Comprehensive genomic analysis of dietary habits in UK Biobank identifies hundreds of genetic loci and establishes causal relationships between educational attainment and healthy eating

Joanne B. Cole[1-3], Jose C. Florez[1,2,4], Joel N. Hirschhorn*[1,3,5,]

[1] Programs in Metabolism and Medical & Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States of America

[2] Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts, United States of America

[3] Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, Massachusetts, United States of America

[4] Department of Medicine, Harvard Medical School, Boston, Massachusetts, United States of America

[5] Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America

*Corresponding author:

Joel N. Hirschhorn

Address: Boston Children's Hospital, 3 Blackfan Circle, CLS 16028, Boston, MA 02115, USA

Phone: (617) 919-2129

Email: Joel.Hirschhorn@childrens.harvard.edu

Abstract

Unhealthy dietary habits are leading risk factors for life-altering diseases and mortality. Large-scale biobanks now enable genetic analysis of traits with modest heritability, such as diet. We performed genomewide association on 85 single food intake and 85 principal component-derived dietary patterns from food frequency questionnaires in UK Biobank. We identified 814 associated loci, including olfactory receptor associations with fruit and tea intake; 136 associations were only identified using dietary patterns. Mendelian randomization suggests a Western vs. prudent dietary pattern is causally influenced by factors correlated with education but is not strongly causal for coronary artery disease or type 2 diabetes.

Unhealthy diet is thought to be the leading risk factor for mortality both globally[1] and in the US.[2] Overall, incidence rates of these dietary risk factors and their related diseases, like obesity and type 2 diabetes (T2D) are rising in parallel worldwide,[3,4] causing global epidemics that require our urgent attention. Describing a biological basis for unhealthy dietary preferences could guide more effective dietary recommendations.

There is a clear, albeit modest, genetic component to diet, such as traditional measures of macronutrient intake (i.e. proportion of carbohydrate, fat, and protein to total energy intake), as demonstrated by significant heritability and individual genetic associations.[5-9] Five genomewide association studies (GWAS) of macronutrient intake have been conducted to date; the most recent multi-trait analysis identified 96 independent genetic loci by combining summary statistics of individual macronutrient GWAS from 24-hour diet recall questionnaires in 283K individuals in UK Biobank (UKB).[6-10] In addition, the Neale Lab conducted GWAS (http://www.nealelab.is/uk-biobank/) across thousands of mostly binary traits analyzed primarily as dichotomous outcomes (i.e. wholemeal bread vs. all others) in 361K unrelated individuals in UKB. The recent GeneAtlas[11] improved power by using linear mixed models in 450K individuals in UKB, but analyzed a smaller set of dietary variables.

Additional measures of dietary intake, including both curated measures of single food intake and multivariate dietary patterns such as those described by principal component (PC) analysis, have also shown significant associations with health outcomes in both epidemiological studies[12,13] and clinical trials.[14] Thus, with the recent advent of large biobank-sized population cohorts with dietary data, we can now perform GWAS with multiple complementary phenotyping approaches to examine a wide array of dietary habits, including previously unstudied single food comparisons (i.e. wholemeal vs. white bread) and dietary patterns. Here, we report heritability and GWAS analysis (using linear mixed models) of both single food intake, analyzed as curated single food intake quantitative traits (FI-QTs), and of PC-derived dietary patterns (PC-DPs) using food frequency questionnaire (FFQ) data in up to 449,210 Europeans from UKB; we

highlight association at biologically interesting loci (olfactory receptor loci associated with tea and fruit consumption) and use Mendelian randomization (MR) analyses to elucidate causal relationships pertaining to specific dietary habits.

Results

Dietary habits have strong phenotypic correlation and most are heritable

We derived 85 curated single FI-QTs from FFQ administered in UKB, using 35 nested and complementary questions (Supplementary Table 1; see Methods). As expected given the nested nature of these questions, many pairs of the 85 FI-QTs were significantly correlated ($P < 0.05/85 = 5.88\times10^{-4}$; Supplementary Figure 1 and Supplementary Table 2). We therefore also conducted PC analysis of these FI-QTs to generate 85 PC-DPs that capture correlation structure among intake of single foods and represent independent components of real-world dietary habits. The top 20 PC-DPs have eigenvalues greater than one, and explain 66.6% of the total variance in the 85 FI-QTs (Supplementary Figure 2).

Overall, 84.1% of dietary habits analyzed (143 of the 170 FI-QTs and PC-DPs) were significantly heritable (as assessed by h_g^2 $P < 0.05 / 170 = 2.9\times10^{-4}$; see Methods, Table 1, Supplementary Table 1, and Supplementary Figure 3) and displayed extensive genetic correlation (r_g; Supplementary Table 3). The relationships between the 83 FI-QTs and the 60 PC-DPs that have significant heritability are illustrated in Supplementary Figure 4. The most heritable FI-QTs fall into a handful of dietary food groups related to milk consumption, alcohol intake, and butter/spread consumption. The first PC-DP (hereafter referred to as PC1) is among the most heritable dietary patterns (PC1 h_g^2=13.6%, Table 1), and is more heritable than all of its individual contributing FI-QTs.

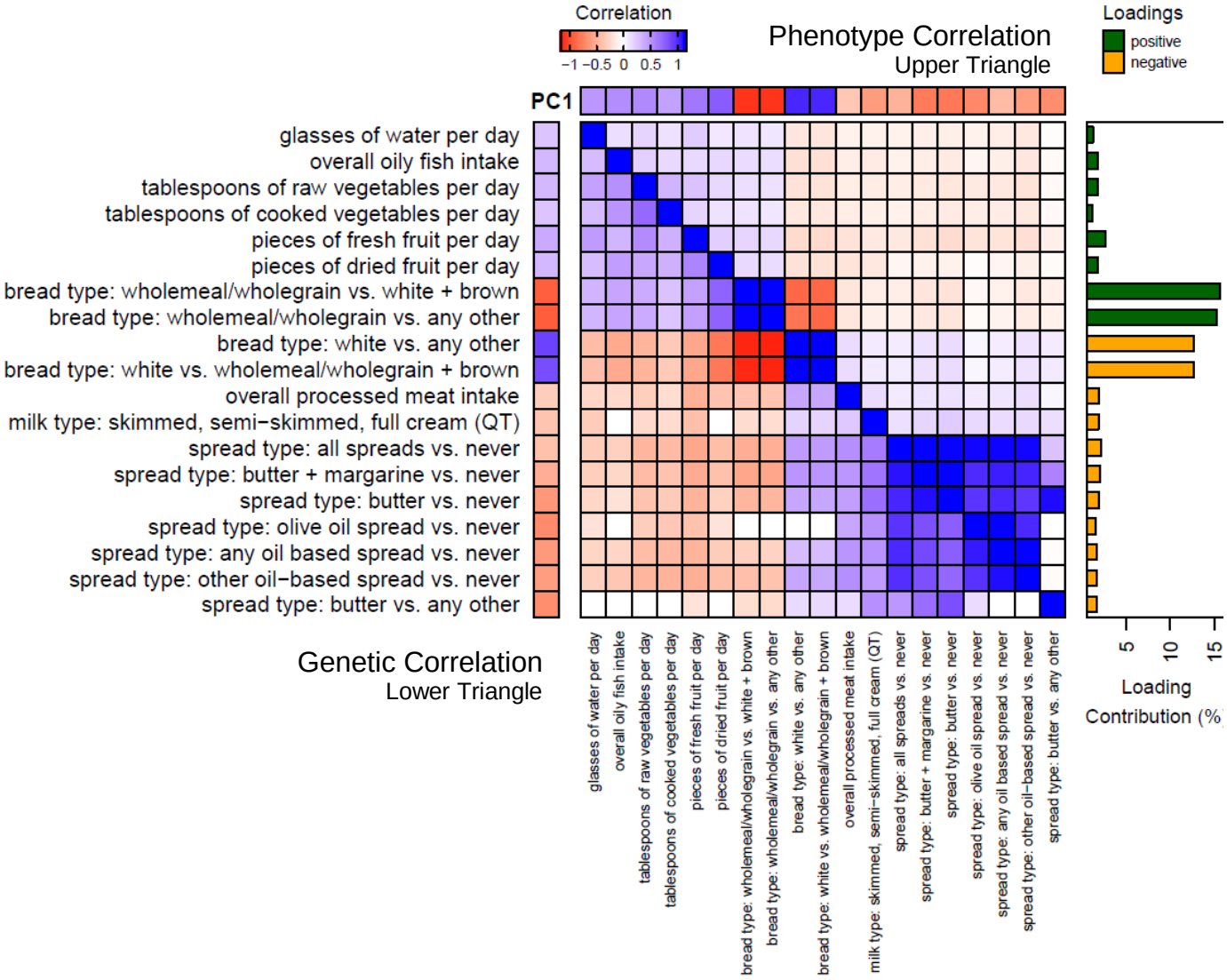
PC1, which explains 8.63% (Supplementary Figure 2) of the total phenotypic variance in FI-QTs, captures previously described “Western” and “prudent” dietary factors[15] and is primarily

defined by the type of bread consumed (wholegrain/wholemeal vs. white bread). Overall, the FI-QTs that have high positive loadings for PC1 include wholemeal/wholegrain bread consumption, increased fruit and vegetable intake, increased oily fish intake, and increased water intake. The FI-QTs that have high negative loadings include white bread consumption, butter and oil spread consumption, increased processed meat intake, and consumption of milk with higher fat content (Figure 1).

Table 1: SNP Heritability Estimates (h_g^2) and Number of Significant GWAS Loci for the Most Heritable 20 Dietary Habits.

Dietary habit	N	h_g^2	SE	P-value	Number of Significant Loci
milk type: soy milk vs. never	31,889	0.282	0.010	5.8×10^{-175}	1
milk type: full cream vs. never	43,995	0.243	0.007	4.8×10^{-264}	1
spread type: tub margarine vs. never	77,738	0.182	0.004	$< 1.00 \times 10^{-300}$	5
among current drinkers, drinks usually with meals: yes vs. no	235,312	0.160	0.001	$< 1.00 \times 10^{-300}$	30
PC1	449,210	0.136	0.001	$< 1.00 \times 10^{-300}$	140
spread type: flora + benecol vs. never	86,823	0.133	0.004	2×10^{-242}	2
milk type: other milk vs. never	20,557	0.128	0.015	1.42×10^{-17}	0
overall alcohol intake	448,623	0.121	0.001	$< 1.00 \times 10^{-300}$	98
PC3	449,210	0.118	0.001	$< 1.00 \times 10^{-300}$	82
glasses of water per day	445,965	0.116	0.001	$< 1.00 \times 10^{-300}$	78
total drinks of alcohol per month	449,210	0.113	0.001	$< 1.00 \times 10^{-300}$	104
spread type: low fat spread vs. never	72,017	0.110	0.004	1.8×10^{-166}	0
coffee type: decaffeinated vs. any other	353,710	0.108	0.001	$< 1.00 \times 10^{-300}$	36
pieces of fresh fruit per day	447,401	0.108	0.001	$< 1.00 \times 10^{-300}$	99
overall cheese intake	438,453	0.108	0.001	$< 1.00 \times 10^{-300}$	60
spread type: butter vs. never	213,549	0.102	0.001	$< 1.00 \times 10^{-300}$	15
frequency of adding salt to food	448,890	0.102	0.001	$< 1.00 \times 10^{-300}$	82
among current drinkers, drinks usually with meals: yes/, it varies, no	357,136	0.101	0.001	$< 1.00 \times 10^{-300}$	29
bread type: white vs. wholemeal/wholegrain + brown	416,312	0.100	0.001	$< 1.00 \times 10^{-300}$	46
milk type: skimmed vs. never	108,035	0.098	0.003	$< 1.00 \times 10^{-300}$	1

Figure 1: Relationships between PC1 and its 19 Significantly Contributing Single Food Intake QT. The heat map depicts the phenotypic (upper triangle) and genetic (lower triangle) correlation between the 19 significantly contributing single FI-QTs with each other and PC1. All correlations with nonsignificant P-values (P>0.05/85) were set to 0. Percent contribution of each of the 19 traits to PC1 is depicted in the correlation matrix bar plot annotation on the right, colored by loading direction.

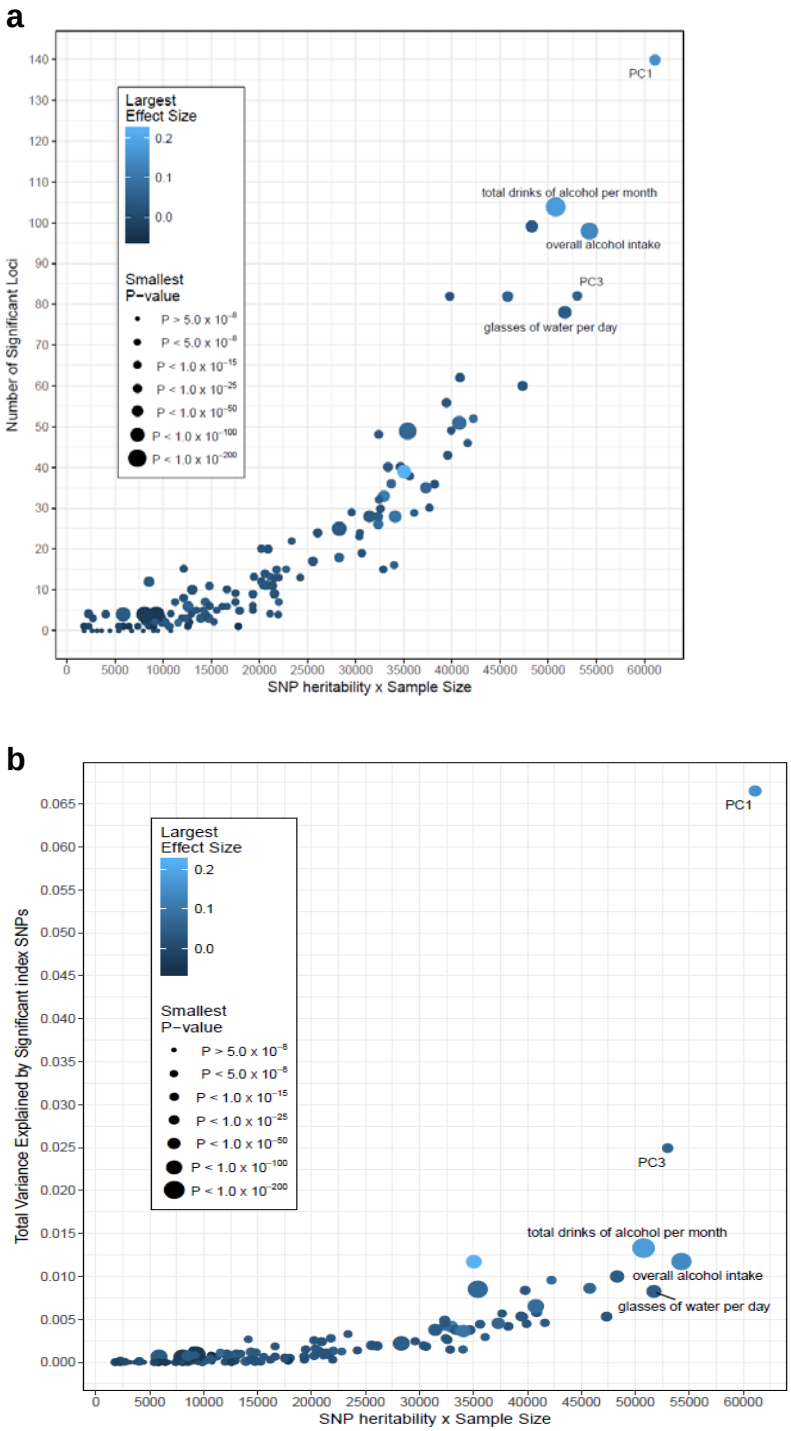


Dietary habit GWAS in UKB identifies 814 independent loci

GWAS on the 143 significantly heritable dietary habits, using linear mixed models in up to 449,210 individuals, identified 814 independent loci (defined as >500 kb apart) surpassing genome-wide significance ($P<5.0\times10^{-8}$). Of these, 309 also surpass a more conservative Bonferroni corrected study-wide significance threshold ($P<5.0\times10^{-8}/143$ traits$=2.9\times10^{-10}$; Supplementary Table 4). Across the 143 dietary traits, there was a clear positive correlation of heritability estimates (h_g^2) with the number of significant loci and the variance explained by these loci (Figure 2), with outliers often explained by smaller sample size (Supplementary Figure 5). Notably, there is a mix of FI-QTs and PC-DPs among the most successful GWAS traits, with the Western vs. prudent PC1 GWAS identifying the most significant loci (M=140) that together explain the largest amount of variance of any dietary habit analyzed (6.65%).

Of the 814 index SNPs, 767 (94%) are common with minor allele frequencies greater than or equal to 5%; similar to previous GWAS findings,[16] the vast majority are either intronic or intergenic (82.2%). Credible set analysis found that more than one-third of our lead signals (289/814) were driven by 10 or fewer SNPs, 124 had 2-5 credible set SNPs and 54 had a single 95% credible set SNP (see Methods). Of the credible sets with single SNPs, 65.2% are intronic or intergenic, while 13.0% are missense variants and the remainder largely regulatory (by comparison, only 1.9% of index SNPs from multi-SNP credible sets are missense). We investigated whether any of our loci were previously reported at genome-wide significance for any trait in either the GWAS catalog[17] or the comprehensive Neale Lab GWAS (http://www.nealelab.is/uk-biobank/) of 4,358 traits in 361,194 unrelated individuals in UKB. Of the 814 dietary habit GWAS loci, 205 have never been previously reported. This can largely be explained by our use of previously unstudied curated FI-QTs and PC-DPs and/or statistical power gained from linear mixed models in nearly 450K individuals.

Figure 2: Relationship between SNP Heritability and GWAS Success. A) Scatter plots of the number of genome-wide significant loci or B) variance explained by genome-wide significant SNPs vs. SNP heritability×sample size for all 143 significantly heritable traits. Points are colored by the largest effect size and sized by the smallest P-value of their significant index SNPs.



Several olfactory receptor loci are associated with intake of specific foods

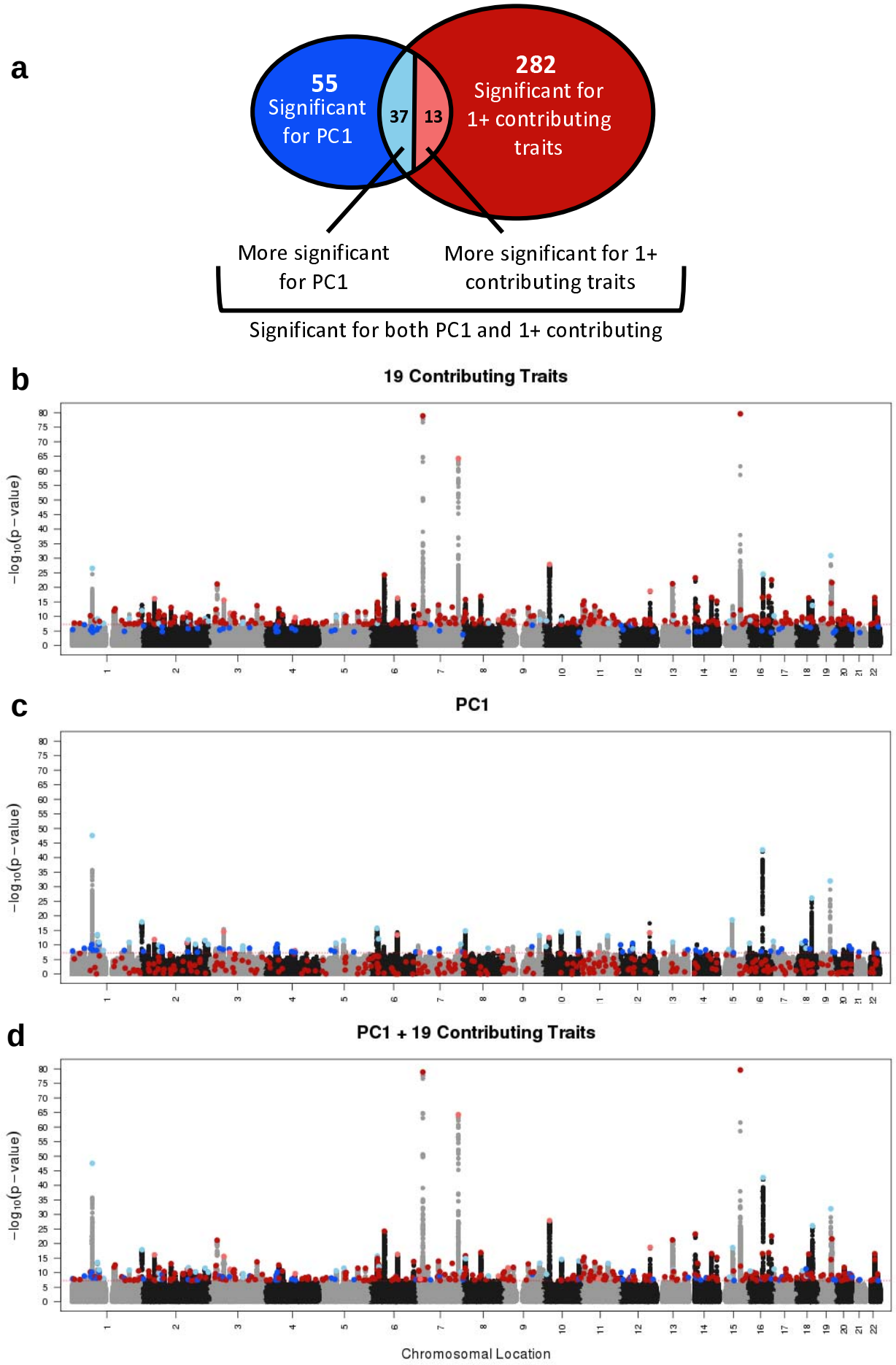
Among the lead loci are several regions containing clusters of olfactory receptor genes. Although our dietary habit GWAS loci are not enriched for olfactory receptor genes (P=0.419), the association of olfactory receptors with specific FI-QTs supports the well-known link between smell and taste. The top signal associated with "pieces of fresh fruit eaten per day" is a region on chromosome 7q35 (lead SNP rs10249294 $P=5.7\times10^{-65}$; Supplementary Figure 6). All 44 SNPs within the 95% credible set cover a 27 kb region containing olfactory receptor gene *OR6B1* (Supplementary Figure 7). Another locus specific to fruit intake, on chromosome 14q (Supplementary Figure 6) is associated with both "pieces of fresh fruit per day" (rs34162196 $P=5.1\times10^{-24}$) and "pieces of dried fruit per day" (rs35260863 $P=6.4\times10^{-18}$). The only two SNPs contained in both 95% credible sets cover a 5.8 kb region (chr14: 22038125-22043949) containing olfactory receptor gene *OR10G3* (Supplementary Figure 7).

Five additional loci have 95% credible sets that overlap olfactory receptor genes and are strongly associated with a single FFQ question. Notably, "cups of tea per day" is associated with a region on chr11q12 (lead SNP rs1453548 $P=3.1\times10^{-9}$) that contains six different olfactory receptor genes including *OR5AN1*, *OR5A2*, *OR5A1*, *OR4D6*, *OR4D10*, and *OR4D11* (Supplementary Figure 7). The 95% credible set contains SNP rs7943953 (chr11:59224144) previously reported for being associated with odor perception, and specifically that of β-ionone (floral) sensitivity, an aroma compound found in high quantities in tea.[18-20] By contrast, "cups of coffee per day" is associated with a nearby yet distinct region on chr11q12 (lead SNP rs643017 $P=1.1\times10^{-8}$) in a dense cluster of olfactory receptor genes containing *OR8U8* and *OR5M3* (Supplementary Figure 7). Other associations with olfactory receptor gene regions include regions containing *OR52J3* and OR52E2 with three dietary habits measuring butter consumption (rs2445249 with "spread type: butter vs. any other" $P=9.1\times10^{-15}$), *OR4K17* with "tablespoons of raw vegetables per day" (rs10140123, $P=4.1\times10^{-10}$), and *OR10A3* and *OR10A6* with "overall cheese intake" (rs757969034, $P=2.4\times10^{-8}$; Supplementary Figure 7).

GWAS of dietary pattern PC1: relationship to individual food intake traits and non-dietary phenotypes

Of the PC-DPs, the "Western vs. prudent" PC1 dietary pattern has the highest heritability and most genome-wide significant loci, and similar dietary patterns have been associated with disease.[21,22] Although PC1 is both phenotypically and genetically correlated with its 19 contributing FI-QTs (absolute $r_{p=}$ 0.25-0.81; absolute r_g =0.29-0.93), the GWAS results provide distinct sets of significant associations. Together, PC1 and its 19 contributing traits are associated with a total of 387 independent genome-wide significant loci, falling into one of four categories (Figure 3): 55 loci significant for PC1 only (dark blue), 282 loci significant for one or more of the 19 contributing FI-QTs only (dark red), 37 loci more significant for PC1 (light blue), and 13 loci more significant for one or more of the 19 QTs (light red). The 55 loci significantly associated with PC1 but not FI-QTs still trend toward association with one or more of the 19 contributing FI-QTs, whereas the reciprocal is not always the case: some FI-QT-associated SNPs display no association with PC1 at all. This observation indicates that the use of PCs can increase power to detect some associations, while others are only detectable through association with specific foods, supporting the use of both of these complementary phenotyping approaches to more effectively define the genetic architecture of dietary intake.

Figure 3: Manhattan Plot of PC1 and its 19 significantly contributing single food intake QT. A) Schematic depiction of 4 groups of significant loci (and corresponding B-C Manhattan plot colors) for various combinations of PC1 and its 19 contributing traits. 55 loci significant for PC1 only (dark blue), 282 loci significant for one or more of the 19 contributing FI-QTs only (dark red), 37 loci more significant for PC1 (light blue), and 13 loci more significant for one or more of the 19 FI-QTs (light red). B) Combined Manhattan plot of the minimum p-value across the 19 FI-QTs that significantly contribute to PC1. C) Manhattan plot of PC1 only. D) Combined Manhattan plot of the minimum P-value across PC1 and the 19 FI-QTs.



In addition to being correlated with its contributing FI-QTs, PC1 displays significant genetic correlation with 248 non-diet related traits, including traits relating to physical activity, educational attainment, socioeconomic status, smoking status, medication codes, and urine biomarkers (Supplementary Table 5). The lead PC1 SNP, rs66495454, is a common indel (MAF=38%) located at chr1:72748567 in the promoter of neuronal growth factor 1 (*NEGR1;* Supplementary Figure 7). SNP rs66495454's deletion allele (-/TCCT) is associated with a decrease in prudent eating (beta=-0.017, P=2.80×10^{-48}) and has been previously reported as associated with a decrease in intelligence, educational attainment and, perhaps surprisingly, BMI.[23-25] We were struck by the observation that 122 PC1-associated loci were already present in the GWAS catalog[17] or Neale Lab GWAS (http://www.nealelab.is/uk-biobank/), and of these, 83 were associated with intelligence or cognitive ability, educational attainment, or obesity-related anthropometric traits, including rs1421085 in *FTO* and rs429358 in *APOE* (Supplementary Table 4). Of these 83, 24 were associated with intelligence or educational attainment and not obesity-related traits; 38 were associated with obesity-related traits and not intelligence or educational attainment. DEPICT pathway analysis of PC1 identified enrichment of 2 gene sets, including axonogenesis, and 22 tissues, of which 21 were brain-related (FDR < 0.05; Supplementary Table 6). The large and broad overlap in both the significant loci and the overall genetic make-up of our dietary habits with other traits, many of which do not have well-established biological links with diet, emphasizes a need for exploring correlation vs. causation.

Therefore, we sought to understand the cause-and-effect relationships between PC1 and educational attainment, fluid intelligence scores, and BMI. We performed additional GWAS of educational attainment, fluid intelligence scores, and BMI in UKB and applied bidirectional MR with these phenotypes and PC1 (see Methods). Both inverse-variance weighted (IVW) random-effects and Weighted Median (WM) MR provided significant evidence of causal effects of educational attainment and fluid intelligence on PC1 (educational attainment WM beta=0.823, 95% Confidence Intervals [CI]: 0.751-0.894, P=6.59×10^{-113}; intelligence WM beta=0.261, 95%

CI: 0.190-0.332, P=7.71×10^{-13}; Table 2 and Supplementary Figure 8). Egger regression intercepts for both intelligence and educational attainment were not significantly different than 0 (Table 2) and causal effect estimates were essentially unchanged after different approaches to variant filtering (See Methods, Table 2). For each 1 standard deviation increase in educational attainment and intelligence, these data estimate that PC1 shifts from Western towards prudent eating by 0.823 and 0.261 standard deviations, respectively. To rule out weak instrument bias in overlapping samples, we validated our educational attainment results in two ways with high congruence. An externally derived educational attainment genetic instrument[25] had an effect estimate on PC1 equal to 0.708 (95% CI: 0.576-0.840 P=6.23×10^{-26}), and splitting UKB into training and testing datasets also yielded similar and significant effect estimates (Supplementary Table 7).

The reverse analysis shows significant but smaller estimates of causal influences of PC1 on educational attainment and fluid intelligence (educational attainment WM beta=0.199, 95% CI: 0.175-0.224, P=7.44×10^{-56}; intelligence WM beta=0.074, 95% CI: 0.045-0.104, P=8.04×10^{-7}). The smaller effects suggest that while there could be pleiotropy undetected by the Egger and WM approaches, or true bidirectional effects, the causal influences are more likely to be in the direction of educational attainment to prudent dietary patterns. Importantly, because the instrumental variables used for educational attainment are not mechanistically linked directly to educational attainment/intelligence, it remains possible that causal influences on PC1 could be due to unmeasured heritable factor(s) that are themselves causal for educational attainment/intelligence. In contrast to the educational attainment analyses, we were unable to provide robust evidence of a causal relationship in either direction between BMI and PC1 due to significant heterogeneous pleiotropic effects leading to inconsistent causal effect estimates (Table 2, Supplementary Figure 8)

Table 2. Bi-directional Mendelian Randomization results for PC1 as both an exposure and outcome with educational attainment (ED), fluid intelligence scores (INT), and BMI.

Pre-SNP filtering																
			Inverse Variance Weighted				**MR Egger**						**Weighted Median**			
Exposure	**Outcome**	**Number of SNPs**	**Est.**	**CI Lower**	**CI Upper**	**P-value**	**Est.**	**CI Lower**	**CI Upper**	**P-value**	**Intercept**	**Int. P-value**	**Est.**	**CI Lower**	**CI Upper**	**P-value**
ED	PC1	309	0.836	0.765	0.906	1.69×10^{-118}	0.970	0.681	1.259	4.88×10^{-11}	-0.002	0.347	0.823	0.751	0.894	6.59×10^{-113}
PC1	ED	140	0.203	0.173	0.232	2.20×10^{-41}	0.182	0.068	0.296	1.70×10^{-3}	0.001	0.715	0.199	0.175	0.224	7.44×10^{-56}
INT	PC1	184	0.264	0.183	0.345	1.82×10^{-10}	0.080	-0.260	0.420	0.644	0.004	0.276	0.261	0.190	0.332	7.71×10^{-13}
PC1	INT	140	0.105	0.071	0.138	8.23×10^{-10}	0.100	-0.029	0.229	0.129	0.000	0.941	0.074	0.045	0.104	8.04×10^{-7}
BMI	PC1	1165	0.000	-0.034	0.035	0.993	0.319	0.223	0.415	6.97×10^{-11}	-0.007	3.30×10^{-12}	0.014	-0.019	0.047	0.403
PC1	BMI	140	0.089	0.012	0.166	2.34×10^{-2}	0.432	0.140	0.724	3.78×10^{-3}	-0.012	0.017	0.007	-0.019	0.033	0.614
Post-SNP filtering																
			Inverse Variance Weighted				**MR Egger**						**Weighted Median**			
Exposure	**Outcome**	**Number of SNPs**	**Est.**	**CI Lower**	**CI Upper**	**P-value**	**Est.**	**CI Lower**	**CI Upper**	**P-value**	**Intercept**	**Int. P-value**	**Est.**	**CI Lower**	**CI Upper**	**P-value**
ED	PC1	288	0.829	0.765	0.893	7.80×10^{-142}	1.066	0.776	1.356	5.46×10^{-13}	-0.004	0.100	0.835	0.763	0.907	2.79×10^{-114}
PC1	ED	121	0.178	0.152	0.204	7.07×10^{-42}	0.199	0.067	0.332	3.14×10^{-3}	-0.001	0.749	0.184	0.159	0.209	1.87×10^{-47}
INT	PC1	172	0.265	0.189	0.340	5.41×10^{-12}	0.218	-0.170	0.607	0.271	0.001	0.812	0.270	0.200	0.341	6.01×10^{-14}
PC1	INT	126	0.086	0.058	0.115	3.72×10^{-9}	0.124	-0.014	0.263	0.079	-0.001	0.582	0.065	0.034	0.095	2.78×10^{-5}
BMI	PC1	1081	-0.064	-0.096	-0.032	8.76×10^{-5}	-0.031	-0.145	0.084	0.601	-0.001	0.552	-0.070	-0.102	-0.037	2.40×10^{-5}
PC1	BMI	110	-0.036	-0.072	0.001	0.055	-0.108	-0.319	0.102	0.314	0.002	0.492	-0.015	-0.041	0.012	0.272

Do dietary habits have a causal relationship with disease and related risk factors?

To test whether the "prudent" PC1 dietary pattern is likely to causally influence disease risk, we repeated bidirectional MR analysis between PC1 with coronary artery disease (CAD) from the mostly European CARDIoGRAMplusC4D GWAS and T2D from the DIAGRAM consortium 2017 GWAS. [26,27] The only association that provides robust evidence of a causal relationship was an increased risk in CAD leading to an increase in PC1 prudent eating with a significant, albeit small effect (WM beta=0.0458, 95% CI: 0.016-0.076, *P*=0.003, Supplementary Table 8), suggesting reverse causation of CAD on diet. Though we found that higher educational attainment increases healthier eating, using our genetic instruments we did not identify causal evidence that eating healthier (PC1) causes a decreased risk for CAD or T2D.

In contrast to the associations with PC1, a single SNP, rs1453548, strongly influences "cups of tea per day" and has a plausible biological mechanism (it is located in an olfactory receptor-dense region and explains >96% of the observed phenotypic variance of β-ionone sensitivity[28]). We therefore performed an MR analysis using rs1453548 on the complete set of traits in UKB using the Neale Lab GWAS (http://www.nealelab.is/uk-biobank/). Using a strict Bonferroni-corrected significance threshold ($P<0.05/4358$ traits $=1.15\times10^{-5}$), we identified a significant causal effect of "cups of tea per day" on smoking status, for which increases in the minor allele T (MAF=34%) that cause an increase in "cups of tea per day" cause a decrease in smoking status (Neale 20160:Ever smoked Wald ratio estimate= -0.51, 95% CI: -0.69 to -0.32, $P=4.326\times10^{-8}$; Supplementary Table 9). However, rs1453548 is directly associated with "ever smoked" smoking status in the Neale Lab GWAS at genome-wide significance ($P=4.329\times10^{-8}$), the 51-SNP "cups of tea per day" genetic instrument excluding rs1453548 has no significant causal effect on smoking status (IVW *P*=0.20), and beta-ionone is also found in tobacco,[29] suggesting the effects of rs1453548 on odor perception of β-ionone may have pleiotropic effects on both smoking status and tea drinking. Together with a lack of additional significant causal relationships between rs1453548 and health outcomes in UKB, our results indicate that drinking

more tea does not have clear effects on health outcomes in UKB, and it's possible that some previous reports on the health benefits of drinking more tea are a result of confounding with smoking status.

Discussion

Understanding the genetic architecture of dietary habits has immense implications for human health, but has been a difficult task, in part due to the low heritability of many dietary traits. The recent advent of large-scale datasets such as UKB, with deep phenotyping on hundreds of thousands of individuals, has now made genetic discovery of traits with relatively low heritability possible. Expansion of phenotyping to include both curated FI-QTs and PC-DPs together with GWAS in nearly 450K individuals allowed our study to make hundreds of new genetic discoveries relating to diet. Our work advances the elucidation of the genetic architecture of multiple correlated dietary habits and helps lay the groundwork for future research on nutrigenomic and other complex multifactorial multivariable datasets.

Our work emphasizes the importance of interrogating the genetics of complementary phenotypes to glean a more complete picture of the genetic architecture of diet. One of the strongest associations we observed was between SNP rs1229984 and the FI-QT "total drinks of alcohol per month" (P=3.8×10^{-248}, Supplementary Figure 6). This SNP has been previously associated with alcohol consumption;[11,30-33] consistent with a recent meta-analysis,[33] our use of a curated and quantitative FI-QT improved power compared with the more categorical "overall alcohol intake" phenotype and individual alcohol subtypes (i.e. "red wine glasses per month"; Supplementary Figures 6 and 9). This increase in power is consistent with the high genetic correlation but low phenotypic correlation between individual questions related to alcohol, indicating that a composite alcohol question is more suitable for genetic discovery (Supplementary Figure 10). Furthermore, using PC1 as an example, we demonstrate that the

genetic architecture of FI-QTs and PC-DPs are distinct, with hundreds of genetic associations more strongly associated with either PC1 or with its contributing FI-QTs. Overall, by using complementary phenotyping approaches, we identified 814 independent genetic associations, of which 205 were completely novel, 311 were uniquely associated with curated FI-QTs, and 136 were uniquely associated with PC-DPs.

As an initial exploration of the implications of genetically-influenced composite dietary patterns, we focused on the strong genetic overlap between PC1 dietary pattern and phenotypes related to educational attainment.[34] While bidirectional MR demonstrates some pleiotropic effects between educational attainment and PC1, the relative strengths of these causal estimates suggests that higher educational attainment and/or correlated phenotypes (such as socieoeconomic status or factors related to school performance) shift eating habits towards a healthier, more prudent diet. While previous observational studies have shown that Western and prudent dietary patterns are associated with CAD and T2D,[12,22] our MR analysis of PC1 on CAD or T2D did not demonstrate a causal effect from diet to disease, but rather a small suggestion of a reverse causal relationship between CAD and diet, (CAD diagnosis leads to a more “prudent” dietary pattern).

The conclusion that a “Western” dietary pattern does not appear to be a causal risk factor for disease must be viewed in the context of several potential limitations of our study. Genetic instruments derived from genome-wide significant variants tend to explain a small fraction of phenotypic variance, which can lead to lack of power to detect potentially true causal effects of diet on outcomes, although this is mitigated by the large sample size of the UKB cohort. Additionally, while the use of overlapping samples in MR could in theory lead to inflated causal estimates,[35] UKB’s large sample size provides robust genetic instruments, and the strength of our causal associations, combined with our validation analysis with an independently ascertained set of instruments for educational attainment, suggest that our results are likely not influenced by weak instrument bias. Furthermore, although we did not detect evidence of

pleiotropy, it remains possible that pleiotropic effects of some of the variants associated with PC1 masked a causal effect on cardiometabolic disease risk. Finally, the aspects of a prudent dietary pattern reflected by PC1 (predominantly driven by wholemeal/wholegrain vs. white bread consumption) may not capture the causal protective features of a prudent dietary pattern. However, it remains possible that there is a stronger correlative than causal relationship between the Western dietary pattern and increased risk of cardiometabolic disease.

We also find several interesting associations between specific FI-QTs (fruit, tea, coffee, vegetables, cheese, and butter) and olfactory receptors. The chr11p15 locus controlling odor perception of beta-ionone,[28] described as smelling of cedar wood but upon dilution (e.g. in tea) a more floral aroma,[36] has pleiotropic effects that both reduce the chances of ever smoking and increase tea intake. While SNPs at chr11p15 have already been shown to be associated with food choice with and without added β-ionone,[28] we highlight here for the first time a link between β-ionone odor perception with smoking status, with potential significant implications for smoking-related health problems. This result also highlights the importance of understanding the pleiotropic consequences of variants used as genetic instruments in Mendelian randomization.

Of note, our dietary habits derived from a shortened FFQ were not adjusted for total energy intake, a measure highly correlated with physical activity and body weight;[37] as such, our dietary habits represent potentially non-isocaloric variations in dietary intake. However, we found minimal phenotypic correlation between any of our FFQ-derived phenotypes and 24-hour recall questionnaire-derived total energy intake (maximum correlation $r = 0.037$). Furthermore, none of our lead 814 SNPs were nominally significant in the Neale Lab total energy intake GWAS ($P>0.05/814$). Nine of our phenotypes, including “slices of bread per week”, “overall cheese intake”, and “glasses of water per day” did show significant genetic correlations with total energy intake, suggesting that the genetic architecture of these traits could be shared with traits that reflect more global lifestyle and dietary patterns (Supplementary Table 5).

Overall, we present the genetic analysis of two complementary phenotypic approaches for dietary habits in a well-powered sample. Our results expand the understanding of the genetic contributors to dietary preferences and highlight the advantages of using complementary and novel approaches to derive carefully curated phenotypes, both for diet and for polygenic traits more generally. Our results also empower investigations of how our eating habits causally relate to disease risk. Comprehensive and rigorous investigation into the causal consequences of different modifiable aspects of diet and lifestyle can potentially have enormous implications for public health.

Online Methods

UK Biobank Genetic Data

UKB is a large prospective cohort with both deep phenotyping and molecular data, including genome-wide genotyping, on over 500,000 individuals ages 40-69 living throughout the UK between 2006-2010.[38] Genotyping, imputation, and initial quality control on the genetic dataset has been described previously.[39,40] Additionally we removed individuals flagged for failing UKBiLEVE genotype quality control, heterozygosity or missingness outliers, individuals with putative sex chromosome aneuploidy, individuals with self-reported vs. genetically inferred sex mismatches, and individuals whom withdrew consent at the time of analysis. Work within was conducted on genetic data release version 3, with imputation to both Haplotype Reference Consortium and 1000 Genomes Project (1KGP)[41], under UK Biobank application 11898.

UK Biobank Phenotype Derivation

All phenotype derivation and genomic analysis was conducted on a homogenous population of individuals of European (EUR) ancestry (N=455,146), as determined by: 1) projection on to 1KGP phase 3 PCA space, 2) outlier detection to identify the largest cluster of individuals using Aberrant R package[42], selecting the lambda in which all clustered individuals fell within 1KGP EUR PC1 and PC2 limits (lambda=4.5), 3) removed individuals who did not self-report as “British”, “Irish”, “Any other white background”, “White”, “Do not know”, or “Prefer not to answer”, as self-identified non-EUR ancestry could confound dietary habits.

Prior to phenotype derivation, we removed individuals that were pregnant, had kidney disease as defined by ICD10 codes, or a cancer diagnosis within the last year (field 40005). The UKB food frequency questionnaire consists of quantitative continuous variables (i.e. field 1289, tablespoons of cooked vegetables per day), ordinal non-quantitative variables depending on overall daily/weekly frequency (i.e. field 1329, overall oily fish intake), food types (i.e. milk,

spread, bread, cereal, or coffee), or foods never eaten (field 6144, dairy, eggs, sugar, and wheat). Supplementary Table 1 provides a list of UKB fields relating to the corresponding FFQ question for each dietary habit, which can be looked up in the UKB Data Showcase (http://biobank.ndph.ox.ac.uk/showcase/). Ordinal variables were ranked and set to quantitative values, while food types or foods never eaten were converted into a series of binary variables. Variables relating to alcoholic drinks per month were derived from a conglomeration of drinks per month and drinks per week questions answered by different individuals depending on their response to overall alcohol frequency (field 1558). All 85 single food intake dietary phenotypes were then adjusted for age in months and sex, followed by inverse rank normal transformation. For individuals with repeated FFQ responses, both the dietary variable and the age in months covariate were averaged over all repeated measures. Principal components were then derived from all 85 FI-QTs after filling in missing data with the median using the prcomp base function in R. All phenotypes after averaging, covariate adjustment, and transformation were normally distributed continuous variables. FI-QTs with percent contribution (squared coordinates) greater than expected under a uniform distribution [1/85*100=1.18%] were included in Figure 1 and Supplementary Figure 4, created using ComplexHeatmap package in R.[43] Phenotype correlation between all 170 dietary habits was estimated using Pearson pair-wise correlation on complete observations in R. All correlations (phenotypic and genetic) with $P > 0.05/85$ were set to 0. The significance threshold here was selected based on a Bonferroni correction for 85 total FI-QTs to maintain stringency for multiple testing and consistency across phenotype and genetic correlation analyses, while allowing for the nested and non-independent nature of the FFQ questions and derived FI-QTs

Heritability, GWAS, and Genetic Correlation Analyses

Measures of heritability were obtained from BOLT-lmm software (v.2.3.2)[44,45] pseudo-heritability measurement representing the fraction of phenotypic variance explained by the

estimated relatedness matrix.[44-46] These estimates were highly correlated with h_g^2 estimates using LD score regression (LDSC)[47] (*r*=0.988 ; Supplementary Figure 11). BOLT-lmm was unable to calculate h_g^2 for "spread type: block margarine vs. never" yielding an "invalid estimate" error, likely indicating no genetic component as BOLTlmm reported h_g^2 equal to 0 for the highly related "spread type: block margarine vs. any other" dietary habit with ten times the sample size. GWAS of all variables was conducted using BOLT-lmm software (v.2.3.2)[44,45] linear mixed model association testing to account for relatedness. Additional covariates included in BOLT-lmm analysis for both heritability and GWAS included genotyping array and the first 10 genetic PCs derived on the subset of unrelated individuals using FlashPCA2,[48] followed by projection of related individuals on to the PC space. The number of associated loci was determined by clumping of signals within 500kb windows. Variance explained for each SNP was calculated from the effect size (beta) and effect allele frequency (f) as follows: beta^2*(1-f)*2f. Pair-wise r_g of all significantly heritable dietary habits were estimated using LDSC.[49] Again, all correlations (phenotypic and genetic) with P> 0.05/85 were set to 0. Index SNP annotations were evaluated using Neale Lab UKB GWAS consequence annotations based on Ensembl's Variant Effect Predictor (VEP).[50] We performed DEPICT gene set and tissue enrichment analysis, which relies on reconstituted gene sets from publicly available gene sets and pathways and gene expression data to evaluate enrichment in Western vs. prudent PC1 GWAS loci.[51]

Ninety-five percent credible sets were estimated for all 2,515 genome-wide significant associations using posterior probabilities as calculated by FINEMAP v1.3[52] using shotgun stochastic search in 500kb windows. We then used LDstore v1.1[53] to calculate LD and identify SNPs in high LD (r^2≥0.80) with any of the 77,229 95% credible set SNPs. The GWAS catalog[17] (downloaded on November 13, 2018) and Neale Lab GWAS v2 conducted in up to 361,194 unrelated individuals (August 2018 release; http://www.nealelab.is/uk-biobank/) was then searched for this list of 339,832 unique SNPs in either any of the 95% credible sets or in high

LD with any SNP in the 95% credible sets for any of the 2,515 significant GWAS signals in 814 independent loci. We considered any locus as being previously reported if any SNP in its 95% credible set or in high LD ($r^2 \geq 0.80$) with any SNP in the 95% credible SNP was associated with any trait at genome-wide significance ($P < 5.0\times10^{-8}$). LocusZoom plots were made using the stand-alone package[54] to include LD information from UKB as determined by LDstore and 95% credible sets from FINEMAP.[52,53]

LDSC r_g was again performed on 143 significantly heritable dietary habits with 4,336 Neale Lab UKB GWAS traits, 3,219 of which had at least one non-missing r_g estimate from LDSC. Using a strict Bonferroni correction threshold for all pairwise tests between 143 dietary habits and 3,219 highly correlated and even overlapping Neale Lab GWAS traits ($P<0.05/460{,}317=1.09\times10^{-7}$), we set all non-significant r_g to 0. Supplementary Table 5 represents a complete pair-wise r_g matrix.

Enrichment for olfactory receptor genes among dietary habits was evaluated using 1,000 sets of null matched SNPs based on minor allele frequency, number of SNPs in LD at various LD thresholds, distance to nearest gene, and gene density using the SNPsnap webtool.[55] We used fisher's test for enrichment of olfactory receptor genes using SNPsnap's nearest gene annotation for 647 dietary habit GWAS index SNPs and for 1,000 sets of null matched SNPs (167 of our index SNPs were excluded for being in the HLA region or had insufficient matches). We based the enrichment analysis contingency table on 842 annotated olfactory receptor genes among 48,903 genes in SNPsnap. Of the 1,000 null enrichment analyses, 419 had a fisher's test estimate equal to or greater than our real data's enrichment estimate.

Mendelian Randomization

Bidirectional Mendelian Randomization was conducted using genome-wide significant index SNPs clumped by 500kb windows from GWAS in UKB on PC1 (M=140), fluid intelligence scores (M=184), educational attainment (M=309), and BMI (M=1165). Fluid intelligence scores

for GWAS in EUR (N=232,601) was derived from both in person and online cognitive tests. Assessment center fluid intelligence scores (field 20016) were averaged for up to 3 visits and adjusted for average age in months, sex, and assessment center. Online fluid intelligence scores (field 20191) were adjusted for age in months, sex, and townsend deprivation index. The final fluid intelligence score was first set to average assessment center fluid intelligence score, and when missing was filled in with the online fluid intelligence score, for which the combination of these scores were then adjusted for collection method, followed by inverse normal transformation. Educational attainment for GWAS in EUR (N=450,884) was derived using a previously published method based on mapping UKB qualifications field 6138 to US years of schooling,[56] following by adjusted for age in months, sex, and assessment center, and inverse normal transformation. BMI was calculated from weight (field 21002) and standing height (field 50) and averaged from up to 3 assessment center visits. Average BMI was adjusted for average age in months, average age in months squared, assessment center, and average measurement year, followed by inverse normal transformation conducted in males and females separately. The combined male and female BMI Z-scores were then used together for genetic association testing. All GWAS were run in BOLT-lmm adjusted for 10 genetic PCs (calculation described above) and genotyping array.

Genetic instruments for each of the three traits consisted of the complete set of index SNPs for each independent genome-wide significant GWAS locus defined by clumping signals in 500 kb windows. In addition to inverse-variance weighted (IVW) MR, we also conducted MR Egger to detect pleiotropic effects and Weighted Median (WM) MR to allow for the inclusion of up to 50% invalid genetic instruments. To test the robustness of any causal association, we repeated MR with filtered genetic instruments using Steiger filtering to remove variants likely influenced by reverse causation and Cook's distance filtering to remove outlying heterogeneous variants.[57] First-pass bidirectional MR included all genome-wide significant SNPs using IVW, MR Egger, and WM using the MendelianRandomization R package.[58] MR sensitivity analysis

was conducted by first Steiger Filtering to remove variants that explained more variance in the outcome than the exposure as determined by the get_r_from_pn command in TwoSampleMR R package[57,59] and then by filtering out Cook's distance outliers using base R functions. IVW, MR Egger, and WM were then repeated on the filtered genetic instruments. Validation of educational attainment MR results were conducted in two ways. First, a completely external GWAS dataset was used for which 74 discovery stage (not including UKB) genome-wide significant index SNP summary statistics were published online.[25] After removing missing and ambiguous SNPs, the external genetic instrument contained 66 variants. Second, we split the UKB sample into two subsets and conducted GWAS for PC1 and educational attainment as described above in each subset: 1/3 (N_{pc1}= 149,212 and N_{ed}= 150,884) of the EUR sample was used for genetic instrument variable identification and 2/3 (N=300,000) of the EUR sample was used for testing. MR of the tea intake SNP rs1453548 was conducted using the wald ratio estimate as calculated by the MendelianRandomization R package[58] using effect size and standard errors from our "cups of tea per day" GWAS and effect size and standard errors from 4,358 traits in the Neale Lab GWAS.

Data Availability

All 170 derived dietary habits will be returned and shared through UK Biobank and all GWAS results for the 143 significantly heritable dietary habits will be made publically available on the Type 2 Diabetes Knowledge Portal (http://www.type2diabetesgenetics.org/) upon publication.

Acknowledgements

J.B.C. is supported by NIH NIDDK T32 training grant (5T32DK110919-02) and American Diabetes Association postdoctoral fellowship grant (1-19-PDF-028). Work was conducted under UK Biobank application 11898.

Author Contributions

J.B.C. conceptualized, analyzed, and interpreted the data and wrote the manuscript. J.N.H. and J.C.F. conceptualized and supervised the analysis and revised the manuscript.

References

1 Lim, S. S. *et al.* A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: A systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2224-2260, doi:https://doi.org/10.1016/S0140-6736(12)61766-8 (2012).

2 U. S. Burden of Disease Collaborators. The state of us health, 1990-2010: Burden of diseases, injuries, and risk factors. *JAMA* **310**, 591-606, doi:10.1001/jama.2013.13805 (2013).

3 Kelly, T., Yang, W., Chen, C. S., Reynolds, K. & He, J. Global burden of obesity in 2005 and projections to 2030. *Int. J. Obes.* **32**, 1431, doi:10.1038/ijo.2008.102. (2008).

4 Boyle, J. P., Thompson, T. J., Gregg, E. W., Barker, L. E. & Williamson, D. F. Projection of the year 2050 burden of diabetes in the US adult population: Dynamic modeling of incidence, mortality, and prediabetes prevalence. *Popul. Health Metr.* **8**, 29 (2010).

5 Reed, D. R., Bachmanov, A. A., Beauchamp, G. K., Tordoff, M. G. & Price, R. A. Heritable variation in food preferences and their contribution to obesity. *Behav. Genet.* **27**, 373-387 (1997).

6 Meddens, S. F. W. *et al.* Genomic analysis of diet composition finds novel loci and associations with health and lifestyle. *bioRxiv*, doi:10.1101/383406 (2018).

7 Chu, A. Y. *et al.* Novel locus including FGF21 is associated with dietary macronutrient intake. *Hum. Mol. Genet.* **22**, 1895-1902 (2013).

8 Tanaka, T. *et al.* Genome-wide meta-analysis of observational studies shows common genetic variants associated with macronutrient intake. *Am. J. Clin. Nutr.* **97**, 1395-1402 (2013).

9 Merino, J. *et al.* Genome-wide meta-analysis of macronutrient intake of 91,114 European ancestry participants from the cohorts for heart and aging research in genomic epidemiology consortium. *Mol. Psychiatry*, doi:10.1038/s41380-018-0079-4 (2018).

10 Merino, J. *et al.* Multi-trait genome-wide association meta-analysis of dietary intake identifies new loci and genetic and functional links with metabolic traits. *bioRxiv*, 623728, doi:10.1101/623728 (2019).

11 Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat. Genet.* **50**, 1593-1599, doi:10.1038/s41588-018-0248-z (2018).

12 Hu, F. B. *et al.* Prospective study of major dietary patterns and risk of coronary heart disease in men. *Am. J. Clin. Nutr.* **72**, 912-921 (2000).

13 Nettleton, J. A., Polak, J. F., Tracy, R., Burke, G. L. & Jacobs, J. D. R. Dietary patterns and incident cardiovascular disease in the Multi-Ethnic Study of Atherosclerosis. *Am. J. Clin. Nutr.* **90**, 647-654, doi:10.3945/ajcn.2009.27597 (2009).

14 Appel, L. J. *et al.* A clinical trial of the effects of dietary patterns on blood pressure. *N. Engl. J. Med.* **336**, 1117-1124 (1997).

15 Hu, F. B. *et al.* Reproducibility and validity of dietary patterns assessed with a food-frequency questionnaire. *Am. J. Clin. Nutr.* **69**, 243-249 (1999).

16 Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106**, 9362-9367, doi:10.1073/pnas.0903103106 (2009).

17 Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005-D1012, doi:10.1093/nar/gky1120 (2018).

18 McRae, J. F. *et al.* Identification of regions associated with variation in sensitivity to food-related odors in the human genome. *Curr. Biol.* **23**, 1596-1600, doi:10.1016/j.cub.2013.07.031 (2013).

19 Ho, C.-T., Zheng, X. & Li, S. Tea aroma formation. *Food Science and Human Wellness* **4**, 9-27, doi:https://doi.org/10.1016/j.fshw.2015.04.001 (2015).
20 Burdock, G. A. *Fenaroli's handbook of flavor ingredients*. (CRC press, 2016).
21 Wu, K. *et al.* Dietary patterns and risk of colon cancer and adenoma in a cohort of men (United States). *Cancer Causes Control* **15**, 853-862, doi:10.1007/s10552-004-1809-2 (2004).
22 van Dam, R. M., Rimm, E. B., Willett, W. C., Stampfer, M. J. & Hu, F. B. Dietary Patterns and Risk for Type 2 Diabetes Mellitus in U.S. Men. *Ann. Intern. Med.* **136**, 201-209, doi:10.7326/0003-4819-136-3-200202050-00008 (2002).
23 Sniekers, S. *et al.* Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat. Genet.* **49**, 1107, doi:10.1038/ng.3869 (2017).
24 Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197-206, doi:10.1038/nature14177 (2015).
25 Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539, doi:10.1038/nature17671 (2016).
26 Nikpay, M. *et al.* A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121, doi:10.1038/ng.3396 (2015).
27 Scott, R. A. *et al.* An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* **66**, 2888-2902, doi:10.2337/db16-1253 (2017).
28 Jaeger, Sara R. *et al.* A Mendelian trait for olfactory sensitivity affects odor experience and food selection. *Curr. Biol.* **23**, 1601-1605, doi:https://doi.org/10.1016/j.cub.2013.07.030 (2013).
29 Wu, L. *et al.* Analysis of the aroma components in tobacco using combined GC-MS and AMDIS. *Anal. Methods* **5**, 1259-1263, doi:10.1039/C2AY26102B (2013).
30 Gelernter, J. *et al.* Genome-wide association study of alcohol dependence: significant findings in African-and European-Americans including novel risk loci. *Mol. Psychiatry* **19**, 41 (2014).
31 Jorgenson, E. *et al.* Genetic contributors to variation in alcohol consumption vary by race/ethnicity in a large multi-ethnic genome-wide association study. *Mol. Psychiatry* **22**, 1359 (2017).
32 Bierut, L. J. *et al.* ADH1B is associated with alcohol dependence and alcohol consumption in populations of European and African ancestry. *Mol. Psychiatry* **17**, 445 (2012).
33 Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237-244, doi:10.1038/s41588-018-0307-5 (2019).
34 Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112-1121, doi:10.1038/s41588-018-0147-3 (2018).
35 Burgess, S., Davies, N. M. & Thompson, S. G. Bias due to participant overlap in two-sample Mendelian randomization. *Genet. Epidemiol.* **40**, 597-608 (2016).
36 O'Neil, M. J. *The Merck index: an encyclopedia of chemicals, drugs, and biologicals*. (RSC Publishing, 2013).
37 Willett, W. & Stampfer, M. J. Total energy intake: implications for epidemiologic analyses. *Am. J. Epidemiol.* **124**, 17-27, doi:10.1093/oxfordjournals.aje.a114366 (1986).
38 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, doi:10.1371/journal.pmed.1001779 (2015).
39 Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv*, 166298 (2017).

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

41 The Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68, doi:10.1038/nature15393 (2015).

42 Bellenguez, C. *et al.* A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics* **28**, 134-135, doi:10.1093/bioinformatics/btr599 (2012).

43 Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-2849 (2016).

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

45 Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906-908, doi:10.1038/s41588-018-0144-6 (2018).

46 Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348-354 (2010).

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

48 Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776-2778, doi:10.1093/bioinformatics/btx299 (2017).

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

50 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122, doi:10.1186/s13059-016-0974-4 (2016).

51 Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nature communications* **6** (2015).

52 Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493-1501, doi:10.1093/bioinformatics/btw018 (2016).

53 Benner, C. *et al.* Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *Am. J. Hum. Genet.* **101**, 539-551 (2017).

54 Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336-2337 (2010).

55 Pers, T. H., Timshel, P. & Hirschhorn, J. N. SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinformatics* **31**, 418-420, doi:10.1093/bioinformatics/btu655 (2014).

56 Ge, T. *et al.* The shared genetic basis of educational attainment and cerebral cortical morphology. doi:10.1093/cercor/bhy216 (2018).

57 Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLOS Genetics* **13**, e1007081, doi:10.1371/journal.pgen.1007081 (2017).

58 Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* **46**, 1734-1739 (2017).

59 Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**, e34408 (2018).