Gene-vegetarianism interactions detected in genome-wide analyses across 30 serum biomarkers

Michael Francis[1*], Kaixiong Ye[1,2*]

[1] Institute of Bioinformatics, University of Georgia, Athens, Georgia, US

[2] Department of Genetics, University of Georgia, Athens, Georgia, US

*Corresponding Author

Email: michaelfrancisphd@protonmail.com

Email: kaixiong.ye@uga.edu

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.

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.

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.” 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 in the gene-environment interaction model with robust standard errors. Marginal effects refer to the association between a genetic effect and phenotype when the environment has been mean-centered. 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].

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-int$_{standard}$ = 4.47e-08; P-int $_{adj\text{-}BMI}$ = 6.29e-08; Figure 3A), while the marginal P-value was high (P-marginal$_{standard}$ = 0.0269; P-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-int$_{standard}$ = 2.58e-07; P-int $_{adj\text{-}BMI}$ = 8.61e-07; P-marginal$_{standard}$ = 0.089; P-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 = 2,328 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 $P < 1e-05$. 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.

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.

Figure captions

Main tables

Table 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.

Main figures

Figure 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 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 3. Calcium gene-vegetarianism interaction at rs72952628 (chr4:146,637,234). (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. 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 $P = 2.75e\text{-}06$). 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 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.

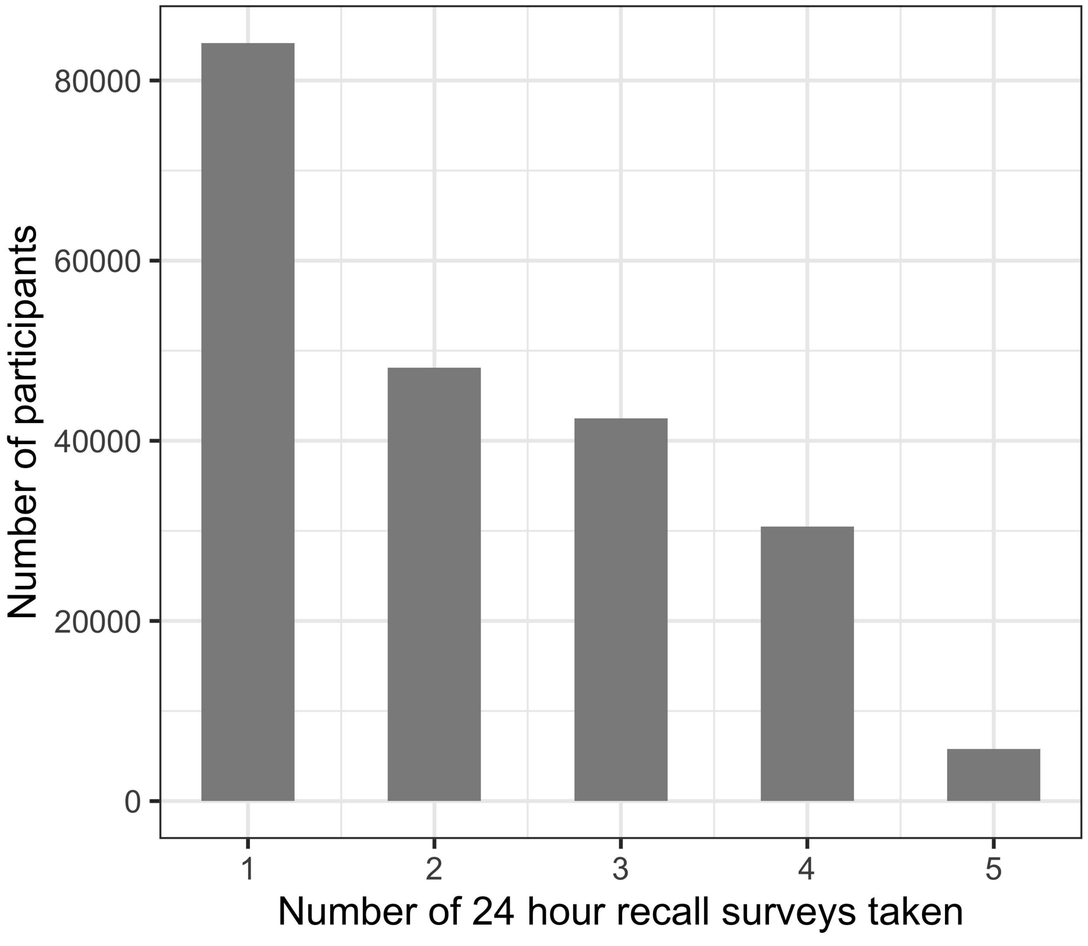
S4. Summarize GWAS/GWIS. Top: Most significant hits for GWAS of vegetarianism as a trait in variant-level and gene-level analysis. Most significant interaction hits across 30 traits in variant-level and gene-level analysis. All traits analyzed in standard and BMI-adjusted models. Start and stop coordinates 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.

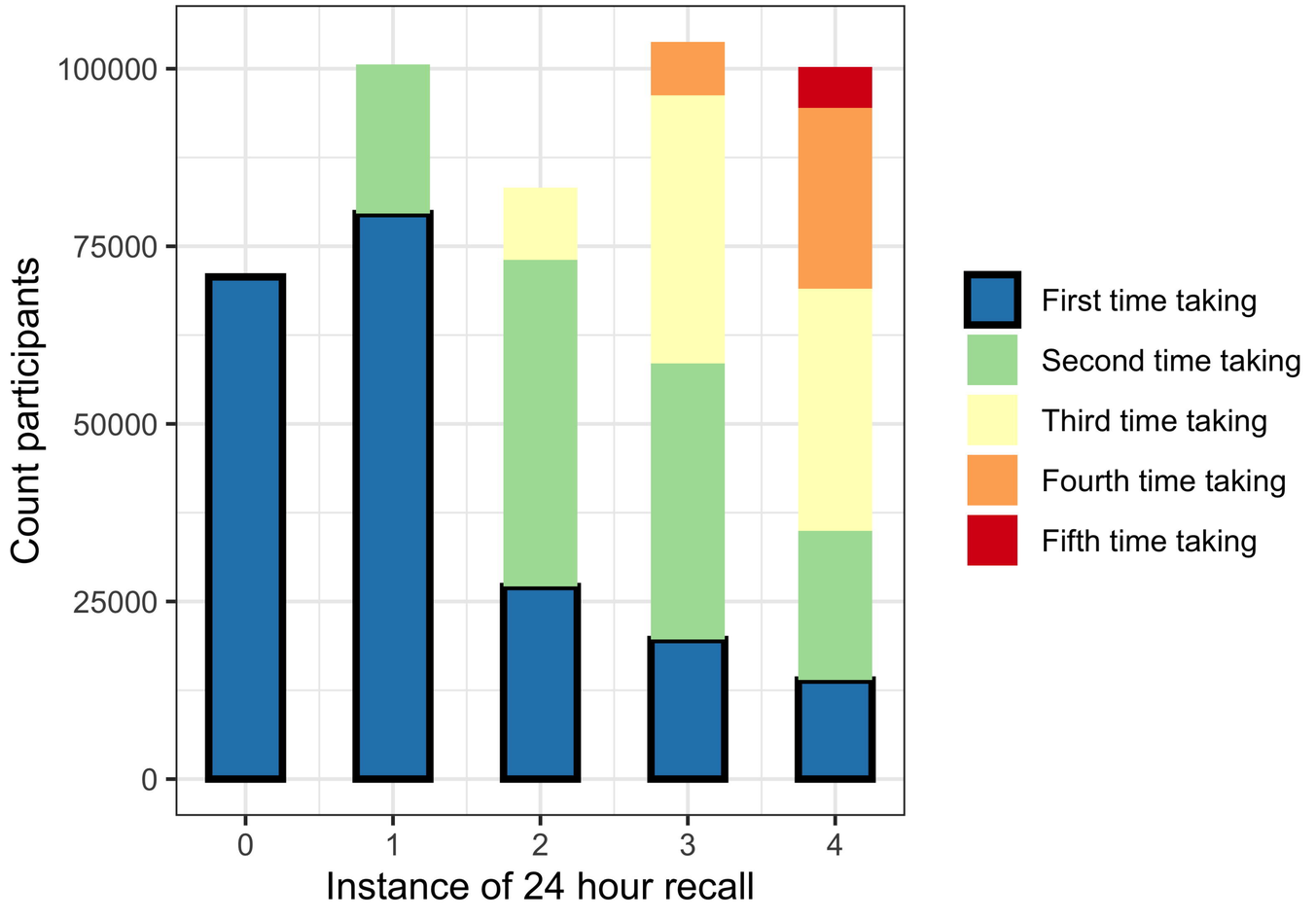
Survey results for 9115 participants who self-identified as vegetarians at least once

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

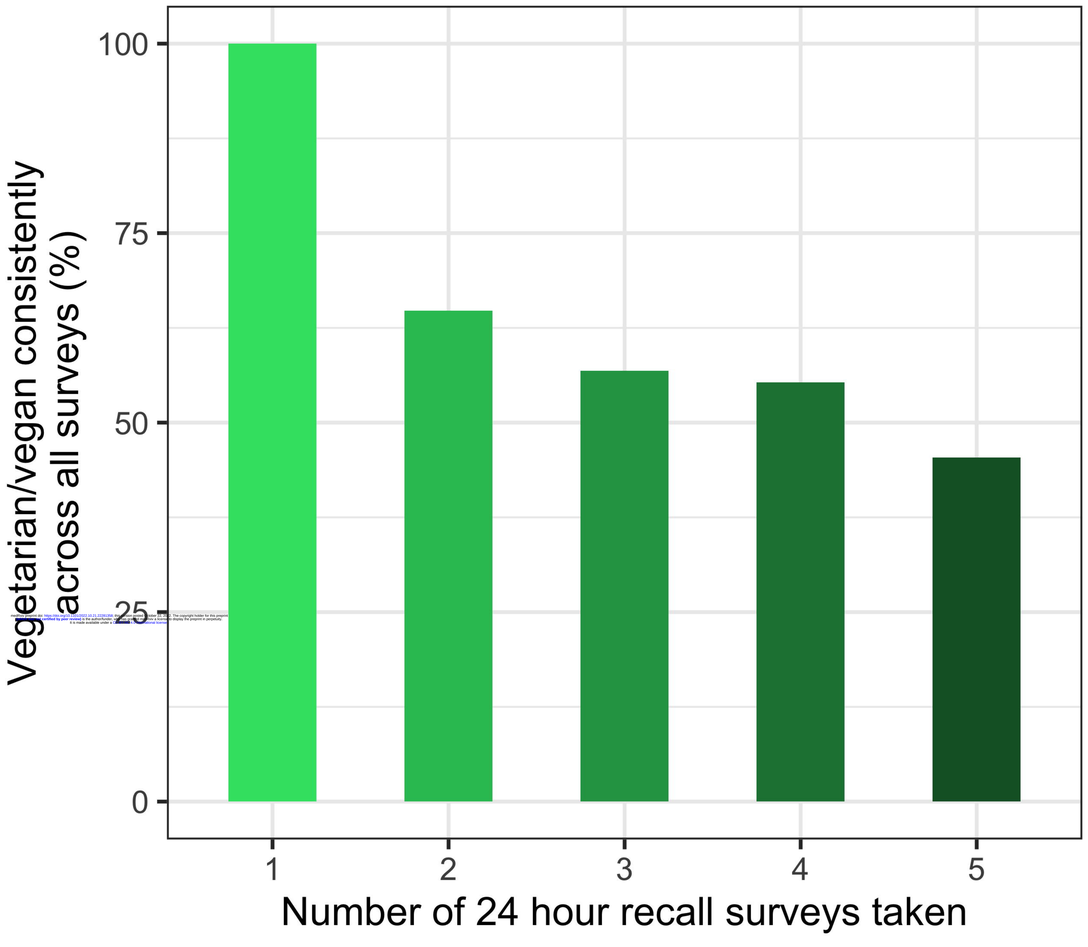
A


Number of participants
80000
60000
40000
20000
0
1
2
3
4
5
Number of 24 hour recall surveys taken

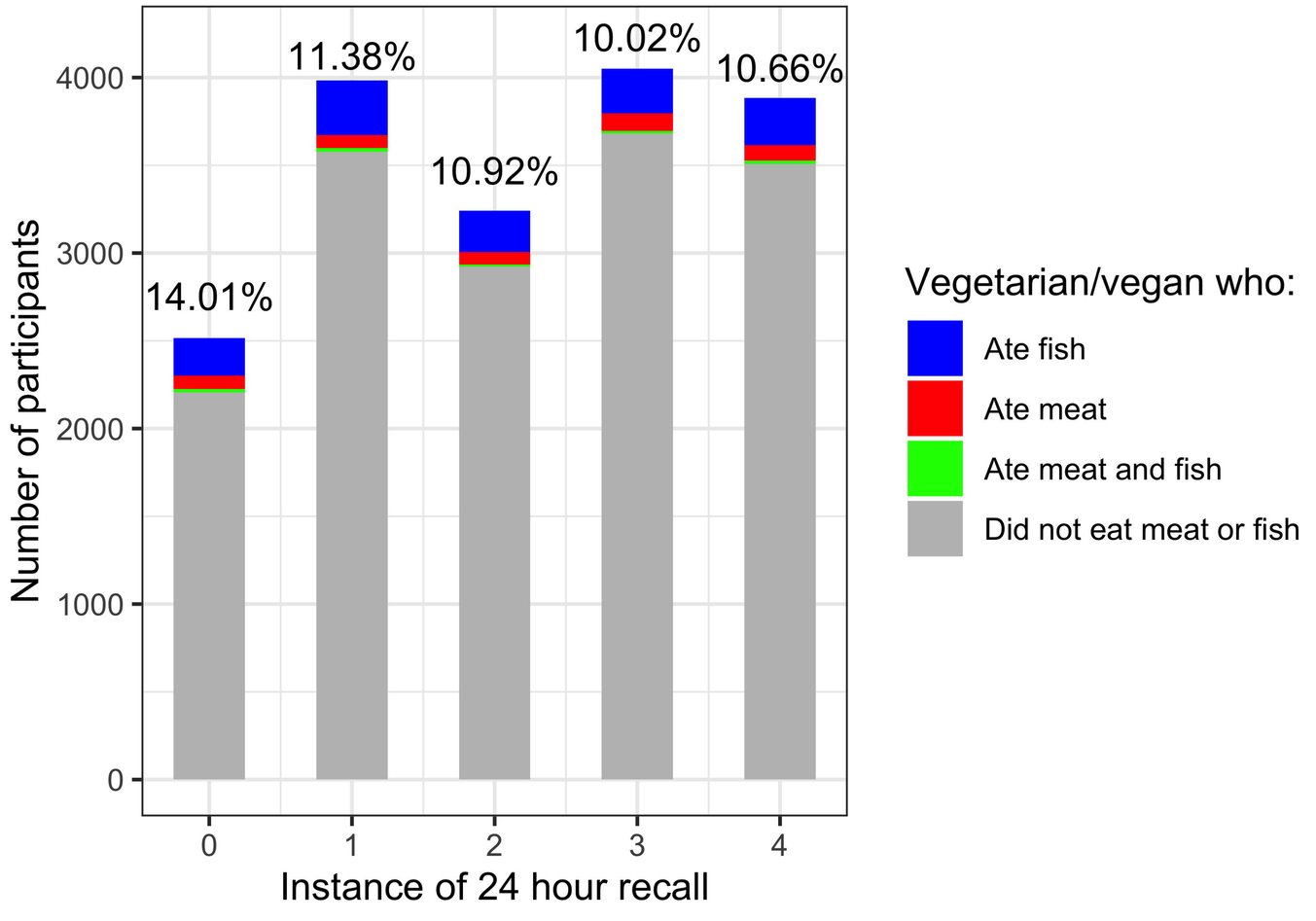
B


Count participants
100000
75000
50000
25000
0
0
1
2
3
4
Instance of 24 hour recall
First time taking
Second time taking
Third time taking
Fourth time taking
Fifth time taking

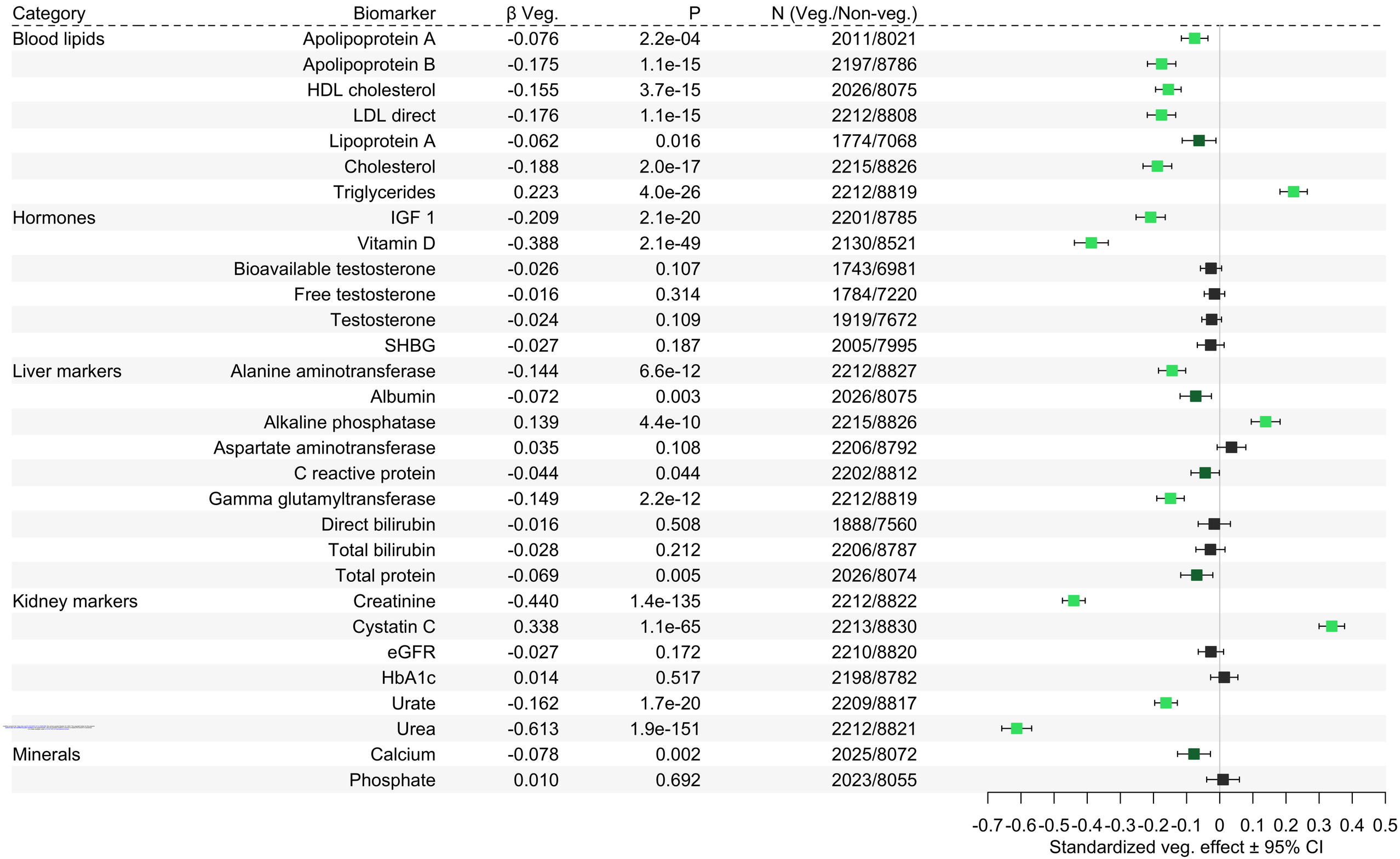
C

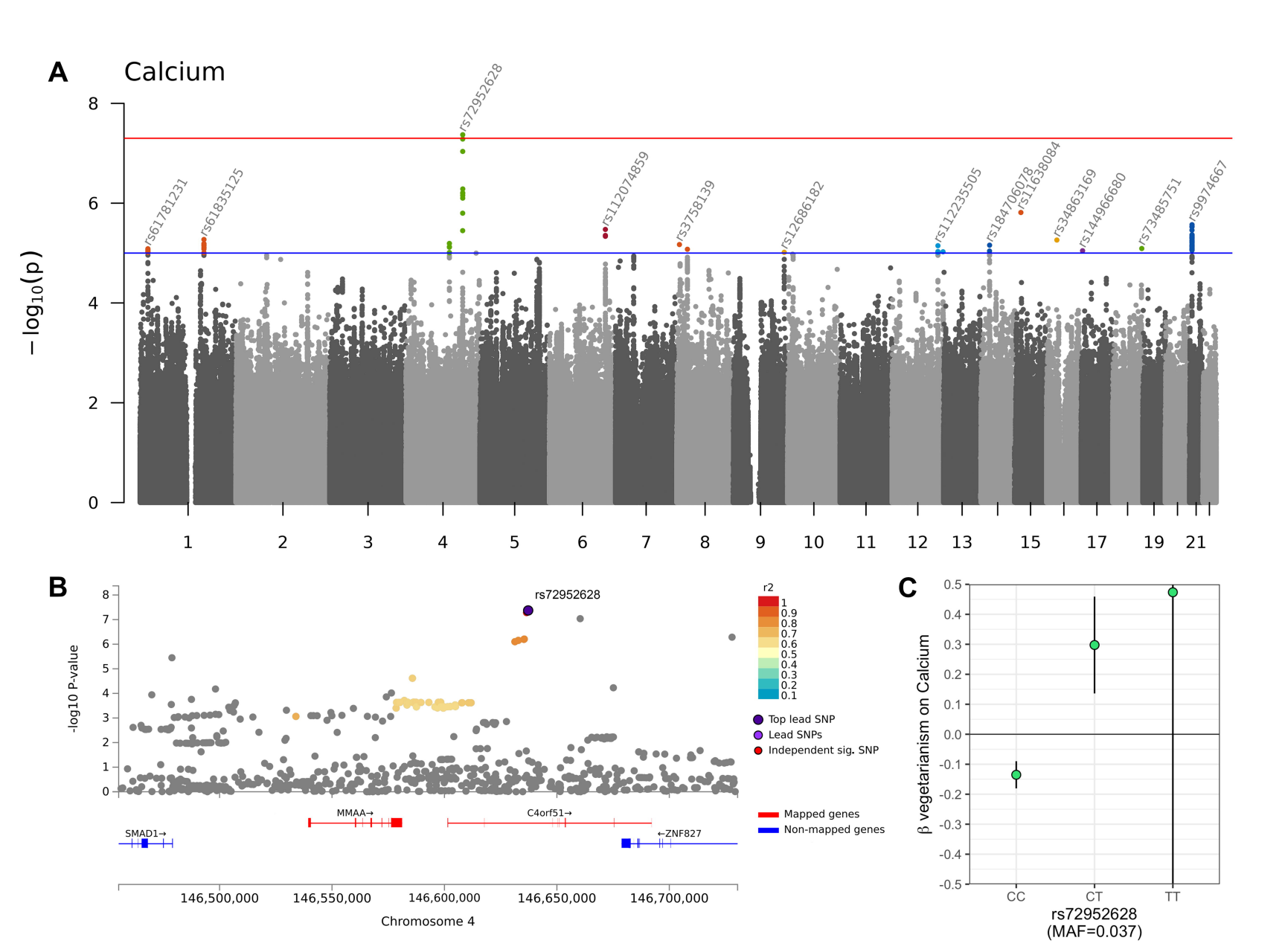

Vegetarian/vegan consistently across all surveys (%)
100
75
50
25
0
1
2
3
4
5
Number of 24 hour recall surveys taken

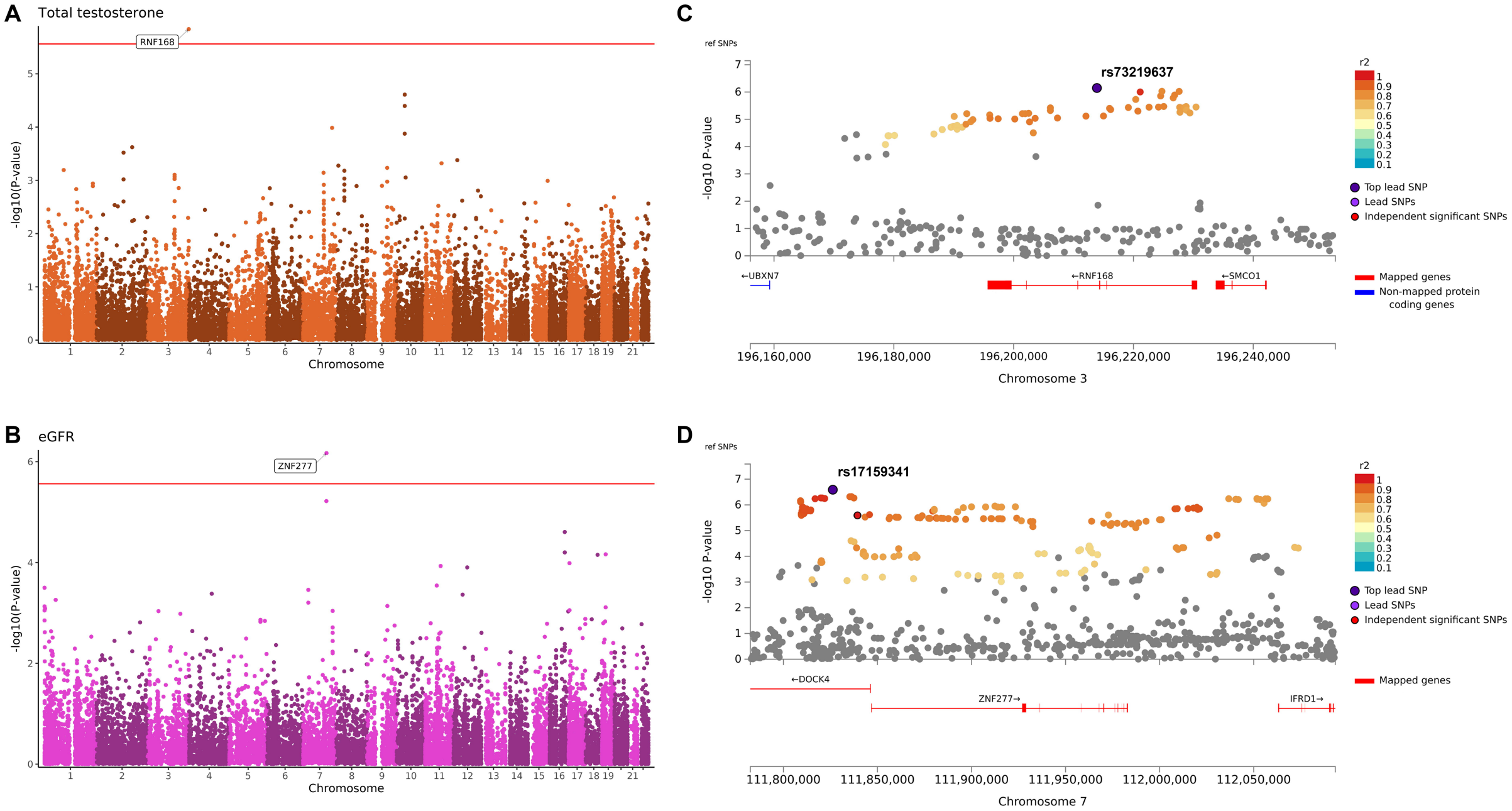
D


Number of participants
4000
3000
2000
1000
0
14.01%
11.38%
10.92%
10.02%
10.66%
0
1
2
3
4
Instance of 24 hour recall
Vegetarian/vegan who:
Ate fish
Ate meat
Ate meat and fish
Did not eat meat or fish

Category	Biomarker	β Veg.	P	N (Veg./Non-veg.)
Blood lipids	Apolipoprotein A	-0.076	2.2e-04	2011/8021
	Apolipoprotein B	-0.175	1.1e-15	2197/8786
	HDL cholesterol	-0.155	3.7e-15	2026/8075
	LDL direct	-0.176	1.1e-15	2212/8808
	Lipoprotein A	-0.062	0.016	1774/7068
	Cholesterol	-0.188	2.0e-17	2215/8826
	Triglycerides	0.223	4.0e-26	2212/8819
Hormones	IGF 1	-0.209	2.1e-20	2201/8785
	Vitamin D	-0.388	2.1e-49	2130/8521
	Bioavailable testosterone	-0.026	0.107	1743/6981
	Free testosterone	-0.016	0.314	1784/7220
	Testosterone	-0.024	0.109	1919/7672
	SHBG	-0.027	0.187	2005/7995
Liver markers	Alanine aminotransferase	-0.144	6.6e-12	2212/8827
	Albumin	-0.072	0.003	2026/8075
	Alkaline phosphatase	0.139	4.4e-10	2215/8826
	Aspartate aminotransferase	0.035	0.108	2206/8792
	C reactive protein	-0.044	0.044	2202/8812
	Gamma glutamyltransferase	-0.149	2.2e-12	2212/8819
	Direct bilirubin	-0.016	0.508	1888/7560
	Total bilirubin	-0.028	0.212	2206/8787
	Total protein	-0.069	0.005	2026/8074
Kidney markers	Creatinine	-0.440	1.4e-135	2212/8822
	Cystatin C	0.338	1.1e-65	2213/8830
	eGFR	-0.027	0.172	2210/8820
	HbA1c	0.014	0.517	2198/8782
	Urate	-0.162	1.7e-20	2209/8817
	Urea	-0.613	1.9e-151	2212/8821
Minerals	Calcium	-0.078	0.002	2025/8072
	Phosphate	0.010	0.692	2023/8055




A
Calcium
−log10(p)
rs72952628
rs61781231
rs61835125
rs112074859
rs3758139
rs12686182
rs112235505
rs184706078
rs116638084
rs34863169
rs144966680
rs73485751
rs9974667
1
2
3
4
5
6
7
8
9
10
11
12
13
15
17
19
21
B
rs72952628
-log10 P-value
r2
Top lead SNP
Lead SNPs
Independent sig. SNP
Mapped genes
Non-mapped genes
MMAA→
C4orf51→
SMAD1→
←ZNF827
146,500,000
146,550,000
146,600,000
146,650,000
146,700,000
Chromosome 4
C
β vegetarianism on Calcium
CC
CT
TT
rs72952628
(MAF=0.037)


A
Total testosterone
RNF168
-log10(P-value)
Chromosome
B
eGFR
ZNF277
-log10(P-value)
Chromosome
C
ref SNPs
rs73219637
-log10 P-value
r2
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Top lead SNP
Lead SNPs
Independent significant SNPs
←UBXN7
←RNF168
←SMCO1
Mapped genes
Non-mapped protein coding genes
196,160,000
196,180,000
196,200,000
196,220,000
196,240,000
Chromosome 3
D
ref SNPs
rs17159341
-log10 P-value
r2
Top lead SNP
Lead SNPs
Independent significant SNPs
←DOCK4
ZNF277→
IFRD1→
Mapped genes
111,800,000
111,850,000
111,900,000
111,950,000
112,000,000
112,050,000
Chromosome 7