Using genetics to disentangle the complex relationship between food choices and health status

Authors

Nicola Pirastu[1†], Ciara McDonnell[1,11*],Eryk J. Grzeszkowiak[1*], Ninon Mounier[2, 3], Fumiaki Imamura[3], Felix R. Day[4], Jie Zheng[5], Nele Taba[6,12], Maria Pina Concas[7], Linda Repetto[1], Katherine A. Kentistou[1,11], Antonietta Robino[7],Tõnu Esko[6,9],Peter K. Joshi[1], Krista Fischer[6], Ken K. Ong[4],Tom R. Gaunt[5], Zoltan Kutalik[2,3], John R. B. Perry[4], James F. Wilson[1,10].

Affiliations

1 Centre for Global Health Research, Usher Institute, University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, Scotland.

2. Center for Primary Care and Public Health, University of Lausanne, Lausanne, Switzerland

3. Swiss Institute of Bioinformatics, Lausanne, Switzerland

4. MRC Epidemiology Unit, Institute of Metabolic Science, Cambridge Biomedical Campus, University of Cambridge School of Clinical Medicine, Box 285, Cambridge, CB2 0QQ, UK

5. MRC Integrative Epidemiology Unit, Bristol Medical School, Bristol, UK

6. Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Riia 23b, 51010, Estonia

7. Institute for Maternal and Child Health - IRCCS “Burlo Garofolo”, Trieste, Italy

8. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Nobels väg 12A, SE-171 77 Stockholm, Sweden

9. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

10. MRC Human Genetics Unit, Institute of Genetic and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, EH4 2XU, Scotland

11. Centre for Cardiovascular Sciences, Queen's Medical Research Institute, University of Edinburgh, Royal Infirmary of Edinburgh, Little France Crescent, Edinburgh EH16 4TJ, Scotland

12. Institute of Molecular and Cell Biology, University of Tartu, Tartu, Riia 23, 51010, Estonia

*Authors contributed equally to this work.

†Correspondence should be addressed to Nicola Pirastu, nicola.pirastu@ed.ac.uk

Abstract.

Despite food choices being one of the most important factors influencing health, efforts to identify individual food groups and dietary patterns that cause disease have been challenging, with traditional nutritional epidemiological approaches plagued by biases and confounding. After identifying 302 (289 novel) individual genetic determinants of dietary intake in 445,779 individuals in the UK Biobank study, we develop a statistical genetics framework that enables us, for the first time, to directly assess the impact of food choices on health outcomes. We show that the biases which affect observational studies extend also to GWAS, genetic correlations and causal inference through genetics, which can be corrected by applying our methods. Finally, by applying Mendelian Randomization approaches to the corrected results we identify some of the first robust causal associations between eating patterns and risks of cancer, heart disease and obesity, distinguishing between the effects of specific foods or dietary patterns.

Introduction

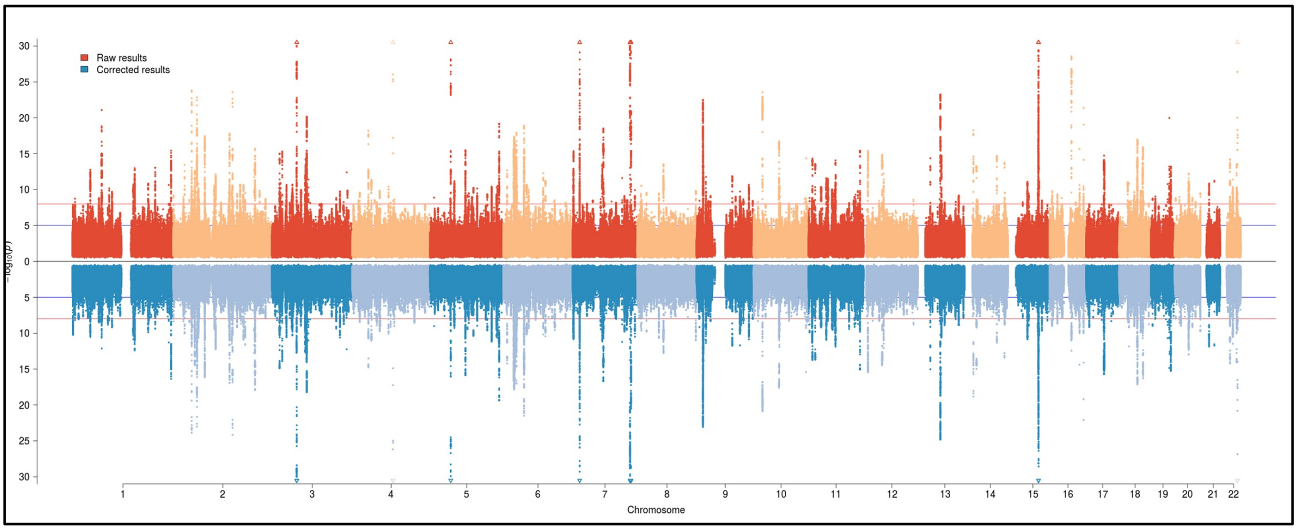
Given their profound impact on human well-being, nutritional choices and their impact on health are one of the most studied human behaviours. Quality and quantity of food consumption are associated with a wide range of medical conditions including metabolic syndrome and cardiovascular disease[1], cancer[1], liver disease[2], inflammatory bowel disease[3] and depression[4]. Food choice is becoming increasingly significant for global health as energy-dense, low fibre western diets proliferate across the globe and an obesity epidemic follows[4]. Despite the extremely high number of studies reporting food/health associations it has been hard to establish causal relationships due to difficulty in measurement, recall bias and confounding.

Recently, causal inference has been improved by a large number of studies which use Mendelian Randomization (MR) to assess the causal relationship between one or more exposures and outcomes. In MR, genetic variants are used as instrumental variables to measure the “life-long exposure” to a risk factor[5]. This technique has proven to be extremely powerful, not influenced by confounding typical of observational studies and many of the results have been mirrored by randomised controlled trials[5]. It is thus appealing to use MR to assess the causal relationship between food and health. Unfortunately, genetic variants predicting dietary consumption has been limited to a few food groups, such as alcoholic beverages[6] , coffee[7], milk[8,9], and existing evidence from dietary MR studies remain unremarkable[10,11]. More importantly, previous studies on a single food group have not accounted for interrelationships between different food groups. We therefore aimed to assess the causal relationship between food and several health outcomes by exploiting consumption patterns of multiple food groups in the UK Biobank (UKB) to create a new set of genetic instruments for MR analysis and then testing the causal effect of food consumption on health.[12]

GWAS of food traits

The first step in MR is to identify those genetic variants which are associated with the exposure of interest (food consumption in our case). We thus conducted a genome-wide association study (GWAS) on 29 food consumption traits, such as “beef” and “cheese” intake, using a mixed linear model in the white European participants of UKB[13] (up to N=445,779), including only sex and age as covariates to avoid collider bias[14] For a full description of the traits see Tables S1 and S2. The GWAS identified 414 phenotype-genotype associations divided into 260 independent loci with $p < 1 \times 10^{-8}$, summarized in Table S3 and Figure 1.

Fig. 1 Miami plot showing 302 independent loci associated with food choices. *Results for both univariate and multivariate analyses are included. For each SNP the lowest p-value for all traits was plotted. The upper panel represents the unadjusted GWAS associations while the lower panel represents the association with food choices, after adjustment for mediating traits, such as health status.*

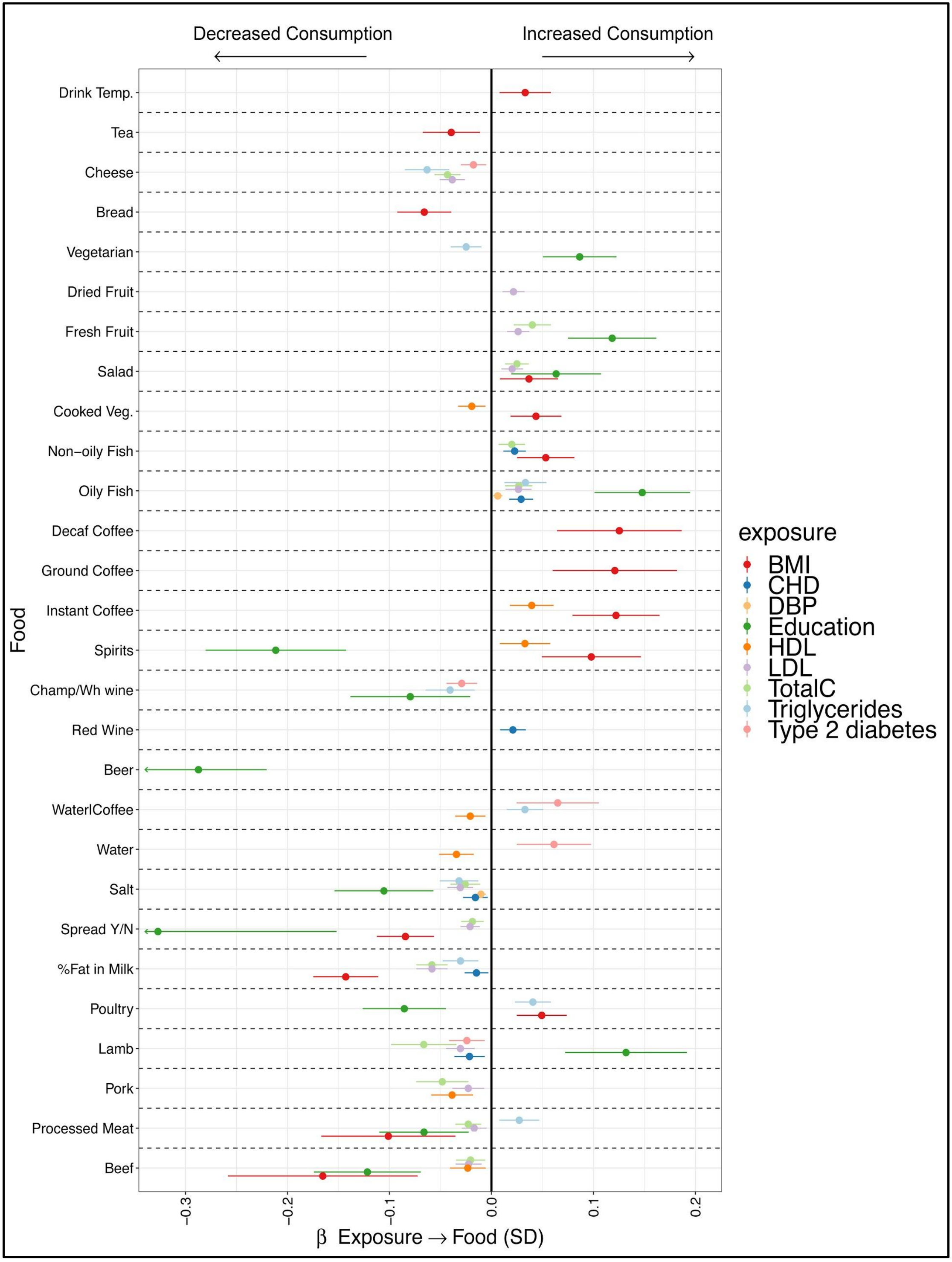


Replication for 23 of the 29 traits was sought in two additional UK based cohorts (EPIC-Norfolk[15] and Fenland[16]) totalling up to 32,779 subjects. Despite relatively limited power, we could nominally replicate 104/325 associations at p<0.05 (one-sided test) (32%; p=9.47x10^{-54}). The direction of effect was consistent with that for discovery in 268 of the 325 associations (82%; p=7.82x10^{-35}, Binomial test; see Table S5). After prioritization of the genes in each locus (see Methods for details and Supp. Table S4 for the prioritized genes), we noticed that for many genes associated with BMI, the BMI-raising allele was associated with lower reported consumption of energy-dense foods such as meat or fat and with higher consumption of lower-calorie foods. Although the exact mechanism of action of many of these genes is unknown, in the case of *MC4R* in mice loss-of-function K314X mutants show an increase in weight, higher intake of calories and higher preference for a high fat diet[17], while we observe a lower intake of fat and higher intake of fresh fruit. We thus wondered if this could be due to the effect of higher BMI on food choices instead of the reverse and if this effect might also occur for a broader range of health-related traits.

Detecting the effects of potential confounders on food frequency data

To test this hypothesis, we first selected nine diseases and risk factors for which dietary advice is usually given and for which GWA summary statistics (from large meta-analyses not including UKB) were available. Educational attainment was also included as a proxy for socioeconomic status. Using MR we identified 81 instances where we had evidence of health-related traits significantly influencing food choice (Fig. 2).

Fig 2. Health status influences reported food choices. *The plot reports only the univariable MR results which were significant at FDR<0.05. For each food outcome the effect estimate (β) is reported in standard deviations of the exposure trait, together with 95% confidence intervals. Each colour represents a different exposure. BMI, body mass index; CHD, coronary heart disease; DBP, diastolic blood pressure; HDL, high density lipoprotein cholesterol; LDL, low density lipoprotein cholesterol; TotalC, total cholesterol. Champ/Wh wine, champagne, white wine. Temp, temperature.*



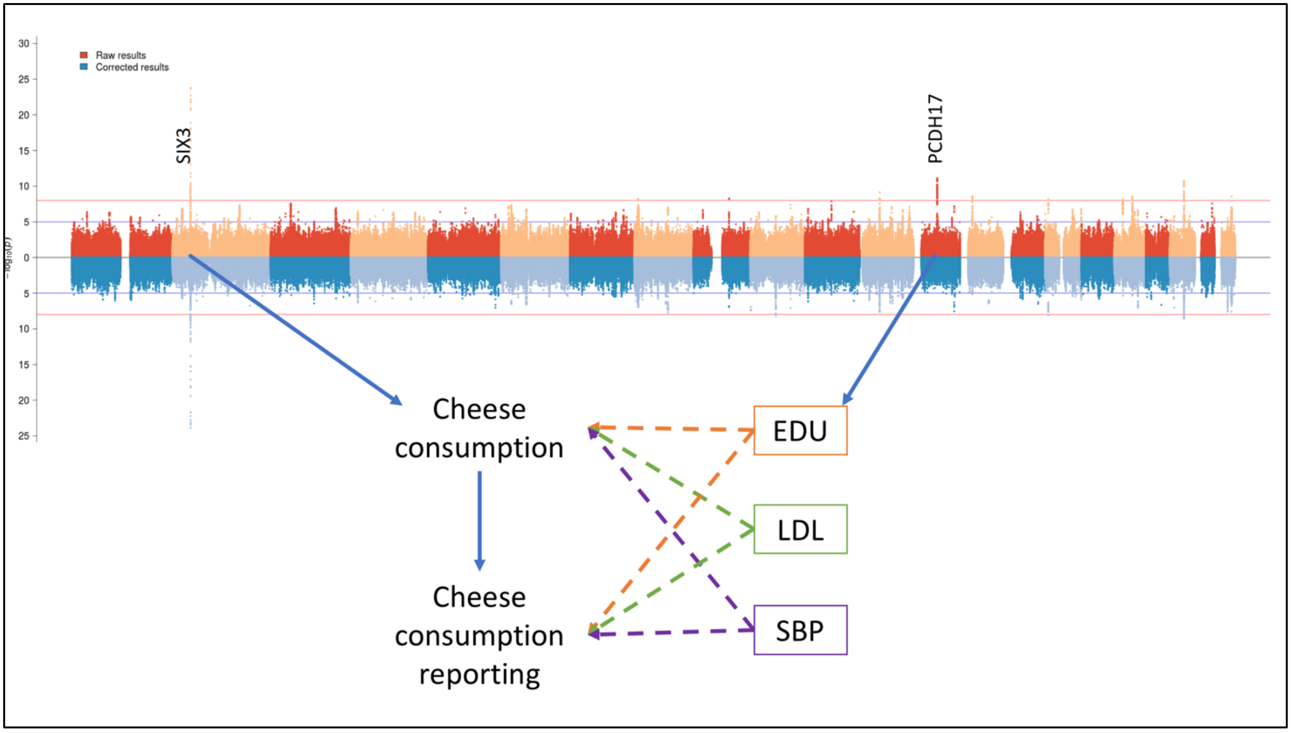
Aside from educational attainment, many associations seem to reflect common nutritional advice. For example, higher genetically-determined BMI associates with higher consumption of poultry, vegetables (both raw and cooked), non-oily fish, (also spirits and coffee); but less beef, processed meat, bread and fatty foods. Similarly, those genetically predisposed to CHD report lower consumption of whole milk, salt and lamb; and higher consumption of fish and red wine. This last case is particularly interesting, reflecting the standard dietary advice (lower intake of fat and salt but higher intake of fish as a means to increase omega-3 fatty acid intake[18]), but also higher consumption of red wine (and not other alcoholic beverages), which is commonly believed to have cardioprotective effects[19,20].

From these MR results, it is clear that some of the loci we have identified in GWAS are not directly associated with food consumption but are the result of the effect of the health-related phenotypes on food consumption. Although we commonly consider the food-health relationship with diet as the exposure and disease as the outcome, we must consider that humans may change their behaviour because of their health status. This reverses the expected cause and effect relationship, making the interpretation of the GWAS results complex.

Correcting biases in dietary GWAS

To address the possibility of mediated effects, it is common to add the potential mediators as covariates in the association model. However, adding heritable covariates may lead to spurious associations due to collider bias (i.e. the false association between two variables induced by including a third variable (the collider) in the regression model, to which both variables of interest are causal)[14]. Moreover, when the causal relationship is bidirectional, adding a covariate will correct for the overall effect and not for the unidirectional effect we actually want to correct for.

Fig. 3 Direct and indirect SNP effects. *The plot shows the causal path of exemplar genes identified for cheese consumption. In the multivariable MR model cheese consumption is causally influenced by educational attainment (EDU), low density lipoprotein cholesterol levels (LDL) and systolic blood pressure (SBP). The effect of PDCH17 and is mediated through educational attainment, while SIX3 has a direct effect on cheese consumption. The mediated effects cannot be used reliably as MR instruments as they could be affecting either consumption or its reporting. Moreover, they could act as confounders in the MR analysis and thus they need to be identified.*



We thus developed a new MR-based approach to correct the effect of each SNP in the dietary GWAS for the effect mediated through other confounding traits. Briefly, our approach consists of two steps: the first is to fit a multivariable MR model to estimate the effects of the traits we would like to test (the health-related traits in our case) on the traits of interest (the food traits). For each SNP, then an expected mediated effect is calculated, based on the effect of the SNP on the mediator traits. The expected effect is then subtracted from the observed one to get an adjusted estimate (see Methods for details). This last step is exactly analogous to estimating the direct effect in mediation analysis[21].

We applied this method to all 29 food traits. As potential mediators, we used the traits tested in the univariate models, to which we added Crohn's disease and ulcerative colitis, as they may impact dietary choices after diagnosis. We also removed total cholesterol to avoid problems due to collinearity with LDL and HDL cholesterol. Looking at the exposure traits selected for the multivariable (MV) causal model of each food trait (Supplementary Fig S3 panel A and Supplementary Table S8), educational attainment plays a fundamental role in shaping food choices, significantly influencing over half of the traits, as does BMI. Looking at the percentage of the genetic variance of the food traits explained by the health-related traits (Supplementary Fig S3 panel B and Supplementary Table S16), it ranges from 42% for cheese to ~0% for fortified wine and white wine/champagne, highlighting the scope these effects have to influence GWAS results. The combined results from all traits before and after adjustment for the effect of health status on food preference are shown in Fig. 1 (see Supplementary file 1 for trait-specific plots). In many loci previously associated with health-related traits, the effect changed dramatically, suggesting that the effect of the SNP on the food traits is mediated through health status. For example, the effect size of the lead *FTO* variant (rs55872725) with percentage fat in milk reduces by three-fold from 0.0045 to 0.0015 log units (p=2x10^{-29} and p=7x10^{-5}, respectively). We observed similar effects for other associations at the same locus, which suggests that in general the associations we are observing near *FTO* are primarily mediated through its strong association with BMI[22]. This insight is crucial to understanding: a naïve approach would interpret that eating less healthy foods and more calorie-dense foods would lead to a lower BMI, while in fact, our analysis suggests that it is having a higher BMI that leads to either having a healthier diet or reporting one. This accords with known biases in a dietary assessment[23]. Unfortunately, we cannot distinguish between a change in behaviour (and thus indication bias) or such reporting bias. These results warrant even greater caution in using SNPs influencing diet in MR or for functional follow up studies. Moreover, most nutritional epidemiological studies have focused only on BMI and socioeconomic status for correction, while we show that the confounding effects extend to many other health traits such as blood pressure and lipids. The widespread effect of education and BMI on dietary choices is especially strong on cheese and percentage fat in milk. This may explain some of the recent epidemiological results linking dairy product consumption to positive health benefits[24].

To further explore the effects of the correction procedure, we compared the correlation patterns between the food traits and 832 phenotypes present in the LD hub[25] database using the raw and corrected results (See Supplementary Data 2.3 and additional table S10). These analyses showed that the correction produced more meaningful food clusters and that in many cases the genetic correlations with other traits changed greatly (see https://npirastu.shinyapps.io/rg_plotter_2/ for a graphical representation of these results). For example, if we look at the relationship of the two fat intake traits (percentage fat in milk and adding spread to bread) and body fat percentage we can see that they both have a seemingly beneficial effect before correction (r_G = -0.43 and -0.10, respectively) which diminishes to near zero (r_G = -0.04 and 0.07) after applying the correction, suggesting that the apparent protective effect is likely due to confounding.

Clustering of food items

To investigate how the mediation procedure affected the genetic correlations amongst the consumption traits and with other traits, we first compared the clustering based on the uncorrected and adjusted genetic correlations. Figure S7 panel A shows the tanglegram comparing the two analyses. The adjusted correlations give more reasonable groupings, showing that some of the unadjusted clusterings are due in part to common confounders (e.g. wine clustering closer to coffee than other alcoholic beverages) than actual common genetic background.

Clustering of the food traits based on their corrected genetic associations using ICLUST identified five different food groups (Fig S7 panel B): one composed of increased meat, fat, salt and decreased vegetarianism (labelled as “Meat/Fat”), one made up of alcoholic beverages and coffee (labelled “Psychoactive drinks”) and one comprised of healthier items such as fish, fruit and vegetables (labelled “Low-Calorie Foods”). Two final groups contained only two items each: drink temperature and tea; and cheese and bread; these were not used for the MV analysis. In order to explore if additional loci influence these groups, we ran a multivariate GWAS using the package MultiABEL, which performs MANOVA on summary statistics. 168 additional associations, including 42 novel loci not identified in the single-trait analysis, were identified in multivariate analysis of the three main food groups (Table S5).

Selection of instruments for MR

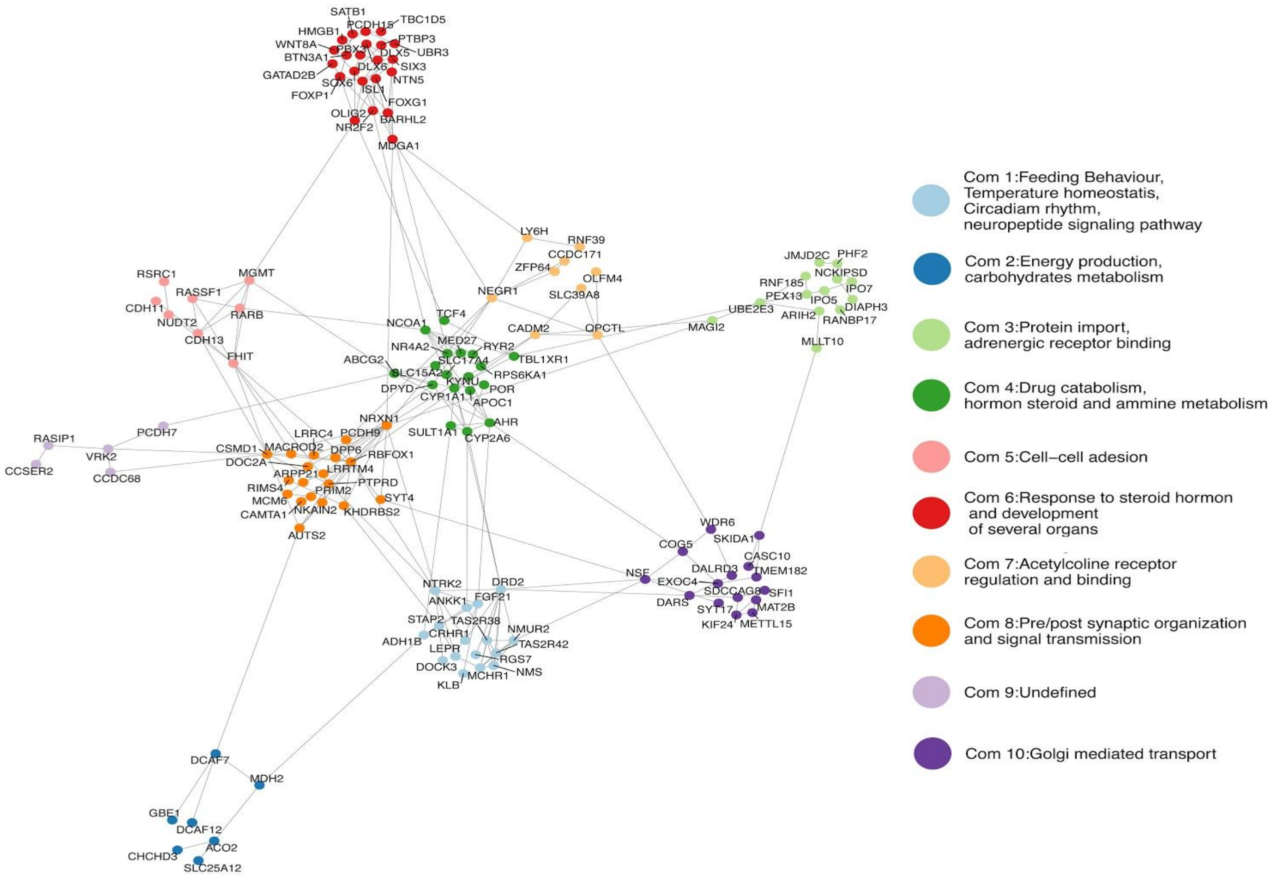
The primary objective of our study is to use MR to assess causal relationships between food choices and health. To achieve this goal we need to be able to identify the SNPs which have only a direct effect on the food trait, which is not mediated through other possible confounders. We hypothesised that if a SNP is biologically associated with a food behaviour - without mediation by health - its effect should not change strongly after the adjustment procedure. To try to distinguish the variants with only a direct effect from those with effects at least partly mediated through other traits, we defined the corrected-to-raw ratio (CRR) as the ratio between the corrected effect and the raw uncorrected one.

Through extensive simulations we estimated that the CRR range between 0.95 and 1.05 maximises this probability, with 88% of the SNPs being directly associated with the trait of interest (see Supplementary Data 2.1 for details on the simulations and Supplementary Data 1.8 for theory). Further evidence comes from variants in alcohol dehydrogenase 1B and the taste and olfactory receptors (for which clear biological pathways can be defined): all have CRR values between 0.95-1.05. We thus defined SNPs with a CRR in this range as “non-mediated”. 387 out of 581 associations corresponding to 208/302 loci (~69%) were categorised as non-mediated associations, although of these 50 showed both mediated and non-mediated effects. The balance of mediated to non-mediated SNP associations varied by foodstuff, ranging from none mediated for tea, spirits and processed meat to all mediated for percentage fat in milk and adding spread to bread (see Table S3). The necessity of using the CRR filtering instead of existing methods is further outlined in additional paragraph 2.7.

Functional annotation of the direct-effect-only loci and tissue enrichment analysis prominently feature brain areas involved in reward (Supplementary Data 2.5). Inference of interaction networks reveals ten communities ranging from feeding behaviour and energy metabolism to steroid response, acetylcholine receptor regulation and synaptic transmission (Supplementary Data 2.6 and Figure. 4).

Fig 4. STRING network of genes in non mediated loci*. Network plot of the genes in the non-mediated loci. After performing community detection we identified ten different clusters of genes each with its particular set of functions and*

expression patterns (see additional paragraph 2.6 for details). Nodes have been colored according to community membership.



Causal inference

We proceeded to perform two-sample MR using the food traits as exposures and 78 traits (see table S17 for a list and description) as outcomes (chosen to include those for which diet could be a causal factor, that were in MR-base and for which full GWAS summary statistics were available). As well as using each single food trait as exposures, we also assessed the effect of 16 different principal components (PC)-derived phenotypes based on the previous clustering of food traits, to quantify the consequences of broader dietary patterns. The relationships between the different traits are reported in figure S2 while loadings for each PC trait are reported in Fig 5 panel A. Traits which had no direct-effect-only SNPs (percentage fat in milk, fortified wine and adding spread to bread) were left out of the analysis. For each exposure-outcome pair, four types of analyses were performed, selecting instrumental variables with or without filtering by CRR or using corrected or uncorrected betas. We considered as the main analysis the CRR-filtered analysis using uncorrected betas and used the others for comparison. Finally we considered as significant the exposure-outcome pairs after multiple test correction of the main analysis using Storey's q-value at q<0.05. Table 1 reports the significant results, while all results can be found in table S18 and are available through a shiny app https://npirastu.shinyapps.io/Food_MR/.

Fig 5. Significant effects of food choice on disease. (a) *Heatmap of the loadings of each food trait on the PC traits. Red reflects a positive loading while blue a negative one.* ***(b)*** *Network representation of all the significant exposure-outcome pairs. The green nodes represent the food traits used as exposures while the yellow ones represent the outcome traits. Arrows represent the causal relationships detected through the MR analysis, they are directed to reflect the exposure -> outcome relationship and the colour reflects the direction of effect: blue, decrease; red, increase. Clearly, All PC1 (which reflects what is generally considered a healthy vs unhealthy diet) is the trait with most putatively causal associations, which range from an improved blood lipid profile to protection from both myocardial infartion and lung cancer. Blood triglyceride (TG) levels seem to be the outcome influenced by the largest number of food traits, being lowered by All PC2 and PC3, Healthy PC1, Fruit PC1, and Oily fish. Abbreviations: WC, waist circumference; Hip, hip circumference; CHD, coronary heart disease; Hb, Hemoglobin concentration; MI, myocardial infarction; LDL, low density lipoproteins; TC, total cholesterol; Serous ovarian cancer (1), High grade and low grade serous ovarian cancer; Serous ovarian cancer (2), Serous ovarian cancer: low grade and low malignant potential; Edu, Educational attainment; BMI, body mass index; WHR, waist to hip ratio; WHR|BMI, waist to hip ratio BMI adjusted; PLT, platelet; Celiac, celiac disorder.*

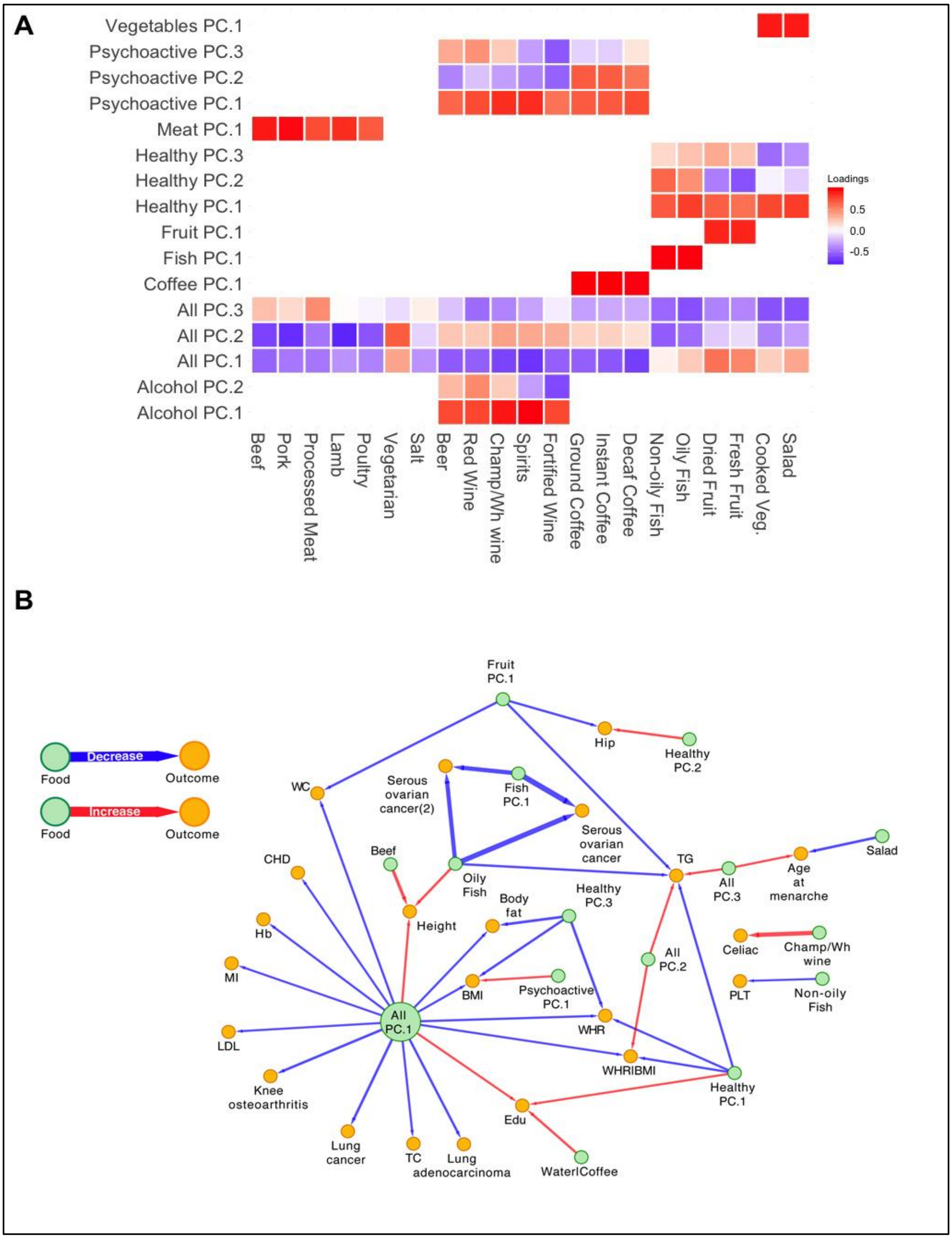


Table 1. Significant Food-Outcome relationships. *Results are presented for the associations with FDR<0.05. The Method column refers to the primary analysis method (either IVW fixed effect (FE) or random effect (RE) or Wald ratio in case of a single SNP IV). The other columns report effect sizes, standard errors and p-values for the main analysis and the two methods used as sensitivity analyses (MR-RAPS and MR Median). Finally the p-value for heterogeneity in the main analysis is reported.*

Exposure	Outcome	Method	N SNPs	IVW (wald ratio)		MR-RAPS		MR Median		Heterogeneity p-value
				beta (se)	p-value	beta (se)	p-value	beta (se)	p-value	
All PC.1	Body fat	IVW (FE)	123	-0.08 (0.022)	3.2E-04	-0.074 (0.028)	7.5E-03	-0.02 (0.035)	5.7E-01	1.1E-03
All PC.1	BMI	IVW (RE)	120	-0.087 (0.021)	8.1E-05	-0.087 (0.021)	4.2E-05	-0.056 (0.022)	1.3E-02	1.5E-12
All PC.1	CHD	IVW (FE)	128	-0.059 (0.016)	2.2E-04	-0.065 (0.019)	5.6E-04	-0.066 (0.027)	1.5E-02	2.2E-02
All PC.1	Hb	IVW (FE)	124	-0.074 (0.021)	6.7E-04	-0.071 (0.027)	8.3E-03	-0.066 (0.035)	6.1E-02	4.5E-03
All PC.1	Height	IVW (RE)	117	0.094 (0.025)	2.2E-04	0.092 (0.028)	9.7E-04	0.122 (0.026)	2.2E-06	2.0E-19
All PC.1	Knee osteoarthritis	IVW (FE)	122	-0.257 (0.067)	1.8E-04	-0.271 (0.078)	5.4E-04	-0.259 (0.105)	1.3E-02	1.9E-01
All PC.1	LDL	IVW (FE)	121	-0.061 (0.017)	6.4E-04	-0.062 (0.02)	1.8E-03	-0.057 (0.029)	4.9E-02	1.7E-01
All PC.1	Lung adenocarcinoma	IVW (FE)	128	-0.176 (0.05)	6.2E-04	-0.188 (0.056)	8.2E-04	-0.133 (0.086)	1.2E-01	1.4E-01
All PC.1	Lung cancer	IVW (FE)	127	-0.278 (0.044)	3.5E-09	-0.287 (0.054)	1.1E-07	-0.275 (0.074)	2.0E-04	1.6E-02
All PC.1	MI	IVW (FE)	128	-0.056 (0.016)	6.7E-04	-0.055 (0.02)	6.0E-03	-0.049 (0.028)	8.3E-02	1.4E-02
All PC.1	TC	IVW (FE)	121	-0.07 (0.017)	6.0E-05	-0.063 (0.019)	1.2E-03	-0.05 (0.028)	7.3E-02	3.7E-02
All PC.1	WC	IVW (RE)	123	-0.113 (0.025)	1.5E-05	-0.122 (0.022)	5.4E-08	-0.071 (0.03)	1.9E-02	1.4E-06
All PC.1	WHR	IVW (RE)	124	-0.104 (0.021)	2.4E-06	-0.109 (0.021)	3.4E-07	-0.074 (0.027)	6.7E-03	2.0E-04
All PC.1	WHR | BMI	IVW (RE)	124	-0.078 (0.022)	4.9E-04	-0.08 (0.022)	3.6E-04	-0.069 (0.026)	8.3E-03	3.5E-06
All PC.1	Edu	IVW (RE)	123	0.086 (0.019)	1.3E-05	0.084 (0.018)	2.7E-06	0.059 (0.022)	7.2E-03	6.0E-05
All PC.2	TG	IVW (FE)	114	0.092 (0.022)	6.2E-05	0.077 (0.031)	1.3E-02	0.023 (0.036)	5.2E-01	6.7E-04
All PC.2	WHR | BMI	IVW (FE)	116	0.116 (0.02)	3.7E-08	0.108 (0.026)	3.6E-05	0.093 (0.03)	2.3E-03	5.6E-03
All PC.3	Age at menarche	IVW (FE)	117	0.118 (0.026)	1.3E-05	0.108 (0.034)	1.5E-03	0.093 (0.041)	2.1E-02	7.0E-04
All PC.3	TG	IVW (FE)	118	0.147 (0.028)	6.3E-07	0.151 (0.037)	3.9E-05	0.15 (0.047)	1.4E-03	4.9E-03
Beef	Height	IVW (FE)	2	0.516 (0.114)	6.4E-06	NA (NA)	NA	NA (NA)	NA	3.8E-01
Champ/Wh wine	Celiac	Wald ratio	1	1.129 (0.326)	5.3E-04	NA (NA)	NA	NA (NA)	NA	NA

Fish PC.1	*Serous ovarian cancer*	*Wald ratio*	*1*	*-1.7 (0.436)*	*9.8E-05*	*NA (NA)*	*NA*	*NA (NA)*	*NA*	*NA*
Fish PC.1	*Serous ovarian cancer(2)*	*Wald ratio*	*1*	*-1.146 (0.327)*	*4.6E-04*	*NA (NA)*	*NA*	*NA (NA)*	*NA*	*NA*
Fruit PC.1	*Hip*	*IVW (FE)*	*31*	*-0.13 (0.034)*	*6.3E-04*	*-0.113 (0.04)*	*4.5E-03*	*-0.093 (0.052)*	*7.4E-02*	*1.2E-01*
Fruit PC.1	*TG*	*IVW (FE)*	*32*	*-0.142 (0.038)*	*7.4E-04*	*-0.154 (0.045)*	*5.7E-04*	*-0.15 (0.057)*	*8.9E-03*	*1.6E-01*
Fruit PC.1	*WC*	*IVW (FE)*	*32*	*-0.162 (0.034)*	*3.8E-05*	*-0.155 (0.046)*	*6.9E-04*	*-0.163 (0.054)*	*2.4E-03*	*1.6E-02*
Healthy PC.1	*TG*	*IVW (FE)*	*58*	*0.143 (0.029)*	*6.6E-06*	*0.14 (0.036)*	*1.2E-04*	*0.095 (0.047)*	*4.3E-02*	*2.6E-02*
Healthy PC.1	*WHR*	*IVW (FE)*	*58*	*0.115 (0.026)*	*3.3E-05*	*0.112 (0.034)*	*8.2E-04*	*0.122 (0.042)*	*4.1E-03*	*1.3E-02*
Healthy PC.1	*WHR | BMI*	*IVW (FE)*	*58*	*0.126 (0.026)*	*8.0E-06*	*0.116 (0.033)*	*3.6E-04*	*0.11 (0.041)*	*7.7E-03*	*1.7E-02*
Healthy PC.1	*Edu*	*IVW (FE)*	*59*	*-0.079 (0.022)*	*7.1E-04*	*-0.072 (0.03)*	*1.5E-02*	*-0.096 (0.038)*	*1.0E-02*	*4.9E-03*
Healthy PC.2	*Hip*	*IVW (FE)*	*58*	*0.197 (0.037)*	*2.3E-06*	*0.174 (0.053)*	*1.0E-03*	*0.141 (0.06)*	*2.0E-02*	*9.4E-04*
Healthy PC.3	*Body fat*	*IVW (FE)*	*57*	*-0.338 (0.089)*	*3.8E-04*	*-0.339 (0.119)*	*4.2E-03*	*-0.282 (0.13)*	*3.0E-02*	*2.5E-02*
Healthy PC.3	*BMI*	*IVW (FE)*	*50*	*-0.197 (0.052)*	*3.8E-04*	*-0.167 (0.074)*	*2.5E-02*	*-0.202 (0.083)*	*1.5E-02*	*5.9E-03*
Healthy PC.3	*WHR*	*IVW (FE)*	*57*	*-0.218 (0.06)*	*5.9E-04*	*-0.195 (0.089)*	*2.8E-02*	*-0.211 (0.095)*	*2.6E-02*	*2.7E-03*
Non-oily Fish	*PLT*	*IVW (FE)*	*2*	*-0.016 (0.004)*	*9.2E-05*	*NA (NA)*	*NA*	*NA (NA)*	*NA*	*5.1E-01*
Oily Fish	*Height*	*IVW (FE)*	*21*	*0.196 (0.035)*	*1.6E-05*	*0.177 (0.054)*	*9.6E-04*	*0.174 (0.053)*	*1.1E-03*	*8.4E-04*
Oily Fish	*Serous ovarian cancer*	*Wald ratio*	*1*	*-1.518 (0.39)*	*9.8E-05*	*NA (NA)*	*NA*	*NA (NA)*	*NA*	*NA*
Oily Fish	*Serous ovarian cancer(2)*	*Wald ratio*	*1*	*-1.02 (0.291)*	*4.6E-04*	*NA (NA)*	*NA*	*NA (NA)*	*NA*	*NA*
Oily Fish	*TRG*	*IVW (FE)*	*21*	*-0.175 (0.042)*	*5.1E-04*	*-0.156 (0.056)*	*5.2E-03*	*-0.084 (0.061)*	*1.7E-01*	*7.3E-02*
Psyco PC.1	*BMI*	*IVW (FE)*	*21*	*-0.064 (0.016)*	*8.5E-04*	*-0.058 (0.024)*	*1.7E-02*	*-0.047 (0.024)*	*5.1E-02*	*2.0E-03*
Salad	*Age at menarche*	*IVW (FE)*	*14*	*-0.298 (0.065)*	*5.3E-04*	*-0.28 (0.079)*	*4.2E-04*	*-0.251 (0.095)*	*8.5E-03*	*1.7E-01*
Water|Coffee	*Edu*	*IVW (FE)*	*24*	*0.162 (0.035)*	*1.3E-04*	*0.169 (0.048)*	*4.4E-04*	*0.162 (0.052)*	*1.7E-03*	*8.4E-03*

Looking at the significant MR results, we detected no sign of directional pleiotropy using the MR-Egger test (results in table S18). In some cases, we did detect strong heterogeneity of effect, especially with All PC1 and in general with PC-food exposures which included several diverse items. Considering more specific results, all PC.1 differentiates those eating more meat and salt while drinking more alcohol and coffee from those who eat more fruit and vegetables, thus it describes a general healthy-unhealthy diet continuum. All PC1 showed the largest number of associations (15; Fig.S22a), with a healthy value of All PC1 lowering most risk factors linked to obesity and lipid profile (and likely consequently lowering cardiovascular disease risk) and having a positive effect on height and education. With the exception of educational attainment, these results may not be surprising as they broadly overlap with general dietary advice. However, when we decompose these effects into food groups or single foods, we detect differences amongst traits. For example, All PC 1 leads to very similar effects across different obesity/adiposity measures : body fat % (β=-0.080,p=3.2x10^{-4}), body mass index (β= -0.087,p=8.1x10^{-5}), waist-to-hip ratio (=-0.104, p=2.4x10^{-6}) and BMI-adjusted waist-to-hip ratio (β=-0.078,p=2.9x10^{-4}). Figure S23 shows the comparative effects of each food on the four obesity measures: generally, the individual foods affect all four in very similar ways showing that the estimates are stable regardless of the outcome. However, there are some exceptions, for example, both Fresh Fruit and Oily Fish affect Body Fat and both waist:hip ratio measures but not BMI, suggesting that their effect is specifically on adiposity and not body size.

As a whole, alcohol does not seem to impact any of the four obesity traits, with a very small effect on waist-to-hip ratios. However, looking at each alcoholic beverage individually, beer has a substantial and specific effect on BMI not seen for the other alcoholic beverages, suggesting that this effect is independent of alcohol content.

Another notable result is the association of oily fish consumption with height (β= 0.2, p=1.76x10^{-8}) (Fig S22c). It is unclear, however, if this is the result of general healthy eating or if it is the effect of a specific food. In particular if we look at the effects of *All PC1-3*, we see that a height-raising of *PC1* (higher healthy foods, less alcohol/coffee and meat β= 0.09, p=1.35x10^{-4}), a height-lowering effect *PC2* (lower healthy food and meat and higher alcohol/coffee β= -0.1, p=1.34x10^{-3}), but no effect of *PC3* (higher meat and less alcohol/coffee and healthy foods β=-0.02, p=0.65) suggesting that the effect on height is lead by healthy foods and alcohol/coffee but independent of meat. Looking at the associations of *Healthy PC1-3*, we see association only with the first which represents the overall consumption of fish, fruit and vegetables. Finally, comparing these three we find that both higher consumption of vegetables and fish are associated with being taller, with similar effect sizes (*Fish PC1*, β=0.17, p=4.99x10^{-4} and *Vegetables PC1*, β=0.15, p=1.30x10^{-3}), while fruit has no effect (β= 0, p= 0.96), which makes the effects of fish and vegetables indistinguishable.

Several associations seem to be masked by the confounding effects, for example if we look at genetically-determined beef intake, the CRR-corrected instruments show a significant association with being taller (β= 0.51 SD adjusted vs. β= -0.01 unadjusted) and with other anthropometric traits such as hip and waist circumference. None of these associations were recovered using the raw instruments with estimated effects extremely close to 0, showing that the problems arising from using the unadjusted set of instruments are not limited to false positive results but also can generate false negatives, depending on the biases involved.

Discussion

Our results emphasise how complicated relationships among dietary traits are. We have clearly shown that the causal path between food and health is not unidirectional and that in fact genes may affect food behaviours in many different and unexpected ways. Understanding the origins of these effects is fundamental not only for prioritizing loci for functional follow up, but also for understanding why genetic correlations and GWAS results change when different datasets or populations are used. In fact, given that many of the effects we see are likely due to confounding, if the health advice in different populations changes this could alter the architecture of the studied trait and thus the GWAS results, which would appear as allelic heterogeneity.

It is unclear whether these effects are limited to dietary phenotypes or if they extend to other traits and further studies are needed to resolve this issue. Recent similar studies[10,11] on the genetic bases of dietary patterns reported having detected no reverse causality. We believe that this difference is due to our novel approach, which is not based on using the potential confounders as covariates, but rather exploits MR, which should be able to distinguish the forward and reverse effects when the causal relationship is bidirectional. Nevertheless, extreme care is required when claiming causal relationships between food and health as the level and complexity of the biases and confounding is so high that it affects even MR, which is known to be more robust than other approaches to these types of effects.

In a classic dietary analysis, investigators evaluate macronutrient compositions. In this study, we did not see similar effects from foods which have similar macronutrient composition. For example, if we look at cheese and meat, which are both relatively high in saturated fat and protein, we see no association of eating either with blood lipid profile (triglycerides, LDL or total cholesterol), while they have opposite effects on BMI (cheese lowering it and meat increasing it) (Fig S22e).). While the findings require further investigations in mechanisms and related behaviours, our genetic evidence lenders the support for the importance of food consumption and dietary patterns, not only intakes of specific nutrients[26].

If we look at which foods have the greatest effect on triglycerides, it is fruit, vegetables and fish; all with lowering effects (Fig S22f), not sources of carbohydrates or alcohol, known drivers of de novo lipogenesis. This seems to be confirmed by looking at the results with the overall PC traits (*All-PC1*, *-PC2*, *-PC3*) in which a higher consumption of fruit, vegetables and fish is always associated with lower triglycerides regardless of the loading on other food groups. It is impossible, however, to separate the effects of fruit, vegetables and fish from each other, in fact, if we look at the *Healthy PC* traits (see fig 5 panel A), only PC1, which summarises a higher consumption of all three is associated with lower triglycerides, suggesting the combined effects of all the three dietary factors or unmeasured correlated dietary behaviours or healthful habits.

This example shows that when considering the effect of food on health it is sometimes hard to separate the effect of single foods (although we have shown some examples) from those which are usually consumed together in a pattern. In this case, although fish and fruit and vegetables have a very different macronutrient composition it is impossible to separate their effect on triglycerides. This has been implied in previous studies including the European study on lactase persistence gene[9]. There, while the MR relating lactase-persistence gene to diabetes incidence supported no causal evidence of milk consumption, the secondary analyses identified the lactase-persistence variant would relate to consumption of potatoes, poultry, and cereals. These pieces of genetic evidence highlight the importance of a dietary pattern rather than single foods or nutrients. Any health claim from observational studies regarding one or the other should always take into account these facts. For further details of specific results, our online app allows exploration of hypotheses.

Our study was limited by the number of items available in the dietary questionnaire in the UK BioBank and thus has not explored the full extent of human nutrition, unfortunately apart from bread consumption no carbohydrate or sugar sources were measured, limiting our ability to explore these macronutrients and thus capture the overall diet. Nonetheless, this limitation is unlikely to turn over the abovementioned cautionary interpretation of the dietary MR results. Another important limitation is that effect sizes could be inflated because of the underestimation of the SNP effects on the food traits which will increase MR estimate effects. This under-estimation is due to the noise in the questionnaire responses, which warrant further statistical investigations. Of note, as we have no rationale to consider non-random measurement error, it is unlikely to hinder the detection of a causal effect or its direction, but further studies are needed to assess the precise effect sizes. Before translation of our findings into policy, more studies using different methodologies will be required.

In conclusion, we have developed an important framework and new tools to help illuminate the effects of nutrition on health and have shown that despite the existing belief that certain dietary assessment provides low-quality data, it is still possible to extract useful information using our methods. It will be interesting to learn to what degree the confounding of food choice reporting by educational attainment and disease risk factors observed here is seen in other settings with different food cultures and social stratification to the UK.

Author Contributions

NP,JFW,JRBP,ZK,EJG,FRD,KKO contributed to the study design. JFW,TE,JRBP,AR,TG,FI,KKO,FRD contributed data. NP,CMD,EJG,NM,FI,JZ,NT,KAK,MPC, performed the statistical analyses. NP, JFW,ZK, JRBP, TE, NT,KF,CMD,LR,EJG,FI,KKO,FRD contributed to the interpretation of the results. All authors contributed to writing and editing of the text.

Acknowledgements

J.F.W. acknowledges support from the MRC Human Genetics Unit quinquennial programme grant “QTL in Health and Disease”. This research has been conducted using the UK Biobank Resource under Application Number 19655. We would like to thank Erin MacDonald-Dunlop, and Pascale Lubbe for help with statistical analyses.
EGCUT was funded by Estonian Research Council Grant IUT20-60, PUT1660 (T.E.), PUT1665 (K.F.), the European Union through the European Regional Development Fund grant no. 2014-2020.4.01.15-0012 GENTRANSMED and 2014-2020.4.2.2, and Estonian and European Research Roadmap grant no.2014-2020.4.01.16-0125. The EPIC-Norfolk study (DOI 10.22025/2019.10.105.00004) has received funding from the Medical Research Council (MR/N003284/1, MC-PC_13048, and MC-UU_12015/1). The Fenland study (DOI: 10.1186/ISRCTN72077169) was funded by the Medical Research Council and the Wellcome Trust (Ref: 074548). J.P., K.O., F.I., and F.R.D. were funded by the UK Medical Research Council Epidemiology Unit core grant (MC-UU_12015/2, MC_UU_12015/5). T.R.G. receives funding from the UK Medical Research Council (MC_UU_00011/4). Z.K. received funding from the Swiss National Science Foundation (31003A_169929). We are grateful to all the participants who have been part of the project and to the many members of the study teams at the University of Cambridge who have enabled this research.

Data Availability

All GWAS results will be made available through GWAS catalog at the time of publication. All results from the MR analyses have been shared in the additional tables.

References

1. Lis, C. G., Gupta, D., Lammersfeld, C. A., Markman, M. & Vashi, P. G. Role of nutritional status in predicting quality of life outcomes in cancer – a systematic review of the epidemiological literature. *Nutrition Journal* **11**, (2012).
2. Simpson, S. J. *et al.* The nutritional geometry of liver disease including non-alcoholic fatty liver disease. *J. Hepatol.* **68**, 316–325 (2018).
3. Misra, A. & Khurana, L. Obesity and the Metabolic Syndrome in Developing Countries. *The Journal of Clinical Endocrinology & Metabolism* **93**, s9–s30 (2008).
4. Misra, A. & Khurana, L. Obesity and the Metabolic Syndrome in Developing Countries. *The*

Journal of Clinical Endocrinology & Metabolism **93**, s9–s30 (2008).

5. Zheng, J. *et al.* Recent Developments in Mendelian Randomization Studies. *Curr Epidemiol Rep* **4**, 330–345 (2017).
6. Millwood, I. Y. *et al.* Conventional and genetic evidence on alcohol and vascular disease aetiology: a prospective study of 500 000 men and women in China. *Lancet* **393**, 1831–1842 (2019).
7. Cornelis, M. C. & Munafo, M. R. Mendelian Randomization Studies of Coffee and Caffeine Consumption. *Nutrients* **10**, (2018).
8. Bergholdt, H. K. M., Nordestgaard, B. G., Varbo, A. & Ellervik, C. Milk intake is not associated with ischaemic heart disease in observational or Mendelian randomization analyses in 98,529 Danish adults. *Int. J. Epidemiol.* **44**, 587–603 (2015).
9. Vissers, L. E. T. *et al.* Dairy Product Intake and Risk of Type 2 Diabetes in EPIC-InterAct: A Mendelian Randomization Study. *Diabetes Care* **42**, 568–575 (2019).
10. Meddens, S. F. W., de Vlaming, R., Bowers, P. & Burik, C. A. P. Genomic analysis of diet composition finds novel loci and associations with health and lifestyle. *bioRxiv* (2018).
11. Cole, J. B., Florez, J. C. & Hirschhorn, J. N. Comprehensive genomic analysis of dietary habits in UK Biobank identifies hundreds of genetic loci and establishes causal relationships between educational attainment and healthy eating. doi:10.1101/662239
12. Lis, C. G., Gupta, D., Lammersfeld, C. A., Markman, M. & Vashi, P. G. Role of nutritional status in predicting quality of life outcomes in cancer – a systematic review of the epidemiological literature. *Nutrition Journal* **11**, (2012).
13. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
14. Day, F. R., Loh, P.-R., Scott, R. A., Ong, K. K. & Perry, J. R. B. A Robust Example of Collider Bias in a Genetic Association Study. *The American Journal of Human Genetics* **98**, 392–393 (2016).
15. Day, N. *et al.* EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. *Br. J. Cancer* **80 Suppl 1**, 95–103 (1999).

16. Heid, I. M. *et al.* Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat. Genet.* **42**, 949–960 (2010).

17. Mul, J. D. *et al.* Melanocortin receptor 4 deficiency affects body weight regulation, grooming behavior, and substrate preference in the rat. *Obesity* **20**, 612–621 (2012).

18. 8 steps to a heart-healthy diet. *Mayo Clinic* (2019). Available at: https://www.mayoclinic.org/diseases-conditions/heart-disease/in-depth/heart-healthy-diet/art-20047702. (Accessed: 26th July 2019)

19. Iriti, M. & Varoni, E. M. Cardioprotective effects of moderate red wine consumption: Polyphenols vs. ethanol. *J. Appl. Biomed.* **12**, 193–202 (2014).

20. Das, S., Santani, D. D. & Dhalla, N. S. Experimental evidence for the cardioprotective effects of red wine. *Exp. Clin. Cardiol.* **12**, 5–10 (2007).

21. VanderWeele, T. J. A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology* **24**, 224–232 (2013).

22. Frayling, T. M. *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889–894 (2007).

23. Paul, D. R., Rhodes, D. G., Kramer, M., Baer, D. J. & Rumpler, W. V. Validation of a food frequency questionnaire by direct measurement of habitual ad libitum food intake. *Am. J. Epidemiol.* **162**, 806–814 (2005).

24. Dehghan, M. *et al.* Association of dairy intake with cardiovascular disease and mortality in 21 countries from five continents (PURE): a prospective cohort study. *Lancet* **392**, 2288–2297 (2018).

25. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).

26. Mozaffarian, D., Rosenberg, I. & Uauy, R. History of modern nutrition science—Implications for current research, dietary guidelines, and food policy. *BMJ* (2018).