OXFORD

ASSOCIATION STUDIES ARTICLE

Genome-wide gene–diet interaction analysis in the UK Biobank identifies novel effects on hemoglobin A1c

Kenneth E. Westerman[1,2,3], Jenkai Miao[4], Daniel I. Chasman[5,6,7,8], Jose C. Florez[2,3,9], Han Chen[10,11], Alisa K. Manning[1,2,3] and Joanne B. Cole[2,4,9,*]

[1]Clinical and Translational Epidemiology Unit, Mongan Institute, Massachusetts General Hospital, Boston, MA 02114, USA, [2]Programs in Metabolism and Medical & Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA, [3]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA, [4]Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA 02115, USA, [5]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA, [6]Division of Genetics, Brigham and Women's Hospital, Boston, MA 02115, USA, [7]Medical and Population Genetics Program, Broad Institute, Cambridge, MA 02142, USA, [8]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA, [9]Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA, [10]Human Genetics Center, Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA and [11]Center for Precision Health, School of Public Health and School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

*To whom correspondence should be addressed at: Broad Institute of MIT and Harvard, 75 Ames Street #10026D, Cambridge, MA, 02142, USA. Tel: 617-714-7359; Email: jcole@broadinstitute.org

Abstract

Diet is a significant modifiable risk factor for type 2 diabetes (T2D), and its effect on disease risk is under partial genetic control. Identification of specific gene–diet interactions (GDIs) influencing risk biomarkers such as glycated hemoglobin (HbA1c) is a critical step towards precision nutrition for T2D prevention, but progress has been slow due to limitations in sample size and accuracy of dietary exposure measurement. We leveraged the large UK Biobank (UKB) cohort and a diverse group of dietary exposures, including 30 individual dietary traits and 8 empirical dietary patterns, to conduct genome-wide interaction studies in ~340 000 European-ancestry participants to identify novel GDIs influencing HbA1c. We identified five variant-dietary trait pairs reaching genome-wide significance ($P < 5 \times 10^{-8}$): two involved dietary patterns (meat pattern with rs147678157 and a fruit & vegetable-based pattern with rs3010439) and three involved individual dietary traits (bread consumption with rs62218803, dried fruit consumption with rs140270534 and milk type [dairy vs. other] with 4:131148078_TAGAA_T). These were affected minimally by adjustment for geographical and lifestyle-related confounders, and four of the five variants lacked genetic main effects that would have allowed their detection in a traditional genome-wide association study for HbA1c. Notably, multiple loci near transient receptor potential subfamily M genes (*TRPM2* and *TRPM3*) interacted with carbohydrate-containing food groups. These interactions were further characterized using non-European UKB subsets and alternative measures of glycaemia (fasting glucose and follow-up HbA1c

measurements). Our results highlight GDIs influencing HbA1c for future investigation, while reinforcing known challenges in detecting and replicating GDIs.

Introduction

Diet is a major modifiable risk factor for type 2 diabetes and related cardiometabolic traits, but a complete understanding of the therapeutic potential of dietary intervention is limited by the high inter-individual variability in glycaemic response (1). The relationship between diet and hyperglycaemia is known to be influenced by genetics (2,3), suggesting that gene–diet interaction (GDI) analysis for glycaemic traits has the potential to uncover novel genetic variants that modify the effect of dietary intake on glycaemic response and diabetes risk. This approach can contribute to our understanding of diet-related health disparities between populations and inform the development of precision nutrition.

Despite much interest in GDIs over the past two decades, there have been few robust discoveries (4). This can be attributed in part to the low statistical power of gene–environment interaction (GEI) analysis in general, which requires approximately four times the sample size of a genetic main effect analysis to detect associations of similar strength (5,6). The recent availability of both cohorts with hundreds of thousands of individuals, such as the UK Biobank (UKB), and computationally efficient software programs for biobank-scale GEI analysis represents an opportunity to overcome power limitations and identify novel interactions of modest effect size (7,8). Recent analyses at loci with genetic main effects in UKB have identified GEIs impacting body mass index (BMI), using a series of exposures including, but not limited to, food frequency questionnaire (FFQ)-based diet measurements (9,10). However, the use of self-reported dietary data as environmental exposures compounds the power limitations of GEI analysis. FFQ-based intake estimates are imprecise and influenced by numerous potentially confounding socio-cultural factors (11), which may bias interaction estimates toward the null hypothesis and further hamper discovery. Empirical dietary patterns that combine correlated single foods or nutrients provide one approach to deriving more meaningful and robust representations of dietary behaviors. For example, a recent genome-wide association study of dietary behaviors in UKB identified genetic signals that were unique to either individual foods or dietary patterns, emphasizing the potential for discovery when using complementary and multivariate dietary traits (12).

Here, we leveraged the large-scale UKB dataset along with a diverse set of dietary exposures to search for novel GDIs impacting glycated hemoglobin (HbA1c), a blood biomarker of hyperglycaemia and diagnostic biomarker for T2D. The primary interaction analysis was conducted genome-wide in a subset of unrelated individuals of European ancestry without diabetes (N ∼ 340 000), using a series of individual diet traits (e.g. cooked vegetable intake and level of fat in milk) as well as principal components analysis (PCA)-based dietary patterns. We identified five loci reaching genome-wide significance ($P < 5 \times 10^{-8}$) and conducted comprehensive characterization of their associations with sensitivity analyses incorporating known confounders, multi-ancestry replication, and analysis using related glycaemic traits.

Results

Dietary behaviors associate with HbA1c

Thirty dietary traits capturing type and overall frequency of a variety of foods and beverages were derived from brief FFQs in an unrelated subset of European-ancestry participants from the UK Biobank (workflow in Fig. 1; UKB population characteristics in Supplementary Material, Table S1). These traits included semi-quantitative estimates of consumption frequency (such as cooked vegetables, bread, coffee, and alcohol), food or drink types (such as milk type [dairy vs. non-dairy milk] and decaffeinated vs. caffeinated coffee), and avoidance of specific foods (such as wheat) (Supplementary Material, Table S2). Moderate correlation between many traits, with a maximum Pearson correlation of 0.43 between beef and lamb/mutton intake (Supplementary Material, Fig. S1), motivated the derivation of both global and food sub-group dietary patterns, combining correlated foods into summary variables using PCA. Eight empirical dietary patterns (dietary principal components, or dPCs) were derived from the first principal component of seven group-specific PCA on related foods (fruit & vegetable, drinks, fish, meat, grains, dairy, and food avoidance; see Supplementary Material, Table S2) and one PCA on the full set of 30 dietary traits (global dPC) (Fig. 2A). The global dPC, which explained 9.4% of the total variance in all dietary traits, remained reflective of a general 'prudent' dietary pattern as has been previously observed in UKB (12).

To first understand the relationship between these dietary traits and HbA1c (collected from individuals on the same day) independently of genetic background, we regressed HbA1c on each of the dietary traits, finding associations at the majority of dietary patterns and individual traits (32 of 38 traits at Bonferroni-corrected $P < 0.05/38 = 0.0013$) (Fig. 2B). The strongest associations were observed for the grains dPC (0.30 mmol/mol HbA1c per dPC standard deviation, 95% CI: 0.29 to 0.31, $P < 1 \times 10^{-300}$), which loads most strongly on bread type (increased white vs. brown or whole grain bread), the 'prudent' global dPC (0.30 mmol/mol HbA1c per standard deviation, 95% CI: 0.28 to 0.31, $P < 1 \times 10^{-300}$), which is also largely influenced by bread type, and alcohol intake (−0.316 mmol/mol HbA1c per standard deviation, 95% CI: −0.32 to −0.29, $P < 1 \times 10^{-300}$). Many of these associations were substantially attenuated by adjustment for geographic and lifestyle covariates, indicating the presence of partial confounding (between 30 and 40% reduction in effect size for the grains dPC, global dPC, and processed meat intake, though these relationships remained highly significant). Others, such as alcohol intake and bread intake, remained equally strong in the adjusted model.

Genome-wide interaction studies

Genome-wide interaction studies (GWIS) were performed for each of the 8 dPCs and 30 individual dietary trait exposures to identify loci whose effect on HbA1c is modified by dietary intake. Across the 38 GWIS, nine variants reached genome-wide significance ($P < 5 \times 10^{-8}$) for their interaction effect,
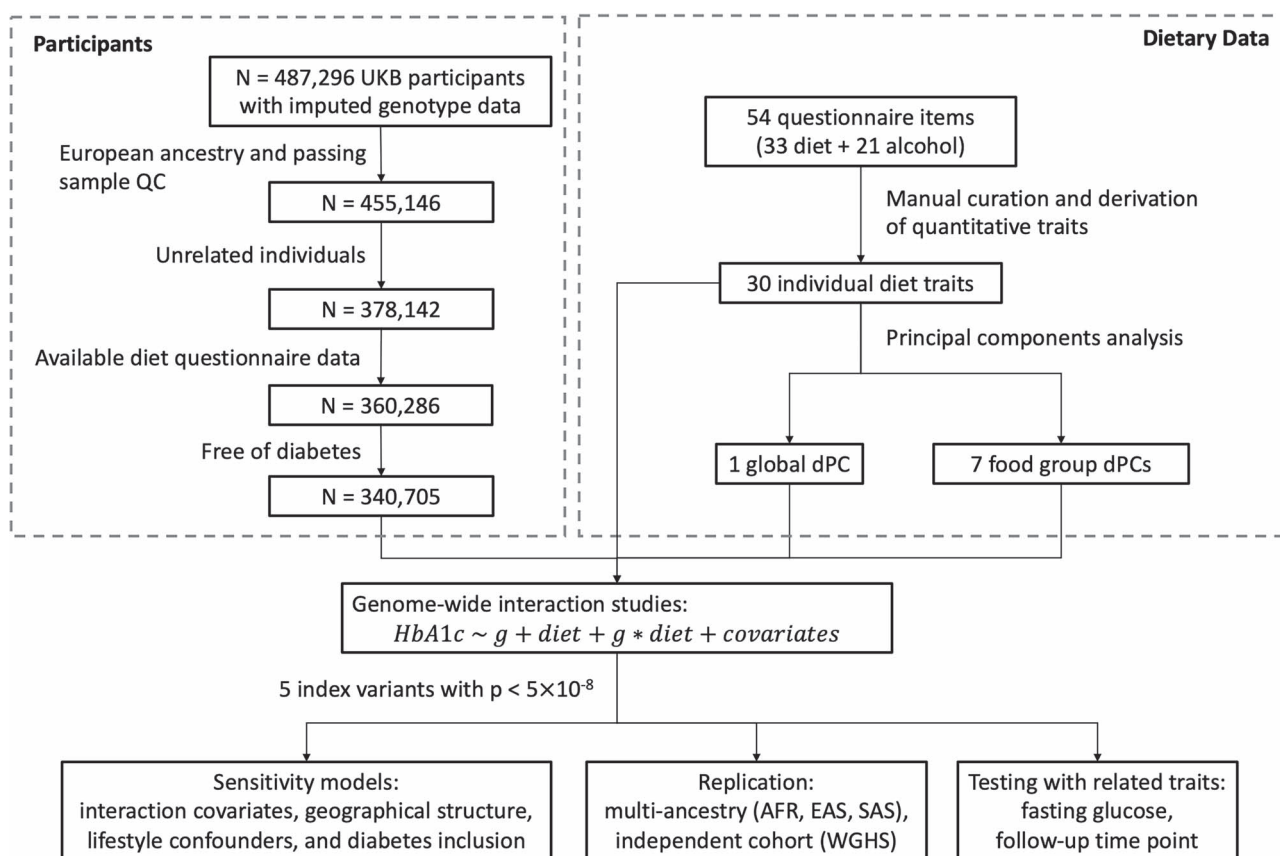
Figure 1. Participant selection and analysis workflow. UK Biobank participants underwent filtering based on genotyping quality control, relatedness, ancestry, availability of dietary data, and diabetes status. Dietary traits were curated from FFQ responses and dietary patterns were derived from principal components analysis. All 38 dietary traits were used in genome-wide interaction studies for HbA1c, with follow-up sensitivity, replication, and within-population analyses using related glycaemic traits.

constituting five independent loci after clumping by 500 kb windows (Fig. 3; Table 1). None of these signals passed a more stringent threshold adjusted for multiple testing of 18.4 effective traits ($P < 5 \times 10^{-8}/18.4 = 2.72 \times 10^{-9}$; see Materials and Methods).

Two of the five lead GDIs involved food group dPCs (meat dPC-rs147678157 and fruit & vegetable dPC-rs3010439) and three involved individual dietary traits (bread consumption-rs62218803, dried fruit consumption-rs140270534, and milk type [dairy vs. other]-4:131148078_TAGAA_T). After visual inspection of the regional association plots (Supplementary Material, Fig. S4–S8), we note that one of these interactions, involving rs147678157, should be evaluated with particular caution due to lower minor allele frequency (0.008), modest imputation quality score (INFO = 0.76), and relative lack of variants in linkage disequilibrium with similar interaction associations. Main effects of the interacting dietary traits on HbA1c were all strong, while genetic effects (as calculated from a model without any interaction term) were absent other than a small association at the meat-interacting rs147678157 (Table 1). To better characterize the interactions at our lead variants, we conducted two sets of stratified analyses: testing the genetic effects on HbA1c after stratifying our dietary exposures into quintiles and testing the dietary effects on HbA1c after stratifying by genotype. Genotypic effects all displayed a qualitative interaction pattern, in which the sign of the genetic effect differed across strata of dietary behaviors (Supplementary Material, Fig. S2). This accounts for the lack of

marginal genetic effect of these variants on HbA1c. In contrast, we found evidence of quantitative interactions, in which the dietary effect changed only magnitude across genotype groups, at three out of the five loci (rs147678157, rs62218803, and rs3010439; Fig. 3B). For example, in those with two copies of the reference allele at rs62218803, each standard deviation of bread consumption associates with an HbA1c increase of 0.22 mmol/mol. The strength of this association decreases by 37% to 0.14 mmol/mol in those with two copies of the alternate allele, meaning that bread consumption may be less detrimental in these individuals.

Of the primary interactions, two were in or near transient receptor potential subfamily M genes, *TRPM2* and *TRPM3*, despite their genomic location in two different chromosomes (21 and 9, respectively). Their protein products, as well as other transient receptor potential (TRP) family proteins, have been implicated in insulin secretion and downstream glycaemia (13). Furthermore, some of these index variants had marginal interaction effects on HbA1c with other dietary traits (such as the meat dPC-interacting rs147678157 with the global dPC and tea intake, as well as the dried fruit-interacting rs140270534 with the fruit & vegetable dPC), whereas other lead GDIs were apparently unique to the dietary exposure used in discovery (such as the milk type-interacting 4q28.3 variant; Supplementary Material, Table S3). Each dPC-interacting variant showed some degree of interaction with at least one of that dPC's constituent traits, indicating that the use of PCA-based dietary patterns increased our power
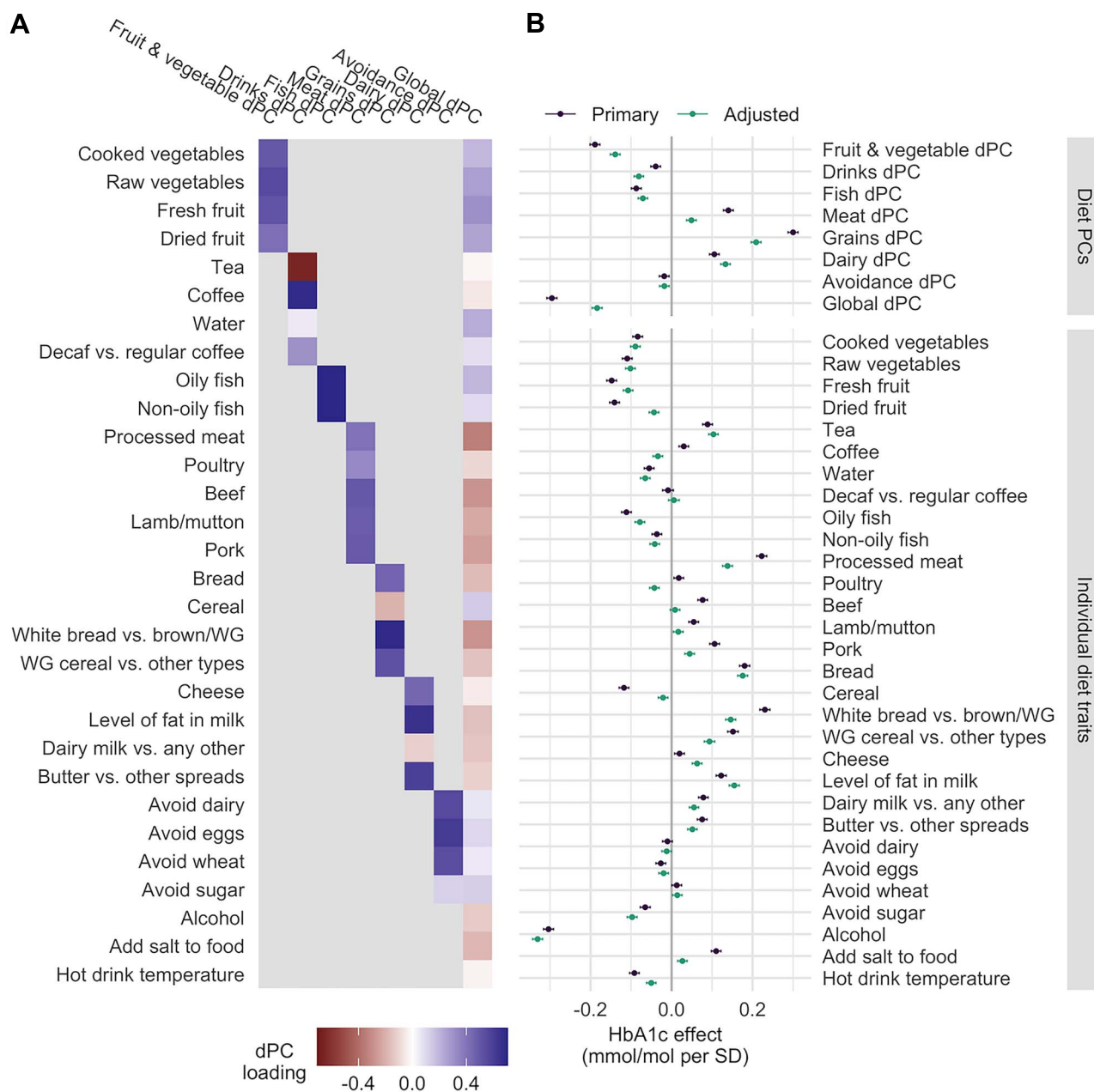
Figure 2. Dietary traits and associations with HbA1c. (**A**) Heatmap displays loadings for each of the 30 dietary traits characterizing the seven food group dPCs and one global dPC. Higher positive loadings denote higher contributions of the trait (*y*-axis) to the associated dPC (*x*-axis). (**B**) Forest plot of coefficients from regression of HbA1c on each dietary trait. Models included either the same set of covariates as in the primary GWIS ('Primary'; purple) or those from sensitivity model 3 ('Adjusted'; adjusted for birthplace, assessment center, BMI, smoking, educational attainment and physical activity; purple). Error bars correspond to 95% confidence intervals.

for the detection of interactions. For example, *P*-values for rs147678157 were between 1×10^{-6} and 1×10^{-3} for four of five individual meat consumption traits, while the meat dPC interaction reached $P = 5.45 \times 10^{-9}$.

Sensitivity models were conducted at all five index variants to assess the potential effects of confounding and interaction model biases (see Methods). Interaction effect estimates were consistent after adjustment for genotype×covariate interactions (sensitivity model 1 [SM1]), birthplace and study center (SM2), and BMI, smoking, physical activity, and educational attainment (SM3) (Supplementary Material, Fig. S3, Supplementary Material, Table S4). The effects were only modestly attenuated when including individuals with diabetes

(SM4), whose diagnosis and treatment would directly affect HbA1c via both behavioral changes and medication use. The small degree of change in the interaction effect estimates across these sensitivity models indicates that the observed relationships are not primarily due to confounding by known correlates of diet such as socioeconomic status or a generally healthy lifestyle.

Replication across ancestries and cohorts

To assess the translation of these GDIs in non-European populations, we conducted a multi-ancestry replication of the five significant interactions in individuals of West African
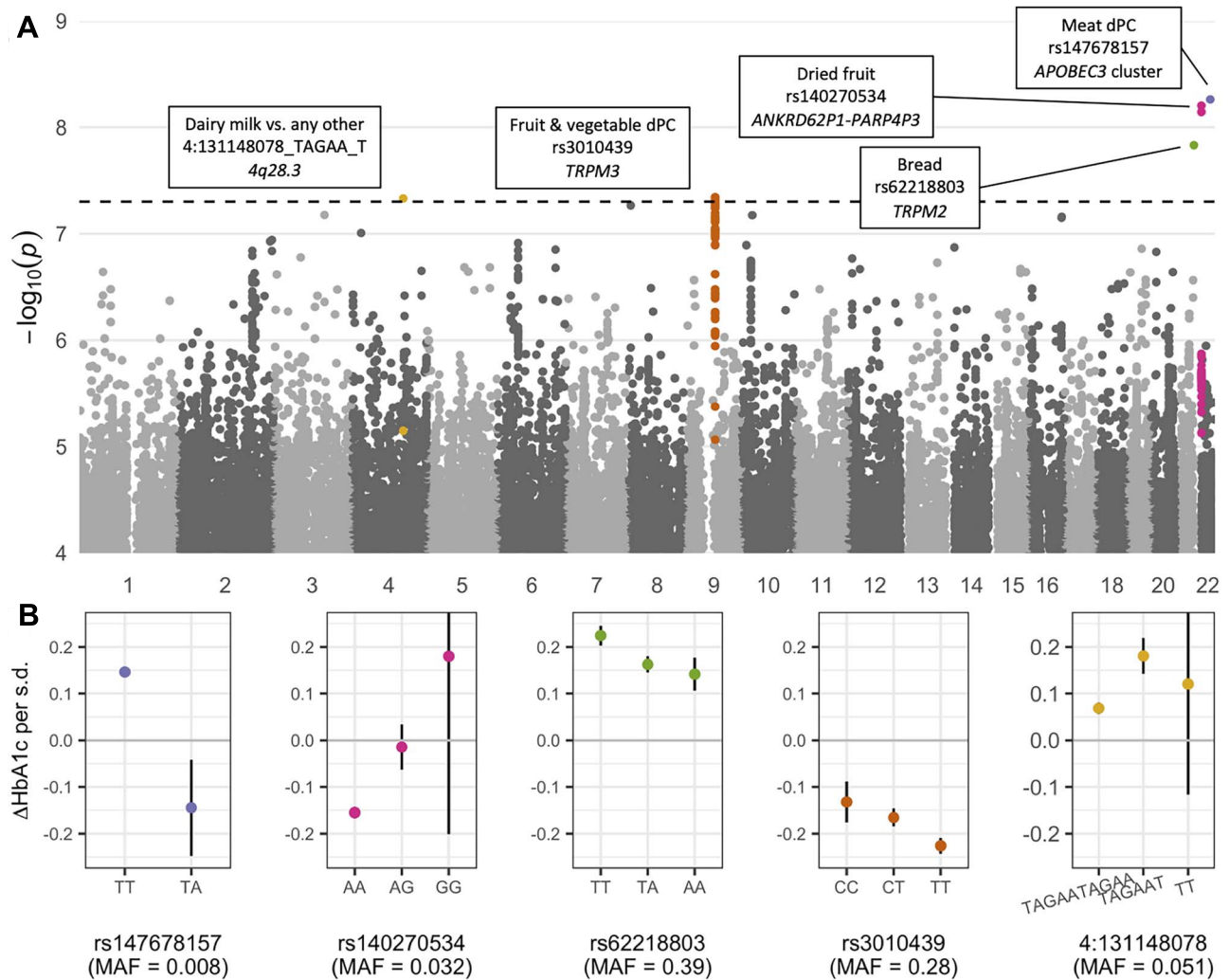
Figure 3. Genome-wide GDI analysis results. (**A**) Manhattan plot shows minimum -log$_{10}$ (*P*-values) as a function of genomic position across all 38 GWIS. Colors correspond to independent loci with index variants having $P < 5 \times 10^{-8}$, and variants are assigned to a peak if within 50 kb of an index variant and having $P < 1 \times 10^{-5}$. Index variants are annotated with the trait, variant rsID (where available), and annotated gene. (**B**) Dietary effect sizes on HbA1c (change in HbA1c associated with a standard deviation increase in the associated dietary trait; y-axis) after stratification by genotype (closest genotype based on dosages; x-axis), using the primary analysis covariates. Points and bars denote stratified main effect estimates and 95% confidence intervals from the primary model. Genotype-stratified effects are shown only for genotype groups containing >100 individuals. Effect estimates are in terms of mmol/mol HbA1c. s.d.: standard deviation.

(AFR; $N = 4763$ after sample QC), East Asian (EAS; $N = 1936$), and South Asian (SAS; $N = 5849$) ancestry based on the same dietary data in UKB. Although we had low statistical power to detect these signals (Supplementary Material, Table S5), evidence for replication across ancestries would provide more robust support for these interactions. One variant-trait pair, rs62218803 and bread consumption (slices per day), was nominally significant (beta = −0.38, $P = 0.022$) and directionally consistent in the East Asian ancestry group. One nominal and directionally consistent interaction was observed for a dPC-interacting variant with the individual contributing dietary trait (Supplementary Material, Table S6), specifically for the fruit & vegetable dPC locus (cooked vegetables-rs3010439 in EAS, beta = −0.24, $P = 0.012$). Overall, the majority of variant-trait-ancestry combinations did not show meaningful replication, though we note the limited power afforded by the much smaller sample size.

To explore this further, we gathered all variant-trait pairs reaching suggestive significance ($P < 1 \times 10^{-5}$) in the primary analysis and tested for an enrichment of shared interaction coefficient signs in the multi-ancestry analysis. As shown in Table 2, we did not find evidence of such an enrichment in any of the ancestries, an observation that could be attributed partially to lower sample sizes, differences in dietary intake and its socio-cultural correlates, and known differences in genetic architecture (e.g. haplotype blocks, allele frequencies, and causal variants).

Follow-up analyses at lead GDIs

To further explore these top loci, we tested for GDIs using follow-up FFQ and HbA1c data (again, collected on the same follow-up visit day) from a subset of UKB individuals from the primary analysis that participated in a repeat assessment center visit between 2012–2013 ($N = 10\,644$) and for GDIs using fasting glucose from the subset of individuals who were fasting at their baseline visit ($N = 10\,879$). The follow-up analysis can provide evidence for the stability of these interactions over time, and

Table 1. Loci reaching genome-wide significance ($P < 5 \times 10^{-8}$) in primary GWIS

Dietary trait	Index variant	Chromosome: Position	Annotated gene	Effect allele	Effect allele frequency	Diet main effect	Genetic main effect	Interaction effect—Primary	Interaction effect—Adjusted
Meat dPC	rs147678157	chr22:39373897	APOBEC3B	A	0.008	0.14 (0.0062); 7.2×10^{-115}	−0.13 (0.054); 0.016	−0.21 (0.035); 5.5×10^{-9}	−0.18 (0.036); 4.7×10^{-07}
Dried fruit	rs140270534	chr22:17141482	ANKRD62P1-PARP4P3	G	0.032	−0.14 (0.0062); 5.1×10^{-115}	0.016 (0.025); 0.53	0.17 (0.03); 6.3×10^{-09}	0.17 (0.03); 2.2×10^{-08}
Bread	rs62218803	chr21:45863307	TRPM2	C	0.392	0.18 (0.0064); 2.1×10^{-173}	−0.014 (0.011); 0.19	−0.061 (0.011); 1.5×10^{-08}	−0.059 (0.011); 7.6×10^{-08}
Fruit & vegetable dPC	rs3010439	chr9:73594857	TRPM3	T	0.72	−0.19 (0.0062); 4.7×10^{-202}	−0.016 (0.0096); 0.11	−0.041 (0.0075); 4.5×10^{-08}	−0.038 (0.0077); 1.1×10^{-06}
Dairy milk vs. any other	4:131148078_TAGAA_T	chr4:131148078	Intergenic	T	0.051	0.079 (0.0061); 2.4×10^{-38}	−0.017 (0.02); 0.4	0.38 (0.07); 4.7×10^{-08}	0.39 (0.07); 3.1×10^{-08}

Effects are shown as: coefficient (standard error); p-value. Effect sizes are shown as mmol/mol HbA1c per allele for genetic effects and mmol/mol HbA1c per standard deviation (or per binary change in milk type) for diet effects. Both main effects come from models without an interaction effect included. Adjusted model includes assessment center, birthplace, BMI, smoking, educational attainment, physical activity, and genotype interaction terms for all covariates (corresponding to sensitivity model 3 described in section Materials and Methods).

the fasting glucose analysis indicates whether these interactions are generalizable to other measures of hyperglycaemia. Though significance levels were weak, possibly due to sample size limitations, the directions of the interaction effects were consistent with the primary analysis for all variant-trait pairs (Supplementary Material, Table S7). The strongest associations included the meat dPC-rs147678157 interaction impacting fasting glucose ($P = 0.02$) and the milk type-4:131148078_TAGAA_T locus interaction evaluated at follow-up ($P = 0.005$). Given the generally lower statistical power for these analyses, we tested for an enrichment of directional consistency compared to the primary analysis across all 959 suggestive variant-trait pairs ($P < 1 \times 10^{-5}$ in any of the 38 GWIS), and found strong evidence for enrichment using a subset of the same European individuals (66% directionally consistent with $P = 8.3 \times 10^{-26}$ for follow-up and 57% directionally consistent with $P = 3.9 \times 10^{-6}$ for fasting glucose), an enrichment we did not find among our non-European replication analyses (Table 2).

To better characterize these loci at a molecular level, we used PhenoScanner to examine cross-trait associations with DNA methylation, gene expression, metabolites, and proteins, including potential proxy variants with $r^2 > 0.8$ in Europeans. Variant rs3010439 (fruit & vegetable dPC-interacting), lies within an intron of TRPM3 and has strong associations with the methylation of nearby CpG sites within TRPM3 (including cg20555507 [$P = 4.42 \times 10^{-14}$] and proxy variant rs3095766 with cg14165911 [$P = 8.61 \times 10^{-286}$]). The other TRPM variant, rs62218803 (bread-interacting), is located less than 400 bp downstream of TRPM2, yet there were no significant ($P < 5 \times 10^{-8}$) expression-quantitative trait loci (eQTLs) in any tissues contained in the associated GTEx dataset. However, approximately 100 kb upstream of rs62218803, and in complete linkage equilibrium, is a missense coding variant in PFKL with no interaction effect ($P = 0.92$) but a strong genetic main effect on HbA1c in our analysis (rs17850433, $P = 7.8 \times 10^{-23}$; Supplementary Material, Fig. S4). Variant rs140270534 (dried fruit-interacting) lies in an intron of the pseudogene ANKRD62P1-PARP4P3 and acts as an eQTL for XKR3 in whole blood ($P = 5.52 \times 10^{-13}$ in the eQTLGen dataset (14)). XKR3 codes for a homolog of the putative membrane transporter and Kell blood group component XK, but has no clear link to glycaemic regulation.

To test the robustness of these interactions in an independent cohort, replication was sought in individuals of European ancestry from the US-based Women's Genome Health Study ($N = 20\,095$). Though we were only able to test for 3 variant-trait pairs surpassing genotype imputation quality filters, no interactions were detected at $P < 0.05$ (Supplementary Material, Table S8). We note that this lack of replication may reflect a variety of factors including heterogeneity in dietary behaviors and measurement as well as lack of statistical power.

Discussion

We have conducted the largest GDI study for HbA1c to-date, uncovering multiple interactions between genotype and dietary traits (both individual diet questionnaire items and empirical diet patterns) that influence HbA1c. We further characterized these loci through multi-ancestry replication and other validation analyses within UKB, and showed that these loci have the potential to have a meaningful impact on the diet-hyperglycaemia relationship.

HbA1c is both a diagnostic and prognostic biomarker for T2D, reflecting cumulative blood glucose over a two to

Table 2. Aggregate replication and validation of suggestive ($P < 1 \times 10^{-5}$) index variant-trait pairs

Dataset	Sample size	# variants tested[a]	% directionally consistent	P[b]
West African	4763	958	52%	0.16
East Asian	1936	486	48%	0.25
South Asian	5849	738	51%	0.96
Follow-up	10 644	959	66%	8.3×10^{-26}
Fasting glucose	10 879	959	57%	3.9×10^{-6}

[a]Variant numbers vary across ancestries due to the MAF > 0.005 criterion.
[b]Chi-square test *P*-value for shared interaction effect signs with primary analysis.

three-month period prior to measurement [13]. Therefore, it provides a long-term quantitative measure of glucose levels, making it an appealing outcome for GDI analysis with FFQ dietary exposures collected at the same study visit and which similarly reflect long term usual intake over a 1-year period. It is important to note that HbA1c is more strongly influenced by postprandial, rather than fasting, glucose blood glucose concentrations (15). In this light, it may not be surprising that our fasting glucose analysis in a subset of the primary population did not strongly validate our top interaction associations. Furthermore, while HbA1c is a valuable continuous measure for hyperglycaemia, it is also independently influenced by red blood cell characteristics, such that its genetic influences are often classified as either glycaemic or erythrocytic [15]. However, based on the dietary traits used as exposures and the candidate genes identified, it does not appear that these interactions are affecting HbA1c via red blood cell traits.

The primary analysis highlighted interactions within or near two independent *TRPM* subfamily genes. *TRPM2* and *TRPM3* are members of a larger family of 28 cation-permeable TRP channels, many of which are modulated by a diverse array of natural and dietary compounds (16). Both *TRPM2* and *TRPM3* have been shown to affect insulin secretion by beta cells in animal models (17,18), highlighting a potential role in glycaemic control. The *TRPM3* variant (rs3010439), which showed an interaction with the fruit and vegetable dPC in European-ancestry individuals and cooked vegetable intake in East Asian-ancestry individuals, is particularly interesting in that the TRPM3 protein product is inhibited by citrus fruit metabolites, specifically naringenin and hesperetin (19). In general, our finding of interactions between *TRPM*-annotated variants and multiple carbohydrate-heavy foods (bread and fruit) provides an avenue for further exploration.

The effect sizes of the interactions identified here may be practically meaningful towards the implementation of personalized nutrition approaches. For example, based on the stratified models, an increase of one standard deviation in bread consumption (approximately 8 slices of bread per week) associates with an HbA1c increase of 0.22 mmol/mol in reference homozygotes at rs62218803 (TRMP2). However, this effect size is decreased by 37% to 0.14 mmol/mol in alternate allele homozygotes. Likewise, the protective effect of the fruit and vegetable-increasing dietary pattern (0.13 mmol/mol per standard deviation in reference homozygotes) is magnified by 70% in alternate allele homozygotes at rs3010439 (*TRPM3*). Still, these effect estimates should be treated with caution, as they are not only susceptible to residual confounding, but also the Winner's Curse phenomenon, in which effects from hypothesis-free studies tend to be overestimated in discovery analyses (20).

The large sample size used in this study along with the use of data-driven dietary patterns improved our statistical power to uncover GDIs of small effect size. For example, the combination of related foods into empirical dietary patterns led to the identification of genome-wide significant interaction effects that were present with only modest significance with the individual foods themselves. However, we were not able to demonstrate replication of these interactions in the independent Women's Genome Health Study, and showed only partial nominal replication of the interactions in the secondary within-UKB analyses. This incongruence could be due to a variety of factors, including differences in dietary variability across ancestries and countries (US vs. UK), differences in the dietary questionnaires used, and differences in the extent to which these food groups act as proxies for underlying traits (such as socioeconomic status or general healthy behaviors). The difficulty in replication of GDIs was highlighted in the null findings of a systematic replication effort for gene–macronutrient interactions influencing type 2 diabetes risk in the EPIC-InterAct study (21). The authors emphasized rigor in reporting and replication in diverse populations as strategies to combat potential false positive findings. While our in-depth replication approach, which included diverse ancestries, independent cohorts, and alternative outcomes, was an important strength of this study, it also reinforces the difficulty in replicating GDIs, even among biologically relevant candidate genes. To address known obstacles to gene–environment interaction detection (22), we suggest that future investigations seek out populations with substantial variability in the dietary behaviors of interest and use dietary datasets with greater granularity. These more precise dietary datasets can complement the dPC-based discovery approach employed here by helping with 'phenotypic fine-mapping' to pinpoint the relevant dietary components.

Additional limitations should be considered when evaluating these results. First, the UKB FFQ contains a limited set of questions on high-level food behaviors and therefore does not enable the calculation of reliable nutrient or total energy intake estimates. While the use of dietary patterns, which reflect higher level patterns of behavior, may be less influenced by total energy intake (TEI), interactions involving these patterns as well as the individual dietary traits may still be susceptible to residual confounding by total energy intake. However, we note that adjustment for both BMI and physical activity, which are highly correlated with TEI (23) had a minimal impact on the results. Second, the individual dietary traits and dPCs used here are likely reflective to some extent of confounding factors such as socioeconomic status, as has been previously observed (12), and are susceptible to reverse causation (impacts of known diabetes risk or family history on dietary behaviors). While our sensitivity analyses aimed to control for socio-cultural effects on dietary intake, it remains possible that our top interactions are actually capturing additional lifestyle exposures, which, given their major

differences across ethnicities and countries, would explain our limited replication.

In summary, we have leveraged the large sample size of the UK Biobank and an empirical dietary pattern-based approach to conduct a well-powered scan for GDIs influencing HbA1c. We uncovered multiple biologically relevant genetic loci that may influence the relationship between diet and hyperglycaemia. Our findings indicate the value of GDI analysis in large-scale datasets and the use of composite dietary exposure variables, while also reinforcing the persistent difficulty in detecting and replicating interactions based on self-reported dietary questionnaires. This study can serve as a basis for future investigations into the interplay of lifestyle and genetics for type 2 diabetes prevention.

Materials and 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 and 2010 (24). Genotyping, imputation, and initial quality control on the genetic dataset have been described previously (25). 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. Furthermore, only genetic variants with minor allele frequency (MAF) > 0.005 and imputation INFO score > 0.5 were included, resulting in a set of approximately 11 100 000 autosomal variants for analysis. Work was conducted on genetic data release version 3, with imputation to both Haplotype Reference Consortium and 1000 Genomes Project (1KGP), under UK Biobank application 27 892. This work was conducted under a Not Human Subjects Research determination (NHSR-4298 at the Broad Institute of MIT and Harvard).

UK Biobank phenotypes

Phenotype data processing and other downstream analyses were conducted using R version 3.6.0 (26). HbA1c (UKB field 30 750) was measured in blood samples through high-performance liquid chromatography (HPLC) using a Bio-Rad VARIANT II Turbo platform. HbA1c values and effect sizes are reported here in units of mmol/mol, as originally provided by UKB. Type 1 and type 2 diabetes were defined by expanding 'probable' and 'possible' definitions of diabetes from a previously developed algorithm (27) to take into account information from repeat assessment center visits and touchscreen diabetes diagnosis (UKB field 2443). After the removal of individuals with type 1 or type 2 diabetes, outliers for HbA1c were removed using a cutoff of three interquartile ranges below or above the 25th and 75th percentiles, respectively.

Dietary data processing was based on the procedure described by Cole and colleagues (12). All dietary phenotype derivation was conducted on a homogenous population of unrelated individuals of European (EUR) ancestry (N = 340 705), as determined by: (1) limiting to unrelated, high quality samples used in ancestry PCA conducted by UKB (Bycroft 2018), (2) projection of genotypes onto 1KGP phase 3 PCA space, (3) outlier detection to identify the largest cluster of individuals in which all clustered individuals fell within 1KGP EUR PC1 and PC2 limits,

and (4) self-report as one of the following: 'British', 'Irish', 'Any other white background', 'White', 'Do not know', or 'Prefer not to answer'.

The UKB FFQ was completed during the same baseline study visit in which the blood draw for HbA1c measurement was taken. It consists of quantitative continuous variables (e.g. field 1289, tablespoons of cooked vegetables per day), ordinal non-quantitative variables depending on overall daily/weekly frequency (e.g. 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 Material, Table S2 provides a list of UKB fields relating to the corresponding FFQ question for each dietary habit, which can be viewed 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. A single high-resolution overall alcohol intake variable was derived by combining questions on either 'drinks per month' or 'drinks per week' for each alcoholic beverage type (answered by different individuals depending on their reported overall alcohol frequency (field 1558)) while median-imputing values for those who reported consuming alcohol but responded either 'Do not know' or 'Prefer not to answer' to specific drink frequency.

Quantitative traits (QTs) were then inverse rank normal transformed and missing data were median-imputed for all traits. Dietary patterns (dietary principal components, or dPCs) represent the first principal component from a series of PCA (using the *prcomp* function in base R), including seven PCA on groups of related dietary traits and one global PCA using all 30 dietary traits (Supplementary Material, Table S2). Given the substantial correlation between many of these traits, we derived an effective number of traits for multiple testing correction: PCA was conducted on the 38 scaled and mean-imputed traits, and the associated eigenvalues λ were used to calculate the number of effective traits as $\left(\sum_{k=1}^{38} \lambda_k \right)^2 / \sum_{k=1}^{38} \lambda_k^2 = 18.4$ (28).

Additional phenotypes were retrieved for use in sensitivity models. Physical activity was calculated in units of excess metabolic equivalents (METs) as previously described (29), with median-imputation used to replace ambiguous values ('Prefer not to answer', 'Do not know', or missing) for physical activity-related questionnaire items. Tobacco smoking behavior was assigned a score based on questionnaire responses: 0 ('No'), 1 ('Only occasionally'), or 2 ('Yes, on most or all days'). Educational attainment was derived in terms of US years of schooling equivalents as previously described (30), with median-imputation used to replace ambiguous values ('Prefer not to answer' or missing).

Genome-wide interaction studies

Basic power calculations were performed using the *genpwr* package for R (31), using parameters: outcome standard deviation = 3.71 (based on the HbA1c standard deviation in the European group), genetic main effect = 0 mmol/mol/allele, diet main effect = 0.15 mmol/mol/s.d., minor allele frequency = 0.25, and assuming an additive genetic model. For each dPC and individual dietary trait (38 exposures in total), a GWIS was performed using the following model:

$$HbA1c \sim g + D + g * D + covariates \tag{1}$$

where g represents the imputed genotype dosage and D represents the dietary exposure. Covariates in the primary model

included sex, age, age^2, and 10 genetic principal components to control for population stratification estimated using Flash-PCA2 (32). Interaction analysis was performed using *GEM* (Gene–Environment interaction analysis in Millions of samples) v1.1 (7) with robust standard error estimates to protect against bias due to heteroscedasticity and mis-specification of the genotype-HbA1c or diet-HbA1c relationship. Summary statistics from the primary GWIS for each of the 38 GWIS have been deposited in the T2D Knowledge Portal (https://t2d.hugeamp.org/).

An extensive set of sensitivity models were then applied to loci reaching genome-wide significance ($P < 5 \times 10^{-8}$). First, additional interaction parameters were included in the model between genotype and all covariates (sensitivity model 1 [SM1]) to adjust for possible confounding by the genotype × covariate interaction effect that is not accounted for by the covariate main effect (33). Second, birthplace and study center (and their interactions with genotype) were included as additional covariates (SM2) to further account for potential coincident ancestry and cultural substructure in the population. Third, BMI, smoking, physical activity, and educational attainment (and their interactions with genotype) were additionally added to the model (SM3) to address confounding by specific socio-cultural factors. Fourth, SM3 was repeated while including individuals with diabetes (SM4) to test for the potential of collider bias via the diabetes exclusion criterion in the primary model.

Top variants were assigned to genes based on the NCBI dbSNP database where available, and otherwise based on the nearest gene within 10 kb (in the case of the rs147678157-APOBEC3B assignment). PhenoScanner v2 (34,35) was used to explore variant associations with phenotypic traits and molecular quantities (e.g. eQTL as retrieved from the GTEx v7 release and methylation-quantitative trait loci (36)). Queries included proxy variants having $r^2 > 0.8$ with the query variant in European populations. Stratified analyses were performed using linear regression models in R and the same set of covariates as in the primary interaction analysis. Separate stratified models were performed with respect to the relevant dietary trait (quintiles) and genotype (three groups, with hard-calls generated by rounding dosages to the nearest genotype [0/1/2]). Stratified analyses were performed in two ways: testing for the effect of hard-called genotypes among each of 5 dietary trait quintiles, and vice versa, testing for the effect of the dietary trait in each of 3 genotype groups. Stratified models were tested using base linear regression models in R with the same set of covariates as in the primary interaction analysis.

Multi-ancestry replication and testing in related glycaemic traits

Multi-ancestry replication was performed using three non-European ancestry groups in UKB (West African, East Asian, and South Asian). Using 1KGP phase 3 as the training dataset, we built probabilistic Gaussian mixture models to represent normally-distributed subpopulations within the overall population, assigning data points to the multivariate normal components that maximize the component posterior probability. Ten-fold cross-validation was used with different initialization states and clustering was evaluated using the adjusted Rand index score. The model with the highest adjusted Rand index score for each ancestry was used to cluster individuals in UKB into ancestry subgroups based on their 1KGP projected PCs and self-reported ethnicities (field 2100). Multi-ancestry dietary and clinical phenotypes were derived as described above for European individuals, and dPCs were derived by projecting

ancestry-specific dietary traits onto the European principal component loadings. Interaction analysis was conducted for top variant-trait pairs under the primary model described above in Europeans.

To test the robustness of the top associations, two GDI analyses using related glycaemic traits (follow-up HbA1c and fasting glucose levels) were performed based on the primary model in Europeans. For the follow-up analysis, all variables were taken from the second assessment center visit (i.e. the first follow-up), with dPCs projected as described above in the multi-ancestry analysis. For the fasting glucose analysis, all phenotypes and analysis parameters were identical to the primary model in Europeans, other than the replacement of HbA1c as the outcome with fasting glucose, defined as random glucose in participants who reported having fasted for at least 8 hours prior to the blood draw (UKB fields 30 740 and 74).

General agreement between the primary results and replication results was tested via an enrichment test for shared interaction effect signs. Suggestive variant-trait pairs, defined as having interaction $P < 1 \times 10^{-5}$ in the primary analysis, were collected and pruned using 500 kb windows. For each within-UKB replication approach, a Chi-square test for shared signs of interaction effects was performed.

Replication was also undertaken in the Women's Genome Health Study, a prospective US-based cohort of ~23 000 females 45 years or older, described in detail elsewhere (37). After quality control procedures, the final interaction analysis was conducted in 20 095 females of European ancestry, free of a history of diabetes, and whose current diet had not 'greatly' changed in comparison to the last 5 years. The HbA1c phenotype was processed in a similar manner as described for UK Biobank, and was collected within an approximately 12-week overlapping period with the WGHS FFQ. Semi-quantitative dietary exposures and fruit and vegetable food groups were converted to continuous variables as described previously (38) and inverse normal transformed. Using an Rsq cutoff of 0.50, three variant-trait pairs were tested for replication. The rs140270534 (dried fruit) interaction was tested with overall fruit intake in WGHS, the rs3010439 (fruit and vegetable dPC) was tested with overall fruit and vegetable intake in WGHS, and the 4:131148078 indel (milk) was tested with milk consumption in WGHS derived by binarizing individuals who drink any amount of skim or whole milk versus those that answer never or rarely for both. Effect sizes on HbA1c were converted from % to mmol/mol.

Supplementary Material

Supplementary Material is available at *HMG* online.

Acknowledgements

Funding

Author Contributions

K.E.W.—conceptualization, data curation, formal analysis, software, writing—original draft preparation, J.M.—data curation, software, D.I.C.—validation, writing—review & editing, J.C.F.—supervision, writing—review & editing, H.C.—methodology, software, A.K.M.—conceptualization, funding acquisition, supervision, writing—review & editing and J.B.C.—conceptualization, data curation, supervision, writing—original draft preparation.

References

1. Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M. *et al.* (2015) Personalized nutrition by prediction of Glycemic responses. *Cell*, **163**, 1079–1094.

2. Barrington, W.T., Wulfridge, P., Wells, A.E., Rojas, C.M., Howe, S.Y.F., Perry, A., Hua, K., Pellizzon, M.A., Hansen, K.D., Voy, B.H. *et al.* (2018) Improving metabolic health through precision dietetics in mice. *Genetics*, **208**, 399–417.

3. Zheng, J.-S., Arnett, D.K., Lee, Y.-C., Shen, J., Parnell, L.D., Smith, C.E., Richardson, K., Li, D., Borecki, I.B., Ordovás, J.M. *et al.* (2013) Genome-wide contribution of genotype by environment interaction to variation of diabetes-related traits. *PLoS One*, **8**, e77442.

4. Franks, P.W. and Merino, J. (2018) Gene-lifestyle interplay in type 2 diabetes. *Curr. Opin. Genet. Dev.*, **50**, 35–40.

5. Thomas, D. (2010) Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Annu. Rev. Public Health*, **31**, 21–36.

6. Gauderman, W.J., Mukherjee, B., Aschard, H., Hsu, L., Lewinger, J.P., Patel, C.J., Witte, J.S., Amos, C., Tai, C.G., Conti, D. *et al.* (2017) Update on the state of the science for analytical methods for gene-environment interactions. *Am. J. Epidemiol.*, **186**, 762–770.

7. Westerman, K.E., Pham, D.T., Hong, L., Chen, Y., Sevilla-González, M., Sung, Y.J., Sun, Y.V., Morisson, A.C., Chen, H. and Manning, A.K. (2020) Scalable and flexible gene-environment interaction analysis in millions of samples. *Hum. Mol. Genet.* (In press).

8. Bi, W., Zhao, Z., Dey, R., Fritsche, L.G., Mukherjee, B. and Lee, S. (2019) A fast and accurate method for genome-wide scale phenome-wide G × E analysis and its application to UK biobank. *Am. J. Hum. Genet.*, **105**, 1182–1192.

9. Tyrrell, J., Wood, A.R., Ames, R.M., Yaghootkar, H., Beaumont, R.N., Jones, S.E., Tuke, M.A., Ruth, K.S., Freathy, R.M., Davey Smith, G. *et al.* (2017) Gene–obesogenic environment interactions in the UK biobank study. *Int. J. Epidemiol.*, **46**, 559–575.

10. Moore, R., Casale, F.P., Jan Bonder, M., Horta, D., Franke, L., Barroso, I. and Stegle, O. (2019) A linear mixed-model approach to study multivariate gene–environment interactions. *Nat. Genet.*, **51**, 180–186.

11. Tucker, K.L., Smith, C.E., Lai, C.-Q. and Ordovas, J.M. (2013) Quantifying diet for nutrigenomic studies. *Annu. Rev. Nutr.*, **33**, 349–371.

12. Cole, J.B., Florez, J.C. and Hirschhorn, J.N. (2020) Comprehensive genomic analysis of dietary habits in UK biobank identifies hundreds of genetic associations. *Nat. Commun.*, **11**, 1467.

13. MacDonald, P.E. (2011) TRP-ing down the path to insulin secretion. *Diabetes*, **60**, 28–29.

14. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S. *et al.* (2018) Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv* October 19, 2018. doi: 10.1101/447367 preprint: not peer reviewed.

15. Ketema, E.B. and Kibret, K.T. (2015) Correlation of fasting and postprandial plasma glucose with HbA1c in assessing glycemic control; systematic review and meta-analysis. *Arch. Public Heal.*, **73**, 43.

16. Vriens, J., Nilius, B. and Vennekens, R. (2008) Herbal compounds and toxins modulating TRP channels. *Curr. Neuropharmacol.*, **6**, 79–96

17. Uchida, K. and Tominaga, M. (2011) TRPM2 modulates insulin secretion in pancreatic β-cells. *Islets*, **3**, 209–211.

18. Wagner, T.F.J., Loch, S., Lambert, S., Straub, I., Mannebach, S., Mathar, I., Düfer, M., Lis, A., Flockerzi, V., Philipp, S.E. *et al.* (2008) Transient receptor potential M3 channels are ionotropic steroid receptors in pancreatic β cells. *Nat. Cell Biol.*, **10**, 1421–1430.

19. Straub, I., Mohr, F., Stab, J., Konrad, M., Philipp, S., Oberwinkler, J. and Schaefer, M. (2013) Citrus fruit and fabacea secondary metabolites potently and selectively block TRPM3. *Br. J. Pharmacol.*, **168**, 1835–1850.

20. Palmer, C. and Pe'er, I. (2017) Statistical correction of the Winner's curse explains replication variability in quantitative trait genome-wide association studies. *PLoS Genet.*, **13**, e1006916.

21. Li, S.X., Imamura, F., Ye, Z., Schulze, M.B., Zheng, J., Ardanaz, E., Arriola, L., Boeing, H., Dow, C., Fagherazzi, G. *et al.* (2017) Interaction between genes and macronutrient intake on the risk of developing type 2 diabetes: systematic review and findings from European prospective investigation into cancer (EPIC)-InterAct. *Am. J. Clin. Nutr.*, **106**, 263–275.

22. Kraft, P. and Aschard, H. (2015) Finding the missing gene–environment interactions. *Eur. J. Epidemiol.*, **30**, 353–355.

23. Willett, W.C., Howe, G.R. and Kushi, L.H. (1997) Adjustment for total energy intake in epidemiologic studies. *Am. J. Clin. Nutr.*, **65**, 1220S–1228S.

24. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M. *et al.* (2015) 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.

25. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J. *et al.* (2018) The UK biobank resource with deep phenotyping and genomic data. *Nature*, **562**, 203–209.

26. R Core Team (2019) *R: A Language and Environment for Statistical Computing*. R: A Language and Environment for Statistical Computing, Vienna, Austria.

27. Eastwood, S. V., Mathur, R., Atkinson, M., Brophy, S., Sudlow, C., Flaig, R., de Lusignan, S., Allen, N. and Chaturvedi, N. (2016) Algorithms for the capture and adjudication of prevalent and incident diabetes in UK biobank. *PLoS One*, **11**, e0162388.

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

29. Guo, W., Bradbury, K.E., Reeves, G.K. and Key, T.J. (2015) Physical activity in relation to body size and composition in women in UK biobank. *Ann. Epidemiol.*, **25**, 406–413.e6.

30. Ge, T., Chen, C.-Y., Doyle, A.E., Vettermann, R., Tuominen, L.J., Holt, D.J., Sabuncu, M.R. and Smoller, J.W. (2019) The shared genetic basis of educational attainment and cerebral cortical morphology. *Cereb. Cortex*, **29**, 3471–3481.

31. Moore, C., Jacobson, S. and Fingerlin, T. (2018) Power and sample size calculations for genetic association studies in the presence of genetic model misspecification. *American Society for Human Genetics*, **84**, 256–271

32. Abraham, G., Qiu, Y. and Inouye, M. (2017) FlashPCA2: principal component analysis of biobank-scale genotype datasets. *Bioinformatics*, **33**, 2776–2778.

33. Keller, M.C. (2014) Gene × environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution. *Biol. Psychiatry*, **75**, 18–24.

34. Kamat, M.A., Blackshaw, J.A., Young, R., Surendran, P., Burgess, S., Danesh, J., Butterworth, A.S. and Staley, J.R. (2019) PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations. *Bioinformatics*, **35**, 4851–4853.

35. Staley, J.R., Blackshaw, J., Kamat, M.A., Ellis, S., Surendran, P., Sun, B.B., Paul, D.S., Freitag, D., Burgess, S., Danesh, J. *et al.* (2016) PhenoScanner: a database of human genotype–phenotype associations. *Bioinformatics*, **32**, 3207–3209.

36. Bonder, M.J., Luijk, R., Zhernakova, D.V., Moed, M., Deelen, P., Vermaat, M., Van Iterson, M., Van Dijk, F., Van Galen, M., Bot, J. *et al.* (2017) Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.*, **49**, 131–138.

37. Ridker, P.M., Chasman, D.I., Zee, R.Y.L., Parker, A., Rose, L., Cook, N.R. and Buring, J.E. (2008) Rationale, design, and methodology of the Women's genome health study: a genome-wide association study of more than 25 000 initially healthy American women. *Clin. Chem.*, **54**, 249–255.

38. Liu, S., Manson, J.E., Lee, I.-M., Cole, S.R., Hennekens, C.H., Willett, W.C. and Buring, J.E. (2000) Fruit and vegetable intake and risk of cardiovascular disease: the Women's health study. *Am. J. Clin. Nutr.*, **72**, 922–928.