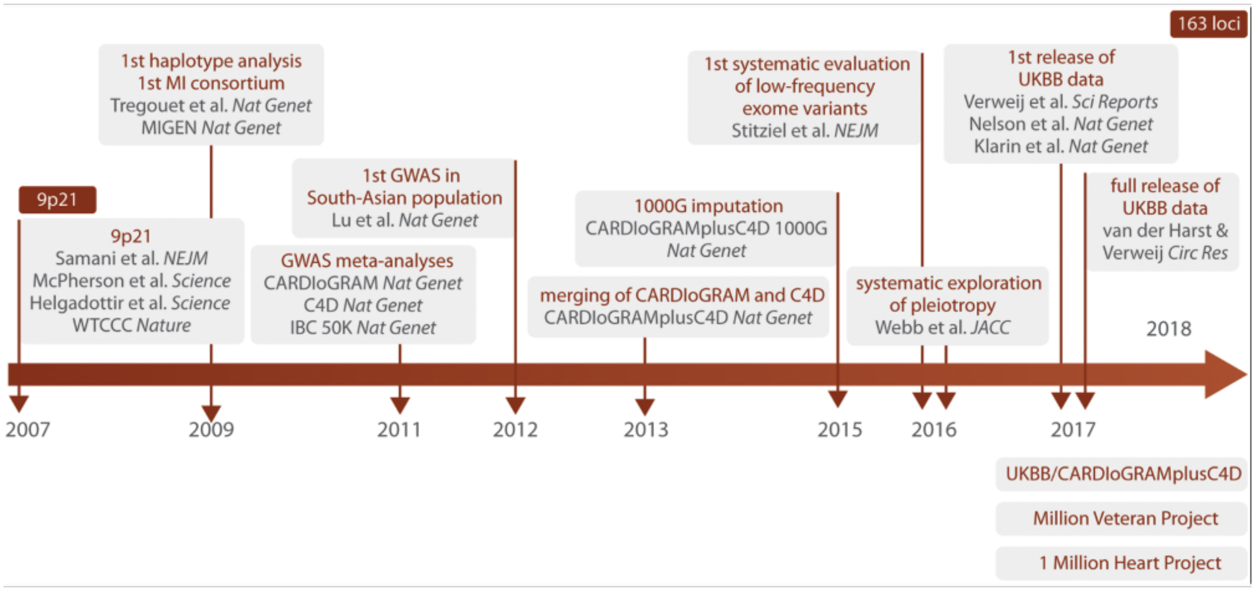


Figure 1.8: Milestones in cardiovascular genome research from 2007 to 2017 and beyond. Figure adapted from Erdmann J et al. 2018

1.4 Summary and motivation for original research

The overarching objective of human genetics is to explore the impact of genetic factors on human traits. The progress in high-throughput sequencing technologies has empowered us to access various types of molecular omics data, while the availability of diverse analysis methods facilitates the exploration of these molecular omics data. This, in turn, provides a comprehensive understanding of the underlying biological mechanisms and pathways involved in complex diseases, which can enable a better diagnosis, prevention and treatments.

Polyunsaturated fatty acids (PUFAs) play vital roles in innate immunity, energy homeostasis, brain development and cognitive function and coronary artery disease (CAD) is a leading cause of death and disability worldwide. Previous genetic studies including genome-wide association study of PUFAs and CAD demonstrate valuable foundation of genetic basis in PUFAs and CAD. However, numerous gaps in knowledge of PUFAs and CAD remain. For example, the impact of population differences in allele frequencies on population-specific risk of traits deficiency/diseases have not beenwell examined; (2), most of the genetic information gathered to date is mainly focus on European ancestry and the paucity of genetic studies in non-European ancestry cohorts may cause the challenges of achieving precise diagnosis and further treatments; (3), although GWAS have identified multiple genetic variants of PUFAs and CAD, the follow-up identification of trait/disease-related causal variants and molecular targets remains elusive. Overall, these limitations have contributed to the relatively limited success of genetic studies in the realm of human health.

In my dissertation work, I aimed to apply various and novel statistical approaches to address these biological questions and limitations using the molecular omics data in the following studies:

(1). In chapter 2, I estimated the global proportion of Amerind ancestry in MESA Hispanic Americans and further investigated the impact of Amerind ancestry and *FADS* genetic variation on polyunsaturated fatty acids.

(2). In chapter 3, I conducted a meta-analysis of GWAS of PUFAs in Hispanic Americans African American and follow-up analysis including fine-mapping and integrative analysis to identity the potential causal variants and genes.

(3). In chapter 4, I leveraged the multi-omics data to prioritize a list of candidate genes associated with CAD and subclinical atherosclerosis by integrative analysis and correlation network analysis.

(4). In chapter 5, I share my ideas and thoughts on future directions for improvement of applying statistical methods on molecular omics data to enable a better understanding of the genetic basis of complex diseases/traits.

1.5 References

1. Passarge, E. Origins of human genetics. A personal perspective. *Eur. J. Hum. Genet.* **29**, 1038–1044 (2021).
2. Neel, J. V. The Inheritance of Sickle Cell Anemia. *Science* **110**, 64–66 (1949).
3. Pauling, L. & Itano, H. A. Sickle cell anemia a molecular disease. *Science* **110**, 543–548 (1949).
4. Ferguson-Smith, M. A. History and evolution of cytogenetics. *Mol. Cytogenet.* **8**, 19 (2015).
5. Polani, P. E. Human and clinical cytogenetics: origins, evolution and impact. *Eur. J. Hum. Genet. EJHG* **5**, 117–128 (1997).
6. Behjati, S. & Tarpey, P. S. What is next generation sequencing? *Arch. Dis. Child. Educ. Pract. Ed.* **98**, 236–238 (2013).
7. Chaitankar, V. *et al.* Next Generation Sequencing Technology and Genomewide Data Analysis: Perspectives for Retinal Research. *Prog. Retin. Eye Res.* **55**, 1–31 (2016).
8. Qin, D. Next-generation sequencing and its clinical application. *Cancer Biol. Med.* **16**, 4–10 (2019).
9. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
10. Ommen, G. van, Bakker, E. & Dunnen, J. den. The human genome project and the future of diagnostics, treatment, and prevention. *The Lancet* **354**, S5–S10 (1999).

11. Musunuru, K. *et al.* Genetic Testing for Inherited Cardiovascular Diseases: A Scientific Statement From the American Heart Association. *Circ. Genomic Precis. Med.* **13**, e000067 (2020).

12. Estimating Trait Heritability | Learn Science at Scitable. https://www.nature.com/scitable/topicpage/estimating-trait-heritability-46889/.

13. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).

14. Liability threshold modeling of case–control status and family history of disease increases association power | Nature Genetics. https://www.nature.com/articles/s41588-020-0613-6.

15. Dahlqwist, E., Magnusson, P. K. E., Pawitan, Y. & Sjölander, A. On the relationship between the heritability and the attributable fraction. *Hum. Genet.* **138**, 425–435 (2019).

16. Falconer, D. S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.* **29**, 51–76 (1965).

17. Benchek, P. H. & Morris, N. J. How meaningful are heritability estimates of liability? *Hum. Genet.* **132**, 10.1007/s00439-013-1334-z (2013).

18. Jorde, L. B. & Bamshad, M. J. Genetic Ancestry Testing What Is It and Why Is It Important? *JAMA* **323**, 1089–1090 (2020).

19. Peterson, R. E. *et al.* Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589–603 (2019).

20. Dries, D. L. Genetic Ancestry, Population Admixture, and the Genetic Epidemiology of Complex Disease. *Circ. Cardiovasc. Genet.* **2**, 540–543 (2009).

21. Peralta, C. A. *et al.* Differences in Albuminuria Between Hispanics and Whites: An Evaluation by Genetic Ancestry and Country of Origin. *Circ. Cardiovasc. Genet.* **3**, 240–247 (2010).

22. Shraga, R. *et al.* Evaluating genetic ancestry and self-reported ethnicity in the context of carrier screening. *BMC Genet.* **18**, 99 (2017).

23. Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D. & Mountain, J. L. The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. *Am. J. Hum. Genet.* **96**, 37–53 (2015).

24. Mersha, T. B. & Abebe, T. Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. *Hum. Genomics* **9**, 1 (2015).

25. Price, A. L. *et al.* Discerning the Ancestry of European Americans in Genetic Association Studies. *PLOS Genet.* **4**, e236 (2008).

26. Manichaikul, A. *et al.* Population structure of Hispanics in the United States: the multi-ethnic study of atherosclerosis. *PLoS Genet.* **8**, e1002640 (2012).

27. Cavalli-Sforza, L. L. The Human Genome Diversity Project: past, present and future. *Nat. Rev. Genet.* **6**, 333–340 (2005).

28. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

29. Lever, J., Krzywinski, M. & Altman, N. Principal component analysis. *Nat. Methods* **14**, 641–642 (2017).

30. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

31. Porras-Hurtado, L. *et al.* An overview of STRUCTURE: applications, parameter settings, and supporting software. *Front. Genet.* **4**, (2013).

32. Wassel, C. L. *et al.* Genetic Ancestry Is Associated With Subclinical Cardiovascular Disease in African-Americans and Hispanics From the Multi-Ethnic Study of Atherosclerosis. *Circ. Cardiovasc. Genet.* **2**, 629–636 (2009).

33. Saab, S., Manne, V., Nieto, J., Schwimmer, J. B. & Chalasani, N. P. Nonalcoholic Fatty Liver Disease in Latinos. *Clin. Gastroenterol. Hepatol. Off. Clin. Pract. J. Am. Gastroenterol. Assoc.* **14**, 5–12; quiz e9-10 (2016).

34. Williams, C. D. *et al.* Prevalence of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis among a largely middle-aged population utilizing ultrasound and liver biopsy: a prospective study. *Gastroenterology* **140**, 124–131 (2011).

35. Mensah, G. A., Mokdad, A. H., Ford, E. S., Greenlund, K. J. & Croft, J. B. State of disparities in cardiovascular health in the United States. *Circulation* **111**, 1233–1241 (2005).

36. Lewis, A. C. F. *et al.* Getting genetic ancestry right for science and society. *Science* **376**, 250–252 (2022).

37. Wollstein, A. & Lao, O. Detecting individual ancestry in the human genome. *Investig. Genet.* **6**, 7 (2015).

38. Farrell, R. M., Pederson, H. & Padia, S. Incorporating Genetic Testing Ancestry Results into Medical Decisions. *AMA J. Ethics* **16**, 428–433 (2014).

39. Borrell, L. N. *et al.* Race and Genetic Ancestry in Medicine — A Time for Reckoning with Racism. *N. Engl. J. Med.* **384**, 474–480 (2021).

40. Uffelmann, E. *et al.* Genome-wide association studies. *Nat. Rev. Methods Primer* **1**, 1–21 (2021).

41. Witte, J. S. Genome-Wide Association Studies and Beyond. *Annu. Rev. Public Health* **31**, 9–20 (2010).

42. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).

43. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).

44. Tam, V. *et al.* Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).

45. Korte, A. & Farlow, A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* **9**, 29 (2013).

46. Cottrell, P. Advantages and Limitation of GWAS. SSRN Scholarly Paper at https://doi.org/10.2139/ssrn.3220662 (2018).

47. Frayling, T. Genome-wide association studies: the good, the bad and the ugly. *Clin. Med.* **14**, 428–431 (2014).

48. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 83 (2017).

49. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinforma. Biol. Insights* **14**, 1177932219899051 (2020).

50. Micheel, C. M. *et al.* Omics-Based Clinical Discovery: Science, Technology, and Applications. in *Evolution of Translational Omics: Lessons Learned and the Path Forward* (National Academies Press (US), 2012).

51. Chen, C. *et al.* Applications of multi-omics analysis in human diseases. *MedComm* **4**, e315 (2023).

52. Kedaigle, A. & Fraenkel, E. Turning omics data into therapeutic insights. *Curr. Opin. Pharmacol.* **42**, 95–101 (2018).

53. Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55–62 (2016).

54. Neagu, A.-N. *et al.* Proteomics-Based Identification of Dysregulated Proteins in Breast Cancer. *Proteomes* **10**, 35 (2022).

55. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).

56. LaPierre, N. *et al.* Identifying causal variants by fine mapping across multiple studies. *PLOS Genet.* **17**, e1009733 (2021).

57. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).

58. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).

59. Ignatieva, E. V. & Matrosova, E. A. Disease-associated genetic variants in the regulatory regions of human genes: mechanisms of action on transcription and

genomic resources for dissecting these mechanisms. *Vavilov J. Genet. Breed.* **25**, 18–29 (2021).

60. Grisel, J. E. & Crabbe, J. C. Quantitative Trait Loci Mapping. *Alcohol Health Res. World* **19**, 220–227 (1995).
61. Fernandes, P. B. Technological advances in high-throughput screening. *Curr. Opin. Chem. Biol.* **2**, 597–603 (1998).
62. Banovich, N. E. *et al.* Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet.* **10**, e1004663 (2014).
63. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinforma. Oxf. Engl.* **28**, 1353–1358 (2012).
64. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinforma. Oxf. Engl.* **32**, 1479–1485 (2016).
65. Hertzberg, L. Commentary: Integration of gene expression and GWAS results supports involvement of calcium signaling in Schizophrenia. *J. Ment. Health Clin. Psychol.* **2**, 5–7 (2018).
66. Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLOS Genet.* **13**, e1006646 (2017).
67. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, e1004383 (2014).

68. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLOS Genet.* **17**, e1009440 (2021).

69. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).

70. Li, B. *et al.* Evaluation of PrediXcan for prioritizing GWAS associations and predicting gene expression. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* **23**, 448–459 (2018).

71. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014).

72. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

73. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

74. Kothapalli, K. S. D., Park, H. G. & Brenna, J. T. Polyunsaturated fatty acid biosynthesis pathway and genetics. implications for interindividual variability in prothrombotic, inflammatory conditions such as COVID-19. *Prostaglandins Leukot. Essent. Fatty Acids* **162**, 102183 (2020).

75. Innis, S. M. Omega-3 fatty acid biochemistry: perspectives from human nutrition. *Mil. Med.* **179**, 82–87 (2014).

76. Serhan, C. N., Chiang, N. & Van Dyke, T. E. Resolving inflammation: dual anti-inflammatory and pro-resolution lipid mediators. *Nat. Rev. Immunol.* **8**, 349–361 (2008).

77. Serhan, C. N. Pro-resolving lipid mediators are leads for resolution physiology. *Nature* **510**, 92–101 (2014).

78. Oscarsson, J. & Hurt-Camejo, E. Omega-3 fatty acids eicosapentaenoic acid and docosahexaenoic acid and their mechanisms of action on apolipoprotein B-containing lipoproteins in humans: a review. *Lipids Health Dis.* **16**, 149 (2017).

79. Lemaitre, R. N. *et al.* Genetic Loci Associated with Plasma Phospholipid n-3 Fatty Acids: A Meta-Analysis of Genome-Wide Association Studies from the CHARGE Consortium. *PLOS Genet.* **7**, e1002193 (2011).

80. Guan, W. *et al.* Genome-wide association study of plasma N6 polyunsaturated fatty acids within the cohorts for heart and aging research in genomic epidemiology consortium. *Circ. Cardiovasc. Genet.* **7**, 321–331 (2014).

81. Guan, W. *et al.* Genome-Wide Association Study of Plasma N6 Polyunsaturated Fatty Acids within the CHARGE Consortium. *Circ. Cardiovasc. Genet.* **7**, 321–331 (2014).

82. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).

83. CDC. Heart Disease Facts | cdc.gov. *Centers for Disease Control and Prevention* https://www.cdc.gov/heartdisease/facts.htm (2023).

84. Ralapanawa, U. & Sivakanesan, R. Epidemiology and the Magnitude of Coronary Artery Disease and Acute Coronary Syndrome: A Narrative Review. *J. Epidemiol. Glob. Health* **11**, 169–177 (2021).

85. Frostegård, J. Immunity, atherosclerosis and cardiovascular disease. *BMC Med.* **11**, 117 (2013).

86. Hansson, G. K. Inflammation, Atherosclerosis, and Coronary Artery Disease. *N. Engl. J. Med.* **352**, 1685–1695 (2005).

87. Khoury, Z. *et al.* Relation of Coronary Artery Disease to Atherosclerotic Disease in the Aorta, Carotid, and Femoral Arteries Evaluated by Ultrasound. *Am. J. Cardiol.* **80**, 1429–1433 (1997).

88. Barrett-Connor, E. L. Obesity, Atherosclerosis, and Coronary Artery Disease. *Ann. Intern. Med.* **103**, 1010–1019 (1985).

89. Fong, I. W. Emerging relations between infectious diseases and coronary artery disease and atherosclerosis. *CMAJ* **163**, 49–56 (2000).

90. Brown, J. C., Gerhardt, T. E. & Kwon, E. Risk Factors For Coronary Artery Disease. in *StatPearls* (StatPearls Publishing, 2020).

91. Hajar, R. Risk Factors for Coronary Artery Disease: Historical Perspectives. *Heart Views Off. J. Gulf Heart Assoc.* **18**, 109–114 (2017).

92. Heart Failure - What Is Heart Failure? | NHLBI, NIH. https://www.nhlbi.nih.gov/health/heart-failure (2022).

93. Velagaleti, R. & Vasan, R. S. Heart Failure in the 21st Century: Is it a Coronary Artery Disease Problem or Hypertension Problem? *Cardiol. Clin.* **25**, 487–v (2007).

94. Cowie, M. R. *et al.* Incidence and aetiology of heart failure; a population-based study. *Eur. Heart J.* **20**, 421–428 (1999).

95. Shah, S. *et al.* Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nat. Commun.* **11**, 163 (2020).

96. Samani, N. J. *et al.* Genomewide association analysis of coronary artery disease. *N. Engl. J. Med.* **357**, 443–453 (2007).

97. McPherson, R. *et al.* A Common Allele on Chromosome 9 Associated with Coronary Heart Disease. *Science* **316**, 1488–1491 (2007).

98. Lieb, W. & Vasan, R. S. Brief Review: Genetics of coronary artery disease. *Circulation* **128**, 10.1161/CIRCULATIONAHA.113.005350 (2013).

99. Preuss, M. *et al.* Design of the Coronary ARtery DIsease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) Study: A Genome-wide association meta-analysis involving more than 22 000 cases and 60 000 controls. *Circ. Cardiovasc. Genet.* **3**, 475–483 (2010).

100. Coronary Artery Disease (C4D) Genetics Consortium. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nat. Genet.* **43**, 339–344 (2011).

101. van der Harst, P. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ. Res.* **122**, 433–443 (2018).

102. Tcheandjieu, C. *et al.* Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. *Nat. Med.* **28**, 1679–1692 (2022).

103. Erdmann, J., Kessler, T., Munoz Venegas, L. & Schunkert, H. A decade of genome-wide association studies for coronary artery disease: the challenges ahead. *Cardiovasc. Res.* **114**, 1241–1257 (2018).

Chapter 2

Impact of Amerind ancestry and FADS genetic variation on omega-3 deficiency and cardiometabolic traits in Hispanic populations

Published in: **Yang C**, Hallmark B, Chai JC, O'Connor TD, Reynolds LM, Wood AC, Seeds M, Chen YI, Steffen LM, Tsai MY, Kaplan RC, Daviglus ML, Mandarino LJ, Fretts AM, Lemaitre RN, Coletta DK, Blomquist SA, Johnstone LM, Tontsch C, Qi Q, Ruczinski I, Rich SS, Mathias RA, Chilton FH, Manichaikul A. Impact of Amerind ancestry and FADS genetic variation on omega-3 deficiency and cardiometabolic traits in Hispanic populations. Commun Biol. 2021 Jul 28;4(1):918. doi: 10.1038/s42003-021-02431-4

2.1 Abstract

Long chain polyunsaturated fatty acids (LC-PUFAs) have critical signaling roles that regulate dyslipidemia and inflammation. Genetic variation in the *FADS* gene cluster accounts for a large portion of interindividual differences in circulating and tissue levels of LC-PUFAs, with the genotypes most strongly predictive of low LC-PUFA levels at strikingly higher frequencies in Amerind ancestry populations. In this study, we examined relationships between genetic ancestry and *FADS* variation in 1,102 Hispanic American participants from the Multi-Ethnic Study of Atherosclerosis. We demonstrate strong negative associations between Amerind genetic ancestry and LC-PUFA levels. The *FADS* rs174537 single nucleotide polymorphism (SNP) accounted for much of the AI ancestry effect on LC-PUFAs, especially for low levels of n-3 LC-PUFAs. Rs174537 was also strongly associated with several metabolic, inflammatory and anthropomorphic traits including circulating triglycerides (TGs) and E-selectin in MESA Hispanics. Our study demonstrates that Amerind ancestry provides a useful and readily available tool to identify individuals most likely to have *FADS*-related n-3 LC-PUFA deficiencies and associated cardiovascular risk.

2.2 Introduction

Human diets in developed countries have changed dramatically over the past 75 years, leading to increased obesity, inflammation, cardio-metabolic disorders and cancer risk, possibly due to interactions between genotype with diet and other factors. Certain racial/ethnic groups carry a disproportionate burden of preventable negative outcomes and associated mortality.[1–3] Hispanic populations represent the largest racial/ethnic US minority where, compared to non-Hispanic whites, they have higher rates of obesity[4], poorly controlled high blood pressure[5], and elevated circulating triglycerides (TGs)[6], Hispanic populations also demonstrate a higher prevalence of diabetes and nonalcoholic fatty liver disease (NAFLD) than other racial/ethnic populations in the United States.[7,8] Hispanic Americans represent a heterogenous group with respect to ancestry, with notable differences in cultural/ lifestyle factors and disease prevalence based on country of origin. In particular, Hispanics identifying with the higher Amerind (AI)-ancestry origin have demonstrated enhanced urine albumin excretion[9], heart failure[10], lupus erythematosus risk[11], and prevalence of NAFLD compared to other Hispanic populations[12], supporting the critical need to conduct studies in these large, rapidly growing populations.

Omega-3 (n-3) and omega-6 (n-6) long chain (20-22 carbon; LC-) polyunsaturated fatty acids (PUFAs) and their metabolites play vital roles in innate immunity, energy homeostasis, brain development and cognitive function.[13–19] LC-PUFAs are critical signaling molecules for immunity and inflammation with most evidence showing that *n*-3 and *n*-6 LC-PUFAs and their metabolic products have different and often opposing effects.[20–24] Metabolites of the *n*-6 LC-PUFA arachidonic acid (ARA) typically act locally to promote inflammatory responses[25–27], while *n*-3 LC-PUFAs, such as eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA) and their metabolites, have anti-inflammatory and pro-resolution properties (meaning that they promote resolution of inflammation)[28,29]. In addition to their effects on inflammation, circulating levels of n-3 LC-PUFAs, including EPA and DHA, are inversely associated with fasting and postprandial serum TG concentrations, largely through attenuation of hepatic very-low-density lipoprotein (VLDL)-TG production.[30,31] Dietary supplementation with these n-3 LC-PUFAs has been shown consistently to reduce fasting circulating TG levels and improve lipid accumulation associated with NAFLD.[32,33]

The biosynthesis of n-3 and *n*-6 LC-PUFAs transpires via alternating desaturation (∆6, ∆5, and ∆4) and elongation enzymatic steps encoded by fatty acid desaturase (*FADS)* cluster genes (*FADS1* and *FADS2*), and fatty acid elongase genes (*ELOVL2* and *ELOVL5)*, and there is a limited capacity for biosynthesis through this pathway.[34–36] As a result, the primary dietary PUFAs that enter this pathway (linoleic acid [18:2n-6; LA], α-linolenic acid [18:3n-3; ALA], and their metabolic intermediates) compete as substrates for the desaturation and elongation steps. Additionally early studies with deuterated substrates indicated there is a saturation point where additional dietary quantities of 18 carbon dietary substrates had no effect on circulating LC-PUFA levels.[37] These studies also estimated that conversion of dietary ALA provided 75-85% of total n-3 LC-PUFAs needed to meet daily requirements.[37]

In 1961, a major effort was initiated to reduce levels of saturated fatty acids and replace them with PUFAs in an attempt to reduce circulating LDL-cholesterol and TGs.[38–40] This in turn led to a dramatic increase in eighteen carbon (18C-) PUFA-containing vegetable oils such as soybean, corn, and canola oils that contain high levels of n-6 LA relative n-3 ALA. It has been estimated that dietary LA increased from 2.79% to 7.21% of energy, whereas there was only a modest elevation in ingested ALA (from 0.39% to 0.72%), resulting in a ~15:1 ratio of LA to ALA entering the LC-PUFA biosynthetic pathway and an estimated 40% reduction in total circulating n-3 LC-PUFA levels.[41] Since LA and ALA compete for the same desaturation and elongation steps and there is a limited capacity for n-6 and n-3 LC-PUFA biosynthesis through the pathway, several human and animal studies suggested that the dramatic shift in quantities and ratios of dietary LA and ALA could lead to imbalances in n-6 to n-3 LC-PUFAs and, potentially, n-3 LC-PUFA deficiencies [42–46] Thus, as certain populations moved from traditional to modern Western diets (MWD), it was suggested excess LA would lead to 'Omega-3 Deficiency Syndrome'.[47]

The rate limiting step of LC-PUFA biosynthesis has long been recognized to be the *FADS*-encoded ∆6 and ∆5 desaturation steps. Over the past decade, GWAS and candidate gene studies have shown that variation in the *FADS* gene locus on human chromosome 11 is strongly associated with plasma levels of ARA and EPA and the efficiency by which LC-PUFA precursors (18C dietary PUFAs) are metabolized to n-6 and n-3 LC-PUFAs.[48,49] *FADS* cluster genetic variation is associated with numerous molecular phenotypes that impact human disease as well as the risk of several diseases, including coronary heart disease[50], diabetes[51–53] and colorectal cancer[54]. *FADS* cluster genetic variation is strongly associated with circulating TG and VLDL concentrations in young healthy Mexicans.[55]

Our previous studies revealed that African (compared to European) ancestry populations had elevated levels of LC-PUFAs, an increased frequency of the associated *FADS* genetic variants and a more efficient LC-PUFA biosynthesis (termed the derived haplotype).[56] In contrast, *FADS* variants associated with more limited capacity to synthesize LC-PUFAs (termed ancestral haplotype) are nearly fixed in Native American and Greenland Inuit populations and found at high frequencies in Amerind (AI) Ancestry Hispanic populations.[56] These distinct patterns of haplotypes have resulted in part from positive selection for the ancestral haplotype among Indigenous American populations.[56]

While the role of *FADS* variation in modulating circulating fatty acid levels has been documented previously[48,49], prior studies have not examined the impact that population differences in *FADS* allele frequencies have in downstream population-specific risk of fatty acid deficiency, The hypothesis tested in this paper is that ancestral *FADS* variation in the context of MWD is associated with low (perhaps inadequate) circulating levels of LC-PUFAs (particularly n-3 LC-PUFAs) in a large proportion of high AI-Ancestry Hispanic populations compared to other Hispanic populations, with downstream effects on numerous cardiometabolic and inflammatory risk factors. To address this question, we first examined the relationship between the genomic proportions of AI ancestry and circulating phospholipid LC-PUFA levels in self-reported Hispanic individuals from the Multi-Ethnic Study of Atherosclerosis (MESA)[57], which includes Hispanic groups with varying levels of AI ancestry. Second, we assessed the extent to which this relationship is explained by genetic variation within the *FADS1/2* locus, and also examined the impact of *FADS* genetic variation on cardiometabolic and inflammatory risk factors (lipids, anthropometric and inflammatory markers). Third, we tested whether these *FADS* genetic associations replicated in two high AI-Ancestry Hispanic cohorts, the Arizona Insulin Resistance (AIR) Registry[58], and the Hispanic Community Health Study/Study of Latinos (HCHS/SOL)[59,60].

2.3 Results

2.3.1 Participant Characteristics

The MESA participants[57,61] included in this analysis comprised 1,102 unrelated individuals aged 45 to 84 years at baseline of self-reported Hispanic race/ethnicity with country-specific classification based on the birthplace of parents and grandparents corresponding to Central American (n=80), Cuban (n=45), Dominican (n=145), Mexican (n=572), Puerto Rican (n=167) and South American (n=93) **(Table 2.1).** MESA Hispanic participants were recruited primarily from three field centers in the United States (Columbia University, University of California – Los Angeles (UCLA) and the University of Minnesota). The global proportions of AI, African, and European genetic ancestry in each individual were estimated using genome-wide SNP data (**Table 2.1**). Higher frequencies of the rs174537 T allele in the *FADS* cluster (corresponding to the ancestral allele) were observed in subjects with country/region-specific origins in Central America (0.59), South America (0.56) and Mexico (0.59) compared to those of Dominican (0.27), Cuban (0.28) or Puerto Rican origin (0.40) (**Table 2.1**).

2.3.2 LC-PUFA levels are associated with Amerind genetic ancestry

Higher proportions of AI genomic ancestry were associated with lower levels of LC-PUFAs in MESA Hispanics participants. **Figure 2.1 (panels a, c and e)** shows levels of EPA, DHA and ARA (expressed as the percentage of total fatty acids here and throughout the entire manuscript) as a function of inferred AI ancestry. Overall, AI ancestry explained 12.32%, 12.30% and 12.48% of total variation in EPA, DHA and ARA, respectively. Each 10% increase in AI ancestry was associated with a decrease of EPA

(0.049), DHA (0.185) and ARA (0.401) in phospholipids**.** Between subjects with the lowest and highest proportions of AI ancestry, the n-3 LC-PUFAs decreased by 60.6% (for EPA) and 46.8% (for DHA) and the n-6 LC-PUFAs decreased by 30.7% (for ARA). Consequently, the nadir in predicted fatty acids levels in plasma phospholipids among those with 100% AI ancestry was ~0.3 and ~2 for EPA and DHA, respectively, compared to ~8.6 for the n-6 LC-PUFA, ARA.

Given the prior evidence that key genetic determinants of LC-PUFAs mapping to the *FADS* locus show strong variation in frequency between populations, we sought to determine the role of *FADS* variation in the relationships between LC-PUFA levels and global AI ancestry. LC-PUFAs were adjusted for rs174537 genotype (**Figure 2.1; panels b, d and f);** rs174537 is selected as a representative proxy SNP for the well documented associations between the *FADS* locus and LC-PUFAs[62,63]. The rs174537 SNP has a strong effect on the ancestry-related decline in all LC-PUFAs. After adjusting for rs174537 genotype, an inverse association remains between global proportion of AI ancestry and EPA (β = -0.30, 95% Confidence Interval [Ci] = [-0.39, -0.22], *P* = 9.05 x 10^{-12} calculated using a two-sided t-test for the regression coefficient derived with n = 1102), DHA (β = -1.42, 95% CI = [-1.76, -1.08], *P* = 6.76 x 10^{-16} calculated using a two-sided t-test for the regression coefficient derived with n = 1102) and ARA (β = -0.99, 95% CI = [-1.59, -0.38], *P* = 0.0015 calculated using a two-sided t-test for the regression coefficient derived with n = 1102)**.** Regression analysis of n-3 and n-6 LC-PUFAs with global proportion of AI ancestry, accounting for covariates: age, sex and fish intake (Model 1), resulted in inverse relationships between the global proportion of AI ancestry with EPA (β = -0.48, *P* = 3.7 x 10^{-23} calculated by a Z-test from inverse variance weighted meta-analysis with a total of n = 1057), DPA (β = -0.18, *P* = 7.6 x 10^{-6} based on a Z-test from inverse variance weighted meta-analysis with a total of n = 1057), DHA (β = -0.63, *P* = 0.0007 based on a Z-test from inverse variance weighted meta-analysis with a total of n = 1057) and ARA (β = -4.06, *P* = 1.3 x 10^{-16} based on a Z-test from inverse variance weighted meta-analysis with a total of n = 1057) **(Supplementary Table 2.1).** These effects were consistent across study sites in MESA, with the largest effects observed at the University of Minnesota field center (**Supplementary Table 2.1**). Accounting for rs174537 genotype (Model 2), there remained an inverse association between the global proportion of AI ancestry with EPA (β = -0.28, *P* = 3.7 x 10^{-08} based on a Z-test from inverse variance weighted meta-analysis with a total of n = 1057) (**Supplementary Table 2.1**), while the relationship of global proportion of AI ancestry with ARA, DPA and DHA was no longer significant (**Supplementary Table 2.1**). In a model further accounting for local AI ancestry in addition to rs174537 (Model 3), EPA continued to be inversely associated with global proportion of AI ancestry (β = -0.34, *P* = 8.4 x 10^{-07} based on a Z-test from inverse variance weighted meta-analysis with a total of n = 1057), while the associations with DPA, DHA and ARA were not statistically significant (**Supplementary Table 2.1**). In additional analysis examining global and local ancestry as potential modifiers of the effect of rs174537 on circulating fatty acid levels, we did not observe statistically significant evidence of interaction **(Supplementary Table 2.2).**

As other studies have suggested different specific variants as potentially functional within the *FADS* region, we further repeated the analysis presented in **Figure 2.1** through sensitivity analysis focused on the *FADS* region variant rs174557 **(Supplementary Figure 2.1)**, a common variant that diminishes binding of *PATZ1*, a transcription factor conferring allele-specific downregulation of *FADS1*.[64] After adjusting for rs174557 genotype, we observed association between global proportion of AI ancestry and LC-PUFA levels (EPA: β = -0.30, *P* = 2.19 x 10^{-11}; DHA: β = -1.39, *P* = 2.54 x 10^{-15}; ARA: β = -1.02, *P* = 0.0012 calculated using a two-sided t-test for the regression coefficient derived with n = 1102) similar to that seen after adjusting for rs174537.

2.3.3 Association of global Amerind ancestry with triglycerides

Higher global proportions of AI ancestry were significantly associated with higher levels of circulating triglycerides (TG) in MESA Hispanic participants (β = 65.40 mg/dL, 95% CI = [42.28, 88.52] *P* = 3.58 x 10^{-8} based on a two-sided t-test for the regression coefficient derived with n = 1101) **(Figure 2.2a).** This relationship was attenuated after adjusting for rs174537 (**Figure 2.2b**), although there remained a significant relationship between global proportions of AI ancestry and TG levels (β = 39.47 mg/dL, 95% CI = ([2.62, 52.05], *P* = 8.16 x 10^{-04} based on a two-sided t-test for the regression coefficient derived with n = 1101). In sensitivity analysis, circulating triglycerides (TG) were adjusted for the variant rs174557. The relationship between global proportion of AI ancestry and TG levels (**Supplementary Figure 2.2**; β = 38.93 mg/dL, *P* = 9.59 x 10^{-04} based on a two-sided t-test for the regression coefficient derived with n = 1101) is similar with the association adjusting for rs174537**.** Examining the unadjusted relationship between triglyceride levels and rs174537 genotype, we observed mean triglyceride levels increased with the number of copies of the rs174537 effect allele T (**Figure 2.3a**). In analysis that incorporated adjustment for age and sex, the rs174537 T allele was significantly associated with higher levels of TG (GT vs GG: β = 21.27 mg/dL, 95% CI = [10.29, 32.25], P = 0.0001, TT vs GG: β = 29.94 mg/dL, 95% CI = [17.98, 41.88], P = 1.01 x 10^{-6} based on a two-sided t-test for the regression coefficient derived with n = 1101) **(Table 2.2 and Figure 2.3c)**.

2.3.4 Association of PUFAs with *FADS* cluster SNPs

We performed genetic association analysis adjusting for rs174537 genotype to determine if there was any residual association in the *FADS* region for the MESA Hispanic participants. In each of the Hispanic subgroups, after accounting for the rs174537 SNP, no additional genetic variants in the region were associated with EPA, DPA, DHA or ARA (**Supplementary Figure 2.3**). The rs174537 SNP is in strong linkage disequilibrium with other *FADS* cluster SNPs; thus, subsequent analyses are focused solely on the rs174537 SNP.

2.3.5 Effects of rs174537 on Inflammatory Biomarkers, Fasting Lipids and Anthropometrics

The effect of the *FADS* cluster SNP rs174537 on height, weight, body mass index (BMI), waist-hip ratio, s-ICAM, E-Selectin and HDL-C was estimated in the MESA Hispanic participants. Initially, we examined unadjusted relationships which showed, for example, that mean E-selectin levels increased with the number of copies of the rs174537 effect allele T (**Figure 2.3b**). In a model adjusted for age and sex, the rs174537 T allele was significantly associated with lower levels of HDL-C, higher waist-hip, lower height and weight, and higher levels of the inflammatory markers E-Selectin and s-ICAM (**Table 2.2,**

Figure 2.3d and Supplementary Table 2.3). In regression analysis with adjustment for principal components of ancestry, the rs174537 T allele remained significantly associated with higher TGs and lower height, while the associations with weight, waist-hip ratio, s-ICAM, E-Selectin and HDL-C were no longer statistically significant (**Supplementary Figure 2.4 and Supplementary Table 2.3**). In sensitivity analysis, we also examined the effect of rs174557 on the same set of phenotypes as examined for rs174537. Similar to the rs174537 T allele, the rs174557 A allele was significantly associated with lower levels of HDL-C, higher waist-hip, lower height and weight, and higher levels of the inflammatory markers s-ICAM **(Supplementary Table 2.4 and Supplementary Figure 2.4).** In regression analysis with adjustment for principal components of ancestry, the rs174557 A allele remained significantly associated with higher TGs and lower height, while the associations with weight, waist-hip ratio, s-ICAM, E-Selectin and HDL-C were no longer statistically significant **(Supplementary Table 2.4)**.

2.3.6 Replication in the AIR registry and HCHS/SOL cohort

We conducted analyses in the AIR registry (n = 497) and HCHS/SOL (n = 12,333) cohorts to examine the genotypic effect of rs174537 on multiple phenotypic traits including TGs and waist-to-hip ratio (**Supplementary Tables 2.5-2.6**). In regression analyses adjusted for age and sex (and inclusion of random effects for household block and unit sharing in HCHS/SOL), the rs174537 T allele was significantly associated with TGs (AIR: β = 10.4 mg/dL*, P* = 0.03*,* HCHS/SOL: $\beta =$ 8.75 mg/dL*, P* = 5.84×10^{-25} based on a two-sided t-test for the regression coefficient derived with n = 12,333) (**Supplementary Table 2.7**). The rs174537 T allele was also significantly associated with reduced height (β = -1.33, *P* = 4.47 x 10^{-56} calculated using a two-sided t-test for the regression coefficient derived with n = 12,333) and weight (β = -1.25, *P* = 2.61 x 10^{-08} calculated using a two-sided t-test for the regression coefficient derived with n = 12,333), and increased waist-to-hip ratio (β = 0.003; *P* = 2.77 x 10^{-05} calculated using a two-sided t-test for the regression coefficient derived with n = 12,333) in the HCHS/SOL cohort. The direction of effect was consistent, but not statistically significant, in the much smaller AIR cohort (**Supplementary Table 2.7**). The association of rs174537 with TGs remained statistically significant after adjustment for principal components of ancestry (β = 4.05 mg/dL, *P* = 1.26 x 10^{-05} calculated using a two-sided t-test for the regression coefficient derived with n = 497) and the effects were consistent across the HCHS/SOL study sites (**Supplementary Table 2.8**). We did not replicate these findings in the smaller AIR registry (**Supplementary Table 2.7**). S-ICAM and E-Selectin were not measured in either AIR or HCHS/SOL and thus could not be evaluated for replication of the findings from MESA.

2.4 Discussion

While prior studies have identified genetic variants within the *FADS* locus with strong impact on fatty acid levels[48,49], prior literature has not examined directly the impact of population differences in allele frequencies on population-specific risk of fatty acid deficiency. In light of dramatic differences in genetic variation within the *FADS* locus across worldwide populations[56] and the marked changes in dietary n-6 and n-3 PUFA levels and ratios over the past 75 years, we carried out a study of to examine genomic proportion of AI ancestry as a predictor of n-3 and n-6 LC-PUFA levels and related cardiometabolic and inflammatory risk in the Hispanic participants from MESA. Our study first illustrates that certain Hispanic populations and particularly high AI-Ancestry populations have high frequencies of the ancestral allele at T at rs174537. Importantly, the frequency of the TT genotype associated with limited LC-PUFA biosynthesis ranges from <1% in African-Ancestry populations including African Americans to 40-55% in high AI-Ancestry Hispanics, and ~11% in European-Ancestry populations.[65] In light of high ancestral frequencies in certain Hispanic populations together with elevated dietary n-6 (LA) to n-3 (ALA) PUFAs ratios (>10:1) from the MWD entering the pathway, we postulated that these populations would be most likely to saturate their capacity to synthesize LC-PUFAs and particularly n-3 LC-PUFAs. Our statistical analyses demonstrated that global proportion of AI ancestry is predictive of reduced LC-PUFA phospholipid levels in the Hispanic population of the United States, accounting for ~12% of total variation in EPA, DHA and ARA. Further, we showed that this relationship can be explained in large part by genetic variation within the *FADS* cluster. Given that many Hispanic individuals will have reasonable knowledge of their AI ancestry, our work suggests a practical way to identify individuals likely to carry the homozygous TT genotypes, and for whom follow-up *FADS* genotyping assays may be warranted.

While both n-6 and n-3 LC-PUFAs are impacted, relatively high levels of ARA (~8.6% of total fatty acids) remain in circulating phospholipids in even the highest AI-Ancestry populations. In contrast, n-3 LC-PUFAs including EPA and DHA are reduced to the low (perhaps inadequate) levels of ~0.3% [EPA] and ~2% [DHA] of total fatty acids in circulating phospholipids in high AI-Ancestry individuals. It is not possible to say with certainty what levels of EPA and DHA or ratio of EPA + DHA/ ARA would be inadequate (deficient) and have pathophysiologic impact, but these are certainly quantitatively very low concentrations and ratios of n-3 LC-PUFAs. It has been recognized that high levels of dietary LA relative to ALA from the modern Western diets (MWD) entering the LC-PUFA biosynthetic pathway are reciprocally related to levels of n-3 LC-PUFAs due to substate saturation of the enzymatic pathway.[66,67] Such a scenario was proposed by both Okuyama and colleagues and Lands and colleagues three decades ago to give rise to Omega-3 Deficiency Syndrome and chronic pathophysiological events.[20,47,68] We propose that a limited LC-PUFA synthetic capacity in a greater proportion of AI-Ancestry Hispanics (due to the ancestral haplotype) in the context of excess dietary LA levels and high LA/ALA ratios renders inadequate n-3 LC-PUFAs more likely in this population.

Our study also suggests that *FADS* variation has large effects on some critical cardiometabolic and inflammatory risk factors. Specifically, the proportion of AI ancestry was positively related to levels of circulating TGs and much of this effect was explained by variation in the *FADS* locus. While other studies have found associations between numerous genetic loci including *FADS* SNPs and circulating TGs[69–78], the high frequency of the ancestral *FADS* alleles (associated with elevated TGs) and their effect size in AI-Ancestry Hispanic populations that suggest that *FADS* variation is particularly relevant to TG levels in this population. The presence of the T allele at rs174537 had a large effect on circulating TG (GT vs GG: β = 21.27 mg/dL, *P* = 0.0002, TT vs GG: β = 29.94 mg/dL, *P* = 1.01 x 10^{-6}) and this genotypic effect was replicated in both the AIR registry and HCHS/SOL cohort. Circulating TG are primarily synthesized in the liver and deficiencies of n-3 LC-PUFAs and imbalances of n-6 relative to n-3 PUFAs have been associated with elevated TGs and NAFLD.[79] Elevating n-3 LC-PUFA by diet or supplementation reduces TG by promoting hepatic fatty acid oxidation and reducing synthesis (via reducing *de novo* lipogenesis and decreasing fatty acid and adipokine release from adipocytes).[80–82] These current data suggest that inadequate levels of n-3 LC-PUFAs in AI-Ancestry Hispanic populations may impact TG formation in the liver resulting in higher levels of circulating TG and potentially NAFLD.

Waist-to-hip ratio, used to describe the distribution of body fat, has been shown to be closely associated with hypertension, diabetes, dyslipidemia and cardiovascular disease.[83] A previous study examined genetic loci associated with BMI and waist-to-hip ratio and found nine BMI and seven central adiposity loci in Hispanic women.[84] To date, variation within *FADS* has not been associated with waist-to-hip ratio. While our study demonstrated that the ancestral rs174537 T allele was strongly associated with a higher waist-to-hip ratio and this risk factor was replicated in HCHS/SOL, the relationship was not statistically significant after adjusting for principal components of ancestry. Thus, waist-hip-ratio is an example of a trait for which the association with AI-Ancestry is not explained in large part by *FADS* variation.

The rs174537 allele T further demonstrated association with reduced height and weight in the large HCHS/SOL cohort (n = 12,333). Fumagalli and colleagues examined indigenous Greenland Inuit and found strong signals of natural selection within the *FADS* cluster.[85] The identified *FADS* variants were also strongly associated with anthropometric traits including body weight and height in the Inuit, and those associations were replicated in Europeans.

A wide variety of biomarkers of inflammation were measured in MESA, and there was a strong association between rs174537 and E-selectin which maintained suggestive evidence of association even after adjustment for population structure using genetic principal components of ancestry. E-Selectin (CD-62E) plays a pivotal role in the activation and adhesion of the migrating leukocytes to the endothelium.[86] These membrane bound adhesion molecules also undergo proteolytic cleavage that generate soluble forms that can be measured in the blood.[87] Serum levels of E-Selectin increase in many pathologies involving chronic inflammation including obesity[88], cardiovascular disease[89], bronchial asthma[90] and cancer[91,92].

Limitations of the study include a focus on primarily urban Hispanic American populations represented by the MESA cohort, potential confounding by diet and lifestyle habits across the six Hispanic subgroups in MESA, and systematic differences in PUFA levels across MESA study sites. To address the observable variation across Hispanic subgroups and study site, we included additional analyses stratified by these factors and demonstrated that our results were consistent across strata. Additionally, we used food frequency questionnaire data to confirm participants included in our analyses did not have self-reported use of fish oil supplements, and we performed analyses adjusted for self-reported fish intake in MESA. Still, we recognize there are inherent limitations with the quality of self-report-based measures of diet and supplement use. Further, we did not consider additional measures of dietary intake of n-3 and n-6 PUFAs in our regression analyses, in part because we determined that we did not have reliable measures available for these parameters in the MESA participants. Therefore, future studies should examine further the impact of dietary differences on the relationship between AI ancestry, *FADS* variation, and LC-PUFA levels.

Despite these limitations, our study reveals that *FADS* variation in AI-Ancestry Hispanic populations is inversely associated with dyslipidemia and inflammation, risk factors for a wide range of pathologies including cardiovascular and metabolic diseases. These associations are observed strongly in these Hispanic populations in part because of the high frequencies of ancestral *FADS* alleles. It may be that LC-PUFAs or their metabolites (eicosanoids, docosanoids, resolvins, protectins, etc.) are responsible for these genetic effects given the direct relationship between *FADS* variation and LC-PUFA levels. Alternatively, we have recently combined genetic and metabolomic analyses to identify the *FADS* locus as a central control point for biologically-active LC-PUFA-containing complex lipids that act as signaling molecules such as the endocannabinoid, 2-AG, and such endocannabinoids are known to impact anthropomorphic and other phenotypic characteristics.[93]

Our results also suggest that targeting recommendations for n-3 and n-6 LC-PUFA intake/supplementation within AI-Ancestry Hispanic populations may be particularly effective. This premise is supported by the fact that numerous mechanistic studies directly link low levels of n-3 LC-PUFAs and high n-6 to n-3 ratios to elevated tissue and circulating TGs and NAFLD, and several recent reviews and meta-analyses suggest that n-3 LC-PUFA supplementation improves circulating and tissue levels of TG and NAFLD.[94,95] Prior research demonstrates that mean proportions of Amerind ancestry vary greatly by self-identified regions of origin among Hispanic Americans, with Mexican, Central American and South American Hispanics showing the greatest proportions, and individuals identifying as Cuban, Dominican and Puerto Rican showing considerably lower proportions.[61,96] While a long term goal of applying precision nutrition may include genotyping of rs174537 (or related *FADS* region variants)_in routine health care screening, current health care practice does not provide adequate resources to genotype most individuals. Thus, a priori information predictive of ancestry such as country or origin or otherwise, may serve as a preliminary tool to prioritize those who are most likely to have low circulating and tissue levels of n-3 LC-PUFA and would benefit from additional screening either through genotyping or screening for n-3 LC-PUFA deficiency. Despite the current limitations of precision nutrition including inadequate genetic testing, the translational implications of this work are to point out that a large proportion of AI-Ancestry Hispanic populations have low (perhaps deficient) levels of n-3 LC-PUFAs and increased related risk factors. Thus, because of *FADS*-related deficiencies, these populations may be particularly responsive to diets or supplements enriched in n-3 LC-PUFAs.

2.5 Methods

2.5.1 Study participants

MESA is a longitudinal cohort study of subclinical cardiovascular disease and risk factors that predict progression to clinically overt cardiovascular disease or progression of subclinical disease [57]. Between 2000 and 2002, MESA recruited 6,814 men and women 45 to 84 years of age from Forsyth County, North Carolina; New York City; Baltimore; St. Paul, Minnesota; Chicago; and Los Angeles. Participants at baseline were 38% White, 28% African American, 22% Hispanic and 12% Asian (primarily Chinese) ancestry. This manuscript focuses on Hispanic American participants from MESA. Among the MESA Hispanic participants, self-reported birthplaces for parents' and grandparents' country/region of origin were used to assign country/region of origin to the following categories Central America, Cuba, the Dominican Republic, Mexico, Puerto Rico and South American origin were assigned for the MESA Hispanic participants.

2.5.2 Fatty Acid measurements

The fatty acids were measured by gas chromatography in EDTA plasma frozen at –70°C.[97]

Lipids were extracted from the plasma using a chloroform/methanol extraction method and the cholesterol esters, triglyceride, phospholipids and free fatty acids are separated by thin layer chromatography. The fatty acid methyl esters were obtained from the phospholipids and were detected by gas chromatography flame ionization. Individual fatty acids were expressed as a percent of total fatty acids. A total of 28 fatty acids were identified. Here, we focus on the following n-3 and n-6 fatty acids: eicosapentaenoic acid

(EPA), docosapentaenoic acid (DPA), docosahexaenoic acid (DHA), and arachidonic acid (ARA).

2.5.3 Additional phenotypes in MESA

We considered additional phenotypes in analysis of the MESA data including lipids (HDL-C and triglycerides), anthropometric (height, weight, waist-hip ratio), and inflammatory markers (soluble E-Selectin and soluble ICAM-1). Details of measurement and treatment of outliers are provided in the **Supplementary Methods** and **Supplementary Figures 5-6**.

2.5.4 Genotyping, genetic association and ancestry analysis

Participants in the MESA cohort who consented to genetic analyses and data sharing (dbGaP) were genotyped using the Affymetrix Human SNP Array 6.0 (GWAS array) as part of the NHLBI SHARe (SNP Health Association Resource) project. Genotype quality control for these data included filter on SNP level call rate < 95%, individual level call rate < 95%, heterozygosity > 53%.[98] The cleaned genotypic data was deposited with MESA phenotypic data into dbGaP (study accession phs000209.v13.p3); 8,224 consenting individuals (2,685 White, 2,588 non-Hispanic African-American, 2,174 Hispanic, 777 Chinese) were included, with 897,981 SNPs passing study specific quality control (QC). SNP coverage from the original GWAS SNP genotyping array was increased through imputation using the 1,000 Genomes Phase 3 integrated variant set completed using the Michigan Imputation Server (https://imputationserver.sph.umich.edu).

Prior studies have highlighted multiple different *FADS* variants for their role in regulation of fatty acid synthesis, including rs174537[48] and rs174557[64]; however, the relevant variants at the primary signal within the *FADS* region exhibit extended linkage disequilibrium across the region.[99] Therefore, we focused our genetic analyses primarily on the variant rs174537, with additional sensitivity analyses using similar models for the variant rs174557. Imputed genotype data were used for genetic association analysis of the rs174537 and rs174557 SNPs (for which the imputation R-squared in MESA Hispanics were both 0.99). Our statistical analyses used genotype dosage information from imputation, except where noted otherwise. For analyses that required us to stratify by genotype, including those presented in **Table 2** and **Figure 3**, we used the estimated most likely genotype from imputation. Principal components of ancestry were computed using genome-wide genotype data.[98] Global proportions of Amerind ancestry were estimated in MESA participants by leveraging reference samples from the 1000 Genomes[100] and the Human Genome Diversity Project (HGDP)[101,102]. Local ancestry for each individual was defined as the genetic ancestry at the position of *FADS* SNP rs174537, where each individual can have 0, 1 or 2 copies of an allele derived from each of the three possible ancestral populations (European, African and Amerind). Local ancestry, was estimated using the RFMix package.[103] Details are provided in the **Supplementary Methods**.

2.5.5 Regression modeling of n-3 and n-6 PUFAs

As we observed a strong effect of study site in regression analysis of all LC-PUFAs (**Supplementary Table 9**), we performed regression analyses stratified by study site and combined by inverse-variance weighted meta-analysis. In order to examine the effect of global Amerind ancestry on the levels of n-3 and n-6 PUFAs in the MESA Hispanic participants, we carried out linear regression analyses using three different models for each of the PUFA levels as follows:

1) PUFA ~ age + sex + fish intake + global proportion of Amerind ancestry,

2) PUFA ~ age + sex + fish intake + global proportion of Amerind ancestry + rs174537 genotype, and

3) PUFA ~ age + sex + fish intake + global proportion of Amerind ancestry + rs174537 genotype + *FADS* region local proportion of Amerind ancestry.

2.5.6 Regression modeling for genotypic effects of *FADS* cluster SNP rs174537 on proximal traits

To examine the effect of *FADS* SNP rs174537 on lipids (HDL-C and triglycerides), anthropometric (height, weight, and waist-to-hip ratio), and inflammatory markers (s-ICAM and E-Selectin) in MESA Hispanic participants, we performed linear regression analysis with covariate adjustment for (1) age and sex, and (2) age, sex and the first four principal components of ancestry.

2.5.7 Replication analysis in the AIR registry and HCHS/SOL cohort

We conducted follow-up regression analyses to examine the association of rs174537 with phenotypic traits in both the AIR registry and the HCHS/SOL cohort. The variant rs174537 was genotyped directly in both AIR and HCHS//SOL. Details are provided in the **Supplementary Methods**.

2.5.8 Statistics and Reproducibility

All of our statistical analyses were carried out on biologically independent samples from MESA (n=1,102), AIR (n=497) and HCHS/SOL (n=12,333). Analyses in MESA were carried out for an unrelated subset of participants constructed by retaining at most one individual from each group of first-degree relatives. We did not perform relationship inference and removal of first -degree relatives in AIR, as there were no genome-wide data available to infer relationship among individuals. In HCHS/SOL, all individuals (both related and unrelated) were included in analyses, as we accounted for their relationships using linear mixed models. Regression analyses presented throughout the manuscript included adjustments for relevant covariates as stated for each model presented in the text.

2.5.9 Ethical review

All MESA participants provided written informed consent for participation at their respective MESA study sites, and the MESA study was also reviewed and approved by the Institutional Review Boards (IRBs) at each of the participating study sites. The current investigation including activities for analysis of LC-PUFA levels in MESA was reviewed and approved by the Institutional Review Board (IRB) at the University of Virginia. The AIR registry was approved by the IRB at the University of Arizona and all subjects gave written informed consent before their participation. The HCHS/SOL was approved by the IRBs at all participating institutions including the Albert Einstein College of Medicine, and all participants gave written informed consent.

2.6 Data availability

Genome-wide genotype data for the Multi-Ethnic Study of Atheroslerosis (MESA)[57,61,98] and the Hispanic Community Health Study / Study of Latinos (HCHS/SOL)[60,96,104] are available by application through dbGaP. The dbGaP accession numbers are: MESA phs000209 and HCHS/SOL phs000810. All other data are available from the corresponding author (or other sources, as applicable) on reasonable request.

2.7 Competing Interests

Floyd H. Chilton is a co-founder of Tyrian Omega Inc. All other authors declare no competing interests.

2.8 References

1. Sankar, P. *et al.* Genetic research and health disparities. *JAMA* **291**, 2985–2989 (2004).
2. Sankar, P., Cho, M. K. & Mountain, J. Race and ethnicity in genetic research. *Am. J. Med. Genet. A.* **143A**, 961–970 (2007).
3. Mensah, G. A., Mokdad, A. H., Ford, E. S., Greenlund, K. J. & Croft, J. B. State of disparities in cardiovascular health in the United States. *Circulation* **111**, 1233–1241 (2005).
4. CDC. Obesity is a Common, Serious, and Costly Disease. *Centers for Disease Control and Prevention* https://www.cdc.gov/obesity/data/adult.html (2020).
5. Campos, C. L. & Rodriguez, C. J. High blood pressure in Hispanics in the United States: a review. *Curr. Opin. Cardiol.* **34**, 350–358 (2019).
6. Carroll, M., Kit, B. & Lacher, D. Trends in elevated triglyceride in adults: United States, 2001-2012. *NCHS Data Brief* 198 (2015).
7. Saab, S., Manne, V., Nieto, J., Schwimmer, J. B. & Chalasani, N. P. Nonalcoholic Fatty Liver Disease in Latinos. *Clin. Gastroenterol. Hepatol. Off. Clin. Pract. J. Am. Gastroenterol. Assoc.* **14**, 5–12; quiz e9-10 (2016).
8. Williams, C. D. *et al.* Prevalence of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis among a largely middle-aged population utilizing ultrasound and liver biopsy: a prospective study. *Gastroenterology* **140**, 124–131 (2011).
9. Brown, L. A. *et al.* Admixture Mapping Identifies an Amerindian Ancestry Locus Associated with Albuminuria in Hispanics in the United States. *J. Am. Soc. Nephrol.* **28**, 2211–2220 (2017).

10. Pereira, F. dos S. C. F. *et al.* A systematic literature review on the European, African and Amerindian genetic ancestry components on Brazilian health outcomes. *Sci. Rep.* **9**, 8874 (2019).

11. Alarcón-Riquelme, M. E. *et al.* Genome-Wide Association Study in an Amerindian Ancestry Population Reveals Novel Systemic Lupus Erythematosus Risk Loci and the Role of European Admixture. *Arthritis Rheumatol. Hoboken NJ* **68**, 932–943 (2016).

12. Fleischman, M. W., Budoff, M., Zeb, I., Li, D. & Foster, T. NAFLD prevalence differs among hispanic subgroups: the Multi-Ethnic Study of Atherosclerosis. *World J. Gastroenterol.* **20**, 4987–4993 (2014).

13. Spector, A. A. Plasma free fatty acid and lipoproteins as sources of polyunsaturated fatty acid for the brain. *J. Mol. Neurosci. MN* **16**, 159–165; discussion 215-221 (2001).

14. Lands, W. E. & Hart, P. Metabolism of Glycerolipids. VI. Specificities of acyl coenzyme A: phospholipid acyltransferases. *J. Biol. Chem.* **240**, 1905–1911 (1965).

15. Gibson, R. A., Muhlhausler, B. & Makrides, M. Conversion of linoleic acid and alpha-linolenic acid to long-chain polyunsaturated fatty acids (LCPUFAs), with a focus on pregnancy, lactation and the first 2 years of life. *Matern. Child. Nutr.* **7 Suppl 2**, 17–26 (2011).

16. McNamara, R. K. & Carlson, S. E. Role of omega-3 fatty acids in brain development and function: potential implications for the pathogenesis and prevention of psychopathology. *Prostaglandins Leukot. Essent. Fatty Acids* **75**, 329–349 (2006).

17. McCann, J. C. & Ames, B. N. Is docosahexaenoic acid, an n-3 long-chain polyunsaturated fatty acid, required for development of normal brain function? An overview of evidence from cognitive and behavioral tests in humans and animals. *Am. J. Clin. Nutr.* **82**, 281–295 (2005).

18. Weiser, M. J., Butt, C. M. & Mohajeri, M. H. Docosahexaenoic Acid and Cognition throughout the Lifespan. *Nutrients* **8**, 99 (2016).

19. Hibbeln, J. R. *et al.* Maternal seafood consumption in pregnancy and neurodevelopmental outcomes in childhood (ALSPAC study): an observational cohort study. *Lancet Lond. Engl.* **369**, 578–585 (2007).

20. Lands, W. E. *et al.* Maintenance of lower proportions of (n - 6) eicosanoid precursors in phospholipids of human plasma in response to added dietary (n - 3) fatty acids. *Biochim. Biophys. Acta* **1180**, 147–162 (1992).

21. James, M. J., Gibson, R. A. & Cleland, L. G. Dietary polyunsaturated fatty acids and inflammatory mediator production. *Am. J. Clin. Nutr.* **71**, 343S–8S (2000).

22. Lands, B. A critique of paradoxes in current advice on dietary lipids. *Prog. Lipid Res.* **47**, 77–106 (2008).

23. Lands, B. Omega-3 PUFAs Lower the Propensity for Arachidonic Acid Cascade Overreactions. *BioMed Res. Int.* **2015**, 285135 (2015).

24. Schmitz, G. & Ecker, J. The opposing effects of n-3 and n-6 fatty acids. *Prog. Lipid Res.* **47**, 147–155 (2008).

25. Smith, W. L. The eicosanoids and their biochemical mechanisms of action. *Biochem. J.* **259**, 315–324 (1989).

26. Smith, W. L., DeWitt, D. L. & Garavito, R. M. Cyclooxygenases: structural, cellular, and molecular biology. *Annu. Rev. Biochem.* **69**, 145–182 (2000).

27. Haeggström, J. Z. & Funk, C. D. Lipoxygenase and leukotriene pathways: biochemistry, biology, and roles in disease. *Chem. Rev.* **111**, 5866–5898 (2011).

28. Serhan, C. N., Chiang, N. & Van Dyke, T. E. Resolving inflammation: dual anti-inflammatory and pro-resolution lipid mediators. *Nat. Rev. Immunol.* **8**, 349–361 (2008).

29. Serhan, C. N. Pro-resolving lipid mediators are leads for resolution physiology. *Nature* **510**, 92–101 (2014).

30. Oscarsson, J. & Hurt-Camejo, E. Omega-3 fatty acids eicosapentaenoic acid and docosahexaenoic acid and their mechanisms of action on apolipoprotein B-containing lipoproteins in humans: a review. *Lipids Health Dis.* **16**, 149 (2017).

31. Shearer, G. C., Savinova, O. V. & Harris, W. S. Fish oil — How does it reduce plasma triglycerides? *Biochim. Biophys. Acta BBA - Mol. Cell Biol. Lipids* **1821**, 843–851 (2012).

32. Scorletti, E. & Byrne, C. D. Omega-3 fatty acids and non-alcoholic fatty liver disease: Evidence of efficacy and mechanism of action. *Mol. Aspects Med.* **64**, 135–146 (2018).

33. Jump, D. B., Depner, C. M., Tripathy, S. & Lytle, K. A. Potential for dietary ω-3 fatty acids to prevent nonalcoholic fatty liver disease and reduce the risk of primary liver cancer. *Adv. Nutr. Bethesda Md* **6**, 694–702 (2015).

34. Park, H. G., Park, W. J., Kothapalli, K. S. D. & Brenna, J. T. The fatty acid desaturase 2 (FADS2) gene product catalyzes Δ4 desaturation to yield n-3

docosahexaenoic acid and n-6 docosapentaenoic acid in human cells. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **29**, 3911–3919 (2015).

35. Zhang, J. Y., Kothapalli, K. S. D. & Brenna, J. T. Desaturase and elongase-limiting endogenous long-chain polyunsaturated fatty acid biosynthesis. *Curr. Opin. Clin. Nutr. Metab. Care* **19**, 103–110 (2016).
36. Chilton, F. H. *et al.* Diet-gene interactions and PUFA metabolism: a potential contributor to health disparities and human diseases. *Nutrients* **6**, 1993–2022 (2014).
37. Emken, E. A., Adlof, R. O. & Gulley, R. M. Dietary linoleic acid influences desaturation and acylation of deuterium-labeled linoleic and linolenic acids in young adult males. *Biochim. Biophys. Acta* **1213**, 277–288 (1994).
38. Dietary fat and its relation to heart attacks and strokes. Report by the Central Committee for Medical and Community Program of the American Heart Association. *JAMA* **175**, 389–391 (1961).
39. Sacks, F. M. *et al.* Dietary Fats and Cardiovascular Disease: A Presidential Advisory From the American Heart Association. *Circulation* **136**, e1–e23 (2017).
40. Miller, M. *et al.* Triglycerides and cardiovascular disease: a scientific statement from the American Heart Association. *Circulation* **123**, 2292–2333 (2011).
41. Blasbalg, T. L., Hibbeln, J. R., Ramsden, C. E., Majchrzak, S. F. & Rawlings, R. R. Changes in consumption of omega-3 and omega-6 fatty acids in the United States during the 20th century. *Am. J. Clin. Nutr.* **93**, 950–962 (2011).
42. Lands, B. Dietary omega-3 and omega-6 fatty acids compete in producing tissue compositions and tissue responses. *Mil. Med.* **179**, 76–81 (2014).

43. Liou, Y. A., King, D. J., Zibrik, D. & Innis, S. M. Decreasing linoleic acid with constant alpha-linolenic acid in dietary fats increases (n-3) eicosapentaenoic acid in plasma phospholipids in healthy men. *J. Nutr.* **137**, 945–952 (2007).

44. Wood, K. E., Mantzioris, E., Gibson, R. A., Ramsden, C. E. & Muhlhausler, B. S. The effect of modifying dietary LA and ALA intakes on omega-3 long chain polyunsaturated fatty acid (n-3 LCPUFA) status in human adults: a systematic review and commentary. *Prostaglandins Leukot. Essent. Fatty Acids* **95**, 47–55 (2015).

45. MacIntosh, B. A. *et al.* Low-n-6 and low-n-6 plus high-n-3 diets for use in clinical research. *Br. J. Nutr.* **110**, 559–568 (2013).

46. Novak, E. M., Dyer, R. A. & Innis, S. M. High dietary omega-6 fatty acids contribute to reduced docosahexaenoic acid in the developing brain and inhibit secondary neurite growth. *Brain Res.* **1237**, 136–145 (2008).

47. Okuyama, H., Kobayashi, T. & Watanabe, S. Dietary fatty acids--the N-6/N-3 balance and chronic elderly diseases. Excess linoleic acid and relative N-3 deficiency syndrome seen in Japan. *Prog. Lipid Res.* **35**, 409–457 (1996).

48. Mathias, R. A. *et al.* The impact of FADS genetic variants on ω6 polyunsaturated fatty acid metabolism in African Americans. *BMC Genet.* **12**, 50 (2011).

49. Lemaitre, R. N. *et al.* Genetic loci associated with plasma phospholipid n-3 fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium. *PLoS Genet.* **7**, e1002193 (2011).

50. Lettre, G. *et al.* Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project. *PLoS Genet.* **7**, e1001300 (2011).

51. Brayner, B., Kaur, G., Keske, M. A. & Livingstone, K. M. FADS Polymorphism, Omega-3 Fatty Acids and Diabetes Risk: A Systematic Review. *Nutrients* **10**, (2018).

52. Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.* **42**, 105–116 (2010).

53. Ingelsson, E. *et al.* Detailed physiologic characterization reveals diverse mechanisms for novel genetic Loci regulating glucose and insulin metabolism in humans. *Diabetes* **59**, 1266–1275 (2010).

54. Zhang, B. *et al.* Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. *Nat. Genet.* **46**, 533–542 (2014).

55. Vazquez-Vidal, I. *et al.* Serum Lipid Concentrations and FADS Genetic Variants in Young Mexican College Students: The UP-AMIGOS Cohort Study. *Lifestyle Genomics* **11**, 40–48 (2018).

56. Harris, D. N. *et al.* Evolution of Hominin Polyunsaturated Fatty Acid Metabolism: From Africa to the New World. *Genome Biol. Evol.* **11**, 1417–1430 (2019).

57. Bild, D. E. *et al.* Multi-Ethnic Study of Atherosclerosis: objectives and design. *Am. J. Epidemiol.* **156**, 871–881 (2002).

58. Shaibi, G. Q., Coletta, D. K., Vital, V. & Mandarino, L. J. The design and conduct of a community-based registry and biorepository: a focus on cardiometabolic health in Latinos. *Clin. Transl. Sci.* **6**, 429–434 (2013).

59. Lavange, L. M. *et al.* Sample design and cohort selection in the Hispanic Community Health Study/Study of Latinos. *Ann. Epidemiol.* **20**, 642–649 (2010).

60. Sorlie, P. D. *et al.* Design and implementation of the Hispanic Community Health Study/Study of Latinos. *Ann. Epidemiol.* **20**, 629–641 (2010).

61. Manichaikul, A. *et al.* Population structure of Hispanics in the United States: the multi-ethnic study of atherosclerosis. *PLoS Genet.* **8**, e1002640 (2012).

62. Tanaka, T. *et al.* Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study. *PLoS Genet.* **5**, e1000338 (2009).

63. Hester, A. G. *et al.* Relationship between a common variant in the fatty acid desaturase (FADS) cluster and eicosanoid generation in humans. *J. Biol. Chem.* **289**, 22482–22489 (2014).

64. Pan, G. *et al.* PATZ1 down-regulates FADS1 by binding to rs174557 and is opposed by SP1/SREBP1c. *Nucleic Acids Res.* **45**, 2408–2422 (2017).

65. Blomquist, S. *et al.* Fatty Acid Desaturase Gene-Induced Omega-3 Deficiency in Amerindian-Ancestry Hispanic Populations. *FASEB J.* **34**, 1–1 (2020).

66. Mohrhauer, H. & Holman, R. T. EFFECT OF LINOLENIC ACID UPON THE METABOLISM OF LINOLEIC ACID. *J. Nutr.* **81**, 67–74 (1963).

67. Gibson, R. A., Neumann, M. A., Lien, E. L., Boyd, K. A. & Tu, W. C. Docosahexaenoic acid synthesis from alpha-linolenic acid is inhibited by diets high in polyunsaturated fatty acids. *Prostaglandins Leukot. Essent. Fatty Acids* **88**, 139–146 (2013).

68. Okuyama, H., Ichikawa, Y., Fujii, Y., Ito, M. & Yamada, K. Changes in dietary fatty acids and life style as major factors for rapidly increasing inflammatory diseases and elderly-onset diseases. *World Rev. Nutr. Diet.* **95**, 52–61 (2005).

69. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).

70. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).

71. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.* **41**, 56–65 (2009).

72. Waterworth, D. M. *et al.* Genetic variants influencing circulating lipid levels and risk of coronary artery disease. *Arterioscler. Thromb. Vasc. Biol.* **30**, 2264–2276 (2010).

73. Spracklen, C. N. *et al.* Association analyses of East Asian individuals and trans-ancestry analyses with European individuals reveal new loci associated with cholesterol and triglyceride levels. *Hum. Mol. Genet.* **26**, 1770–1784 (2017).

74. Bentley, A. R. *et al.* Multi-ancestry genome-wide gene-smoking interaction study of 387,272 individuals identifies new loci associated with serum lipids. *Nat. Genet.* **51**, 636–648 (2019).

75. de Vries, P. S. *et al.* Multiancestry Genome-Wide Association Study of Lipid Levels Incorporating Gene-Alcohol Interactions. *Am. J. Epidemiol.* **188**, 1033–1054 (2019).

76. Kulminski, A. M. *et al.* Strong impact of natural-selection-free heterogeneity in genetics of age-related phenotypes. *Aging* **10**, 492–514 (2018).

77. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).

78. Hoffmann, T. J. *et al.* A large electronic-health-record-based genome-wide study of serum lipids. *Nat. Genet.* **50**, 401–413 (2018).

79. Scorletti, E. & Byrne, C. D. Omega-3 fatty acids, hepatic lipid metabolism, and nonalcoholic fatty liver disease. *Annu. Rev. Nutr.* **33**, 231–248 (2013).

80. Botolin, D., Wang, Y., Christian, B. & Jump, D. B. Docosahexaneoic acid (22:6,n-3) regulates rat hepatocyte SREBP-1 nuclear abundance by Erk- and 26S proteasome-dependent pathways. *J. Lipid Res.* **47**, 181–192 (2006).

81. Lambert, J. E., Ramos-Roman, M. A., Browning, J. D. & Parks, E. J. Increased de novo lipogenesis is a distinct characteristic of individuals with nonalcoholic fatty liver disease. *Gastroenterology* **146**, 726–735 (2014).

82. Gnoni, A. & Giudetti, A. M. Dietary long-chain unsaturated fatty acids acutely and differently reduce the activities of lipogenic enzymes and of citrate carrier in rat liver. *J. Physiol. Biochem.* **72**, 485–494 (2016).

83. Huxley, R., Mendis, S., Zheleznyakov, E., Reddy, S. & Chan, J. Body mass index, waist circumference and waist:hip ratio as predictors of cardiovascular risk--a review of the literature. *Eur. J. Clin. Nutr.* **64**, 16–22 (2010).

84. Graff, M. *et al.* Generalization of adiposity genetic loci to US Hispanic women. *Nutr. Diabetes* **3**, e85 (2013).

85. Fumagalli, M. *et al.* Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* **349**, 1343–1347 (2015).

86. Muller, W. A. Leukocyte-endothelial cell interactions in the inflammatory response. *Lab. Investig. J. Tech. Methods Pathol.* **82**, 521–533 (2002).

87. Pigott, R., Dillon, L. P., Hemingway, I. H. & Gearing, A. J. Soluble forms of E-selectin, ICAM-1 and VCAM-1 are present in the supernatants of cytokine activated cultured endothelial cells. *Biochem. Biophys. Res. Commun.* **187**, 584–589 (1992).

88. Glowinska, B., Urban, M., Peczynska, J. & Florys, B. Soluble adhesion molecules (sICAM-1, sVCAM-1) and selectins (sE selectin, sP selectin, sL selectin) levels in children and adolescents with obesity, hypertension, and diabetes. *Metabolism* **54**, 1020–1026 (2005).

89. O'malley, T., Ludlam, C. A., Riemermsa, R. A. & Fox, K. a. A. Early increase in levels of soluble inter-cellular adhesion molecule-1 (sICAM-1). Potential risk factor for the acute coronary syndromes. *Eur. Heart J.* **22**, 1226–1234 (2001).

90. Kobayashi, T. *et al.* Elevation of serum soluble intercellular adhesion molecule-1 (sICAM-1) and sE-selectin levels in bronchial asthma. *Clin. Exp. Immunol.* **96**, 110–115 (1994).

91. Lawson, C. & Wolf, S. ICAM-1 signaling in endothelial cells. *Pharmacol. Rep. PR* **61**, 22–32 (2009).

92. Merendino, R. A. *et al.* Serum Levels of Interleukin-18 and sICAM-1 in Patients Affected by Breast Cancer: Preliminary Considerations. *Int. J. Biol. Markers* **16**, 126–129 (2001).

93. Reynolds, L. M. *et al.* FADS genetic and metabolomic analyses identify the ∆5 desaturase (FADS1) step as a critical control point in the formation of biologically important lipids. *Sci. Rep.* **10**, 15873 (2020).

94. de Castro, G. S. & Calder, P. C. Non-alcoholic fatty liver disease and its treatment with n-3 polyunsaturated fatty acids. *Clin. Nutr. Edinb. Scotl.* **37**, 37–55 (2018).

95. Guo, X.-F., Yang, B., Tang, J. & Li, D. Fatty acid and non-alcoholic fatty liver disease: Meta-analyses of case-control and randomized controlled trials. *Clin. Nutr. Edinb. Scotl.* **37**, 113–122 (2018).

96. Conomos, M. P. *et al.* Genetic Diversity and Association Studies in US Hispanic/Latino Populations: Applications in the Hispanic Community Health Study/Study of Latinos. *Am. J. Hum. Genet.* **98**, 165–184 (2016).

97. Cao, J., Schwichtenberg, K. A., Hanson, N. Q. & Tsai, M. Y. Incorporation and clearance of omega-3 fatty acids in erythrocyte membranes and plasma phospholipids. *Clin. Chem.* **52**, 2265–2272 (2006).

98. Manichaikul, A. *et al.* Association of SCARB1 variants with subclinical atherosclerosis and incident cardiovascular disease: The multi-ethnic study of atherosclerosis. *Arterioscler. Thromb. Vasc. Biol.* (2012) doi:10.1161/ATVBAHA.112.249714.

99. Mathias, R. A., Pani, V. & Chilton, F. H. Genetic Variants in the FADS Gene: Implications for Dietary Recommendations for Fatty Acid Intake. *Curr. Nutr. Rep.* **3**, 139–148 (2014).

100. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

101. Li, J. Z. *et al.* Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science* **319**, 1100–1104 (2008).

102. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, (2020).

103. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* (2013) doi:10.1016/j.ajhg.2013.06.020.

104. Qi, Q. *et al.* Objectively Measured Sedentary Time and Cardiometabolic Biomarkers in US Hispanic/Latino Adults: The Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *Circulation* **132**, 1560–1569 (2015).

2.9 Acknowledgments

This work was supported by NCCIH R01 AT008621 and USDA ARZT-1361680-H23-157. *The Multi-Ethnic Study of Atherosclerosis:* MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420, UL1-TR-001881, and DK063491. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278. Genotyping was performed at Affymetrix (Santa Clara, California, USA) and the Broad Institute of Harvard and MIT (Boston, Massachusetts, USA) using the Affymetrix Genome-Wide Human SNP Array 6.0. *The Hispanic Community Health Study/Study of Latinos:* HCHS/SOL is a collaborative study supported by contracts from the National Heart, Lung, and Blood Institute (NHLBI) to the University of North Carolina (HHSN268201300001I / N01-HC-65233), University of Miami (HHSN268201300004I / N01-HC-65234), Albert Einstein College of Medicine (HHSN268201300002I / N01-HC-65235), University of Illinois at Chicago (HHSN268201300003I / N01-HC-65236 Northwestern Univ), and San Diego State University (HHSN268201300005I / N01-HC-65237). The following Institutes/Centers/Offices have contributed to the HCHS/SOL through a transfer of funds to the NHLBI: National Institute on Minority Health and Health Disparities, National Institute on Deafness and Other Communication Disorders, National Institute of Dental and Craniofacial Research, National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Neurological Disorders and Stroke, and NIH Institution-Office of Dietary Supplements. The authors thank the staff and participants of HCHS/SOL for their important contributions. A complete list of staff and investigators has been provided by Sorlie P., et al. in Ann Epidemiol. 2010 Aug;20: 642-649 and is also available on the study website http://www.cscc.unc.edu/hchs/. Other funding sources for this study include R01HL060712, R01HL140976, and R01HL136266 from the NHLBI;

and R01DK119268, R01DK120870 and the New York Regional Center for Diabetes Translation Research (P30 DK111022) from the NIDDK. The Genetic Analysis Center at the University of Washington was supported by NHLBI and NIDCR contracts (HHSN268201300005C AM03 and MOD03). Genotyping efforts were supported by NHLBI HSN 26220/20054C, NCATS CTSI grant UL1TR000123, and NIDDK Diabetes Research Center (DRC) grant DK063491. *The Arizona Insulin Resistance Registry:* The AIR registry was supported by Health Research Alliance Arizona and the Center for Metabolic Biology at Arizona State University. Data management support was provided by a grant (UL1 RR024150) from the Mayo Clinic to utilize Research Electronic Data Capture (REDCap).

2.10 Author contributions:

YIC, AMF, SAB, LMJ, and CT generated the data. CY, BH, JCC, QQ and AM analyzed the data. CY, BH, JC, TDO, LMR, ACW, MS, MYT, RNL, DKC, LMJ, QQ, IR, SSR, RAM, FHC and AM contributed to interpretation of the data. FHC, AM, RAM and SSR conceptualized and designed the study. YIC, LMS, MYT, RCK, MLD, LJM, DKC, and SSR provided critical oversight to data collection and study coordination. CY, BH, AM and FHC wrote the manuscript. All authors contributed to critical editing of the manuscript.

Table 2.1: Participant Characteristics for individuals of self-identified Hispanic origin from the MESA cohort.

Characteristics	Self-reported Hispanic country/region of origin						Total (N=1102)
	Cuba (N=45)	Dominican (N=145)	Puerto Rico (N=167)	South Amer. (N=93)	Central Amer. (N=80)	Mexico (N=572)	
Sex (Female)	42.2%	53.1%	52.7%	54.8%	58.8%	48.3%	50.6%
Age (years)	69.8 (9.1)	58.8 (10.1)	59.3 (9.4)	62.9 (10.3)	58.7 (8.1)	61.8 (10.2)	61.2 (10.1)
Study Site: Columbia University	73.33%	99.31%	86.24%	60.22%	20%	0.54%	35.93%
Study Site: University of Minnesota	15.56%	0.69%	11.37%	15.05%	13.75%	38.98%	24.95%
Study Site: UCLA	11.11%	0	2.39%	24.73%	66.25%	60.48%	39.12%
Height (cm)	163.1 (9.7)	163.3 (9.4)	162.6 (9.3)	160.4 (8.9)	159.6 (9.2)	161.7 (9.5)	161.8 (9.4)
Weight (kg)	75.3 (13.9)	75.4 (14.3)	79.5 (17.4)	71.4 (12.3)	74.9 (15.3)	77.7 (15.8)	76.8 (15.6)
Waist-to-hip ratio	0.98 (0.06)	0.93 (0.08)	0.94 (0.08)	0.94 (0.07)	0.96 (0.06)	0.97 (0.07)	0.96 (0.07)
BMI (kg/m^2)	28.3 (5.3)	28.2 (4.6)	29.9 (5.7)	27.7 (4.1)	29.3 (5.2)	29.6 (5.2)	29.3 (5.1)
HDL-C (mg/dl)	49.8 (18.1)	47.2 (10.7)	49.7 (14.2)	50.8 (13.9)	48.0 (12.1)	45.9 (12.5)	47.4 (13.0)
LDL-C (mg/dl)	121.1 (26.1)	124.7 (35.5)	118.0 (33.3)	115.8 (29.6)	120.9 (38.3)	119.4 (32.8)	119.8 (33.2)
Triglycerides (mg/dl)	154.2 (100.4)	132.9 (69.9)	134.4 (72.1)	151.5 (168.8)	144.1 (75.5)	173.6 (113.4)	157.5 (107.6)
s-ICAM (ng/ml)	311.7 (71.8)	262.4 (88.2)	307.6(110.6)	276.4 (72.3)	286.6 (69.4)	298.7 (80.6)	293.2 (86.2)
E-Selectin (ng/ml)	64.05 (32.8)	54.15 (17.9)	63.65 (27.4)	57.96 (29.8)	62.74 (26.1)	67.07 (29.3)	63.06 (27.4)
Fish intake (servings/day)	0.19 (0.28)	0.21 (0.22)	0.21 (0.25)	0.20 (0.28)	0.22 (0.23)	0.15 (0.21)	0.18 (0.23)
EPA (% of total fatty acids)	0.90 (0.55)	0.87 (0.68)	0.76 (0.55)	0.72 (0.43)	0.58 (0.38)	0.52 (0.29)	0.64 (0.46)
DPA (% of total fatty acids)	0.98 (0.24)	0.95 (0.26)	0.90 (0.20)	0.89 (0.22)	0.83 (0.18)	0.84 (0.19)	0.88 (0.21)
DHA (% of total fatty acids)	3.71 (1.51)	4.15 (1.31)	3.58 (1.23)	3.76 (1.25)	3.24 (1.15)	2.69 (0.90)	3.19 (1.23)
ARA (% of total fatty acids)	12.18 (2.58)	12.84 (2.52)	12.00 (2.65)	10.53 (2.26)	11.04 (2.40)	10.64 (2.36)	11.22 (2.56)
Global Proportion of Amerind ancestry	0.06	0.06	0.12	0.33	0.39	0.41	0.30
Global Proportion of African ancestry	0.19	0.41	0.23	0.09	0.16	0.04	0.14
Global Proportion of European ancestry	0.75	0.53	0.65	0.58	0.45	0.55	0.56
rs174537 frequency* of effect allele T (versus G allele)	0.28`	0.27	0.40	0.56	0.59	0.59	0.51

Table 2.1 shows the phenotypic descriptive statistics are presented as percentages for dichotomous variables and mean (standard deviation) for continuous variables.

* For comparison, the rs174537 effect allele frequencies were 0.007, 0.328 and 0.858 in the 1000 Genomes AFR, EUR and AMR populations, respectively, where the allele frequency calculation was restricted to the cleaned set of samples that were included in the reference set for local ancestry analysis (see Supplementary Methods for details).

Table 2.2. Genotypic effects of rs174537 on fasting lipids, anthropometrics and inflammatory traits.

			Beta	P value
Fasting Lipids	Triglycerides (mg/dL)	GT	21.27	0.0001
		TT	29.94	1.01 x 10^{-06}
	HDL-C (mg/dL)	GT	-1.30	0.141
		TT	-2.48	0.010
Anthropometrics	waist-hip ratio	GT	0.006	0.152
		TT	0.013	8.94 x 10^{-03}
	Height (cm)	GT	-1.36	0.002
		TT	-3.46	6.59 x 10^{-12}
	Weight (kg)	GT	-1.89	0.077
		TT	-3.12	7.48 x 10^{-03}
	BMI (kg/m^2)	GT	-0.25	0.50
		TT	-0.002	0.99
Inflammatory	s-ICAM (ng/mL)	GT	30.64	0.002
		TT	26.09	0.018
	E-Selectin (ng/mL)	GT	10.00	0.048
		TT	11.50	0.032

Table 2.2 shows the regression analysis results for the effect of rs174537 genotype on Fasting Lipids, Anthropometrics and inflammatory with adjustment for age and sex. For the effect sizes, the effects of GT and TT are in reference to GG. The sample size is 1102 (GG: 293; GT: 484; TT: 325) for waist-hip ratio; height; weight and BMI, 1101 (GG: 293 ;GT: 483 ;TT: 325) for triglycerides and HDL-C, 439 (GG: 112; GT: 194; TT: 133) for s-ICAM and 183 (GG: 48; GT: 76; TT: 59) for E-Selectin. *P*-values are calculated using two-sided t test for the regression coefficient.

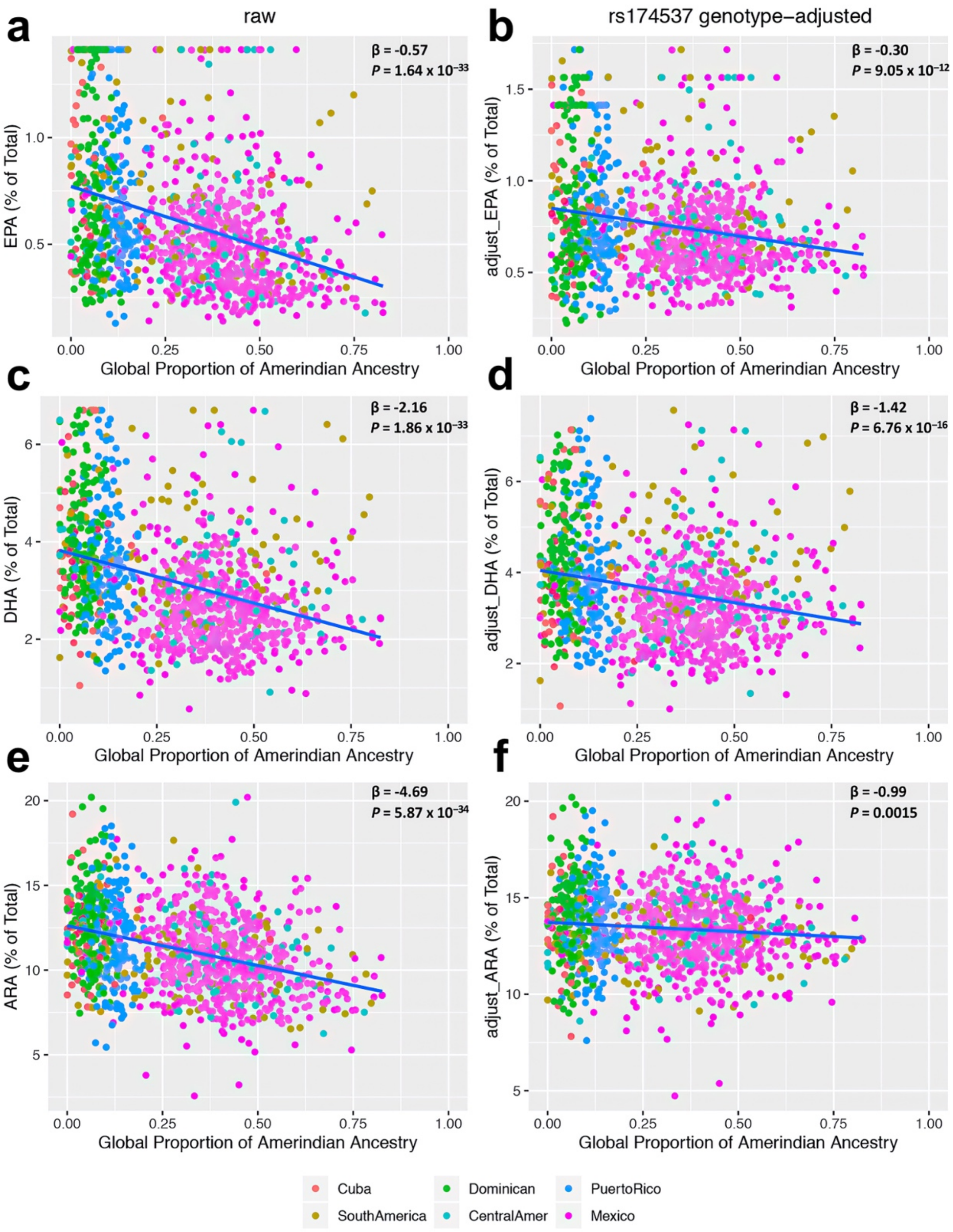


Figure 2.1. Relationship of LC-PUFA levels with global proportion of Amerind ancestry before and after adjustment for rs174537 genotype. The regression effect estimates (β expressed as % of total fatty acids) and *P*-values are shown in the upper right corner of each panel. The relationship of LC-PUFA levels with Global Proportion of Amerind Ancestry as estimated from genome-wide SNP data is shown for (a) EPA - raw, (b) EPA – genotype-adjusted, (c) DHA – raw, (d) DHA – genotyped-adjusted, (e) ARA – raw, and (f) ARA – genotype-adjusted. Here, the rs174537 genotype-adjusted LC-PUFA levels were obtained as residuals after regression against rs174537 genotype and re-centered around the raw means. *P*-values are calculated using a two-sided t-test for the regression coefficient derived with n = 1102 biologically independent samples.

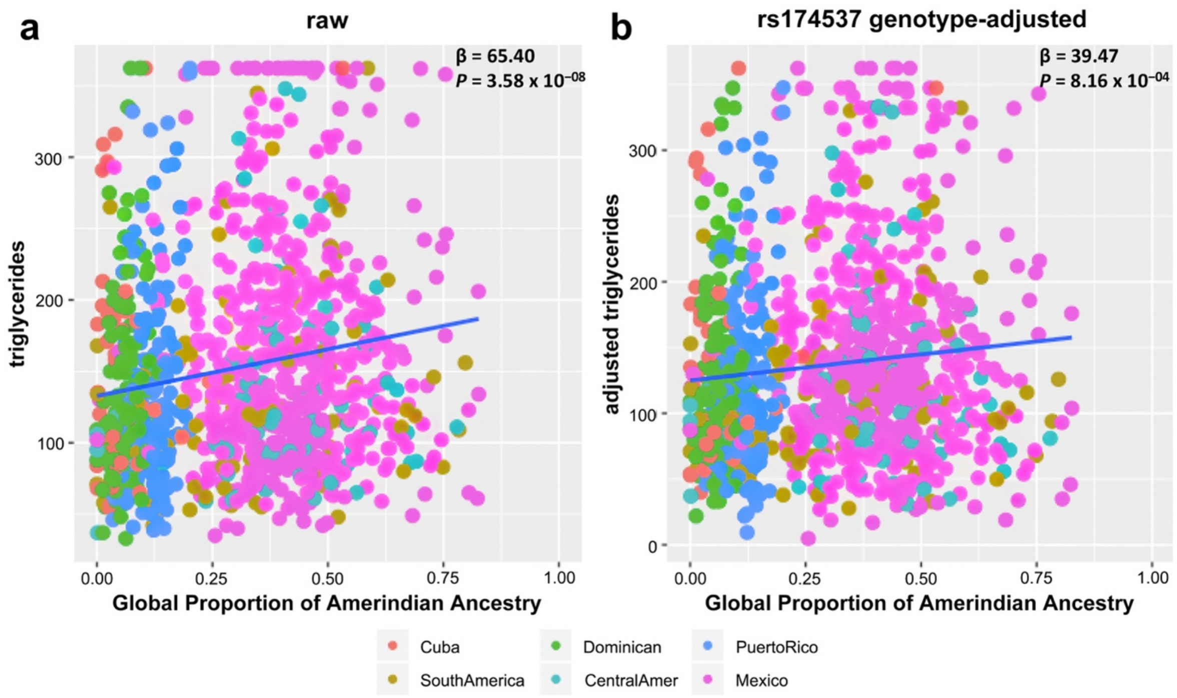


Figure 2.2. Relationship of triglycerides with global proportion of Amerind ancestry before and after adjustment for rs174537 genotype. The regression effect estimates (β in mg/dL) and *P*-values are shown in the upper right corner of each panel. The relationships are shown for each of (a) raw triglyceride levels, and (b) genotype-adjusted triglyceride levels with Global Proportion of Amerind Ancestry. Here, rs174537 genotype-adjusted triglyceride levels were obtained as residuals from regression accounting for rs174537 genotype, and re-centered around the raw means. *P*-values are calculated using a two-sided t-test for the regression coefficient derived with n = 1101 biologically independent samples.

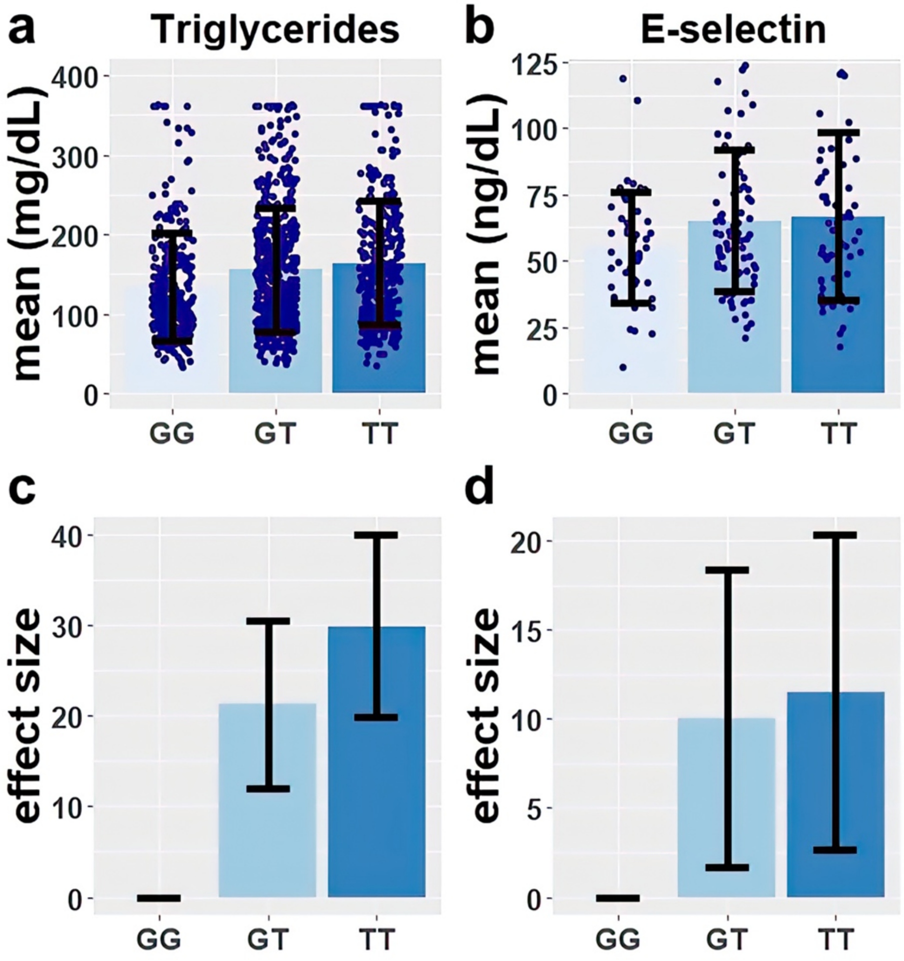


Figure 2.3: Genotypic effects of rs174537 on triglycerides and E-selectin. The mean and standard deviation are shown for (a) triglycerides, and (b) E-selectin stratified by rs174537 genotypes. The estimated effect and standard error among participants carrying one or two copies of the ancestral allele T (compared to the reference of zero), after covariate-adjustment for age and sex are shown for (c) triglycerides, and (d) E-selectin. Numbers of independent samples for the analyses presented are 1101 (GG: 293 ;GT: 483 ;TT: 325) for triglycerides and 183 (GG: 48; GT: 76; TT: 59) for E-selectin. Source data for the figure are provided in Supplementary Data 1.

Chapter 3

Genome-wide association studies and fine-mapping identify genomic loci for n-3 and n-6 polyunsaturated fatty acids in Hispanic American and African American cohorts

Published in: **Yang C**, Veenstra J, Bartz TM, Pahl MC, Hallmark B, Chen YI, Westra J, Steffen LM, Brown CD, Siscovick D, Tsai MY, Wood AC, Rich SS, Smith CE, O'Connor TD, Mozaffarian D, Grant SFA, Chilton FH, Tintle NL, Lemaitre RN, Manichaikul A. Genome-wide association studies and fine-mapping identify genomic loci for n-3 and n-6 polyunsaturated fatty acids in Hispanic American and African American cohorts. Commun Biol. 2023 Aug 16;6(1):852. doi: 10.1038/s42003-023-05219-w.

3.1 Abstract

Omega-3 (n-3) and omega-6 (n-6) polyunsaturated fatty acids (PUFAs) play critical roles in human health. Prior genome-wide association studies (GWAS) of n-3 and n-6 PUFAs in European Americans from the CHARGE Consortium have documented strong genetic signals in/near the *FADS* locus on chromosome 11. We performed a GWAS of four n-3 and four n-6 PUFAs in Hispanic American (n = 1454) and African American (n = 2278) participants from three CHARGE cohorts. Applying a genome-wide significance threshold of $P < 5 \times 10^{-8}$, we confirmed association of the *FADS* signal and found evidence of two additional signals (in *DAGLA* and *BEST1*) within 200 kb of the originally reported *FADS* signal. Outside of the *FADS* region, we identified novel signals for arachidonic acid (AA) in Hispanic Americans located in/near genes including *TMX2*, *SLC29A2*, *ANKRD13D* and *POLD4,* and spanning a >9 Mb region on chromosome 11 (57.5Mb ~ 67.1Mb). Among these novel signals, we found associations unique to Hispanic Americans, including rs28364240, a *POLD4* missense variant for AA that is common in CHARGE Hispanic Americans but absent in other race/ancestry groups. Our study sheds light on the genetics of PUFAs and the value of investigating complex trait genetics across diverse ancestry populations.

3.2 Introduction

Omega-3 (n-3) and omega-6 (n-6) polyunsaturated fatty acids (PUFAs) are critical structural components of cell membranes, which can influence cellular activities by promoting the fluidity, flexibility, and the permeability of a membrane.[1–3] Additionally, PUFAs affect a variety of other biological processes and molecular pathways, including modulating membrane channels and proteins, regulating gene expression through nuclear receptors and transcription factors, and conversion of the PUFAs themselves into bioactive metabolites.[4] Levels of circulating PUFAs and long chain (≥20 carbons) PUFAs (LC-PUFAs) are associated with reduced risk of cardiovascular disease[5,6], type 2 diabetes mellitus[7], cognitive decline[8], Alzheimer's disease[9], metabolic syndrome[10] and breast cancer[11], as well as all-cause mortality.[12]

PUFAs and LC-PUFAs are characterized by the position of the first double bond from the methyl terminal (omega; ω; or n−FAs) and fall into two primary families, n-3 and n-6. The most abundant n-3 PUFAs are alpha-linolenic acid (ALA), eicosapentaenoic acid (EPA), docosapentaenoic acid (DPA) and docosahexaenoic acid (DHA), while the primary n-6 PUFAs are linoleic acid (LA), gamma-linolenic acid (GLA), dihomo-γ-linolenic acid (DGLA) and arachidonic acid (AA). ALA and LA are essential n-3 and n-6 PUFAs consumed from the diet and these then can be converted to more unsaturated LC-PUFAs through a set of desaturation and elongation enzymatic steps. For example, DGLA and AA can be synthesized from LA, while EPA, DPA and DHA can be produced from ALA. The precursors LA and ALA are essential fatty acids that must be provided by the diet. Due to the lower abundance of ALA in Western diets and the inefficiency of conversion of ALA to longer chain n-3 LC-PUFAs such as EPA and DHA, dietary intake of these via fatty fish or marine oil supplementation is often recommended.[13,14]

Previous studies have shown that African ancestry populations have higher circulating levels of LC-PUFAs compared to European Americans.[15] These large differences can be explained in part by variation in the allele frequencies of *FADS* variants associated with different biosynthetic efficiencies in these two populations.[16] Mathias *et al.* also revealed that African Americans have significantly higher levels of AA and lower levels of the AA precursor DGLA, and that *FADS1* variants were significantly associated with AA, DGLA and the AA/DGLA ratio in a sample of fewer than 200 African Americans from the GeneSTAR study.[15] In addition, African ancestry populations have higher frequencies of the derived *FADS* haplogroup (represented by the variant rs174537 allele G)[17] that is associated with more efficient conversion for PUFAs.[16] In contrast, Amerind ancestry Hispanic populations have higher frequencies of the ancestral *FADS* haplogroup (represented by rs174537 allele T) that has a reduced capacity to synthesize PUFAs. Accordingly, we demonstrated that higher global proportions of Amerind ancestry are associated with lower levels of PUFAs in Hispanic populations.[17]

Genome-wide association studies (GWAS) of n-3 and n-6 PUFAs were performed by the CHARGE consortium in European ancestry (EUR) participants.[18–20] The CHARGE GWAS of n-3 PUFAs in 8,866 European Americans identified genetic variants in/near *FADS1* and *FADS2* associated with higher levels of ALA and lower levels of EPA and DPA, as well as SNPs in *ELOVL2* associated with higher EPA and DPA and lower DHA. The CHARGE GWAS of n-6 PUFAs in 8,631 European Americans confirmed that variants in the *FADS* gene cluster were associated with LA and AA, and it revealed that variants near *NRBF2* were associated with LA and those in *NTAN1* were associated with LA, GLA, DGLA, and AA (**Figure 1**). In the Framingham Heart Offspring Study, variants in/near *PCOLCE2*, *LPCAT3*, *DHRS4L2*, *CALN1 FADS1/2*, and *ELOVL2* were associated with PUFAs in European ancestry participants.[21,22] Collectively, these studies played an important role in identifying the genetic associations of n-3 and n-6 PUFAs in European ancestry populations.

To address the paucity of GWAS of PUFAs in non-European ancestry cohorts, we performed a meta-analysis of genome-wide association studies for n-3 and n-6 PUFAs for Hispanic American (HIS) and African American (AFA) participants from three CHARGE consortium cohorts: the Multi-Ethnic Study of Atherosclerosis (MESA), the Cardiovascular Health Study (CHS) and the Framingham Heart Study (FHS) Omni cohort. The major goals of the study were (1) to examine whether the major loci identified in European Americans are shared across race/ancestry groups, and (2) to examine evidence for genetic association unique to HIS and AFA populations. As GWAS approaches are not sufficient to identify the causal variants and determine the number of independent signals, especially in the context of long stretches of linkage disequilibrium (LD) within the *FADS* locus[15,23], we conducted statistical fine-mapping[24] to identify the most likely causal variants within each n-3 and n-6 PUFA-associated locus. We performed cross-ancestry replication analysis in CHARGE and MESA, with validation using the multi-ancestry GWAS of lipids from the Global Lipids Genetics Consortium (GLGC).[25] Subsequently, we performed integrative analysis leveraging gene expression data from MESA[26,27] and the Genotype-Tissue Expression (GTEx) project[28] to identify genes that could contribute to our identified genetic association results. Finally, we examined open chromatin defined by ATAC-seq to determine the impact and physical contact of the identified variants with nearby genes **(Figure 2)**. Our study demonstrates the vital importance of diverse ancestry genetic studies for the study of complex traits, and particularly for metabolites that have been subject to evolutionary pressures and are closely regulated by specific protein-coding genes.

3.3 Results

3.3.1 Participant characteristics

The participants in the meta-analysis of GWAS for PUFAs included 1,454 HIS and 2,278 AFA unrelated participants **(Table 3.1;** fatty acid levels are expressed as the percentage of total fatty acids throughout the entire manuscript**)**. There were some differences in the distributions of fatty acid levels observed across cohorts, which were likely due to the sources of biospecimens for the assays (plasma phospholipids for MESA and CHS versus erythrocytes for FHS). For example, mean levels of DPA varied from 0.85% (CHS: plasma phospholipids) to 2.54% of total fatty acids (FHS: erythrocytes) in AFA and AA from 11.01% (MESA: plasma phospholipids) to 16.56% (FHS: erythrocytes) in HIS **(Table 3.1)**. In addition, n-6 PUFAs, especially LA and AA, have relatively higher mean levels than n-3 PUFAs in all cohorts **(Table 3.1**).

Regardless of whether the fatty acids were measured in plasma phospholipids or erythrocytes, AFA populations had higher levels of AA and elevated ratios of AA to DGLA and AA to LA relative to Hispanic populations. This result would be expected given the frequency differences in the derived (efficient) to ancestral (inefficient) *FADS* haplogroups between these two populations. As expected, due to the lower levels of dietary ALA relative to LA entering the biosynthetic pathway, levels of n-3 LC-PUFAs including EPA, DPA and DHA were significantly lower than the n-6 LC-PUFA, AA. Additionally, African Americans had higher levels of n-3 LC-PUFAs than Hispanic Americans, again likely due to differences in the ratio of the derived to ancestral *FADS* haplogroups. These differences are similar to those observed examining the same

PUFAs and LC-PUFAs and ratios when comparing African Americans and European Americans.[15,29]

3.3.2 Confirmation of top variants identified in prior CHARGE EUR GWAS of PUFAs

We began by examining associations of seven known PUFA-associated signals from CHARGE EUR (summarized in **Figure 3.1**) in our current study of CHARGE HIS and AFA. Multiple variants identified by previous CHARGE EUR GWAS meta-analyses[19,20] were also identified in CHARGE HIS (*FADS1/2* region: rs174547 and rs174538, *PDXDC1* variant: rs16966952 and *GCKR* variant: rs780094) and AFA (*FADS1/2* region: rs174547, *PDXDC1* variant: rs16966952, *GCKR* variant: rs780094 and *ELOVL2* variant: rs3734398) after adjusting for multiple testing for the number of variants examined across the eight PUFAs ($P < 0.05/8 = 0.006$) (**Supplementary Data 3.1**). The directions of effect observed in HIS and AFA for these variants were consistent with those reported for European ancestry populations in prior CHARGE GWAS meta-analyses of n-3 and n-6 PUFAs (**Supplementary Data 3.1**).

3.3.3 GWAS and fine-mapping identify novel PUFA-associated genetic signals in CHARGE HIS and AFA

Based on a genome-wide significance threshold of $P < 5 \times 10^{-8}$, our complete GWAS of n-3 and n-6 PUFAs identified associations on chromosomes 11, 15 and 16 in CHARGE HIS (**Table 3.2, Supplementary Figure 3.1 and Supplementary Figure 3.2**) and chromosomes 6, 7, 10 and 11 in CHARGE AFA (**Table 3.3, Supplementary Figure 3.3 and Supplementary Figure 3.4**). For regions with more than one genome-wide significant variant, we applied statistical fine-mapping to identify the independent putative causal signals (credible sets) for each genome-wide significant locus. We carried out these analyses separately for our CHARGE HIS and CHARGE AFA GWAS meta-analysis results.

We identified multiple independent putative causal signals for the PUFA traits [AA: 8 signals (credible sets); ALA: 1; DGLA: 5, DPA: 2; EPA: 1; GLA: 1; LA: 6] in HIS and [AA: 5; DGLA: 2, DPA: 2, LA: 1] in AFA (**Table 3.2, Table 3.3, Supplementary Data 3.2 and Supplementary Data 3.3).** We examined the overlap of signals identified from fine-mapping in HIS versus AFA. We observed that the credible sets were generally smaller in AFA (average number of variants in credible set: HIS:3.4; AFR:2.2) possibly driven by the lower average LD in AFA.

Among the independent credible sets identified, most were novel associated signals within a +/- 5 Mb region of the previously reported *FADS* signal on chromosome 11 (**Tables 3.2-3.3**). Examining all the signals for PUFAs in HIS and AFA, we observed that the lead signal (reflecting the strongest evidence of association) on chromosome 11 represents the *FADS* signal reported in the previous GWAS.[20] For example, rs174547, the *FADS1* variant reported in the previous CHARGE EUR GWAS, is one of the variants in the first credible set for AA in HIS.[19,20] In addition to the known *FADS* signals, we also observed multiple novel independent signals at other regions of chromosome 11 for PUFAs [AA: 6 novel signals (credible sets) and LA: 3] in HIS, for example, in/near *ANKRD13D*, *TMX2*, *POLD4* and *SLC29A2* and spanning a long range (57.5Mb ~ 67.1Mb) on chromosome 11 for AA in HIS (**Table 3.2**). Additionally, we observed several novel independent signals on other chromosomes showing associations with the PUFA traits in AFA [AA: 1 novel signal on chromosome 7 and DPA: 1 on chromosome 6] (**Table 3.3**).

3.3.4 Additional independent PUFA-associated signals on chromosome 11 demonstrate chromatin contacts with *FADS* and other genes

While prior studies have represented the *FADS* signal as primarily one signal,[19,20] our study demonstrates numerous independent signals within the region (**Table 3.2**). For example, for AA we report signals intronic to *BEST1* and *DAGLA* within the *FADS* region (+/- 1Mb of the lead variant, rs102274**; Figure 3.3a)**. We examined this region to identify the subset of variants that may affect cis-regulatory elements in physical contact with nearby genes. Four variants within the credible sets in this region were located in regions of open chromatin defined by ATAC-seq and were in contact with gene promoters defined by Promoter Capture C in multiple metabolic-relevant cell types (human mesenchymal stem cells [hMSC], adipocytes derived from in vitro from the hMSC [hMSC_Adipocytes], induced pluripotent stem cell derived Hepatocytes [iPSC_Hepatocytes], embryonic stem cell derived Hypothalamic Neurons [hESC_HypothalamicNeurons], Enteroids, and HepG2s). Almost all of the interactions we detected were bait-to-bait interactions, meaning that they reflected physical contact between promoters of two different genes (**Supplementary Data 3.4**). For example, the region surrounding rs2668898 near *BEST1* showed evidence of physical contact with the *TMEM258*, *FADS1* and *FADS2* region in multiple cell types and *TMEM258* region also showed evidence of physical contact with the *FADS1* and *FADS2* region (**Figure 3.4a and Supplementary Data 3.4**). Besides the *FADS* region, we further found evidence of physical contact between *POLD4* and *ANKRD13D* **(Figure 3.4b and Supplementary Data 3.4)**, corresponding to the regions surrounding two signals identified in fine-mapping of AA in HIS **(Figure 3.3a)**.

3.3.5 Novel signals on chromosome 11 identified in HIS show evidence of cross-ancestry replication or validation

We investigated evidence of cross-ancestry replication for signals identified in our present GWAS of CHARGE HIS and AFA by examining evidence of genetic association in European Americans (CHARGE EUR[19,20] and MESA EUR), African Americans (CHARGE AFA), Hispanic Americans (CHARGE HIS) and Chinese Americans (MESA CHN). Replication analysis was performed with multiple testing correction (HIS: P < 0.05/19 signals = 0.0026 and AFA: P < 0.05/11 signals= 0.004; **Supplementary Data 3.5 and Supplementary Data 3.6**).

As noted previously, the first credible set identified in our present GWAS of HIS and AFA for each trait (reflecting the strongest evidence of association) generally coincided with the region of chromosome 11 reported in prior CHARGE GWAS efforts. These signals showed evidence of genetic association in European Americans, as well as across race/ancestry groups. For example, rs102274 for AA was replicated in the MESA EUR, CHARGE AFA and MESA CHN groups (MESA EUR: P = 1.04 x 10^{-151}, CHARGE AFA: P = 2.36 x 10^{-47}, MESA CHN: P = 8.75 x 10^{-92}) (**Supplementary Data 3.5**).

Additionally, three novel signal were also replicated across race/ancestry groups (**Table 3.4**). Specifically, the *DAGLA* variant rs198434 and *MYRF* variant rs198461 in credible sets 3 and 4, respectively, for DGLA were replicated in analysis of MESA EUR (rs198434: P = 2.54 x 10^{-03} and rs198461: P = 7.37 x 10^{-09}). *TMX2* variant rs518894 in credible set 6 for LA was replicated in CHARGE EUR (P = 2.50 x 10^{-03}).

Some of the novel signals without cross-ancestry replication demonstrated large differences in allele frequencies across groups. For example, the effect allele frequency of rs28364240, a *POLD4* missense variant in credible set 3 for AA in Hispanics, is 0.204 in our CHARGE HIS group, but close to zero in the other race/ancestry groups examined (EUR: 0.003, AFR: 0.007, CHN: 0.005) (**Figure 3.3b, Supplementary Data 3.5 and 3.7**) and the effect allele frequency of rs142068305, a *ANKRD13D* intron variant, is 0.196 in our CHARGE HIS group while 0.007, 0.004 and 0.005 in AFR, EUR and CHN, respectively. These results suggest evidence of genetic association signals unique to HIS or other groups carrying Amerindian ancestry or admixture.

As some variants could not be interrogated using independent GWAS of PUFA traits, given those studies' focus on specific race/ancestry groups which may not include our variants of interest and/or limited sample sizes, we performed validation analyses using the results of multi-ancestry GWAS of lipid levels from the GLGC[25] including ~1.65 million individuals from five genetic ancestry groups (admixed African or African, East Asian, European, Hispanic and South Asian). We focused on the most significant putative causal variants from each credible set and applied multiple testing correction for the number of validated variants (HIS: $P < 0.05/19 = 0.0026$ and AFA: $P < 0.05/11 = 0.004$). Interestingly, we observed that multiple novel signals without cross-ancestry replication did demonstrate association with one or more lipid levels. For example, the LA associated *LRP4* variant rs11039018 was validated based on its association with

HDL and Triglycerides (HDL: P = 2.85 x 10^{-74} and Triglycerides: P = 4.50 x 10^{-43}), while the LA associated *MARK2* intron variant rs10751002 was validated based on its association with LDL and Total Cholesterol (LDL: P = 3.31 x 10^{-12} and Total Cholesterol: P = 5.74 x 10^{-09}) (**Table 3.4, Supplementary Data 3.8 and Supplementary Data 3.9**).

3.3.6 Integrative analyses identify putative causal genes and pathways for the PUFA loci

Using colocalization with eQTL resources, we identified candidate genes underlying the genetic association signals for the PUFA traits. In HIS, we found colocalization with expression of the genes *MED19, TMEM258, PACS1, RAD9A, C11orf24, CTTN* on chromosome 11 and *PDXDC1* on chromosome 16 based on MESA multi-ancestry eQTL resources[26] (**Table 3.5 and Supplementary Data 3.10**). In further analysis using eQTL resources from GTEx whole blood[28], we confirmed colocalization with *TMEM258* and *MED19* identified using the MESA multi-ancestry eQTLs, and also identified colocalization with *FADS1*, *RPS4XP13, AP001462.2, PGA5, PGA5, TPCN2, MEN1* on chromosome 11 and *RP11-156C22.5* on chromosome 16. (**Table 3.5 and Supplementary Data 3.11**).

We also performed complementary integrative analysis using PrediXcan, identifying significant associations for predicted expression of *TMEM258* with AA, ALA, DGLA, DPA, EPA, GLA and LA (after multiple testing correction for all genes examined: P < 0.05/4470 = 0.00001), based on integration with eQTL from both MESA and GTEx. PrediXcan also identified *TMEM109*, *ZBTB3*, *TTC9C, POLD4, INCENP* and *FERMT3* on chromosome 11 and *PDXDC1* on chromosome 16 as putative genes associated with

PUFAs in HIS (**Table 3.5, Supplementary Data 3.12 and Supplementary Data 3.13**). For AFA, colocalization and PrediXcan analyses did not identify any genes of interest that met our pre-specified thresholds for statistical significance.

Incorporating the prior chromatin contacts identified (**Supplementary Data 3.4**), we found that several of our GWAS regions had physical contact with one or more genes identified by integration with eQTL resources. For example, *RAD9A* was supported by colocalization with MESA eQTL[26] and also showed chromatin contact with *POLD4* in nearly all cell types examined (**Figure 3.4b**). In addition, *INCENP* was supported by PrediXcan using both MESA[26] and GTEx[30] resources and also showed chromatin contact with *TMEM258*, *FADS1* and *FADS2* in nearly all cell types examined (**Figure 3.4a**). We further observed that *CLCF1*, *RAD9A*, *FADS2*, *TMEM258*, *INCENP*, *FADS1* identified from colocalization or PrediXcan were additionally supported by chromatin contact analyses (**Figure 3.4, Supplementary Data 3.4**).

To follow-up on the genes of interest identified by colocalization and PrediXcan analyses, we examined their co-expression with *FADS1* using GTEx whole blood gene expression[28] with multiple testing correction for the number of genes under consideration (HIS: $P < 0.05/39 = 0.0012$). In both unadjusted and age/sex-adjusted regression models, multiple genes showed statistically significant co-expression with *FADS1*, for example, *TMEM258*, *MED19*, *POLD4*, *RAD9A* and *SSH3* (**Supplementary Data 3.14**), suggesting these genes have shared patterns of expression.

We further applied gene set enrichment analysis to the set of genes identified by our integrative colocalization and PrediXcan analyses using the Molecular Signatures Database (MSigDB)[31–33] gene sets **(Supplementary Data 3.15)**. The most significantly enriched gene set (NIKOLSKY_BREAST_CANCER_11Q12_Q14_AMPLICON) comprised the set of genes within amplicon 11q12-q14 identified in a copy number alterations study of 191 breast tumor samples[34] (P = 6.71 x 10^{-17}), which included twelve genes from among those identified by the integrative follow-up analyses of our GWAS results: *RAD9A*, *CTTN*, *PGA5*, *TPCN2*, *TMEM109*, *POLD4*, *CLCF1*, *SSH3*, *TBC1D10C*, *CCS*, *BBS1*, and *DPP3.* The second most significantly enriched gene set (PEA3_Q6) represents the set of genes having at least one occurrence of the motif ACWTCCK in the regions spanning 4 kb centered on their transcription starting sites (P = 3.25 x 10^{-09}), which included eight genes from among those identified in our integrative analyses: *TMEM258, C11orf24, FERMT3, POLD4, TBC1D10C, CCDC88B, MAP4K2* and *DPP3.*

3.4 Discussion

To address the relative lack of prior studies examining genetics of PUFA levels in non-European ancestry populations, we carried out a meta-analysis of GWAS of n-3 and n-6 PUFAs in HIS and AFA across three cohorts: MESA, CHS and FHS. Examining genetic variants identified in prior CHARGE GWAS of the same traits in European Americans[19,20], we demonstrated evidence of association with n-3 and n-6 PUFAs for the signals in/near *FADS1/2* on chromosome 11, *PDXDC1* on chromosome 16, and *GCKR* on chromosome 2 in both HIS and AFA from our current CHARGE GWAS, as well as for *ELOVL2* on chromosome 6 in AFA only.

Through genome-wide analysis and subsequent statistical fine-mapping of our ancestry-specific results, we demonstrated evidence of multiple independent novel signals within the *FADS1/2* locus in both HIS and AFA, and in/near *ELOVL2* in AFA*.* Among these independent novel signals, we found three signals identified in HIS demonstrated evidence of replication in AFA based on association with the same PUFA traits in MESA and CHARGE (LA: rs518804 intronic to *TMX2* [Thioredoxin related transmembrane protein 2]; DGLA: rs198461 intronic to *MYRF* [Myelin regulatory factor] and rs198434 intronic to *DAGLA* [Diacylglycerol lipase alpha]). Additionally, multiple novel signals without cross-ancestry replication did show evidence of validation based on association with lipid levels in GLGC[25]. For example, rs11039018, a *LRP4* (LDL receptor related protein) intron variant associated with AA and LA was validated based on its association with HDL and Triglycerides. This finding is supported by animal studies showing that deficiency of *Lrp4* in adipocytes increased glucose and insulin tolerance and reduced serum fatty acids.[35] Prior studies from the FORCE consortium have shown that LA is associated with lower risk of diabetes, thus it is possible that the association of *LRP4* on diabetes risk factors is mediated by LA.[36] In addition, a *MARK2 (*microtubule affinity regulating kinase 2*)* intron variant rs10751002 associated with LA was validated based on its association with LDL and total cholesterol. We chose to perform validation analysis using association results for lipid levels from the GLGC[25] due to (1) the large sample size (>1 million) which made our validation effort very well powered to detect associations with the selected lipid traits, and (2) the association between fatty acids and lipid traits, for example, fish oil supplements lowering triglycerides[37] and dietary linoleic acid lowering cholesterol[38,39].

While we identified three signals in HIS with evidence of cross-ancestry replication, we also found a large number of signals in HIS that could not be replicated across race/ancestry groups (European, African American and Chinese), in part to differences in allele frequencies. For example, the chromosome 11 *POLD4* (DNA polymerase delta 4, accessory subunit) missense variant rs28364240 and *ANKRD13D* (ankyrin repeat domain 13D) intron variant rs142068305 identified in association with AA have minor allele frequencies of 0.204 and 0.196 in HIS, compared to frequencies close to zero in other race/ancestry groups.

Examining the distance between the putative causal variants in different credible sets identified in HIS, we observed that the signals on chromosome 11 span a long range (57.5Mb ~ 67.1Mb). The extended physical distance covered by these independent PUFA-associated variants, combined with their subsequent validation in association with selected lipid traits, suggests there may be long-range chromatin interactions or other forms of physical interaction that may have yielded distinct genetic associations across this region.[40] Interestingly, prior studies have reported the *FADS* signal on chromosome 11 as primarily just one genetic signal.[19,20] However, our study provides evidence of two more independent signals (*BEST1* and *DAGLA*) within this *FADS* region. In order to understand the chromatin interactions of the *FADS* region on chromosome 11, we used ATAC-seq peaks and chromatin loops to perform the chromatin contact analyses. We identified multiple genes from colocalization or PrediXcan also supported by chromatin contacts, including *CLCF1*, *RAD9A*, *FADS2*, *TMEM258*, *INCENP* and *FADS1*, providing support for the role of our identified genetic signals in regulating these genes. In addition, we observed evidence of chromatin contacts among multiple distinct credible sets identified based on our fine-mapping of genetic signals on chromosome 11. For example, the region surrounding rs2668898 near *BEST1* also showed evidence of physical contact with the *TMEM258, FADS1 and FADS2* region in multiple cell types and *TMEM258* also showed evidence of physical contact with the *FADS1 and FADS2* region. This support for physical contact among some of the multiple independent signals within the *FADS* region opens the possibility of coordinated regulation among these distinct genetic signals. Besides the *FADS* region, *POLD4* also showed evidence of physical contact with the *ANKRD13D* region in multiple cell types. The cell types examined for chromatin interaction correspond to pancreas, liver, and other cell types that could play a role in synthesis and regulation of fatty acids. While the cell types used to examine chromatin interactions are distinct from those used for our integrative eQTL analyses, the chromatin interaction results do provide support for the plausible role of the genes identified by colocalization and PrediXcan.

Through integrative analyses, including colocalization analysis and PrediXcan, that examined overlap of our GWAS of PUFA levels with selected eQTL resources[26,28], we identified putative candidate genes that may shed light on the functional mechanisms of our identified genetic association signals. On chromosome 11 containing the *FADS* genes, we identified overlap with eQTL for multiple other genes including *MED19* (Mediator Complex Subunit 19), *TMEM258* (Transmembrane Protein 258), *PACS1* (Phosphofurin Acidic Cluster Sorting Protein 1), *RAD9A* (RAD9 Checkpoint Clamp Component A) and *CTTN* (Cortactin) suggesting additional complexity within this region beyond the *FADS* genes. For the signals on chromosome 16 identified based on analyses of DGLA in HIS and AFA, in/near *NTAN1* and *PDXDC1*, our integrative PrediXcan analyses identified *PDXDC1* (Pyridoxal Dependent Decarboxylase Domain Containing 1) (but not *NTAN1*) as a putative gene for DGLA. Additionally, having identified association with AA in HIS for the *POLD4* missense variant rs28364240, our subsequent identification of *POLD4* (DNA Polymerase Delta 4, Accessory Subunit) based on the PrediXcan analyses brings additional support for this gene. To follow-up on the genes of interest identified by colocalization and PrediXcan analyses, we examined their co-expression with *FADS1* using GTEx whole blood gene expression. Multiple genes on chromosome 11 identified in our integrative analyses combining the GWAS of PUFAs with whole blood expression from GTEx showed evidence of co-expression with *FADS1*, for example, *TMEM258*, *POLD4*, *TMEM109* and *ZBTB3.* This finding suggests some genomic regions at a considerable distance from *FADS1* may play a role in regulating its expression, and ultimately influence circulating PUFA levels. Downstream pathway analysis of the genes identified by our integrative analyses further highlighted common features of these genes, including their regulation by transcription factors[41] and their relevance to breast cancer[34], UV radiation[42], and cell states or perturbations within the immune system. [43,44] As a recent Mendelian randomization study highlighted the relationship between genetically elevated PUFA levels and risk of cancer,[45] our current work provides further support for that connection.

While our genetic association study of PUFA levels in HIS and AFA provides novel insights, our work has several limitations. First, while we have combined data from multiple CHARGE cohorts, the overall sample size of our study is still relatively small for a GWAS. Second, as we began this GWAS effort some years ago, our work makes use of older imputation panels based on the 1000 Genomes. We expect future work could leverage newer resources including imputation based on the Trans-omics for Precision Medicine (TOPMed) reference panel or newer whole genome sequence data from TOPMed[46]. Third, the circulating PUFA levels examined in this study are derived from heterogeneous sources (plasma phospholipids in MESA and CHS vs. erythrocytes in FHS), which could have resulted in heterogeneity of genetic associations across the included studies and overall loss of power. Finally, while our integration of GWAS with eQTL proved useful in some cases, our efforts were driven in part by the available resources. We made use of multi-ancestry eQTL resources based on purified monocytes in MESA, as we knew these resources were well-matched with our GWAS cohorts in terms of LD structure, although purified monocytes were likely not the most relevant cell type for our study. We complemented those efforts with whole blood eQTL from GTEx through which we were able to capture colocalization of *FADS1* that was not observed in MESA due to the lack of significant cis-eQTL for *FADS1.* This limitation underscores the need for more diverse ancestry eQTL resources across a wider range of tissues and cell types.

In summary, working with the CHARGE Consortium, we conducted a consortium-based GWAS of circulating PUFA levels in HIS and AFA cohorts. Our study demonstrated evidence of shared genetic influences on PUFA levels across race/ancestry groups, and demonstrated a large number of distinct genetic association signals within a neighborhood of the well documented *FADS* region on chromosome 11.[19,20] Our findings provide insight into the complex genetics of circulating PUFA levels that reflect, in part, their response to evolutionary pressures across the course of human history.[47,48] Overall, our study demonstrates the value of investigating complex trait genetics in diverse ancestry populations and highlights the need for continued efforts for expanded genetic association efforts in cohorts with genetic ancestry that reflects that of the general population within the United States and worldwide. In future work, genetic loci identified in this study could be leveraged to examine gene x fatty acid interactions on disease outcomes, or to construct more precise genetic predictors of sub-optimal or deficient fatty acid levels which could be central to efforts in precision nutrition.[17,49] Additionally, we anticipate the results from this work could help to motivate downstream studies focused on fatty acids as a mediator of specific genes' influences on identified pathways, including cancer and immune responses, as well as the long-range regulation of gene function by other genes located in distinct and distant portions of the same chromosome.

3.5 Methods

3.5.1 Study participants

The participants in this study were recruited from three population-based cohorts: the Multi-Ethnic Study of Atherosclerosis (MESA), the Cardiovascular Health Study (CHS) and the Framingham Heart Study (FHS). This manuscript focuses on HIS participants from MESA (N = 1,243) and FHS (N = 211) and AFA participants from MESA (N = 1472), CHS (N = 603) and FHS (N = 203).

MESA is a longitudinal cohort study of subclinical cardiovascular disease and risk factors that predict progression to clinically overt cardiovascular disease or progression of subclinical disease.[50] Between 2000 and 2002, MESA recruited 6,814 men and women 45 to 84 years of age from Forsyth County, North Carolina; New York City; Baltimore; St. Paul, Minnesota; Chicago; and Los Angeles. Participants at baseline were 38% White, 28% African American, 22% Hispanic and 12% Asian (primarily Chinese) ancestry.

CHS is a population-based cohort study of risk factors for coronary heart disease and stroke in adults ≥65 years conducted across four field centers.[51] The original predominantly European ancestry cohort of 5,201 persons was recruited in 1989-1990 from random samples of the Medicare eligibility lists; subsequently, an additional predominantly African-American cohort of 687 persons was enrolled in 1992-1993 for a total sample of 5,888. Analyses were limited to those with available DNA who consented to genetic studies.

FHS is a population-based longitudinal study of families living in Framingham, Massachusetts which originated in 1948 and consisted of individuals of predominantly

European ancestry.[52] In 1994, the Omni Cohort 1 enrolled 507 men and women of African-American, Hispanic, Asian, Indian, Pacific Islander and Native American origins, who at the time of enrollment were residents of Framingham and the surrounding towns.

3.5.2 Fatty Acid measurements

Circulating PUFA levels were quantified from plasma phospholipids in MESA and CHS, and from erythrocytes in FHS. Measurements were taken from biologically independent distinct samples.

MESA: The fatty acids were measured in EDTA plasma, frozen at –70°C.[53] Lipids were extracted from the plasma using a chloroform/methanol extraction method and the cholesterol esters, triglyceride, phospholipids and free fatty acids are separated by thin layer chromatography. The fatty acid methyl esters were obtained from the phospholipids and were detected by gas chromatography flame ionization. Individual fatty acids were expressed as a percent of total fatty acids. A total of 28 fatty acids were identified.

CHS: Blood was drawn after a 12-hour fast and stored at –70°C. Total lipids were extracted from plasma using methods of Folch[54], and phospholipids separated from neutral lipids by one-dimensional TLC. Fatty-acid-methyl-ester (FAME) samples were prepared by direct transesterification using methods of Lepage and Roy[55], and separated using gas chromatography (Agilent5890 gas- chromatograph-FID-detector; Supelco fused-silica 100m capillary column SP-2560; initial 160°C 16 min, ramp 3.0°C/min to 240°C, hold 15 min).[56] Identification, precision, and accuracy were continuously evaluated using model mixtures of known FAMEs and established inhouse controls, with identification confirmed by GC-MS at USDA (Peoria, IL). A total of 42 fatty acids were identified. Fatty acid levels were expressed as percent of total fatty acids. CVs were <3% for most fatty acids.

FHS: Red blood cells (RBCs) were isolated from blood drawn after a 10–12 h fast and frozen at −80 °C immediately after collection. RBC fatty acid composition was analyzed by gas chromatography (GC) with flame ionization detection.[57] Unwashed, packed RBCs were directly methylated with boron trifluoride and hexane at 100 °C for 10 min. The fatty acid methyl esters thus generated were analyzed using a GC2010 Gas Chromatograph (Shimadzu Corporation, Columbia, MD) equipped with an SP2560, fused silica capillary column (Supelco, Bellefonte, PA). Fatty acids were identified by comparison with a standard mixture of fatty acids characteristic of RBC (GLC 727, NuCheck Prep, Elysian, MN) which was also used to determine individual fatty acid response factors. The omega-3 index is defined as the sum of EPA and DHA expressed as a percent of total identified fatty acids. The coefficients of variation were 6.2% for EPA, 4.4% for DHA and 3.2% for the omega-3 index. All fatty acids present at >1% abundance had CVs of ≤7%.

3.5.3 Genotyping and imputation

Each of the participating cohorts had genome-wide genotype data based on a GWAS array, followed by imputation based on the 1000 Genomes Phase 1 v3 (for CHS) or Phase 3 (for MESA and FHS) Cosmopolitan reference panel.[58]

MESA: Participants in the MESA cohort who consented to genetic analyses and data sharing (dbGaP) were genotyped using the Affymetrix Human SNP Array 6.0

(GWAS array) as part of the NHLBI CARe (Candidate gene Association Resource) and SHARe (SNP Health Association Resource) projects. Genotype quality control for these data included filter on SNP level call rate < 95%, individual level call rate < 95%, heterozygosity > 53%.[59] The cleaned genotypic data was deposited with MESA phenotypic data into dbGaP (study accession phs000209.v13.p3); 8,224 consenting individuals (2,685 White, 2,588 non-Hispanic African-American, 2,174 Hispanic, 777 Chinese) were included, with 897,981 SNPs passing study specific quality control (QC). SNP coverage from the original GWAS SNP genotyping array was increased through imputation using the 1,000 Genomes Phase 3 integrated variant set completed using the Michigan Imputation Server.[60,61]

CHS: DNA was extracted from blood samples drawn on all participants at their baseline examination. In 2010, genotyping was performed at the General Clinical Research Center's Phenotyping/Genotyping Laboratory at Cedars-Sinai using the Illumina HumanOmni1-Quad_v1 BeadChip system on African-American CHS participants who consented to genetic testing, and had DNA available for genotyping. Genotyping was attempted in 844 participants, and was successful in 823. Participants were excluded if they had a call rate<=95% or if their genotype was discordant with known sex or prior genotyping (to identify possible sample swaps). Genotype quality control excluded SNPs with a call rate < 97%, HWE P < $1x10^{-5}$, > 1 duplicate error or Mendelian inconsistency (for reference CEPH trios), heterozygote frequency = 0, which resulted in a final set of 963,248 SNPs (940,567 autosomal). Imputation to the 1,000 Genomes Phase I integrated variant set was completed using IMPUTE version 2.2.2. Variants with insufficient effective minor alleles are filtered prior to analysis, with a threshold set at 5 effective alleles resulting in 14,191,388 variants for analysis.

FHS: Direct genotypes were obtained using the Affymetrix 500K and MIPS 50K chips, and were analyzed at the Affymetrix Core Laboratory. Genotype quality control for these data included filter on SNP level call rate < 95%, individual level call rate < 95%, HWE P < 10-5, and genotypes with Mendel errors were set to missing. The cleaned genotypic data consisted of N= 414 (211 Hispanic, 203 African-American) with 628,076 SNPs passing study specific quality control (QC). SNP coverage from the original GWAS SNP genotyping array was increased through imputation using the 1,000 Genomes Phase 3 integrated variant set completed using the Michigan Imputation Server[60,61].

3.5.4 Data transformation and detection of outliers in measured PUFA Levels

After examining the raw phenotype distributions for each of the phenotypes of interest, we applied variable transform for traits exhibiting deviation from normality. Log-transformation was applied for ALA, EPA and GLA. In addition, outliers for all of the PUFA levels were identified by the limits of median +/- 3.5 * MAD', where MAD' is computed with a scale factor constant of 1.4826 [default for the mad() function in R]. The value of MAD' = 1.4826 * MAD0 where MAD0 is the raw value of median absolute deviation (MAD). For all the PUFAs, outliers were winsorized to the value of (median +/- 3.5 * MAD').

3.5.5 Genome-wide Association Study (GWAS) and Meta-analysis

Participants who were not in the self-reported African American or Hispanic

American groups of interest to this manuscript were excluded from the primary GWAS analyses. To construct clean race/ancestry groups for stratified GWAS analyses, self-reported race/ethnicity groups were cleaned by removing outliers for principal components (PCs) of ancestry based on the limited of mean +/- 3.5 * standard deviation, separately for each of the participating cohorts. GWAS was then carried out separately in each cohort and stratified by race/ancestry with covariate adjustment for age, sex, study site and PCs of ancestry. Cohort-specific GWAS results were filtered using EasyQC based on minor allele count (MAC) > 6 and imputation R-squared > 0.3. Cohort-specific results were combined using weighted sum of z-score meta-analysis in METAL[62] and filtered using Effective Heterozygosity Filter (effHET) > 60. A threshold of $P < 5 \times 10^{-8}$ was applied to identify genome-wide significant loci.

3.5.6 Statistical fine-mapping using SuSiE

For each chromosome with more than one genome-wide significant variant (at $P < 5 \times 10^{-8}$), we carried out statistical fine-mapping to identify the putative causal variants and estimate the number of independent signals. We used Sum of Single Effect model (SuSiE)[24] to identify the credible set of putative causal variants, providing as input all variants with $P < 5 \times 10^{-8}$ from the meta-analysis results . For fine-mapping of signals identified in our meta-analysis of HIS and AFA, we used imputed genotype dosage for the same set of variants in MESA HIS and AFA, respectively. To select the parameter L (prior number of independent signals) for fine-mapping in SuSiE, DAP-G (Deterministic Approximation of Posteriors)[63] was conducted to provide a starting value for L based on the number of credible sets that the threshold of posterior inclusion probability was greater than 0.95.

3.5.7 Identification of Novel versus Previously Reported Signals

To distinguish novel versus previously reported signals, we used the results from our previously published CHARGE GWAS n-3 (n=8,866)[19] and n-6 (n=8,631)[20] PUFAs in European ancestry to define the set of known signals. For each trait in the present GWAS effort, credible sets that included at least one variant reported in the previous CHARGE GWAS of the same trait in European ancestry were considered known, while the remaining signals were considered novel in the current study.

3.5.8 Cross-ancestry Replication analysis

Following statistical fine-mapping, cross-ancestry replication analyses were conducted for the most highly supported putative causal variant from each credible set using data on n-3 and n-6 PUFAs from other race/ancestry groups. To do so, we examined results from the prior CHARGE GWAS meta-analysis of European American cohorts (CHARGE EUR), as well as GWAS results of HIS (CHARGE HIS) and AFA (CHARGE AFA) from the present study. As prior GWAS were performed using earlier imputation panels, we further used available measures of n-3 and n-6 PUFAs in self-reported European American (MESA EUR) and Chinese Americans (MESA CHN) from MESA as an additional source of replication having genotype imputation based on 1000 Genomes Phase 3, for consistency with our current work. The resources used for replication analyses were as follows. **European Americans (MESA EUR and CHARGE EUR):** 2344 self-reported European American participants from MESA (using

1000 Genomes Phase 3 imputation, for comparison with the current study), as well as summary statistics from the previously published CHARGE GWAS meta-analysis of n-3 (n=8,866)[19] and n-6 (n=8,631)[20] PUFAs based on imputation from the HapMap Phase I and II; **African Americans (CHARGE AFA):** summary statistics from the present GWAS of PUFAs in AFA to examine cross-ancestry replication of variants identified in the present GWAS of HIS; **Hispanic Americans (CHARGE HIS):** summary statistics from the present GWAS of PUFAs in HIS to examine cross-ancestry replication of variants identified in the present GWAS of AFA; and **Chinese Americans (MESA CHN):** 649 self-reported Chinese American participants from MESA (using 1000 Genomes Phase 3 imputation, for comparison with the current study).

The genetic association analyses performed for replication in each of these studies included covariate adjustment for age, sex, study site (where appropriate) and PCs of ancestry. Multiple testing correction was applied to account for the number of variants examined in cross-ancestry replication (HIS: $P < 0.05/19 = 0.0026$ and AFA: $P < 0.05/11 = 0.004$).

3.5.9 Validation Analysis

Given the limited number of cohorts available for ancestry-specific and cross-ancestry replication of PUFA traits, additional validation analyses were conducted for the same set of variants using multi-ancestry genetic association with lipid traits (HDL, LDL, total cholesterol and triglycerides) from the Global Lipids Genetics Consortium (GLGC).[25] The GLGC aggregated GWAS results of lipid traits from 1,654,960 individuals from 201 primary studies. The genetic ancestry groups include admixed

African or African, East Asian, European, Hispanic and South Asian. The genetic analyses performed by GLGC included covariate adjustment for age, age^2, PCs of ancestry and any necessary study-specific covariates. Multiple testing correction was applied to account for the number of variants examined in cross-ancestry validation (HIS: $P < 0.05/19 = 0.0026$ and AFA: $P < 0.05/11 = 0.004$).

3.5.10 Bayesian colocalization analysis

Bayesian colocalization analysis has proven an effective approach for identification of downstream genes underlying GWAS loci.[35] We used the R/coloc package to conduct Bayesian colocalization analysis[64] to identify the putative gene(s) corresponding to each credible set of variants using MESA multi-ancestry eQTL data from purified monocytes[26] and GTEx multi-ancestry whole blood tissue eQTL data.[65] Bayesian colocalization analysis tested the following hypotheses: H0. neither GWAS of PUFAs nor eQTL has a genetic association in the region (within 1 Mb of the transcription start site); H1. only GWAS of PUFAs has a genetic association in the region; H2. only eQTL has a genetic association in the region; H3. both GWAS of PUFAs and eQTL are associated, but with different causal variants; H4. both GWAS of PUFAs and eQTL are associated and share a single causal variant. Colocalization for variants in credible sets was defined by (1) a posterior colocalization probability of hypothesis 4 (PP.H4) > 0.80, or (2) a PP.H4 > 0.50 *and* the ratio of PP.H4 / PP.H3 > 5.

3.5.11 PrediXcan model.

PrediXcan, a gene-based association method focused on identifying trait-associated genes by quantifying the effect of gene expression on the phenotype on interest.[66] In this study, we applied summary-statistics based PrediXcan (S-PrediXcan)[30] using reference gene expression prediction models from MESA purified monocytes[26] and GTEx multi-ancestry whole blood[30]. S-PrediXcan associations were considered genome-wide significant if they passed the multiple testing correction for all genes (MESA: P < 0.05/4470 = 0.00001 and GTEx: P < 0.05/4350 = 0.00001).

3.5.12 Chromatin Contact Analysis

To identify variants located in open chromatin regions in contact gene promoters, we used GenomicRanges (v. 1.46.1; R version 4.1.1) to intersect the genomic coordinates (hg19) of the variants contained in the credible sets with the open chromatin peaks (called using the ENCODE pipeline) in significantly enriched contact with gene promoter determined by Promoter Capture C (Chicago Score > 5). We queried chromatin accessibility and promoter contacts in human mesenchymal stem cells (hMSC) and Adipocytes differentiated in vitro from these (hMSC_Adipocytes), embryonic stem cell derived hypothalamic neurons (hESC Hypothalamic Neurons), induced pluripotent-dervived Heptocytes (IPS-Hepatocytes), Enteroids, and the hepatic carcinoma HepG2 cell line.[67–72]

3.5.13 Gene Co-expression Analysis.

We used the GTEx whole blood gene expression version 8 TPM dataset to examine co-expression with *FADS1* for genes identified by integrative analyses, including colocalization and PrediXcan. Two models for gene co-expression analysis were used for each expression trait of interest: **Model 1** - an unadjusted model *FADS1* ~ gene expression; and **Model 2** - a covariate adjusted model *FADS1* ~ age + gender + gene expression.

Gene co-expression associations were considered statistically significant if they passed the multiple testing correction for all genes examined from colocalization and PrediXcan ($P < 0.05/39 = 0.0012$).

3.5.14 Gene set enrichment analysis

We applied gene set enrichment analysis for the combined set of genes identified by our integrative analyses (colocalization and PrediXcan) using the Molecular Signature Database (MSigDB) including hallmark gene sets (H), curated gene sets (C2), regulatory target gene sets (C3), computational gene sets (C4), ontology gene sets (C5), oncogenic signature gene sets (C6), immunologic signature gene (C7), cell type signature gene sets (C8). [31–33]

3.5.15 Statistics and Reproducibility

Throughout the manuscript, statistical analyses and reported sample sizes reflect the number of biologically independent samples, with no single individual (person) contributing more than one data point to any given analysis. All *P*-values are presented based on two-sided statistical tests.

3.5.16 Ethical Review

All relevant ethical regulations were followed for the study of human participants.

All MESA, FHS and CHS participants provided written informed consent for participation at their respective study sites, including consent to participate in genetic studies. The MESA, FHS and CHS studies were also reviewed and approved by the Institutional Review Boards (IRBs) at each of the participating study sites. The current investigation including genetic analysis of n-3 and n-6 PUFA levels was reviewed and approved by the Institutional Review Boards (IRB) at the University of Virginia, the University of Washington and the Fatty Acid Research Institute.

3.5.17 Data availability

Genome-wide genotype data for the Multi-Ethnic Study of Atherosclerosis (MESA), the Framingham Heart Study (FHS) and the Cardiovascular Health Study (CHS) are available by application through dbGaP. The dbGaP accession numbers are: MESA phs000209, FHS phs000007 and CHS phs000287. Summary statistics resulting from our GWAS meta-analysis as presented in this manuscript will be available on the CHARGE Summary Results site by application through dbGaP under the accession number phs000930. Summary statistics from the prior CHARGE GWAS of n-3 and n-6 fatty acids[19,20] were obtained from the CHARGE Consortium Results site[73]. Summary statistics from the GLGC GWAS of lipid levels[25] are available publicly[74]. All other data are available from the corresponding author (or other sources, as applicable) on reasonable request.

3.5.18 Code availability

Statistical fine mapping of GWAS loci was conducted using SuSiE[24] as implemented using susieR version 0.12.27.[76] DAP-G[63] was used to choose the starting values for SuSiE and implemented using DAP-G version 1.0.0.[77] Bayesian colocalization analysis[64] was implemented using R/coloc version 5.1.0.1.[78] S-PrediXcan analysis was implemented using S-PrediXcan version 0.6.11.[79] Gene set enrichment analysis was implemented using MSigDB v7.5.1.[80]

3.6 References

1. Calder, P. C., Yaqoob, P., Harvey, D. J., Watts, A. & Newsholme, E. A. Incorporation of fatty acids by concanavalin A-stimulated lymphocytes and the effect on fatty acid composition and membrane fluidity. *Biochemical Journal* **300**, 509–518 (1994).
2. Los, D. A. & Murata, N. Structure and expression of fatty acid desaturases. *Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism* **1394**, 3–15 (1998).
3. Stubbs, C. D. & Smith, A. D. The modification of mammalian membrane polyunsaturated fatty acid composition in relation to membrane fluidity and function. *Biochimica et Biophysica Acta (BBA) - Reviews on Biomembranes* **779**, 89–137 (1984).
4. Mozaffarian, D. & Wu, J. H. Y. Omega-3 fatty acids and cardiovascular disease: effects on risk factors, molecular pathways, and clinical events. *J Am Coll Cardiol* **58**, 2047–2067 (2011).
5. Aung, T. *et al.* Associations of Omega-3 Fatty Acid Supplement Use With Cardiovascular Disease Risks: Meta-analysis of 10 Trials Involving 77 917 Individuals. *JAMA Cardiology* **3**, 225–233 (2018).
6. Simopoulos, A. P. The importance of the omega-6/omega-3 fatty acid ratio in cardiovascular disease and other chronic diseases. *Exp Biol Med (Maywood)* **233**, 674–688 (2008).

7. Krachler, B. *et al.* Fatty acid profile of the erythrocyte membrane preceding development of Type 2 diabetes mellitus. *Nutrition, Metabolism and Cardiovascular Diseases* **18**, 503–510 (2008).

8. Conquer, J. A., Tierney, M. C., Zecevic, J., Bettger, W. J. & Fisher, R. H. Fatty acid analysis of blood plasma of patients with alzheimer's disease, other types of dementia, and cognitive impairment. *Lipids* **35**, 1305–1312 (2000).

9. Söderberg, M., Edlund, C., Kristensson, K. & Dallner, G. Fatty acid composition of brain phospholipids in aging and in Alzheimer's disease. *Lipids* **26**, 421 (1991).

10. Warensjö, E., Sundström, J., Lind, L. & Vessby, B. Factor analysis of fatty acids in serum lipids as a measure of dietary fat quality in relation to the metabolic syndrome in men. *Am J Clin Nutr* **84**, 442–448 (2006).

11. Pizer, E. S. *et al.* Inhibition of Fatty Acid Synthesis Induces Programmed Cell Death in Human Breast Cancer Cells. *Cancer Res* **56**, 2745–2747 (1996).

12. Harris, W. S. *et al.* Blood n-3 fatty acid levels and total and cause-specific mortality from 17 prospective studies. *Nat Commun* **12**, 2329 (2021).

13. Brenna, J. T. Efficiency of conversion of alpha-linolenic acid to long chain n-3 fatty acids in man. *Curr Opin Clin Nutr Metab Care* **5**, 127–132 (2002).

14. Plourde, M. & Cunnane, S. C. Extremely limited synthesis of long chain polyunsaturates in adults: implications for their dietary essentiality and use as supplements. *Appl Physiol Nutr Metab* **32**, 619–634 (2007).

15. Mathias, R. A. *et al.* The impact of FADS genetic variants on ω6 polyunsaturated fatty acid metabolism in African Americans. *BMC Genet.* **12**, 50 (2011).

16. Harris, D. N. *et al.* Evolution of Hominin Polyunsaturated Fatty Acid Metabolism: From Africa to the New World. *Genome Biol Evol* **11**, 1417–1430 (2019).

17. Yang, C. *et al.* Impact of Amerind ancestry and FADS genetic variation on omega-3 deficiency and cardiometabolic traits in Hispanic populations. *Commun Biol* **4**, 918 (2021).

18. Psaty, B. M. *et al.* Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ Cardiovasc Genet* **2**, 73–80 (2009).

19. Lemaitre, R. N. *et al.* Genetic Loci Associated with Plasma Phospholipid n-3 Fatty Acids: A Meta-Analysis of Genome-Wide Association Studies from the CHARGE Consortium. *PLOS Genetics* **7**, e1002193 (2011).

20. Guan, W. *et al.* Genome-Wide Association Study of Plasma N6 Polyunsaturated Fatty Acids within the CHARGE Consortium. *Circ Cardiovasc Genet* **7**, 321–331 (2014).

21. Tintle, N. L. *et al.* A genome-wide association study of saturated, mono- and polyunsaturated red blood cell fatty acids in the Framingham Heart Offspring Study. *Prostaglandins Leukot Essent Fatty Acids* **94**, 65–72 (2015).

22. Kalsbeek, A. *et al.* A genome-wide association study of red-blood cell fatty acids and ratios incorporating dietary covariates: Framingham Heart Study Offspring Cohort. *PLoS One* **13**, e0194882 (2018).

23. Buckley, M. T. *et al.* Selection in Europeans on Fatty Acid Desaturases Associated with Dietary Changes. *Mol Biol Evol* **34**, 1307–1318 (2017).

24. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**, 1273–1300 (2020).

25. Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675–679 (2021).

26. Mogil, L. S. *et al.* Genetic architecture of gene expression traits across diverse populations. *PLoS Genet* **14**, e1007586 (2018).

27. Liu, Y. *et al.* Methylomics of gene expression in human monocytes. *Hum Mol Genet* **22**, 5065–5074 (2013).

28. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

29. Sergeant, S. *et al.* Differences in arachidonic acid levels and fatty acid desaturase (FADS) gene variants in African Americans and European Americans with diabetes or the metabolic syndrome. *Br J Nutr* **107**, 547–555 (2012).

30. Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun* **9**, 1825 (2018).

31. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 15545–15550 (2005).

32. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

33. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).

34. Nikolsky, Y. *et al.* Genome-wide functional synergy between amplified and mutated genes in human breast cancer. *Cancer Res* **68**, 9532–9540 (2008).

35. Kim, S. P. *et al.* Lrp4 expression by adipocytes and osteoblasts differentially impacts sclerostin's endocrine effects on body composition and glucose metabolism. *J Biol Chem* **294**, 6899–6911 (2019).

36. Wu, J. H. Y. *et al.* Omega-6 fatty acid biomarkers and incident type 2 diabetes: pooled analysis of individual-level data for 39 740 adults from 20 prospective cohort studies. *Lancet Diabetes Endocrinol* **5**, 965–974 (2017).

37. Bornfeldt, K. E. Triglyceride lowering by omega-3 fatty acids: a mechanism mediated by N-acyl taurines. *J Clin Invest* **131**, e147558, 147558 (2021).

38. Yuan, X. *et al.* The effects of dietary linoleic acid on reducing serum cholesterol and atherosclerosis development are nullified by a high-cholesterol diet in male and female apoE-deficient mice. *Br J Nutr* **129**, 737–744 (2023).

39. Farvid, M. S. *et al.* Dietary Linoleic Acid and Risk of Coronary Heart Disease: A Systematic Review and Meta-Analysis of Prospective Cohort Studies. *Circulation* **130**, 1568–1578 (2014).

40. Fadason, T., Schierding, W., Lumley, T. & O'Sullivan, J. M. Chromatin interactions and expression quantitative trait loci reveal genetic drivers of multimorbidities. *Nat Commun* **9**, 5198 (2018).

41. Yevshin, I., Sharipov, R., Kolmykov, S., Kondrakhin, Y. & Kolpakov, F. GTRD: a database on gene transcription regulation-2019 update. *Nucleic Acids Res* **47**, D100–D105 (2019).

42. Takao, J., Ariizumi, K., Dougherty, I. I. & Cruz, P. D. Genomic scale analysis of the human keratinocyte response to broad-band ultraviolet-B irradiation. *Photodermatol Photoimmunol Photomed* **18**, 5–13 (2002).

43. Lund, R., Aittokallio, T., Nevalainen, O. & Lahesmaa, R. Identification of novel genes regulated by IL-12, IL-4, or TGF-beta during the early polarization of CD4+ lymphocytes. *J Immunol* **171**, 5328–5336 (2003).

44. Ochiai, K. *et al.* Transcriptional regulation of germinal center B and plasma cell fates by dynamical control of IRF4. *Immunity* **38**, 918–929 (2013).

45. Fatty Acids in Cancer Mendelian Randomization Collaboration *et al.* The association between genetically elevated polyunsaturated fatty acids and risk of cancer. *EBioMedicine* **91**, 104510 (2023).

46. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

47. Mathias, R. A. *et al.* Adaptive evolution of the FADS gene cluster within Africa. *PLoS ONE* **7**, e44926 (2012).

48. Ameur, A. *et al.* Genetic adaptation of fatty-acid metabolism: a human-specific haplotype increasing the biosynthesis of long-chain omega-3 and omega-6 fatty acids. *Am J Hum Genet* **90**, 809–820 (2012).

49. Chilton, F. H. *et al.* Interpreting Clinical Trials With Omega-3 Supplements in the Context of Ancestry and FADS Genetic Variation. *Front Nutr* **8**, 808054 (2021).

50. Bild, D. E. *et al.* Multi-Ethnic Study of Atherosclerosis: objectives and design. *Am. J. Epidemiol.* **156**, 871–881 (2002).

51. Fried, L. P. *et al.* The Cardiovascular Health Study: design and rationale. *Ann Epidemiol* **1**, 263–276 (1991).

52. Mahmood, S. S., Levy, D., Vasan, R. S. & Wang, T. J. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet* **383**, 999–1008 (2014).

53. Cao, J., Schwichtenberg, K. A., Hanson, N. Q. & Tsai, M. Y. Incorporation and clearance of omega-3 fatty acids in erythrocyte membranes and plasma phospholipids. *Clin. Chem.* **52**, 2265–2272 (2006).

54. Folch, J., Lees, M. & Sloane Stanley, G. H. A simple method for the isolation and purification of total lipides from animal tissues. *J Biol Chem* **226**, 497–509 (1957).

55. Lepage, G. & Roy, C. C. Direct transesterification of all classes of lipids in a one-step reaction. *J Lipid Res* **27**, 114–120 (1986).

56. Mozaffarian, D. *et al.* Circulating palmitoleic acid and risk of metabolic abnormalities and new-onset diabetes. *Am J Clin Nutr* **92**, 1350–1358 (2010).

57. Harris, W. S., Pottala, J. V., Vasan, R. S., Larson, M. G. & Robins, S. J. Changes in erythrocyte membrane trans and marine fatty acids between 1999 and 2006 in older Americans. *J Nutr* **142**, 1297–1303 (2012).

58. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

59. Manichaikul, A. *et al.* Association of SCARB1 variants with subclinical atherosclerosis and incident cardiovascular disease: the multi-ethnic study of atherosclerosis. *Arterioscler Thromb Vasc Biol* **32**, 1991–1999 (2012).

60. Fuchsberger, C., Forer, L., Schonherr, S., Das, S. & Abecasis, G. Michigan Imputation Server. https://imputationserver.sph.umich.edu.

61. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284–1287 (2016).

62. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

63. Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient Integrative Multi-SNP Association Analysis via Deterministic Approximation of Posteriors. *Am J Hum Genet* **98**, 1114–1129 (2016).

64. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet* **10**, e1004383 (2014).

65. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).

66. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**, 1091–1098 (2015).

67. Chesi, A. *et al.* Genome-scale Capture C promoter interactions implicate effector genes at GWAS loci for bone mineral density. *Nat Commun* **10**, 1260 (2019).

68. Pahl, M. C. *et al.* Cis-regulatory architecture of human ESC-derived hypothalamic neuron differentiation aids in variant-to-gene mapping of relevant complex traits. *Nat Commun* **12**, 6749 (2021).

69. Lasconi, C. *et al.* Variant-to-Gene-Mapping Analyses Reveal a Role for the Hypothalamus in Genetic Susceptibility to Inflammatory Bowel Disease. *Cell Mol Gastroenterol Hepatol* **11**, 667–682 (2021).

70. Hammond, R. K. *et al.* Biological constraints on GWAS SNPs at suggestive significance thresholds reveal additional BMI loci. *Elife* **10**, e62206 (2021).

71. Çalışkan, M. *et al.* Genetic and Epigenetic Fine Mapping of Complex Trait Associated Loci in the Human Liver. *Am J Hum Genet* **105**, 89–107 (2019).

72. Ramdas, S. *et al.* A multi-layer functional genomic analysis to understand noncoding genetic variation in lipids. *Am J Hum Genet* **109**, 1366–1387 (2022).

73. CHARGE Consortium. CHARGE Consortium Results. https://www.chargeconsortium.com/main/results.

74. Global Lipids Genetics Consortium. Trans-ancestry GWAS summary statistics for HDL-C, LDL-C, nonHDL-C, TC and C. https://csg.sph.umich.edu/willer/public/glgc-lipids2021/results/trans_ancestry/.

75. Mogil, L. S. *et al.* Genetic architecture of gene expression traits across diverse populations. *PLoS Genet.* **14**, e1007586 (2018).

76. Wang, G, Carbonetto, P, Zou, Y, Zhang, Kaiqian & Stephens, M. susieR (version 0.12.27). https://github.com/stephenslab/susieR/releases/tag/v0.12.27.

77. Wen, X. DAP-G v1.0.0. https://github.com/xqwen/dap/releases/tag/v1.0.0.

78. Wallace, C. R/coloc. https://github.com/chr1swallace/coloc.

79. Im, H. MetaXcan / S-PrediXcan software. https://github.com/hakyimlab/MetaXcan.

80. Liberzon, A. MSigDB 7.5.1. https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/MSigDB_v7.5.1_Release_Notes.

3.7 Acknowledgments

This work was supported by R01 HL105756 to the CHARGE Consortium, NCCIH R01 AT008621 (to FHC), and by the Daniel B. Burke Endowed Chair for Diabetes Research (to SFG). The authors acknowledge Research Computing at The University of Virginia for providing computational resources and technical support that have contributed to the results reported within this publication.

The Multi-Ethnic Study of Atherosclerosis: MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420, UL1-TR-001881, and DK063491. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278. Genotyping was performed at Affymetrix (Santa Clara, California, USA) and the Broad Institute of Harvard and MIT (Boston, Massachusetts, USA) using the Affymetrix Genome-Wide Human SNP Array 6.0.

The Cardiovascular Health Study: This research was supported by contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, 75N92021D00006, and grants U01HL080295 and U01HL130114 from the National Heart, Lung, and Blood Institute (NHLBI), with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org. Fatty acid measurements were supported by R01 HL085710. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR000124, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.. Additional support for infrastructure was provided by R01HL105756 and additional genotyping among the African-American cohort was supported in part by R01HL085251. DNA handling and genotyping at Cedars-Sinai Medical Center was supported in part by National Center for Research Resources grant UL1RR033176, now at the National Center for Advancing Translational Technologies CTSI grant UL1TR000124; in addition to the National Institute of Diabetes and Digestive and Kidney Diseases grant P30DK063491 to the Southern California Diabetes Endocrinology Research Center.

The Framingham Heart Study: This research was partially supported by the National Human Genome Research Institute (R15HG006915; R01HL152215; NT). The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (Contract No. N01-HC-25195, HHSN268201500001I and 75N92019D00031). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. Data for this project was obtained via dbGaP (phs000007.v32.p1).

Funding to support the Omni cohort recruitment, retention and examination was provided by NHLBI Contract N01-HC-25195, HHSN268201500001I and 75N92019D00031, as well as NHLBI grants R01-HL070100, R01-HL076784, R01-HL49869, and U01-HL-053941. Funding for SHARe Affymetrix genotyping was provided by NHLBI Contract N02-HL64278. SHARe Illumina genotyping was provided under an agreement between Illumina and Boston University. Funding for Affymetrix genotyping of the FHS Omni cohorts was provided by Intramural NHLBI funds from Andrew D. Johnson and Christopher J. O'Donnell. Funding support for the Framingham Red Blood Cell Fatty Acids dataset was provided by NHLBI grant R01 HL089590.

3.8 Author Contributions

YDIC, LMS, CDB, DS, MYT, SSR, DM, and SFAG generated the data. CY, JV, TMB, MCP, JW, LMS, SSR, CES, TDO, DM, SFAG, NLT, RNL, and AM analyzed the data. CY, LMS, SSR, CES, DM, NLT, RNL and AM conceptualized and designed the study. CDB, MYT, ACW, SSR, CES, DM, SFAG, NLT, RNL and AM provided critical oversight to data collection and study coordination. CY, MCP, BH, SSR, TDO, FHC, NLT, RNL and AM wrote the manuscript. All authors contributed to critical editing of the manuscript.

3.9 Competing Interests

The authors declare no competing interests.

Table 3.1. CHARGE cohort descriptives.

	MESA/Hispanic Americans	**FHS/Hispanic Americans**	**MESA/African Americans**	**CHS/African Americans**	**FHS/African Americans**
Participant characteristics					
No. subjects	1243	211	1472	603	203
Women	629 (50.6)	129 (61.1)	788 (53.5)	390 (64.7)	130 (64.0)
Age, years	61 [53, 69]	53 [44, 60]	63 [53, 70]	74 [71, 79]	58 [50, 67]
n-3 Polyunsaturated Fatty Acids					
ALA (% of total fatty acids)	0.16 [0.12, 0.20]	0.21 [0.16, 0.27]	0.15 [0.12, 0.19]	0.13 [0.11, 0.17]	0.18 [0.15, 0.23]
EPA	0.53 [0.37, 0.74]	0.57 [0.47, 0.78]	0.68 [0.51, 0.98]	0.53 [0.39, 0.67]	0.68 [0.48, 1.01]
DPA	0.86 [0.73, 1.00]	2.49 [2.13, 2.79]	0.93 [0.80, 1.07]	0.85 [0.75, 0.97]	2.54 [2.25, 2.89]
DHA	2.96 [2.29, 3.77]	4.21 [3.45, 5.13]	4.05 [3.25, 4.95]	3.46 [2.87, 4.17]	5.23 [4.21, 6.47]
n-6 Polyunsaturated Fatty Acids					
LA	20.92 [18.87, 23.07]	14.32 [12.24, 16.76]	18.88 [17.12, 20.84]	17.84 [16.46, 19.40]	12.53 [10.88, 15.16]
GLA	0.11 [0.08, 0.14]	0.15 [0.10, 0.18]	0.10 [0.08, 0.13]	0.07 [0.05, 0.09]	0.10 [0.07, 0.15]
DGLA	3.57 [3.04, 4.13]	1.95 [1.63, 2.35]	2.89 [2.47, 3.33]	2.76 [2.39, 3.24]	1.51 [1.32, 1.78]
AA	11.01 [9.37, 12.84]	16.56 [15.17, 17.74]	13.21 [11.65, 14.82]	12.64 [11.57, 13.86]	17.17 [15.95, 18.48]

Table 3.1 shows the participant characteristics of the Hispanic Americans and African Americans from each cohort (MESA, CHS and FHS). Data are presented as n (%) for binary measures or median [IQR] for continuous measures. Summary statistics are reported for the subset of individuals with data available for at least one of the fatty acid traits examined in genetic analyses. Fatty acids were measured in plasma phospholipids in MESA and CHS and in erythrocytes in FHS.

Table 3. 2. Genome-wide significant signals (Credible sets) for PUFAs in CHARGE Hispanic Americans.

	Lead variant (Chr:Pos:EFF:OTH)	EAF	Zscore	P-value	Cluster	# Of SNP	Novel/ Known	*Nearest Gene*
AA	rs102274 (11:61557826:C:T)	0.506	-24.26	5.1E-130	1	7	Known	*TMEM258*
	rs142068305 (11:67065755:T:G)	0.196	-7.06	1.63E-12	2	1	Novel	*ANKRD13D*
	rs28364240 (11:67120530:G:C)	0.204	-7.04	1.88E-12	3	1	Novel	*POLD4*
	rs2668898 (11:61725498:G:A)	0.402	-5.83	5.32E-09	4	1	Known	*BEST1*
	rs180792704 (11:67325239:C:G)	0.199	-7.56	3.81E-14	5	1	Novel	NA
	rs198434 (11:61483417:A:G)	0.710	-8.97	2.80E-19	6	1	Novel	*DAGLA*
	rs518804 (11:57494487:C:A)	0.420	-7.73	1.01E-14	7	1	Novel	*TMX2*
	rs3177514 (11:66130358:G:T)	0.699	-5.60	2.06E-08	8	1	Novel	*SLC29A2*
ALA	rs174562 (11:61585144:G:A)	0.503	7.84	4.30E-15	1	23	Known	*FADS1*
DGLA	rs174538 (11:61560081:A:G)	0.488	14.70	6.03E-49	1	1	Known	*TMEM258*
	rs174585 (11:61611694:A:G)	0.274	9.82	8.72E-23	2	1	Known	*FADS2*
	rs198434 (11:61483417:A:G)	0.710	6.27	3.57E-10	3	1	Novel	*DAGLA*
	rs198461 (11:61524366:C:A)	0.363	-5.95	2.54E-09	4	1	Novel	*MYRF*
	rs57112407 (15:78088914:T:C)	0.255	-5.86	4.46E-09	NA	NA	Novel	*LINGO1*
	rs4985155 (16:15129459:G:A)	0.524	-7.72	1.16E-14	1	25	Known	*PDXDC1*
DPA	rs1535 (11:61597972:G:A)	0.520	-11.31	1.07E-29	1	18	Known	*FADS2*
	rs198434 (11:61483417:A:G)	0.710	-6.26	3.67E-10	2	1	Novel	*DAGLA*
EPA	rs102274 (11: 61557826:C:T)	0.506	-11.56	6.18E-31	1	17	Known	*TMEM258*
GLA	rs174576 (11: 61603510:A:C)	0.546	-7.73	1.07E-14	1	19	Known	*FADS2*
LA	rs174564 (11:61588305:G:A)	0.520	15.11	1.23E-51	1	10	Known	*FADS2*
	rs10751002 (11:63617634:G:T)	0.664	6.06	1.36E-09	2	1	Novel	*MARK2*
	rs2668898 (11:61725498:G:A)	0.402	5.54	2.99E-08	3	1	Known	*BEST1*
	rs28364240 (11:67120530:G:C)	0.204	5.90	3.44E-09	4	1	Novel	*POLD4*
	rs11039018 (11:46909524:A:C)	0.67	-6.10	1.01E-09	5	1	Novel	*LRP4*
	rs518804 (11:57494487:C:A)	0.420	6.03	1.62E-09	6	1	Novel	*TMX2*

Table 3.2 shows the signals (credible sets) of putative causal variants identified for each of the PUFAs by fine mapping using SuSiE in HIS (n = 1,454). All variant positions are presented based on Human Genome Build 37. Variants previously documented in the CHARGE GWAS meta-analysis of n-3 and n-6 PUFAs were considered known prior to the current meta-analysis. The remaining variants were considered novel in the current study. There was only one genome-wide significant variant on

chromosome 15 for DGLA (rs57112407) in HIS, and this signal was not carried forward for fine-mapping. *P*-values are calculated using a two-sided test for the z-score derived by meta-analysis including a total of n = 1454 biologically independent samples.

Table 3.3. Genome-wide significant signals (Credible sets) for PUFAs in CHARGE African Americans.

	Lead variant (Chr:Pos:EFF:OTH)	EAF	Zscore	P-value	Clus ter	# Of SNP	Novel/ Known	Nearest Gene
AA	rs174585 (11:61611694:A:G)	0.060	-9.32	1.08E-20	1	1	Known	*FADS2*
	rs174607 (11:61627321:C:G)	0.078	-6.49	8.47E-11	2	1	Known	*FADS2*
	rs174564 (11:61588305:G:A)	0.133	-14.85	6.43E-50	3	1	Known	*FADS2*
	rs174559 (11:61581656:A:G)	0.078	-13.68	1.27E-42	4	1	Known	*FADS1*
	rs17161592 (7:9388418:C:G)	0.085	-6.31	2.75E-10	1	2	Novel	NA
DGLA	rs174560 (11:61581764:C:T)	0.216	9.12	7.51E-20	1	1	Known	*FADS1*
	rs1136001 (16:15131974:T:G)	0.220	-6.11	9.69E-10	2	17	Known	*PDXDC1*
DPA	rs717894 (6:22119292:A:G)	0.250	-5.48	4.11E-08	1	1	Novel	*CASC15*
	rs9295741 (6:10997166:T:C)	0.223	5.54	2.89E-08	2	2	Known	*ELOVL2*
DHA	rs114622288 (10:14663844:A:G)	0.050	-5.71	1.16e-08	NA	NA	Novel	*FAM107B*
LA	rs1535 (11:61597972:G:A)	0.163	7.88	3.14E-15	1	2	Known	*FADS2*

Table 3.3 shows the signals (credible sets) of putative causal variants identified for each of the PUFAs by fine-mapping using SuSiE in AFA (n = 2,278). All variant positions are presented based on Human Genome Build 37. Variants previously documented in the CHARGE GWAS meta-analysis of n-3 and n-6 PUFAs were considered known prior to the current meta-analysis. The remaining variants were considered novel in the current study. There was only one genome-wide significant variant on chromosome 10 for DHA (rs114622288) in AFA, and this signal was not carried forward for fine-mapping. *P*-values are calculated using a two-sided test for the z-score derived by meta-analysis including a total of n = 2278 biologically independent samples.

Table 3.4. Novel PUFA-associated signals (credible sets) from analysis of HIS with external cross-ancestry replication or multi-ancestry validation evidence.

Traits	Variants (chr:pos:effect:other)	Discovery	Replication	Validation	Direction	Nearest Gene
AA	rs518804 (11:57494487:C:A)	HIS: *P* = 1.01E-14	NS	HDL: *P* = 1.96E-06 logTG: *P* = 0.001	HIS: (-) HDL: (-) logTG: (+)	*TMX2*
	rs198434 (11:61483417:A:G)	HIS: *P* = 2.80E-19	NS	logTG: *P* = 1.65E-03	HIS: (-) logTG: (+)	*DAGLA*
DGLA	rs198461 (11:61524366:C:A)	HIS: *P* = 2.54E-09	EUR: *P* = 7.37E-09	HDL: *P* = 4.81E-13 LDL: *P* = 1.92E-13 logTG: *P* = 1.19E-18 TC: *P* = 5.63E-14	HIS: (-) EUR: (-) HDL: (+) LDL: (+) logTG: (-) TC: (+)	*MYRF*
	rs198434 (11:61483417:A:G)	HIS: *P* = 3.57E-10	EUR: *P* = 2.54E-03	logTG: *P* = 1.65E-03	HIS: (+) EUR: (+) logTG: (+)	*DAGLA*
DPA	rs198434 (11:61483417:A:G)	HIS: *P* = 3.67E-10	NS	logTG: *P* = 1.65E-03	HIS: (-) logTG: (+)	*DAGLA*
LA	rs518804 (11:57494487:C:A)	HIS: *P* = 1.62E-09	EUR: P = 2.50E-03	HDL: *P* = 1.96E-06 logTG: *P* = 0.001	HIS: (+) EUR: (-) HDL: (-) logTG: (+)	*TMX2*
	rs10751002 (11:63617634:G:T)	HIS: *P* = 1.36E-09	NS	LDL: *P* = 3.31E-12 TC: *P* = 5.74E-09	HIS: (+) LDL: (+) TC: (+)	*MARK2*
	rs1039018 (11:46909524:A:C)	HIS: *P* = 1.01E-09	NS	HDL: *P* = 2.85E-74 logTG: *P* = 4.5E-43	HIS: (+) HDL: (+) logTG: (-)	*LRP4*

Table 3.4 shows the novel putative causal variants in each signal (credible set) identified from fine-mapping for PUFAs with replication and validation evidence in HIS (n = 1,454). All variant positions are presented based on Human Genome Build 37. Variants that were not previously documented in the CHARGE GWAS meta-analysis of n-3 and n-6 PUFAs and were not in LD with known GWAS variants were considered novel in the current study. *P*-values corresponding to discovery (in HIS) and replication (in EUR) are calculated using a two-sided test for the z-score derived by meta-analysis including a total of n = 1,454 or n=2,344 biologically independent samples, respectively. Validation *P*-values are extracted directly from the GWAS summary statistics corresponding to the GLGC publication.[133]

Table 3.5. Integrative analysis (Colocalization and PrediXcan) in the Hispanic Americans using multi-ancestry resources from MESA and GTEx.

	Colocalization Analysis		**PrediXcan**	
	MESA multi-ancestry eQTLs	**GTEx eQTLs**	**MESA**	**GTEx**
AA	Chromosome 11			
	MED19, TMEM258, PACS1, RAD9A	*RPS4XP13, AP001462.6*	*TMEM258, TMEM109, ZBTB3, TTC9C, FERMT3, MED19, POLD4, CLCF1, INCENP, MADD, SSH3, C11orf24, PRPF19, TBC1D10C, BANF1, CCDC86, NXF1, MS4A6E, CCS, COX8A, CCDC88B, ACP2, MAP4K2*	*TMEM258, TMEM223, NXF1, INCENP, MUS81, C11orf84, MED19, MEN1, BBS1, NEAT1, DPP3, SSH3, ACP2, ASRGL1, RNASEH2C*
ALA	Chromosome 11			
	TMEM258, MED19	*MED19, PGA5, TMEM258*	*TMEM258, TMEM109*	*TMEM258*
DGLA	Chromosome 11			
	TMEM258		*TMEM258, ZBTB3*	*TMEM258, FADS1, FADS2*
	Chromosome 16			
	PDXDC1	*RP11-426C22.5*	*PDXDC1*	*NPIPA2*
DPA	Chromosome 11			
	TMEM258, C11orf24, RAD9A	*PGA5*	*TMEM258, TMEM109*	*TMEM258, SSH3, TMEM223*
EPA	Chromosome 11			
	TMEM258	*TPCN2*	*TMEM258, FERMT3, TMEM109*	*TMEM258, SSH3, TMEM223*
GLA	Chromosome 11			
	TMEM258	*MEN1*	*TMEM258*	*TMEM258*
	Chromosome 11			
LA	*MED19, CTTN, C11orf24, RAD9A*	*MED19, TPCN2, FADS1, RPS4XP13, AP001462.6*	*TMEM258, TMEM109, FERMT3, ZBTB3, COX8A, MADD, POLD4, TBC1D10C, INCENP, TTC9C, MED19, CLCF1, SSH3, ACP2*	*TMEM258, INCENP, SSH3, C11orf84, TMEM223, GIF, NXF1, MED19, MUS81, ACP2*

Table 3.5 shows the results of integrative analysis including colocalization analysis and PrediXcan in HIS by using MESA and GTEx eQTL data. For colocalization analysis, eQTL resources include MESA multi-ancestry eQTL from purified monocytes and GTEx European ancestry whole blood tissue eQTL. GWAS signals with posterior colocalization probability of hypothesis 4 (PP.H4) > 0.80, or PP.H4 > 0.50 and the ratio of PP.H4 / PP.H3 > 5 were considered colocalized with eQTL. For PrediXcan, reference gene expression prediction models include MESA purified monocytes and GTEx European ancestry whole blood, and multiple testing correction was applied across all genes tested (MESA: $P < 0.05/4470 = 0.00001$ and GTEx: $P < 0.05/4350 = 0.00001$).

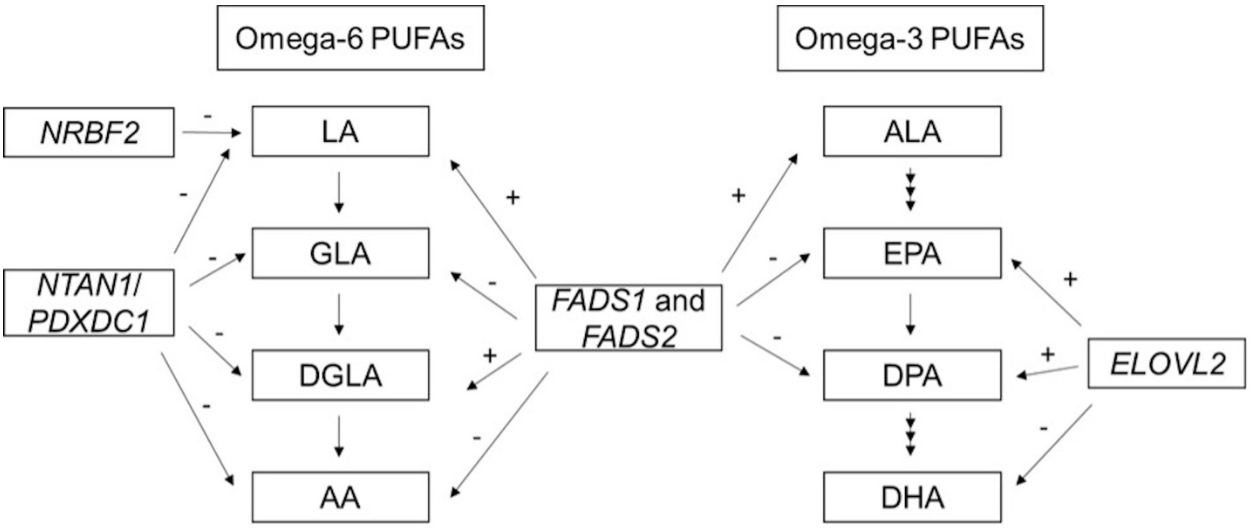


Figure 3.1 PUFAs metabolic pathway and summary of genome-wide association from previous CHARGE GWAS of n-3 and n-6 PUFAs in European Americans. The summary of results from previous CHARGE GWAS of n-3 and n-6 PUFAs in European Americans. + and − signs indicate the direction of the associations for the minor allele of the most significant variant at each locus. The variants used to determine the directions of effect at each locus are as follows:
FADS1 and *FADS2*: rs174547 (ALA, DPA, LA, GLA, DGLA and AA); rs174538 (EPA)
ELOVL2: rs780094 (DPA); rs3798713 (EPA); rs2236212 (DHA)
NTAN1/PDXDC1: rs16966952 (LA, GLA, DGLA and AA)
NRBF2: rs10740118 (LA).

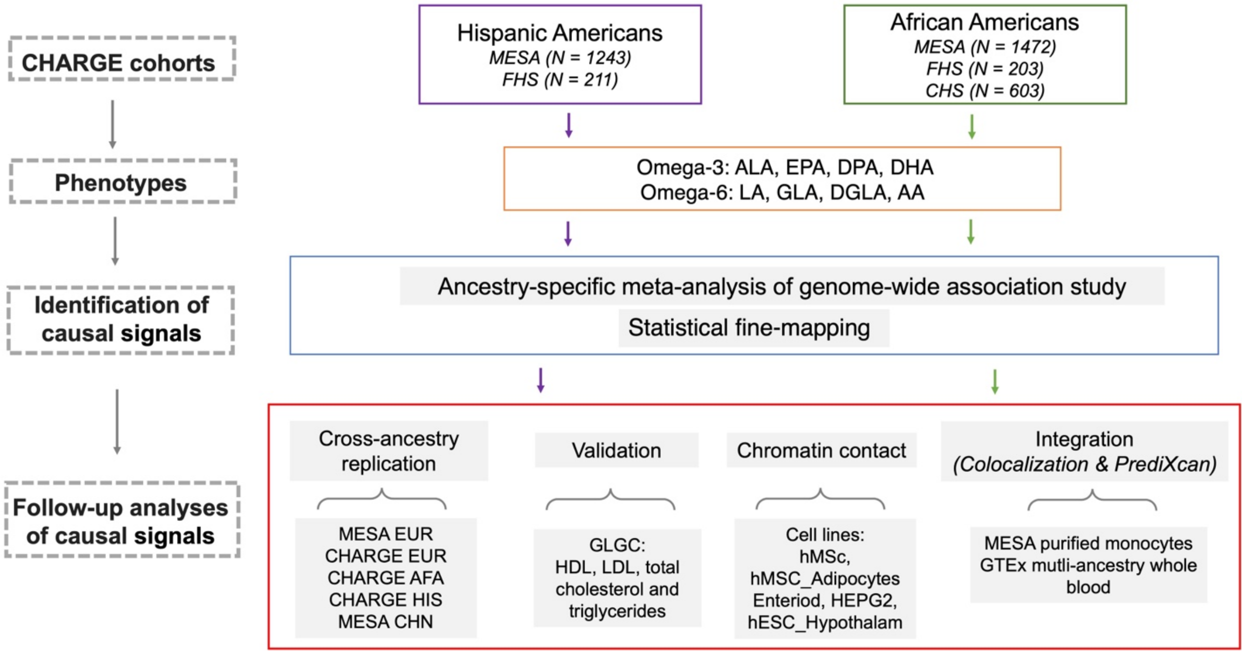

CHARGE cohorts
Phenotypes
Identification of causal signals
Follow-up analyses of causal signals
Hispanic Americans
MESA (N = 1243)
FHS (N = 211)
African Americans
MESA (N = 1472)
FHS (N = 203)
CHS (N = 603)
Omega-3: ALA, EPA, DPA, DHA
Omega-6: LA, GLA, DGLA, AA
Ancestry-specific meta-analysis of genome-wide association study
Statistical fine-mapping
Cross-ancestry replication
Validation
Chromatin contact
Integration (Colocalization & PrediXcan)
MESA EUR
CHARGE EUR
CHARGE AFA
CHARGE HIS
MESA CHN
GLGC:
HDL, LDL, total cholesterol and triglycerides
Cell lines:
hMSc,
hMSC_Adipocytes
Enteriod, HEPG2,
hESC_Hypothalam
MESA purified monocytes
GTEx mutli-ancestry whole blood

Figure 3.2. Study design. GWAS of PUFAs was applied for each cohort stratified by HIS and AFA. Ancestry-specific GWAS meta-analysis and statistical fine-mapping were applied separately for HIS and AFA to identify the potential causal signals. Multiple follow-up analyses were conducted for the causal signals, including cross-ancestry replication, validation, chromatin contact analysis and integrative analyses.

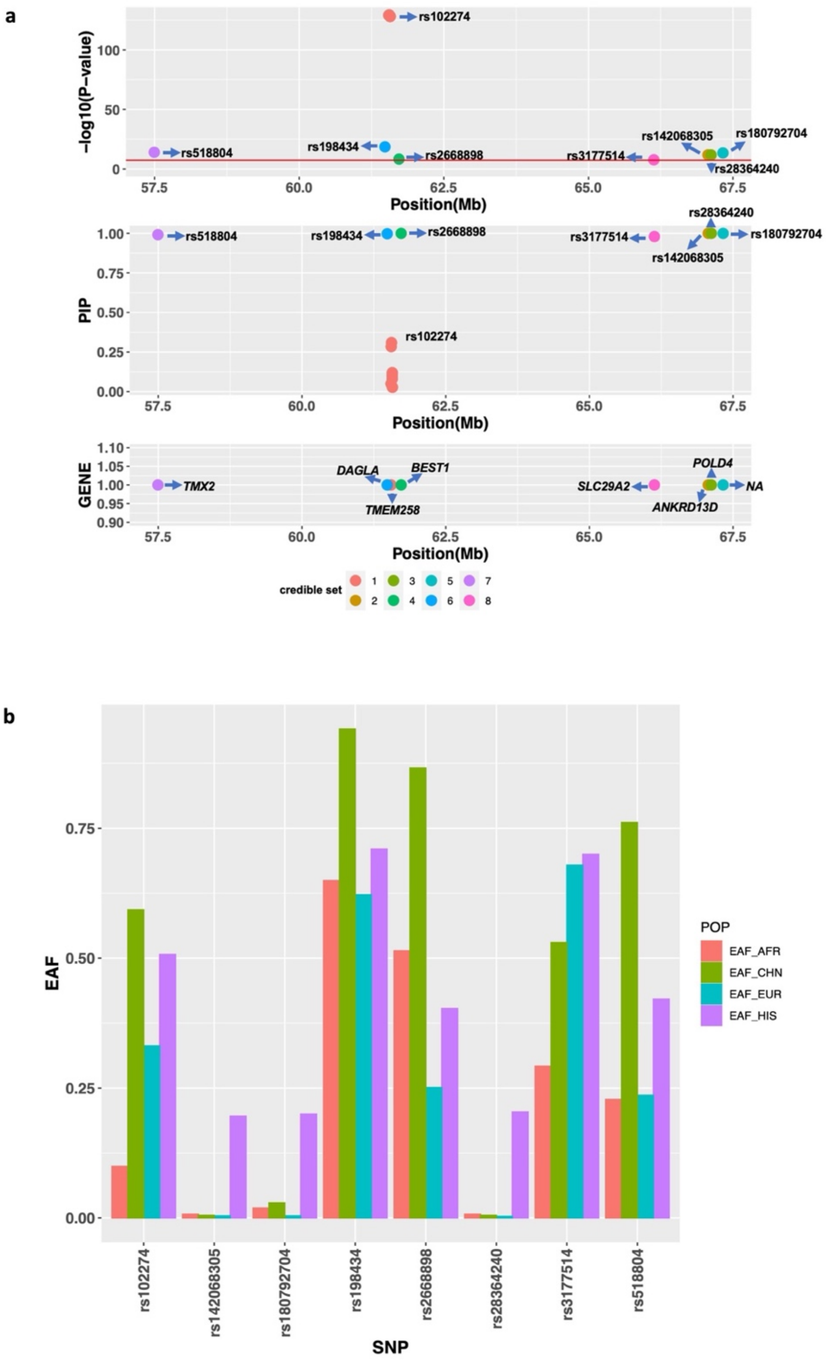


Figure 3.3. Summary of signals (credible sets) identified in association with AA on chromosome 11 in Hispanic Americans. Panel (a) shows detailed information for the identified signals. The upper display shows the *P*-value of the putative causal variants of each signal (credible set) on chromosome 11 from GWAS based on data for a total of n = 1,454 biologically independent samples; middle display shows the Posterior Inclusion Probability (PIP) of the putative causal variants from statistical fine-mapping using SuSIE; bottom display shows the Gene near/in the putative causal variants of each signal. Panel (b) shows the effect allele frequencies (EAF) in MESA across four self-reported race/ethnic groups (African American [n = 2,278], Chinese [n = 648], Hispanic American [n = 1,454], and European ancestry [n = 2,344]) for the most significant putative causal variant from each signal (credible set). Source data for the figure are provided in Supplementary Data 7.

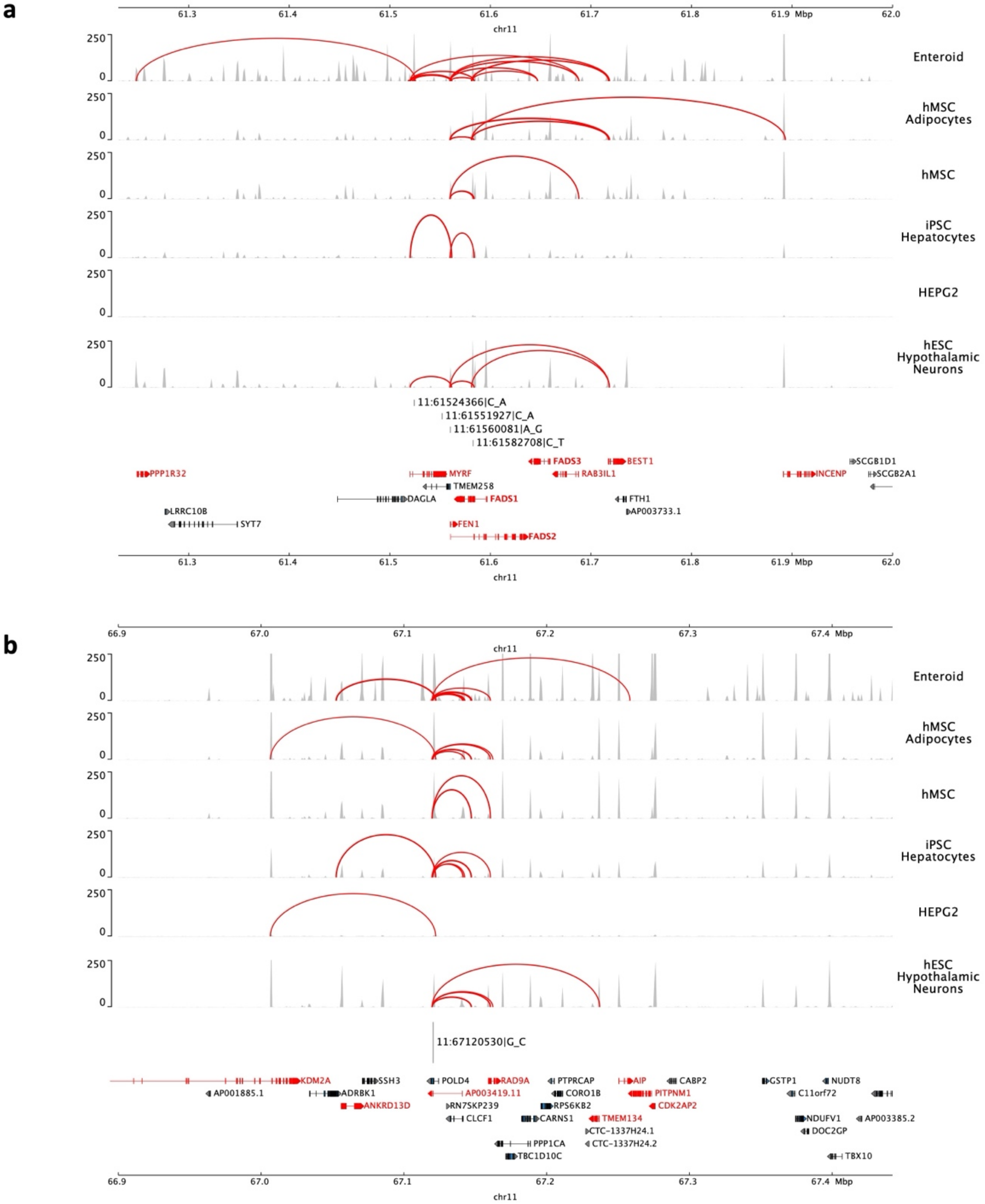


Figure 3.4. Chromatin contact analysis of selected genome-wide significant variants identified on Chromosome 11. The chromatin contacts for the putative causal variants within the selected signals (Figure 4a: *FADS* region and Figure 4b: *POLD4* region) located in open chromatin defined by ATAC-seq with gene promoters defined by Promoter Capture C (implicated genes highlighted in red) in multiple metabolic-relevant cell types. The cell types examined include: human mesenchymal stem cells (hMSC), which were also differentiated in vitro to adipocytes (hMSC_Adipocytes), induced pluripotent stem cell derived Hepatocytes (iPSC_Hepatocytes), embryonic stem cell derived Hypothalamic Neurons (hESC_HypothalamicNeurons), Enteroids, and HepG2s. The y axis shows the ATAC-seq read density normalized using the reads per genomic content (RPGC) method. All variant positions are presented based on Human Genome Build 37.

Chapter 4

Systematic Integration of Multi-omics Data for the Study of Coronary Artery Disease and Subclinical Atherosclerosis

Chaojie Yang, Francois Aguet, Kristin Ardlie, Robert Gerszten, Wendy S. Post, Heather Wheeler, Kent Taylor, Silva Kasela, Tuuli Lappalainen, Hae Kyung Im, Peter Durda, Craig Johnson, Xiuqing Guo, Yongmei Liu, Joseph Polak, David Herrington, Clary Clish, David Van Den Berg, Russell P. Tracy, Elaine Cornell, Tom Blackwell, George Papanicolaou, Stefan Bekiranov, Coleen A. McNamara, Clint L. Miller, Jerome I. Rotter, Stephen S. Rich, Ani Manichaikul

This project was award for American Heart Association Predoctoral Fellowship and was selected for oral presentation in CHARGE Seattle meeting and poster presentation in ASHG meeting.

The text for this chapter is pending co-author review and submission to a journal.

4.1 Abstract

Coronary artery disease (CAD) is a leading cause of death and disability worldwide. Prior genome-wide association studies (GWAS) of CAD have identified over 300 independent loci. However, the molecular consequences of the multi-ancestry GWAS variants and their relevance to subclinical atherosclerosis have not been explored comprehensively in human cohorts. In this study, we applied Bayesian colocalization analysis to overlap results from multi-ancestry CAD GWAS with transcriptomic data from the Trans-Omics for Precision Medicine (TOPMed) Multi-Ethnic Study of Atherosclerosis (MESA). Combined with additional follow-up validation approaches, we prioritized 6 candidate genes (*DDX59, CAMSAP2, AC018816.3, BICC1, EIF2B2* and *PLEKHJ1*) associated with CAD and subclinical atherosclerosis traits (CAC, IMT and carotid plaque). We also conducted weighted gene co-expression network analysis (WGCNA) on the TOPMed MESA transcriptomics data followed by regression analyses to identify 9 modules related to subclinical atherosclerosis. These modules were enriched in pathways related to the up-regulation of genes in response to IFNG and alpha interferon proteins, as well as up-regulation by IL6 via STAT3 during acute phase response. We further identified one module from WGCNA significantly enriched for genes colocalized CAD GWAS signals. One of colocalized genes, *RBC*, was also identified as the only plaque-related hub gene for this module. Our findings enhance our understanding of genetic mechanisms and pathways implicated for CAD and subclinical atherosclerosis and consequently provide valuable insights into potential therapeutic interventions and treatments for CAD.

4.2 Introduction

Coronary artery disease (CAD) is a common complex disease with both genetic and environmental determinants and is a leading cause of death and disability worldwide.[1,2] The primary cause of CAD is atherosclerosis, characterized by the accumulation of plaque within the coronary arteries. Atherosclerotic plaque is composed of a complex mixture, including cholesterol, fatty deposits, calcium deposits, and the clot-making substance fibrin.[3–7] In addition, the identification of atherosclerotic lesions within the carotid arteries demonstrates a strong correlation with CAD.[8–10] Notably, the quantification of coronary artery calcifications (CAC) through a calcium score serves as a widely acknowledged marker indicative of CAD. Multiple studies demonstrated that the presence of carotid plaques and increased carotid intima-media thickness (IMT) significantly associated with the concurrent presence of CAC.[8,11,12] Risk factors contributing to CAD susceptibility can be categorized into two primary categories: non-modifiable and modifiable determinants. Non-modifiable risk factors include male sex, a familial history marked by instances of heart disease and advanced age. In contrast, modifiable risk factors are amendable to intervention through lifestyle adjustments and medical management. These factors include cigarette smoking, elevated blood pressure, excess body weight or obesity, and dietary patterns characterized by an unhealthy composition.[13–15] Understanding and addressing these risk factors are essential in CAD prevention, diagnosis, and management, as they significantly influence the disease's onset, progression, and prognosis. Primary prevention and treatment strategies are designed to reduce the modifiable risk factors of CAD, including smoking cessation, promoting healthy dietary habits, and encouraging regular physical activity.[16,17]

Investigating the genetic mechanisms and pathways of CAD not only enhances our understanding of disease etiology but also provides crucial insights into disease susceptibility, progression, and potential therapeutic targets.[18] Previous Genome-wide association studies (GWAS) of CAD have yielded numerous significant genetic loci. Notably, the transatlantic Coronary ARtery DIsease Genome Wide Replication and Meta-analysis (CARDIoGRAM) consortium was established in 2011 to conduct a meta-analysis GWAS of CAD comprising 22,233 cases and 64,762 controls and identified 13 novel genetic loci exhibiting significant associations with CAD.[19] In 2017, Pim van der Harst et al. conducted a meta-analysis GWAS of CAD leveraging the resources of CARDIoGRAMplusC4D and the UK and identified 64 novel genetic loci of CAD.[20] More recently, in 2022, Tcheandjieu, C. et al. performed a large scale GWAS of CAD in genetically diverse population including European, African American, Hispanics and Asian by using the resources of Million Veteran Program (MVP), UK Biobank, CARDIoGRAMplusC4D and Biobank Japan and identified 95 novel genetic loci of CAD.[21] Collectively, these GWAS endeavors have identified over 300 independent genetic loci of CAD and substantially broadened our knowledge of the genetic basis of CAD.[19–22]

Despite the significant advances made in identifying genetic variants and causal genes associated with CAD through GWAS and follow-up colocalization analysis in European ancestry,[23,24] a comprehensive exploration of the molecular consequences of novel multi-ancestry GWAS variants of CAD and their relevance to subclinical atherosclerosis remains largely uncharted in human cohorts. To bridge this knowledge gap, it is imperative to perform integrative analyses by leveraging multi-omics data including transcriptomics, proteomics, metabolomics, and epigenomics, which is crucial for identification of the molecular targets causally linked to CAD and further provide critical guidance in enhancing the accuracy of disease diagnosis and prognosis.

In this study, we aimed to prioritize CAD-related molecular targets and investigate their relationship with subclinical atherosclerosis using molecular 'omics data from the Trans-Omics for Precision Medicine (TOPMed) Multi-Ethnic Study of Atherosclerosis (MESA). To achieve this goal, we first applied Bayesian colocalization to overlap TOPMed MESA eQTL resources with a large scale multi-ancestry GWAS of CAD[21,25–27], identifying genes of interest for the CAD genetic loci. We further performed several follow-up validation analyses to prioritize a list of candidate causal genes of CAD. Second, we applied the weighted gene co-expression network analysis (WGCNA) to the TOPMed MESA transcriptomic data to identify gene expression modules and investigate their association with subclinical atherosclerosis traits in MESA. Third, we examined the modules identified by WGCNA for enrichment of the colocalized genes identified by Bayesian colocalization analysis.

4.3 Results

4.3.1 Identification of cell type specific colocalized genes underlying GWAS of CAD.

There were 47, 26 and 29 colocalized genes identified by Bayesian colocalization analysis leveraging GWAS of CAD and eQTL resources from the TOPMed MESA derived based on PBMCs, T cells and monocytes, respectively **(Figure 4.1a)**. Among these, 6 genes were identified across all three cell types, including *DDX59*, *CCDC30*, *NHSL1*, *ZNF100*, *VN1R84P* and *NRIP1*. However, it is important to note that several colocalized genes were unique to specific cell types (PBMC: 26 unique colocalized genes, T cell: 10, Monocyte: 10), which can be attributed to variations in sample sizes of the eQTL data **(Figure 4.1b)** and differences in gene expression profiles across distinct cell types.

Statistical fine-mapping (using SuSiE[28,29]) was incorporated in our colocalization analysis, which enhanced the identification of shared causal variants from GWAS of CAD and eQTL. The example of *PLEKHJ1*, a gene identified by colocalization with CAD GWAS, highlights the benefit of incorporating fine-mapping in the colocalization analysis. Fine-mapping successfully identified a total of 3 potential causal variants associated with CAD and 27 potential causal variants linked to *PLEKHJ1* gene expression. Two of these variants are overlapping and exhibit robust associations with both CAD and the expression levels of *PLEKHJ1* (**Figure 4.1c**). This finding suggests a potential mechanistic link between these causal genetic variants, CAD risk, and the regulation of *PLEKHJ1*. In another example, *SCARB1* is a colocalized gene that would have remained unidentified without the incorporation of statistical fine-mapping. There were multiple credible sets from both the GWAS of CAD and the eQTL signals within the region of *SCARB1.* After performing colocalization analysis based on each pair of credible sets, one shared causal signal demonstrated strong associations with both CAD risk and expression of *SCARB1* (**Figure 4.1d**).

4.3.2 Follow-up validation analyses prioritize the casual genes of CAD and subclinical atherosclerosis.

Several follow-up validation analyses focusing on the identified colocalized genes were conducted to prioritize 6 causal genes of CAD and subclinical atherosclerosis (*DDX59, CAMSAP2, AC018816.3, BICC1, EIF2B2* and *PLEKHJ1*) **(Table 4.1)**. Follow-up validation analyses included (a) exploration of the colocalized genes in GWAS of subclinical atherosclerosis, (b) investigation of colocalized genes in Artery tissue from GTEx, (c) examination of causal CAD variants for evidence of pQTL and mQTL associations for the corresponding colocalized genes in TOPMed MESA, (d) examination of the association of colocalized genes with subclinical atherosclerosis traits in MESA, (e) study of the colocalized genes in mouse genome.

Colocalized genes in GWAS of subclinical atherosclerosis: Colocalization analysis leveraging GWAS of subclinical atherosclerosis (CAC[30] and cIMT[31]) and the TOPMed MESA PBMC eQTL resource was performed to investigate the relationship between CAD colocalized gene and subclinical atherosclerosis. Based on this analysis, *PLEKHJ1*, one of the colocalized genes for CAD, also showed evidence of colocalization with cIMT (PPH4 = 87.7%). From this result, we found that multiple shared causal variants (rs2301798, rs76064118 and rs191615952) demonstrated significant associations with CAD, cIMT, as well as expression of *PLEKHJ1* **(Figure 4.1a)**.

Colocalized genes in Artery tissue: The CAD colocalized genes were also carried forward for colocalization analyses examining overlap of CAD GWAS signals with eQTL from GTEx artery tissues, including coronary, aorta and tibial. Multiple CAD colocalized genes, including *ZHF100, DDX59, OPRL1, RBM23, EIF2B2, KIAA0753, CCDC30, CAMSAP2, DHDDS* and *LIPA,* identified using the TOPMed MESA PBMC eQTL, also demonstrated colocalization with eQTL from artery tissues **(Table S4.1)**. For example, *DDX59* **(Figure 4.2b)**, one CAD colocalized gene identified using the TOPMed MESA PBMC eQTL was also identified based on colocalization with eQTL from GTEx artery tissues (coronary: PPH4 = 0.984 and aorta: 0.983).

pQTL and mQTL associations of colocalized genes: Further, the overlapping causal variants, identified by statistical fine-mapping in GWAS of CAD and eQTL, were carried forward to examine their evidence as pQTL or mQTL for the proteins and CpG sites corresponding to CAD-colocalized genes. From this analysis, we found rs11213945, one of the causal variants associated with CAD and gene expression of *LAYN*, demonstrated the strong association with the corresponding protein (LAYN) and CpG site (cg21703322) as a pQTL and mQTL, respectively **(Figure 4.2c)**.

Association of colocalized genes with subclinical atherosclerosis: Linear regression analysis was performed to examine the association between measured expression of CAD colocalized genes and subclinical atherosclerosis traits (CAC, cIMT and plaque) in MESA revealed nominal significant relationships for multiple genes, including *CAMSAP2, HMGN2, CARCRL, VN1R84P, EIF2B2, FDX, CNPY2, PARP12, BICC1, MLX* and others **(Table S4.2)**.

Colocalized genes in mouse genome: Additionally, three CAD colocalized genes (*SCARB1*, *BICC1* and *PAN2*) showed significant associations with heart and cardiovascular-related functions in the mouse genome, as evidenced by data from the International Mouse Phenotyping Consortium (IMPC). For example, mouse knockouts for *SCARB1* demonstrated decreased heart rate, abnormal sinus arrhythmia and cardiovascular system traits. The cardiovascular system traits included the observable morphological and physiological characteristics of the mammalian heart, blood vessels, or circulatory system that are manifested through development and lifespan.

Considered together, our discovery analyses and additional validation analyses identified six colocalized genes, including *DDX59, CAMSAP2, AC018816.3, BICC1, EIF2B2,* and *PLEKHJ1*, as candidate causal variants associated with CAD and subclinical atherosclerosis. These genes were selected based on the criterion that they were colocalized with the CAD GWAS in our primary discovery analysis, and also passed our specified thresholds in at least three of the follow-up validation analyses **(Table 4.1).**

4.3.3 Identification of subclinical atherosclerosis related modules using WGCNA

WGCNA for the TOPMed MESA PBMC transcriptomics data identified 31 modules of highly corelated genes **(Figure 4.3a).** Among these, 9 modules were associated with subclinical atherosclerosis in MESA **(Table S4.3)**. For example, the ‘darkolivegreen4’ module was nominally associated with CAC and cIMT. Pathway enrichment analysis using Molecular Signature Database (MSigDB) was showed the ‘darkolivegreen4’ module was enriched in pathways related to the up-regulation of genes in response to IFNG, up-regulation in response to alpha interferon proteins, up-regulation by IL6 via STAT3 during acute phase response, regulation by NF-kB in response to TNF, and inflammatory response.

4.3.4 Module identified from WGCNA significantly enriched for colocalized gene from colocalization analysis.

Enrichment analysis was performed to investigate whether the co-expression modules identified by WGCNA were enriched for genes colocalized with CAD GWAS signals. One module demonstrated significant enrichment for colocalized genes **(**‘maroon’: *P* = 0.011; **Figure 4.3b)**. Pathway enrichment analysis demonstrated the genes in the ‘maroon’ module are enriched for the G2/M checkpoint, as in progression through the cell division cycle, cell cycle related targets of E2F transcription factors and protein secretion pathway **(Table S4.4).**

Additionally, module membership (MM) was measured to identify 577 hub genes in the ‘maroon’ module based on the threshold of module membership great than 0.8. Regression analysis was performed to identify each hub gene’s relationship with subclinical atherosclerosis traits in MESA. Interestingly, *RDX*, one of the CAD colocalized genes, was also a hub gene in the ‘maroon’ module (MM > 0.8) and nominally associated with one subclinical atherosclerosis trait, plaque, in MESA.

4.4 Discussion

While previous GWAS of CAD have made significant progress by identifying over 400 independent genetic loci associated with CAD, there remains a need for systematic prioritization of CAD-related genes and a comprehensive investigation into their potential relationship with subclinical atherosclerosis. This critical gap in research has prompted the current study to explore and elucidate the genetic mechanisms linking CAD and subclinical atherosclerosis in greater detail. In an effort to bridge this knowledge gap, our study employed Bayesian colocalization analysis, followed by multiple validation analyses, to prioritize 6 candidate genes (*DDX59, CAMSAP2, AC018816.3, BICC1, EIF2B2,* and *PLEKHJ1*) associated with CAD and subclinical atherosclerosis by leveraging multi-omics data from TOPMed MESA. Additionally, we utilized WGCNA to identify 9 modules of genes exhibiting nominal association with subclinical atherosclerosis, allowing us to explore potential pathways linked to these genes in greater depth. Additionally, we conducted the enrichment analysis to demonstrate one module ('maroon') demonstrated significant enrichment for colocalized genes and we found that *RDX*, a plaque-related hub gene in this module, was also one of the CAD colocalized genes.

Our study illustrates the significance of incorporating statistical fine-mapping into the Bayesian colocalization analysis.[32] This approach allowed us to address a common scenario, where multiple independent signals are present within a specific region (typically defined as the transcript start site +/- 1Mb). Such a situation often arises when conducting GWAS and eQTL mapping, as we aim to establish associations between genetic variants, phenotypes, and gene expression levels. By incorporating statistical fine-mapping, our colocalization analysis systematically evaluated all potential pairs of credible sets between GWAS and eQTL results. This strategic approach enhanced our ability to identify colocalized genes in cases where both the GWAS and eQTL datasets exhibit multiple independent signals. For example, *SCARB1,* one of colocalized gene of CAD, was identified by incorporating statistical fine-mapping into colocalization analysis. Within the region of *SCARB1,* statistical fine-mapping identified multiple credible sets for both the GWAS of CAD and the eQTL signals. Following colocalization analysis involving each pair of credible sets revealed one shared causal signal showing .strong associations with both CAD risk and expression of *SCARB1* It is worth noting that in situations where the sample size or the statistical power of GWAS and eQTL does not allow for the confident detection of credible sets by SuSiE, basic colocalization analysis can still play a valuable role in identifying colocalized genes. Two validation analyses in our study, colocalization analysis using GWAS of subclinical atherosclerosis and colocalization analysis using eQTL from GTEx artery tissues, were performed by basic colocalization analysis under the single variant assumption, as the relatively small sample size limited our ability to apply statistical fine-mapping.

Our study further performed several validation analyses focusing on the colocalized genes of CAD to prioritize a list of candidate causal genes of CAD and subclinical atherosclerosis. Among the prioritized genes, *PLEKHJ1* (Pleckstrin Homology Domain Containing J1), involved in cellular processes associated with endosome dynamics and intracellular trafficking and implicated in the organization of endosome and the recycling of receptors, is a colocalized gene that can be identified using GWAS of both CAD and cIMT. This result suggests that *PLEKHJ1* influences the shared pathophysiological mechanisms of both IMT thickening and CAD.

EIF2B2 (Eukaryotic Translation Initiation Factor 2B Subunit Beta), is involved in protein synthesis and exchanges GDP and GTP for its activation and deactivation. This colocalized gene was identified using both MESA PBMC eQTL and well as GTEx Aorta and tibial eQTL. Prior studies demonstrated *EIF2B2,* a candidate causal CAD gene, is a key driver of gene regulatory co-expression network (GRN) in subcutaneous fat tissue in Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task (STARNET) and the top phenotypic associations related to this GRN include body mass index (BMI), waist-to-hip ratio (WHR), high-density lipoprotein cholesterol (HDL-C), and low-density lipoprotein cholesterol (LDL-C).[23] Additionally, *EIF2B2* has been identified as a gene target in CAD using Mendelian randomization (MR), suggesting a potential role for *EIF2B2* in the pathogenesis of CAD and related cardiovascular traits.[33]

DDX59 (DEAD-Box Helicase 59) and *CAMSAP2* (Calmodulin Regulated Spectrin Associated Protein Family Member 2) are two genes identified through colocalization analysis using both TOPMed MESA PBMCs and GTEx artery tissue data. Each of these genes have also shown associations with subclinical atherosclerosis traits in MESA. *DDX59* is predicted to enable both RNA binding activity and RNA helicase activity, suggesting its involvement in crucial cellular processes related to RNA metabolism. Notably, dysregulation or mutations in the *DDX59* gene have been linked to various medical conditions. Among these are Orofaciodigital Syndrome V, a rare genetic disorder characterized by facial and digital abnormalities, and Oliver Syndrome, a condition characterized by developmental anomalies.[34,35] These associations underscore the potential significance of *DDX59* in normal physiological development and highlight its relevance as a candidate gene for investigating and understanding these specific pathologies. *CAMSAP2* is predicted to facilitate microtubule minus-end binding activity. The depletion of CAMSAP2 has been associated with a notable escalation in microtubule dynamics, particularly evident during the intricate process of dendritic development.[36,37] Additionally, these two genes are located in close proximity to one another within the genome, with *DDX59* located in the region chr1:200623896-200669969 and CAMSAP2 in the region chr1:200739558-200860704. The proximity of these genes suggests the possibility of chromatin contacts, implying potential interactions or regulatory relationships between them. Prior research has indicated that both *DDX59* and *CAMSAP2* encode proteins that are expressed in smooth muscle cells. This observation raises the hypothesis that biological processes occurring in the arterial wall, particularly those involving smooth muscle cells, may be of significant importance in coronary atherogenesis, which is a key factor in CAD development.[38]

BICC1 (BicC family RNA binding protein 1) is a colocalized gene was additionally validated using eQTL of artery tissues. *BICC1* demonstrated that its association with abnormal heart morphology and abnormal heart looping in the mouse genome from IMPC. *BICC1*, a gene encoding an RNA-binding protein, plays a crucial role in the intricate regulation of gene expression by modulating protein translation, particularly during the critical period of embryonic development. Notably, previous studies have provided compelling evidence linking the knockout of *Bicc1* in mice to the onset of polycystic kidney disease (PKD) and cystic renal dysplasia, shedding light on its significant implications in renal health.[39,40] Further investigations into the role of *BICC1* in osteoblastogenesis revealed that the reduction in *Bicc1* expression correlates with decreased areal bone mineral density (BMD). This effect is primarily attributed to diminished cortical thickness and cortical tissue mineral density.[41] These findings underscore the importance of *BICC1* in maintaining the delicate balance of cellular processes critical for embryonic development, renal function, and skeletal health.

LAYN (Layilin) is a protein coding gene involved in hyaluronic acid binding activity. Our analyses revealed that multiple genetic variants identified in GWAS of CAD and eQTL of *LAYN* also demonstrated significant associations as pQTL and mQTL for the corresponding protein and CpG sites, respectively. This result suggests these CAD GWAS variants may have downstream effects on gene expression, protein functionality, and DNA methylation patterns.[42] *LAYN* is *a protein coding gene with the capacity* for enabling carbohydrate and hyaluronic acid binding. Previous investigations have underscored the pivotal role of LAYN in the pathogenesis of colorectal and gastric cancers and tumor-immune interactions.[43] Interestingly, CAD is a chronic inflammatory disease that marked by accumulation of atherosclerotic plaques containing immune cells exhibiting diverse states of activation and differentiation.[44] Prior studies demonstrated that *LAYN* is associated with the suppressive function of tumor Regulatory T (Treg) cells and exhausted CD8 T cells.[45,46]

Another distinguishing feature of our study is the utilization of multi-ancestry GWAS of CAD and multi-ancestry MESA eQTL resources, with both resources having been constructed using data sets spanning individuals of European, African, Hispanic and Asian race/ancestry. To date, most statistical genetic studies have focused primarily on a single ancestry group, predominantly emphasizing European ancestry. This single ancestry approach can be underpowered to detect certain genetic signals due to differences in allele frequencies, LD patterns and effect sizes across ancestries.[47–50] The incorporation of multi-ancestry resources in our study conferred several noteworthy advantages. For example, use of multi-ancestry resources allowed for increased sample sizes by incorporating participants from different ancestry groups, which further enhance the statistical power in subsequent fine-mapping and colocalization analyses. Moreover, use of multi-ancestry resources can reduce the risk of missing important genetic variants that may be prevalent in specific ancestry groups and may play a significant role in CAD susceptibility or treatment response. *PLEKHJ1*, one of our candidate genes of CAD and cIMT, was identified through the utilization of multi-ancestry GWAS of CAD and multi-ancestry MESA eQTL resources. The reason *PLEKHJ1* had not been identified in previous GWAS and integrative analysis within the European population can be attributed to the difference in allele frequencies of the causal variants across various ancestry groups. For instance, two shared causal variants (rs76064118 and rs191615952) of CAD risk and expression of *PLEKHJ1*, have allele frequencies of approximately 0.05 in European ancestry, whereas they are closed to 0.12 in Asian ancestry.

Given the limited number of genes identified through Bayesian colocalization, which posed challenges in examination of pathways, our study applied Correlation network analysis (WGCNA) leveraging the TOPMed MESA transcriptomics data to identify modules related to subclinical atherosclerosis and further utilized Molecular Signature Database (MSigDB) to explore the pathways among genes within the modules of interest. We found one plaque-related hub gene, *RDX*, in the ‘maroon’ module, which was also one of the genes identified by colocalization analysis. *RDX* (Radixin) is a cytoskeletal protein that plays a potentially crucial role in connecting actin filaments to the plasma membrane. Our study demonstrates the value of contextualizing genes implicated by GWAS using correlation network approaches to identify the key drivers.

Despite the numerous strengths, our study also faces several limitations. Our discovery analysis for identification of colocalized genes of CAD relied on eQTL resources derived from circulating cells. Although blood and circulating cells can provide valuable insights, they may not represent the most directly relevant sources for studying CAD. To address this limitation, we performed follow-up colocalization analysis leveraging the eQTL resources from GTEx to investigate the effects of CAD colocalized genes in artery tissue. This analysis revealed multiple colocalized genes with overlapping evidence based on eQTL from circulating cells as well as artery tissues. However, to further enhance the relevance and comprehensiveness of our discovery analysis, it would be advantageous to leverage eQTL data from other disease-related tissues available in existing databases. For instance, the Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task (STARNET) study represents a valuable eQTL resource for our primary analysis. STARNET recruited 600 well-characterized CAD patients and sequenced RNA isolated from various tissues, including blood, atherosclerotic-lesion-free internal mammary artery (MAM), atherosclerotic aortic root (AOR), subcutaneous fat (SF), visceral abdominal fat (VAF), skeletal muscle (SKLM), and liver (LIV), which identified ~8 million cis-eQTLs.[51] Incorporating data from these resources could further enrich our genetic studies and provide a more targeted and comprehensive understanding of the genetic factors associated with CAD. Finally, we had limited power for some of the follow-up validation analyses examining associations with subclinical atherosclerosis in MESA. Thus, we made use of a nominal p-value threshold (*P*<0.05) for these validations, which did not meet the most rigorous standards afforded by formal correction for multiple comparisons.

In summary, our study employed a diverse array of statistical analyses leveraging the multi-omics data from TOPMed MESA to prioritize candidate genes associated with CAD and subclinical atherosclerosis. The follow-up experimental validation focusing on the candidate genes of CAD we identified can be conducted to advance understanding of these prioritized candidates in the future. Collectively, these findings can provide a better understanding of genetic mechanisms and pathways implicated by GWAS of CAD and subclinical atherosclerosis and consequently provide valuable insights into potential therapeutic interventions and treatments.

4.5 Methods

4.5.1 Study participants

MESA is a longitudinal cohort study of subclinical cardiovascular disease and risk factors that predict progression to clinically overt cardiovascular disease or progression of subclinical disease. Between 2000 and 2002, MESA recruited 6,814 men and women 45 to 84 years of age from Forsyth County, North Carolina; New York City; Baltimore; St. Paul, Minnesota; Chicago; and Los Angeles. Participants at baseline were 38% White, 28% African American, 22% Hispanic and 12% Asian (primarily Chinese) ancestry.[52]

4.5.2 Primary analysis for identification of CAD colocalized gene

4.5.2.1 TOPMed MESA transcriptomics and eQTL

MESA transcriptomics and cis-eQTL were obtained through TOPMed Freeze 1RNA. Freeze 1RNA TOPMed cis-eQTL results were generated in a collaboration between the TOPMed Informatics Research Center, TOPMed Multi-Omics working group, and the TOPMed parent studies contributing RNA-seq and distributed to TOPMed investigators. The cis-eQTL mapping was performed using tensorQTL[53] for PBMCs (n = 1256), T cells (n = 368) and monocytes (n = 352), and for each gene, genetic variants within 1Mb of the gene TSS were tested. The covariates for cis-eQTL mapping included sex, 15 genotype PCs and 30 gene expression PCs. Samples for the cis-eQTL scan were selected via the following procedure: (1) exclude samples that did not pass the PC-based outlier filter, (2). exclude samples without a known WGS match, (3). exclude samples where the subject (WGS match) is not in the unrelated subject set, and (4) exclude samples with unclear sex based on gene expression.

4.5.2.2 GWAS of coronary artery disease (CAD)

We leveraged a recently published publicly available large-scale multi-ancestry of GWAS of CAD comprising of 243,392 cases and 849.686 controls.[41] This study used METAL to conduct a fixed-effect inverse variance-weighted meta-analysis for the clinical CAD phenotype, including Million Veteran Program (MVP) European participants MVP Black participants, MVP Hispanic participants and Biobank Japan, CARDIoGRAMplusC4D 1000G study, the UK Biobank CAD study and Biobank Japan.

4.5.2.3 Statistical Fine-mapping

Statistical fine-mapping, an approach for the identification of the casual variants from GWAS, was incorporated into Bayesian colocalization analysis. Sum of Single Effect model (SuSiE[28,29]) was performed to identify the credible sets of putative causal variants by using the summary statistic results including effect size, standard error and minor allele frequencies in GWAS of CAD and eQTL with L = 10 (SuSiE default). Each credible set is constructed to have high probability to contain a signal with non-zero effect, while at the same time being as small as possible. The follow-up colocalization analysis systematically evaluated all potential pairs of credible sets between GWAS and eQTL results.

4.5.2.4 Bayesian colocalization analysis

Bayesian colocalization analysis[32,54] was used to identify the downstream genes of CAD leveraging MESA eQTL and GWAS of CAD by using R/coloc.susie package. Bayesian colocalization analysis tested the following hypotheses: H0. neither GWAS nor eQTL has a genetic association in the region; H1. only GWAS has a genetic association in the region; H2. only eQTL has a genetic association in the region; H3. both GWAS and eQTL are associated, but with different causal variants; H4. both GWAS and eQTL are associated and share a single causal variant. A posterior colocalization probability of hypothesis 4 (PP.H4) > 0.80 was used as the threshold of colocalization.

4.5.3 Follow-up validation approaches

4.5.3.1 Colocalization analysis for identification of subclinical atherosclerosis colocalized genes

Colocalization analysis focusing on the CAD colocalized genes from primary analysis was performed to identify the downstream genes of subclinical atherosclerosis leveraging MESA eQTL and GWAS of subclinical atherosclerosis (CAC[30] and cIMT[31]) by using R/coloc package. For GWAS of CAC, we leveraged a published publicly available meta-analysis GWAS of CAC comprising 9961 participants from 5 independent community-based cohorts (Age, Gene/Environment Susceptibility–Reykjavik Study [AGES-Reykjavik], the Framingham Heart Study [FHS], the Rotterdam Study I [RS I], and the Rotterdam Study II [RS II], Genetic Epidemiology Network of Arteriopathy Study [GENOA]).[30] For GWAS of cIMT, we leveraged a published publicly available meta-analyses of GWAS of cIMT in 71,128 individuals of European ancestry from 31 studies for cIMT.[31]

4.5.3.2 Investigation of CAD colocalized genes in Artery tissue from GTEx

Colocalization analysis focusing on the CAD colocalized genes from primary analysis was performed to identify the downstream genes of CAD in artery tissues leveraging GTEx eQTL[55] and GWAS of CAD by using R/coloc package. The tissues of GTEx eQTL we used in colocalization analysis included artery-coronary tissue (n = 213), artery-aorta tissue (n = 387) and artery-tibial (n = 584). GTEx eQTL can be found and downloaded at https://www.gtexportal.org/home/downloads/adult-gtex.

4.5.3.3 Examination of causal CAD variants for evidence of pQTL and mQTL associations for the corresponding colocalized genes

Statistical fine-mapping in GWAS of CAD and MESA eQTL from the primary analysis identified shared causal variants. These shared causal variants were further investigated to assess their potential impact as MESA pQTL or MESA mQTL for the proteins and CpG sites corresponding to CAD-colocalized genes. TOPMed MESA multi-ancestry pQTL resources included 1,305 proteins measured by a SOMAscan assay and 971 unique individuals (African American [n = 183], Chinese [n = 71], European [n = 416], and Hispanic/Latino [n = 301]).[56] TOPMed MESA multi-ancestry mQTL resources included the whole blood DNA methylation (DNAm) data for 747,868 CpG sites (CpG sites passing QC - 740,291) and 900 unique individuals.[57]

4.5.3.4 Association of colocalized genes with subclinical atherosclerosis

CAD colocalized genes identified from primary analysis were carried forward to examine the association with subclinical atherosclerosis traits (CAC, IMT and carotid plaque) in MESA exam 5 using linear regression model and the covariates included age, gender and study sites. A nominal p-value (0.05) was used as the threshold of association. In MESA exam5, CAC was measured with either electron-beam computed tomography (EBT) at 3 field centers or multidetector computed tomography (MDCT) at 3 field centers.[52,58,59] The amount of calcium was quantified with the Agatston scoring method. IMT was defined as the intima-media thickness measured as the mean of the mean left and right mean far wall distal CCA wall thicknesses. Carotid plaque was defined as a discrete, focal wall thickening ≥1.5 cm or focal thickening at least 50% greater than the surrounding IMT.

4.5.3.5 Investigation of CAD colocalized genes in mouse genome

The International Mouse Phenotyping Consortium (IMPC)[64] was leveraged to study the biological function of CAD colocalized genes in mouse genome. IMPC is an international effort by 21 research institutions, consisting of 85M data points and over 95,000 statistically significant phenotype hits mapped to human disease, to identify the function of every protein-coding gene in the mouse genome.

https://www.mousephenotype.org

4.5.4 Primary analysis for identification of modules using correlation network

4.5.4.1 Weighted gene co-expression network analysis (WGCNA)

Weighted gene co-expression network analysis (WGCNA)[65] is an approach to study biological networks of genes, which can find the clusters (modules) of highly correlated genes and summarize the node profiles using the module eigengene or an intramodular hub node. The rationale behind correlation network methodology is to use network language to describe the pairwise relationships (correlations) among the molecular omics traits. Our study applied WGCNA on MESA PBMC transcriptomic data including 1,256 unique individuals and 24,410 genes to identify the modules of highly correlated genes. The steps for conducting WGCNA include: (1). Identification of soft thresholding power based on the criterion of approximate scale-free topology; (2) Calculation the adjacencies based on the soft thresholding power (Power = 5); (3) Transformation of the adjacency into Topological Overlap Matrix and calculation of the corresponding dissimilarity; (4). Generation of a hierarchical clustering tree (dendrogram) of genes. Module membership (MM) was measured to identify hub genes for each module by examining the correlation between the module eigengene and the gene expression profile. Hub genes are highly correlated with many other module genes and have been shown to be important in disease and in controlling module behavior.[67] Additionally, hub genes were further carried forward to identify subclinical atherosclerosis traits related hub genes by investigating the relationship with subclinical atherosclerosis traits (CAC, IMT and carotid plaque) in MESA exam 5.

4.5.4.2 Pathway enrichment analysis for identification of biological pathways

We applied pathway enrichment analysis to investigate the **biological pathways** for the genes within the module of interest identified by our WGCNA using the Molecular

Signature Database (MSigDB).[68,69] MSigDB includes multiple categories, for example, hallmark gene sets (H), curated gene sets (C2), regulatory target gene sets (C3), computational gene sets (C4), ontology gene sets (C5), oncogenic signature gene sets (C6), immunologic signature gene (C7), cell type signature gene sets (C8).

4.6 Reference

1. Khera, A. V. & Kathiresan, S. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat. Rev. Genet.* **18**, 331–344 (2017).
2. Ralapanawa, U. & Sivakanesan, R. Epidemiology and the Magnitude of Coronary Artery Disease and Acute Coronary Syndrome: A Narrative Review. *J. Epidemiol. Glob. Health* **11**, 169–177 (2021).
3. Frostegård, J. Immunity, atherosclerosis and cardiovascular disease. *BMC Med.* **11**, 117 (2013).
4. Hansson, G. K. Inflammation, Atherosclerosis, and Coronary Artery Disease. *N. Engl. J. Med.* **352**, 1685–1695 (2005).
5. Khoury, Z. *et al.* Relation of Coronary Artery Disease to Atherosclerotic Disease in the Aorta, Carotid, and Femoral Arteries Evaluated by Ultrasound. *Am. J. Cardiol.* **80**, 1429–1433 (1997).
6. Barrett-Connor, E. L. Obesity, Atherosclerosis, and Coronary Artery Disease. *Ann. Intern. Med.* **103**, 1010–1019 (1985).
7. Fong, I. W. Emerging relations between infectious diseases and coronary artery disease and atherosclerosis. *CMAJ* **163**, 49–56 (2000).
8. Polak, J. F., Tracy, R., Harrington, A., Zavodni, A. E. H. & O'Leary, D. H. Carotid artery plaque and progression of coronary artery calcium: the multi-ethnic study of atherosclerosis. *J. Am. Soc. Echocardiogr. Off. Publ. Am. Soc. Echocardiogr.* **26**, 548–555 (2013).

9. Nowak, J., Nilsson, T., Sylvén, C. & Jogestrand, T. Potential of Carotid Ultrasonography in the Diagnosis of Coronary Artery Disease. *Stroke* **29**, 439–446 (1998).

10. Kallikazaros, I., Tsioufis, C., Sideris, S., Stefanadis, C. & Toutouzas, P. Carotid Artery Disease as a Marker for the Presence of Severe Coronary Artery Disease in Patients Evaluated for Chest Pain. *Stroke* **30**, 1002–1007 (1999).

11. Liu, D., Du, C., Shao, W. & Ma, G. Diagnostic Role of Carotid Intima-Media Thickness for Coronary Artery Disease: A Meta-Analysis. *BioMed Res. Int.* **2020**, 9879463 (2020).

12. Kablak-Ziembicka, A. *et al.* Association of increased carotid intima-media thickness with the extent of coronary artery disease. *Heart Br. Card. Soc.* **90**, 1286–1290 (2004).

13. Hajar, R. Risk Factors for Coronary Artery Disease: Historical Perspectives. *Heart Views Off. J. Gulf Heart Assoc.* **18**, 109–114 (2017).

14. Ghasemzadeh, G., Soodmand, M. & Moghadamnia, M. T. The Cardiac Risk Factors of Coronary Artery Disease and its relationship with Cardiopulmonary resuscitation: A retrospective study. *Egypt. Heart J.* **70**, 389–392 (2018).

15. Alpert, J. S. New Coronary Heart Disease Risk Factors. *Am. J. Med.* **136**, 331–332 (2023).

16. Pencina, M. J. *et al.* Quantifying Importance of Major Risk Factors for Coronary Heart Disease. *Circulation* **139**, 1603–1611 (2019).

17. Assmann, G., Cullen, P., Jossa, F., Lewis, B. & Mancini, M. Coronary Heart Disease: Reducing the Risk. *Arterioscler. Thromb. Vasc. Biol.* **19**, 1819–1824 (1999).

18. Padmanabhan, S., Hastie, C., Prabhakaran, D. & Dominczak, A. F. Genomic approaches to coronary artery disease. *Indian J. Med. Res.* **132**, 567–578 (2010).

19. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43**, 333–338 (2011).

20. van der Harst, P. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ. Res.* **122**, 433–443 (2018).

21. Tcheandjieu, C. *et al.* Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. *Nat. Med.* **28**, 1679–1692 (2022).

22. Erdmann, J., Kessler, T., Munoz Venegas, L. & Schunkert, H. A decade of genome-wide association studies for coronary artery disease: the challenges ahead. *Cardiovasc. Res.* **114**, 1241–1257 (2018).

23. Hao, K. *et al.* Integrative Prioritization of Causal Genes for Coronary Artery Disease. *Circ. Genomic Precis. Med.* **15**, e003365 (2022).

24. Zhong, Y. *et al.* Integration of summary data from GWAS and eQTL studies identified novel risk genes for coronary artery disease. *Medicine (Baltimore)* **100**, e24769 (2021).

25. Chen, C. *et al.* Applications of multi-omics analysis in human diseases. *MedComm* **4**, e315 (2023).

26. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 83 (2017).

27. Kedaigle, A. & Fraenkel, E. Turning omics data into therapeutic insights. *Curr. Opin. Pharmacol.* **42**, 95–101 (2018).

28. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).

29. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the “Sum of Single Effects” model. *PLOS Genet.* **18**, e1010299 (2022).

30. O’Donnell, C. J. *et al.* Genome-wide association study for coronary artery calcification with follow-up in myocardial infarction. *Circulation* **124**, 2855–2864 (2011).

31. Franceschini, N. *et al.* GWAS and colocalization analyses implicate carotid intima-media thickness and carotid plaque loci in cardiovascular outcomes. *Nat. Commun.* **9**, 5141 (2018).

32. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLOS Genet.* **17**, e1009440 (2021).

33. Chignon, A. *et al.* Enhancer promoter interactome and Mendelian randomization identify network of druggable vascular genes in coronary artery disease. *Hum. Genomics* **16**, 8 (2022).

34. Shamseldin, H. E. *et al.* Mutations in DDX59 Implicate RNA Helicase in the Pathogenesis of Orofaciodigital Syndrome. *Am. J. Hum. Genet.* **93**, 555–560 (2013).

35. Salpietro, V. *et al.* A loss-of-function homozygous mutation in DDX59 implicates a conserved DEAD-box RNA helicase in nervous system development and function. *Hum. Mutat.* **39**, 187–192 (2018).

36. Cao, Y. *et al.* Microtubule Minus-End Binding Protein CAMSAP2 and Kinesin-14 Motor KIFC3 Control Dendritic Microtubule Organization. *Curr. Biol. CB* **30**, 899-908.e6 (2020).

37. Yau, K. W. *et al.* Microtubule minus-end binding protein CAMSAP2 controls axon specification and dendrite development. *Neuron* **82**, 1058–1073 (2014).

38. Howson, J. M. M. *et al.* Fifteen new risk loci for coronary artery disease highlight arterial wall-specific mechanisms. *Nat. Genet.* **49**, 1113–1119 (2017).

39. Cogswell, C. *et al.* Positional cloning of jcpk/bpk locus of the mouse. *Mamm. Genome Off. J. Int. Mamm. Genome Soc.* **14**, 242–249 (2003).

40. Kraus, M. R.-C. *et al.* Two mutations in human BICC1 resulting in Wnt pathway hyperactivity associated with cystic renal dysplasia. *Hum. Mutat.* **33**, 86–90 (2012).

41. Mesner, L. D. *et al.* Bicc1 is a genetic determinant of osteoblastogenesis and bone mineral density. *J. Clin. Invest.* **124**, 2736–2749 (2014).

42. Wu, Y. *et al.* Joint analysis of GWAS and multi-omics QTL summary statistics reveals a large fraction of GWAS signals shared with molecular phenotypes. *Cell Genomics* **3**, 100344 (2023).

43. Pan, J. *et al.* LAYN Is a Prognostic Biomarker and Correlated With Immune Infiltrates in Gastric and Colon Cancers. *Front. Immunol.* **10**, (2019).

44. Human Coronary Plaque T Cells Are Clonal and Cross-React to Virus and Self | Circulation Research. https://www.ahajournals.org/doi/10.1161/CIRCRESAHA.121.320090.

45. Zheng, C. *et al.* Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. *Cell* **169**, 1342-1356.e16 (2017).

46. De Simone, M. *et al.* Transcriptional Landscape of Human Tissue Lymphocytes Unveils Uniqueness of Tumor-Infiltrating T Regulatory Cells. *Immunity* **45**, 1135–1147 (2016).

47. Ishigaki, K. *et al.* Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis. *Nat. Genet.* **54**, 1640–1651 (2022).

48. Chen, M.-H. *et al.* Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* **182**, 1198-1213.e14 (2020).

49. Koyama, S. *et al.* Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease. *Nat. Genet.* **52**, 1169–1177 (2020).

50. Khunsriraksakul, C. *et al.* Multi-ancestry and multi-trait genome-wide association meta-analyses inform clinical risk prediction for systemic lupus erythematosus. *Nat. Commun.* **14**, 668 (2023).

51. Cardiometabolic Risk Loci Share Downstream Cis- and Trans-Gene Regulation Across Tissues and Diseases - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5534139/.

52. Bild, D. E. *et al.* Multi-Ethnic Study of Atherosclerosis: objectives and design. *Am. J. Epidemiol.* **156**, 871–881 (2002).

53. Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (2019).

54. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, e1004383 (2014).

55. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

56. Schubert, R. *et al.* Protein prediction for trait mapping in diverse populations. *PLOS ONE* **17**, e0264341 (2022).

57. Kasela, S. *et al.* Interaction molecular QTL mapping discovers cellular and environmental modifiers of genetic regulatory effects. *bioRxiv* 2023.06.26.546528 (2023) doi:10.1101/2023.06.26.546528.

58. McClelland, R. L. *et al.* 10-Year Coronary Heart Disease Risk Prediction Using Coronary Artery Calcium and Traditional Risk Factors: Derivation in the MESA (Multi-Ethnic Study of Atherosclerosis) With Validation in the HNR (Heinz Nixdorf Recall) Study and the DHS (Dallas Heart Study). *J. Am. Coll. Cardiol.* **66**, 1643–1653 (2015).

59. Carr, J. J. *et al.* Calcified coronary artery plaque measurement with cardiac CT in population-based studies: standardized protocol of Multi-Ethnic Study of Atherosclerosis (MESA) and Coronary Artery Risk Development in Young Adults (CARDIA) study. *Radiology* **234**, 35–43 (2005).

60. McClelland, R. L., Chung, H., Detrano, R., Post, W. & Kronmal, R. A. Distribution of coronary artery calcium by race, gender, and age: results from the Multi-Ethnic Study of Atherosclerosis (MESA). *Circulation* **113**, 30–37 (2006).

61. Agatston, A. S. *et al.* Quantification of coronary artery calcium using ultrafast computed tomography. *J. Am. Coll. Cardiol.* **15**, 827–832 (1990).

62. Stein, J. H. *et al.* Use of carotid ultrasound to identify subclinical vascular disease and evaluate cardiovascular disease risk: a consensus statement from the American Society of Echocardiography Carotid Intima-Media Thickness Task Force. Endorsed by the Society for Vascular Medicine. *J. Am. Soc. Echocardiogr. Off. Publ. Am. Soc. Echocardiogr.* **21**, 93–111; quiz 189–190 (2008).

63. Tattersall, M. C. *et al.* Predictors of Carotid Thickness and Plaque Progression Over a Decade: The Multi-Ethnic Study of Atherosclerosis (MESA). *Stroke J. Cereb. Circ.* **45**, 3257–3262 (2014).

64. Groza, T. *et al.* The International Mouse Phenotyping Consortium: comprehensive knockout phenotyping underpinning the study of human disease. *Nucleic Acids Res.* **51**, D1038–D1045 (2022).

65. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

66. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

67. Liu, Y. *et al.* Identification of Hub Genes and Key Pathways Associated With Bipolar Disorder Based on Weighted Gene Co-expression Network Analysis. *Front. Physiol.* **10**, 1081 (2019).

68. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinforma. Oxf. Engl.* **27**, 1739–1740 (2011).

69. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).

4.7 Acknowledgments

This work was supported by a Predoctoral Fellowship from the American Heart Association (C.Y.). The authors acknowledge Research Computing at The University of Virginia for providing computational resources and technical support that have contributed to the results reported within this publication.

The Multi-Ethnic Study of Atherosclerosis

The Multi-Ethnic Study of Atherosclerosis (MESA) projects are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420, UL1TR001881, DK063491, and R01HL105756. The authors thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutes can be found at http://www.mesa-nhlbi.org.

Whole Genome Sequencing (WGS) for “NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA)” (phs001416) was performed at the Broad Institute of MIT and Harvard (3U54HG003067-13S1 and HHSN268201500014C). RNA-seq for the NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA)” (phs001416.v1.p1) was performed at the Northwest Genomics Center (HHSN268201600032I) and at the Broad Institute Genomics Platform (HHSN268201600034I). Genome-wide methylation for NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA)" (phs001416.v1.p1) was performed at Keck Molecular Genomics Core Facility (HHSN268201600034I). SOMAscan proteomics for NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA)" (phs001416.v1.p1) was performed at the Broad Institute and Beth Israel Proteomics Platform (HHSN268201600034I). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1, contract HHSN268201800001I), and TOPMed MESA Multi-Omics (HHSN2682015000031/HSN26800004). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services. A full list of investigators for the NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium is provided at https://topmed.nhlbi.nih.gov/topmed-banner-authorship.

RNA-Seq eQTL

Freeze 1RNA TOPMed cis-eQTL results were generated in a collaboration between the TOPMed Informatics Research Center, TOPMed Multi-Omics working group, and the TOPMed parent studies contributing RNA-seq and distributed to TOPMed investigators. We acknowledge the contributing cohorts, sequencing centers, and the TOPMed IRC.

Figure 4.1. Identification of colocalized genes for CAD across different cell types.

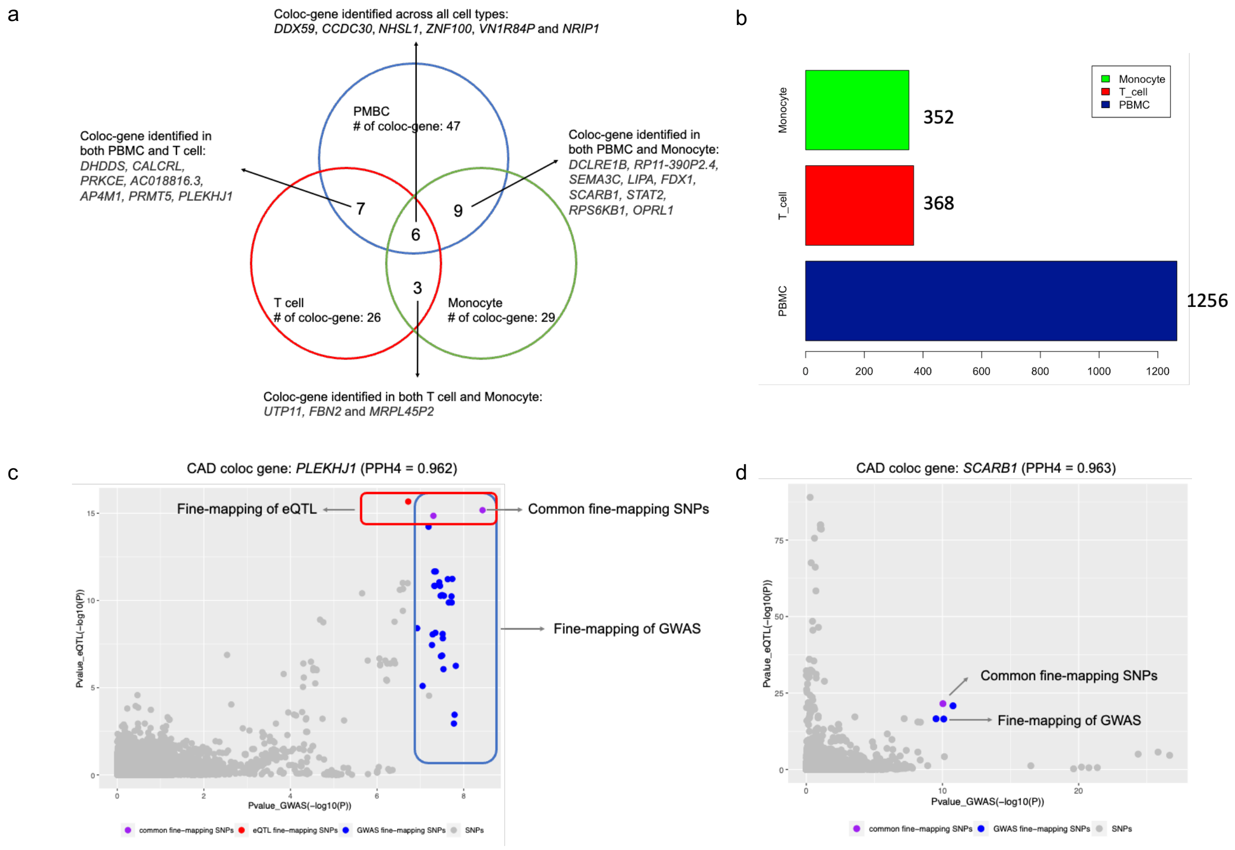


Figure 4.1: Identification of colocalized genes for CAD across different cell types. Panel a shows the colocalized genes for CAD identified using eQTL from TOPMed MESA for PBMCs, T cells and monocytes. Panel b shows the sample size of eQTL for PBMCs, T cells and monocytes. Panel c and d show two examples of colocalized genes (PLEKHJ1 and SCARB1).

Figure 4.2. Prioritization of causal genes for CAD and subclinical atherosclerosis.

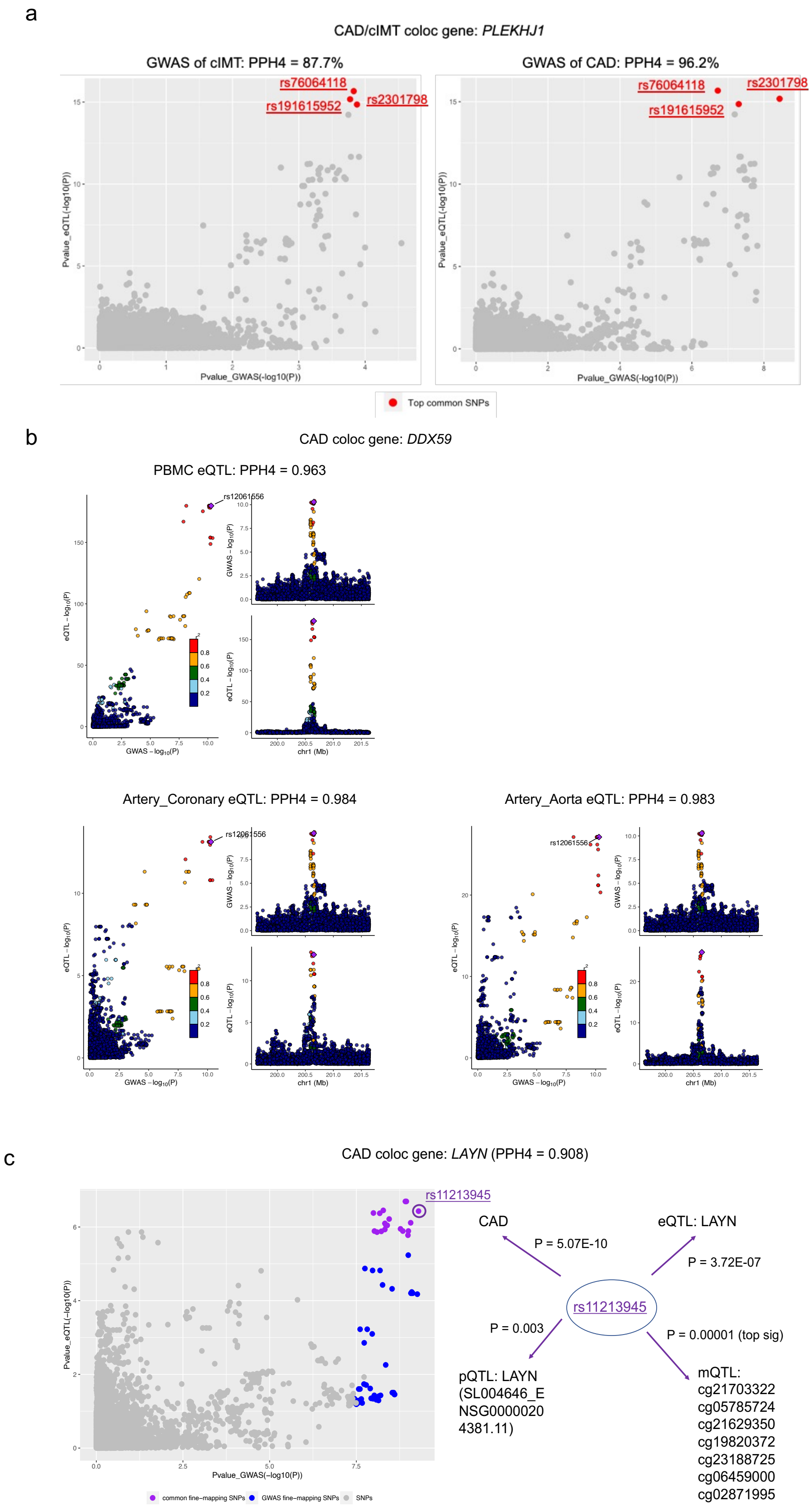
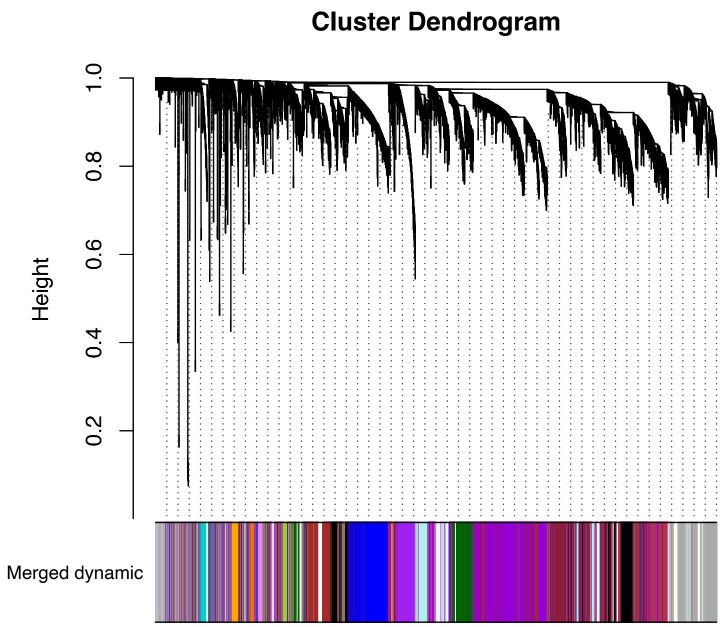


Figure 4.2: Prioritization of causal genes for CAD and subclinical atherosclerosis using different validation analyses. Panel a shows *PLEKHJ1* was identified as a colocalized gene using both GWAS of CAD (left) and cIMT (right). Panel b shows *DDX59* was identified as a colocalized gene using both eQTL of PBMC (top) and eQTL of coronary (bottom left) and aorta (bottom right). Panel c shows the shared causal variant has strong associations in GWAS of CAD, eQTL, pQTL and mQTL for *LAYN*.

Figure 4.3. Module identification using MESA PBMC transcriptomics data.

a



b

Module	# of colocalized gene	Module	# of colocalized gene
maroon	12	brown4	2
darkviolet	11	antiquewhite4	1
darkgreen	5	ivory	1
darkgrey	4	purple	1
brown	3	paleturquoise	1
lightcyan1	2	darkolivegreen4	1
black	2		

	maroon	non-maroon
colo	12	2985
non-colo	35	21378

p-value =0.012

Figure 4.3: Module identification using MESA PBMC transcriptomics data. Panel a shows the hierarchical clustering tree (dendrogram) of the MESA PBMC transcriptomics data. Panel b shows the distribution of colocalized genes in each module identified by WGCNA (left) and the enrichment analysis of genes within the maroon module for colocalized genes.

Table 4.1. Prioritization of causal gene of CAD and subclinical atherosclerosis

	Discovery	Follow-up validation analysis				
Gene	PPH4	Coloc_Athero	Coloc_Artery	pQTL/mQTL	Asso_Athero	MousePhe
DDX59	0.963		●	●	●	
CAMSAP2	0.965		●	●	●	
AC018816.3	0.991		●	●	●	
BICC1	0.836		●	●		●
EIF2B2	0.905		●	●	●	
PLEKHJ1	0.962	●		●	●	

Table 4.1 shows the prioritization of the causal gene list of CAD and subclinical atherosclerosis. Coloc_Athero: colocalization analysis using GWAS of subclinical atherosclerosis. Asso_Athero: association analysis between colocalized gene and subclinical atherosclerosis in MESA. Coloc_Artery: colocalization analysis using eQTL from GTEx artery tissues. MousePhe: function of colocalized gene in mouse genome using IMPC. pQTL/mQTL: association study of shared causal variants from GWAS of CAD and eQTL in pQTL and mQTL.

Chapter 5

Discussion and Future Directions

5.1 Summary and conclusion

This thesis has provided a deeper insight into the genetic and biological mechanisms underlying fatty acid metabolism and coronary artery disease by leveraging different types of molecular omics data (genomics, transcriptomics, proteomics, and methylation) and using various statistical approaches (genome-wide association study, statistical fine-mapping and integrative analysis). Additionally, there are multiples distinguishing features of this thesis: (1). Investigation of global and local proportions of genetic ancestry for Hispanic Americans; (2). Involvement of non-European ancestry (Hispanic Americans and African Americans) for genome-wide association study and follow-up validation analysis; and (3). Integration of different types of molecular ‘omics data for identification of disease-causing genes.

In chapter 2, we estimated the global proportions of Amerind ancestry for MESA Hispanic participants using ADMIXTURE and further carried out linear regression analysis to reveal a significant association between higher proportions of Amerind ancestry and lower levels of LC-PUFAs in MESA Hispanic participants. Additionally, we demonstrated that *FADS* variation rs174537 SNP has a strong effect on the ancestry-related decline in all LC-PUFAs. Furthermore, we also showed that the *FADS* cluster SNP rs174537 T allele was significantly associated with lower HDL-C levels, higher waist-hip ratio, reduced height and weight, and elevated levels of the inflammatory markers E-Selectin and s-ICAM. In conclusion, our research underscores the utility of Amerind ancestry as a readily accessible tool for identifying individuals who are at a higher risk of FADS-related deficiencies in n-3 LC-PUFAs and the associated cardiovascular risks.

In chapter 3, we performed meta-analysis of GWAS for n-3 and n-6 PUFAs in Hispanic Americans and African Americans. Our study confirmed that the genetic variants identified in prior CHARGE GWAS of PUFAs in European ancestry could be identified in Hispanic Americans and African Americans (*FADS1/2*, *PDXDC1*, *GCKR* and *ELOVL2*). Indeed, as established, the *FADS* region is well-documented for its significant association with PUFAs, and our study further revealed a considerable number of independent genetic association signals within neighborhood of *FADS* region on chromosome 11. Our study also found a large number of signals in Hispanic Americans that could not be replicated across race/ancestry groups, which can be attributed to dramatic differences in allele frequencies. For example, the chromosome 11 *POLD4* (DNA polymerase delta 4, accessory subunit) missense variant rs28364240 identified in association with AA have minor allele frequencies of 0.204 in Hispanic Americans, compared to frequencies close to zero in other race/ancestry groups. In conclusion, our findings offer valuable insights into the complex genetics of PUFA levels that reflect, in part, their response to evolutionary pressures across the course of human history. Overall, our study underscores the importance of exploring the genetics of complex traits within diverse ancestry populations. Our study also emphasizes the necessity for ongoing and expanded genetic association research in cohorts with genetic ancestry that reflects that of the general population within the United States and worldwide.

In chapter 4, we performed Bayesian colocalization analysis and correlation network analysis (WGCNA) to prioritize 7 candidate genes (colocalization analysis: *DDX59, CAMSAP2, AC018816.3, BICC1, EIF2B2* and *PLEKHJ1,* WGCNA: *RBC*) of

CAD and subclinical atherosclerosis using multi-omics data from the TOPMed Multi-Ethnic Study of Atherosclerosis (MESA). In conclusion, our study illustrated the value of incorporating statistical fine-mapping into the Bayesian colocalization analysis, which allow us to identify colocalized genes in cases where both the GWAS and eQTL datasets exhibit multiple independent signals. Our study further demonstrated the importance for utilization of multi-ancestry GWAS of CAD and multi-ancestry MESA eQTL resources, which can reduce the risk of missing important genetic variants that may be prevalent in specific ancestry groups. Overall, these findings can provide a better understanding of genetic mechanisms and pathways implicated by GWAS of CAD and subclinical atherosclerosis and consequently provide valuable insights into potential therapeutic interventions and treatments.

5.2 Limitations and future directions

Although this thesis has provided deeper insights into the genetic and biological mechanisms underlying fatty acid metabolism and coronary artery disease, it is essential to acknowledge certain limitations. Future studies should aim to address these limitations and further enhance these lines of research.

In chapter 2, our study primarily focuses on urban Hispanic American populations, as represented by the MESA cohort, which include six major countries/regions of origin: Central America, Cuba, the Dominican Republic, Mexico, Puerto Rico, and South America. The differences of diet and lifestyle habits across the six Hispanic subgroups in MESA could be potential confounding factors in studying the effect of global proportion of Amerind ancestry on PUFAs. Additionally, it is important to note that we did not incorporate additional measures of dietary intake of n-3 and n-6 PUFAs in our regression analyses due to the unavailability of reliable measures for these dietary parameters among the MESA participants. However, measuring and quantifying dietary intake of n-3 AND PUFAs is potentially important for the further examination of the impact of dietary differences on the relationship between Amerind ancestry, *FADS* variation, and LC-PUFA levels.[1–3] For future research directions, it is advisable to develop a professional and detailed questionnaire aimed at assessing the dietary intake of n-3 and n-6 PUFAs. This questionnaire should include detailed information, such as a list of specific seafood and fish varieties available in the areas under investigation. Additionally, it should encompass items related to the consumption of walnuts, flaxseed, flaxseed oil, cod liver oil, and canola oil. Furthermore, the questionnaire should consider portion sizes and the frequency of consumption as crucial factors for accurately quantifying dietary intake of n-3 and n-6 PUFAs. By implementing such a questionnaire, future studies could achieve a more robust and accurate assessment of the dietary factors influencing PUFA metabolism and their associations with health outcomes.

In the chapter 3, while we have incorporated the GWAS results from multiple CHARGE cohorts (MESA, CHS and FHS), the overall sample size of our study is still relatively small for a meta-analysis of GWAS, especially for Hispanics Americans (Hispanic Americans: N = 1,454 and African Americans: N = 2,278). For the future work, we may incorporate other cohorts that include (1) Hispanic Americans/African Americans and (2) measurements of PUFAs, which can increase our sample size for GWAS of PUFAs. For instance, within our collaborative team, we have the expertise of Dr. Ski Chilton, who examines how genetic and epigenetic variations interact with human diets (especially the modern Western diet) to drive inflammation and inflammatory disorders (including cardiovascular disease and cancer). Dr. Chilton's work includes the measurement of PUFAs in Arizona, a region of particular significance due to its self-reported Hispanic and Amerind ancestry populations. It is worth noting that among all U.S. Hispanic populations, those residing in states in the Southwest, particularly those bordering Mexico, such as Arizona, exhibit the highest levels of Native American ancestry, which are not well represented in MESA.[4] Through collaboration with Dr. Chilton, we may be able to significantly increase our sample size for Hispanic Americans, which holds the potential to substantially enhance the statistical power of our genetic studies related to PUFAs. By leveraging a larger and more diverse dataset, we can enhance the robustness of our research and provide deeper insights into the genetic factors influencing PUFA metabolism within the Hispanic American population.

In our GWAS of PUFAs, our primary focus has been on studying the Hispanic American and African American populations. However, we have yet to explore the genetic impacts of PUFAs within the Asian population. It is important to note that African ancestry and certain South Asian populations exhibit a high frequency of a derived haplotype associated with efficient PUFA biosynthesis.[5–7] Furthermore, **Figure 5.1** illustrates a substantial degree of variability in the frequency of ancestral and derived *FADS* variants within Asian populations. Notably, the frequency of the derived haplotype exhibits a wide range, approximately 0.4 in East Asian populations to around 0.8 in South Asian populations. In the South Asian population, Kothapalli et al. conducted research that revealed positive selection for an insertion-deletion mutation (rs66698963) in *FADS2*, which leads to a more efficient biosynthesis of highly PUFAs.[8,9] These findings demonstrate that it may be worthwhile to extend our genetic studies to include some biobanks of Asian in the future, for example, Biobank Japan[10] and China Kadoorie Biobank[11]. Such an expansion would not only enrich our understanding of the genetic factors underpinning PUFA metabolism but also enable us to explore potential genetic links and variation across diverse ancestral backgrounds.

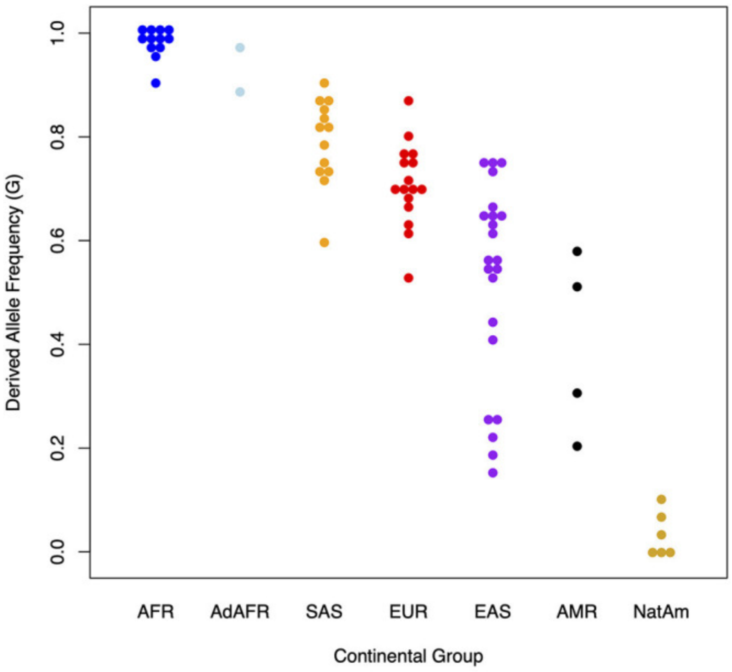


Figure 5.1: Presence and absence of the derived allele (G) at rs174537 in 80 globally diverse populations. Figure adapted from Chilton FH, 2022.

In the chapter 4, our study included Bayesian colocalization analysis by leveraging a large scale of multi-ancestry GWAS of CAD. It is worth noting, however, that the sample size of the eQTL resource we utilized in this analysis was somewhat limited. For instance, the TOPMed MESA included approximately 1200 samples for PBMCs and around 300 samples for each of T-cells and monocytes. Indeed, sample sizes can significantly impact the power to detect true associations, with smaller sample sizes leading to higher rates of false negatives. This limitation not only diminishes the statistical power of our study but also reduces the overall reliability and robustness of the research findings. To address the issue of limited sample size, the TOPMed MESA Principal Investigators, Dr. Stephen Rich and Dr. Jerome Rotter have worked towards expanding our sample size for the QTL analyses in MESA. Once these updated and larger resources become available, I will be able to conduct more comprehensive statistical analyses. Such efforts would enable us to compare the results obtained with the expanded sample size to the findings from our current research, potentially yielding deeper insights from our genetic investigations.

It is worth noting that machine learning and deep learning techniques are gaining increasing popularity in the field of genomics. These approaches offer multiple applications, including the imputation of molecular omics data. The utilization of machine learning and deep learning algorithms can enhance our ability to impute and analyze complex molecular data, providing valuable tools for advancing genetic research.[12] For example, a novel deep learning model known as the Sparse Convolutional Denoising Autoencoder (SCDA) has been developed for genotype imputation, eliminating the necessity for a reference panel. The SCDA model employs convolutional layers within the general autoencoder framework to effectively capture local data correlations and it addresses the challenges posed by high-dimensional genomic data through the incorporation of model sparsity.[13] Additionally, DeepImpute, a deep neural network-based imputation algorithm, was proposed for the imputation of single-cell RNA-seq data. This approach incorporates dropout layers and specialized loss functions to effectively learn and capture patterns within single-cell RNA-seq data, which significantly enhances the accuracy and power of imputation, making it a valuable tool for researchers working with single-cell RNA-seq data.[14] Single omics imputation methods serve as effective tools for mitigating the challenges posed by missing data within molecular omics datasets. For instance, variations in the number of proteins quantified using different versions of the SomaScan assay may lead to disparities, resulting in the absence of certain proteins in a particular SomaScan version.

In addition to single omics imputation methods, there is a growing trend towards integrative imputation techniques that leverage multi-omics data. For example, EpiXcan[15] for integrating epigenomic and transcriptomic data and cTP-net (single cell Transcriptome to Protein prediction with deep neural network)[16] for integrating transcriptomic and proteomic data. These multi-omics imputation methods are similar with PrediXcan, which is to predict the missing gene expression values leveraging large reference panels, including both genotype and gene expression information. Indeed, the application of these novel machine learning and deep learning techniques may help to address the issues of limited sample sizes in molecular omics data, providing a valuable foundation for to further understanding of molecular and genetic mechanisms underlying complex disease phenotypes.

Another notable limitation of our study is the use of eQTL resources based on blood tissue for the discovery analysis. While blood offers valuable insights, it may not be the tissue most directly related to CAD. To further enhance the relevance of our research, it would be advantageous to leverage eQTL data from disease-related tissues available in existing databases. For instance, GTEx provides a broad eQTL resource covering 53 human tissues, including Aorta, Coronary, and Tibial tissues, which are more closely associated with CAD. Additionally, the Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task (STARNET) study represents another valuable eQTL resource that could be incorporated in future efforts to augment our primary analysis. STARNET recruited 600 well-characterized CAD patients and sequenced RNA isolated from various tissues, including blood, atherosclerotic-lesion-free internal mammary artery (MAM), atherosclerotic aortic root (AOR), subcutaneous fat (SF), visceral abdominal fat (VAF), skeletal muscle (SKLM), and liver (LIV), identifying ~8 million cis-eQTLs.[17] Incorporating data from these resources could further enrich our genetic studies and provide a more refined understanding of the genetic factors associated with CAD.

Moreover, the colocalization analysis within this study has been primarily directed towards the intersection of expression quantitative trait loci (eQTL) and genome-wide association studies (GWAS). Nevertheless, it is noteworthy that the TOPMed MESA dataset encompasses multiple molecular quantitative trait loci (QTL) resources, including eQTL, protein QTL (pQTL), and methylation QTL (mQTL) data. The integration of GWAS summary data with more than one molecular QTL resources simultaneously provided the capacity of the identification of regulatory effects at GWAS risk loci.[18,19] For example, integration of GWAS, eQTL and mQTL is able to identify the candidate genes that influence the diseases under investigation through methylation. An efficacious Bayesian statistical framework, denoted as Multiple-trait-coloc (moloc), has been employed in this context. Moloc serves to quantify the degree of evidence in support of a common causal variant within a specific risk region, across a multitude of traits (molecular traits or complex disease traits) by utilizing summary-level information derived from genetic association datasets.[19] **(Figure 5.2)** Through this stratification of data, moloc not only enhances statistical power but also furnishes valuable insights into the functional significance of implicated genes, which improve our understanding of the genetic determinants of complex diseases.

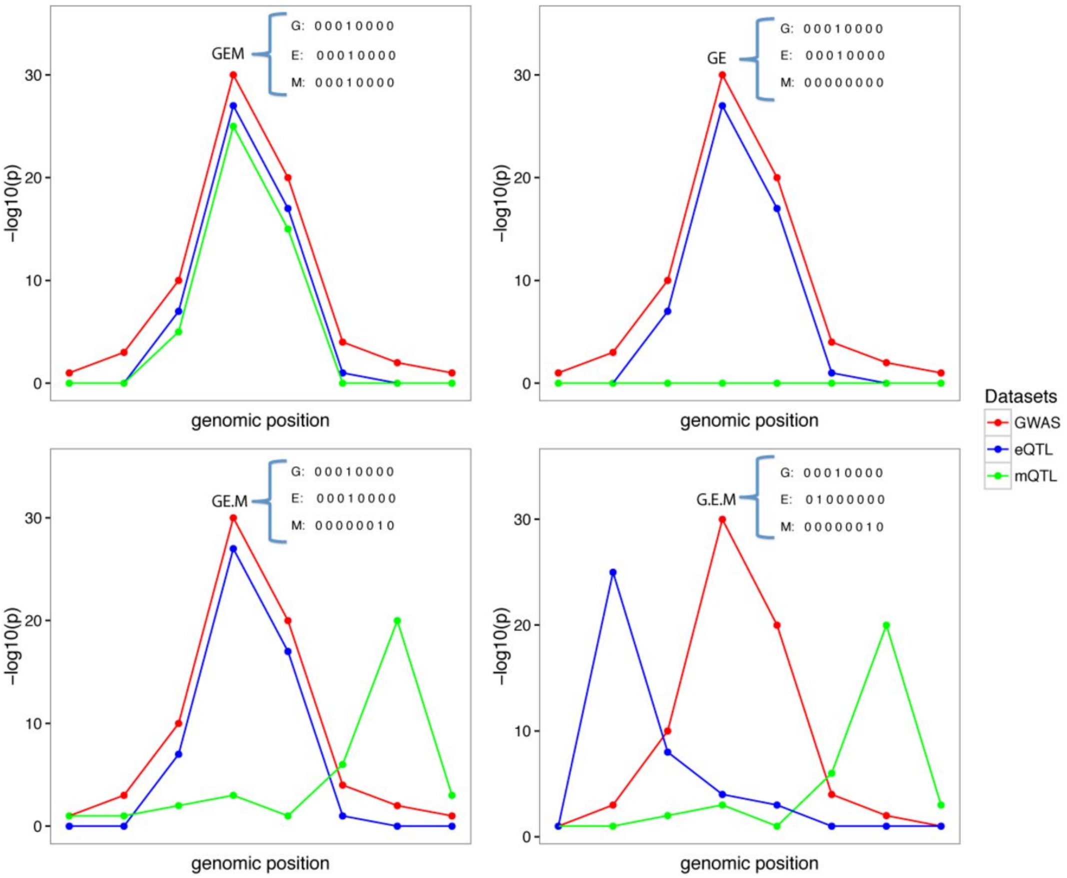


Figure 5.2: Multiple-trait-coloc (moloc): Graphical representation of four possible configurations at a locus with eight SNPs in common across three traits. Figure adapted from Giambartolomei C, 2018.

Furthermore, the current study did not undertake a comprehensive exploration of the impact of the admixed genome structure on eQTL mapping. This limitation has the potential to introduce an increased occurrence of false positive eQTL associations, not attributed to the genetic variants under investigation, but rather to their associations with specific local ancestral backgrounds.[20,21] An illustrative example of this significance is found in the work of Zhong Y et al., who introduced a robust statistical framework known as Joint-GaLA-QTLM. This approach incorporates variant-level local ancestry as a covariate in eQTL mapping, shedding light on the impact of uncertainty in local ancestry estimation on the regulatory effects of genetic variations on gene expression.

Incorporating local ancestry can increase the power of eQTL mapping, leading to the detection of more genuine genetic associations and the reduction of the risk of spurious associations arising from different local ancestral backgrounds.[20] In essence, the integration of local ancestry information not only refines the precision of eQTL mapping but also advances our comprehension of how population structure exerts its influence on genetic associations. Furthermore, it enriches our understanding of the intricate genetic foundations underlying complex traits in the context of diverse and admixed populations.[20,22–24]

Additionally, this study primarily employed Bayesian colocalization analysis for the identification of causal genes associated with CAD and subclinical atherosclerosis. However, it is noteworthy that there exist alternative approaches to further prioritize the causal genes implicated in these complex traits. One such approach is Mendelian Randomization, which is specifically designed for the investigation of functionally relevant genes within the loci identified in GWAS for complex traits.[25,26] Mendelian Randomization leverages genetic variants as instrumental variables, enabling the rigorous assessment of causal relationships between exposures, such as specific genes, and outcomes, in this case, CAD. The steps of Mendelian Randomization include (1). Identification of genetic variants associated with the gene of interest, which is believed to be a potential driver of CAD; (2). It is important to ensure that the selected genetic variants strongly associated with the gene of interest but not significantly associated with any known confounding factors. (3). Investigating whether the gene of interest has a causal effect on CAD risk by assessing the relationship between the genetic variants (linked to the gene of interest) and CAD outcomes using Mendelian

Randomization techniques, such as two-stage least squares regression or the inverse-variance weighted analysis.[27] The application of Mendelian Randomization can provide robust evidence of causality, helping to establish whether specific genes have a causal role in the development or progression of CAD. Furthermore, the findings of Mendelian Randomization have the capacity to serve as a guidance for the development of precise therapeutic interventions designed to modulate the activity of these genes, with the ultimate aim of diminishing CAD risk.[28–31]

It is crucial to emphasize the importance of extensive collaboration between computational scientists and experimental scientists in the pursuit of scientific excellence. As a computational biologist, my role involves applying statistical methods to identify various disease-associated variants and biomarkers. However, the full understanding of the biological background and significance of these findings often requires experimental validation, a task that falls within the expertise of experimental scientists.

In the context of our research outlined in Chapter 4, I have identified a promising list of candidate causal genes for CAD and subclinical atherosclerosis using Bayesian colocalization analysis and correlation network analysis. To advance understanding of these prioritized candidates, we need to work further to foster collaborations with experts in experimental cardiovascular research, including Dr. Clint Miller, Dr. Coleen McNamara, and Dr. Weibin Shi. These distinguished investigators possess specialized knowledge and experience in cardiovascular diseases. By collaborating closely with these experts for follow-up experimental validation, we have the potential to gain in-depth insight into the genetic mechanisms underlying CAD and subclinical atherosclerosis. Overall, such collaborative efforts would foster connections between computational and experimental sciences, ultimately advancing our understanding of complex biological processes and contributing to scientific progress in the field of complex diseases.

In conclusion, this PhD work demonstrates insights that can be gained through generation of a large scale of molecular ‘omics data coupled with development and application of a statistical and computational methods for interpretation of these data, Our studies enhance the identification of causal disease-related variants and biomarkers, and further deepen our understanding of genetic mechanisms and pathways for the diseases under investigation. Ultimately, these collective efforts hold the potential to make substantial contributions to the development and advancement of prevention strategies and therapeutic treatments.

5.3 References

1. Sublette, M. E. *et al.* Validation of a food frequency questionnaire to assess intake of n-3 polyunsaturated fatty acids in subjects with and without Major Depressive Disorder. *J. Am. Diet. Assoc.* **111**, 117-123.e2 (2011).
2. Herter-Aeberli, I. *et al.* Validation of a Food Frequency Questionnaire to Assess Intake of n-3 Polyunsaturated Fatty Acids in Switzerland. *Nutrients* **11**, 1863 (2019).
3. Ahn, Y. *et al.* Validation and reproducibility of food frequency questionnaire for Korean genome epidemiologic study. *Eur. J. Clin. Nutr.* **61**, 1435–1441 (2007).
4. Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D. & Mountain, J. L. The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. *Am. J. Hum. Genet.* **96**, 37–53 (2015).
5. Ameur, A. *et al.* Genetic adaptation of fatty-acid metabolism: a human-specific haplotype increasing the biosynthesis of long-chain omega-3 and omega-6 fatty acids. *Am. J. Hum. Genet.* **90**, 809–820 (2012).
6. Mathias, R. A. *et al.* Adaptive evolution of the FADS gene cluster within Africa. *PloS One* **7**, e44926 (2012).
7. Chilton, F. H. *et al.* Precision Nutrition and Omega-3 Polyunsaturated Fatty Acids: A Case for Personalized Supplementation Approaches for the Prevention and Management of Human Diseases. *Nutrients* **9**, (2017).
8. Kothapalli, K. S. D. *et al.* Positive Selection on a Regulatory Insertion–Deletion Polymorphism in FADS2 Influences Apparent Endogenous Synthesis of Arachidonic Acid. *Mol. Biol. Evol.* **33**, 1726–1739 (2016).

9. Chilton, F. H. *et al.* Interpreting Clinical Trials With Omega-3 Supplements in the Context of Ancestry and FADS Genetic Variation. *Front. Nutr.* **8**, 808054 (2021).

10. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).

11. Chen, Z. *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.* **40**, 1652–1666 (2011).

12. Song, M. *et al.* A Review of Integrative Imputation for Multi-Omics Datasets. *Front. Genet.* **11**, 570255 (2020).

13. Chen, J. & Shi, X. Sparse Convolutional Denoising Autoencoders for Genotype Imputation. *Genes* **10**, 652 (2019).

14. Arisdakessian, C., Poirion, O., Yunits, B., Zhu, X. & Garmire, L. X. DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data. *Genome Biol.* **20**, 211 (2019).

15. Zhang, W. *et al.* Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits. *Nat. Commun.* **10**, 3834 (2019).

16. Zhou, Z., Ye, C., Wang, J. & Zhang, N. R. Surface protein imputation from single cell transcriptomes by deep neural networks. *Nat. Commun.* **11**, 651 (2020).

17. Cardiometabolic Risk Loci Share Downstream Cis- and Trans-Gene Regulation Across Tissues and Diseases - PMC. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5534139/.

18. Foley, C. N. *et al.* A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nat. Commun.* **12**, 764 (2021).

19. Giambartolomei, C. *et al.* A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34**, 2538–2545 (2018).

20. Zhong, Y., Perera, M. A. & Gamazon, E. R. On Using Local Ancestry to Characterize the Genetic Architecture of Human Traits: Genetic Regulation of Gene Expression in Multiethnic or Admixed Populations. *Am. J. Hum. Genet.* **104**, 1097–1115 (2019).

21. Storey, J. D. *et al.* Gene-Expression Variation Within and Among Human Populations. *Am. J. Hum. Genet.* **80**, 502–509 (2007).

22. Wang, X. *et al.* Adjustment for local ancestry in genetic association analysis of admixed populations. *Bioinformatics* **27**, 670–677 (2011).

23. Li, B. *et al.* Incorporating local ancestry improves identification of ancestry-associated methylation signatures and meQTLs in African Americans. *Commun. Biol.* **5**, 401 (2022).

24. Gay, N. R. *et al.* Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. *Genome Biol.* **21**, 233 (2020).

25. Sanderson, E. *et al.* Mendelian randomization. *Nat. Rev. Methods Primer* **2**, 6 (2022).

26. Davey Smith, G. & Ebrahim, S. ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease?*. *Int. J. Epidemiol.* **32**, 1–22 (2003).

27. Davies, N. M., Holmes, M. V. & Davey Smith, G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *The BMJ* **362**, k601 (2018).

28. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).

29. Smith, G. D. & Ebrahim, S. Mendelian Randomization: Genetic Variants as Instruments for Strengthening Causal Inference in Observational Studies. in *Biosocial Surveys* (National Academies Press (US), 2008).

30. Wang, K. *et al.* Mendelian randomization analysis of 37 clinical factors and coronary artery disease in East Asian and European populations. *Genome Med.* **14**, 63 (2022).

31. Jansen, H., Samani, N. J. & Schunkert, H. Mendelian randomization studies in coronary artery disease. *Eur. Heart J.* **35**, 1917–1924 (2014).

Appendix A

Supplementary Data

All supplementary data are available at:

Yang C, Hallmark B, Chai JC, O'Connor TD, Reynolds LM, Wood AC, Seeds M, Chen YI, Steffen LM, Tsai MY, Kaplan RC, Daviglus ML, Mandarino LJ, Fretts AM, Lemaitre RN, Coletta DK, Blomquist SA, Johnstone LM, Tontsch C, Qi Q, Ruczinski I, Rich SS, Mathias RA, Chilton FH, Manichaikul A. Impact of Amerind ancestry and FADS genetic variation on omega-3 deficiency and cardiometabolic traits in Hispanic populations. Commun Biol. 2021 Jul 28;4(1):918. doi: 10.1038/s42003-021-02431-4

Yang C, Veenstra J, Bartz TM, Pahl MC, Hallmark B, Chen YI, Westra J, Steffen LM, Brown CD, Siscovick D, Tsai MY, Wood AC, Rich SS, Smith CE, O'Connor TD, Mozaffarian D, Grant SFA, Chilton FH, Tintle NL, Lemaitre RN, Manichaikul A. Genome-wide association studies and fine-mapping identify genomic loci for n-3 and n-6 polyunsaturated fatty acids in Hispanic American and African American cohorts. Commun Biol. 2023 Aug 16;6(1):852. doi: 10.1038/s42003-023-05219-w.

https://zenodo.org/record/8428481

Supplementary Data 2.1. Regression of n-3 and n-6 PUFAs on global ancestry, also accounting for local ancestry and rs174537 SNP genotype.

Supplementary Data 2.2. Interaction of rs174537 with global and local ancestry

Supplementary Data 2.3. Genotypic effects of rs174537 on fasting lipids, anthropometric and inflammatory traits

Supplementary Data 2.4. Genotypic effects of rs174557 on Fasting Lipids, Anthropometrics and Inflammatory Traits.

Supplementary Data 2.5. Participant characteristics for the AIR registry.

Supplementary Data 2.6. Participant characteristics for the HCHS/SOL cohort.

Supplementary Data 2.7. Analysis of rs174537 effects on triglycerides and waist-hip ratio in AIR registry and HCHS/SOL.

Supplementary Data 2.8. Stratified analysis of rs174537 effects on triglycerides and waist-hip ratio HCHS/SOL, with additional adjustment for principal components of ancestry.

Supplementary Data 2.9. Regression of n-3 and n-6 PUFAs on Study Site in MESA Hispanic Participants

Supplementary Data 3.1. Examination of variants identified in prior CHARGE European American GWAS of PUFAs.

Supplementary Data 3.2. Credible sets of putative causal variants for each of the PUFAs in Hispanic Americans

Supplementary Data 3.3. Credible sets of putative causal variants for eachof the PUFAs in African Americans.

Supplementary Data 3.4. Chromatin Contact Analysis for Understanding the Chromatin Interaction on Chromosome 11.

Supplementary Data 3.5. Cross-ancestry Replication for lead variants from credible sets of putative causal variants in Hispanic Americans.

Supplementary Data 3.6. Cross-ancestry Replication for lead variants from credible sets of putative causal variants in African Americans.

Supplementary Data 3.7. Source Data for Figure 3.3b

Supplementary Data 3.8. Association with lipid traits for lead variants from credible sets of putative causal variants in Hispanic Americans.

Supplementary Data 3.9. Association with lipid traits for lead variants from credible sets of putative causal variants in African Americans.

Supplementary Data 3.10. Colocalization analysis using MESA multi-ancestry eQTL in Hispanic Americans.

Supplementary Data 3.11. Colocalization analysis using GTEx multi-ancestry eQTL in Hispanic Americans.

Supplementary Data 3.12. PrediXcan results using MESA expression prediction models in Hispanic Americans.

Supplementary Data 3.13. PrediXcan results using GTEx expression prediction models in Hispanic Americans.

Supplementary Data 3.14. Gene set enrichment analysis for genes implicated by colocalization and PrediXcan analysis.

Supplementary Data 4.1. Identification of colocalized genes in Artery tissues

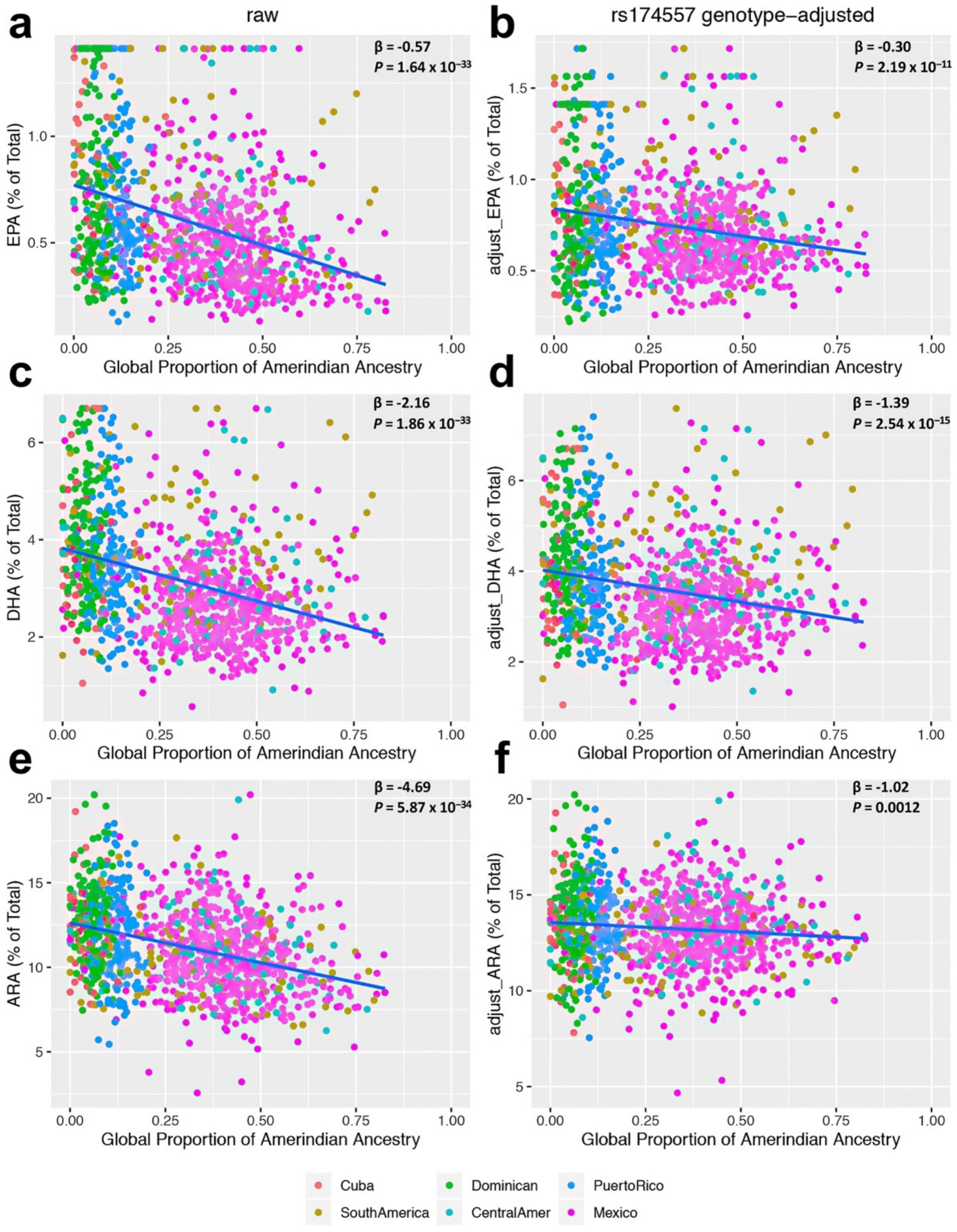
Supplementary Data 4.2. Association of colocalized genes with subclinical atherosclerosis

Supplementary Data 4.3. Association of modules from WGCNA with subclinical atherosclerosis

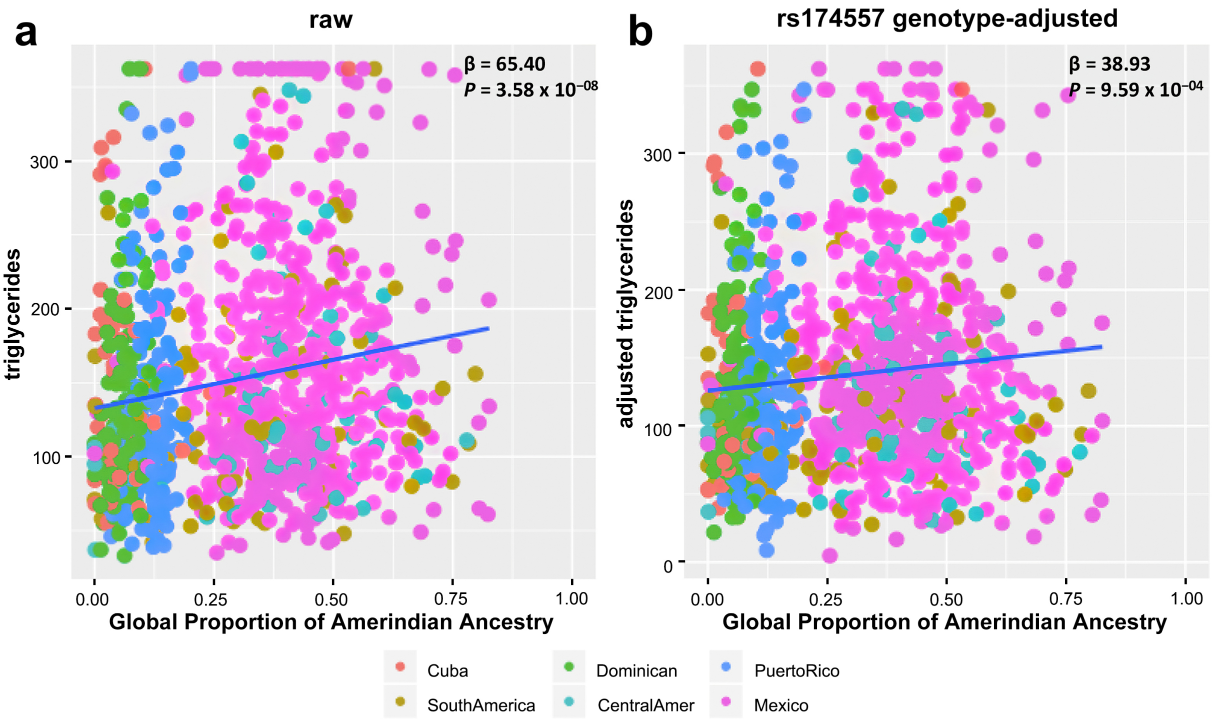
Supplementary Data 4.4. Pathway enrichment analysis of subclinical atherosclerosis related modules.

Appendix B

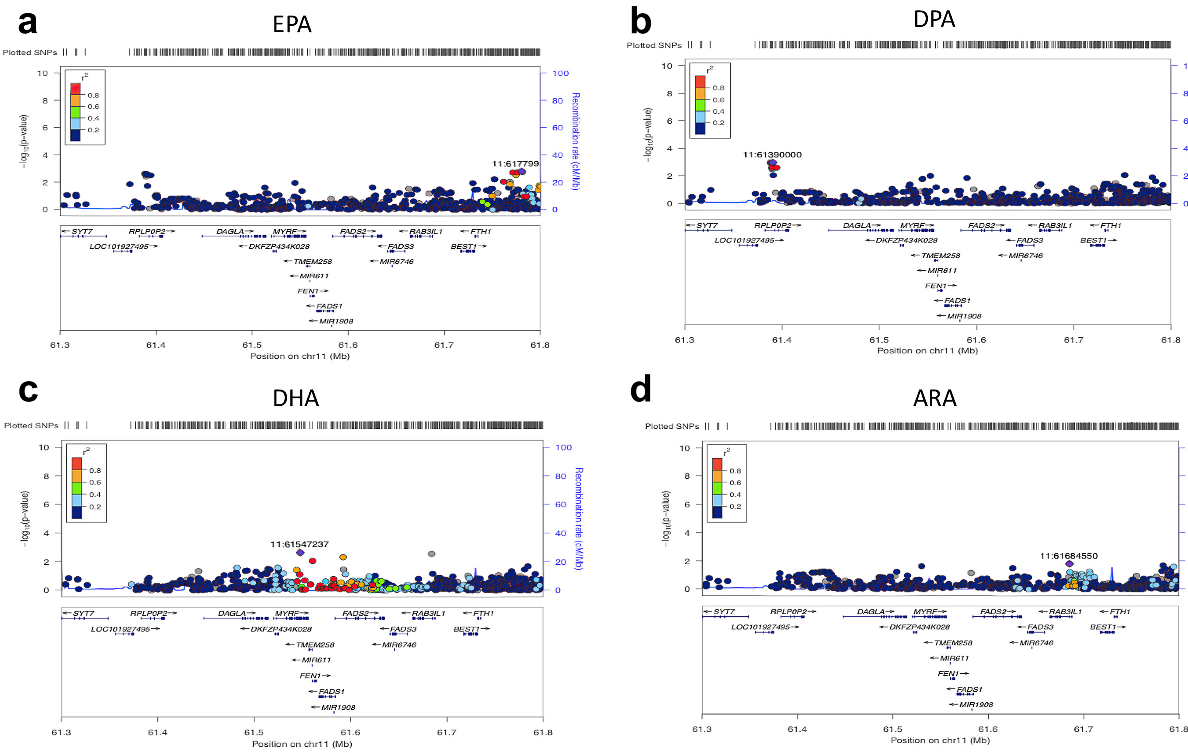
Supplementary Figures



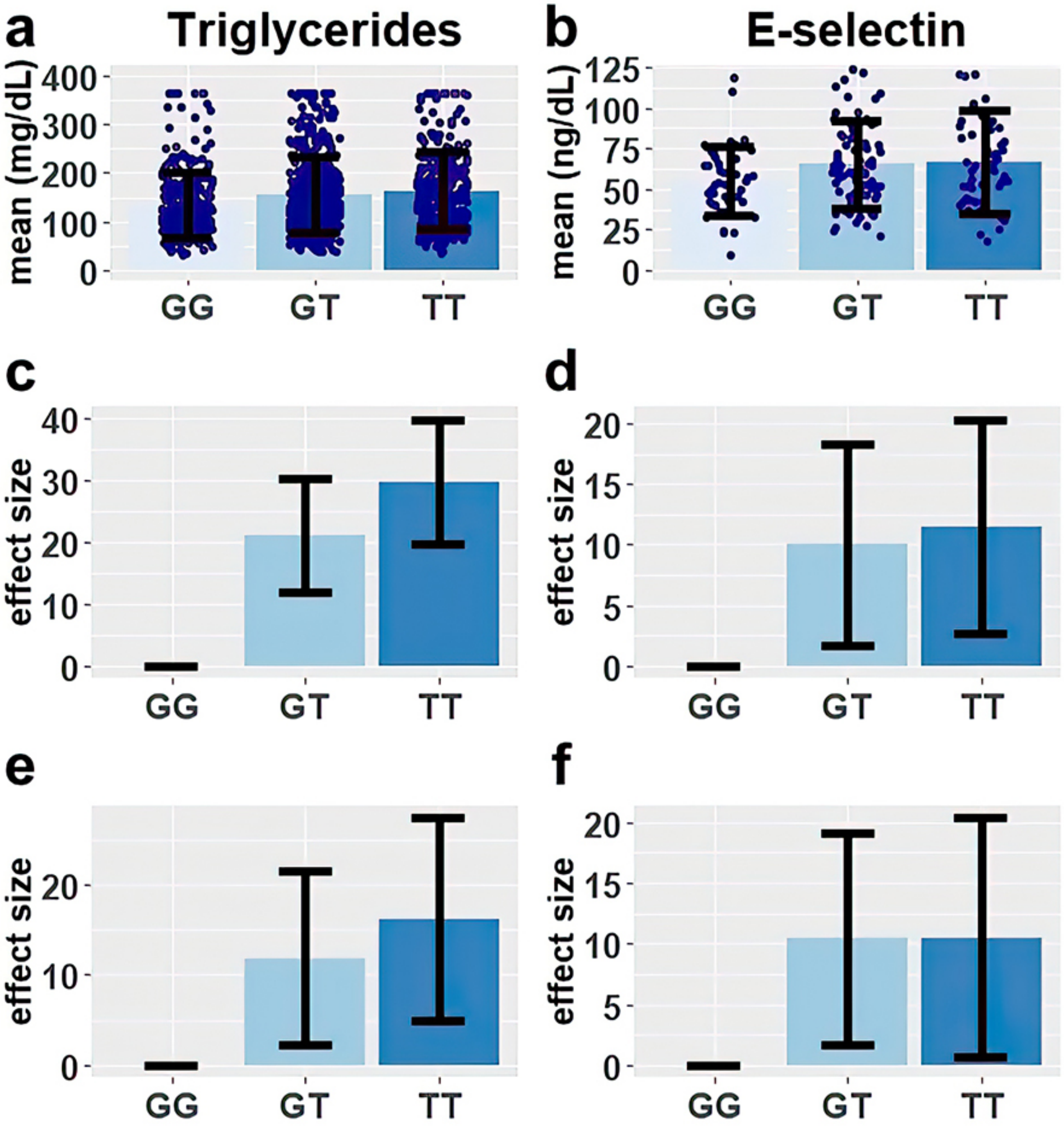
Supplementary Figure 2.1 Relationship of LC-PUFA levels with Global Proportion of Amerind Ancestry before and after adjustment for rs174557 genotype The regression effect estimates (β expressed as % of total fatty acids) and *P*-values are shown in the upper right corner of each panel. The relationship of LC-PUFA levels with Global Proportion of Amerind Ancestry as estimated from genome-wide SNP data is shown for (a) EPA - raw, (b) EPA – genotype-adjusted, (c) DHA – raw, (d) DHA – genotyped-adjusted, (e) ARA – raw, and (f) ARA – genotype-adjusted. Here, the rs174557 genotype-adjusted LC-PUFA levels were obtained as residuals after regression against rs174557 genotype and re-centered around the raw mean. *P*-values are presented based on two-sided t-tests for each regression coefficient derived with n = 1102 biologically independent samples.



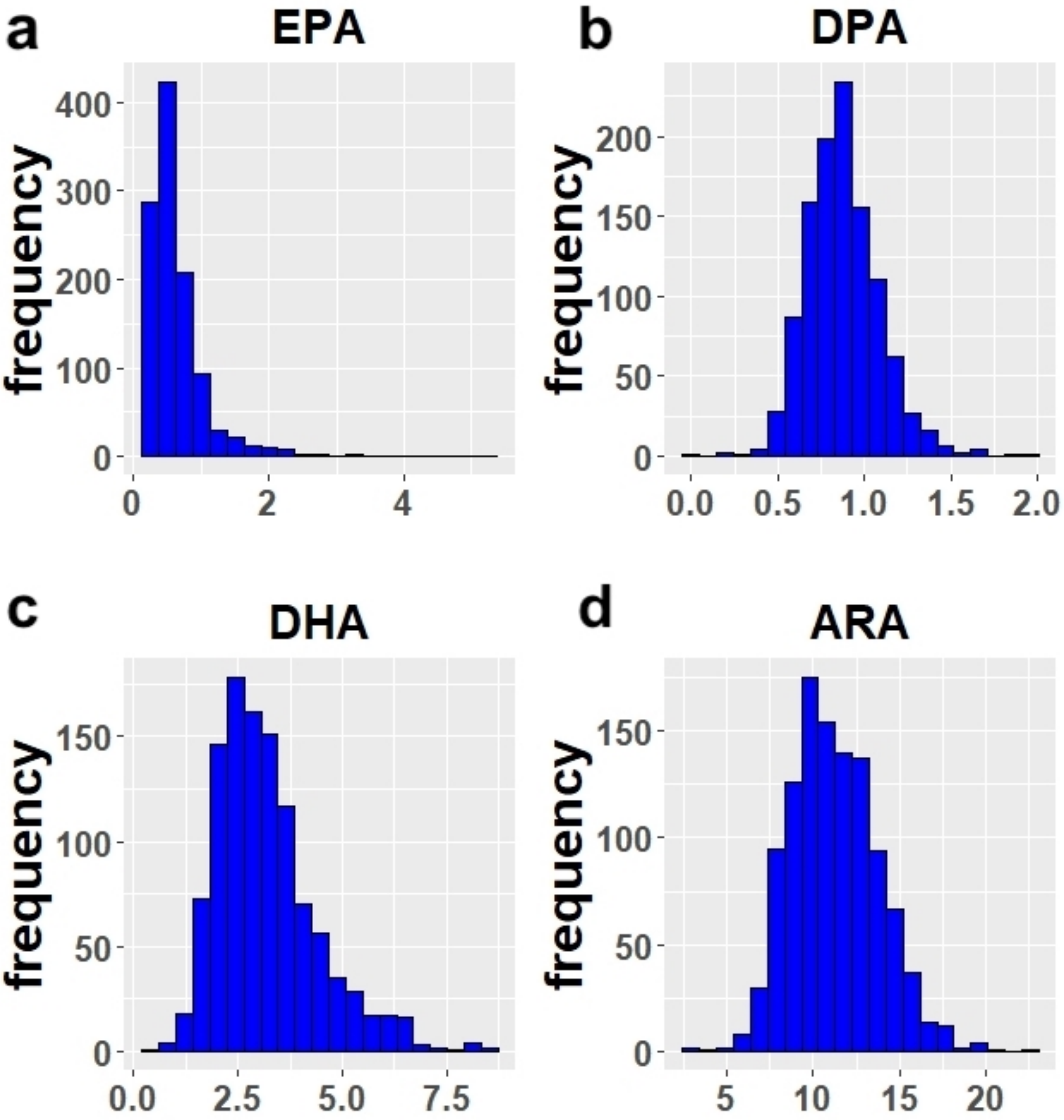
Supplementary Figure 2.2 Relationship of triglycerides with Global Proportion of Amerind Ancestry before and after adjustment for rs174557 genotype. The regression effect estimates (β in mg/dL) and *P*-values are shown in the upper right corner of each panel. The relationships are shown for each of (a) raw triglyceride levels, and (b) genotype-adjusted tryglyceride levels with Global Proportion of Amerind Ancestry. Here, rs174557 genotype-adjusted triglyceride levels were obtained as residuals from regression accounting for rs174557 genotype, and re-centered around the raw means. *P*-values are presented based on two-sided t-tests for each regression coefficient derived with n = 1101 biologically independent samples.



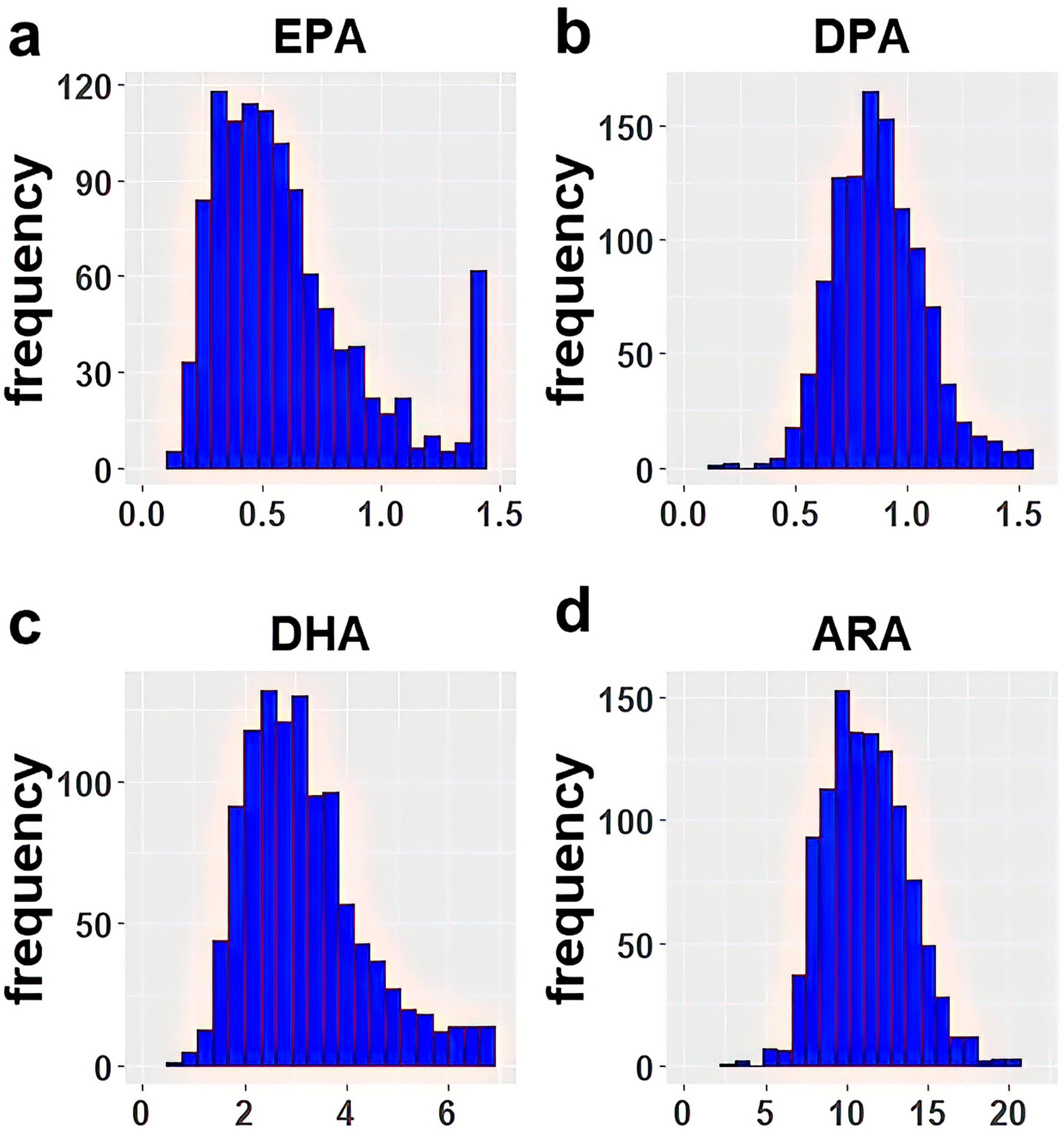
Supplementary Figure 2.3 Conditional local association plots for n-3 and n-6 PUFAs in the FADS1/2 region, accounting for the rs174537 SNP. The plots present -log10 p-values for the association of each genetic variant within the region shown with the fatty acid traits. *P*-values were derived based on two-sided t-tests for the genetic additive effects with n = 1,102 biologically independent samples. Genetic coordinates on the x-axis are denoted based on human genome Build 37.



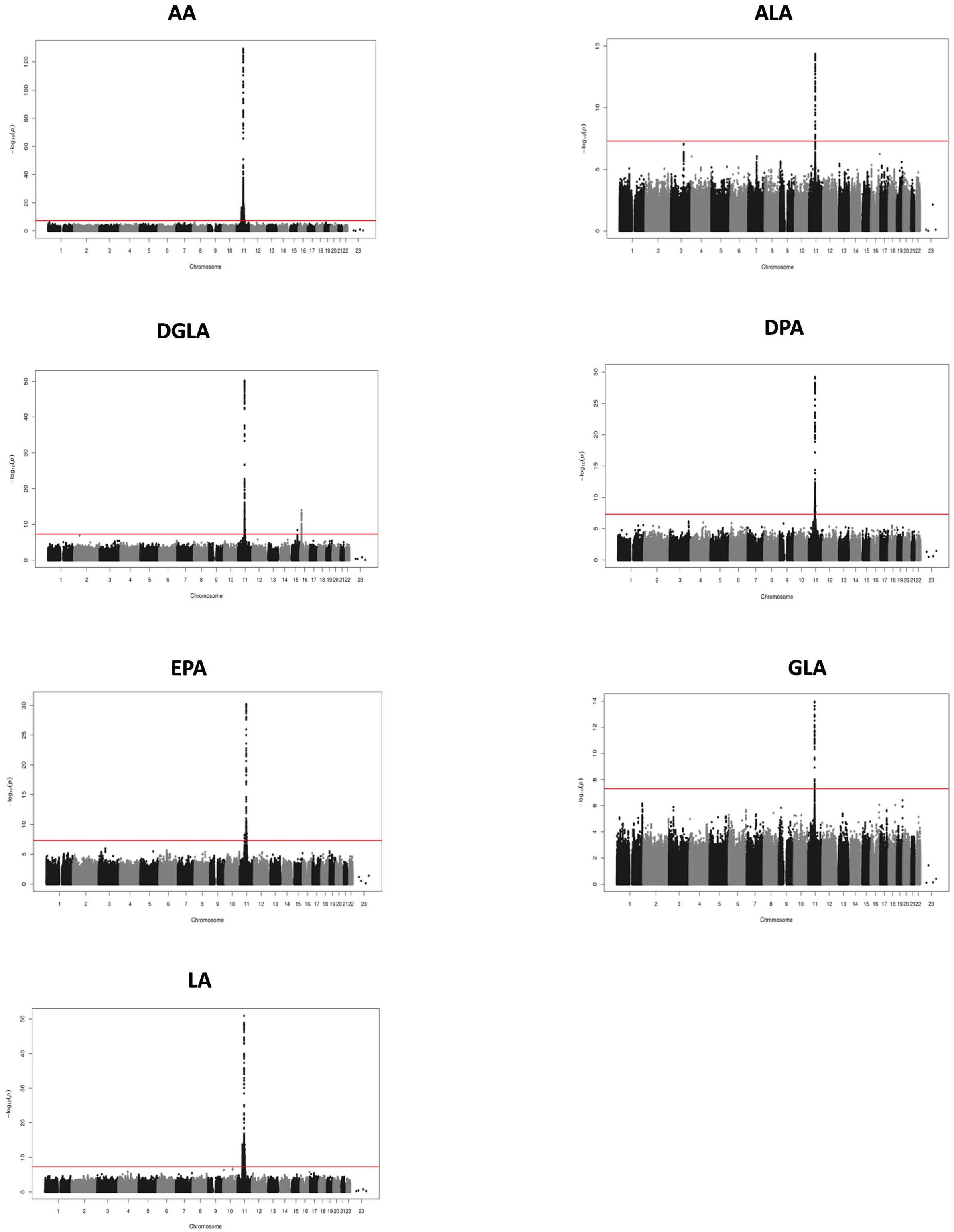
Supplementary Figure 2.4 Genotypic effects of rs174537 on triglycerides and E-selectin. Mean and standard deviation of (a) triglycerides, and (b) E-selectin stratified by the genotypes of rs174537. Estimated effect and standard error among participants carrying one or two copies of the ancestral allele T (compared to the reference of zero) for (c) triglycerides, and (d) E-selectin after adjustment for age and sex. Estimated effect and standard error among participants carrying one or two copies of the ancestral allele T (compared to a reference of zero) for (e) triglycerides; and (f) E-selectin, after adjustment for age, sex and principal components of ancestry. The sample sizes are 1101 (GG: 293 ;GT: 483 ;TT: 325) for triglycerides and 183 (GG: 48; GT: 76; TT: 59) for E-Selectin.



Supplementary Figure 2.5 Raw distribution of n-3 and n-6 PUFAs in MESA Hispanic participants. Values for (a) EPA, (b) DPA, (c) DHA, and (d) ARA are presented in units of % of total fatty acids.



Supplementary Figure 2.6: Distribution of n-3 and n-6 PUFAs in MESA Hispanic participants after winsorizing at median +/- 3.5 Median Absolute Deviation (MAD). Values for (a) EPA, (b) DPA, (c) DHA, and (d) ARA are presented in units of % of total fatty acids.


AA

ALA

DGLA

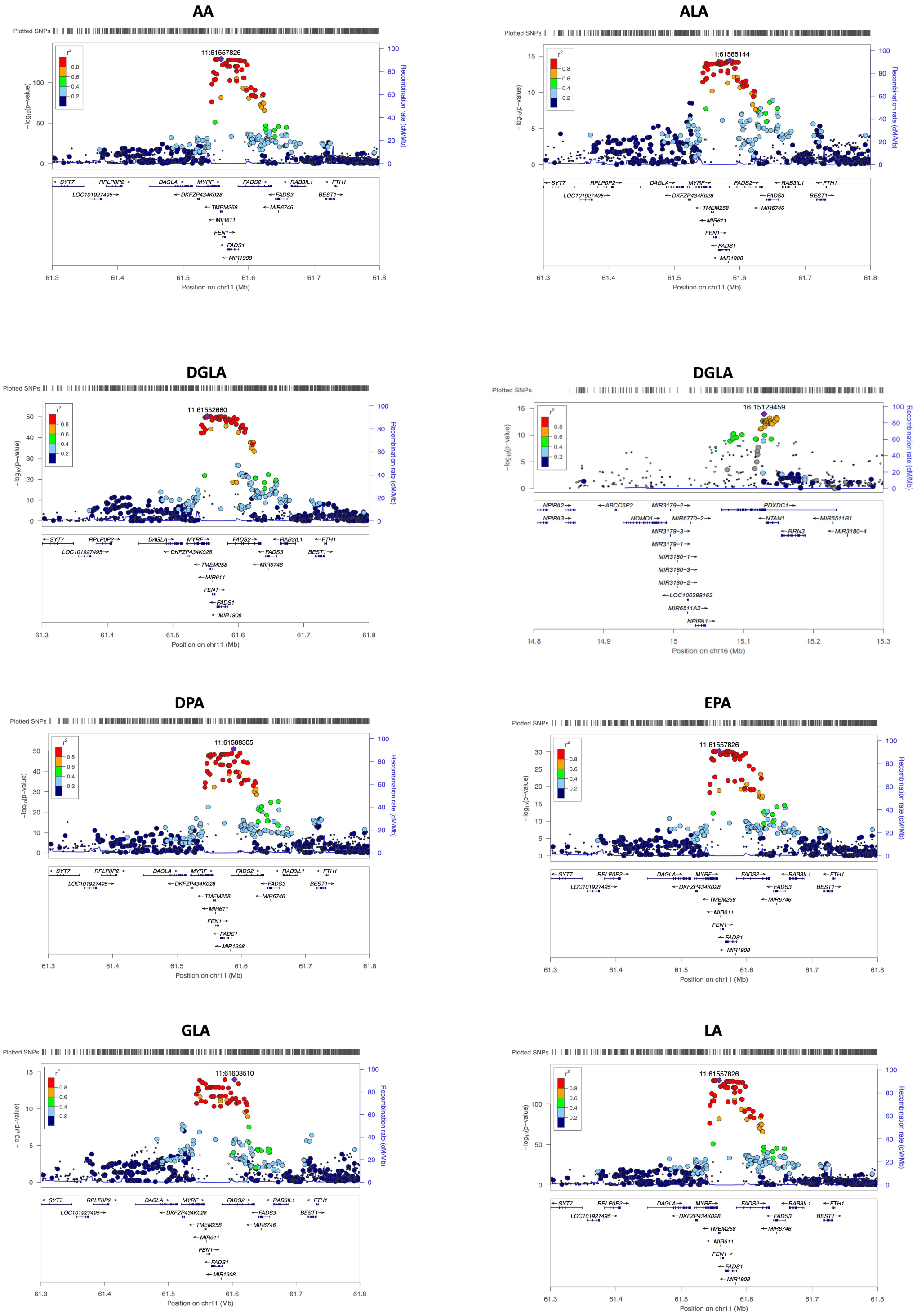
DPA

EPA

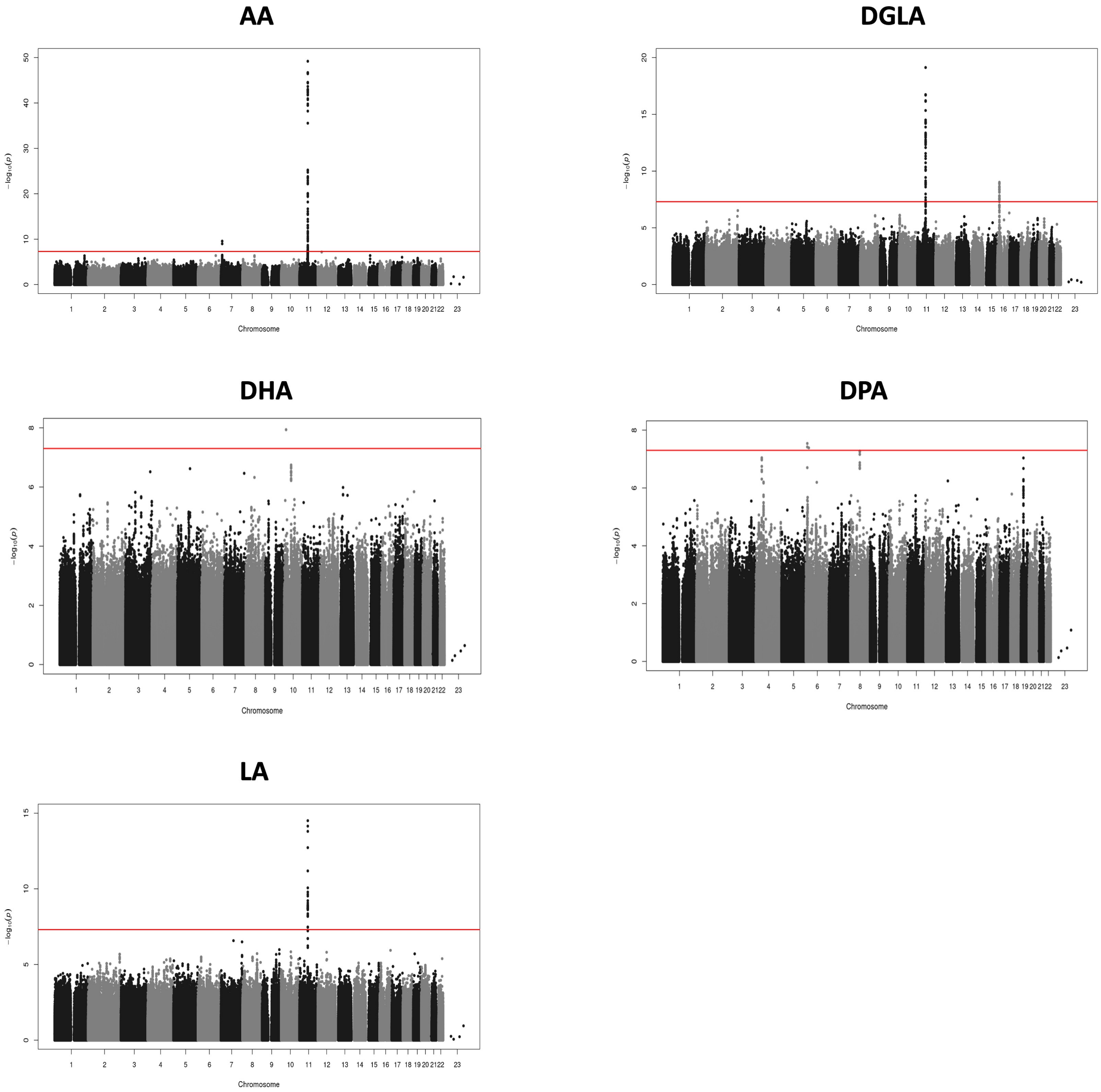
GLA

LA

Supplementary Figure 3.1 Manhattan plot of meta-analysis of GWAS for PUFAs in the Hispanic Americans. Manhattan plots for PUFAs exhibiting one or more genome-wide significant association in the Hispanic Americans. The red line shows the genome-wide significance threshold of (–log10(5 x 10^{-8})).



Supplementary Figure 3.2. Local association plot of most significant for each PUFAs on each chromosome in the Hispanic Americans. Local association plots of the most significant region for each PUFA on each chromosome harboring at least one genome-wide significant signal in the Hispanic Americans. Reference panel used to color LD is 1000 Genomes Ad Mixed American (AMR) and the color scheme is red for strong linkage disequilibrium (LD; r2≥0.8) and blue color for lower LD.



AA

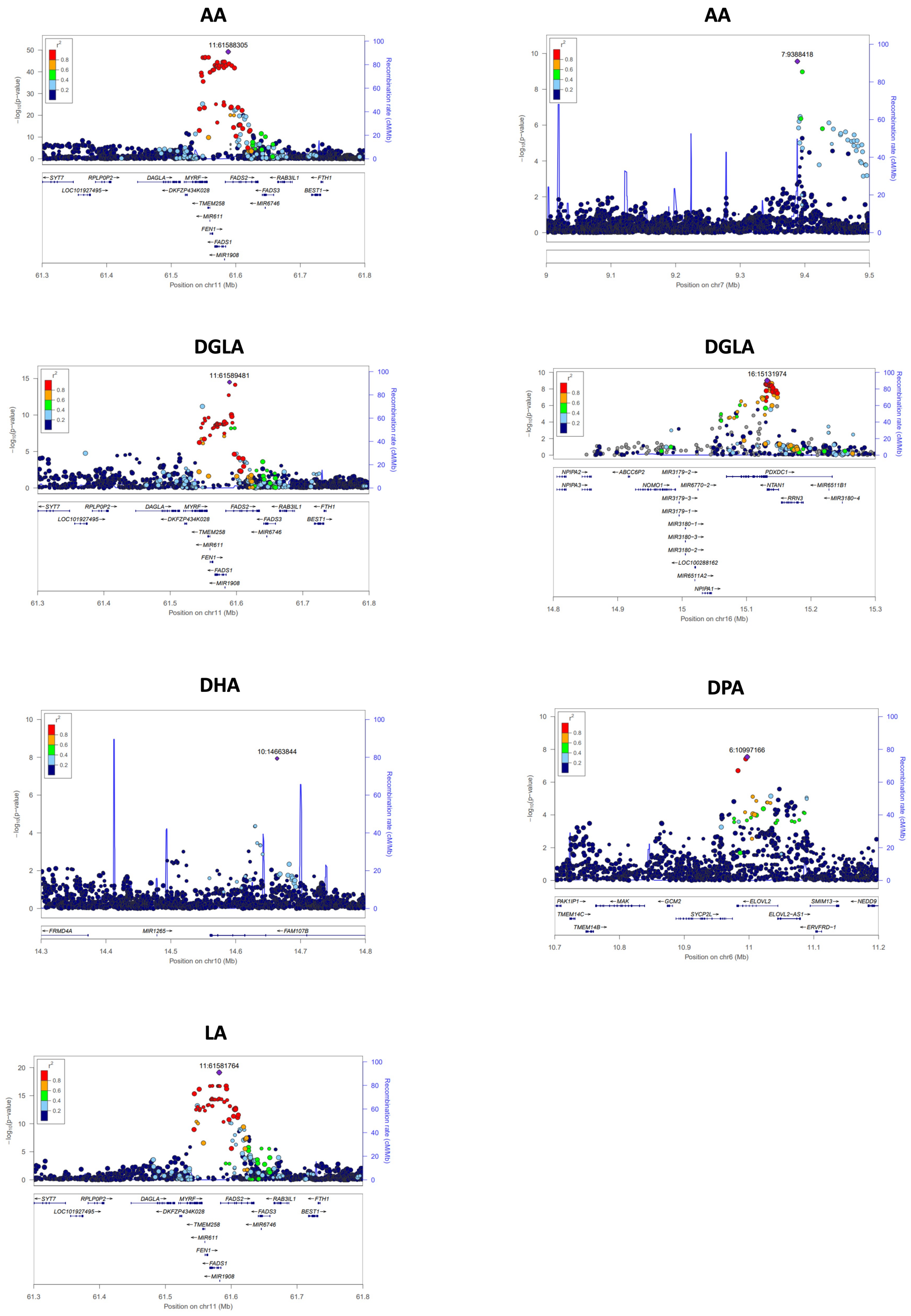
DGLA

DHA

DPA

LA

Supplementary Figure 3.3 Manhattan plot of meta-analysis of GWAS for PUFAs in the African American population. Manhattan plots for PUFAs exhibiting one or more genome-wide significant association in the African Americans. The red line shows the genome-wide significance threshold of (–log10(5 x 10^{-8})).


AA

AA

DGLA

DGLA

DHA

DPA

LA

Supplementary Figure 3.4 Local association plot of most significant for each PUFAs on each chromosome in the African Americans. Local association plots of the most significant region for each PUFA on each chromosome harboring at least one genome-wide significant signal in the African Americans. Reference panel used to color LD is 1000 Genomes African (AFR) and the color scheme is red for strong linkage disequilibrium (LD; r2≥0.8) and blue color for lower LD.