NUTRIGENETICS OF POLYUNSATURATED FATTY ACIDS AND VEGETARIANISM

by

MICHAEL FRANCIS

(Under the Direction of Kaixiong Ye)

ABSTRACT

Nutrigenetics and precision nutrition provide the means to create targeted dietary therapy for disease treatment and prevention, based on an individual's unique genetic composition. Genome-wide association and interaction studies (GWAS and GWIS) are statistical techniques that can help identify nutrigenetic variants of interest; three of these studies are collected herein. In Chapter 2, we performed a GWIS for interactions of genotype with fish oil on blood lipids. We found that rs112803755 had significant interaction with triglycerides ($P = 5.65e{-}10$), in a meta-analysis of UK Biobank (UKB) and Atherosclerosis Risk in Communities (ARIC) cohorts. Specifically, there was a triglyceride-lowering effect of fish oil supplementation on rs112803755 heterozygotes. This locus was significantly associated with higher *GJB2* expression of connexin 26 in adipose tissue; connexin activity is known to change upon exposure to omega-3 fatty acids. In Chapter 3, we performed a GWAS to identify variants associated with circulating polyunsaturated and monounsaturated fatty acid (PUFA and MUFA) levels, which have strong heritable components. We meta-analyzed GWAS from three European studies, UKB, FinMetSeq, and Kettunnen et al., and identified 87 novel loci, 51 of which were replicated. Functional analyses enabled selection of candidate genes for these loci. In Chapter 4, we performed traditional and genetic epidemiology analysis on the effects of vegetarianism. We designed a rigorous definition of vegetarianism using data from two surveys in the UK Biobank to identify a cohort of "reliable" vegetarians. Significant effects of vegetarianism were found in 15 of 30 serum biomarkers; all cholesterol measures plus Vitamin D were significantly lower in vegetarians, while triglycerides were higher. GWIS detected evidence of gene-vegetarianism interaction with one genome-wide significant variant at rs72952628 ($P = 4.47e{-}08$) on calcium. rs72952628 is located in *MMAA*, which is part of the B_{12} metabolism pathway; B_{12} has a high deficiency potential in vegetarians. Gene-based aggregation of interaction *P*-values revealed two additional significant genes, *RNF168* in testosterone ($P = 1.45e{-}06$), and *DOCK4* in eGFR ($P = 6.76e{-}07$), which have previously been associated with testicular and renal traits, respectively. These studies illustrate nutrigenetic variants are being rapidly discovered, and underscore the need for personalized nutrition.

INDEX WORDS: Nutrigenetics, Nutrigenomics, Genome-wide association study (GWAS), Genome-wide gene-environment interaction study (GWIS), Gene-diet interaction, Genetic epidemiology, Quantitative genetics, Fish oil, Polyunsaturated fatty acids, Vegetarian, Vegetarianism

NUTRIGENETICS OF POLYUNSATURATED FATTY ACIDS AND VEGETARIANISM

by

MICHAEL FRANCIS

B.S., State University of New York College of Environmental Science and Forestry, 2014

M.S., University of Georgia, 2019

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2022

NUTRIGENETICS OF POLYUNSATURATED FATTY ACIDS AND VEGETARIANISM

by

MICHAEL FRANCIS

Major Professor: Kaixiong Ye

Committee: Changwei Li
Rob Pazdro
Shaying Zhao

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2022

DEDICATION

This dissertation is dedicated to my grandmother Dolores, who provided endless support and love, and without whom this would not have been possible.

ACKNOWLEDGEMENTS

I could not have asked for a better Ph. D. supervisor than Dr. Kaixiong Ye. He and my lab mates were a pleasure to work with and I immensely enjoyed being a member of Dr. Ye's team.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v
LIST OF TABLES viii
LIST OF FIGURES ix
CHAPTER
1 INTRODUCTION AND LITERATURE REVIEW 1
Part 1: Nature vs. Nurture vs. Carbon Dioxide 1
Part 2: Dissertation overview 3
Literature Review: New Connections Between Genes and Diet 4
References 7
2 GENOME-WIDE ASSOCIATION STUDY OF FISH OIL SUPPLEMENTATION ON LIPID TRAITS IN 81,246 INDIVIDUALS REVEALS NEW GENE-DIET INTERACTION LOCI 13
Abstract 14
Author Summary 15
Acknowledgments 15
Introduction 16
Results 17
Discussion 22
Methods 26
References 32
Supplementary Figure and Table legends 45
3 FIFTY-ONE NOVEL AND REPLICATED GWAS LOCI FOR POLYUNSATURATED AND MONOUNSATURATED FATTY ACIDS IN 124,024 EUROPEANS 47
Abstract 48
Acknowledgements 49
Competing interests 49
Author contributions 49
Introduction 50
Results 52
Discussion 61
Methods 64
Data availability 73
Code availability 73
References 73
Supplementary Figure and Table legends 86

4 GENE-VEGETARIANISM INTERACTIONS DETECTED IN GENOME-WIDE ANALYSES ACROSS 30 SERUM BIOMARKERS 90

Abstract 91
Acknowledgements 92
Competing interests 92
Author contributions 92
Introduction 93
Results 95
Discussion 103
Methods 111
Data availability 116
References 116
Supplementary Figure and Table legends 129

5 CONCLUSION 133

LIST OF TABLES

Page

Table 2.1: Loci with significant interaction between fish oil supplementation and blood lipid levels. 43
Table 2.2: Loci with significant 2df joint test between fish oil supplementation and blood lipid levels..... 44
Table 4.1: Selecting high quality vegetarians for analysis. 124

LIST OF FIGURES

Page

Figure 2.1: Overview of the analysis performed in this study. 39
Figure 2.2: LocusZoom for genome-wide significant ($P < 5 \times 10^{-8}$) replicated gene-fish oil interaction loci. 40
Figure 2.3: Significant results for the replicated interaction locus with lead SNP rs112803755. 41
Figure 2.4: The gene-fish oil interaction locus with lead SNP rs112803755 overlaps eQTLs of GJB2. 42
Figure 3.1: Overview of analyses. 81
Figure 3.2: Circular Manhattan plot of five meta-analyzed PUFA and MUFA traits. 82
Figure 3.3: Regional Manhattan plots for selected novel loci. 83
Figure 3.4: Results from genetic and phenotypic correlations, heritability, and S-MultiXcan. 84
Figure 3.5: Gene sets mapped to significant meta-analysis loci. 85
Figure 4.1: Identifying vegetarians. 125
Figure 4.2: Forest plot of estimated vegetarianism effects. 126
Figure 4.3: Calcium gene-vegetarianism interaction at rs72952628. 127
Figure 4.4: Significant gene-level gene-vegetarianism interactions. 128

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Part 1: Nature vs. Nurture vs. Carbon Dioxide

Can the age-old conflict of nature versus nurture ever be reconciled? The old guard of Nutrition scientists, who are experts of how we are "nurtured" by diet, are now met with an abundance of evidence demonstrating the Genetic "nature" of our metabolic machinery [1-3]. Nutrigenetic analyses increasingly reveal that the future of nutrition science is in adopting a synthesized perspective of nature and nurture, leading to optimized dietary recommendations at the level of the individual, i.e. Precision (or "Personalized") Nutrition. Nutrigenetics and nutrigenomics have been increasingly active research areas since the mid-2000s [2, 4], and we present three new studies here. Still, public-facing institutions of nutrition and dietetics, who set dietary recommendations, do not recognize the role of genetics. The low-resolution bits of nutrigenetics that the public does encounter, such as "African-Americans are at higher risk of type 2 diabetes than Caucasians," always beg the same unaddressed question: "But why?"

The sluggish transformation in nutrition science of integrating genetics and ancestry into metabolic disease treatment strategies is understandable, as there are still challenges to face regarding the practical implementation of this information [5]. But, it is highly disconcerting that the scientific and political establishment seems to be moving aggressively in the opposite direction, pushing *en masse* recommendations, and ostensibly ignoring the emerging insights from nutrigenomics. This is especially true regarding the (current hot) topic of meat-eating and vegetarian diets. The most illustrative example is the EAT-*Lancet* Commission's "Food in the Anthropocene," also called the "Great Food Transformation" (2019) [6, 7]. This report

recommends the global adoption of what they have termed “universal healthy diets”—in other words, *de-personalized* nutrition. And their recommended diet is “healthy” only if your optimal diet requires less than their recommended limit of 28 g of meat per day (about one tablespoonful) [6, 7]. The World Economic Forum is also advocating for vegetarianism as part of their Great Reset Agenda, for example by promoting research on how to subliminally “nudge” meat-eaters to plant-eating compliance [8]. Similarly, the most recent position of the Academy of Nutrition and Dietetics is to recommend a vegetarian diet indiscriminately as a health benefit for all people [9].

Genetics and personalized nutrition are conspicuously missing from the official dietary guidelines of these international and policy-influencing organizations. Seemingly in the place of DNA, their dietary position statements contain lengthy and prominent sections dedicated to sustainability and carbon footprints. This situation in which nutrigenetics takes lower priority than climate change in informing dietary recommendations is short-sighted and absurd. Directly comparing the CO_2 and methane emissions of an acre of livestock versus an acre of plants ignores the downstream effects of diet. Because if, for example, vegetarianism causes a higher risk of depression, as they have been associated in many studies [10], shouldn’t the carbon footprint of the subsequently prescribed SSRIs be included in these calculations? The same is true for the environmental impact of all synthetic drugs and disease treatments necessitated by universal-diet-induced nutrient deficiencies, which will occur more often in genetically predisposed individuals.

This new round of dietary recommendations is reminiscent of the misguided Food Pyramid [11], whereby “The Science” has again led us to ignore time-tested dietary wisdom, the validity of which is being demonstrated in nutrigenetics. But now, instead of eleven daily servings of bread and pasta, we are being recommended a scoop of climate consciousness with every bite. Will the field of nutrition science forever be subjugated by special interests?

Part 2: Dissertation overview

To perform our quantitative nutrigenetics research, we relied mainly upon two related statistical analysis techniques. Genome-wide association studies (GWAS) can be used to identify genetic variants associated with phenotypes. GWAS are essentially a series of linear regression models applied to millions of genotypes, to find the effects of a genetic variant upon a trait. GWAS study design has been characterized as "hypothesis-free," although this is not technically accurate, since they are implicitly dependent on the design of genotyping arrays and the choice of statistical models involved [12]. However, calling them hypothesis-free is still a useful shorthand nomenclature, because they are indeed unbiased by pathophysiological assumptions [12]. GWAS also have the benefit of low rates of false positives and high rates of rediscovery across studies, owing to the fact that they typically incorporate large sample sizes and built-in replication stages [13]. In Chapter 3, we performed a GWAS for quantitative levels of polyunsaturated and monounsaturated fatty acids (PUFAs and MUFAs) in blood as measured by NMR. In that study, we identified over 50 novel variants significantly associated with PUFA and MUFA traits that were replicated in external studies.

In Chapters 2 and 4, our main analyses were genome-wide interaction studies (GWIS), which are similar to GWAS, but include a statistical interaction term in the models, to find gene-environment interactions (GEI or GxE). Specifically, we used dietary exposures as the environmental term to infer gene-diet interactions (GDI). The interactions between genotype and dietary exposure can account for departures from standard additive regression models. Chapter 2 is a genome-wide gene-fish-oil interaction study, where we examined the interaction effects of fish oil supplementation on levels of blood lipids (LDL-C, HDL-C, triglycerides, and total cholesterol). We found a replicated genetic variant (rs112803755) is associated with significant differences in

triglyceride levels, based on whether a person reports use of fish oil. In Chapter 4, we report the first gene-vegetarianism interactions, in which a vegetarian diet affected levels of serum biomarkers in a genotype-dependent manner. At the variant rs72952628, we found the heterozygote is associated with higher calcium in vegetarians. Additionally, we found gene-level vegetarianism interactions at *RNF168* in testosterone and *ZNF277* in estimated glomerular filtration rate.

Literature Review: New Connections Between Genes and Diet

Nutrigenetic and nutrigenomic studies that analyze the relationship between nutrition and genetic factors have been conducted at a dramatically increased rate through the past decade [2]. The availability of large-scale datasets like the UK Biobank [14], which include both genetic and dietary information, will enable even more of these analyses to occur in coming years. In our lab, we have utilized genome-wide association studies (GWAS) and genome-wide interaction studies (GWIS) to identify genetic polymorphisms that are statistically associated with specific metabolic traits. We then use these variants to guide downstream functional analyses [2]. Because there is overlap in the definitions of nutrigenetics and nutrigenomics, "nutrigenetics" will be defined here as the process of identifying and characterizing genetic variants that are associated with dietary intake.

GWAS and related nutrigenetics analyses identify candidate variants and genes associated with the etiology of disease, especially chronic disease related to metabolism such as obesity, cardiovascular disease (CVD), type 2 diabetes (T2D), and cancer. As a subset of precision medicine, the goal of nutrigenetics is to prevent and treat disease through targeted nutritional therapies, tailored to an individual's unique genetic composition [15]. For example, in the case of

dyslipidemia, a common risk factor for CVD and T2D, the use of fish oil supplements can produce comparable benefits to pharmaceuticals, but with fewer side effects [16, 17]. GWAS can be an efficient first step in the nutrigenetics discovery pipeline, in narrowing down potential causal variants, and providing statistical associations that can inform subsequent clinical trials.

Dietary lipids and polyunsaturated fatty acids (PUFAs) have been the subject of many nutrigenetic studies, both in our lab (e.g. Chapters 2 and 3 of this dissertation), and in general [18]. Polyunsaturated fatty acids (PUFAs) are dietary fats containing two to six double bonds along linear carbon chains from 14 to 22 carbons in length. Imbalances of tissue PUFAs are involved in the pathophysiology of a broad array of diseases, including cardiovascular disease, cancer, depression, and dementia [19-21]. Omega-3 long-chain PUFAs (n-3) have been consistently shown to improve aspects of metabolic syndrome related to the risk factors of cardiovascular disease and obesity, such as insulin resistance, hypertension, and dyslipidemia [19, 22, 23]. n-3 supplementation has been demonstrated to increase the expression of genes that reduce inflammation [24]. In turn, inflammation reduction has been shown to slow or reverse the progression of most diseases, particularly metabolic disease and cancer [25-27].

Levels of the most abundant LCPUFA, omega-6 (n-6), have been associated with both positive and negative health outcomes [28]. Excess n-6 linoleic acid (LA) suppresses tissue and circulating n-3 LCPUFAs, due to common enzymes operating on both PUFA families; balance in dietary n-6 and n-3 is necessary to avoid suppression of the functional n-3 LCPUFAs, eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA) [29]. Modest overall dietary PUFAs and an n-6/n-3 ratio up to 4/1 have been recommended, while the typical modern industrialized diet has a ratio approximating 15/1 [30, 31].

Heritability analyses in twin studies and large cohorts have indicated that substantial genetic components contribute to determining circulating PUFA levels [32-34]. Additionally, n-3 LCPUFA metabolic genes have been shown to undergo local dietary adaptation and exhibit population-specific allele frequency patterns [35, 36]. GWAS can allow us to identify the variants that contribute to this phenotypic variance.

Diet is one of the most important environmental factors in understanding human disease progression. There have been many studies measuring the effects of diet on particular health outcomes, but recently we have begun to identify how diet interacts with the genetics of an individual, and the downstream health implications of these interactions. Genome-wide interaction studies (GWIS) are a valuable analysis technique for identifying gene-environment interactions (GxEs or GEIs). GEIs are defined as a departure of the effect of a genetic polymorphism from an additive association model (commonly used in GWAS), based on the modification of an environmental exposure variable, like diet. For example, a genetic variant can influence the enzymatic metabolism of food or toxin, or modify the risk/benefit of an environmental exposure, or conversely, an environmental exposure can modify gene expression epigenetically [37]. GEIs can help us better understand biological processes leading to disease, idenitify people for whom risk factors are most relevant, and improve the accuracy of epidemiological risk models [37]. In the case of n-3 supplementation, GEIs between n-3s and genetic variants may help explain both missing heritability in lipid biomarker traits [38], and heterogeneity of individual lipid response to fish oil supplementation [39-43].

Dietary environmental exposure can be viewed through the lens of a single nutrient, as is the case with n-3 fish oil supplementation, or can be defined in terms of broad-scale dietary patterns. One such pattern is vegetarianism. Vegetarianism is a superordinate of several animal-

restricted dietary practices; most commonly the term refers to lacto-ovo vegetarianism, which permits all plant-based food plus dairy and eggs, and excludes meat and seafood [44]. Recent estimates indicate interest and adherence to plant-based diets in Western countries have increased substantially over the past decade [45-48]. This shift has occurred for a variety of reasons, including health benefits, taste preferences, ethical concerns with slaughtering animals and factory farming, environmental concerns related to pollution and greenhouse gas emissions, and perceived moral accreditation [48, 49]. For these reasons, it is now common for nutritionists to recommend vegetarianism to the general public [48, 50, 51]. However, the existing literature on the health benefits of vegetarianism has not considered GEIs, and we have demonstrated the first such gene-diet interactions of vegetarianism in Chapter 4. These statiscially significant interactions are the first to implicate that vegetarianism may be better recommended as part of a personalized nutrition program for genetically compatible individuals, in contrast to the current *en masse* recommendations discussed in the previous Introduction section.

It is important to note that results of GWAS and GWIS do not necessarily reveal the precise causal variants of disease. However, they provide a framework for experimental validation by future researchers [52]. The emerging paradigms of precision medicine and precision nutrition suggest that genetic makeup should be utilized to help inform optimal disease treatment strategies [53].

References

1. Drabsch, T. and C. Holzapfel *A Scientific Perspective of Personalised Gene-Based Dietary Recommendations for Weight Management*. Nutrients, 2019. **11**, DOI: 10.3390/nu11030617.

2. Marcum, J.A., *Nutrigenetics/nutrigenomics, personalized nutrition, and precision healthcare.* Current nutrition reports, 2020. **9**(4): p. 338-345.

3. Goodarzi, M.O., *Genetics of obesity: what genetic association studies have taught us about the biology of obesity and its complications.* Lancet Diabetes Endocrinol, 2018. **6**(3): p. 223-236.

4. Mutch, D.M., W. Wahli, and G. Williamson, *Nutrigenomics and nutrigenetics: the emerging faces of nutrition.* The FASEB Journal, 2005. **19**(12): p. 1602-1616.

5. Ordovas, J.M., et al., *Personalised nutrition and health.* Bmj, 2018. **361**: p. bmj.k2173.

6. Willett, W., et al., *Food in the Anthropocene: the EAT–Lancet Commission on healthy diets from sustainable food systems.* The Lancet, 2019. **393**(10170): p. 447-492.

7. Lucas, T. and R. Horton, *The 21st-century great food transformation.* The Lancet, 2019. **393**(10170): p. 386-387.

8. Torkington, S., *Here's a simple way to convince people to eat less meat.* 2022.

9. Melina, V., W. Craig, and S. Levin, *Position of the Academy of Nutrition and Dietetics: Vegetarian Diets.* J Acad Nutr Diet, 2016. **116**(12): p. 1970-1980.

10. Iguacel, I., et al., *Vegetarianism and veganism compared with mental health and cognitive outcomes: a systematic review and meta-analysis.* Nutrition Reviews, 2021. **79**(4): p. 361-381.

11. Soliman, G.A., *Dietary Cholesterol and the Lack of Evidence in Cardiovascular Disease.* Nutrients, 2018. **10**(6).

12. Kitsios, G.D. and E. Zintzaras, *Genome-wide association studies: hypothesis-"free" or "engaged"?* Translational Research, 2009. **154**(4): p. 161-164.

13. Marigorta, U.M., et al., *Replicability and prediction: lessons and challenges from GWAS.* Trends in Genetics, 2018. **34**(7): p. 504-517.

14. Bycroft, C., et al., *The UK Biobank resource with deep phenotyping and genomic data.* Nature, 2018. **562**(7726): p. 203-209.

15. Guasch-Ferré, M., H.S. Dashti, and J. Merino, *Nutritional Genomics and Direct-to-Consumer Genetic Testing: An Overview.* Adv Nutr, 2018. **9**(2): p. 128-135.

16. Thaipitakwong, T. and P. Aramwit, *A Review of the Efficacy, Safety, and Clinical Implications of Naturally Derived Dietary Supplements for Dyslipidemia.* American Journal of Cardiovascular Drugs, 2017. **17**(1): p. 27-35.

17. Scicchitano, P., et al., *Nutraceuticals and dyslipidaemia: Beyond the common therapeutics.* Journal of Functional Foods, 2014. **6**: p. 11-32.

18. Bordoni, L., et al., *Nutrigenomics of Dietary Lipids.* Antioxidants, 2021. **10**(7).

19. Harris, W.S., et al., *Blood n-3 fatty acid levels and total and cause-specific mortality from 17 prospective studies.* Nature Communications, 2021. **12**(1): p. 2329.

20. Lin, P.-Y., et al., *A meta-analytic review of polyunsaturated fatty acid compositions in dementia.* The Journal of clinical psychiatry, 2012. **73**(9): p. 0-0.

21. Grosso, G., et al., *Dietary n-3 PUFA, fish consumption and depression: A systematic review and meta-analysis of observational studies.* Journal of Affective Disorders, 2016. **205**: p. 269-281.

22. Innes, J.K. and P.C. Calder, *Marine Omega-3 (N-3) Fatty Acids for Cardiovascular Health: An Update for 2020.* International journal of molecular sciences, 2020. **21**(4): p. 1362.

23. Lorente-Cebrián, S., et al., *Role of omega-3 fatty acids in obesity, metabolic syndrome, and cardiovascular diseases: a review of the evidence.* Journal of physiology and biochemistry, 2013. **69**(3): p. 633-651.

24. Bouwens, M., et al., *Fish-oil supplementation induces antiinflammatory gene expression profiles in human blood mononuclear cells.* The American Journal of Clinical Nutrition, 2009. **90**(2): p. 415-424.

25. Baker, R.G., M.S. Hayden, and S. Ghosh, *NF-κB, Inflammation, and Metabolic Disease.* Cell Metabolism, 2011. **13**(1): p. 11-22.

26. Libby, P., *Inflammation and cardiovascular disease mechanisms.* The American Journal of Clinical Nutrition, 2006. **83**(2): p. 456S-460S.

27. Coussens, L.M. and Z. Werb, *Inflammation and cancer.* Nature, 2002. **420**(6917): p. 860-867.

28. 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.* Circulation: Cardiovascular Genetics, 2014. **7**(3): p. 321-331.

29. Brenna, J.T. and K.S.D. Kothapalli, *New understandings of the pathway of long-chain polyunsaturated fatty acid biosynthesis.* Curr Opin Clin Nutr Metab Care, 2022. **25**(2): p. 60-66.

30. Kothapalli, K.S.D., H.G. Park, and J.T. Brenna, *Polyunsaturated fatty acid biosynthesis pathway and genetics. implications for interindividual variability in prothrombotic, inflammatory conditions such as COVID-19.* Prostaglandins, Leukotrienes and Essential Fatty Acids, 2020. **162**.

31. Sinclair, A.J., *High Linoleic Acid in the Food Supply Worldwide-What are the Consequences?* Science and Technology of Cereals, Oils and Foods, 2022. **30**(3).

32. Locke, A.E., et al., *Exome sequencing of Finnish isolates enhances rare-variant association power.* Nature, 2019. **572**(7769): p. 323-328.

33. Shin, S.Y., et al., *An atlas of genetic influences on human blood metabolites.* Nat Genet, 2014. **46**(6): p. 543-550.

34. Kettunen, J., et al., *Genome-wide association study identifies multiple loci influencing human serum metabolite levels.* Nat Genet, 2012. **44**(3): p. 269-76.

35. Ye, K., et al., *Dietary adaptation of FADS genes in Europe varied across time and geography.* Nature Ecology & Evolution, 2017. **1**(7): p. 0167.

36. Kothapalli, K.S.D., et al., *Positive Selection on a Regulatory Insertion-Deletion Polymorphism in FADS2 Influences Apparent Endogenous Synthesis of Arachidonic Acid.* Molecular biology and evolution, 2016. **33**(7): p. 1726-1739.

37. Dudbridge, F. and O. Fletcher, *Gene-environment dependence creates spurious gene-environment interaction.* The American Journal of Human Genetics, 2014. **95**(3): p. 301-307.

38. Klarin, D., et al., *Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program.* Nature genetics, 2018. **50**(11): p. 1514-1523.

39. Madden, J., et al., *The Impact of Common Gene Variants on the Response of Biomarkers of Cardiovascular Disease (CVD) Risk to Increased Fish Oil Fatty Acids Intakes.* Annual Review of Nutrition, 2011. **31**(1): p. 203-234.

40. Aung, T., et al., *Associations of Omega-3 Fatty Acid Supplement Use With Cardiovascular Disease Risks: Meta-analysis of 10 Trials Involving 77 917 IndividualsMeta-analysis of Associations of Omega-3 Fatty Acids and Cardiovascular RiskMeta-analysis of Associations of Omega-3 Fatty Acids and Cardiovascular Risk.* JAMA Cardiology, 2018. **3**(3): p. 225-233.

41. Zheng, J., et al., *Fish consumption and CHD mortality: an updated meta-analysis of seventeen cohort studies.* Public Health Nutr, 2012. **15**(4): p. 725-37.

42. Martinelli, N., et al., *FADS genotypes and desaturase activity estimated by the ratio of arachidonic acid to linoleic acid are associated with inflammation and coronary artery disease.* Am J Clin Nutr, 2008. **88**(4): p. 941-9.

43. Bokor, S., et al., *Single nucleotide polymorphisms in the FADS gene cluster are associated with delta-5 and delta-6 desaturase activities estimated by serum fatty acid ratios.* J Lipid Res, 2010. **51**(8): p. 2325-33.

44. Oussalah, A., et al., *Health outcomes associated with vegetarian diets: An umbrella review of systematic reviews and meta-analyses.* Clinical Nutrition, 2020. **39**(11): p. 3283-3307.

45. Hess, J.M., *Modeling Dairy-Free Vegetarian and Vegan USDA Food Patterns for Nonpregnant, Nonlactating Adults.* The Journal of Nutrition, 2022: p. nxac100.

46. Janssen, M., et al., *Motives of consumers following a vegan diet and their attitudes towards animal agriculture.* Appetite, 2016. **105**: p. 643-651.

47. Wickramasinghe, K., et al., *The shift to plant-based diets: are we missing the point?* Global Food Security, 2021. **29**: p. 100530.

48. Leitzmann, C., *Vegetarian nutrition: past, present, future.* The American Journal of Clinical Nutrition, 2014. **100**(suppl_1): p. 496S-502S.

49. Piazza, J., et al., *Rationalizing meat consumption. The 4Ns.* Appetite, 2015. **91**: p. 114-128.

50. Melina, V., W. Craig, and S. Levin, *Position of the Academy of Nutrition and Dietetics: vegetarian diets.* Journal of the Academy of Nutrition and Dietetics, 2016. **116**(12): p. 1970-1980.

51. Radnitz, C., B. Beezhold, and J. DiMatteo, *Investigation of lifestyle choices of individuals following a vegan diet for health and ethical reasons.* Appetite, 2015. **90**: p. 31-36.

52. Schaid, D.J., W. Chen, and N.B. Larson, *From genome-wide associations to candidate causal variants by statistical fine-mapping.* Nature reviews. Genetics, 2018. **19**(8): p. 491-504.

53. De Toro-Martín, J., et al., *Precision Nutrition: A Review of Personalized Nutritional Approaches for the Prevention and Management of Metabolic Syndrome.* Nutrients, 2017. **9**(8): p. 913.

CHAPTER 2

GENOME-WIDE ASSOCIATION STUDY OF FISH OIL SUPPLEMENTATION ON LIPID TRAITS IN 81,246 INDIVIDUALS REVEALS NEW GENE-DIET INTERACTION LOCI[1]

[1] Francis M, Li C, Sun Y, Zhou J, Brenna JT, Li X, Ye K. (2021) Genome-wide association study of fish oil supplementation with continuous lipid traits in 81,246 individuals reveals new interaction loci. PLOS Genetics 17(3): e1009431. doi:10.1371/journal.pgen.1009431.

Abstract

Fish oil supplementation is widely used for reducing serum triglycerides (TAGs) but has mixed effects on other circulating cardiovascular biomarkers. Many genetic polymorphisms have been associated with blood lipids, including high- and low-density-lipoprotein cholesterol (HDL-C, LDL-C), total cholesterol, and TAGs. Here, the gene-diet interaction effects of fish oil supplementation on these lipids were analyzed in a discovery cohort of up to 73,962 UK Biobank participants, using a 1-degree-of-freedom (1df) test for interaction effects and a 2-degrees-of-freedom (2df) test to jointly analyze interaction and main effects. Associations with $P < 1\times10^{-6}$ in either test (26,157; 18,300 unique variants) were advanced to replication in up to 7,284 participants from the Atherosclerosis Risk in Communities (ARIC) Study. Replicated associations reaching 1df $P < 0.05$ (2,175; 1,763 unique variants) were used in meta-analyses. We found 13 replicated and 159 non-replicated (UK Biobank only) loci with significant 2df joint tests that were predominantly driven by main effects and have been previously reported. Four novel interaction loci were identified with 1df $P < 5\times10^{-8}$ in meta-analysis. The lead variant in the *GJB6-GJB2-GJA3* gene cluster, rs112803755 (A>G; minor allele frequency = 0.041), shows exclusively interaction effects. The minor allele is significantly associated with decreased TAGs in individuals with fish oil supplementation, but with increased TAGs in those without supplementation. This locus is significantly associated with higher *GJB2* expression of connexin 26 in adipose tissue; connexin activity is known to change upon exposure to omega-3 fatty acids. Significant interaction effects were also found in three other loci in the genes *SLC12A3* (HDL-C), *ABCA6* (LDL-C), and *MLXIPL* (LDL-C), but highly significant main effects are also present. Our study identifies novel gene-diet interaction effects for four genetic loci, whose effects on blood lipids are modified

by fish oil supplementation. These findings highlight the need and possibility for personalized nutrition.

Author Summary

We are utilizing the unprecedentedly large genotype and phenotype dataset in the UK Biobank to perform a genome-wide association study (GWAS) which accounts for the interplay between genotype and dietary intake. We examined the interaction effects of fish oil supplementation on levels of blood lipids (LDL-C, HDL-C, triglycerides, and total cholesterol). Our findings were replicated in the Atherosclerosis Risk in Communities (ARIC) Study. We found that the genetic variant rs112803755 is associated with significant differences in triglyceride levels, based on whether a person reports use of fish oil. We further analyzed rs112803755 with functional genomics data from the Genotype-Tissue Expression (GTEx) project to identify potential target genes, and found a connexin coding gene which has been previously reported to respond to cellular omega-3 levels. This research suggests that inter-personal variation in triglyceride response to fish oil supplementation is in part explained by genotype, and that fish oil dose adjustment based on genotype should be investigated as a means to protect against cardiovascular disease risk.

Acknowledgments

The authors would like to thank all UK Biobank participants and administrators for data access. We also thank all Ye lab members for helpful discussions.

Introduction

Dyslipidemia, characterized by imbalances in low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), and triglycerides (TAGs), is a common predictive factor for metabolic conditions such as cardiovascular disease and type 2 diabetes [1,2]. Use of dietary supplements in lieu of xenobiotic pharmaceuticals for the management of dyslipidemia may produce comparable benefits with fewer side effects [1,2]. In particular, the omega-3 long-chain polyunsaturated fatty acids (n-3 LCPUFAs) eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA) supplied by fish oil supplements are an effective treatment for hypertriglyceridemia, though results are mixed for LDL-C and HDL-C [3–5]. Genetic polymorphisms have been consistently associated with intra- and inter-population differences in levels of LDL-C, HDL-C, total cholesterol, and TAGs [6,7]. Gene-environment interactions (GEIs; specifically, gene-diet interactions) between n-3 LCPUFAs and genetic variants have been reported, though few have been replicated, likely due to small sample sizes and inconsistencies in study designs such as study length and supplement dosage [8]. Studies of GEIs may reveal novel genetic loci that are otherwise obscure in conventional main-effect-only association studies, and may identify genetic loci whose phenotypic effects are modifiable by specific environmental exposures. Further identification of these GEIs may help explain both missing heritability in lipid biomarker traits [9], and heterogeneity of individual lipid response to fish oil supplementation [8,10–13].

To identify genomic factors which interact with n-3 LCPUFAs supplementation to affect levels of blood lipids, we performed a genome-wide association study (GWAS) among participants of the large UK Biobank cohort [14]. We used only participants whose genetic ethnic grouping is Caucasian, the largest sample available, to avoid population stratification [15]. The focus on

single-ancestry groups is particularly important in studies related to LCPUFAs because their metabolic genes have been shown to undergo genetic adaptation to local diets in multiple geographical regions, and exhibit population-specific allele frequency patterns [16,17]. We used both the traditional 1-degree-of-freedom (1df) interaction test and a 2-degrees-of-freedom (2df) joint test to evaluate interactions between genetic variants and fish oil supplementation on blood lipid phenotypes. The 2df joint test evaluates single nucleotide polymorphism (SNP) main effects and interaction effects jointly, and therefore has higher power to detect SNPs with moderate main effects and moderate interaction effects that would otherwise be missed in the 1df test [18–20]. This method has recently been employed to examine the GEIs of these lipid traits with smoking [21], and sleep duration [22]. We further confirmed promising UK Biobank findings in a US cohort, the Atherosclerosis Risk in Communities Study (ARIC). Replicated SNPs were utilized in a meta-analysis of these studies to reveal new gene-diet interaction loci.

Results

Cohort demographics

Stage 1 discovery analyses were performed in up to 73,962 genetically Caucasian UK Biobank participants (S1 Table). Approximately 15.8% of these participants answered yes to taking fish oil supplements on dietary questionnaires at two time points taken between one and five years apart. The percentage male, mean age and BMI of UK Biobank participants were ~46.6%, 55.6 ± 7.9 (±1 SD) years old, and 27.0 ± 4.6 kg/m^2, respectively. Stage 2 (replication) analyses were performed in up to 7,284 white participants in the ARIC cohort study. Approximately 1.4% of these participants answered yes to taking fish oil at one time point during

their primary assessment. The percentage male, mean age and BMI of ARIC participants were ~47.0%, 54.3 ± 5.7 years old, and 26.9 ± 4.7 kg/m^2, respectively.

Gene-diet interaction GWAS

A three-stage discovery, replication, and meta-analysis approach for identification of significant GWAS loci was adopted for the blood lipid phenotypes LDL-C, HDL-C, total cholesterol, and TAGs (Fig 1). Genomic control (GC) correction was applied during Stage 1 2df P-value calculation and Stage 3 2df P-value calculation; lambda (λ) values after GC correction were 1. GC values of 1df P-values for Stage 1 and Stage 3 were < 1, therefore GC correction was not necessary.

Variants with 1df or 2df $P < 1\times10^{-6}$ in a gene-fish-oil interaction GWAS model (Eq (1)) were selected for replication (S2 Table). For the four lipid traits, LDL-C, HDL-C, total cholesterol, and TAGs, 26,157 associations (18,300 unique variants) met this criterion (S1 and S2 Figs).

Stage 2 replication analyses were performed in up to 7,284 white participants in the ARIC cohort (S1 Table). A gene-fish-oil interaction GWAS model was performed. Variants passed from Stage 1 with 1df $P < 0.05$ were considered as replicated. Of the 26,157 associations from Stage 1, a total of 2,175 associations (1763 unique variants) for the lipid traits were replicated (S3 Table) and passed to the meta-analysis step. There were also 17,259 associations (12,440 unique variants, 85 unique loci; S4 and S5 Tables) which reached genome-wide significance in Stage 1 ($P < 5 \times 10^{-8}$), but were not replicated in Stage 2, and therefore not sent to meta-analysis. All of the 85 lead variants had significant 2df joint test *P*-values, and none of their 1df interaction *P*-values approached significance, suggesting these variants influence lipid traits predominantly through main effects.

Meta-analysis of Stage 1 and Stage 2 results for both 1df and 2df tests were performed for each blood lipid phenotype. Significant variants were defined in meta-analysis as those meeting the genome-wide significance threshold in their 1df or 2df tests ($P < 5 \times 10^{-8}$). This revealed 16 novel and significant 1df associations (4 unique loci; Table 1) and 53 significant 2df associations (11 unique loci; Table 2 and S6 Table). One variant, rs112803755 (*GJB6*; A>G; minor allele frequency (MAF) = 0.0410) had a significant 1df interaction term and no significant 2df or main effects terms. In the discovery cohort, the minor allele of SNP rs112803755 is associated with a strong decrease in TAGs among those taking fish oil supplements ($\beta_{G(E=1)}$ = -0.12 mmol/L, $P = 5.59\times10^{-5}$), but is suggestively associated with a mild increase in TAGs for those without supplementation ($\beta_{G(E=0)}$ = 0.030 mmol/L, P = 0.024), resulting in a significant interaction effect (1df $P = 1.95\times10^{-7}$). There is no association between the SNP main effects and TAGs in the UK Biobank (β_G = 0.0063 mmol/L, P = 0.60) if not considering the interaction effect. Meta-analysis revealed that the interaction effect at this SNP reaches genome-wide significance (1df $P = 5.65 \times 10^{-10}$). Three additional variants have both significant 1df interaction and 2df joint test *P*-values in the meta-analysis: rs799157 (*MLXIPL*; C>T; MAF = 0.0407) with LDL-C, rs77542162 (*ABCA6*; A>G; MAF = 0.0218) with LDL-C, and rs148931404 (*SLC12A3*; G>A; MAF = 0.0221) with HDL-C (Table 1 and Fig 2). In the discovery cohort, the minor allele of SNP rs799157 is associated with an increase in LDL-C (β_G = 0.057 mmol/L, $P = 3.33\times10^{-8}$) after adjusting for fish oil supplementation status and other covariates. Less significant associations were observed in the stratified groups with fish oil supplementation ($\beta_{G(E=1)}$ = 0.087 mmol/L, $P = 8.14\times10^{-4}$) and in those without ($\beta_{G(E=0)}$ = 0.052 mmol/L, $P = 5.36\times10^{-6}$). Meta-analysis confirmed the presence of main effect and revealed an interaction effect (1df $P = 1.92\times10^{-11}$, 2df $P = 1.93\times10^{-33}$). Similarly, SNP rs77542162 is associated with an increase in LDL-C in the overall discovery cohort (β_G =

1.41 mmol/L, $P = 5.40\times10^{-23}$), in those without ($\beta_{G(E=0)} = 1.50$, $P = 4.24\times10^{-21}$) and with ($\beta_{G(E=1)} = 1.11$, $p = 1.55\times10^{-3}$) fish oil supplementation. Meta-analysis revealed genome-wide significance in both tests (1df $P = 4.48\times10^{-9}$, 2df $P = 6.58\times10^{-63}$). For HDL-C, there is only one SNP, rs148931404, that reaches genome-wide significant 1df P-value (1.82×10^{-16}) in the meta-analysis. It is associated with an increase in HDL-C in the overall discovery cohort ($\beta_G = 0.049$ mmol/L, $P = 2.70\times10^{-16}$), in those without ($\beta_{G(E=0)} = 0.045$ mmol/L, $P = 5.67\times10^{-12}$) and with ($\beta_{G(E=1)} = 0.071$ mmol/L, $P = 3.49\times10^{-6}$) fish oil supplementation. The three variants with both significant 1df and 2df *P*-values are mainly driven by main effects, as reflected by the much more significant 2df *P*-values and the consistent associations across subgroups in UK Biobank. All four loci have been previously found to be associated with the corresponding lipid. Overall, we unraveled novel gene-fish oil interaction effects for four previously known lipid-associated genetic loci.

There are 11 unique genetic loci whose 2df joint test *P*-values reached the genome-wide significance cutoff ($P < 5 \times 10^{-8}$) but their 1df interaction test *P*-values did not. For instance, a SNP upstream of *LPL* rs117860853 is associated with a decrease in HDL-C in the overall discovery cohort ($\beta_G = -0.078$ mmol/L, $P = 5.72\times10^{-24}$), in those without ($\beta_{G(E=0)} = -0.081$ mmol/L, $P = 3.47\times10^{-22}$) and with fish oil supplementation ($\beta_{G(E=1)} = -0.063$ mmol/L, $P = 3.00\times10^{-3}$). Meta-analysis revealed that this SNP has a significant main effect but no interaction effect (1df $P = 0.015$, 2df $P = 5.46\times10^{-28}$). Notably, two loci have 1df interaction test *P*-values that are close to the genome-wide significance level. SNP rs141844019, downstream of *HAPLN4*, has a suggestive interaction effect on TAGs ($\beta_{G\times E} = 1.64$ mmol/L, $P = 1.64\times10^{-6}$), while SNP rs77542162, a missense variant of *ABCA6*, may have an interaction effect on total cholesterol ($\beta_{G\times E} = -1.59$ mmol/L, $P = 4.58\times10^{-7}$). All these significant 2df replicated loci (Tables 1 and 2)

were within 1 Mb of one or more previously reported loci associated with the same blood lipid phenotype and are therefore not reported as novel.

rs112803755 modifies the effect of fish oil on TAGs

Using TAG levels as a phenotype, the locus of 11 significant variants whose lead SNP is rs112803755 (*GJB6*: 5650 bp downstream; A>G; MAF = 0.0410) has a significant 1df interaction *P*-value (5.65×10^{-10}), while its 2df joint P-value is not significant ($P = 0.0124$) (Fig 3A). Its fish-oil adjusted main effects model SNP term is not significant ($P = 0.600$), and in a stratified analysis the *P*-value is lower in the fish-oil supplementation exposure group ($P = 5.59 \times 10^{-5}$) than the non-supplementing group ($P = 2.42 \times 10^{-2}$) (Fig 3A). This evidence suggests that this locus is involved predominantly with interaction effects but not main effects.

The rs112803755 locus has significant TAG-lowering effect in those who supplement fish oil versus those who do not when considering AA vs. AG genotypes (Fig 3B and S7 Table). Since this variant has low MAF (~4.1%), homozygous individuals of GG genotype are rare. TAG levels were significantly higher in AG heterozygotes who did not take fish oil ($\Delta\bar{x}_{TAGs} = +0.04\ \text{mmol/L}$) versus those who did ($\Delta\bar{x}_{TAGs} = -0.111\ \text{mmol/L}$). However, with respect to rs112803755, while fish oil supplementation is associated with lower TAGs in heterozygous individuals, it has a slight opposite effect in AA homozygotes ($\Delta\bar{x}_{TAGs} = +0.0197\ \text{mmol/L}$; $P = 0.0258$).

rs112803755 eQTL mapping

To evaluate if regulation of gene expression is an underlying molecular mechanism for the interaction locus whose lead SNP is rs112803755 (Table 1), we interrogated the association of these genetic markers with expression levels of nearby genes using data from the Genotype-Tissue

Expression (GTEx) project. For the 11 genetic markers in this locus with genome-wide significance of interaction with fish oil, all of them are exclusively associated with the expression of *GJB2*. Expression quantitative trait loci (eQTLs) for *GJB2* were found in multiple tissues but the strongest signals were observed in subcutaneous adipose, which overlap with the significant interaction signals (Fig 4A). rs112803755 is associated with *GJB2* expression in subcutaneous adipose ($P = 7.7 \times 10^{-14}$; Fig 4B), while another interaction SNP in this locus, rs7987144 (G>A; MAF = 0.0375), has an even stronger association ($P = 2.6 \times 10^{-25}$; Fig 4C). Both of these SNPs show increased *GJB2* expression with increased minor allele dosage. These eQTLs results indicate that regulatory variants of *GJB2* are likely responsible for the interaction signals at this locus.

Discussion

In this gene-diet interaction GWAS, we identified and replicated novel interaction loci, in which fish oil supplementation affected levels of continuous lipid traits in a large Caucasian cohort. We found one locus, rs112803755, with a significant interaction effect but a non-significant main effect, suggesting that the presence of minor alleles at this locus can enhance the TAG-lowering effects of fish oil supplementation. We found three additional new significant interaction loci related to LDL-C and HDL-C levels, though these appear predominantly influenced by main effects (Table 1).

rs112803755 is found 5.65 kb downstream from *GJB6*, or alternatively 23.3 kb upstream from GJB2. It is also in high LD with variants found in the other genes in the *GJB6-GJB2-GJA3* gene cluster at 13q12.11 (Fig 2A). *GJB6*, *GJB2*, and *GJA3* are connexin (Cx) gap junction protein-coding genes that encode Cx30, Cx26, and Cx46, respectively. Cxs are responsible for forming hemichannels across gap junctions to enable the exchange of messenger molecules between

adjacent cells. An n-6 LCPUFA, linoleic acid, has been shown to increase hemichannel activity of Cx26 in HeLa cells [23], and n-3 LCPUFAs lowered the expression of another connexin, Cx43, in rats with hypertriglyceridemia [24]. Genetic polymorphisms in another Cx gene are associated with protective effects on cardiovascular disease [25]. It is therefore plausible that changes in n-3 LCPUFA status induced by fish oil supplementation could interact with one Cx in this cluster to affect TAG levels. Although our analysis supports the likely presence of a regulatory variant, we also cannot rule out the existence of a causal coding variant.

rs799157 is a synonymous variant in exon 6 of *MLXIPL*, whose gene product is known as Carbohydrate-responsive element-binding protein (ChREBP). We found this variant has a significant interaction effect of fish oil on LDL-C. Variants in *MLXIPL* have previously been associated with changes in LDL-C and TAGs [26,27]. Intracellular levels of PUFAs are known to suppress ChREBP transactivity, though the molecular basis for this is not defined [28,29]. GTEx reveals that SNPs in this locus are significantly associated with increased expression of TYW1B. Specifically, increased minor allele dosage at rs799157, which we found to be associated with lower LDL-C levels, is most significantly associated with higher TYW1B expression in subcutaneous adipose tissue. TYW1B is a tRNA-yW synthesizing protein coding pseudogene involved in wybutosine synthesis, whose characteristics are not well-studied. This evidence suggests biological support for the ChREBP coding variant, while the regulatory variant for TYW1B is unlikely to be the causal variant.

rs148931404 is an intron variant of *SLC12A3* which we found to be associated with lower HDL-C levels. This gene has previously been associated with HDL-C in a large multi-ethnic GWAS [6]. SLC12A3 encodes the sodium-chloride symporter protein. We did not find any plausible underlying biological mechanism for this variant. rs77542162 is a missense variant in

ABCA6 that we found to be associated with LDL-C. This variant has been reported in several GWAS studies in relation to LDL-C levels [6,7] and also in a 2df test joint GEI test with alcohol consumption [30]. Therefore, it is likely that this SNP is driven by main effects as *ABCA6* is thought to be regulated with macrophage lipid homeostasis [31].

Fish oil supplementation for treatment of hypertriglyceridemia has long been recognized [32]. Recent studies suggest that EPA and DHA have differential effects on HDL-C subfractions, but their overall effects on cardiometabolic lipid risk markers remain unresolved [33], despite dozens of human trials. Nearly all studies to date ignored genetic variants and focused on random cross sections of the population. Our unbiased study identified a variant modulating TAG levels, the only one of the lipid biomarker traits examined that is known to be clearly related to fish oil intake. Further, we identified variants modulating HDL-C and LDL-C, though these effects require further study. Overall, our study found no strong variants that may modulate LDL-C or HDL-C differentially between individuals based on fish oil supplementation status, thus supporting the hypothesis that EPA and DHA effects on these biomarkers are well represented by clinical trials that do not consider interaction with genotype. Our findings emphasize that a one-size-fit-all recommendation of fish oil supplementation to reduce TAG may not be appropriate. While individuals who are heterozygous (AG) at SNP rs112803755 experience a reduction in blood TAG when taking fish oil supplements, homozygotes of AA actually experience an increase. Based on the strong relationship between TAG and cardiovascular diseases, it is natural to hypothesize that the same genetic locus at *GJB2* might interact with n-3 LCPUFAs intakes to have differing effects on the risk of cardiovascular diseases. This is a promising hypothesis calling for direct tests in future studies.

Our study has several strengths and weaknesses. One strength granted by the UK Biobank is a large sample size with two data points taken several years apart for fish oil supplementation. This makes our discovery dataset quite robust and reduces the measurement error of our environmental exposure, which is an important consideration for GEI studies [34]. The ARIC data is less reliable, with only one fish oil data point. A weakness that we recognize is that other dietary quantities of n-3 and n-6 PUFAs are difficult to ascertain, and may interfere with the effects of fish oil. Another limitation of this study is that the ratio of samples in the discovery and replication cohorts is about 10:1. Currently, datasets which provide participant genotype data, fish oil supplementation use, and blood lipid measurements, are rare. Despite the difference in sample size between the UK Biobank and ARIC datasets, each is sufficiently powered to identify significant variants, with the exception of those which are rare or have low effect sizes. Previous gene-diet interaction studies of fish oil have had participants in the hundreds [8], and this is the largest fish oil interaction GWAS to date. One additional weakness is that there may be heterogeneity in the dosage of n-3 LCPUFAs provided by fish oil supplements. These limitations of exact nutrient quantification are present in most nutritional studies which rely on food frequency questionnaires and/or 24-hour recall surveys. Lastly, as in any other association study, ours is associative in nature and could not pinpoint the causal environmental exposure or the genetic variant [35]. We only examine one environmental exposure in this study, fish oil supplementation, which is correlated with many other lifestyle factors [36]. It is possible that other unexamined but correlated environmental factors drive the observed interaction effects, highlighting the need to perform interaction analysis with more environmental factors. However, our novel results make biological sense and many can be placed in a plausible mechanistic context. Finding significant interactions

associated with the genes GJB2 and MLXIPL, which have been shown to be regulated by PUFAs, is a validation of our approach.

Our study unravels novel gene-diet interaction effects for four genetic loci, whose effects on blood lipids are modified by fish oil supplementation. Such results lend further support to the practice of precision nutrition to catalyze nutrition science into meaningful and clinically relevant dietary suggestions [37]. Personalizing and optimizing fish oil supplementation recommendations based on a person's unique genetic composition can improve our understanding of nutrition, and lead to significant improvements in human health and well-being. Once validated, these variants in *GJB2*, *SLC12A3*, *ABCA6*, and *MLXIPL*, will contribute to our understanding of how accounting for genetic differences can allow every person to implement their optimal nutrient intake. Accounting for interaction effects can also help us better understand biological processes leading to disease, and improve the accuracy of future risk prediction models.

Methods

Ethics statement

Use of participant data was approved by the University of Georgia Institutional Review Board, UK Biobank (Project ID 48818), and the National Center for Biotechnology Information. Participants of UK Biobank and the Atherosclerosis Risk in Communities Study (ARIC) have signed written consent forms authorizing the use of their medical and genetic data for use in research studies. All methods were performed securely and in accordance with ethical guidelines and regulations.

Participants

UK Biobank is a prospective cohort study which recruited > 500,000 volunteer participants between 2006 and 2010 in England, Scotland and Wales. Biochemical, clinical, and genotype data were collected. ARIC is a prospective cohort study conducted in four U.S. communities, which began in 1987 and continued to 2007. ARIC participants were randomly selected from pre-defined populations to have medical, social, and demographic data collected. All participants were 40 to 70 years of age at the time of assessment. Participant characteristics can be found in S1 Table.

Participants were quality controlled on the following criteria: genetic ethnicity is Caucasian, used in PCA analysis, not an outlier for heterogeneity and missing genotype rate, no sex chromosome aneuploidy, does not have high degree of genetic kinship (ten or more third-degree relatives identified), and self-reported sex matches genetic sex. Additionally, we removed the minimum number of participants to eliminate all related pairs.

Phenotypes

All continuous blood lipid measures are reported and analyzed in mmol/L. For stage 1 participants, lipid measures were collected during the UK Biobank Assessment Centres initial assessment from 2006–2010. Blood lipids were analyzed by direct aliquot assays in UK Biobank participants using a Beckman Coilter AU5800. LDL-C was measured by enzymatic protective selection analysis; HDL-C was measured by enzyme immunoinhibition analysis; total cholesterol was measured by CHO-POD analysis; TAGs were measured by GPO-POD analysis.

For ARIC participants, plasma was ultracentrifuged to obtain VLDL-free infranate. LDL-C was precipitated by addition of dextran sulfate and Mg2+ to separate an HDL-C supernate. HDL-C was re-precipitated with dextran sulfate and Mg2+, and separated by centrifugation. LDL-C levels were calculated using the Friedewald equation. TAGs and total cholesterol were processed

and their levels measured by spectrophotometry as described in the ARIC manual for Lipid and Lipoprotein Determinations [38].

LDL-C was adjusted for those who self-reported the use of statins or lipid-lowering drugs as described in [19]; this adjustment was performed in 9,951 UK Biobank participants and 316 ARIC participants. No adjustments were made for other lipids.

Fish oil supplementation status

Blood LCPUFA levels were not taken in UK Biobank or ARIC cohort studies. Because omega-3 content in dietary intake can vary significantly depending on animal feed quality (e.g. egg laying hens fed an omega-3 rich diet), as well as source (e.g. wild or farmed raised salmon) [39–42], and since neither dietary questionnaire specifies these details, we use fish oil consumption as a minimally confounded contributor to EPA and DHA consumption [43].

Dietary intake data for UK Biobank participants was taken at two time points approximately 3–4 years apart. Participants were asked of their supplement use, including fish oil, in their health and medical history questionnaire at the initial assessment, "Do you regularly take any of the following? (You can select more than one answer)" (f.6179). An online follow-up assessment which included the Oxford WebQ, a digital 24-hour dietary recall questionnaire, was completed by UK Biobank participants on a voluntary basis between 2011–2012 [44,45]. Participants self-reported their use of dietary supplements from the preceding 24 hours (f.20084). Those who answered yes to fish oil supplementation at both time points were coded as 1, those who answered no at both points were coded as 0, and those with different answers were excluded from our analysis (S3 Fig).

ARIC participants indicated their fish oil supplementation status at one time point during their primary assessment. Participants were asked “Do you regularly take fish oil? (Including

omega-3 fatty acids, EPA, cod liver oil).” in the “Vitamin Survey Form” at the date of their primary assessment between 1985–2007.

Covariates

Covariates used in our association analyses were age, sex, body mass index (BMI), weekly servings of oily fish, socioeconomic status measured by Townsend deprivation index, and the first ten genetic principal components. BMI is measured in kg/m2, and was transformed using ordered quantile normalization for ARIC participants. Weekly servings of oily fish were converted to ordinal variables ranging from 0 (none) to 5 (more than one serving per day). Genetic principal components were provided in the original genotype data of both cohorts.

Genotype data

The first 50,000 UK Biobank participants of the full study cohort were genotyped using the Affymetrix UK BiLEVE Axiom array, and the remaining 450,000 participants were genotyped using the Affymetrix UK Biobank Axiom array; the two arrays are more than 95% similar in their variant content. Imputation and initial quality control of UK Biobank SNPs were performed by a collaborative group headed by the Wellcome Trust Centre for Human Genetics. We excluded autosomal SNPs with imputation quality score < 0.5, minor allele frequency (MAF) < 1%, missing genotype per individual > 5%, missing genotype per variant > 2%, or Hardy-Weinberg equilibrium (HWE) $P < 1\times10-6$. After quality control, a total of 7,954,107 autosomal variants among 73,962 participants were included in the analyses. Our quality control and genotype file format conversions were performed using PLINK2 alpha-v2.3 [46–48].

ARIC participants were genotyped using the Affymetrix GeneChip SNP Array 6.0. Before imputation, quality control removed variants with missing rate > 10%, or MAF < 1%, and individuals with missing genotype rate > 80%. After quality control, genotypes were imputed to

the ALL ancestry panel of the 1000 Genome Phase III integrate Release Version 5 [49] using MiniMac software [50]. After imputation, SNPs with r2 < 0.50, MAF < 1%, or HWE P < 1×10−6 were removed.

Stage 1 analysis

Stage 1 analysis included up to 73,962 UK Biobank participants and up to 7,954,107 variants after quality control (S1 Table).

Interaction regression was performed for each variant using QuickTest (v1.2) according to the following fixed effects GWAS interaction model:

$$Y = \beta_0 + \beta_G G + \beta_E E + \Sigma\, \beta_{Ck} C_k + \beta_{G\times E} G\times E + \varepsilon \quad (1)$$

where Y is a measure of lipid traits (LDL-C, HDL-C, total cholesterol, and TAGs), G is the effect variant count (0/1/2), E is a binary variable representing fish oil supplementation status (0/1), Ck are covariates, and G×E is the GEI term (S4 Fig). Regression coefficients and P-values were calculated using QuickTest normal mean method for expected genotype dosages; this method is implemented to reduce false positives [51]. Robust Huber sandwich estimates of the variance-covariance matrix were generated.

Main effects adjusted by E were calculated according to the fixed effects model:

$$Y = \beta_0 + \beta_G G + \beta_E E + \Sigma\, \beta_{Ck} C_k + \varepsilon \quad (2)$$

Main additive variant effects, and variant effects stratified by (E) were also calculated using the generalized fixed effects model:

$$Y = \beta_0 + \beta_G G + \Sigma\, \beta_{Ck} C_k + \varepsilon \quad (3)$$

These main effects models were performed using the same QuickTest normal mean method.

Joint *P*-values of main and interaction effects (β_G and $\beta_{G\times E}$) were calculated according to a 2df χ^2 distribution which corrected for the determinant of the covariance matrix between these two

terms [18]. Genomic control was applied to Stage 1 2df joint *P*-values for each lipid phenotype. Variants reaching $P < 1 \times 10^{-6}$ in either the 1df interaction test or 2df joint test were advanced to replication in Stage 2.

Stage 2 analysis

Stage 2 analysis included up to 7,284 ARIC participants, and 48,608,505 variants (S1 Table). Participants were filtered on the basis of their ethnicity (white) only. Additional quality control on samples and genomic data (as in Stage 1) was not conducted, because these filters are meant to reduce the rate of false positives, which was not relevant for Stage 2 replication. Regression coefficients and P-values were calculated using QuickTest normal mean method. Variants advanced from Stage 1 which also had a $P < 0.05$ in the 1df interaction term in the ARIC cohort were advanced to joint meta-analysis between the two cohorts in Stage 1+2.

Meta-analysis of stage 1+2

METAL meta-analysis software (2010-02-08) [52] was used to perform a meta-analysis of those associations with $P < 1 \times 10^{-6}$ in 1df interaction and/or 2df joint tests in Stage 1, and $P < 0.05$ in 1df interaction test in Stage 2 (patch provided by A. Manning to enable 2df GEI testing [18]; genome.sph.umich.edu/wiki/Meta_Analysis_of_SNPxEnvironment_Interaction). Stage 1+2 meta-analyses were performed using a weighted z-statistic by sample size [52]. Genomic control was applied to all meta-analyses as implemented by METAL. Associations exceeding the genome-wide significance threshold of $P < 5 \times 10^{-8}$ were passed to FUMA to identify the lead SNP for each locus.

Identifying lead SNPs

Variants exceeding the genome-wide significance threshold of $P < 5 \times 10^{-8}$ were inputted to FUMA to identify independent loci and their lead SNPs [53]. Lead SNPs are defined as the SNP

within a locus having the lowest *P*-value. UK Biobank release 2b 10k White British was used as the reference panel population. The maximum *P*-value cutoff was set to 0.05, and a first threshold of $r^2 \geq 0.6$ and second threshold of $r^2 \geq 0.1$ were used to define independent significant SNPs. The maximum distance between LD blocks to merge into a locus was < 1Mb.

Identifying novel variants

For replicated and non-replicated variants with joint meta-analysis $P < 5 \times 10^{-8}$, GWAS Catalog [54] was used to identify novel variants. Gene-fish-oil interaction variants were checked in a literature search for their novelty. Variants within 1Mb from previously published variants associated with the same trait were considered to be non-novel.

Additional analyses

The R package qqman v 0.1.4 was used to generate Manhattan plots and QQ plots [55]. Regional loci plots were made using LocusZoom [56]. Data analysis was conducted in R v3.6.1 [57]. The Genotype-Tissue Expression Project (GTEx) data used were obtained from the GTEx Portal on 04/29/20 [58].

References

1. Thaipitakwong T, Aramwit P. A Review of the Efficacy, Safety, and Clinical Implications of Naturally Derived Dietary Supplements for Dyslipidemia. Am J Cardiovasc Drug. 2017;17(1):27-35. doi: 10.1007/s40256-016-0191-2.

2. Scicchitano P, Cameli M, Maiello M, Modesti PA, Muiesan ML, Novo S, et al. Nutraceuticals and dyslipidaemia: Beyond the common therapeutics. J Funct Food. 2014;6:11-32. doi: https://doi.org/10.1016/j.jff.2013.12.006.

3. Eslick GD, Howe PRC, Smith C, Priest R, Bensoussan A. Benefits of fish oil supplementation in hyperlipidemia: a systematic review and meta-analysis. Int J Cardiol. 2009;136(1):4-16. doi: https://doi.org/10.1016/j.ijcard.2008.03.092.

4. Lombardo YB, Chicco AG. Effects of dietary polyunsaturated n-3 fatty acids on dyslipidemia and insulin resistance in rodents and humans. A review. J Nutr Biochem. 2006;17(1):1-13. doi: https://doi.org/10.1016/j.jnutbio.2005.08.002.

5. Goldberg RB, Sabharwal AK. Fish oil in the treatment of dyslipidemia. Curr Opin Endocrinol, Diabetes and Obesity. 2008;15(2):167-74. doi: 10.1097/MED.0b013e3282f76728.

6. Hoffmann TJ, Theusch E, Haldar T, Ranatunga DK, Jorgenson E, Medina MW, et al. A large electronic-health-record-based genome-wide study of serum lipids. Nat Genet. 2018;50(3):401-13. Epub 2018/03/05. doi: 10.1038/s41588-018-0064-5.

7. Klarin D, Damrauer SM, Cho K, Sun YV, Teslovich TM, Honerlaw J, et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat Genet. 2018;50(11):1514-23. Epub 2018/10/03. doi: 10.1038/s41588-018-0222-9.

8. Madden J, Williams CM, Calder PC, Lietz G, Miles EA, Cordell H, et al. The Impact of Common Gene Variants on the Response of Biomarkers of Cardiovascular Disease (CVD) Risk to Increased Fish Oil Fatty Acids Intakes. Annu Rev Nutr. 2011;31(1):203-34. doi: 10.1146/annurev-nutr-010411-095239.

9. Klarin D, Damrauer SM, Cho K, Sun YV, Teslovich TM, Honerlaw J, et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat Genet. 2018;50(11):1514-23. Epub 2018/10/01. doi: 10.1038/s41588-018-0222-9.

10. Aung T, Halsey J, Kromhout D, Gerstein HC, Marchioli R, Tavazzi L, et al. Associations of Omega-3 Fatty Acid Supplement Use With Cardiovascular Disease Risks: Meta-analysis of 10 Trials Involving 77 917 IndividualsMeta-analysis of Associations of Omega-3 Fatty Acids and Cardiovascular RiskMeta-analysis of Associations of Omega-3 Fatty Acids and Cardiovascular Risk. JAMA Cardiol. 2018;3(3):225-33. doi: 10.1001/jamacardio.2017.5205.

11. Zheng J, Huang T, Yu Y, Hu X, Yang B, Li D. Fish consumption and CHD mortality: an updated meta-analysis of seventeen cohort studies. Public Health Nutr. 2012;15(4):725-37. Epub 2011/09/15. doi: 10.1017/s1368980011002254.

12. Martinelli N, Girelli D, Malerba G, Guarini P, Illig T, Trabetti E, et al. FADS genotypes and desaturase activity estimated by the ratio of arachidonic acid to linoleic acid are associated with inflammation and coronary artery disease. Am J Clin Nutr. 2008;88(4):941-9. Epub 2008/10/10. doi: 10.1093/ajcn/88.4.941.

13. Bokor S, Dumont J, Spinneker A, Gonzalez-Gross M, Nova E, Widhalm K, et al. Single nucleotide polymorphisms in the FADS gene cluster are associated with delta-5 and delta-6 desaturase activities estimated by serum fatty acid ratios. J Lipid Res. 2010;51(8):2325-33. Epub 2010/04/30. doi: 10.1194/jlr.M006205.

14. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203-9. doi: 10.1038/s41586-018-0579-z.

15. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. PLOS Comput Biol. 2012;8(12):e1002822. doi:10.1371/journal.pcbi.1002822.

16. Ye K, Gao F, Wang D, Bar-Yosef O, Keinan A. Dietary adaptation of FADS genes in Europe varied across time and geography. Nat Ecol Evol. 2017;1(7):0167. doi: 10.1038/s41559-017-0167.

17. Kothapalli KSD, Ye K, Gadgil MS, Carlson SE, O'Brien KO, Zhang JY, et al. Positive Selection on a Regulatory Insertion-Deletion Polymorphism in FADS2 Influences Apparent Endogenous Synthesis of Arachidonic Acid. Mol Biol Evol. 2016;33(7):1726-39. Epub 2016/03/29. doi: 10.1093/molbev/msw049.

18. Manning AK, LaValley M, Liu C-T, Rice K, An P, Liu Y, et al. Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP × environment regression coefficients. Genet Epidemiol. 2011;35(1):11-8. doi: 10.1002/gepi.20546.

19. Rao DC, Sung YJ, Winkler TW, Schwander K, Borecki I, Cupples LA, et al. Multiancestry Study of Gene-Lifestyle Interactions for Cardiovascular Traits in 610 475 Individuals From 124 Cohorts: Design and Rationale. Circ Cardiovasc Genet. 2017;10(3). Epub 2017/06/18. doi: 10.1161/circgenetics.116.001649.

20. Psaty BM, O'Donnell CJ, Gudnason V, Lunetta KL, Folsom AR, Rotter JI, 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. 2009;2(1):73-80. Epub 2009/12/25. doi: 10.1161/circgenetics.108.829747.

21. Sung YJ, de las Fuentes L, Winkler TW, Chasman DI, Bentley AR, Kraja AT, et al. A multi-ancestry genome-wide study incorporating gene–smoking interactions identifies multiple new loci for pulse pressure and mean arterial pressure. Hum Mol Genet. 2019;28(15):2615-33. doi: 10.1093/hmg/ddz070.

22. Noordam R, Bos MM, Wang H, Winkler TW, Bentley AR, Kilpeläinen TO, et al. Multi-ancestry sleep-by-SNP interaction analysis in 126,929 individuals reveals lipid loci stratified by sleep duration. Nat Comm. 2019;10:5121. doi: 10.1101/559393.

23. Figueroa V, Saez PJ, Salas JD, Salas D, Jara O, Martinez AD, et al. Linoleic acid induces opening of connexin26 hemichannels through a PI3K/Akt/Ca(2+)-dependent pathway. Biochim Biophys Acta. 2013;1828(3):1169-79. Epub 2012/12/25. doi: 10.1016/j.bbamem.2012.12.006.

24. Dlugosova K, Weismann P, Bernatova I, Sotnikova R, Slezak J, Okruhlicova L. Omega-3 fatty acids and atorvastatin affect connexin 43 expression in the aorta of hereditary hypertriglyceridemic rats. Can J Physiol Pharmacol. 2009;87(12):1074-82. Epub 2009/12/24. doi: 10.1139/y09-104.

25. Brisset AC, Isakson BE, Kwak BR. Connexins in vascular physiology and pathology. Antioxid Redox Signal. 2009;11(2):267-82. doi: 10.1089/ars.2008.2115.

26. Kooner JS, Chambers JC, Aguilar-Salinas CA, Hinds DA, Hyde CL, Warnes GR, et al. Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. Nat Genet. 2008;40(2):149-51. doi: 10.1038/ng.2007.61.

27. Zeng X-N, Yin R-X, Huang P, Huang K-K, Wu J, Guo T, et al. Association of the MLXIPL/TBL2 rs17145738 SNP and serum lipid levels in the Guangxi Mulao and Han populations. Lipids Health Dis. 2013;12(1):156. doi: 10.1186/1476-511X-12-156.

28. Iizuka K. The transcription factor carbohydrate-response element-binding protein (ChREBP): A possible link between metabolic disease and cancer. BBA-Mol Basis Dis. 2017;1863(2):474-85. doi: https://doi.org/10.1016/j.bbadis.2016.11.029.

29. Jump DB, Tripathy S, Depner CM. Fatty acid-regulated transcription factors in the liver. Annu Rev Nutr. 2013;33:249-69. Epub 2013/03/22. doi: 10.1146/annurev-nutr-071812-161139.

30. de Vries PS, Brown MR, Bentley AR, Sung YJ, Winkler TW, Ntalla I, et al. Multiancestry Genome-Wide Association Study of Lipid Levels Incorporating Gene-Alcohol Interactions. Am J Epidemiol. 2019;188(6):1033-54. doi: 10.1093/aje/kwz005.

31. Kaminski WE, Wenzel JJ, Piehler A, Langmann T, Schmitz G. ABCA6, a novel a subclass ABC transporter. Biochem Biophys Res Commun. 2001;285(5):1295-301. Epub 2001/08/02. doi: 10.1006/bbrc.2001.5326.

32. Harris WS. Fish oils and plasma lipid and lipoprotein metabolism in humans: a critical review. J Lipid Res. 1989;30(6):785-807. Epub 1989/06/01.

33. Innes JK, Calder PC. The Differential Effects of Eicosapentaenoic Acid and Docosahexaenoic Acid on Cardiometabolic Risk Factors: A Systematic Review. Int J Mol Sci. 2018;19(2). Epub 2018/02/10. doi: 10.3390/ijms19020532.

34. Greenwood DC, Gilthorpe MS, Cade JE. The impact of imprecisely measured covariates on estimating gene-environment interactions. BMC Med Res Methodol. 2006 4;6:21. doi: 10.1186/1471-2288-6-21.

35. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet. 2018;19(8):491-504. doi: 10.1038/s41576-018-0016-z.

36. Cundiff DK, Lanou AJ, Nigg CR. Relation of omega-3 Fatty Acid intake to other dietary factors known to reduce coronary heart disease risk. Am J Cardiol. 2007 1;99(9):1230-3. doi: 10.1016/j.amjcard.2006.12.032.

37. Rodgers GP, Collins FS. Precision Nutrition—the Answer to "What to Eat to Stay Healthy". JAMA. 2020;324(8):735–736. doi:10.1001/jama.2020.13601.

38. The National Heart, Lung, and Blood Institute (NHLBI) of the National Institutes of Health. Manual 8: Lipid and Lipoprotein Determinations. ARIC Protocol. 1987.

39. Cladis DP, Kleiner AC, Freiser HH, Santerre CR. Fatty Acid Profiles of Commercially Available Finfish Fillets in the United States. Lipids. 2014;49(10):1005-18. doi: 10.1007/s11745-014-3932-5.

40. Bostock J, McAndrew B, Richards R, Jauncey K, Telfer T, Lorenzen K, et al. Aquaculture: global status and trends. Philos T R Soc B. 2010;365(1554):2897-912. doi: 10.1098/rstb.2010.0170.

41. Henriques J, Dick JR, Tocher DR, Bell JG. Nutritional quality of salmon products available from major retailers in the UK: content and composition of n-3 long-chain PUFA. Brit J Nutr. 2014;112(6):964-75. Epub 2014/07/14. doi: 10.1017/S0007114514001603.

42. Sprague M, Dick JR, Tocher DR. Impact of sustainable feeds on omega-3 long-chain fatty acid levels in farmed Atlantic salmon, 2006–2015. Sci Rep-UK. 2016;6(1):21892. doi: 10.1038/srep21892.

43. Tur JA, Bibiloni MM, Sureda A, Pons A. Dietary sources of omega 3 fatty acids: public health risks and benefits. Brit J Nutr. 2012;107 Suppl 2:S23-52. Epub 2012/05/25. doi: 10.1017/s0007114512001456.

44. Bradbury KE, Young HJ, Guo W, Key TJ. Dietary assessment in UK Biobank: an evaluation of the performance of the touchscreen dietary questionnaire. J Nutr Sci. 2018;7:e6-e. doi: 10.1017/jns.2017.66.

45. Liu B, Young H, Crowe FL, Benson VS, Spencer EA, Key TJ, et al. Development and evaluation of the Oxford WebQ, a low-cost, web-based method for assessment of previous 24 h dietary intakes in large-scale prospective studies. Public health Nutr. 2011;14(11):1998-2005. Epub 2011/07/07. doi: 10.1017/s1368980011000942.

46. Purcell S, Chang CC. PLINK 1.9-beta3. Available from: www.cog-genomics.org/plink/1.9/.

47. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4(1). doi: 10.1186/s13742-015-0047-8.

48. Purcell S, ChangCC. PLINK 2.3 alpha. 2020. Available from: www.cog-genomics.org/plink/2.0/.

49. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74. doi: 10.1038/nature15393.

50. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48(10):1284-7. Epub 2016/08/29. doi: 10.1038/ng.3656.

51. Kutalik Z, Johnson T, Bochud M, Mooser V, Vollenweider P, Waeber G, et al. Methods for testing association between uncertain genotypes and quantitative traits. Biostatistics. 2010;12(1):1-17. doi: 10.1093/biostatistics/kxq039.

52. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190-1. doi: 10.1093/bioinformatics/btq340.

53. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;8(1):1826. Epub 2017/12/01. doi: 10.1038/s41467-017-01261-5.

54. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005-d12. Epub 2018/11/18. doi: 10.1093/nar/gky1120.

55. Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. bioRxiv. 2014:005165. doi: 10.1101/005165.

56. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010;26(18):2336-7. Epub 2010/07/17. doi: 10.1093/bioinformatics/btq419.

57. Team R Consortium. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.

58. Battle A, Brown CD, Engelhardt BE, Montgomery SB. Genetic effects on gene expression across human tissues. Nature. 2017;550(7675):204-13. Epub 2017/10/13. doi: 10.1038/nature24277.

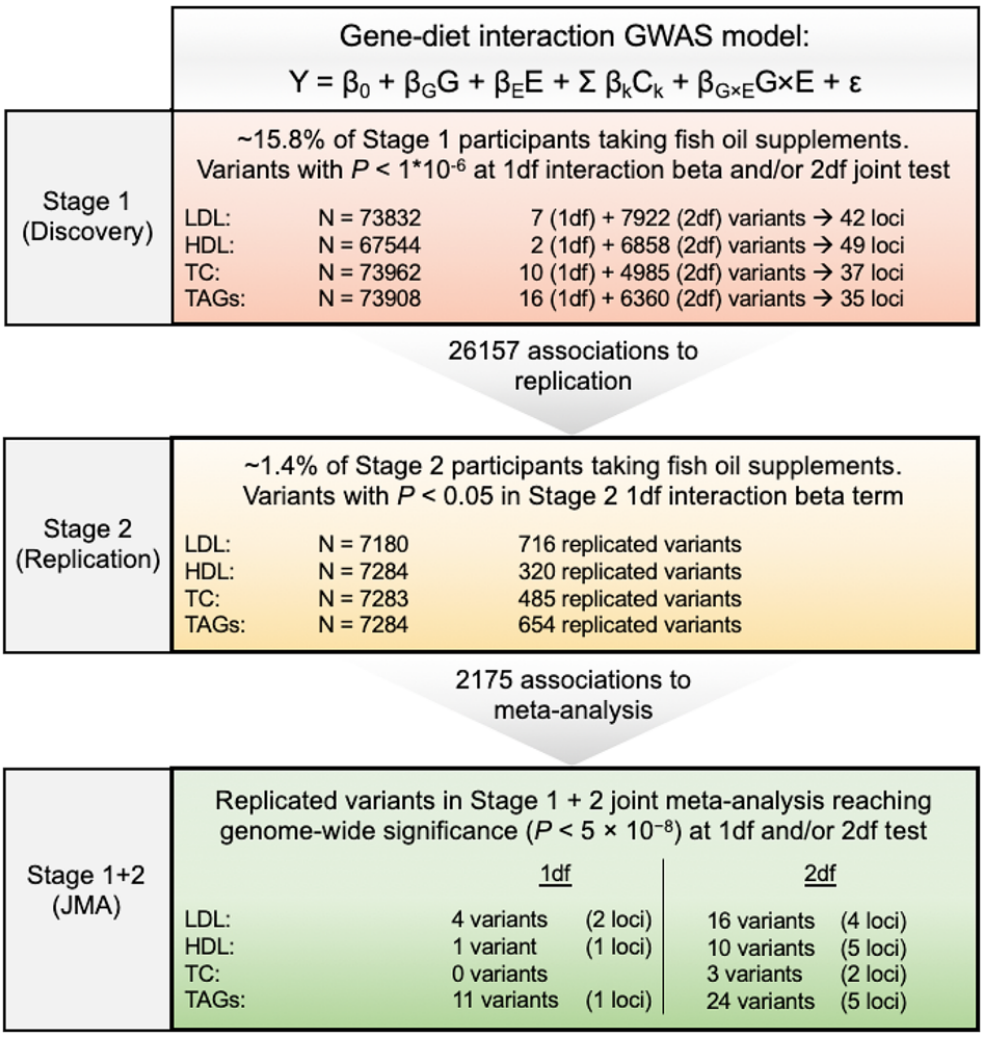

Figure 2.1. Overview of the analysis performed in this study.

A three-stage discovery, replication, and meta-analysis process was used to identify significant variants. Stage 1 revealed 26,157 associations with 1df and/or 2df P < 1×10−6 in a cohort of up to 73,962 participants. Of these associations, 2,175 were replicated in a cohort of up to 7,284 participants. In meta-analysis, 4 1df loci (Table 1) and 16 2df loci (13 additional loci, Table 2) reached the genome-wide significance of P < 5 × 10−8. TC, total cholesterol; TAGs, triglycerides.

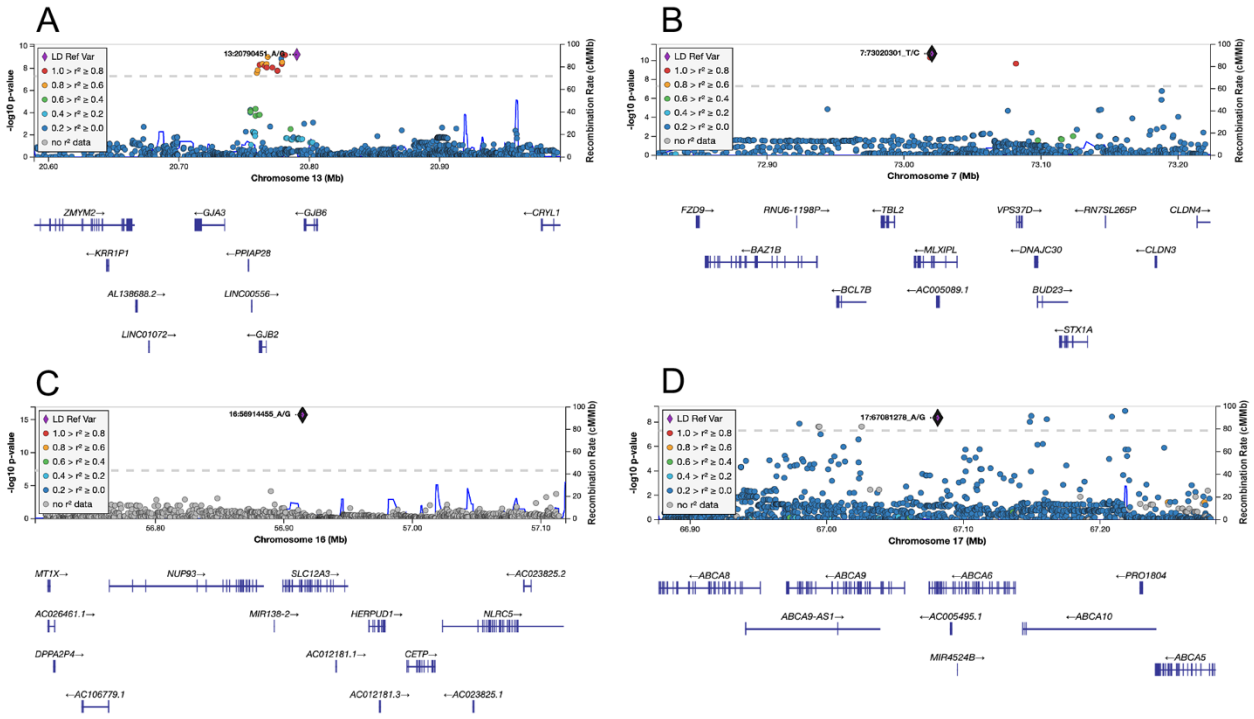

Figure 2.2. LocusZoom for genome-wide significant ($P < 5 \times 10{-}8$) replicated gene-fish oil interaction loci.

(A) rs112803755 × fish oil and TAGs, stage 1 + 2 1df tests (n = 81,192). (B) rs799157 × fish oil and LDL-C, stage 1 + 2 1df tests (n = 81,012). (C) rs148931404 × fish oil and HDL-C, stage 1 + 2 1df tests (n = 74,824). (D) rs77542162 × fish oil and LDL-C, stage 1 + 2 1df tests (n = 81,012).

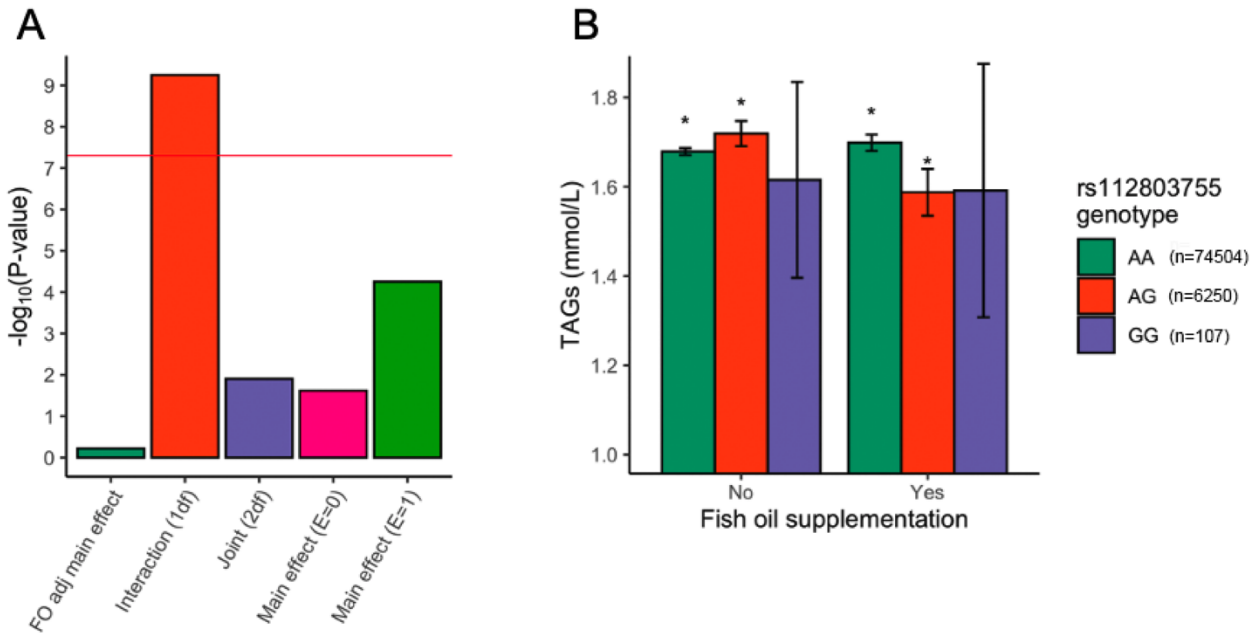

Figure 2.3. Significant results for the replicated interaction locus with lead SNP rs112803755.

(A) rs112803755 P-values in five regression models. The red line is the negative log10-transformed genome-wide significance of 5 × 10−8. (B) Triglyceride lowering effect of fish oil supplementation on rs112803755 heterozygotes. Levels of TAGs stratified by genotypes at rs112803755 and fish oil supplementation status. Error bars show 95% confidence intervals. Exact numbers and sample sizes can be found in S7 Table.

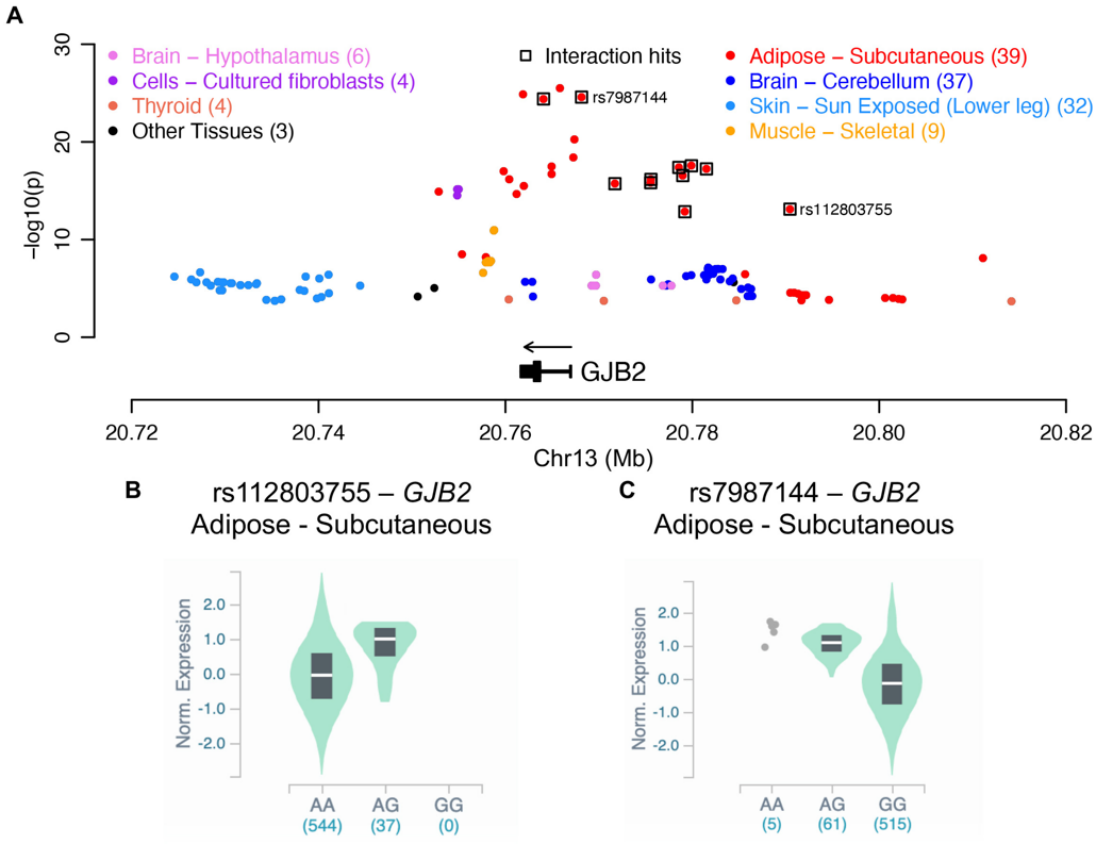

Figure 2.4. The gene-fish oil interaction locus with lead SNP rs112803755 overlaps eQTLs of *GJB2*.

(A) Genetic variants significantly associated with the expression of *GJB2* as detected in the GTEx project. Colors indicate the tissues or cells. For variants with significant association in more than one tissues, the most significant p value is shown. The association of (B) rs112803755 and (C) rs7987144 with the expression of *GJB2* in subcutaneous adipose tissues.

Table 2.1. Loci with significant interaction between fish oil supplementation and blood lipid levels.

Listed variants represent the lead association within a 1 Mb region for 1df tests of variant × fish oil interaction after meta-analysis. The name of the nearest gene is listed with each lead variant. Bold *P*-values indicate meeting the genome-wide significance threshold of $P < 5 \times 10^{-8}$. Main effect *P*-values are calculated using Stage 1 (UK Biobank) participants only, and without interaction (Eq (2); stratified for exposure groups as in Eq (3)). Effect, beta coefficient of the minor allele dose term (β_G in Eq (1)); MAF, minor allele frequency; SE, standard error; Int effect, beta coefficient of the interaction term ($\beta_{G\times E}G\times E$ in Eq (1)). Lipid traits were measured in mmol/L.

Stage 1 + 2 Meta-analysis genome-wide significant interaction loci													**Stage 1**			
hg19 chr: pos	**rsID (nearest gene)**	**Minor (effect) allele (avg freq)**	**Reference allele**	**Stage1 MAF / Stage 2 MAF**	**Lipid trait**	**n**	**Effect**	**SE**	**Int Effect**	**Int SE**	**1df interaction *P*-value**	**2df joint *P*-value**	**n**	**Adj. Main effect *P*-value**	**Main effect *P*-value (E = 0)**	**Main effect *P*-value (E = 1)**
13:20790451	rs112803755 (*GJB6*: ~5650 bp downstream)	g (0.0410)	a	0.0416 / 0.0352	TAGs	81192	0.04272	0.0447	-0.3084	0.1048	**5.65E-10**	0.01247	73908	0.6007	2.42E-02	5.59E-05
7:73020301	rs799157 (*MLXIPL*: Synonymous Variant)	t (0.0407)	c	0.0435 / 0.0112	LDL	81012	0.07162	0.05678	1.459	0.1371	**1.92E-11**	**1.93E-33**	73832	**3.33E-08**	5.36E-06	8.14E-04
17:67081278	rs77542162 (ABCA6: Missense Variant)	g (0.0218)	a	0.0226 / 0.0134	LDL	81012	0.1741	0.06495	-1.900	0.1224	**4.48E-09**	**6.58E-63**	73832	**5.40E-23**	**4.24E-21**	1.55E-03
16:56914455	rs148931404 (*SLC12A3*: Intron Variant)	a (0.0221)	g	0.0226 / 0.0172	HDL	74824	0.05594	0.02328	0.7734	0.04719	**1.82E-16**	**5.04E-91**	67544	**2.70E-16**	**5.67E-12**	3.49E-06

https://doi.org/10.1371/journal.pgen.1009431.t001

Table 2.2. Loci with significant 2df joint test between fish oil supplementation and blood lipid levels.

Listed variants represent the lead association within a 1-Mb region for 2df tests of variant × fish oil interaction after meta-analysis. The name of the nearest gene is listed with each lead variant. Bold *P*-values indicate meeting the genome-wide significance threshold of $P < 5 \times 10^{-8}$. Main effect *P*-values are calculated using Stage 1 (UK Biobank) participants only, and without interaction (Eq (2); stratified for exposure groups as in Eq (3)). Effect, beta coefficient of the minor allele dose term (β_G in Eq (1)); MAF, minor allele frequency; SE, standard error; Int effect, beta coefficient of the interaction term ($\beta_{G\times E}G\times E$ in Eq (1)). Lipid traits were measured in mmol/L.

Stage 1 + 2 Meta-analysis replicated genome-wide significant loci													Stage 1			
hg19 chr:pos	**rsID (nearest gene)**	**Minor (effect) allele (avg freq)**	**Reference allele**	**Stage1 MAF / Stage 2 MAF**	**Lipid trait**	**n**	**Effect**	**SE**	**Int Effect**	**Int SE**	**1df interaction *P*-value**	**2df joint *P*-value**	**n**	**Adj. Main effect *P*-value**	**Main effect *P*-value (E = 0)**	**Main effect *P*-value (E = 1)**
6:34094919	rs115675705 (*GRM4*: Intron Variant)	g (0.0357)	a	0.0367 / 0.0268	HDL	74824	-0.03882	0.01795	-0.3348	0.04479	0.00873	**1.23E-19**	67544	**3.98E-09**	**3.43E-09**	2.24E-01
8:19722204	rs117860853 (*LPL*: ~217kb upstream)	a (0.0135)	g	0.0137 / 0.0116	HDL	74824	-0.08003	0.02691	-0.4139	0.05311	1.48E-02	**5.46E-28**	67544	**5.72E-24**	**3.47E-22**	3.00E-03
11:116916060	rs144018203 (*SIK3*: Intron Variant)	c (0.0102)	g	0.0107 / 0.0052	HDL	74824	-0.05472	0.03791	0.4387	0.05912	0.04122	**1.03E-16**	67544	**6.21E-17**	**1.16E-13**	1.00E-04
17:42061277	rs147438979 (*PYY*: Intron Variant)	c (0.011)	g	0.0116 / 0.006	HDL	74824	-0.06327	0.02952	-0.3336	0.05753	4.18E-02	**2.75E-16**	67544	**3.03E-09**	**1.78E-08**	5.58E-02
1:55583210	rs530804537 (USP24: Intron Variant)	a (0.011)	g	0.011 / 0.0012	LDL	81012	-0.07453	0.08869	-1.133	0.1360	0.1325	**2.68E-31**	73832	**2.01E-22**	**6.23E-21**	4.72E-03
19:45198060	rs112952132 (*LOC107985305*: Intron Variant)	t (0.01)	c	0.01 / 0.0101	LDL	81012	-0.2051	0.08381	-1.055	0.1329	2.77E-02	**7.58E-34**	73832	**9.49E-21**	**8.44E-18**	1.45E-01
7:72921771	rs117788606 (*BAZ1B*: Intron Variant)	c (0.0112)	t	0.0120 / 0.0032	TAGs	81192	-0.06165	0.09183	1.012	0.1464	2.67E-04	**3.93E-16**	73908	6.68E-07	7.64E-07	2.67E-01
8:19768150	rs142084074 (*LOC107986921*: Intron Variant)	a (0.015)	g	0.0150 / 0.0148	TAGs	81192	-0.1182	0.07241	0.911	0.1319	2.72E-02	**3.153E-12**	73908	**1.95E-15**	**1.42E-12**	2.84E-04
11:116916060	rs144018203 (*SIK3*: Intron Variant)	c (0.0103)	g	0.0108 / 0.0052	TAGs	81192	0.2531	0.1359	-0.9079	0.1735	0.008127	**2.48E-09**	73908	**3.50E-56**	**1.31E-50**	3.69E-07
15:43820717	rs55707100 (*MAP1A*: Missense Variant)	t (0.0249)	c	0.0247 / 0.0273	TAGs	81192	0.1201	0.05043	0.6116	0.1122	0.01044	**2.36E-13**	73908	**5.77E-21**	**2.03E-17**	5.18E-05
19:19365178	rs141844019 (*HAPLN4*: 500B Downstream Variant)	t (0.0104)	c	0.0109 / 0.0054	TAGs	81192	-0.06922	0.09157	1.637	0.1366	1.64E-06	**4.27E-53**	73908	3.91E-07	1.59E-06	1.06E-01
1:55583210	rs530804537 (*USP24*: Intron Variant)	a (0.0109)	g	0.0113 / 0.0073	Tot. Chol.	81245	-0.05886	0.1084	-1.057	0.1603	0.1002	**2.24E-20**	73962	**1.19E-17**	**9.04E-17**	1.88E-02
17:67081278	rs77542162 (*ABCA6*: Missense Variant)	g (0.0218)	a	0.0226 / 0.0135	Tot. Chol.	81245	0.1599	0.06831	-1.587	0.1261	4.58E-07	**1.80E-41**	73962	**3.06E-11**	**4.45E-11**	1.10E-01

https://doi.org/10.1371/journal.pgen.1009431.t002

Supplementary Figure and Table legends

Supplementary data can be downloaded from: https://doi.org/10.1371/journal.pgen.100943.

Supplementary Figures

S1 Fig. Manhattan plots for Stage 1 1df interaction term *P*-values and 2df joint test *P*-values for lipid traits. Plots show post-genomic control values.

S2 Fig. QQ plots for Stage 1 1df interaction term *P*-values and 2df joint test *P*-values for lipid traits. Plots show post-genomic control values.

S3 Fig. Fish oil supplementation taken at two time points. The number of UK Biobank participants who responded yes/yes, no/no, yes/no, and no/yes to the two dietary assessment time points at the initial assessment and in the 24-hour follow-up questionnaire are shown. Numbers reflect the total number of participants who answered in both assessments, but not the number of participants used in this study after quality control.

S4 Fig. Visualization of the G×E interaction regression model. $Y = \beta_0 + \beta_G G + \beta_E E + \Sigma\, \beta_k C_k + \beta_{G\times E} G\times E + \varepsilon$, where Y = phenotype, G = minor variant dosage (0/1/2 coding), E = environmental exposure, C_k = covariates, and G×E = interaction term. In this study, Y is a continuous lipid trait, and E is a binary variable representing the presence or absence of self-reported dietary fish oil supplementation.

Supplementary Tables

S1 Table. Participant characteristics. Participant characteristics, by blood lipid phenotype, for those included in GEI analyses for Stage 1 (UK Biobank) and Stage 2 (ARIC). Mean and standard deviation values are shown for blood lipid phenotypes and for applicable covariates.

S2 Table. Numbers of stage 1 significant variants. Variants which passed a significance threshold of $P <$ 1e-06 in Stage 1 (UK Biobank) are counted here. Significance was assessed for both 1df interaction terms and 2df joint terms. Variant count and number of independent loci are shown, as well as unique variants with 1df and 2df tests.

S3 Table. Numbers of replicated variants. Variants which reached Stage 1 $P <$ 1e-06 (in either 1df or 2df) and were found to have 1df $P < 0.05$ in Stage 2 interaction models.

S4 Table. Numbers of genome-wide significance loci in only Stage 1. Counts of variants and loci which met the significance threshold of P< 5e-08 in Stage 1 (in either 1df or 2df) but which were not replicated in Stage 2. Note that no Stage 1 1df *P*-values reached this threshold so all variants in this table refer to their 2df joint test *P*-values.

S5 Table. Non-replicated genome-wide significant Stage 1 variants. Full details for the loci counted in Table S4. Effect, beta coefficient of the minor allele dose term (β_G in Eq (1)); MAF,

minor allele frequency; SE, standard error; Int effect, beta coefficient of the interaction term ($\beta_{G\times E}$G×E in Eq (1)). Lipid traits were measured in mmol/L. All *P*-values are calculated using Stage 1 (UK Biobank) participants only.

S6 Table. Numbers of genome-wide significance loci after meta-analyses. Counts of replicated results reaching genome-wide significance ($P < 5e0{-}8$) in Stage 1+2 meta-analyses. Significant variants determined by 1df *P*-values (top) and 2df *P*-values (bottom).

S7 Table. Data used in Fig 3B. Fish oil status, number of G alleles at rs112803755, mean triglycerides, sample size, standard deviation of triglycerides, and 95% confidence interval for combined participants from Stage 1 and Stage 2.

CHAPTER 3

FIFTY-ONE NOVEL AND REPLICATED GWAS LOCI FOR POLYUNSATURATED AND MONOUNSATURATED FATTY ACIDS IN 124,024 EUROPEANS[2]

[2] Francis M, Sun Y, Xu H, Brenna JT, Ye K. (2022) Fifty-one novel and replicated GWAS loci for polyunsaturated and monounsaturated fatty acids in 124,024 Europeans. medRxiv 2022.05.27.22275343; doi: https://doi.org/10.1101/2022.05.27.22275343.

Reprinted here with permission of the publisher.

Abstract

Circulating polyunsaturated and monounsaturated fatty acid (PUFA and MUFA) levels, whose imbalances co-occur with human metabolic diseases, have strong heritable components. We performed the largest genome-wide association study (GWAS) to-date on fourteen PUFA and MUFA phenotypes, measured by nuclear magnetic resonance in plasma. We identified 612 significant locus-phenotype associations (115 unique loci; $P < 1.678*10^{-8}$) in a European cohort from UK Biobank (UKB-EUR; n=101,729). Replication of five phenotypes (omega-3, omega-6, DHA, LA, MUFAs) was conducted in two external European studies: FinMetSeq (n=8,751) and a meta-analysis by Kettunnen *et al.* (n=3,644-13,544). Meta-analysis of these three studies yielded 254 significant locus-phenotype associations (109 unique loci; $P < 2.439*10^{-8}$); we identified 87 novel loci, 51 of which were replicated. A transcriptome-wide association study of the UKB-EUR cohort revealed an additional twelve novel loci. This study improves our understanding of the genetic architecture of unsaturated fatty acids and can inform future genotype-based dietary interventions.

Acknowledgements

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institute of Health under award numbers T32GM007103 (MF) and R35GM143060 (KY). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Special thanks to the Georgia Advanced Computing Resource Center (GACRC) at the University of Georgia for supporting our data analyses.

Competing interests

The authors declare no competing interests.

Author contributions

KY conceived and supervised the study. KY and MF designed the analyses. MF led the data analyses, with assistance from YS and HX. MF, KY, and JTB interpreted the results. MF and KY wrote the manuscript. MF created data visualizations. All authors reviewed, revised, and approved the final paper.

Introduction

Polyunsaturated fatty acids (PUFAs) are dietary fats containing two to six double bonds along linear carbon chains from 14 to 22 carbons in length. Imbalances of tissue PUFAs are involved in the pathophysiology of a broad array of diseases, including cardiovascular disease, cancer, depression, and dementia [1-3]. Omega-3 long-chain PUFAs (n-3 LCPUFAs) have been consistently shown to improve aspects of metabolic syndrome related to the risk factors of cardiovascular disease and obesity, such as insulin resistance, hypertension, and dyslipidemia [1,4,5]. Omega-6 (n-6) LCPUFAs have been associated with both positive and negative health outcomes [6]. Excess n-6 linoleic acid (LA) suppresses tissue and circulating n-3 LCPUFAs, due to common enzymes operating on both PUFA families; balance in dietary n-6 and n-3 is necessary to avoid suppression of the functional n-3 LCPUFAs, eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA) [7]. Modest overall dietary PUFAs and an n-6/n-3 ratio up to 4/1 have been recommended, while the typical modern industrialized diet has a ratio approximating 15/1 [8,9].

Dietary PUFA intake strongly influences circulating fatty acid levels [10]. Heritability analyses in twin studies and large cohorts have indicated that substantial genetic components also contribute to determining circulating PUFA levels [11-13]. Previous genome-wide association studies (GWAS) have identified 37 unique genomic loci related to PUFAs and monounsaturated fatty acids (MUFAs; Supplementary Table 1). However, they collectively only explain a small fraction of the phenotypic variance [12,14], suggesting more loci may be found in large-sample GWAS. High-throughput nuclear magnetic resonance spectroscopy (NMR) enables rapid large-scale quantification of metabolic biomarkers [15]. The UK Biobank (UKB) has recently released NMR data of plasma fatty acid levels for over 110,000 participants, presenting an opportunity to identify novel genetic loci associated with PUFAs and MUFAs.

Here, we perform a linear mixed model (LMM) GWAS to identify the genetic variants associated with the fourteen available NMR-derived plasma unsaturated fatty acid phenotypes in UKB participants: n-3 LCPUFAs, n-6 LCPUFAs, DHA, LA, total PUFAs, total MUFAs, the percentages of each of these fatty acid groups per the total amount of fatty acids, as well as the ratios of n-6/n-3 and PUFA/MUFA. Our discovery cohort consisted of UKB participants with NMR PUFA and MUFA trait measurements, who were determined to be genetically European (EUR) by the Pan-UK Biobank project [16] (n=101,729). Additional multi-ancestry replication analyses were performed in African (AFR), Central and South Asian (CSA), and East Asian (EAS) UKB participants (n=4,400). Two external EUR studies were used in our replication and meta-analysis: the Finnish Metabolic Sequencing (FinMetSeq) study [11] (n=8,751); and a meta-analysis of 14 datasets derived from ten EUR GWAS studies by Kettunnen et al. [14] (n=3,644 to 13,544). In our five meta-analyzed traits (n-3, DHA, n-6, LA, MUFAs) we identified 254 significant locus-trait associations, 102 of which were novel and replicated. These consisted of 51 unique, novel and replicated loci across traits (overview in Figure 1). We also performed a transcriptome-wide association analysis in the fourteen discovery traits, which revealed an additional 12 novel and significant loci that were not found in GWAS. We have provided follow-up analyses and functional interpretations to put these significant associations into plausible biological context, and provide a contemporary description of the genetics involved in circulating PUFA and MUFA levels.

Results

Discovery analysis

We performed a GWAS of NMR-measured polyunsaturated fatty acid (PUFA) and monounsaturated fatty acid (MUFA) traits in individuals of European ancestry (EUR). An overview of our three-stage discovery, replication, and meta-analysis approach can be seen in Figure 1. First, we performed a GWAS discovery analysis using a UKB-EUR discovery cohort composed of 101,729 participants with NMR data, who were designated as genetically EUR by the Pan UKBB project [16]. Mean age of participants was 56.8 years old and 45.98% were male (Supplementary Figure 1; Supplementary Table 2). Two discovery stage sensitivity models were compared, M1 and M2. The *P*-values were highly correlated between the two models and no residual confounding was observed with either model (Supplementary Figure 2; Supplementary Table 3). The ranges of LD score regression (LDSC) intercepts and genomic control (λ) were 1.00-1.041 and 1.1475-1.2545 for M1 and 0.99-1.045 and 1.1475-1.2 for M2, respectively (Supplementary Table 3). The number of significant variants identified in M2 was always comparable to or greater than the same trait of M1 despite no inflation identified in M2; we inferred this was because our inclusion of relevant covariates reduced residual variability and enhanced the statistical power for variant discovery in M2 [17]. All results are therefore based on M2.

A total of 15,578,593 variants were tested for associations with all fourteen available PUFA and MUFA traits in our discovery analysis. We found a total of 146,982 significant associations (35,869 unique variants across traits; Supplementary Figure 3; Supplementary Table 3) at the significance threshold corrected for the effective number of traits ($P < 1.678\times10^{-8}$). Conditional and joint analysis (COJO) identified 968 independent significant associations in this cohort (471 unique variants; Supplementary Tables 4 and 5). We used FUMA to group significant associations

into LD blocks and merged loci < 250Kb apart; this yielded 612 genomic risk loci (115 unique loci across traits; Supplementary Tables 4 and 6). There were 404 novel locus-trait associations (95 unique loci across phenotypes) identified in the discovery stage.

Replication and meta-analysis

Our primary replication analysis utilized two external EUR GWAS studies: FinMetSeq and Kettunen et al. (Supplementary Table 11). These studies contained five out of the fourteen traits analyzed in the discovery stage: n-3, n-6, DHA, LA, and MUFAs. After munging all three EUR studies to ensure high quality alleles and to harmonize alleles to the reference genome, ~8.7 million variants overlapped between Kettunen et al. and UKB-EUR, and 209,509 variants overlapped between FinMetSeq and UKB-EUR. Across the five available phenotypes, there were 19,929 UKB-EUR associations (8,543 unique variants) replicated at $P < 0.05$ in one of the two external EUR studies. Of these, 615 associations (266 unique variants) were replicated in both studies (Supplementary Table 11).

We performed additional replication analyses across the three UKB multi-ancestry groups with adequate sample sizes, to evaluate the reproducibility of associations found in the UKB-EUR cohort. No significant phenotypic differences were found between UKB ancestry groups in any of fourteen PUFA and MUFA traits (Supplementary Figure 1; Supplementary Table 2). Mixed linear model-based association analysis (MLMA) was performed in UKB African (UKB-AFR), Central and South Asian (UKB-CSA) and East Asian (UKB-EAS) cohorts (Supplementary Figure 4). Counts of UKB-EUR associations replicated ($P < 0.05$) were: UKB-AFR 5,327 (2,358 unique variants), UKB-CSA 16,560 (5,179 unique variants), and UKB-EAS 5,466 (2,113 unique variants) (Supplementary Table 12). Out of the 612 significant loci for UKB-EUR associations, 170 were replicated in ≥ 1 UKB multi ancestry group (Supplementary Tables 6 and 12). Interestingly, despite

having a smaller sample size, UKB-EAS contained more replicated UKB-EUR loci than UKB-AFR (46 vs. 31).

Meta-analysis was performed on variants which appeared in at least two out of three EUR studies, and ~10,200,000 variants were tested. Across the five PUFA and MUFA traits there were 38,344 significant associations ($P < 2.439\times10^{-8}$; 17,301 unique variants across traits; Figure 2; Supplementary Table 13). LDSC intercepts for the five meta-analyzed traits ranged from 1.013 (SE = 0.0079) for DHA to 1.04 (0.011) for MUFAs, indicating that there was not inflation or residual confounding (Supplementary Table 13). COJO revealed 402 independent significant associations (258 unique variants; Supplementary Tables 13 and 14). Of these 402 significant COJO associations, 265 were replicated in at least one external EUR study ($P < 0.05$). When grouping all significant meta-analysis variants into loci, we found 254 significant loci (109 unique across traits; Supplementary Tables 13 and 15). Of the 254 grouped meta-analysis loci, 171 were replicated in at least one external EUR study (Supplementary Table 18).

Our literature review found a total of 210 previously reported significant PUFA and MUFA GWAS associations (106 unique variants, 37 loci based on 1Mb grouping; Supplementary Table 1). We compared these known loci with our significant meta-analysis loci. Of our 254 meta-analysis loci-trait associations, 173 were novel (87 unique across traits), and 102 of our novel loci were replicated in at least one external EUR study. This yielded a total of 51 unique loci across traits (Supplementary Table 13).

Notable associated genes

Among the 109 unique genomic loci identified in our meta-analysis of five PUFA and MUFA traits, thirteen loci were associated with all five traits (Figure 2; Supplementary Table 15), nine of which were identified in previous GWAS. These loci spanned genes that are well-known

in lipid metabolism, including the apolipoprotein gene clusters at chr11q23 (*APOA5*, *APOA4*, *APOC3*, and *APOA1*) and chr19q13 (*APOE, APOC1, APOC4*, and *APOC2*), plus *APOB*, LDL receptor adaptor protein 1 (*LDLRAP1)*, LDL receptor (*LDLR*), lipase C (*LIPC*), and lysophosphatidic acid receptor 2 (*LPAR2*). Another notable known gene is glucokinase regulator (*GCKR*), which has been associated with docosapentaenoic acid (DPA) and palmitoleic acid [18-20]. Of the four novel loci associated with all five traits, the locus of chr18q21 covers a candidate gene of lipase G (*LIPG*). The candidate genes at loci chr1p13 and chr8q24 include *PSRC1*, *SORT1, TRIB1*, and *SQLE*. Both *PSRC1* and *TRIB1* have been previously associated with familial hypercholesterolemia [21]. *SQLE* encodes squalene epoxidase, a rate-limiting enzyme catalyzing the first oxygenation step in sterol biosynthesis.

When excluding MUFAs and only considering n-3, DHA, n-6, and LA, there are three loci associated with these four traits, all of which are novel and externally replicated. These are located at chr2q21 (the *LCT* locus), chr4q13, and chr7p22 (Supplementary Table 15). The chr4q13 locus (Figure 3A) encompasses multiple members of the UDP-glycosyltransferase family (*UGT2B17; UGT2B10; UGT2B11; UGT2A1; UGT2A2;* and *UGT2A1*), which play important roles in bile acid (BA) detoxification by catalyzing the glucuronidation of BA substrates, and impact dietary lipid absorption [22]. One candidate gene at chr7p22 is *CYP2W1* (Figure 3B), a member of the cytochrome P450 superfamily, which encodes monooxygenases and oxidizes steroids, fatty acids, and xenobiotics [23].

Nine genomic loci are only associated with n-3, DHA, or both, eight of which are novel. The one known locus has at least two relevant candidate genes, choline kinase alpha (*CHKA*) and carnitine palmitoyltransferase 1A (*CPT1A*). *CHKA* encodes the initial enzyme which catalyzes the phosphorylation of ethanolamine in the CDP-choline pathway for phosphatidylcholine

biosynthesis. *CPT1A* locates in the outer membrane of mitochondria and catalyzes the transport of long chain fatty acids from cytosol into mitochondria, enabling beta-oxidation. One novel locus at chr10q23 has a cluster of genes in the cytochrome P450 superfamily (*CYP2C18, CYP2C19, CYP2C9,* and *CYP2C8*). Another notable novel locus at chr11q24 has a candidate gene of *ST3GAL4*, which is involved in the terminal sialylation of glycolipids. There are 35 loci associated with only n-6, LA, or both; 33 of these are novel. Multiple novel candidate genes are implicated in lipid metabolism, such as LDL receptor related protein 2 (*LRP2*), NPC1 like intracellular cholesterol transporter 1 (*NPC1L1*), scavenger receptor class B member 1 (*SCARB1*), phospholipase C gamma 1 (*PLCG1*), and lipin 3 (*LPIN3*). Three additional novel loci carry cytochrome P450 genes, including chr2q33 (*CYP20A1*), chr8q12 (*CYP7A1)*, and chr19q13 (*CYP2A6*). Another two notable candidate genes are arachidonate 5-lipoxygenase (*ALOX5*) and peroxisome proliferator activated receptor delta (*PPARD*). ALOX5 catalyzes conversion of DHA to signaling molecules [24]; supplemental DHA modulates PPARD in adult men [25].

The two key sets of genes which catalyze LCPUFA biosynthesis are fatty acid desaturase (*FADS*) and elongase protein family genes (*ELOVL*) [7]. We have re-confirmed the primary importance of *FADS* genes in n-3 LCPUFA genetics, as these genes had the most significant *P*-values in our meta-analysis, at lead SNPs rs174528 (DHA, $P < 1E\text{-}300$; MAF = 0.39) and rs509360 (n-3, $P < 1E\text{-}300$; MAF = 0.33). Both of these variants were mapped to *FADS1*, *FADS2*, and *FADS3*. The *ELOVL* gene family has been rarely associated in GWAS; of the seven *ELOVL* genes, only *ELOVL2* (chr6:10,980,992-11,044,547) has been previously associated with PUFAs, specifically with the n-3 traits DHA, EPA, and DPA (Supplementary Table 1, locus 13). We found an association in UKB-EUR with the *ELOVL2* locus surpassing the suggestive significance threshold ($P < 5e\text{-}05$) for DHA (rs9380082, $P = 2.6e\text{-}05$), but did not find this locus associated at

genome-wide significance. We did identify a novel, unreplicated association for MUFAs to total fatty acids percentage close by to *ELOVL6* (Supplementary Table 6). The variant rs114816312 (chr4:110,578,226; P = 1.9e-08) is ~330Kbp downstream from *ELOVL6*. This tentative trait association would be consistent with previous findings that demonstrate the primary role of *ELOVL6* gene product in elongating MUFAs [26]. This variant rs114816312 is also found within phospholipase A2 group XIIA (*PLA2G12A*); the primary function of PLA2 enzymes is to remove arachidonic acid (AA) from phospholipids for the production of eicosanoids.

ACSL6 is part of the Acyl-CoA synthetase (ACS) family of enzymes which catalyze the formation of acyl-CoAs from free fatty acids [27]. We report two novel and externally replicated associations (Supplementary Table 15): n-3 with rs273913 (MAF = 0.39; locus start = chr5:131,407,493; Supplementary Figure 5), and DHA with rs166635 (MAF = 0.31; locus start = chr5:131,590,114), that are ~60Kbp and ~242Kbp upstream from the *ACSL6* gene (chr5:131,142,683-131,347,936; reverse strand). It should be noted with regard to novelty, this locus was previously reported to be associated with AA at rs274559 [6], but their *P*-value at 3.81e-06 did not reach genome-wide significance. *ACSL6* expression has been previously linked to DHA enrichment in the brain. Our finding of significant associations with only n-3 and DHA, not n-6, LA, or MUFAs, is consistent with previous experimental reports [27,28].

Trait correlations and heritability

We evaluated the shared genetic basis across PUFA and MUFA traits using genetic correlations (r_g). For the fourteen traits in the discovery analysis, the levels of phenotypic (r_p) and genetic correlations were broadly consistent across all 91 trait-pairs, with slightly stronger genetic correlations (Figures 4A, 4B; Supplementary Table 7). Among the 78 trait-pairs that had both nominally significant phenotypic and genetic correlations (adj$P$ < 0.05), 57 had stronger

correlations at the genetic level (binomial test p = 2.79e-5). Interestingly, the genetic correlations between the absolute concentrations and their relative percentages of total fatty acids were not always high, ranging from 0.89 for MUFAs, 0.84 for omega-3, 0.73 for DHA, -0.56 for omega-6, -0.32 for PUFAs, to 0.017 for LA. These medium to low genetic correlations suggest the involvement of different biological mechanisms and emphasize the need to perform separate GWAS for absolute concentrations and relative percentages. Moreover, the correlation between n-3 and n-6 was moderate with the absolute concentrations ($r_g = 0.67$; $r_p = 0.45$), and low using their respective percentages of total fatty acids ($r_g = -0.11$; $r_p = -0.12$), indicating that there is substantial unique genetic basis for either trait.

In the fourteen traits tested in the discovery cohort, SNP-based heritability (h^2) calculated in LDSC ranged from 0.12 (SE = 0.018) for LA percentage to 0.20 (0.032) for MUFA percentage (Supplementary Table 4). Using individual-level genotype data in the discovery cohort, BOLT-REML found the h^2 of the six traits measured in absolute concentration units (n-3, n-6, DHA, LA, PUFAs, MUFAs) to range from 0.16 (0.0065) for LA to 0.22 (0.0066) for MUFAs (Supplementary Table 7). The lower range of LDSC when compared to BOLT-REML is consistent with our expectation that LDSC reports the lower bound of h^2 estimates. SNP-based h^2 was similar in meta-analysis, ranging from 0.12 (0.022) for DHA to 0.16 (0.023) for MUFAs (Figure 4C; Supplementary Table 13). Meta-analysis FUMA-defined significant loci explained between 5.09 to 8.12% of variance in the five traits examined; the novel loci we identified contributed between 0.63-2.95% of that variance. COJO independent variants explained between 8.47-10.78% of trait variance; the amount of SNP heritability explained ranged from 55.77% for MUFAs to 79.65% for n-3 (Fig 4C; Supplementary Table 13). Our independent association signals capture majority of the common variants underlying these five PUFA and MUFA traits.

Transcriptome-wide association analysis

To identify genes with expression associated with PUFA and MUFA traits, S-PrediXcan was used to integrate GTEx (v8) eQTL (expression quantitative trait loci) data from 49 tissues and UKB-EUR cohort GWAS summary statistics. Across fourteen traits in the discovery stage, 24,666 Bonferroni-corrected significant gene-trait associations comprised of 527 unique genes were identified (Supplementary Table 8). We then used S-MultiXcan to find joint effects of gene expression associations across tissues. We found 2,818 associations (601 unique genes), of which 392 unique genes have not been found in previous GWAS for PUFA and MUFA traits (Supplementary Figure 6; Supplementary Tables 9, 10).

Since there was a high degree of overlap between TWAS and GWAS results, we searched for novel gene-trait associations in S-MultiXcan that had not been found in our discovery or meta-analysis GWAS analyses. We found 55 genes, spanning 12 loci, that were identified exclusively in TWAS and are novel though unreplicated (Figure 4D). Many of these genes (44) are found in a cluster at 6p21. These 44 genes are significantly enriched for "immune system process" (GO: GO:0002376; 17 genes; FDR = 7.41E-04) and "regulation of immune system process" (GO:0002682; 17 genes; FDR = 2.09E-06). Three other novel genes, *F2* (MUFAs), *WDR81* (LA), and *PTK2* (PUFAs), are involved in "regulation of lipid kinase activity" (GO:0043550).

Gene set enrichment analysis

MAGMA tissue expression analysis for sets of positionally mapped genes from each of the five meta-analyzed PUFA and MUFA traits revealed that liver was exclusively the significantly enriched tissue type (Supplementary Figure 7). Because of this, we sent genes mapped from GTEx liver eQTLs and HiC liver chromatin data, in addition to positionally mapped genes, to GENE2FUNC for gene set enrichment (Supplementary Table 16).

Across the five traits, the most significant gene sets in the categories of Curated gene set, Positional gene set, Gene Ontology (GO): Biological process, GO: Cellular component, GO: Molecular function, Cancer modules, Canonical pathways, Computational gene sets, KEGG pathways, and TF targets, were all driven by genes in the major histocompatibility complex (MHC). However, the MHC region is the most polymorphic in the human genome, associated with the most disease traits, and determining causal variants in this region is highly prone to confounding [29], so we have excluded these from subsequent enrichment analyses. The next most significant positional gene set in all five PUFA and MUFA traits is chr6p22, corresponding to the *GCKR* locus; this locus has been identified in several previous GWAS (Supplementary Table 1). We found that significantly enriched gene sets in Wikipathways included statin pathway, histone modifications, pathways of LDL, HDL, and triglycerides, and metabolism of several nutrients, including zinc, copper, Vitamin B12, folate, and Vitamin A (Fig 5A). In the gene sets defined by GWAS catalog, the most significant enrichments are in genes that have previously been associated with blood lipids, including total cholesterol, LDL cholesterol, and triglycerides, as well as several traits related to mental characteristics such as "autism spectrum disorder or schizophrenia" (n-6, LA, and MUFAs only), "Bipolar disorder (I and II)" and "General factor of neuroticism" (Fig 5B).

To identify new relationships between novel genes associated with our meta-analyzed traits and previously reported traits in GWAS Catalog, we stratified GENE2FUNC analysis based on novelty (Figure 5C; Supplementary Table 17). The second most significant gene set enrichment of non-lipid GWAS catalog traits for n-3 and DHA (after "Handedness") was "Alcohol use disorder (total score)" (n-3: adjP_{novel} = 5.95E-10, adjP_{known} = 0.0052, adjP_{all} = 3.38E-09; DHA: adjP_{novel} = 1.57E-10, adjP_{known}=0.075, adjP_{all} = 1.73E-10; Figure 5D). Additionally, many alcoholism-adjacent traits were found to be significantly enriched for n-6, LA, and MUFAs, such as

"Triglyceride/LDL/HDL levels in current drinkers", "Response to alcohol consumption (flushing response", and "Alcohol consumption (max-drinks)." There was a total of 72 genes mapped to novel PUFA and MUFA variants that were significantly enriched for fifteen GWAS Catalog alcohol-related traits. These were located at 46 loci across 19 chromosomes, indicating the enrichment signal was not driven by a few gene clusters. Alcohol-related traits have been experimentally linked to PUFAs in multiple studies [30-34] (more in discussion). Therefore, our gene set enrichment analysis highlights a possible role of PUFAs and MUFAs in mental health, especially related to alcohol usage.

Discussion

Here we report the largest GWAS to-date of PUFA and MUFA phenotypes, performed in European ancestry cohorts (EUR; N_{EUR} = 124,024), consisting of fourteen traits in the discovery stage, five of which were replicated and meta-analyzed in external EUR cohorts. The discovery cohort, those designated EUR in UK Biobank (UKB-EUR), is the largest publicly available human dataset with measures of these traits in genotyped participants ($N_{UKB\text{-}EUR}$=101,729 after QC). We have identified 51 novel and externally replicated loci, as well as 36 loci that were not replicated, but have not been reported in previous GWAS (Supplementary Tables 1, 15). Considering that only 37 genomic risk loci were previously reported in relation to these traits, this study greatly increases our scope of understanding the genetic architecture of PUFAs and MUFAs. Of the 37 previously reported loci, we have replicated 23 loci in our discovery analysis (UKB-EUR) and 22 loci in our EUR meta-analysis (Supplementary Table 1). We have included plausible biological mechanistic explanations for many of our novel loci, and we have also added context to previously

identified GWAS loci, to provide the most comprehensive functional analysis of variants associated with PUFA and MUFA traits to-date.

In our follow-up analysis of genes mapped to loci from our meta-analysis, we found notable differences between novel and known gene enrichment *P*-values for the GWAS catalog trait "Alcohol use disorder (total score)" (AUD) (Supplementary Table 15, Figure 5D). Across the phenotypes n-3, n-6, DHA, and LA, there are two novel clusters of PUFA-gene associations that have previously been associated with AUD. These genes are *PLEKHM1*, *CRHR1*, *SPPL2C*, *MAPT*, *STH*, and *KANSL1*, *NSF*, and *WNT3* at chr17q21.31, mapped to n-3 and DHA; and *FUT2*, *MAMSTR*, *RASIP1*, and *IZUMO1* at chr19q13.33, mapped to n-6 and LA. The inversion at chr17q21.31 has recently been associated with alcohol intake in a GWAS of ~127,000 European participants from the Million Veterans Program cohort [30]. The association of the gene cluster at chr19q13.33 with AUD was reported as a novel association in an analysis of ~435,000 European participants of UKB [31].

In addition to the shared genetic variants between PUFAs and AUD, such as variants in *SNX17* and *GCKR*, variability in PUFA levels has been associated directly with AUD. The direction of causality between these traits has not been clearly disentangled. DHA has a neuroprotective effect against binge alcohol drinking, and is depleted with alcohol exposure [35]. In the opposite causal direction, high alcohol consumption was associated with lower fatty acid intake measured by 24 hour recall in the 2001-2002 National Health and Nutrition Examination Survey in 4,168 adults [32]. Deficiencies in n-3s are associated with bipolar disorder [36], which can lead to higher cravings for alcohol. Additionally, alcohol abuse has been characterized by an increase in oleic acid / LA ratio; Teubert et al. demonstrated a shift back to higher LA during alcohol

detoxification in a small study of 45 alcoholic patients [34]. Overall, the data on this topic are sparse, and more research should be done to elucidate this relationship.

Our study has several limitations. First, the PUFA and MUFA traits that we were able to investigate are limited to those reported by the UKB NMR metabolomics panel. We cannot resolve, for instance, differences in specific PUFAs that are often reported with higher resolution metabolite analyses, such as the difference in effects associated with DHA and other n-3s, notably eicosapentaenoic acid (EPA). Second, UKB is known to have volunteer bias, which can skew results, as has previously been shown [37].

Next, our analysis is mostly limited to determining the genetic associations of PUFA and MUFA traits in EUR populations. We recognize that an overwhelming number of genomic analyses to date have been conducted on EUR populations [38], to the detriment of understanding other ancestry groups. Further, our replication analysis shows that of 115 discovery loci in UKB-EUR, only 47 were replicated at $P < 0.05$ in one or more of the AFR, CSA, or EAS multi-ancestry groups. While power calculation shows that the difference was mainly driven by small sample sizes of non-EUR samples, it may also be that there is a distinct set of variants associated with PUFA and MUFA traits in non-EUR groups. We hope that the results of this analysis can be meta-analyzed with ancestrally diverse participant groups in future studies.

Another limitation is possibly introduced by the quantification of PUFA and MUFA traits using nuclear magnetic resonance spectroscopy (NMR). NMR has advantages and disadvantages as compared to the gold standard methods for quantitative fatty acid analysis, specifically high-resolution capillary gas chromatography coupled to flame ionization detection (FID) or mass spectrometry (GC-MS), or alternatively, liquid chromatography mass spectrometry (LC-MS) [39]. First, the speed and cost advantages of NMR over GC or MS are advantageous in biobank-scale

sample quantification [40]. NMR is also a non-destructive technique, meaning samples can be stored and re-measured in the future. However, NMR is of reduced sensitivity and selectivity compared to GC-based techniques. GC resolves all fatty acids at picogram levels, compared to NMR which operates at minimum on milligram scale [41]. GC-MS is also able to perform fatty acid analysis of high selectivity and completely resolve analytes [42]; NMR resolution is limited, and great care must be taken to ensure confounding overlapping signals are avoided, particularly in complex mixtures [41]. Nevertheless, as discussed above, our most significant results are congruent with biochemical expectations and with previous GWAS studies, including those studies which used MS-based quantification (Supplementary Table 1). This adds confidence to our usage of NMR-measured phenotypes and strengthens our novel findings.

Finally, as with any GWAS, significant associations are no more than candidates for mechanistic processes that, when altered, will have a reproducible influence on traits and ultimately human health. Replication of the associations and detailed investigation in experimental models and in randomized control trials are required to lead to clinical and precision nutrition applications. This study adds to a growing body of genomics literature that may help realize these applications in relation to PUFA and MUFA traits [43].

Methods

Ethics

Participant data use was approved by UK Biobank (UKB; Project ID 48818). UKB participants have consented to the use of their medical and genetic data in research studies. This research was performed on a University of Georgia (UGA) computing cluster with strict data protection protocols and two-factor authentication. The UGA Institutional Review Board (IRB)

approved the use of human subject data in this study. Additional datasets use publicly available summary statistics from previous GWAS, and approval was not required.

Participants

The full UKB consists of > 500,000 volunteer participants between ages 40 and 70 that were recruited between 2006 and 2010 in England, Scotland, and Wales. Approximately 120,000 participants had metabolic traits measured from plasma samples taken at recruitment using NMR between June 2019 and April 2020. Participants used in this study were removed on the following criteria: withdrawn consent, mismatches between self-reported and genetic sex, poor quality genotyping as flagged by UKB, sex chromosome aneuploidy, or poor-quality NMR measurement flagged by UKB. After quality control (QC) and stratification by ancestry using Pan UKBB designations [16], counts of UKB participants included in our analyses were: 101,729 European (EUR); 1,564 African (AFR); 2,203 Central South Asian (CSA); and 633 East Asian (EAS). UKB participant characteristics can be found in Supplementary Table 2.

For replication and meta-analysis, we included two external EUR studies. First, the Finnish Metabolic Sequencing (FinMetSeq) study [11], which consists of a combination of FINRISK and METSIM cohorts. METSIM participants were 10,197 men from Kuopio, Eastern Finland, aged 45 to 73 years during initial examinations from 2005 to 2010. FINRISK participants were recruited every five years from 1972 to 2012, and consisted of random population samples of men and women aged 30-59 years. FinMetSeq used 10,192 participants from 1992-2007 FINRISK surveys who had a residence in northeastern Finland. Pregnant women, type 1 and 2 diabetics, and those fasting less than eight hours were excluded from this cohort. Of the approximately 19,000 participants in FinMetSeq, 8,751 had NMR metabolomics data available for the fatty acid phenotypes of interest and were used in this study.

We also utilized a meta-analysis consisting of fourteen genotyped datasets derived from ten EUR studies performed by Kettunnen et al. in 2016 [14] in our replication and meta-analysis. The number of participants contributed by the Kettunnen et al. summary statistics in this analysis ranges from 3,644-13,544 participants (from six to ten studies), depending on the variant; this participant range is consistent across the five traits meta-analyzed. There is an overlap of 225 participants between the FinMetSeq and Kettunnen et al. cohorts, which we determined would not affect type I error in a meaningful way [44].

Fatty acid phenotypes

UKB EDTA (ethylenediaminetetraacetic acid) plasma samples were taken at the baseline recruitment timepoint and measured between June 2019 and April 2020 by the metabolic biomarker profiling platform of Nightingale Health Ltd., as described previously [40]. We analyzed fourteen quantitative polyunsaturated fatty acid (PUFA) and monounsaturated fatty acid (MUFA) phenotypes in UKB cohorts, specifically, n-3 LCPUFAs, n-6 LCPUFAs, docosahexaenoic acid (DHA), linoleic acid (LA), total PUFAs, and total monounsaturated fatty acids (MUFAs), all reported in mmol/L, the percentages of n-3 LCPUFAs, n-6 LCPUFAs, DHA, LA, MUFAs, and PUFAs out of the total amount of fatty acids (designated "pct"), as well as n-6/n-3 ratio, and PUFA/MUFA ratio. We regressed all PUFA and MUFA traits on selected covariates in each model (described below) and applied rank-based inverse normal transformation (indirect INT) to the residuals for use in all analyses [45]; this was consistent with transformations performed in our external GWAS replication studies.

Genotype data

Genotype data was initially QCed and imputed with Haplotype Reference Consortium (HRC) and 1000 Genomes variants by UKB (v3) as previously described [46]. For discovery

analyses, we excluded variants with imputation quality (INFO) score < 0.3, minor allele frequency (MAF) < 0.1%, missing genotype per individual > 5%, missing genotype per variant > 5%, or Hardy-Weinberg equilibrium (HWE) $P < 1\times10^{-8}$. After quality control, a total of 15,587,898 variants among 101,729 participants were included in the UKB-EUR discovery cohort. QC and genotype file format conversions were performed using PLINK2 alpha-v2.3 [47,48]. All genomic positions in this study refer to autosomal chromosomes in the Genome Reference Consortium Human Build 37 (GRCh37), also known as hg19.

Generating a pruned variant set

A pruned set of variants were inputted as PLINK-format genotypes to BOLT-LMM for model-fitting in the UKB-EUR discovery GWAS [49]. After filtering for only participants included in this analysis, the exclusion criteria for variants in the pruned set were INFO score < 0.8, MAF < 1%, missing genotype per variant > 1%, or HWE $P < 1\times10^{-8}$. A hard-call threshold of 0.1 was applied to the filtered variants. The lactase locus on chromosome 2, the major histocompatibility complex (MHC) on chromosome 6, and inversions on chromosomes 8 and 17 were excluded. Linkage disequilibrium (LD) pruning was performed at $r^2 = 0.2$, (plink2 --indep-pairwise 50 5 0.2). After pruning, 821,405 variants remained.

Discovery model selection

Two sensitivity models were used in our initial discovery stage analysis. Model 1 (M1) included the covariates sex, age, age^2, genotyping array, and assessment center. Model 2 (M2) included the M1 covariates, plus body mass index (BMI), lipid medication usage, and socioeconomic status as measured by Townsend deprivation index (Supplementary Table 3). The first twenty principal components for study participants as calculated by PLINK2 (randomized

algorithm) [47,48] were also included as covariates in both models. We compared our summary statistics from M1 and M2 using Spearman's rank correlation (two-sided test).

Identification of significant GWAS signals

BOLT-LMM v2.3.6 [49] was used to perform linear mixed-effects model association analyses on fourteen PUFA and MUFA traits in the UKB-EUR discovery stage analysis. The provided 1000G European LD scores [50] were used to calibrate the BOLT-LMM statistic. Covariates and pruned variant sets were included in models as described above. Non-infinitesimal BOLT-LMM *P*-values ("P_BOLT_LMM") were used in all reporting and downstream analyses. Because the fourteen PUFA and MUFA traits were highly related, we calculated the effective number of traits to use for Bonferroni multiple testing correction. Eigenvalues (λ) for the fourteen traits were used to calculate the number of effective traits as: $\left(\sum_{k=1}^{14} \lambda_k\right)^2 \Big/ \sum_{k=1}^{14} \lambda_k^2 = 2.98$ [51]. The threshold of $P < (5 \times 10^{-8} / 2.98) = 1.678 \times 10^{-8}$ was used to designate significant variant-trait associations in the discovery stage.

Transcriptome-wide association analysis

Transcriptome-wide association analysis (TWAS) was performed on UKB-EUR discovery stage summary statistics using S-PrediXcan [52]. This method used Genotype-Tissue Expression (GTEx) v8 [53] expression quantitative trait loci (eQTL) data across 49 available tissues. Summary statistics were harmonized and imputed to GTEx models. A total of 601,176 gene-tissue pairs were analyzed across fourteen PUFA and MUFA traits, using a Bonferroni corrected significance threshold of $P < (0.05 / (601{,}176 \times 2.98)) = 2.791 \times 10^{-8}$. S-MultiXcan was used to integrate tissue-level associations and increase association detection. The cutoff condition number of eigenvalues was set to 30 for truncating singular-value decomposition components. S-MultiXcan was run

across 21,846 genes, using a Bonferroni corrected significance threshold of $P < (0.05 / (21{,}846 \times 2.98)) = 7.68 \times 10^{-7}$.

Replication and meta-analysis

Variant-trait associations from the discovery stage for five traits were sent to replication and meta-analysis steps: omega-3 fatty acids (n-3 LCPUFAs), omega-6 fatty acids (n-6 LCPUFAs), docosahexaenoic acid (DHA), linoleic acid (LA), and total monounsaturated fatty acids (MUFAs). Summary statistics were obtained from the publicly available Finnish Metabolic Sequencing (FinMetSeq) study [11] and the meta-analysis of 10 studies by Kettunen et al. [14]. FinMetSeq summary statistics as provided had been adjusted by age, age^2, sex, cohort year, BMI, sex hormones, and lipid medications. Kettunnen et al. had adjusted for sex, age and ten genetic principal components in their analyses. Discovery stage variants were considered replicated in the external cohorts on a per-variant basis at $P < 0.05$.

Meta-analyses were performed using the METAL [54] software using the STDERR scheme, which weights effect size estimates using the inverse of the corresponding standard errors. The meta-analysis of this study consists of the three EUR participant studies: UKB-EUR+ FinMetSeq+ Kettunen et al. datasets (N=114,124 to 124,024). MungeSumstats [55] was used in pre-processing to harmonize effect alleles from separate cohorts to the reference genome. UKB-EUR was used to estimate the number of effective traits for the three EUR meta-analysis cohorts. Number of effective traits was calculated from phenotype eigenvalues as 2.05 using the formula shown above, and the multiple-testing corrected threshold of $P < (5 \times 10^{-8} / 2.05) = 2.439 \times 10^{-8}$ was used to designate significant meta-analysis variant-trait associations.

Heritability and LD score regression

Restricted maximum likelihood (REML) estimates for genetic correlation and multi-trait heritability (h^2) were calculated using BOLT-REML [49]. The six phenotypes measured in absolute concentration units (mmol/L) from the discovery UKB-EUR cohort (n-3, n-6, DHA, LA, PUFAs, MUFAs) were inputted with the covariates age, age^2, sex, assessment center, genotype batch, and the first twenty PCs. The provided 1000G European LD scores [50] were used to calibrate the BOLT statistic. The refinement step was skipped (--remlNoRefine) to increase computational efficiency.

LD score regression (LDSC) [56] was used to calculate LD Score regression intercept, genomic control (λ), and non-partitioned SNP-based h^2. The 1000 Genomes European set was used as the LD reference panel [50]. MungeSumstats [55] was used to harmonize alleles and convert summary statistics to LDSC format for this and subsequent steps. Variance explained was calculated by the formula ($2 \times MAF \times (1 - MAF) \times \beta^2$). We calculated variance explained for variants identified as independent significant associations, as well as for lead variants of genomic risk loci. Pairwise genetic correlations (r_g) were computed from munged summary statistics using LDSC [57]. Pairwise phenotypic correlations (r_p) were calculated as Pearson's correlation coefficient (two-sided). *P*-values for r_g and r_p were adjusted for false discovery rate (FDR).

Identifying genetic loci

Lead variants for each independent genomic risk loci were defined in both the discovery (UKB-EUR) and meta-analysis cohorts (UKB-EUR + Kettunen et al. + FinMetSeq) by inputting summary statistics to the Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) web server [58]. The UKB release2b 10k European set was used as the LD reference panel. The maximum *P*-value cutoff was set to 0.05, and a first LD threshold of $r^2 \geq 0.6$ and second threshold of $r^2 \geq 0.1$ were used to define loci and lead SNPs. SNPs not available in the GWAS

input but contained in the reference panel were included in output. The maximum distance between LD blocks to merge into a locus was 250 Kb. Summary statistics *P*-values were set with a lower cap of P = 1e-300 to resolve FUMA processing errors that we identified related to minimum Python float size limit. Variants from meta-analysis were annotated to genes with SNP2GENE using positional mapping (maximum distance 10 Kb), eQTL mapping from GTEx v8 liver tissue, and chromatin interaction mapping using built-in data from Hi-C (GSE87112) liver tissue. All other FUMA mapping settings were kept as default.

Identifying novel loci

A table of previously reported ("known") PUFA- and MUFA- associated lead variants was prepared from full summary statistics (where available) or from significance tables found within previous GWAS publications (Supplementary Table 1). All reported genomic coordinates were set to hg19 using liftOver [59]. Genomic risk loci coordinates were identified in each study by *P*-values of reported variants using FUMA (setting LD reference by ancestry of study, otherwise default settings). These loci were grouped together within a ± 500 Kb window prior to checking for novelty of our results, regardless of the ancestry of the study cohorts. Trait names were harmonized across studies. LDtrait [60] was used to cross-check our novelty table; no additional loci were found using this method.

Multi-ancestry replication in non-EUR UKB Cohorts

Genome-wide Complex Trait Analysis mixed linear model-based association analysis (GCTA-MLMA) [61] was used to perform mixed-model GWAS analyses in the UKB African (AFR), Central/South Asian (CSA), and East Asian (EAS) cohorts (Supplementary Tables 2, 12). A genetic relatedness matrix (GRM) was generated for each population using GCTA-GRM [62]. Covariates used in these models were age, age^2, sex, and the first ten principal components.

Genotype quality filtering parameters were INFO < 0.3, MAF < 1 %, missing genotype per individual > 5 %, missing genotype per variant > 5%, or HWE P < 1×10^{-8}.

Gene sets and pathway analysis

FUMA GENE2FUNC [58] was performed on genes mapped from SNP2GENE using parameters described above including all background gene-sets in hypergeometric tests, and using expression data from all GTEx v8 datasets. Benjamini-Hochberg (FDR) was used as the gene set enrichment multiple test correction method. Gene Ontology (GO) was used to categorize sets of genes in downstream analyses [63].

Conditional and joint association analysis

Genome-wide Complex Trait Analysis Conditional and Joint Association Analysis (GCTA-COJO) with stepwise model selection to identify conditionally independent variants was performed using discovery and meta-analysis summary statistics (--cojo-slct) [64]. A random set of 20,000 unrelated UKB-EUR participants were used as the LD reference (--bfile input). Variants with MAF < 1% were removed. COJO was run per chromosome with significance thresholds based on effective trait Bonferroni corrections (described above), using default settings for collinearity and window size. Summary statistics standard error (SE) values were re-calculated with higher precision based on effect size and *P*-values prior to input to COJO, to ensure GCTA-COJO output columns matched input.

Visualizing results

CMplot [65] was used to generate the circular Manhattan plot in Figure 2. The regional Manhattan plots in Figure 3 were generated using karyoploteR [66]. The correlation plot in Figure 4A was generated using ggcorrplot2 [67]. The qqman R package [68] was used to generate Manhattan and QQ plots in Supplementary Figures 3 and 4. S-MultiXcan gene-based Manhattan plots in

Supplementary Figure 6 were generated using the Manhattan R package [69]. Bar plots and scatterplots were generated using ggplot2 [70] in R v4.1.0. Color palettes in all figures were optimized for accessibility with three major types of color blindness (deuteranopia, protanopia, and tritanopia) using https://color.adobe.com/create/color-accessibility.

Data availability

Full summary statistics can be found on GWAS Catalog, using the accession codes provided in Supplementary Table 18.

Code availability

Script repository for this analysis can be found at https://github.com/michaelofrancis/PUFA_GWAS.

References

1 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, doi:10.1038/s41467-021-22370-2 (2021).

2 Lin, P.-Y., Chiu, C.-C., Huang, S.-Y. & Su, K.-P. A meta-analytic review of polyunsaturated fatty acid compositions in dementia. *J Clin Psychiat* **73**, 0-0 (2012).

3 Grosso, G. *et al.* Dietary n-3 PUFA, fish consumption and depression: A systematic review and meta-analysis of observational studies. *J Affect Disorders* **205**, 269-281, doi:https://doi.org/10.1016/j.jad.2016.08.011 (2016).

4 Innes, J. K. & Calder, P. C. Marine Omega-3 (N-3) Fatty Acids for Cardiovascular Health: An Update for 2020. *Int J Mol Sci* **21**, 1362, doi:10.3390/ijms21041362 (2020).

5 Lorente-Cebrián, S. *et al.* Role of omega-3 fatty acids in obesity, metabolic syndrome, and cardiovascular diseases: a review of the evidence. *J Physiol Biochem* **69**, 633-651 (2013).

6 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 Gene* **7**, 321-331 (2014).

7 Brenna, J. T. & Kothapalli, K. S. D. New understandings of the pathway of long-chain polyunsaturated fatty acid biosynthesis. *Curr Opin Clin Nutr Metab Care* **25**, 60-66, doi:10.1097/mco.0000000000000810 (2022).

8 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**, doi:10.1016/j.plefa.2020.102183 (2020).

9 Sinclair, A. J. High Linoleic Acid in the Food Supply Worldwide-What are the Consequences? *Sci Technol of Cereals, Oils and Foods (粮油食品科技)* **30** (2022).

10 Hodson, L., Skeaff, C. M. & Fielding, B. A. Fatty acid composition of adipose tissue and blood in humans and its use as a biomarker of dietary intake. *Prog Lipid Res* **47**, 348-380 (2008).

11 Locke, A. E. *et al.* Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature* **572**, 323-328, doi:10.1038/s41586-019-1457-z (2019).

12 Shin, S. Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat Genet* **46**, 543-550, doi:10.1038/ng.2982 (2014).

13 Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat Genet* **44**, 269-276, doi:10.1038/ng.1073 (2012).

14 Kettunen, J. *et al.* Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun* **7**, 11122, doi:10.1038/ncomms11122 (2016).

15 Würtz, P. *et al.* Quantitative serum nuclear magnetic resonance metabolomics in large-scale epidemiology: a primer on-omic technologies. *Am J Epidemiol* **186**, 1084-1096 (2017).

16 Pan-UKB team. https://pan.ukbb.broadinstitute.org (2020).

17 Mefford, J. & Witte, J. S. The Covariate's Dilemma. *PLoS Genet* **8**, e1003096, doi:10.1371/journal.pgen.1003096 (2012).

18 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, doi:10.1371/journal.pgen.1002193 (2011).

19 Hu, Y. *et al.* Genome-wide meta-analyses identify novel loci associated with n-3 and n-6 polyunsaturated fatty acid levels in Chinese and European-ancestry populations. *Hum Mol Genet* **25**, 1215-1224, doi:10.1093/hmg/ddw002 (2016).

20 Wu, J. H. *et al.* Genome-wide association study identifies novel loci associated with concentrations of four plasma phospholipid fatty acids in the de novo lipogenesis pathway: results from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium. *Circ Cardiovasc Genet* **6**, 171-183, doi:10.1161/circgenetics.112.964619 (2013).

21 Sánchez Muñoz-Torrero, J. F. *et al.* Multivariate analysis for coronary heart disease in heterozygote familial hypercholesterolemia patients. *Per Med* **15**, 87-92, doi:10.2217/pme-2017-0075 (2018).

22 Perreault, M. *et al.* The Human UDP-glucuronosyltransferase UGT2A1 and UGT2A2 enzymes are highly active in bile acid glucuronidation. *Drug Metab Dispos* **41**, 1616-1620, doi:10.1124/dmd.113.052613 (2013).

23 Arnold, C., Konkel, A., Fischer, R. & Schunck, W. H. Cytochrome P450-dependent metabolism of omega-6 and omega-3 long-chain polyunsaturated fatty acids. *Pharmacol Rep* **62**, 536-547, doi:10.1016/s1734-1140(10)70311-x (2010).

24 Barden, A. E., Mas, E. & Mori, T. A. n-3 Fatty acid supplementation and proresolving mediators of inflammation. *Curr Opin Lipidol* **27**, 26-32 (2016).

25 Dawson, K. *et al.* Modulation of blood cell gene expression by DHA supplementation in hypertriglyceridemic men. *J Nutr Biochem* **23**, 616-621 (2012).

26 Wang, Z. *et al.* Fatty acid desaturase 2 (FADS2) but not FADS1 desaturates branched chain and odd chain saturated fatty acids. *Biochim Biophys Acta Mol Cell Biol Lipids* **1865**, 158572, doi:10.1016/j.bbalip.2019.158572 (2020).

27 Fernandez, R. F. *et al.* Acyl-CoA synthetase 6 is required for brain docosahexaenoic acid retention and neuroprotection during aging. *JCI Insight* **6**, doi:10.1172/jci.insight.144351 (2021).

28 Marszalek, J. R., Kitidis, C., DiRusso, C. C. & Lodish, H. F. Long-chain acyl-CoA synthetase 6 preferentially promotes DHA metabolism. *J Biol Chem* **280**, 10817-10826 (2005).

29 Trowsdale, J. & Knight, J. C. Major histocompatibility complex genomics and human disease. *Annu Rev Genom Hum G* **14**, 301-323, doi:10.1146/annurev-genom-091212-153455 (2013).

30 Gelernter, J. *et al.* Genome-wide Association Study of Maximum Habitual Alcohol Intake in >140,000 U.S. European and African American Veterans Yields Novel Risk Loci. *Biol Psychiat* **86**, 365-376, doi:https://doi.org/10.1016/j.biopsych.2019.03.984 (2019).

31 Zhou, H. *et al.* Genome-wide meta-analysis of problematic alcohol use in 435,563 individuals yields insights into biology and relationships with other traits. *Nat Neurosci* **23**, 809-818, doi:10.1038/s41593-020-0643-5 (2020).

32 Kim, S. Y., Breslow, R. A., Ahn, J. & Salem Jr, N. Alcohol consumption and fatty acid intakes in the 2001–2002 National Health and Nutrition Examination Survey. *Alcohol Clin Exp Res* **31**, 1407-1414 (2007).

33 Fogaça, M. N., Santos-Galduróz, R. F., Eserian, J. K. & Galduróz, J. C. F. The effects of polyunsaturated fatty acids in alcohol dependence treatment - a double-blind, placebo-controlled pilot study. *BMC Clin Pharmacol* **11**, 10, doi:10.1186/1472-6904-11-10 (2011).

34 Teubert, A., Thome, J., Büttner, A., Richter, J. & Irmisch, G. Elevated oleic acid serum concentrations in patients suffering from alcohol dependence. *J Mol Psychiat* **1**, 13, doi:10.1186/2049-9256-1-13 (2013).

35 Collins, M. Alcohol abuse and docosahexaenoic acid: Effects on cerebral circulation and neurosurvival. *Brain Circ* **1**, 63-68, doi:10.4103/2394-8108.162533 (2015).

36 Balanza-Martinez, V. *et al.* Therapeutic use of omega-3 fatty acids in bipolar disorder. *Expert Rev Neurother* **11**, 1029-1047 (2011).

37 Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M. & Davey Smith, G. Collider scope: when selection bias can substantially influence observed associations. *Int J Epidemiol* **47**, 226-235, doi:10.1093/ije/dyx206 (2018).

38 Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161-164, doi:10.1038/538161a (2016).

39 Brenna, J. T., Plourde, M., Stark, K. D., Jones, P. J. & Lin, Y. H. Best practices for the design, laboratory analysis, and reporting of trials involving fatty acids. *Am J Clin Nutr* **108**, 211-227, doi:10.1093/ajcn/nqy089 (2018).

40 Soininen, P., Kangas, A. J., Würtz, P., Suna, T. & Ala-Korpela, M. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ-Cardiovasc Gene* **8**, 192-206 (2015).

41 Emwas, A.-H. *et al.* NMR Spectroscopy for Metabolomics Research. *Metabolites* **9**, doi:10.3390/metabo9070123 (2019).

42 Wang, D. H., Wang, Z., Chen, R. & Brenna, J. T. Characterization and Semiquantitative Analysis of Novel Ultratrace C10–24 Monounsaturated Fatty Acid in Bovine Milkfat by Solvent-Mediated Covalent Adduct Chemical Ionization (CACI) MS/MS. *J Agric Food Chem* **68**, 7482-7489 (2020).

43 Francis, M. *et al.* Genome-wide association study of fish oil supplementation on lipid traits in 81,246 individuals reveals new gene-diet interaction loci. *PLoS Genet* **17**, e1009431, doi:10.1371/journal.pgen.1009431 (2021).

44 Lin, D.-Y. & Sullivan, P. F. Meta-analysis of genome-wide association studies with overlapping subjects. *Am J Hum Genet* **85**, 862-872, doi:10.1016/j.ajhg.2009.11.001 (2009).

45 McCaw, Z. R., Lane, J. M., Saxena, R., Redline, S. & Lin, X. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics* **76**, 1262-1272, doi:https://doi.org/10.1111/biom.13214 (2020).

46 Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209, doi:10.1038/s41586-018-0579-z (2018).

47 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7, doi:10.1186/s13742-015-0047-8 (2015).

48 Purcell, S. & Chang, C. PLINK v. 2.3 alpha. www.cog-genomics.org/plink/2.0/ (2020).

49 Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**, 284-290, doi:10.1038/ng.3190 (2015).

50 Consortium, G. P. A global reference for human genetic variation. *Nature* **526**, 68 (2015).

51 Bretherton, C. S., Widmann, M., Dymnikov, V. P., Wallace, J. M. & Bladé, I. The effective number of spatial degrees of freedom of a time-varying field. *J Clim* **12**, 1990-2009 (1999).

52 Barbeira, A. N. *et al.* Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet* **15**, e1007889, doi:10.1371/journal.pgen.1007889 (2019).

53 Consortium, G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330, doi:10.1126/science.aaz1776 (2020).

54 Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-2191, doi:10.1093/bioinformatics/btq340 (2010).

55 Murphy, A. E., Schilder, B. M. & Skene, N. G. MungeSumstats: a Bioconductor package for the standardization and quality control of many GWAS summary statistics. *Bioinformatics* **37**, 4593-4596, doi:10.1093/bioinformatics/btab665 (2021).

56 Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-295, doi:10.1038/ng.3211 (2015).

57 Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nature Genetics* **47**, 1236-1241, doi:10.1038/ng.3406 (2015).

58 Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1-11 (2017).

59 Hinrichs, A. S. *et al.* The UCSC genome browser database: update 2006. *Nucleic Acids Res* **34**, D590-D598 (2006).

60 Myers, T. A., Chanock, S. J. & Machiela, M. J. LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations. *Front Genet* **11**, 157, doi:10.3389/fgene.2020.00157 (2020).

61 Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genet* **46**, 100-106 (2014).

62 Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genet* **42**, 565-569 (2010).

63 The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* **49**, D325-d334, doi:10.1093/nar/gkaa1113 (2021).

64 Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369-375, s361-363, doi:10.1038/ng.2213 (2012).

65 Yin, L. *et al.* rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated tool for Genome-Wide Association Study. *Genomics Proteomics Bioinf*, doi:https://doi.org/10.1016/j.gpb.2020.10.007 (2021).

66 Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088-3090, doi:10.1093/bioinformatics/btx346 (2017).

67 ggcorrplot2: Visualize a Correlation Matrix using ggplot2 (2022).

68 Turner, S. D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv*, 005165, doi:10.1101/005165 (2014).

69 Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nature Genetics* **51**, 768-769, doi:10.1038/s41588-019-0404-0 (2019).

70 Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer, 2016).

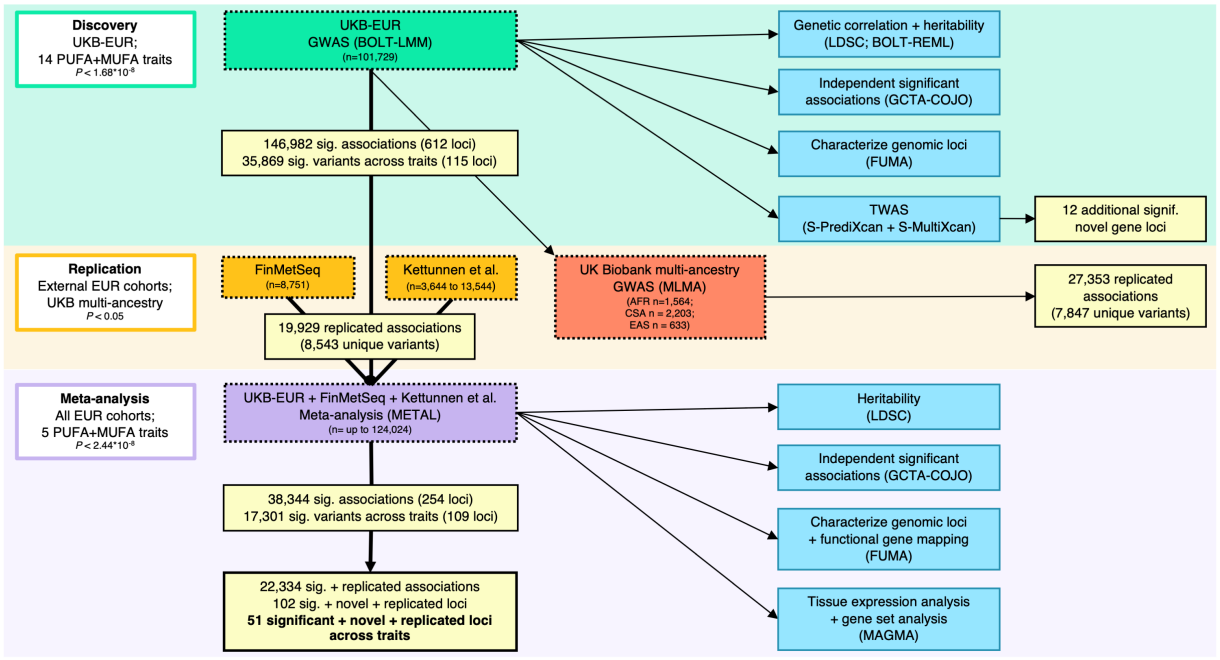

Figure 3.1. Overview of analyses.

Top left moving down: fourteen polyunsaturated fatty acid (PUFA) and monounsaturated (MUFA) traits were analyzed in the UK Biobank European discovery cohort (UKB-EUR). Significant associations were sent to replication in external European cohorts FinMetSeq and Kettunnen et al. for five available PUFA and MUFA traits. These three EUR studies were meta-analyzed, and 22,334 significant and replicated associations were identified across the five traits. Across these meta-analysis results we identified 51 unique, novel, and significant replicated loci. Middle: additional replication was also performed in UKB multi-ancestry cohorts. Right: additional software analyses are shown in blue. FUMA: Functional Mapping and Annotation of Genome-Wide Association Studies; COJO: Genome-wide Complex Trait Analysis Conditional and Joint Association Analysis; LDSC: Linkage Disequilibrium Score Regression; TWAS: transcriptome-wide association analysis.

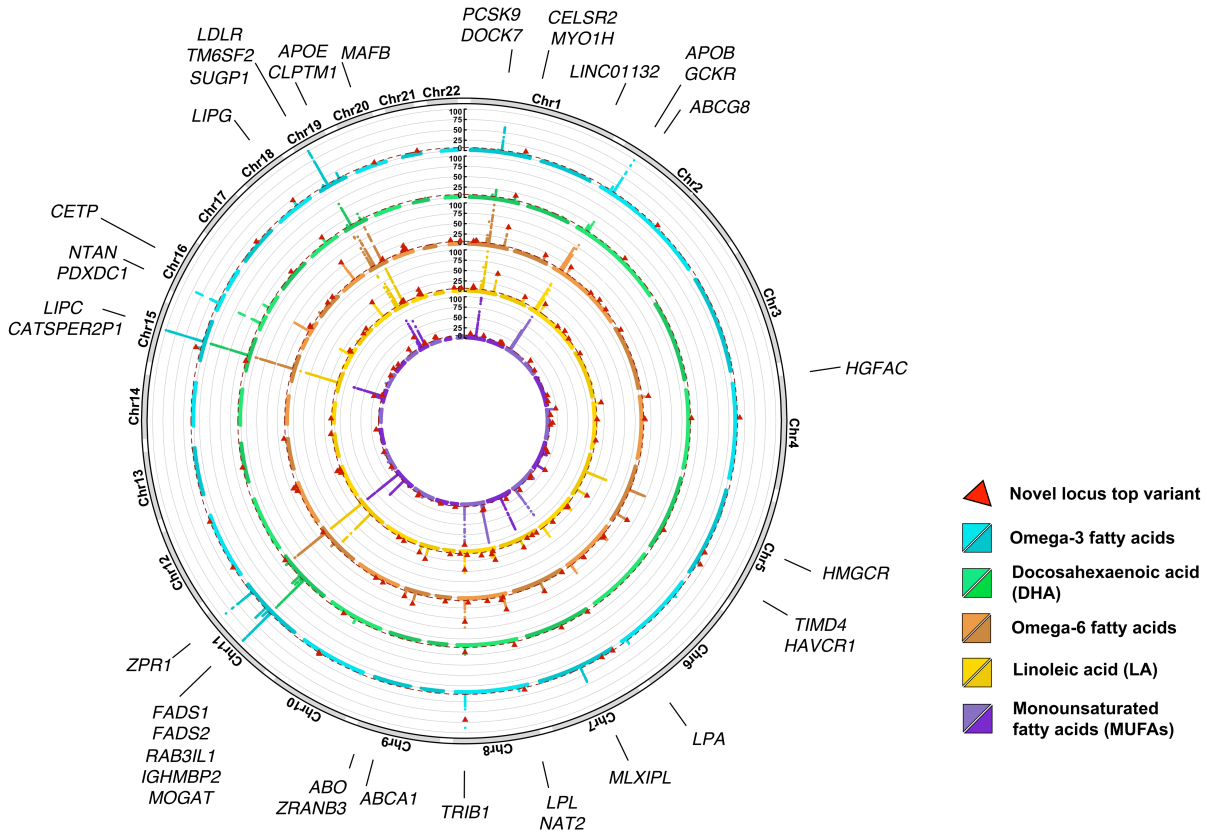

Figure 3.2. Circular Manhattan plot of five meta-analyzed PUFA and MUFA traits.

Plots show the $-\log_{10}P$ of meta-analyzed GWAS for polyunsaturated fatty acid (PUFA) and monounsaturated (MUFA) traits. Red triangles designate the lead variant of a novel locus associated with a trait in our analysis. Red dotted lines at $P < 2.439\times10^{-8}$ indicate the genome-wide significance threshold corrected for effective number of traits. Alternating color shades within each ring designate breaks between chromosomes. Genes corresponding to loci with $P <$ 1e-20 are labeled. All *P*-values were constrained to an upper limit of 1e-100 for visualization. Rings from outer to inner: omega-3 fatty acids, docosahexaenoic acid, omega-6 fatty acids, linoleic acid, and monounsaturated fatty acids.

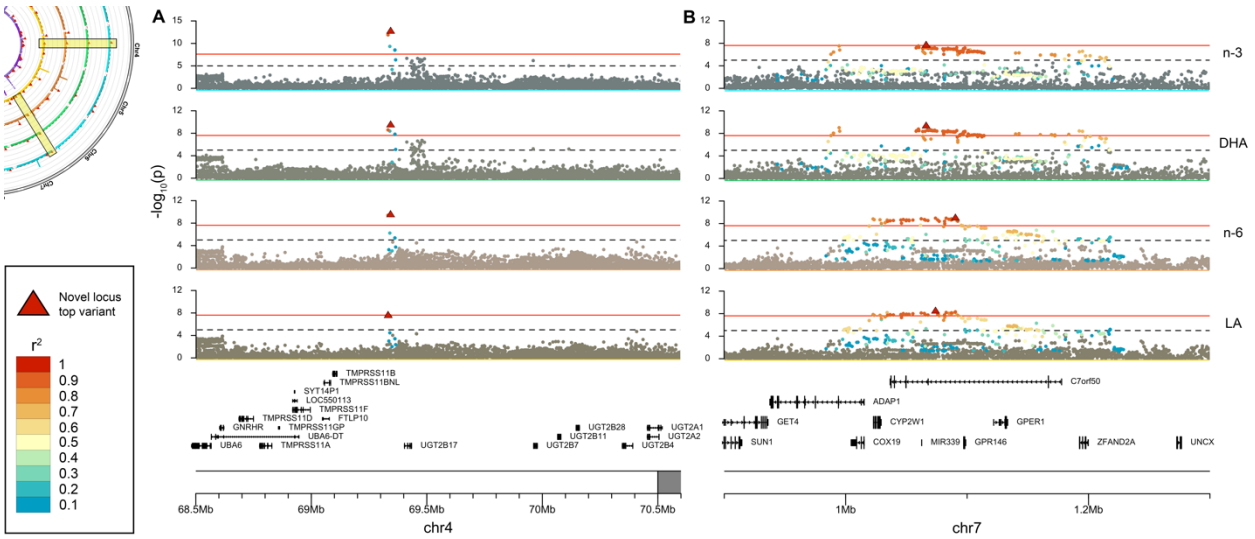

Figure 3.3. Regional Manhattan plots for selected novel loci.

Local association plots of significant loci at (**A**) chr4q13 and (**B**) chr7p22. These loci have novel, replicated associations with all four meta-analyzed polyunsaturated fatty acid (PUFA) traits (from top to bottom: omega-3 fatty acids, docosahexaenoic acid, omega-6 fatty acids, linoleic acid). Genes in each region are shown below the Manhattan plots. Red triangles designate the lead variant of each locus-trait association. Variants in linkage disequilibrium with the lead variant are color-coded according to their r^2 values.

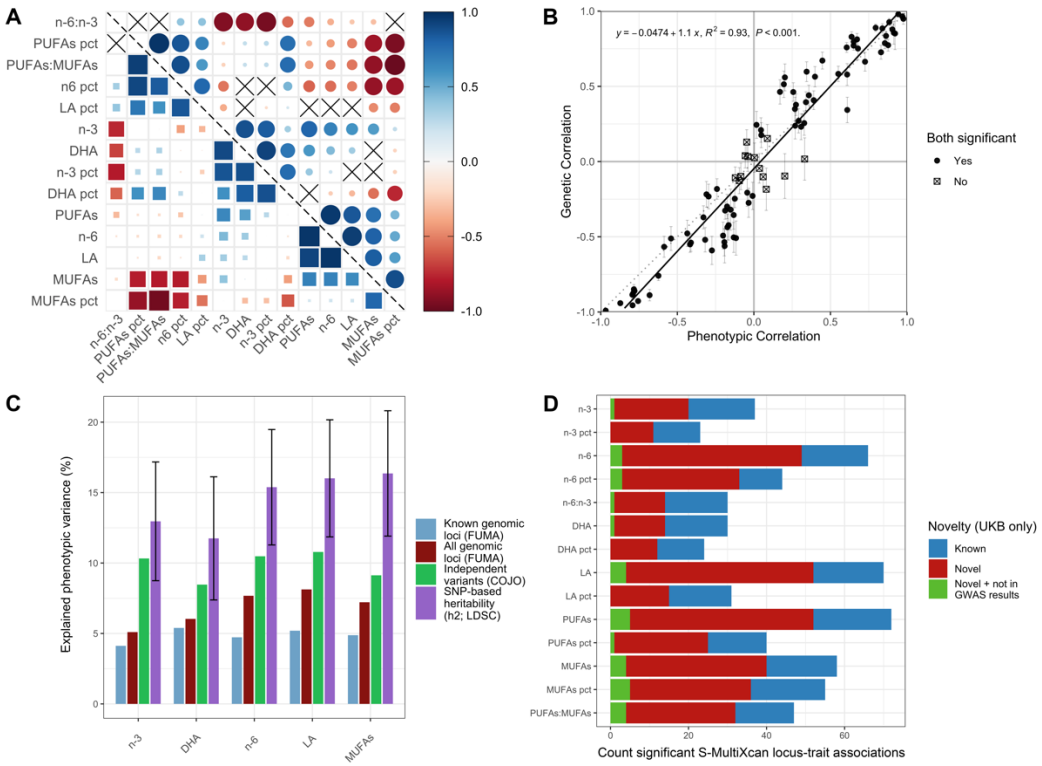

Figure 3.4. Results from genetic and phenotypic correlations, heritability, and S-MultiXcan.

(A) Genetic and phenotypic correlations (r_g and r_p) across all UK Biobank discovery cohort (UKB-EUR) polyunsaturated fatty acid (PUFA) and monounsaturated fatty acid (MUFA) traits. Above diagonal (circles) are genetic correlations; below diagonal (squares) are phenotypic correlations. Color and shape size both correspond to direction and strength of correlation. An "X" signifies a non-significant adjusted *P*-value for correlation coefficient. **(B)** Plot of genetic correlation vs. phenotypic correlation coefficients for 91 trait pairs. Error bars designate s.e.m. of correlation coefficient. Points with "X" did not reach significance for phenotypic correlation, genetic correlation, or both. **(C)** Explained phenotypic variance by different variant-grouping methods. SNP-based heritability shown with 95% confidence intervals. **(D)** Counts of significant locus-trait associations identified by S-MultiXcan for each trait in UKB-EUR, colored by novelty status.

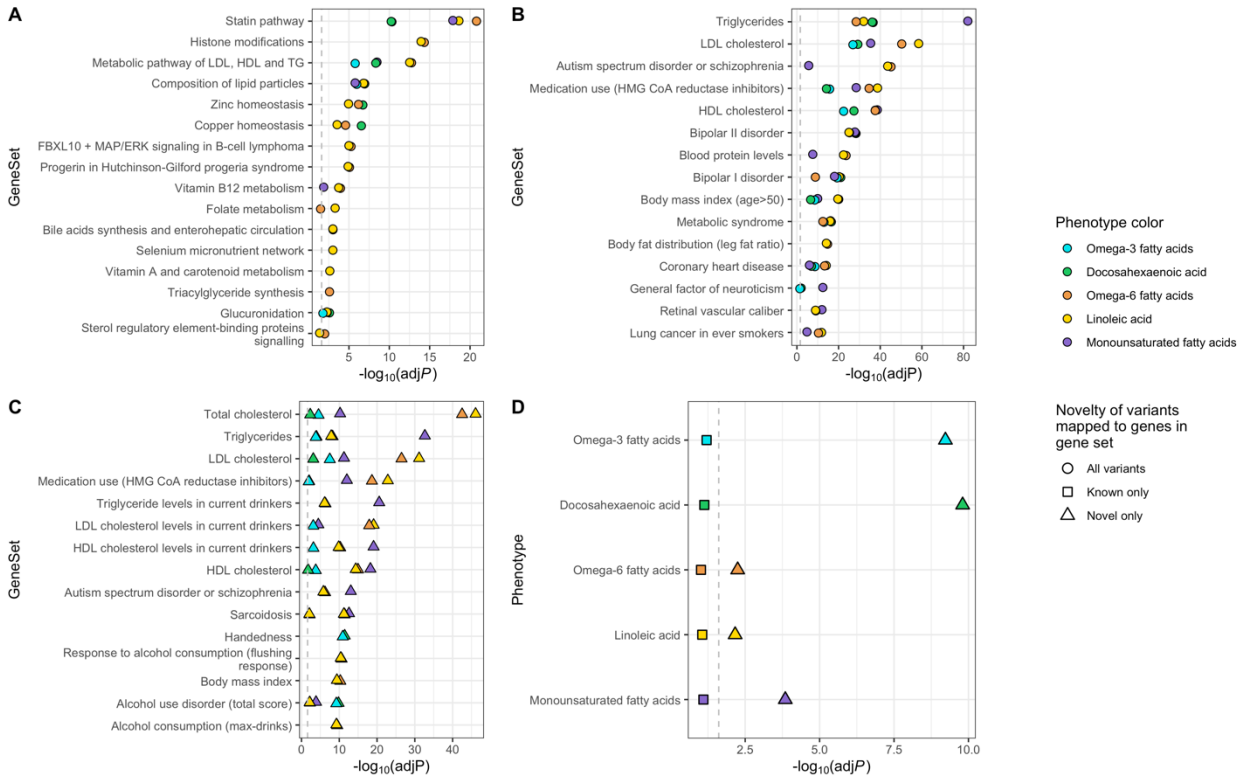

Figure 3.5. Gene sets mapped to significant meta-analysis loci.

Enrichment $-\log_{10}(\text{adj}P)$ for gene sets mapped to significant variants for five meta-analyzed polyunsaturated fatty acid (PUFA) and monounsaturated fatty acid (MUFA) traits. FDR adjusted *P*-values shown by trait for: **(A)** Wikipathways **(B)** GWAS Catalog **(C)** GWAS Catalog (novel variants only) **(D)** Variants mapped to the GWAS catalog trait "alcohol use disorder (total score)," stratified by novelty.

Supplementary Figure and Table legends

Supplementary figures and tables can be downloaded from:

https://www.medrxiv.org/content/10.1101/2022.05.27.22275343v2.supplementary-material

Supplementary Figures

S1. Participant characteristics for PUFA and MUFA traits of four ancestries in UK Biobank. (**A**) Mean and s.d. of polyunsaturated fatty acid (PUFA) and monounsaturated fatty acid (MUFA) traits that were measured in absolute concentration units (mmol/L). (**B**) Left: mean percentage and s.d. of each trait measured in percentage of total amount of fatty acids. Right: Mean of ratio-based traits and s.d. of these values.

S2. Correlation plot comparing *P*-values of our Models 2 versus Model 1. GWAS $-\log_{10}(P)$ between two sensitivity models were compared. Each plot is one of fourteen traits analyzed in the UK Biobank European discovery stage. Each point represents one variant. Spearman's Rho (R) and correlation *P*-value shown.

S3. Manhattan and QQ plots of UK Biobank discovery (EUR) dataset. Left: Manhattan plots showing the $-\log_{10}(P)$ of associations in each of fourteen polyunsaturated fatty acid (PUFA) and monounsaturated fatty acid (MUFA) traits in the UK Biobank European discovery cohort. Alternating point color shades indicate associations across 22 chromosomes. Red line at $P=1.678\times10^{-8}$ indicates genome-wide significance threshold corrected for number of effective traits. Right: Quantile-quantile (QQ) plots showing observed versus expected distributions of association *P*-values for each trait. Genomic control (λ) and Linkage Disequilibrium Score Regression intercept are shown. $N_{UKB\text{-}EUR} = 101{,}729$.

S4. Manhattan and QQ plots of three UK Biobank multi-ancestry cohorts. Manhattan plots showing the $-\log_{10}(P)$ across 22 chromosomes for associations in each of fourteen polyunsaturated fatty acid (PUFA) and monounsaturated fatty acid (MUFA) traits in three UK Biobank multi-ancestry cohorts: African (AFR); Central and South Asian (CSA); and East Asian (EAS). Red line at $P < 1.678\times10^{-8}$ shows genome-wide significance threshold corrected for number of effective traits. Right: Quantile-quantile (QQ) plots showing observed versus expected distributions of association *P*-values for each trait. $N_{UKB\text{-}AFR} = 1{,}564$; $N_{UKB\text{-}CSA} = 2{,}203$; $N_{UKB\text{-}EAS} = 633$.

S5. ACLS6 novel association with omega-3 fatty acids. Regional Manhattan plot showing local associations between genes in close proximity to *ACLS6* and omega-3 fatty acids in the meta-analysis GWAS of three European cohorts.

S6. S-MultiXcan gene-based Manhattan plots. Manhattan plots showing $-\log_{10}(P)$ of significant associations between gene expression levels and summary statistics from fourteen polyunsaturated fatty acid (PUFA) and monounsaturated fatty acid (MUFA) traits in the UK Biobank discovery cohort. The red line indicates the corrected significance threshold at $P < 7.68\times10^{-7}$. The most

significant genes for each 5Mb window of significant associations are labeled. Alternating color shades designate breaks between chromosomes.

S7. Tissue expression analysis for meta-analyzed traits. Significant tissue-expression specificity by tissue type for five meta-analyzed polyunsaturated fatty acid (PUFA) and monounsaturated fatty acid (MUFA) traits. Plots were created by MAGMA. Liver is the only significant tissue type identified in these traits.

Supplementary Tables

S1. Known PUFA loci. Previously reported lead variants from significant genome-wide association studies of polyunsaturated fatty acids (PUFA) and monounsaturated fatty acids (MUFA). Each row represents a genomic risk locus identified by inputting summary statistics from previous publications into FUMA SNP2GENE. Phenotype abbreviations: AA: arachidonic acid; AdrA: adrenic acid; ALA: alpha-linolenic acid; DGLA: dihomo-gamma-linolenic acid; DHA: docosahexaenoic acid; DPA: *cis*-7,10,13,16,19-docosapentaenoic acid; DPAn6: *cis*-4,7,10,13,16-docosapentaenoic acid; EDA: eicosadienoic acid; EPA: eicosapentaenoic acid; FAw3: omega-3 fatty acids; FAw6: omega-6 fatty acids; FAw67: omega-6 and -7 fatty acids; GLA: gamma-linolenic acid; LA: linoleic acid; MUFA: monounsaturated fatty acids; OA: oleic acid; otPUFA: polyunsaturated fatty acids (other than 18:2); POA: palmitoleic acid; PUFA: polyunsaturated fatty acids.

S2. Participant characteristics table for UK Biobank cohorts. Phenotype and covariate data for UK Biobank cohorts. Continuous variables are represented as: mean (standard deviation). BMI: body mass index. EUR: European; AFR: African; CSA: Central and South Asian; EAS: East Asian.

S3. Comparison of discovery models one and two. Counts of genome-wide significant variants and compare Spearman correlation coefficient between *P*-values for variants in UKB-EUR discovery analysis sensitivity Models 1 and 2. See Supplementary Figure 2 for plots. Number of variants in analysis, genomic control (λ) and Linkage Disequilibrium Score Regression (LDSC) intercepts are shown for each trait of each model.

S4. Discovery stage GWAS summary. Number of significant variants, independent significant variants (from COJO), and significant loci for each of fourteen polyunsaturated fatty acid (PUFA) and monounsaturated fatty acid (MUFA) traits tested in the UK Biobank European discovery cohort. Novel loci for each trait and unique novel loci across are shown. SNP-based heritability (h^2) and standard error (SE) are reported.

S5. Discovery GCTA-COJO results. Output from Genome-wide Complex Trait Analysis Conditional and Joint Analysis (GCTA-COJO) using summary statistics from fourteen polyunsaturated fatty acid (PUFA) and monounsaturated fatty acid (MUFA) traits in the the UK Biobank European discovery cohort. RefA: effect allele; freq: frequency of the effect allele in the original data; b: effect size; se: standard error; p: *p*-value from original GWAS; n: estimated effective sample size; freq_geno: frequency of the effect allele in the reference sample; bJ: effect

size from joint analysis of selected SNPs; bJ_se: standard error from joint analysis of selected SNPs; pJ: *p*-value from joint analysis of selected SNPs; LD_r: LD correlation between the SNP i and SNP i + 1 for the SNPs on the list.

S6. Discovery cohort significant loci. Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) genomic risk loci from the UK Biobank (UKB) European discovery cohort summary statistics. Corresponding summary statistics from UKB multi-ancestry African (AFR), Central and South Asian (CSA), and East Asian (EAS) cohorts are also provided.

S7. Genetic and phenotypic correlation matrices. Top: Coefficients, standard error, and FDR adjusted *P*-values for genotypic and phenotypic correlations of fourteen polyunsaturated fatty acid (PUFA) and monounsaturated fatty acid (MUFA) traits in the UK Biobank European discovery cohort (UKB-EUR). Above diagonal are genetic correlations calculated using LDSC using UKB GWAS summary statistics. Below diagonal are phenotypic correlations by Pearson correlation coefficient. Bottom: Output table from BOLT-REML multi-trait heritability correlations for the six traits in the discovery UKB-EUR cohort measured in absolute concentration units (mmol/L). The diagonal represents heritability explained by genotyped SNPs, other values are genetic correlations between traits.

S8. S-PrediXcan results. Gene-trait associations by tissue type from S-PrediXcan, which reached the Bonferroni corrected significance threshold $P < 2.791 \times 10^{-8}$ (0.05/(601,176×2.98). Gene: gene ID; gene_name: HUGO gene name; Zscore: S-PrediXcan association result for the gene; Pvalue: P-value of Zscore; var_g: variance of the gene expression, calculated as W' × G × W (where W is the vector of SNP weights in a gene's model, W' is its transpose, and G is the covariance matrix); n_snps_used: number of SNPs in the covariance matrix; n_snps_in_model: number of SNPs in the model.

S9. S-MultiXcan results. Significant genes across tissue types from S-MultiXcan, which reached the Bonferroni corrected significance threshold of $P < 7.68 \times 10^{-7}$ (0.05/(21,846×2.98). Gene: gene ID; gene_name: HUGO gene name; pvalue: significance p-value of S-MultiXcan association; n: number of "tissues" available for this gene; n_indep: number of independent components of variation kept among the tissues' predictions. (Synthetic independent tissues); p_i_best: best p-value of single-tissue S-PrediXcan association; t_i_best: name of best single-tissue S-PrediXcan association; p_i_worst: worst p-value of single-tissue S-PrediXcan association; t_i_worst: name of worst single-tissue S-PrediXcan association.

S10. Summarize S-MultiXcan results. Number of significant associations from S-PrediXcan and S-MultiXcan results for fourteen polyunsaturated fatty acid (PUFA) and monounsaturated fatty acid (MUFA) traits in the UK Biobank European discovery cohort. Novelty of results also shown.

S11. UKB-EUR external replication. Number of variants in common between UK Biobank European discovery cohort (UKB-EUR) and the external European cohorts FinMetSeq and Kettunnen et al., after munging. Counts of variants from UKB-EUR that were replicated at $P < 0.05$.

S12. UKB multi-ancestry replication. Number of variants in common between the UK Biobank European discovery cohort (UKB-EUR) and UKB multi-ancestry African (AFR), Central and South Asian (CSA), and East Asian (EAS) cohorts after quality control protocol. Counts of variants from UKB-EUR that were replicated at $P < 0.05$.

S13. Meta-analysis stage GWAS summary. Number of significant variants, independent significant variants, and significant loci for each of five meta-analyzed polyunsaturated fatty acid (PUFA) and monounsaturated fatty acid (MUFA) traits. Novel loci for each trait and unique novel loci are counted here. Linkage disequilibrium score regression (LDSC) intercept, genomic control, variance explained (%) and heritability (h^2) are shown.

S14. Meta-analysis GCTA-COJO results. Genome-wide Complex Trait Analysis Conditional and Joint Analysis (GCTA-COJO) output using summary statistics from five meta-analyzed polyunsaturated fatty acid (PUFA) and monounsaturated fatty acid (MUFA) traits. RefA: effect allele; freq: frequency of the effect allele in the original data; b: effect size; se: standard error; p: *p*-value from original GWAS; n: estimated effective sample size; freq_geno: frequency of the effect allele in the reference sample; bJ: effect size from joint analysis of selected SNPs; bJ_se: standard error from joint analysis of selected SNPs; pJ: *p*-value from joint analysis of selected SNPs; LD_r: LD correlation between the SNP i and SNP i + 1 for the SNPs on the list.

S15. Meta-analysis significant loci. Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) genomic risk loci from meta-analysis of UK Biobank European discovery cohort, FinMetSeq, and Kettunen et al. studies. Corresponding summary statistics from each study are shown.

S16. Gene set enrichment. GWAS Catalog gene set enrichment of significant variant associations for five meta-analyzed polyunsaturated fatty acid (PUFA) and monounsaturated fatty acid (MUFA) traits. Genes in query gene sets mapped from significant meta-analysis associations by position, GTEx liver eQTLs, and HiC liver chromatin data using GENE2FUNC implemented by Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA).

S17. Gene set enrichment by novelty. GWAS Catalog gene set enrichment, stratified by novelty of significant variant associations for five meta-analyzed polyunsaturated fatty acid (PUFA) and monounsaturated fatty acid (MUFA) traits. Genes in query gene sets mapped from significant meta-analysis associations by position, GTEx liver eQTLs, and HiC liver chromatin data using GENE2FUNC implemented by Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA).

S18. GWAS Catalog Accessions. GWAS Catalog accession codes for all summary statistics generated in this study.

CHAPTER 4

GENE-VEGETARIANISM INTERACTIONS DETECTED IN GENOME-WIDE ANALYSES ACROSS 30 SERUM BIOMARKERS[3]

[3] Francis M & Ye K. (2022) Gene-vegetarianism interactions detected in genome-wide analyses across 30 serum biomarkers. medRxiv 2022.10.21.22281358; doi: https://doi.org/10.1101/2022.10.21.22281358.
Reprinted here with permission of the publisher.

Abstract

Large cohort studies showing health impacts of vegetarianism have not considered differences in genetics. We designed a rigorous definition of vegetarianism using data from two surveys in the UK Biobank to identify a reliable cohort of vegetarians. Vegetarians were matched 1:4 with non-vegetarians, revealing significant effects of vegetarianism in 15 of 30 serum biomarkers. Notably, all cholesterol measures plus Vitamin D (P = 2.1e-49) were significantly lower in vegetarians, while triglycerides were higher (P = 4.0e-26). We performed a genome-wide association study and found no significant associations with vegetarianism as a trait. Finally, we performed the first ever genome-wide gene-vegetarianism interaction analyses for 30 biomarker traits (N = 147,253). We detected evidence of gene-vegetarianism interaction with one genome-wide significant variant at rs72952628 (P = 4.47e-08), where the heterozygous genotype was associated with higher calcium in vegetarians. rs72952628 is located in *MMAA*, which is part of the B_{12} metabolism pathway; B_{12} has a high deficiency potential in vegetarians. Gene-based aggregation of interaction P-values revealed two additional significant genes, *RNF168* in testosterone (P = 1.45e-06), and *DOCK4* in eGFR (P = 6.76e-07), which have previously been associated with testicular and renal traits, respectively. These nutrigenetic findings suggest differences in genotype may play a role in moderating the benefits a vegetarian diet.

Acknowledgements

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institute of Health under award numbers T32GM007103 (MF) and R35GM143060 (KY). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Special thanks to the Georgia Advanced Computing Resource Center (GACRC) at the University of Georgia for supporting our data analyses.

Competing interests

The authors declare no competing interests.

Author contributions

MF designed and performed the analysis. KY supervised the project.

Introduction

Vegetarianism is a superordinate term for a variety of animal-restricted dietary practices, typically referring to lacto-ovo vegetarianism, which permits plant-based food plus dairy and eggs, and excludes meat, fish and seafood [1]. Estimates indicate that in Western countries, interest in and adherence to plant-based diets have increased over the past decade [2-5]. This has occurred for several reasons, including health benefits, taste preferences, ethical concerns with slaughtering animals and factory farming, environmental concerns related to pollution and greenhouse gas emissions, and perceived moral accreditation [5-7]. It is now typical for nutritionists to recommend vegetarianism to the general public *en masse* [5, 8-10].

Recent large meta-analyses have found health benefits associated with vegetarianism, such as improved blood lipids, and reductions in body mass index (BMI), heart disease, type 2 diabetes, and certain cancers, though no significant differences have been found in all-cause mortality [1, 11-13] . As the authors of these meta-analyses have pointed out, many vegetarian observational studies are confounded by information and selection biases [1, 11, 12]. We have attempted to find ways to address the most commonly occurring biases from these studies.

Heterogenous and imprecise questionnaire design in defining vegetarianism is an important source of information bias. Self-reported vegetarians vary widely in their strictness of following a diet that contains no meat or fish [14]. There are issues of trustworthiness in dietary questionnaire response, particularly in the direction of over-reporting "healthy" behaviors [15, 16]. Using multiple dietary assessment surveys to define variables is one way to significantly improve the quality of measurement as compared to using a single question [17-19].

Vegetarians may also be more health conscious in general than omnivores, which introduces a selection bias that has been called the "healthy user effect" [20]. When lifestyle factors

adjacent to vegetarianism are not properly controlled for, it can lead to overestimating the effect of vegetarianism. One outstanding example of this bias in vegetarianism studies, specifically those conducted in the US, has been an over-generalization of results from Seventh Day Adventists (SDAs) [1, 11, 12, 21, 22], who in addition to vegetarianism, observe many healthy lifestyle practices, such as increased emphasis on exercise, and avoidance of all tobacco, drugs, and alcohol. Meta-analyses revealed that non-SDA vegetarians consistently show less health benefits than SDAs [1, 11, 12]. Matching participants on relevant characteristics can help alleviate this issue [23]. Large-scale databases like the UK Biobank (UKB) offer an opportunity to match vegetarians to omnivores while still maintaining sufficient analysis power.

In addition to the aforementioned biases, there has been no consideration of genetics in large epidemiological studies of vegetarianism. Genetics and ancestry are known to play an important role in metabolic processes, i.e., nutrigenetics [24, 25]. There are two aspects of genetics we consider in this analysis. First, we asked whether there is a genetic component to vegetarianism status. Heritable components have been associated with plant-eating dietary preferences [26, 27]. Significant variants have been associated with quantitative measures of plant-eating [28, 29], though a recent GWAS of vegetarianism as a trait found none [30].

Perhaps more meaningful than finding a genetic predisposition towards certain dietary habits, is identifying how a diet relates to our personal genetics. This question is at the heart of the "nature plus nurture" approach of nutrigenetics. Gene-diet interactions (GDI) are a type of gene by environment interaction (GEI) where diet is the environmental exposure. GDIs are defined as a departure of the effect of a genetic polymorphism from the typical additive association model, based on differences in diet. GDIs have been identified using exposures of overall dietary patterns for some serum biomarkers [31], but gene-vegetarianism interactions have not yet been reported.

This study consists of four parts. First, by utilizing both dietary surveys administered to UKB participants, we defined a high-quality cohort of vegetarians that were most likely to be vegetarian at the time of the serum biomarker collection. Participants' vegetarianism status was based on four criteria: self-identified as vegetarian on first 24-hour recall survey (24HR), did not eat meat or fish on first 24HR, did not eat meat or fish on initial assessment, and had no major dietary changes over the past 5 years. Second, we estimated exposure effects of vegetarianism in a matched sample of vegetarian and nonvegetarian Europeans across 30 serum biomarkers. Third, we performed a genome-wide association study (GWAS) to search for variants that may explain vegetarianism preference on a genetic level. Finally, we performed the first genome-wide gene-diet interaction study (GWIS) of vegetarianism across 30 biomarkers, and identified genome-wide significant gene-vegetarianism interactions on calcium, testosterone, and estimated glomerular filtration rate (eGFR). This study provides evidence that genetic factors play a role in differential phenotypic outcomes across vegetarians, and suggests that the current trend of universal vegetarianism recommendations may be premature.

Results

Identifying a reliable sample of vegetarians

We searched the UK Biobank (UKB) to find a reliable subset of participants that were most likely to be vegetarian at the initial assessment (IA), when blood samples were collected for biomarker measurement. Two separate dietary surveys were part of UKB data collection, one at the IA which was taken by all UKB participants (N=502,413), and one in the 24-hour recall survey (24HR), which was administered after the IA in five waves or "instances", between April 2009

and June 2012 (N=210,967 unique participants; Figure 1A). Participants were invited to take the 24HR between one and five times on a voluntary basis (Figure 1B).

We used four criteria to designate participants as vegetarian. Our first criterion was whether a participant indicated they routinely followed a vegetarian or vegan diet; this question was only asked on the 24HR. A total of 9,115 participants self-identified in at least one 24HR that they were either vegetarian or vegan (hereafter collectively referred to as "vegetarian"). We found an inverse relationship between the percentage of participants who consistently self-identified as vegetarian in every 24HR they took, and the number of times participants took the 24HR (Figure 1C). For example, of the participants who identified as vegetarian at least once and participated in two instances of the 24HR, only 64.8% self-identified as vegetarian both times (1,380 of 2,130); for participants that took the 24HR in all five instances, only 45.4% consistently identified as vegetarian or vegan every time (168 of 370). Because we were interested in biomarker levels at the IA time point, we considered identification as vegetarian/vegan the earliest instance taken of the 24HR as sufficient for passing this criterion.

Next, the 24HR asked whether a participant ate meat or fish yesterday. To find intra-survey discrepancies of vegetarianism status, we identified those who identified as vegetarian and also self-reported eating meat or fish on the same instance of the 24HR. The percentage of these participants ranged from 10.02-14.01% per survey instance (Figure 1D). Participants who reported eating meat on their first 24HR were disqualified from our "reliable" vegetarianism status. Similarly, as our third criterion, we disqualified vegetarians who did not answer "Never" to questions asking their frequency of eating meat or fish on the IA.

Finally, because of the high amount of dietary fluctuation we found in self-identified vegetarians, we also required vegetarians to have answered "No" to the question on the IA which

asked whether they had any major dietary changes over the past five years. Overall, out of 9,115 UKB participants who self-identified as vegetarian or vegan on at least one 24HR, we found 3,205 met our criteria of not reporting eating meat on IA, nor on the nearest 24HR to the blood draw time point, plus had not reported major dietary changes (Table 1).

Sample matching and estimating vegetarianism effects on serum biomarkers

After quality controlling participants and keeping only those who were part of the largest ancestry group, European, using Pan UKBB designations [34], 2,328 vegetarians and 153,047 non-vegetarians remained (Supplementary Table 1). Raw (untransformed) values for 30 traits were plotted, and some exhibited apparent differences between vegetarians and non-vegetarians (Supplementary Figure 1). However, the covariates selected for our effects estimation model (age, sex, BMI, alcohol use frequency, previous smoker, current smoker, Townsend index, and the first five genetic principal components) were highly imbalanced between the two groups (Supplementary Table 2). For example, the average ages of non-vegetarians and vegetarians were 56.5 (7.9) and 52.7 (7.8), respectively. Similarly, non-vegetarians were 54.1% female, compared to 66.2% in vegetarians. The covariates with highest standardized mean differences (SMD) between the two groups were age (-0.482) and BMI (-0.501). Therefore, prior to estimating the effects of vegetarianism across the 30 traits, we matched each vegetarian to four non-vegetarians along these covariates. After matching, the absolute SMD (ASMD) in all model covariates were < 0.05 S.D. (Supplementary Figure 2). The variance ratio of the distance of propensity scores between unmatched and matched vegetarians was improved from 2.1203 to 1.0216. Similarly, the maximum empirical cumulative density function (eCDF) difference (also known as the Kolmogorov-Smirnov statistic, D_n) was improved from 0.3038 to 0.0013 (Supplementary Table 2). These measures indicate that good balance was achieved between matched vegetarians and

non-vegetarians. The untransformed trait values for matched participants was included in the initial plot (Supplementary Figure 1).

Participants were filtered for those who had complete covariate data. The standardized effect of vegetarianism was estimated across 30 serum biomarker traits with rank-based inverse normal transformation in 2,312 vegetarians and 9,248 matched non-vegetarians. Fifteen of these traits had significant effects at the Bonferroni corrected *P*-value threshold of 0.05/30 = 0.0017, while five additional trait effects were nominally significant ($P < 0.05$). (Figure 2, Supplementary Table 3). Effects of vegetarianism were significant and negative across all cholesterol measures, including total cholesterol, low-density lipoprotein cholesterol (LDL), high-density lipoprotein cholesterol (HDL), plus Apolipoproteins A and B (ApoA, ApoB); while lipoprotein (a) (Lp (a)) was nominally significant. A significant positive effect of vegetarianism was associated with triglycerides ($\beta = 0.223$; $P = 4.0e\text{-}26$).

Vegetarianism had a significant negative effect on the steroid hormone Vitamin D ($\beta = -0.388$; $P = 2.1e\text{-}49$), and with the growth hormone-regulating Insulin-like growth factor 1 (IGF-1). Sex-related hormone measures of testosterone (total, bioavailable-T, and free-T) and sex hormone binding globulin (SHBG) were not significant in the combined nor sex-stratified effects estimation (Supplementary Figure 3).

Alanine aminotransferase (ALT) and gamma-glutamyl transferase (GGT) were associated with significant negative effects of vegetarianism, while a positive effect was observed with alkaline phosphatase (ALP). Effects for other liver-associated markers such as albumin, aspartate aminotransferase, C-reactive protein, direct bilirubin, total bilirubin and total serum protein, were not significant.

Kidney markers associated with protein metabolism and breakdown, such as creatinine, urate and urea, displayed negative effects from vegetarianism, while cystatin C was associated with a strong positive effect, and eGFR did not have significant effects. HbA1c (glycated haemoglobin) was not significantly associated. Vegetarianism had a negative effect on serum calcium that nearly reached the Bonferroni significance threshold (P = 0.002), while phosphate was not significantly associated.

Sex-stratification revealed effects signals were driven by only one sex in three traits. ApoA was significant only in males, ALP and Lp (a) were significant only in females; C-reactive protein was nearly significant in females (Supplementary Figure 3; Supplementary Table 3).

Genome-wide association study

A total of 7,918,739 variants were tested in a GWAS of 152,764 European UK Biobank participants, using vegetarianism as a binary trait as defined in Table 1 in a standard model and BMI-adjusted model. P-values were highly correlated between the two models (R = 0.97; Supplementary Figure 4). No variants were significantly associated with vegetarianism at the genome-wide significance threshold (P < 5e-08; Supplementary Figure 5). Potential inflation from imbalanced case:control (2,312 vegetarians, 152,764 non-vegetarians) was properly adjusted for by regenie (λ = 1.032 in both models). The two most significant variants in both models were indels, at 4:183448129_AT_A ($P_{standard}$ = 1.645e-07; $P_{adj\text{-}BMI}$ = 1.358e-07) and 11:870094_CG_C ($P_{standard}$ = 1.612e-07; $P_{adj\text{-}BMI}$ = 2.101e-07).

Variant P-values were aggregated into genic regions using MAGMA. No genes achieved significance. The most significant genes in each model were Major Histocompatibility Complex, Class II, DP Beta 1 (*HLA-DPB1*; $P_{standard}$ = 1.12e-05; $P_{adj\text{-}BMI}$ = 5.73e-05) and Tyrosine 3-

Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein Zeta (*YWHAZ*; $P_{standard}$ = 1.44e-04; $P_{adj\text{-}BMI}$ = 4.66e-05).

Genome-wide gene-vegetarianism interactions

Variant level

Gene-environment interactions using vegetarianism status (Table 1) as the environmental exposure was performed across 30 serum biomarker traits (N = 117,356-147,253) using standard and BMI-adjusted models (Supplementary Table 4). For each GWIS, 7,934,157 variants were tested for marginal effects, interaction effects (1 degree of freedom), and joint main and interaction effects (2 degrees of freedom). We were specifically interested in interaction effects and their corresponding *P*-values, as these would most directly demonstrate the interaction of vegetarianism with genetic variants. Genomic control (λ) using non-robust standard errors ranged from 0.895-1.255, likely due to heteroskedasticity, therefore robust standard errors as implemented by GEM were used for all models; λ for robust *P*-values ranged from 0.985-1.024 (Supplementary Table 4).

Across the 30 traits analyzed for gene-vegetarianism interactions, only one variant was significant at the genome-wide significance threshold, and no variants reached significance at a stricter threshold Bonferroni corrected for the number of traits (5e-08 / 30 = 1.67e-09) (Supplementary Figure 6; Supplementary Table 4). For calcium, rs72952628 (chr4:146,637,234**)** passed the genome-wide significance threshold in the standard model and nearly in the BMI-adjusted model ($P\text{-int}_{standard}$ = 4.47e-08; $P\text{-int}_{adj\text{-}BMI}$ = 6.29e-08; Figure 3A), while the marginal P-value was high ($P\text{-marginal}_{standard}$ = 0.0269; $P\text{-marginal}_{adj\text{-}BMI}$ = 0.0233), indicating predominately interaction effects at this locus. This variant is located in the intron of Chromosome 4 Open Reading Frame 51 (*C4orf51*), and is also in moderate linkage disequilibrium (r^2: 0.605-0.719) with

variants in exon 7 of Metabolism of Cobalamin Associated A (*MMAA*) (Figure 3B). In a genotype-stratified model using the standard analysis covariates, vegetarianism effect was associated with a 0.135-unit decrease (standard deviation of calcium level) in those homozygous for the major allele CC, while the heterozygote was associated with a 0.298 increase (Figure 3C).

The gene product MMAA is a GTPase involved in one-carbon metabolism of vitamin B_{12} (B_{12}; also known as cobalamin). Specifically, MMAA helps mediate the transport of cobalamin (Cbl) into mitochondria for the final steps of adenosylcobalamin (AdoCbl) synthesis. The most prominent cause of B_{12} deficiency is inadequate dietary intake, and this is especially common among vegetarians and vegans since the majority of dietary B_{12} is derived from animal sources [53]. GTEx single-tissue eQTL data for rs72952628 showed an exclusive and significant association with *MMAA* gene expression in four tissue types, and nearly reaching the GTEx multiple testing significance threshold in liver tissue (P = 4.77e-04), where the heterozygote CT is consistently associated with higher expression of this gene (Supplementary Figure 7). GTEx bulk tissue gene expression of *MMAA* is highest in the liver (median TPM = 5.195; Supplementary Figure 8A).

There are fifteen or more gene products involved in B_{12} transport and processing [53]; of these, two have calcium-binding domains, cubilin, and CD320. In the distal ileum, binding of the IF-B_{12} complex to the cubilin receptor is calcium-dependent [54]. However, a closer candidate in the B_{12} pathway for calcium involvement is CD320. In the liver, CD320 receptor mediates transcobalamin-bound B_{12} cellular uptake, a process which is Ca^{2+} dependent [55]. This would occur in the same cells where MMAA is active in the mitochondria, including but not exclusive to liver cells.

Gene level

Interaction *P*-values of GWIS variants were aggregated into genic regions using MAGMA for each of the 30 biomarkers. Variants were mapped to 18,208 genes, making the significant *P*-value threshold corrected for the number of genes as (0.05 / 18,208 = 2.75e-06), and that threshold additionally corrected for the number of traits as (2.75e-06 / 30 = 9.15e-08). Genomic control (λ) for these aggregated models ranged from 0.898-1.098 (Supplementary Figure 9; Supplementary Table 4).

Two genes in two traits were significant at the threshold corrected for the number of genes: Ring finger protein 168 (*RNF168*) in total testosterone ($P_{standard}$ = 1.45e-06, $P_{adj\text{-}BMI}$ = 1.03e-06; Figure 4A), and Zinc finger protein 277 (*ZNF277*) in eGFR ($P_{standard}$ = 6.76e-07, $P_{adj\text{-}BMI}$ = 9.28e-06; Figure 4B). No genes in the analysis were significant at the more conservative significance level correcting for the number of traits (Supplementary Table 4).

RNF168 had the highest expression levels in the testis in GTEx (median TPM = 45.10; Supplementary Figure 8B). *RNF168* has previously been associated with testosterone levels at the top variant rs5855544 in multiple UKB GWAS (main effects) [35, 36]; but, rs5855544 exceeded our genotype missingness threshold and therefore was not included in this analysis. Our top interaction variant at this gene locus was rs73219637 (P-int$_{standard}$ = 1.46e-07; P-int $_{adj\text{-}BMI}$ = 2.31e-07; P-marginal$_{standard}$ = 0.435; P-marginal$_{adj\text{-}BMI}$ = 0.338; Figure 4C). The rs73219637 heterozygote (TC) was associated with an increased expression of *RNF168* in the testis (P = 4.81e-3), though this did not pass the GTEx multiple testing significance cutoff. The RNF168 protein is involved in the repair of DNA double-strand breaks. Mutation of this gene is associated with Riddle syndrome, symptoms of which include increased radiosensitivity, immunodeficiency, motor control and learning difficulties, facial dysmorphism, and short stature. A mouse model of Riddle

syndrome found RNF168 deficiency caused decreased spermatogenesis, and *RNF168* was identified as a candidate gene as a tumor suppressor in testicular embryonal carcinomas [56].

While *ZNF277* contains a number of variants with suggestive interaction *P*-values, the lead variant in this region is rs17159341 ($P\text{-int}_{standard}$ = 2.58e-07; $P\text{-int}_{adj\text{-}BMI}$ = 8.61e-07; $P\text{-marginal}_{standard}$ = 0.089; $P\text{-marginal}_{adj\text{-}BMI}$ = 0.102; Figure 4D), found in the first intron of Dedicator of Cytokinesis 4 (*DOCK4*). *DOCK4* appears to be a more relevant candidate gene than *ZNF277*. Though *DOCK4* has not been directly associated with eGFR in GWAS studies, it has been associated with several traits related to kidney health, such as diastolic blood pressure, type 2 diabetes, dehydroepiandrosterone sulphate measurement (a marker for adrenal disorders) and "Water consumption (glasses per day)." A recent study demonstrated that *in vivo* and *in vitro* DOCK4 expression was found to increase with high-glucose, and that DOCK4 could reverse USP36-induced epithelial-to-mesenchymal transition effect, which is involved in diabetic renal fibrosis and nephropathy [57].

Discussion

In this study we developed a multi-step approach of evaluating the health impacts of the vegetarian dietary pattern, using both traditional and genetic epidemiological methods, the latter of which has rarely been applied to vegetarianism. First, we applied a quality control procedure to identify "reliable" vegetarians (N = 2,312 European vegetarians), based on four criteria from two dietary questionnaires (Table 1; Figure 1). Next, we matched vegetarians to non-vegetarians, and estimated the effects of vegetarianism on thirty serum biomarkers in a traditional model that did not consider genetic effects. We found vegetarianism had significant effects on fifteen of these biomarker traits after multiple testing correction (Figure 2). Third, we conducted a GWAS, and

found no genetic variants that had statistically significant effects on whether a participant was vegetarian or not (Supplementary figure 5). Finally, we performed GWIS across thirty biomarkers, and identified significant gene-vegetarianism interactions in three traits: a variant-level interaction in calcium (Figure 3), and gene-level interactions in testosterone and eGFR (Figure 4). These represent the first gene-vegetarianism interactions identified to-date.

Because of the heterogeneity in the way vegetarianism is defined, plus evidence showing that single-survey self-reported dietary data is often inaccurate [15-19], we assessed the quality of the 9,115 participants who self-reported as vegetarian in one or more 24HR instances. We found several patterns in the UKB participant data that indicated rigorous quality control was necessary. For example, at each instance of the 24HR, we found that about 10-14% of self-identified vegetarians indicated eating fish, or less often meat, or both, on that same instance of the dietary survey (Figure 1D). Additionally, of 1,229 participants who indicated they "have never eaten meat in [their] lifetime," (IA, Field 3680), 132 (10.7%) also indicated on the same dietary questionnaire that they occasionally eat oily fish, with 83 participants (6.8%) indicating they eat oily fish once a week or more (Supplementary Figure 10). The simplest explanation for this discrepancy is that many people consider fish eating to be compatible with vegetarianism, despite this contradicting the common usage of that term. We also observed that in the three-year period of 24HR administration between April 2009 and June 2012, many participants either stopped identifying as vegetarian, or began identifying as one (Figure 1C). Duration of vegetarianism adherence is an important consideration that has been shown to impact effects on traits in multiple studies (i.e., vegetarianism is a "time-dependent exposure") [11, 22, 58-60]. Overall, these results show that self-identification of vegetarianism should be treated with caution in dietary surveys, and single-

criterion designations of vegetarianism can increase noise and potentially result in spurious associations [59].

The majority of results from our effects estimation (Figure 2) can be understood within the context of the restricted dietary cholesterol, increased dietary fiber, and differences in amino acid profiles found in the plant-based components of vegetarian diets. Vegetarianism had significant negative effects on serum levels of total cholesterol, all lipoproteins (LDL, HDL, Lp (a), ApoA, ApoB), and Vitamin D, which is synthesized from cholesterol. Although serum cholesterol is mainly derived from *de novo* synthesis in the liver, our results suggest that intake of animal protein can make a significant difference in serum levels of cholesterol and related molecules. These differences could also be explained by higher levels of fiber in plant-based diets, which has been shown to reduce cholesterol as well as overall inflammation [61]. Interestingly, vegetarianism had a significant and moderate positive effect on triglycerides. This finding adds further evidence that a vegetarian diet may actually raise triglycerides [62, 63], though recent large meta-analyses had opposite findings [1, 11]. This positive effect on triglycerides may be explained by low Vitamin D [64], or a higher dietary intake of simple carbohydrates [61]. Conversely, without considering genetic differences, vegetarianism did not have significant effects on the cholesterol-derived sterol hormone testosterone, nor on the two calculated testosterone traits (bioavailable-T, and free-T), nor on the testosterone inhibitor SHBG; this was observed in the full and sex-stratified effects estimations, and is consistent with previous findings [65].

Our results did not clearly indicate benefit nor harm of vegetarianism on biomarkers commonly associated with liver function. For example, we found that vegetarianism had a significant negative effect on ALT and GGT, lower levels of which are associated with healthier liver function. Conversely, we observed a significant positive effect on ALP. Increased levels of

ALP have been observed in the context of chronic kidney disease (CKD) and Vitamin D deficiency. Several studies have shown a decrease in ALP can be achieved by administering activated Vitamin D compounds [66].

Improved kidney biomarkers have been associated with increased plant protein intake [61]. Creatinine and urea, byproducts of protein metabolism, had a significant negative effect of vegetarianism. This can be explained by lower overall protein intake, amino acid composition, or increased fiber intake in vegetarian diets [61]. Vegetarianism also had a significant negative effect on urate (AKA "uric acid"), which can cause gout, kidney stones, and kidney injury in high amounts, but is also a serum antioxidant. Urate infusion has been shown to reduce neurological injury after stroke [67]. Higher consumption of fiber in plant-based diets has been associated with higher eGFR and a lower risk of developing CKD [61]. The effect of vegetarianism on serum calcium was small, negative, and marginally significant (β = -0.078; P = 0.002). Serum calcium is regulated by calcitriol (1,25-dihydroxycholecalciferol), the active form of Vitamin D made in the kidneys. Calcitriol increases serum calcium by increasing the uptake of calcium from the intestines, and may also increase calcium excretion via decreased parathyroid synthesis [68]. Calcium deficiency is a risk in vegetarian diets, though it can be mediated by increased dairy consumption [69]. Serum calcium is also indirectly dependent on intake of sodium, caffeine, and total protein [69].

It is noteworthy that two of three traits with significant gene-vegetarianism interactions, eGFR and calcium, are closely related to kidney function. This is likely due to the major differences in levels, composition, and bioavailability of proteins and minerals, plus the higher overall alkalinity, found in vegetarian diets which directly impact kidney function [61]. Meanwhile, testosterone, the third trait with significant interactions, and Vitamin D, whose

activated form regulates serum calcium, are steroids synthesized from cholesterol. None of the three traits found to have gene-vegetarianism interactions showed significant effects of vegetarianism (at $P < 0.0017$) in the traditional, non-genetic epidemiological analysis. This emphasizes the importance of genetic interaction models in understanding the phenotypic effects of an exposure.

We did not find a so-called "vegetarianism gene," nor any single variant that was significantly associated with one group being vegetarian. This null finding is similar to a recent GWAS in a Japanese cohort [30]. Variants in *HLA-DPB1*, the most significant hit in the gene-based test, have been previously associated with cognitive empathy [70], which could potentially be involved with one's decision to become vegetarian. This connection, while interesting, is highly speculative, and more evidence is necessary. We also found that for all three significant interaction loci, at rs72952628, *RNF168* and *DOCK4*, there were no vegetarianism GWAS main effects, nor marginal effects in the trait interaction analyses, that reached the suggestive genome-wide threshold of $P < 1e\text{-}05$. This strengthens the likelihood that gene-vegetarianism interaction effects are responsible for the signals at these loci.

Only one single nucleotide polymorphism (SNP) (rs72952628) had a variant-level interaction with vegetarianism at the genome-wide significance level (P-int = 4.47e-08). This SNP was found to be significantly associated with expression changes in *MMAA*, a protein in the B_{12} metabolism pathway. B_{12} deficiency is the highest nutritional concern in vegetarians, and dietary intake plays a primary role in B_{12} availability [1, 53]. And, though we did not directly query B_{12} levels, its metabolism pathway was implicated in our results. We have suggested *CD320* as a calcium-dependent candidate gene; *CD320* serves as the cellular gateway for transcobalamin-bound B_{12} to the cell [55]. Similarly, we have proposed *RNF168* and *DOCK4* as the most likely

candidate genes based on gene expression and experimental evidence related to testosterone and eGFR, respectively. More experimental evidence is needed to validate these proposals, and there may be less direct mechanisms involved in these interaction.

We made several decisions when designing our GWIS models. First, we consider 1 degree of freedom (df) interaction results more compelling and indicative of true interaction than 2df joint effects, so we did not interpret these joint effects, though they are reported in our summary statistics. We also did not perform a pre-screening filter for variants with significant main effects; this two-step approach has been used to reduce multiple testing burdens in GWIS [52]. Because the marginal effects at our significant interaction loci were weak, and in some cases not significant, it is possible some of our significant interactions may have been lost by this pre-screening. Third, in our preliminary models, we observed a high degree of inflation (some traits with $\lambda > 1.2$) presumably caused by heteroskedasticity. By correcting for this inflation using robust standard errors, the genomic control values we reported for these GWIS do not exceed 1.024, indicating type I error was properly controlled [71].

Our study was not without limitations. First, we performed a one-stage analysis (discovery only) without replication. The UKB is among the first datasets which contain dietary data and are sufficiently powered for a GWIS. The benefit in our study of being able to utilize the multiple dietary surveys and criteria in defining vegetarians from UKB, also caused us to be unable to produce an equally rigorous set of vegetarians for replication. This characterization of reliable vegetarians was also important in the context of performing GWIS [59]. Nonetheless, we consider these one-stage results valuable, for several reasons. We have clearly demonstrated the noise in a single-criterion definition of vegetarianism in UKB; this may also be broadly applicable to interpreting other studies. Next, our use of algorithmic matching in our traditional epidemiological

analysis, which simulates the experimental design of a large-scale randomized control trial, achieved a greater balance between vegetarians and non-vegetarians than has been achieved in previous effects analyses in observational studies [1, 11, 12]. Finally, despite being unreplicated, our GWIS results are valuable in the exploratory context of our analysis. We have found the first evidence of a gene-vegetarianism interaction, a new phenomenon which, once further validated, represents a highly relevant nutrigenetic finding. We hope that future researchers can use our results, analysis protocol, and open computational pipeline in future studies to conduct replications and meta-analyses, and to inform clinical trials.

The next limitation of this study is that although we found significant interactions that passed the genome-wide multiple testing correction thresholds ($P < 5e-08$ for variant-level analysis and $P < 2.75e-06$ for gene-level analysis), none of these met a threshold further corrected for thirty traits ($P < 1.67e-09$ and $P < 9.15e-08$, respectively). In this multi-trait GWIS, the multiple testing burden was high, on top of the already strict genome-wide significance threshold. GWIS have sample size requirements which require approximately four times more participants to achieve the same power as in a GWAS with comparable effect sizes [72, 73]. We suspect that future studies with larger sample sizes would produce a higher number of significant loci. This is supported by several interactions, for example, *BRINP3* in Vitamin D ($P = 3.88e-06$), and *INTU* in SHBG ($P = 3.93e-06$) which nearly reached the gene-level significance threshold of $P < 2.75e-06$.

In contrast with increasingly common recommendations that vegetarianism is universally beneficial for all people [5, 8-10], we found several significant biomarker signals of potentially worse health in vegetarians. On its face, vegetarianism is a broad category which is not specific enough to determine whether a given diet is “healthy” either overall or in specific mediative contexts. For example, vegetarian diets which are too high in carbohydrates and added sugars have

been associated with higher cardiometabolic disease risk [61]. Our traditional epidemiological analysis showed a triglyceride-raising effect of vegetarianism; raised triglycerides are a symptom of metabolic syndrome and commonly understood as a risk factor for heart disease and stroke. Lower Vitamin D and higher ALP were also observed, both of which have been associated with negative health outcomes as described above. Two traits, urate, which was significantly lower in vegetarians, as well as testosterone which had gene-vegetarianism interaction effects, have been associated with depression [67, 74]. Depression has been repeatedly associated with vegetarianism in observational studies [75]. It should also be reiterated that these results are most relevant for those who are in the same age range as our study cohort, i.e. 40 to 70 years old. Vegetarian and vegan diets for children [76] and pregnant women [77] come with serious risks of malnutrition, and should be meticulously structured if no alternative is possible.

The emerging paradigms of precision medicine and precision nutrition (also called nutrigenetics) suggest that genetic makeup should help inform optimal disease treatment strategies [78]. Gene-environment interactions are able to indicate potential differences in molecular mechanisms and pathways utilized among individuals in different exposure groups; these pathways may be different even in cases where small phenotypic variance is observed [73]. These gene-vegetarianism interactions can also help explain inconsistencies observed in previous observational studies, especially across ancestral groups [73]. We proposed three novel gene-vegetarianism interactions in this study and used available functional analyses to put these interactions into plausible biological context. But as in any genome-wide study, these statistically significant interactions must be externally replicated and verified experimentally.

Methods

Ethics

UK Biobank (UKB) approved use of medical and genetic data under Project ID 48818. Data analysis was performed on a University of Georgia high performance computing server with strict data protection protocols and two-factor authentication. Institutional Review Board (IRB) approval was obtained for human data use in this study. Participants that withdrew their consent as of Feb. 22nd, 2022 were removed (N=114).

Vegetarianism designation

UKB is a prospective cohort study containing > 500,000 participants between ages 40 and 70, who were recruited in England, Scotland, and Wales between 2006 and 2010. All UKB Field and Category references can be located in their publicly available data dictionary (https://biobank.ndph.ox.ac.uk/ukb/). Dietary data was collected in two separate surveys. All participants answered the touchscreen questionnaire on "Diet" during their initial visit to the Assessment Centre (Category ID 100052). Additionally, the "Diet by 24-hour recall" section of the "Online follow-up questionnaire" (24HR; Category ID 100090) was administered to a subset of participants on a voluntary basis, during the last phase of the initial assessment (Instance 0; N=70,689) and subsequently via email, for a total of up to five rounds between April 2009 and June 2012 (N=210,966 unique participants) [32, 33].

Our goal was to identify a subset of participants most likely to have been consistently following a strict vegetarian or vegan diet at the time of the blood draw for biomarker measurement at the Initial Assessment. Vegetarians and vegans were grouped together in all analyses because of the limited number of vegans. In this study vegetarians/vegans were defined as meeting all four of the following criteria. First, in a participant's first instance taking the 24HR, in response to the

question “Do you routinely follow a special diet?” (Field 20086), they must have indicated “Vegetarian diet (no meat, no poultry and no fish)” and/or “Vegan diet” (**Supplementary Figure**: screenshot of f.20086). Next, on that same first instance taken of the 24HR, a participant must have also answered “No” to "Did you eat any meat or poultry yesterday? Think about curry, stir-fry, sandwiches, pie fillings, sausages/burgers, liver, pate or mince," (Field 103000) as well as to "Did you eat any fish or seafood yesterday? e.g. at breakfast, takeaway with chips, smoked fish, fish pate, tuna in sandwiches." (Field 103140). Third, on the initial dietary assessment survey, participants must have answered “Never” to all of the questions asking how often meat or fish was eaten (Fields 1329, 1339, 1349, 1359, 1369, 1379, and 1389). Finally, on the Initial Assessment, participants must have answered “No” to the question “Have you made any major changes to your diet in the last 5 years?” (Field 1538).

Participants

Only participants designated as having European (EUR) ancestry by the Pan UKBB project [34] were used in analyses to avoid population stratification. Participants were removed on the following quality control parameters: mismatches between self-reported and genetic sex, poor quality genotyping as flagged by UKB, sex chromosome aneuploidy, and/or having a high degree of genetic kinship (ten or more third-degree relatives identified). Additionally, we removed the minimum number of participants to eliminate all related pairs.

Phenotype data

Continuous serum biochemistry markers were obtained from Category 17518. Oestradiol and rheumatoid factor (Fields 30800, 30820) were excluded due to limited participant data (<20% of participants). Glucose (Field 30740) was excluded due to inconsistencies in fasting times among participants, and a limited number of participants with fasting times larger than 7h. Total

cholesterol, LDL-C, and apolipoprotein B were divided by an adjustment factor (0.749, 0.684, and 0.719, respectively) for those who self-reported use of statins [35]. Three derived traits were also included. Free testosterone was calculated with the Vermeulen equation bioavailable testosterone was calculated with the Morris equation [36-38]. The CKD-EPI Creatinine-Cystatin Equation (2021) was used to calculate estimated glomerular filtration rate (eGFR) [39]. All traits were transformed using direct rank-based inverse normal transformation with random separation of ties.

Genotype data

Genotype data was provided with initial QC and imputation with Haplotype Reference Consortium (HRC) and 1000 Genomes variants by UKB (v3) as previously described [40]. Additionally, we removed variants with imputation quality score (INFO) < 0.5, minor allele frequency (MAF) < 1%, missing genotype per individual > 5%, missing genotype per variant > 2%, or Hardy-Weinberg equilibrium (HWE) $P < 1\times10^{-6}$. Variant filtering and genotype file format conversions were performed using PLINK2 alpha-v2.3 [41, 42]. After quality control, 7,918,739 variants remained. All genomic positions in this study refer to the Genome Reference Consortium Human Build 37 (GRCh37), also known as hg19.

Sample matching and estimating vegetarianism effects

To select controls for the analysis of vegetarianism exposure effects, cases were pre-processed to match four controls with nearest-neighbor (greedy) matching without replacement, using MatchIt v4.4.0.9004 [43]. Matching distance between participants was calculated by general linearized model, and was performed on the basis of age, sex, body mass index (BMI; kg/m^2), alcohol use frequency (<3 drinks/week or ≥ 3 drinks/week), previous smoker (yes/no), current smoker (yes/no), standardized Townsend deprivation index, and the first five genetic principal

components. Sixteen vegetarians with incomplete covariate information were excluded, leaving a total of 2,312 European vegetarians (**Supplementary table 2**).

Matching was followed by regression using the same vector of covariates; using the same covariates is recommended to reduce the dependence of regression estimates on modeling decisions, increase precision, reduce bias, and increase robustness of the effect estimate [23, 44]. Vegetarianism marginal effects estimates were computed by linear model in R (v4.2.1) with cluster-robust standard errors implemented by Sandwich v3.0-2 [45]. Sex-stratified models of the same matched participants were also run. Forestplot (v3.0.0) was used to make forest plots.

Genome-wide association

Genome-wide association study (GWAS) was performed using regenie (v3.1.2) [46]. Vegetarianism status as defined above was used as a binary trait. A whole genome regression model was fit at a subset of genetic markers from non-imputed UKB genotype calls. Variants used in model fitting were filtered in PLINK2 alpha-v2.3 [41, 42] by these criteria: minor allele frequency < 0.01, minor allele count < 100, genotype missingness < 0.1, Hardy-Weinberg equilibrium exact test *P*-value < 1e-15. Covariates used for both model fitting and GWAS (standard model) were: age, sex, genotyping batch, alcohol use frequency, previous smoker (yes/no), current smoker (yes/no), standardized Townsend deprivation index, and the first ten genetic principal components as provided by UKB. A BMI-adjusted model was separately run to compare sensitivity models for confounding effects of BMI. Firth correction was applied for $P < 0.01$ to reduce the bias in the maximum-likelihood estimates using a penalty term from Jeffrey's Prior as described previously [47]. Genomic control (λ) was calculated for *P*-values using the median of the chi-squared test statistics divided by the expected median of the chi-squared distribution.

Genome-wide interactions with vegetarianism

GEM (Gene–Environment interaction analysis in Millions of samples) v1.4.3 [48] was used to perform genome-wide interaction study (GWIS) of 30 continuous biomarker traits, using vegetarianism status as a binary exposure variable. Covariates used in GWIS were age, sex, genotyping batch, alcohol use frequency, previous smoker (yes/no), current smoker (yes/no), standardized Townsend deprivation index, and the first ten genetic principal components. Robust SE correction as implemented by GEM was performed in all models to correct for initially observed heteroskedasticity. Interaction effects and P-values refer to 1 degree of freedom test of variant effect with robust standard errors. Marginal effects refer to the genetic effect of a variant in the absence of gene-environment interaction terms. A BMI-adjusted model was separately run for all traits. Correlation between standard and BMI-adjusted models was assessed using a two-sided Spearman's rank correlation coefficient.

Variants were queried for associations with gene expression levels in tissues using Genotype-Tissue Expression (GTEx) Project (GTEx) Analysis Release V8 (dbGaP Accession phs000424.v8.p2). Fastman v0.1.0 was used to generate Manhattan plots [49]. Hudson (v1.0.0) was used to create interactive Manhattan plots [50].

Gene-based analyses

MAGMA v.1.10 [51] was used to aggregate *P*-values from individual variant associations (for vegetarianism) and 1 df interactions (for 30 biomarkers) to genic regions. Variants were mapped to a total of 18,208 genes using a window of +2 kb upstream and -1 kb downstream of the transcription start and stop sites to allow for the inclusion of proximal regulatory variants. Linkage disequilibrium was estimated using reference data from the 1000 Genomes British population of

European ancestry. The "multi model" method of aggregation was used to apply both "mean" and "top" models and select the one with the best fit [52].

Data availability

Full and annotated code used in this analysis, gene-level summary statistics, and interactive Manhattan plots are publicly available at https://michaelofrancis.github.io/VegetarianGDI/. Summary statistics for GWAS and GWIS at GWAS Catalog (https://www.ebi.ac.uk/gwas/). The corresponding accession numbers can be found in Supplementary table S5.

References

1. Oussalah, A., et al., *Health outcomes associated with vegetarian diets: An umbrella review of systematic reviews and meta-analyses.* Clinical Nutrition, 2020. 39(11): p. 3283-3307.

2. Hess, J.M., *Modeling Dairy-Free Vegetarian and Vegan USDA Food Patterns for Nonpregnant, Nonlactating Adults.* The Journal of Nutrition, 2022: p. nxac100.

3. Janssen, M., et al., *Motives of consumers following a vegan diet and their attitudes towards animal agriculture.* Appetite, 2016. 105: p. 643-651.

4. Wickramasinghe, K., et al., *The shift to plant-based diets: are we missing the point?* Global Food Security, 2021. 29: p. 100530.

5. Leitzmann, C., *Vegetarian nutrition: past, present, future.* The American Journal of Clinical Nutrition, 2014. 100(suppl_1): p. 496S-502S.

6. Piazza, J., et al., *Rationalizing meat consumption. The 4Ns.* Appetite, 2015. 91: p. 114-128.

7. Rosenfeld, D.L. and A.L. Burrow, *Vegetarian on purpose: Understanding the motivations of plant-based dieters.* Appetite, 2017. 116: p. 456-463.

8. Melina, V., W. Craig, and S. Levin, *Position of the Academy of Nutrition and Dietetics: vegetarian diets.* Journal of the Academy of Nutrition and Dietetics, 2016. 116(12): p. 1970-1980.

9. Radnitz, C., B. Beezhold, and J. DiMatteo, *Investigation of lifestyle choices of individuals following a vegan diet for health and ethical reasons.* Appetite, 2015. 90: p. 31-36.

10. Willett, W., et al., *Food in the Anthropocene: the EAT–Lancet Commission on healthy diets from sustainable food systems.* The Lancet, 2019. 393(10170): p. 447-492.

11. Dinu, M., et al., *Vegetarian, vegan diets and multiple health outcomes: a systematic review with meta-analysis of observational studies.* Critical reviews in food science and nutrition, 2017. 57(17): p. 3640-3649.

12. Kwok, C.S., et al., *Vegetarian diet, Seventh Day Adventists and risk of cardiovascular mortality: a systematic review and meta-analysis.* International journal of cardiology, 2014. 176(3): p. 680-686.

13. Huang, T., et al., *Cardiovascular disease mortality and cancer incidence in vegetarians: a meta-analysis and systematic review.* Ann Nutr Metab, 2012. 60(4): p. 233-40.

14. Rosenfeld, D.L., *Psychometric properties of the Dietarian Identity Questionnaire among vegetarians.* Food Quality and Preference, 2019. 74: p. 135-141.

15. van de Mortel, T.F., *Faking It: Social Desirability Response Bias in Self-report Research.* The Australian Journal of Advanced Nursing, 2008. 25(4): p. 40-48.

16. Tucker, K.L., et al., *Quantifying diet for nutrigenomic studies.* Annu Rev Nutr, 2013. 33: p. 349-71.

17. Burton, P.R., et al., *Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology.* International Journal of Epidemiology, 2009. 38(1): p. 263-273.

18. Grandits, G.A., G.E. Bartsch, and J. Stamler, *Method issues in dietary data analyses in the Multiple Risk Factor Intervention Trial.* The American journal of clinical nutrition, 1997. 65(1): p. 211S-227S.

19. Francis, M., et al., *Genome-wide association study of fish oil supplementation on lipid traits in 81,246 individuals reveals new gene-diet interaction loci.* PLOS Genetics, 2021. 17(3): p. e1009431.

20. Shrank, W.H., A.R. Patrick, and M. Alan Brookhart, *Healthy User and Related Biases in Observational Studies of Preventive Interventions: A Primer for Physicians.* Journal of General Internal Medicine, 2011. 26(5): p. 546-550.

21. Orlich, M.J., et al., *Vegetarian Dietary Patterns and Mortality in Adventist Health Study 2.* JAMA Internal Medicine, 2013. 173(13): p. 1230-1238.

22. Key, T.J., et al., *Mortality in vegetarians and nonvegetarians: detailed findings from a collaborative analysis of 5 prospective studies.* The American Journal of Clinical Nutrition, 1999. 70(3): p. 516s-524s.

23. Ho, D.E., et al., *Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference.* Political Analysis, 2007. 15(3): p. 199-236.

24. Marcum, J.A., *Nutrigenetics/Nutrigenomics, Personalized Nutrition, and Precision Healthcare.* Current Nutrition Reports, 2020. 9(4): p. 338-345.

25. Goodarzi, M.O., *Genetics of obesity: what genetic association studies have taught us about the biology of obesity and its complications.* Lancet Diabetes Endocrinol, 2018. 6(3): p. 223-236.

26. Çınar, Ç., et al., *Sex differences in the genetic and environmental underpinnings of meat and plant preferences.* Food Quality and Preference, 2022. 98: p. 104421.

27. Smith, A.D., et al., *Genetic and environmental influences on food preferences in adolescence.* The American journal of clinical nutrition, 2016. 104(2): p. 446-453.

28. Niarchou, M., et al., *Genome-wide association study of dietary intake in the UK biobank study and its associations with schizophrenia and other traits.* Translational Psychiatry, 2020. 10(1): p. 51.

29. Matoba, N., et al., *GWAS of 165,084 Japanese individuals identified nine loci associated with dietary habits.* Nature Human Behaviour, 2020. 4(3): p. 308-316.

30. Nakamura, Y., et al., *A genome-wide association study on meat consumption in a Japanese population: the Japan Multi-Institutional Collaborative Cohort study.* Journal of Nutritional Science, 2021. 10: p. e61.

31. Abdullah, M.M.H., et al. *Common Genetic Variations Involved in the Inter-Individual Variability of Circulating Cholesterol Concentrations in Response to Diets: A Narrative Review of Recent Evidence*. Nutrients, 2021. 13, DOI: 10.3390/nu13020695.

32. Bradbury, K.E., et al., *Dietary assessment in UK Biobank: an evaluation of the performance of the touchscreen dietary questionnaire.* Journal of nutritional science, 2018. 7.

33. Liu, B., et al., *Development and evaluation of the Oxford WebQ, a low-cost, web-based method for assessment of previous 24 h dietary intakes in large-scale prospective studies.* Public health nutrition, 2011. 14(11): p. 1998-2005.

34. team, P.-U. 2020; Available from: https://pan.ukbb.broadinstitute.org.

35. Sinnott-Armstrong, N., et al., *Genetics of 35 blood and urine biomarkers in the UK Biobank.* Nature Genetics, 2021. 53(2): p. 185-194.

36. Ruth, K.S., et al., *Using human genetics to understand the disease impacts of testosterone in men and women.* Nature Medicine, 2020. 26(2): p. 252-258.

37. Vermeulen, A., L. Verdonck, and J.M. Kaufman, *A critical evaluation of simple methods for the estimation of free testosterone in serum.* J Clin Endocrinol Metab, 1999. 84(10): p. 3666-72.

38. Morris, P.D., et al., *A mathematical comparison of techniques to predict biologically available testosterone in a cohort of 1072 men.* Eur J Endocrinol, 2004. 151(2): p. 241-9.

39. Inker, L.A., et al., *New Creatinine- and Cystatin C–Based Equations to Estimate GFR without Race.* New England Journal of Medicine, 2021. 385(19): p. 1737-1749.

40. Bycroft, C., et al., *The UK Biobank resource with deep phenotyping and genomic data.* Nature, 2018. 562(7726): p. 203-209.

41. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets.* Gigascience, 2015. 4: p. 7.

42. Purcell, S.C.C., *PLINK* 2020: www.cog-genomics.org/plink/2.0/.

43. Ho, D., et al., *MatchIt: Nonparametric Preprocessing for Parametric Causal Inference.* Journal of Statistical Software, 2011. 42(8): p. 1 - 28.

44. Abadie, A. and J. Spiess, *Robust Post-Matching Inference.* Journal of the American Statistical Association, 2022. 117(538): p. 983-995.

45. Zeileis, A., S. Köll, and N. Graham, *Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R.* Journal of Statistical Software, 2020. 95(1): p. 1 - 36.

46. Mbatchou, J., et al., *Computationally efficient whole-genome regression for quantitative and binary traits.* Nature Genetics, 2021. 53(7): p. 1097-1103.

47. Firth, D., *Bias reduction of maximum likelihood estimates.* Biometrika, 1993. 80(1): p. 27-38.

48. Westerman, K.E., et al., *GEM: scalable and flexible gene–environment interaction analysis in millions of samples.* Bioinformatics, 2021. 37(20): p. 3514-3520.

49. Paria, S.S., S.R. Rahman, and K. Adhikari, *fastman: A fast algorithm for visualizing GWAS results using Manhattan and Q-Q plots.* bioRxiv, 2022: p. 2022.04.19.488738.

50. Lucas, A., A. Verma, and M.D. Ritchie, *hudson: A User-Friendly R Package to Extend Manhattan Plots.* bioRxiv, 2022: p. 2022.01.25.474274.

51. de Leeuw, C.A., et al., *MAGMA: Generalized Gene-Set Analysis of GWAS Data.* PLOS Computational Biology, 2015. 11(4): p. e1004219.

52. Werme, J., et al., *Genome-wide gene-environment interactions in neuroticism: an exploratory study across 25 environments.* Translational psychiatry, 2021. 11(1): p. 1-13.

53. Nielsen, M.J., et al., *Vitamin B12 transport from food to the body's cells—a sophisticated, multistep pathway.* Nature reviews Gastroenterology & hepatology, 2012. 9(6): p. 345-354.

54. Ahmed, M.A., *Metformin and vitamin B12 deficiency: where do we stand?* Journal of Pharmacy & Pharmaceutical Sciences, 2016. 19(3): p. 382-398.

55. Alam, A., et al., *Structural basis of transcobalamin recognition by human CD320 receptor.* Nature Communications, 2016. 7(1): p. 12100.

56. Cheung, H.-H., et al., *Hypermethylation of genes in testicular embryonal carcinomas.* British Journal of Cancer, 2016. 114(2): p. 230-236.

57. Zhu, S., et al., *USP36-Mediated Deubiquitination of DOCK4 Contributes to the Diabetic Renal Tubular Epithelial Cell Injury via Wnt/β-Catenin Signaling Pathway.* Frontiers in Cell and Developmental Biology, 2021. 9.

58. Haghighatdoost, F., et al., *Association of vegetarian diet with inflammatory biomarkers: a systematic review and meta-analysis of observational studies.* Public Health Nutrition, 2017. 20(15): p. 2713-2721.

59. Thomas, D., *Gene--environment-wide association studies: emerging approaches.* Nat Rev Genet, 2010. 11(4): p. 259-72.

60. Key, T.J., et al., *Mortality in British vegetarians: results from the European Prospective Investigation into Cancer and Nutrition (EPIC-Oxford).* The American Journal of Clinical Nutrition, 2009. 89(5): p. 1613S-1619S.

61. Carrero, J.J., et al., *Plant-based diets to manage the risks and complications of chronic kidney disease.* Nature Reviews Nephrology, 2020. 16(9): p. 525-542.

62. Yokoyama, Y., S.M. Levin, and N.D. Barnard, *Association between plant-based diets and plasma lipids: a systematic review and meta-analysis.* Nutr Rev, 2017. 75(9): p. 683-698.

63. Viguiliouk, E., et al., *Effect of vegetarian dietary patterns on cardiometabolic risk factors in diabetes: A systematic review and meta-analysis of randomized controlled trials.* Clinical Nutrition, 2019. 38(3): p. 1133-1145.

64. Cheng, Y.-L., et al. *Sex and Age Differences Modulate Association of Vitamin D with Serum Triglyceride Levels*. Journal of Personalized Medicine, 2022. 12, DOI: 10.3390/jpm12030440.

65. Allen, N., et al., *Hormones and diet: low insulin-like growth factor-I but normal bioavailable androgens in vegan men.* British Journal of Cancer, 2000. 83(1): p. 95-97.

66. Kalantar-Zadeh, K. and C.P. Kovesdy, *Clinical outcomes with active versus nutritional vitamin D compounds in chronic kidney disease.* Clinical Journal of the American Society of Nephrology, 2009. 4(9): p. 1529-1539.

67. Kim, W.-J., et al., *Low levels of serum urate are associated with a higher prevalence of depression in older adults: a nationwide cross-sectional study in Korea.* Arthritis Research & Therapy, 2020. 22(1): p. 104.

68. Letavernier, E. and M. Daudon *Vitamin D, Hypercalciuria and Kidney Stones*. Nutrients, 2018. 10, DOI: 10.3390/nu10030366.

69. Weaver, C.M., W.R. Proulx, and R. Heaney, *Choices for achieving adequate dietary calcium with a vegetarian diet.* The American Journal of Clinical Nutrition, 1999. 70(3): p. 543s-548s.

70. Warrier, V., et al., *Genome-wide meta-analysis of cognitive empathy: heritability, and correlates with sex, neuropsychiatric conditions and cognition.* Mol Psychiatry, 2018. 23(6): p. 1402-1409.

71. Rao, T.J. and M.A. Province, *A framework for interpreting type I error rates from a product-term model of interaction applied to quantitative traits.* Genetic epidemiology, 2016. 40(2): p. 144-153.

72. Gauderman, W.J., et al., *Update on the State of the Science for Analytical Methods for Gene-Environment Interactions.* Am J Epidemiol, 2017. 186(7): p. 762-770.

73. Laville, V., et al., *Gene-lifestyle interactions in the genomics of human complex traits.* European Journal of Human Genetics, 2022. 30(6): p. 730-739.

74. Zarrouf, F.A., et al., *Testosterone and Depression: Systematic Review and Meta-Analysis.* Journal of Psychiatric Practice®, 2009. 15(4).

75. Iguacel, I., et al., *Vegetarianism and veganism compared with mental health and cognitive outcomes: a systematic review and meta-analysis.* Nutrition Reviews, 2021. 79(4): p. 361-381.

76. Hovinen, T., et al., *Vegan diet in young children remodels metabolism and challenges the statuses of essential nutrients.* EMBO Molecular Medicine, 2021. 13(2): p. e13492.

77. Sebastiani, G., et al. *The Effects of Vegetarian and Vegan Diet during Pregnancy on the Health of Mothers and Offspring*. Nutrients, 2019. 11, DOI: 10.3390/nu11030557.

78. De Toro-Martín, J., et al., *Precision Nutrition: A Review of Personalized Nutritional Approaches for the Prevention and Management of Metabolic Syndrome.* Nutrients, 2017. 9(8): p. 913.

Table 4.1. Selecting high quality vegetarians for analysis.

Vegetarians were selected on four criteria: self-identifying as vegetarian on first 24-hour recall survey (24HR) that they participated in, no eating meat or fish on first 24HR, no eating meat or fish on initial assessment, and no major dietary changes over the past 5 years. A total of 3,205 UK Biobank participants met these criteria (top row; green highlight). This table shows counts of participants from all UK Biobank participants who took the 24HR (N = 210,967). After filtering by ancestry, the total of 3,205 became 2,312 European vegetarians, the number used in the analyses that follow.

Survey results for 9115 participants who self-identified as vegetarians at least once

Veg. on first 24HR taken	Ate meat/fish on first 24HR taken	Ate meat/fish on initial assessment	Major dietary changes past 5 years	N
Yes	No	No	No	3205
Yes	No	No	Yes	1136
Yes	Yes	No	No	11
Yes	Yes	No	Yes	11
Yes	No	Yes	No	1628
Yes	No	Yes	Yes	932
Yes	Yes	Yes	No	490
Yes	Yes	Yes	Yes	369
Yes	No	NA	NA	6
No				1327

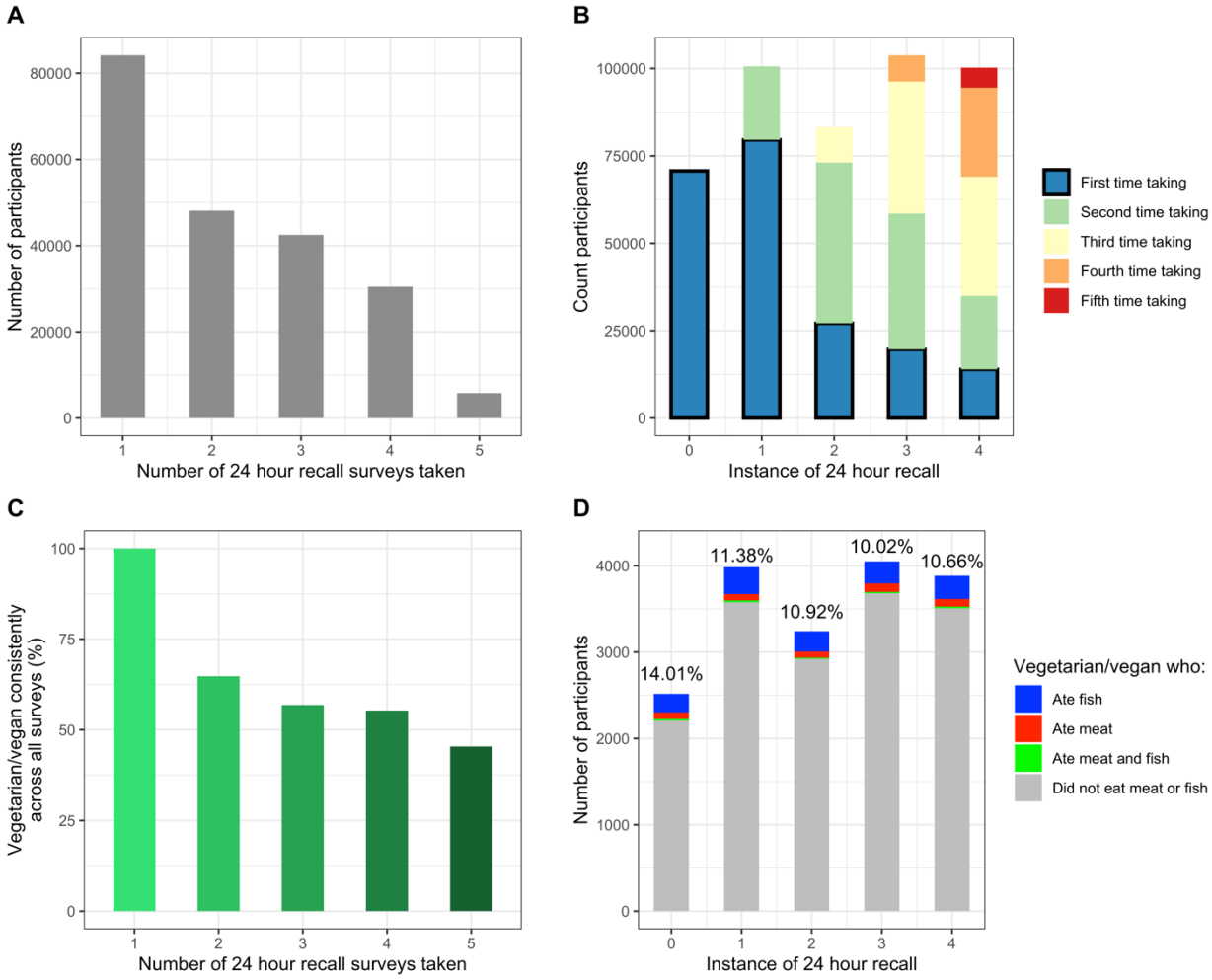

Figure 4.1. Identifying vegetarians.

(A) Participants were invited to take the 24-hour recall survey (24HR) between one and five times on a voluntary basis. (B) The 24HR we considered were restricted to the first time participants took that survey, because this was the closest time point to the blood draw of biomarkers at the initial assessment. (C) For participants who self-identified as vegetarian in at least one 24HR and took multiple 24HRs, they were less likely to self-identify as vegetarian in all surveys. (D) The percentage of vegetarians who indicated eating meat or fish on the same 24HR as identifying as vegetarian ranged between 10.02-14.01%.

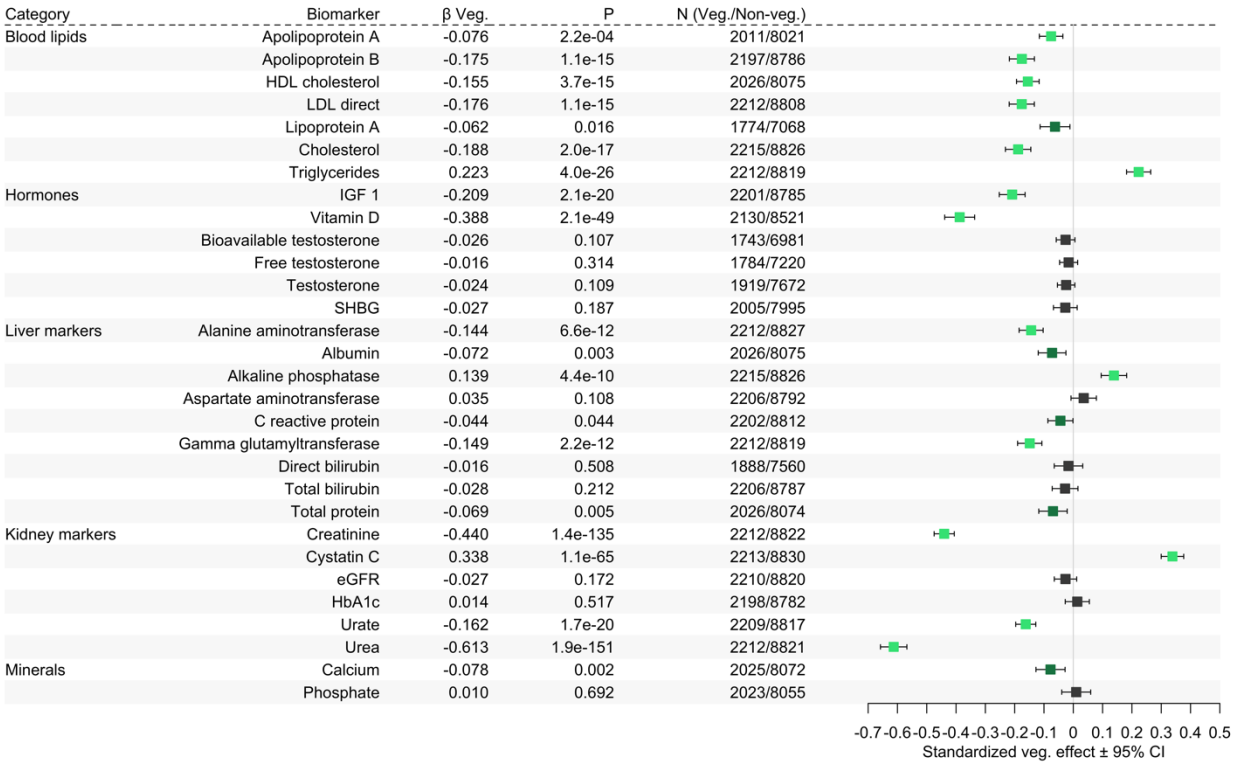

Figure 4.2. Forest plot of estimated vegetarianism effects.

Vegetarians were matched 1:4 with non-vegetarians and effects of vegetarianism were estimated across thirty biomarkers. Error bars show 95% confidence intervals. Light green dots indicate Bonferroni-corrected significance $P < 0.0017$), dark green show nominally significant $P < 0.05$, and black dots are not significant.

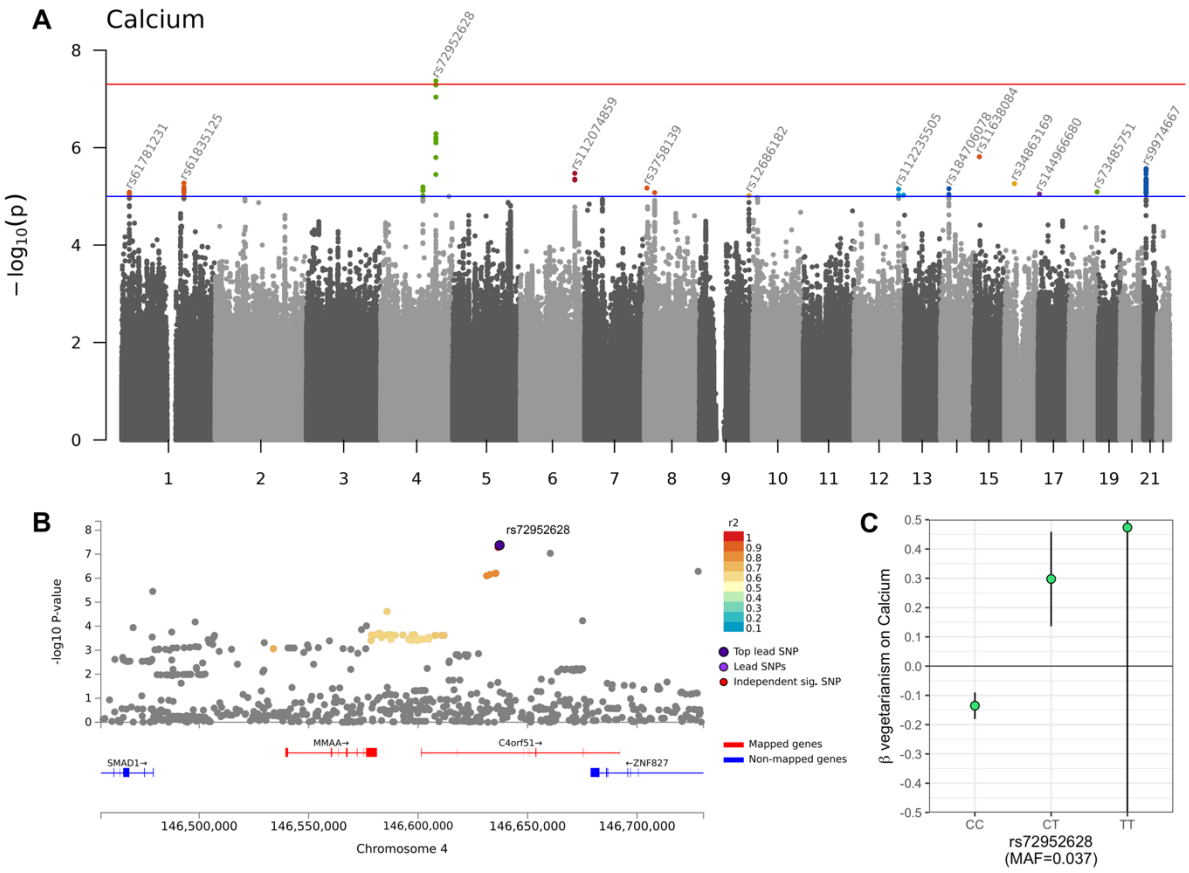

Figure 4.3. Calcium gene-vegetarianism interaction at rs72952628.

(A) Manhattan plot of *P*-values for gene-vegetarianism interaction on calcium. One variant, rs72952628 (chr4:146,637,234**)**, passed the genome-wide significance threshold of $P < 5e\text{-}08$. (B) The regional Manhattan plot of rs72952628 shows rs72952628 is in linkage disequilibrium with variants in *C4orf51* and *MMAA*. (C) The effect of vegetarianism on calcium, stratified by genotype. The homozygous minor genotype, TT, has large error because of its infrequency in our sample (n = 207). Error bars show 95% confidence interval. Units of calcium are s.d.

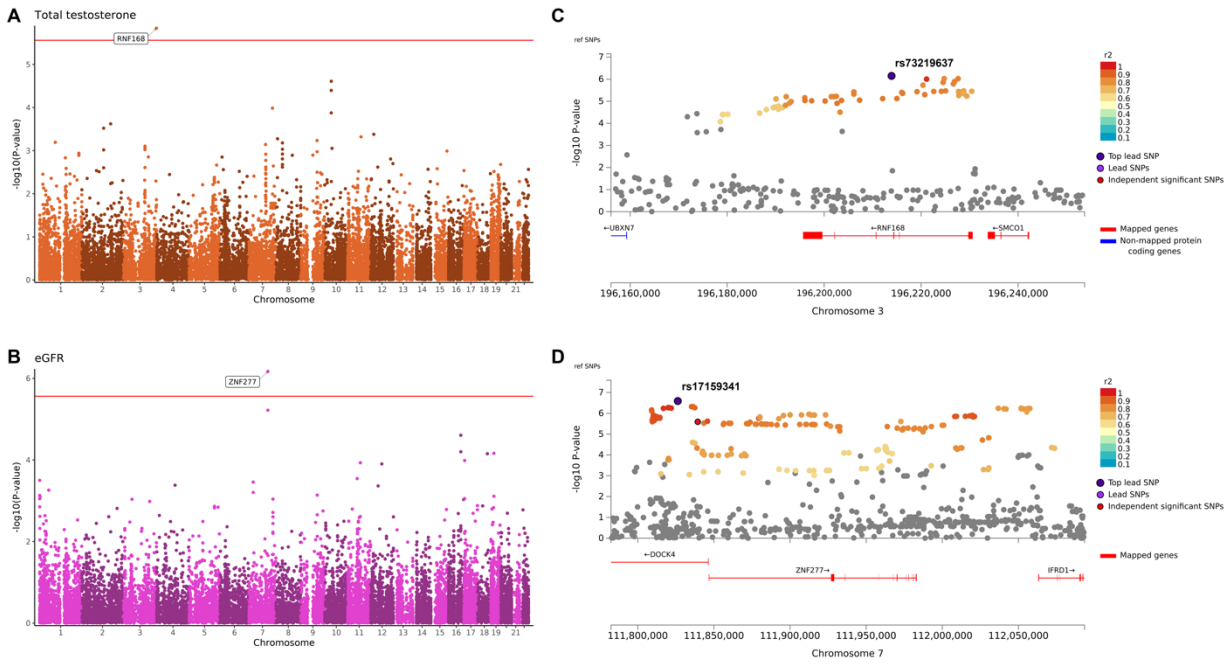

Figure 4.4. Significant gene-level gene-vegetarianism interactions.

Gene-level Manhattan plots for two traits, (A) testosterone and (B) eGFR, which had gene-vegetarianism interactions that reached significance at a level corrected for the number of genes tested (red line at P = 2.75e-06). Local Manhattan plots show the top variant-level interactions at (C) *RNF168* in testosterone and (D) *ZNF277* / *DOCK4* in eGFR. Red genes indicate these genes were positionally mapped to the locus of significant interaction variants. Variants in linkage disequilibrium with the top lead variant are color-coded according to their r^2 values.

Supplementary Figure and Table legends

Supplementary figures and tables can be downloaded from: https://www.medrxiv.org/content/10.1101/2022.10.21.22281358v1.supplementary-material

Supplementary figures

S1. Boxplots of unadjusted trait values. Comparing raw values of vegetarians (as defined by Table 1) and non-vegetarians across 30 biomarker traits. Boxplots show first decile, first quartile, median, third quartile, and last decile. Dot and label refer to mean. Units of each biomarker ("value") can be found in table S1. Matched cohort details are found in table S2 and S3. (A) Full cohort. (B) Stratified by sex.

S2. Love plot of covariates before and after matching. Plot shows the absolute standardized mean difference of model covariates in non-vegetarians before and after matching with vegetarians for effects estimation. After matching, the ASMD in all model covariates were < 0.05 standardized units. BMI=body mass index; AlcoholFreq = frequency of alcohol usage (<3 drinks/week or ≥ 3 drinks/week); zTownsend = standardized Townsend deprivation index; PCA = genetic principal component; distance = matching distance between participants was calculated by general linearized model.

S3. Sex-stratified forest plot. Effects estimation for vegetarianism in (BMI adjusted) model. Participants stratified by male or female. Error bars indicate 95% confidence interval. Bonferroni-corrected significance threshold at P = 0.0017. Full data is shown in **S3 table**.

S4. Correlation plot comparing *P*-values of BMI-adjusted model. Vegetarianism GWAS $-\log_{10}(P)$ between BMI-adjusted versus standard (without BMI) models were compared. Each point represents one variant. Spearman's Rho (R) and correlation *P*-value shown. Correlation coefficients for interaction analysis are found in **S4 table**.

S5. Vegetarianism genome-wide association Manhattan plots. Manhattan plots and QQ plots showing the $-\log_{10}(P)$ of genetic main effects with vegetarianism as a binary trait outcome. Genomic control (λ) for each model is shown in the QQ plots. Plots correspond to (A) Variant-level GWAS, (B) variant-level (BMI adjusted) GWAS, (C) gene-level GWAS where *P*-values were aggregated by MAGMA and (D) gene-level GWAS (BMI-adjusted). Top variants in a 60 Mb window that exceeded the genome-wide suggestive threshold (P = 1e-05; blue line) were annotated. Top genes (P < 1e-04) in a 5 Mb window were annotated. No variants or genes for vegetarianism as a trait were significant.

S6. Variant-level gene-vegetarianism interaction Manhattan plots. Manhattan plots and QQ plots showing the variant-level $-\log_{10}(P)$ of genome-wide gene-vegetarianism interaction effects in thirty serum biomarker traits. The blue line corresponds to the genome-wide suggestive threshold (P < 1e-05). In the standard interaction model (A), one trait, calcium, had a significant variant above the genome-wide significance threshold (P < 5e-08; red line). No variants were significant in the BMI-adjusted model (B).

S7. eQTLs for MMAA and rs72952628. Violin plots showing expression quantitative trait loci (eQTLs) in four tissues which were significant at the GTEx multiple testing threshold (adipose: subcutaneous, colon: sigmoid, muscle: skeletal, and cells: cultured fibroblasts) plus liver tissue, which nearly reached significance. In all five of these tissues, the heterozygote (CT) shows higher median normalized expression.

S8. Bulk tissue gene expression for interaction genes. Candidate genes for significant interactions with vegetarianism in either the variant-level or gene-level analyses. Transcripts per million (TPM) shown in tissues ranked from low to high for the genes (A) *MMAA,* (B) *RNF168*, and (C) *DOCK5*.

S9. Gene-level gene-vegetarianism interaction Manhattan plots. Manhattan plots and QQ plots showing the gene-level $-\log_{10}(P)$ of genome-wide gene-vegetarianism interaction effects in thirty serum biomarker traits. The red line corresponds to the genome-wide significance threshold (). In the standard interaction model (A) two traits, estimated glomerular filtration rate (eGFR) and testosterone, had a significant gene above the genome-wide significance threshold ($P < 2.75e-06$; red line). Testosterone had one significant gene in the BMI-adjusted model (B).

S10. Fish eating frequency of those who have "never eaten meat" in their lifetime. Bar plot shows non-oily fish and oily fish eating frequency, reported at Initial Assessment, for those who reported on that same dietary survey that they had "never eaten meat in [their] lifetime" (N=1,230).

Supplementary tables

S1. Participant characteristics. Categorical covariate (top), continuous covariate (middle) and phenotype data for 155,375 European UK Biobank participants used in analyses. Continuous variables are represented as: mean (standard deviation). Values are shown as full cohort and stratified by vegetarianism as defined in Table 1.

S2. Matching summary. Top: matchit function call used for 1:4 matching of vegetarian and non-vegetarian participants for use in effects estimation analysis. Middle: Summary of balance for all data and for matched data shows the results of matching on relevant lifestyle factors and genetic principal components one through five. Bottom: Sample sizes in control (non-vegetarian) and treated (vegetarian) samples before and after matching. Std. Mean Diff. = standardized mean difference; Var. Ratio = variance ratio; eCDF Mean = empirical cumulative density functions to assess imbalance across entire covariate distribution; eCDF Max = maximum eCDF difference, also known as the Kolmogorov-Smirnov statistic.

S3. Estimate effects. Effects of vegetarianism across 30 traits in full and sex-stratified matched groups. Left = BMI-adjusted models, right = models without BMI. BetaVeg = the effect of vegetarianism; SE = standard error; BMI = body mass index; M = male only; F = female only.

S4. Summarize GWAS/GWIS. Top: Most significant hits for GWAS of vegetarianism as a trait in variant-level and gene-level analysis. Most significant interaction hits across 30 traits in variant-level and gene-level analysis. All traits analyzed in standard and BMI-adjusted models. Start and stop coordinates of genes represent the +2 Kbp upstream and -1 Kbp downstream window of variant P-value aggregation. GC λ = genomic control; A0 = non effect allele; A1 = effect allele; A1Freq = frequency of effect allele; P interaction = P-value of 1df interaction test for variant calculated with robust standard errors; NSNPS = number of SNPs annotated to the top gene; NPARAM = number of relevant parameters used in model; P interaction (MULTI) = gene P-value for best fit of "mean" and "top" models.

S5. GWAS catalog accessions. GWAS Catalog accession codes for all variant-level GWAS and GWIS summary statistics generated in this study.

CHAPTER 5

CONCLUSION

The studies contained in Chapters 2, 3, and 4 contain novel contributions to nutrigenetics. These GWAS and GWIS analyses have narrowed down, from the millions of possible genetic variants, a small number of candidate regions which are statistically associated with the nutrient-related traits we queried. In all three chapters, we have provided functional interpretations of our results, and placed the significant findings into biological context. In Chapters 2 and 3, our analyses also had built-in replication and meta-analysis stages, to further verify our results.

These studies all use the UK Biobank (UKB) dataset as the discovery cohort. UKB is the largest biobank-scale dataset to-date which, in addition to genetic data, also includes extensive dietary information from participants. The quantitative genetics discoveries that can be made with this dataset is only beginning to become apparent. However, one downside of UKB is that the results are only properly generalized to Europeans. More ancestrally diverse datasets such as All of Us and Million Veteran Program (MVP), which will be publicly available in coming years, should help to alleviate this Caucasian/European bias.

As our analyses became more sophisticated over time, so too did our data-sharing approach. For the study in Chapter 2, we only shared key scripts. This is regrettable in hindsight; however, our individual protocols and the protocols of our lab had not yet matured. But in Chapter 3, we shared all analysis scripts. And in Chapter 4, we improved even that, by developing an interactive website to host our annotated analysis protocols. Additionally, summary statistics for all genome-wide analyses from all three chapters are publicly available and hosted by GWAS Catalog. It is hopeful that the user-friendly website style we used in Chapter 4 will eventually

become the standard of data-sharing in the fields of bioinformatics and quantitative genetics. As datasets get larger, the possibility for errors increases as well, and false positives can linger in the literature for decades or longer. Clear and transparent analysis protocols can help alleviate this issue.

Nutrigenetics has in the past decade finally begun to demonstrate and solidify a concept which should have perhaps already been apparent: your optimal diet is dependent upon your genetic makeup. Instead of recommending universally applicable diets, leaders in the field of nutrition ought to move towards individually-optimized diets. We hope the results presented by these three studies can be experimentally validated, so that they can inform precision nutrition recommendations and, ultimately, help improve human health.