bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genomic analysis of diet composition finds novel loci and associations with health and lifestyle

All authors and their affiliations are listed at the end of the manuscript.

ABSTRACT

We conducted genome-wide association study (GWAS) meta-analyses of relative caloric intake from fat, protein, carbohydrates and sugar in over 235,000 individuals. We identified 21 approximately independent lead SNPs. Relative protein intake exhibits the strongest relationships with poor health, including positive genetic associations with obesity, type 2 diabetes, and heart disease ($r_g \approx 0.15 - 0.5$). Relative carbohydrate and sugar intake have negative genetic correlations with waist circumference, waist-hip ratio, and neighborhood poverty ($|r_g| \approx 0.1 - 0.3$). Overall, our results show that the relative intake of each macronutrient has a distinct genetic architecture and pattern of genetic correlations suggestive of health implications beyond caloric content.

MAIN TEXT

Understanding the effects of nutrition on health is a priority given the ongoing worldwide obesity epidemic[1–5]. The health impacts of many aspects of dietary intake have been studied, but the effects of macronutrient composition (i.e., relative intake from fat, protein, and carbohydrate) have been especially controversial. There is still no consensus on whether macronutrients exert specific health effects beyond their caloric value[6–8]. Despite a lack of robust empirical evidence from randomized trials on the long-term effects of macronutrient restriction on body weight and health[2,6,9,10], dietary recommendations have

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

shifted from low-fat to low-sugar and, more recently, lower animal-protein diets[11–17]. Observational studies have found inconsistent phenotypic correlations between macronutrient proportions, body mass index (BMI) and related health outcomes (e.g., [18–20]), and the mechanisms underlying these relationships are not well understood.

Insights from genetics may help to elucidate the connections between nutrition and health outcomes. Twin studies suggest that diet composition is moderately heritable, with h^2 estimates ranging from 27% to 70% for the different macronutrients' contributions to total energy intake[21–23]. Previous GWAS on relative caloric intake from protein, fat, and carbohydrates (up to N = 91,114) have identified three genome-wide significant SNPs in or near *RARB*, *FTO* and *FGF21*, each of which captures only a miniscule part of trait heritability (R^2 < 0.06%)[24–26]. These results suggest that diet composition is a genetically complex phenotype and that most associated genetic variants have not yet been identified. Furthermore, no large-scale genome-wide association study (GWAS) has investigated relative sugar intake.

Here we report GWAS results for diet composition, and we use the results to conduct bioinformatics analyses and to calculate genetic correlations with a range of other phenotypes. For the GWAS, we expand the samples used in earlier work from N = 91,114[24–26] to 268,922 for relative intake of PROTEIN, CARBOHYDRATE, and FAT. Furthermore, we report GWAS results for SUGAR (N = 235,391), which is a subcomponent of CARBOHYDRATE and captures relative intake of both naturally-occurring and added sugars.

RESULTS

Phenotype definition

All cohorts used self-report questionnaires containing ≥70 food items, with average estimated intakes showing strong similarity across cohorts (**Supplementary Table 1.2**). Using these self-reports, we calculated the relative contributions of FAT, PROTEIN, CARBOHYDRATE and SUGAR to total caloric intake

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(we do not study total caloric intake because it is mainly determined by body size and physical activity[27] and because systematic underreporting of total food intake is correlated with BMI[28]). Since macronutrient intake may not scale linearly with total caloric intake, we developed and applied a method that adjusts for the observed non-linear relationships (**Supplementary Information 2.6, Extended Data Figure 1**). Consistent with the satiating properties of protein[29], we find that at higher levels of total caloric intake, relative protein intake declines, while relative fat intake increases, and relative sugar and carbohydrate intake remain roughly constant (**Supplementary Table 2.2**).

Main results

We began by assessing the SNP-based heritability of our phenotypes. We calculated GREML[30] estimates using a random N = 30,000 subsample of conventionally unrelated UK Biobank (UKB) individuals. The estimates range from 2.1% for PROTEIN to 7.9% for CARBOHYDRATE (**Extended Data Figure 2** and **Supplementary Information 7**).

GWAS were performed in individuals of European ancestry. When possible, we excluded individuals on calorie- or macronutrient-restricted diets (**Supplementary Table 1.3**). Our discovery sample was the subset of the UKB with survey data on dietary intake (N = 175,253). The replication phase consisted of a meta-analysis of GWAS summary statistics from 14 additional cohorts that followed our analysis plan (N = 60,138) and summary statistics from DietGen[25] (for FAT, PROTEIN and CARBOHYDRATE, N = 33,531). DietGen[25] assumed a linear scaling of macronutrients with total energy intake. Since the genetic correlations between DietGen and our replication cohorts is not significantly different from 1 (**Supplementary Table 6.1**), we added DietGen to our meta-analysis.

Association statistics underwent rigorous quality control (**Supplementary Information 3.3**). The discovery stage identified 21 approximately independent genome-wide-significant lead SNPs (see **Supplementary Information 3.3.5** for a description of the clumping algorithm): 4 for FAT, 5 for PROTEIN, 5 for SUGAR, and 7 for CARBOHYDRATE (**Supplementary Table 4.1**). These lead SNPs partially overlap across phenotypes and reside in 14 unique loci. In the replication stage, all 21 lead SNPs had the

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

anticipated signs and comparable effect sizes (**Extended Data Figure 3**), and 15 reached statistical significance at $P < 0.05$ (**Supplementary Table 4.1**). This empirical replication record matches or exceeds theoretical predictions that take into account the statistical winner's curse, sampling variation, and statistical power[31] (**Supplementary Information 4.1**).

In order to maximize statistical power, all follow-up analyses that now follow are based on results from the combined analyses of discovery and replication samples (N = 235,391 to 268,922). The quantile-quantile plots of exhibit substantial inflation (λ_{GC} = 1.12 to 1.19, **Extended Data Figure 4**). The estimated intercepts from LD Score regressions[32] (LDSC) suggest that the vast majority of this inflation is due to polygenic signal, and only a small share is attributable to population stratification (max ~6% for FAT, *n.s.* different from 0%; **Supplementary Table 3.4**). The number of approximately independent lead SNPs is 36 (pairwise $r^2 < 0.01$), including 6 for FAT, 7 for PROTEIN, 10 for SUGAR, and 13 for CARBOHYDRATE (**Table 1**, **Figure 1**). These 36 lead SNPs reside in 21 unique loci (**Supplementary Table 5.2**). The SNP effect sizes range from 0.015 to 0.098 phenotypic standard deviations per allele. The phenotypic variance explained per SNP, expressed in terms of coefficient of determination (R^2), ranged from 0.011% to 0.054%, comparable to other genetically complex traits such as BMI and educational attainment (**Extended Data Figure 5**).

MAGMA[33] analyses of our GWAS summary statistics identified 81 unique genes (**Extended Data Figure 6** and **Supplementary Table 5.4**). While the majority of these genes were near our lead SNPs, MAGMA also identified 33 genomic regions harboring 44 unique genes that are physically distant (> 1 Mb) from our lead SNPs.

We constructed polygenic scores for the macronutrient intakes by applying LDpred[34] to our GWAS summary statistics. We assessed the scores' out-of-sample predictive accuracy in two holdout cohorts: the Health and Retirement Study (N = 2,344) and the Rotterdam Study (N = 3,585). The scores predicted the macronutrient intakes with R^2 ranging between 0.08% (P = 0.088) and 0.71% ($P = 9.11\times10^{-7}$; **Supplementary Table 8.1**).

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We estimated pairwise genetic correlations between the macronutrients with bivariate LDSC[35]. All are statistically distinguishable from zero at $P < 0.05$ (except FAT and PROTEIN), as well as from one and negative one (**Table 2**). These results indicate that intake of each macronutrient has a different genetic architecture, consistent with previous work from animal studies showing distinct biological mechanisms involved in macronutrient-specific appetites[36].

Discussion of lead SNPs from combined meta-analysis

Seven of the 21 lead SNPs have not been (directly or via LD partners, $r^2 \geq 0.6$ and distance < 250 kb) associated with any other traits in the NHGRI-EBI GWAS Catalog[37] (**Supplementary Table 5.5**). Each of these seven SNPs is located in or near genes that have not been studied in depth to date.

Five lead SNPs are located in (or near) genes that have well characterized biological functions in nutrient metabolism or homeostasis but have not previously been associated with food intake. First, a missense variant in *APOE* (rs429358) was associated with FAT, SUGAR, and CARBOHYDRATE, where the allele that decreases Alzheimer's risk is associated with greater FAT intake, and vice-versa for SUGAR and CARBOHYDRATE. *APOE* is not only strongly associated with Alzheimer's disease[38] but is also involved in fatty acid metabolism. We explored whether this association may be driven by sample selection. Specifically, older people with dementia may be systematically missing from the UKB, and unaffected elderly people may have different eating habits than younger people. We found that the association was greatly reduced in the subsample of UKB participants aged below 60, but the 95% confidence intervals of the effect sizes still overlapped with those of the older sample (**Supplementary Table 5.3**).

Second, a well-known missense variant (rs1229984 in *ADH1B*) that limits alcohol metabolism was positively associated with FAT intake. The association was weaker in a sample of UKB alcohol abstainers (N = 39,679; **Supplementary Table 5.3**), suggesting that it may be partially driven by substitution of fat for alcohol.

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Third, a PROTEIN lead SNP (rs13146907) was found in *KLB*, an essential cofactor to *FGF21*[39,40] which influences sweet and alcohol taste preference via the liver-brain-endocrine axis[41–43]. *KLB* was only associated with PROTEIN, while variants in (or near) *FGF21* were strongly associated with all four macronutrients. With MAGMA, we also identified *MLXIPL* (only for FAT), a gene that acts as a transcription factor to *FGF21*[44]. This might imply that different genes involved in the same pathway are important for directing intake of different macronutrients.

Fourth, an intergenic variant (rs2472297) linked to higher caffeine consumption[45,46] was associated with lower CARBOHYDRATE intake. There are various possible explanations, such as interrelated lifestyle choices pertaining to food and caffeinated drinks.

Fifth, an intronic variant in *GCKR* (rs780094), a carbohydrate-metabolism gene, is associated with PROTEIN. The lead SNP is in almost perfect LD with a missense variant that has been associated with lipid levels[47].

Bioinformatic analyses

Animal studies indicate that the brain and peripheral organs interact in directing macronutrient intake[36,48]. A question that arises is whether the "periphery", which digests and metabolizes macronutrients, plays a larger role than the brain, for instance by determining how the brain assigns reward values to macronutrients. (For example, this is partially the case with alcohol, where mutations that limit metabolic capacity render alcohol consumption unpleasant[49,50].) To examine to what extent genetic variation in the brain and the periphery contributes to macronutrient intake in humans, we used stratified LDSC[51,52] to identify in which tissues diet-composition-associated SNPs are likely to be expressed (**Supplementary Information 9.1**). We performed two stratified LDSC analyses, which partitioned SNP heritability according to (i) 10 broadly-defined tissues, which were ascertained with LDSC reference data from chromatin data[53] and (ii) 53 tissues (including 14 brain regions), as ascertained with LDSC reference data from sets of Specifically Expressed Genes in GTEx (known as LDSC-SEG)[52]. To correct for multiple

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

testing across tissues, we applied Bonferroni adjustments for the number of tested tissues ($P_{\mathrm{Bonf}} = 10 \cdot P$ and $P_{\mathrm{Bonf}} = 53 \cdot P$, respectively).

We found that genetic variation related to the central nervous system plays a major role for intake of all macronutrients ($P_{\mathrm{Bonf}} < 0.015$ for the regression coefficients; **Figure 4**), with the proportions of explained heritability ranging from 44% (FAT and SUGAR) to 55% (PROTEIN). Within the central nervous system, we found broad involvement of the brain, including (frontal) cortex (FAT and SUGAR), the basal ganglia (FAT), limbic system (FAT and SUGAR), cerebellum (PROTEIN), and hypothalamus and substantia nigra for FAT and PROTEIN (and SUGAR suggestively: $P_{\mathrm{Bonf}} = 0.06$). The confidence intervals for the coefficients overlap across brain regions so we cannot draw conclusions about the specificity of brain regions for intake of particular macronutrients.

For FAT, genetic variation related to adrenals and/or pancreas tissue is estimated to explain 37% of the heritability. Because the adrenals play a role in lipid metabolism, and the pancreas is crucial for digestion, either tissue may plausibly affect fat intake. We caution, however, that in the LDSC-SEG analyses of 53 tissues, all non-brain regions had *P* values above 0.05 even before Bonferroni adjustment (**Figure 5**).

To gain insight into the putative functions of the top associated loci, we queried the 81 genes identified by the MAGMA analyses in Gene Network[54], which predicts Reactome[55] functions for genes (**Supplementary Information 9.3**). In addition to neural functioning (e.g., axon guidance), we found that the MAGMA genes were predicted to be involved in growth factor signaling and the immune system (**Supplementary Table 9.6**). These results may imply a more pronounced role for peripheral gene functions than our stratified LDSC results, which mainly implicated the brain.

Relationships with health, lifestyle and socioeconomic status

Using bivariate LDSC[35,56], we estimated genetic correlations between our diet-composition phenotypes and 19 preselected relevant medical and lifestyle phenotypes for which well-powered GWAS results were available. We also included four additional phenotypes for which GWAS results became available after

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

our study was underway, as well as Alzheimer's disease, motivated by the association we found between *APOE* with macronutrient intake. To control for multiple testing, we again calculated Bonferroni-adjusted *P* values ($P_{\mathrm{Bonf}} = 24 \cdot P$).

PROTEIN showed the strongest genetic correlations with poor health outcomes, including obesity ($r_g = 0.35$), type 2 diabetes ($r_g = 0.45$), fasting insulin ($r_g = 0.41$), and coronary artery disease ($r_g = 0.16$), as well as BMI ($r_g = 0.40$) (**Figure 2**, **Supplementary Table 10.1**). FAT, SUGAR, and CARBOHYDRATE had negative, non-significant genetic correlations with BMI (r_g between -0.06 and -0.02). For comparison, we estimated phenotypic associations between diet composition and BMI in four independent cohorts (combined *N* = 173,353) and meta-analyzed the results (**Figure 3**). PROTEIN (standardized $\hat{\beta}$ = 0.09) and FAT (standardized $\hat{\beta}$ = 0.06) are positively associated with BMI, while SUGAR and CARBOHYDRATE are negatively associated with BMI (standardized $\hat{\beta}$ = -0.09 and -0.09, respectively, **Supplementary Table 10.2**). Thus, the genetic correlation between PROTEIN and BMI stands out as large relative to the phenotypic correlation.

Despite their relatively weak genetic correlations with BMI, SUGAR and CARBOHYDRATE have significant negative genetic correlations with waist circumference ($r_g = -0.13$ and -0.14) and waist-hip ratio ($r_g = -0.15$ and -0.18). All the macronutrients have negative genetic correlations with alcohol consumption (r_g between -0.61 and -0.11), as expected since alcohol is included in energy intake and our phenotype measures are shares of energy intake.

Next, we computed genetic correlations with indicators of socioeconomic status[31,57,58], which are known to be phenotypically associated with food access, dietary choices, and health[59–63]. We found that FAT is negatively genetically correlated with educational attainment ($r_g = -0.13$). SUGAR and CARBOHYDRATE are negatively genetically correlated with the Townsend deprivation index ($r_g = -0.23$ and -0.30), which is constructed from the rates of unemployment, non-ownership of cars and houses, and overcrowding of the neighborhood in which individuals live[64,58], with higher scores indicating more

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

severe socioeconomic deprivation. These genetic correlations are suggestive of environmental channels that affect macronutrient intake.

Finally, we estimate the genetic correlations between diet composition and physical activity. Because physical activity is known to have health benefits[65], its genetic correlations with diet composition may provide clues about mechanisms underlying relationships between diet composition and health. In these genetic correlation analyses, we used unpublished physical activity GWAS summary statistics from a sample of research participants from 23andMe (N = 123,983). The physical activity phenotype is a composite measure based on self-reported activities from leisure, occupation, and commuting. We found a negative genetic correlation of physical activity with FAT ($r_g = -0.20$) and a positive genetic correlation with SUGAR ($r_g = 0.22$). The genetic correlations with PROTEIN and CARBOHYDRATE are positive but not statistically distinguishable from zero (0.11 and 0.06, respectively).

Discussion

A possible role for PROTEIN in the etiology of metabolic dysfunction is implicated by the genetic correlation between PROTEIN and obesity, waist-hip ratio, fasting insulin, type 2 diabetes, HDL cholesterol, and heart disease, as well as by the BMI-increasing *FTO* allele associating with increased protein intake. This conclusion coincides with a growing (but often overlooked[66]) body of evidence that links protein intake to obesity and insulin resistance[67–75]. The positive genetic link between PROTEIN and BMI could reflect a causal effect of relative protein intake. There is some evidence from randomized trials with infants, which found a causal relationship between high-protein baby formula and infant body fat[76]. While the underlying biological mechanisms are unclear, high consumption of protein or certain types of amino acids (i.e., building blocks of protein) can induce insulin resistance[77–79], rapamycin signaling[72], and growth factor signaling[80], thereby increasing metabolic dysfunction and early mortality risk.

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We caution, however, that the strong and consistent links between PROTEIN and poor health outcomes might also be consistent with alternative explanations. Causation could run in the reverse direction: overweight individuals may have higher protein needs, or use high-protein diets as a weight-loss strategy. The associations might also be caused by other, unmeasured variables such as unhealthy lifestyle factors or co-consumed ingredients. However, we find that the phenotypic association between PROTEIN and BMI is robust to controls for educational attainment and household income. Furthermore, the genetic correlation between PROTEIN and physical activity is statistically indistinguishable from zero but positive. These findings weigh against socioeconomic status or physical activity being confounders of the positive genetic correlation between PROTEIN and BMI.

For SUGAR, the phenotypic and genetic correlations we found with BMI and other health outcomes are consistent with observations from systematic reviews and meta-analyses of phenotypic relationships. Together, this body of evidence suggests that dietary sugar, beyond its caloric value, does not have negative health effects[81–85], contrary to some popular beliefs (e.g., [17]). Another possibility is that exercise offsets negative metabolic effects of high sugar intake[86,87]. Those with a higher predisposition to be physically active may tend to consume more sugar, as sugar is a metabolically convenient source of energy during exercise[88] and may enhance endurance[89]. If so, the positive genetic correlation between SUGAR and physical activity might partially explain the lack of genetic correlations between SUGAR and poor health.

For FAT and CARBOHYDRATE, we also found no consistent pattern of genetic and phenotypic associations with poor metabolic health. Taken together, our results complement the findings of phenotypic analyses from a large, multinational study by the EPIC-PANACEA consortium (N = 373,803), which found that only calories from protein are associated with prospective weight gain[18] – a finding that was consistent across 10 countries.

While the phenotypic associations between dietary intake and health and lifestyle factors have been extensively explored in prior work, the large-scale genetic study of dietary intake is new. Overall, our

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

results show that the relative intake of each macronutrient has a distinct genetic architecture, and the pattern of genetic correlations is suggestive of health implications beyond caloric content. Moreover, our genetic correlation and bioinformatics analyses suggest a number of novel hypotheses regarding the causes and consequences of dietary intake that can be explored in future work.

Online Methods

Materials and methods are described in detail in the online Supplementary Information. Upon publication, GWAS summary statistics for the four macronutrients can be downloaded from the SSGAC website (https://thessgac.org/data).

References

1. Katz, D. L. Diets, diatribes, and a dearth of data. *Circ. Cardiovasc. Qual. Outcomes* **7,** 809–811 (2014).
2. Katz, D. L. & Meller, S. Can we say what diet is best for health? *Annu. Rev. Public Health* **35,** 83–103 (2014).
3. Piper, M. D. W., Partridge, L., Raubenheimer, D. & Simpson, S. J. Dietary restriction and ageing: a unifying perspective. *Cell Metab.* **14,** (2014).
4. Flatt, J.-P. Macronutrient Composition and Food Selection. *Obes. Res.* **9,** 256S–262S (2001).
5. Bellissimo, N. & Akhavan, T. Effect of macronutrient composition on short-term food intake and weight loss. *Am. Soc. Nutr. Adv. Nutr.* **6,** 3025–3085 (2015).
6. Hall, K. D. & Guo, J. Obesity energetics: body weight regulation and the effects of diet composition. *Gastroenterology* **152,** 1718–1727 (2017).
7. Buchholz, A. C. & Schoeller, D. A. Is a calorie a calorie? *Am. J. Clin. Nutr.* **79,** 899S–906S (2004).
8. Feinman, R. D. & Fine, E. J. 'A calorie is a calorie' violates the second law of thermodynamics. *Nutr. J.* **3,** (2004).
9. Atallah, R. *et al.* Long-term effects of 4 popular diets on weight loss and cardiovascular risk factors: A systematic review of randomized controlled trials. *Circ. Cardiovasc. Qual. Outcomes* **7,** 815–827 (2014).
10. Howard, B. V. *et al.* Low-fat dietary pattern and weight change over 7 years: The Women's Health Initiative Dietary Modification Trial. *J. Am. Med. Assoc.* **295,** 39–49 (2006).
11. La Berge, A. F. How the ideology of low fat conquered America. *J. Hist. Med. Allied Sci.* **63,** 139–177 (2008).

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

12. WHO. Information note about intake of sugars recommended in the WHO guideline for adults and children. (2015). at <http://www.who.int/nutrition/publications/guidelines/sugar_intake_information_note_en.pdf>

13. U.S. Department of Health and Human Services and U.S. Department of Agriculture. *2015-2020 Dietary Guidelines for Americans*. (2015). at <http://health.gov/dietaryguidelines/2015/guidelines/>

14. Mozaffarian, D. & Ludwig, D. S. The 2015 US Dietary Guidelines. *JAMA* **313,** 2421 (2015).

15. Johns, D. M. & Oppenheimer, G. M. Was there ever really a "sugar conspiracy"? *Science (80-.).* **359,** 747–750 (2018).

16. Koletzko, B., Demmelmair, H., Grote, V., Prell, C. & Weber, M. High protein intake in young children and increased weight gain and obesity risk. *Am. J. Clin. Nutr.* **103,** 303–304 (2016).

17. Lustig, R. H., Schmidt, L. A. & Brindis, C. D. The toxic truth about sugar. *Nature* **482,** 27–29 (2012).

18. Vergnaud, A.-C. *et al.* Macronutrient composition of the diet and prospective weight change in participants of the EPIC-PANACEA study. *PLoS One* **8,** e57300 (2013).

19. Anderson, J. J. *et al.* Adiposity among 132 479 UK Biobank participants; contribution of sugar intake vs other macronutrients. *Int. J. Epidemiol.* dyw173 (2016). doi:10.1093/ije/dyw173

20. Austin, G. L., Ogden, L. G. & Hill, J. O. Trends in carbohydrate, fat, and protein intakes and association with energy intake in normal-weight, overweight, and obese individuals. *Am J Clin Nutr* **93,** 836–43 (2011).

21. Wade, J., Milner, J. & Krondl, M. Evidence for a physiological regulation of food selection and nutrient intake in twins. *Am. J. Clin. Nutr.* **34,** 143–7 (1981).

22. De Castro, J. M. Heritability of diurnal changes in food intake in free-living humans. *Nutrition* **17,** 713–720 (2001).

23. Hasselbalch, A. L., Heitmann, B. L., Kyvik, K. O. & Sørensen, T. I. A. Studies of twins indicate that genetics influence dietary intake. *J. Nutr.* **138,** 2406–12 (2008).

24. 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).

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

26. 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* 1 (2018). doi:10.1038/s41380-018-0079-4

27. Mifflin, M. D. *et al.* A new predictive equation for resting energy expenditure in healthy individuals. *Am. J. Clin. Nutr.* **51,** 241–247 (1990).

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

28. Poslusna, K., Ruprich, J., De Vries, J. H. M., Jakubikova, M. & Van 't Veer, P. Misreporting of energy and micronutrient intake estimated by food records and 24 hour recalls, control and adjustment methods in practice. *Br. J. Nutr.* **101,** S73–S85 (2009).

29. Halton, T. L. & Hu, F. B. The effects of high protein diets on thermogenesis, satiety and weight loss: A critical review. *J. Am. Coll. Nutr.* **23,** 373–385 (2004).

30. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88,** 76–82 (2011).

31. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533,** 539–542 (2016).

32. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47,** 291–295 (2015).

33. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11,** e1004219 (2015).

34. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97,** 576–592 (2015).

35. Bulik-Sullivan, B. K. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47,** 1236–1241 (2015).

36. Berthoud, H. R., Münzberg, H., Richards, B. K. & Morrison, C. D. Neural and metabolic regulation of macronutrient intake and selection. *Proc. Nutr. Soc.* **71,** 390–400 (2012).

37. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001-6 (2014).

38. Liu, C.-C., Kanekiyo, T., Xu, H. & Bu, G. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nat. Rev. Neurol.* **9,** 106–118 (2013).

39. Ogawa, Y. *et al.* betaKlotho is required for metabolic activity of fibroblast growth factor 21. *Proc. Natl. Acad. Sci.* **104,** 7432–7437 (2007).

40. Kurosu, H. *et al.* Tissue-specific expression of βklotho and Fibroblast Growth Factor (FGF) receptor isoforms determines metabolic activity of FGF19 and FGF21. *J. Biol. Chem.* **282,** 26687–26695 (2007).

41. Schumann, G. *et al.* KLB is associated with alcohol drinking, and its gene product β-Klotho is necessary for FGF21 regulation of alcohol preference. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 14372–14377 (2016).

42. Von Holstein-Rathlou, S. *et al.* FGF21 mediates endocrine control of simple sugar intake and sweet taste preference by the liver. *Cell Metab.* **23,** 335–343 (2016).

43. Talukdar, S. *et al.* FGF21 regulates sweet and alcohol preference. *Cell Metab.* **23,** 344–349 (2016).

44. Adams, A. C. & Gimeno, R. E. The sweetest thing: Regulation of macronutrient preference by FGF21. *Cell Metab.* **23,** 227–228 (2016).

45. Cornelis, M. C. *et al.* Genome-wide meta-analysis identifies regions on 7p21 (AHR) and 15q24 (CYP1A2) as determinants of habitual caffeine consumption. *PLoS Genet.* **7,**

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

e1002033 (2011).

46. Coffee and Caffeine Genetics Consortium, C. and C. G. *et al.* Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption. *Mol. Psychiatry* **20,** 647–656 (2015).

47. Chasman, D. I. *et al.* Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. *PLoS Genet.* **5,** e1000730 (2009).

48. Efeyan, A., Comb, W. C. & Sabatini, D. M. Nutrient-sensing mechanisms and pathways. *Nature* **517,** 302–310 (2015).

49. Whitfield, J. B. & Martin, N. G. Aversive reactions and alcohol use in europeans. *Alcohol. Clin. Exp. Res.* **17,** 131–134 (1993).

50. Harada, S., Agarwal, D. P., Goedde, H. W., Tagaki, S. & Ishikawa, B. Possible protective role against alcoholism for aldehyde dehydrogenase isozyme deficiency in Japan. *Lancet (London, England)* **2,** 827 (1982).

51. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47,** 1228–1235 (2015).

52. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50,** 621–629 (2018).

53. Finucane, H. K. *et al.* Partitioning heritability by functional category using GWAS summary statistics. *Nat. Genet.* **47,** 1228–1235 (2015).

54. Fehrmann, R. S. N. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **47,** 115–25 (2015).

55. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42,** D472-7 (2014).

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

57. Rietveld, C. A. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340,** 1467–71 (2013).

58. Hill, W. D. *et al.* Molecular genetic contributions to social deprivation and household income in UK Biobank. *Curr. Biol.* **26,** 3083–3089 (2016).

59. Beaulac, J., Kristjansson, E. & Cummins, S. A systematic review of food deserts, 1966-2007. *Prev. Chronic Dis.* **6,** A105 (2009).

60. Handbury Ilya Rahkovsky Molly Schnell, J. *et al.* Is the focus on food deserts fruitless? Retail access and food purchases across the socioeconomic spectrum. *NBER Work. Pap.* (2015). at <http://www.nber.org/papers/w21126>

61. Adler, N. E. *et al.* Socioeconomic status and health. The challenge of the gradient. *Am. Psychol.* **49,** 15–24 (1994).

62. Marmot, M. G. & Wilkinson, R. G. *Social determinants of health: The solid facts*. (World

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Health Organization, 2003).

63. Stringhini, S. *et al.* Socioeconomic status and the 25 × 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1·7 million men and women. *Lancet* **389,** 1229–1237 (2017).

64. Townsend, P. Deprivation. *J. Soc. Policy* **16,** 125 (1987).

65. Reiner, M., Niermann, C., Jekauc, D. & Woll, A. Long-term health benefits of physical activity – a systematic review of longitudinal studies. *BMC Public Health* **13,** 813 (2013).

66. Campbell, T. C. A plant-based diet and animal protein: questioning dietary fat and considering animal protein as the main cause of heart disease. *J. Geriatr. Cardiol.* **14,** 331–337 (2017).

67. Pimpin, L., Jebb, S., Johnson, L., Wardle, J. & Ambrosini, G. L. Dietary protein intake is associated with body mass index and weight up to 5 y of age in a prospective cohort of twins. *Am. J. Clin. Nutr.* **103,** 389–397 (2016).

68. Günther, A. L., Remer, T., Kroke, A. & Buyken, A. E. Early protein intake and later obesity risk: which protein sources at which time points throughout infancy and childhood are important for body mass index and body fat percentage at 7 y of age? *Am. J. Clin. Nutr.* **86,** 1765–1772 (2007).

69. Voortman, T. *et al.* Protein intake in early childhood and body composition at the age of 6 years: The Generation R Study. *Int. J. Obes.* **40,** 1018–1025 (2016).

70. Trichopoulou, A. *et al.* Lipid, protein and carbohydrate intake in relation to body mass index. *Eur. J. Clin. Nutr.* **56,** 37–43 (2002).

71. Koletzko, B. *et al.* Lower protein in infant formula is associated with lower weight up to age 2 y: A randomized clinical trial. *Am. J. Clin. Nutr.* **89,** 1836–1845 (2009).

72. Solon-Biet, S. M. *et al.* The ratio of macronutrients, not caloric intake, dictates cardiometabolic health, aging, and longevity in ad libitum-fed mice. *Cell Metab.* **19,** 418–430 (2014).

73. Hörnell, A., Lagström, H., Lande, B. & Thorsdottir, I. Protein intake from 0 to 18 years of age and its relation to health: a systematic literature review for the 5th Nordic Nutrition Recommendations. *Food Nutr. Res.* **57,** 21083 (2013).

74. Van Nielen, M. *et al.* Dietary protein intake and incidence of type 2 diabetes in Europe: The EPIC-InterAct case-cohort study. *Diabetes Care* **37,** 1854–1862 (2014).

75. Weber, M. *et al.* Lower protein content in infant formula reduces BMI and obesity risk at school age: follow-up of a randomized trial. *Am J Clin Nutr* **99,** 1041–51 (2014).

76. Patro-Gołąb, B. *et al.* Nutritional interventions or exposures in infants and children aged up to 3 years and their effects on subsequent risk of overweight, obesity and body fat: a systematic review of systematic reviews. *Obes. Rev.* **17,** 1245–1257 (2016).

77. Newgard, C. B. *et al.* A branched-chain amino acid-related metabolic signature that differentiates obese and lean humans and contributes to insulin resistance. *Cell Metab.* **9,** 311–326 (2009).

78. Lynch, C. J. & Adams, S. H. Branched-chain amino acids in metabolic signalling and

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

insulin resistance. *Nat. Rev. Endocrinol.* **10,** 723–736 (2014).

79. Fontana, L. *et al.* Decreased consumption of branched-chain amino acids improves metabolic health. *Cell Rep.* **16,** 520–530 (2016).

80. Levine, M. E. *et al.* Low protein intake is associated with a major reduction in IGF-1, cancer, and overall mortality in the 65 and younger but not older population. *Cell Metab.* **19,** 407–417 (2014).

81. Sadler, M. J., McNulty, H. & Gibson, S. Sugar-sat seesaw: A systematic review of the evidence. *Crit. Rev. Food Sci. Nutr.* **55,** 338–356 (2015).

82. Reid, M. & Hammersley, R. Sugars and obesity: Meta-analysis establishes the strength of the correlation, not the cause. *Nutr. Bull.* **39,** 153–156 (2014).

83. Te Morenga, L., Mallard, S. & Mann, J. Dietary sugars and body weight: systematic review and meta-analyses of randomised controlled trials and cohort studies. *BMJ* **346,** e7492 (2012).

84. Khan, T. A. & Sievenpiper, J. L. Controversies about sugars: results from systematic reviews and meta-analyses on obesity, cardiometabolic disease and diabetes. *Eur. J. Nutr.* **55,** 25–43 (2016).

85. Tappy, L. & Mittendorfer, B. Fructose toxicity: Is the science ready for public health actions? *Curr. Opin. Clin. Nutr. Metab. Care* **15,** 357–361 (2012).

86. Egli, L. *et al.* Exercise performed immediately after fructose ingestion enhances fructose oxidation and suppresses fructose storage. *Am. J. Clin. Nutr.* **103,** 348–355 (2016).

87. Bidwell, A. J. *et al.* Physical activity offsets the negative effects of a high-fructose diet. *Med. Sci. Sports Exerc.* **46,** 2091–2098 (2014).

88. Tappy, L. & Rosset, R. Fructose metabolism from a functional perspective: Implications for athletes. *Sport. Med.* **47,** 23–32 (2017).

89. Rowlands, D. S. *et al.* Fructose–glucose composite carbohydrates and endurance performance: Critical review and future perspectives. *Sport. Med.* **45,** 1561–1576 (2015).

Acknowledgements

This research was carried out under the auspices of the Social Science Genetic Association Consortium (SSGAC, https://www.thessgac.org/). The research has also been conducted using the UK Biobank Resource under Application Number 11425. The study was supported by funding from the Ragnar Söderberg Foundation (E9/11 and E42/15), the Swedish Research Council (421-2013-1061), The Jan Wallander and Tom Hedelius Foundation, an ERC Consolidator Grant to Philipp Koellinger (647648 EdGe), the Pershing Square Fund of the Foundations of Human Behavior, The Open Philanthropy Project

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(2016-152872), and the NIA/NIH through grants P01-AG005842, P01-AG005842-20S2, P30-AG012810, and T32-AG000186-23 to NBER, and R01-AG042568-02 to the University of Southern California. CCC was supported by the Intramural Research Program of the NIH/NIDDK. We thank the DietGen and CHARGE consortia for sharing diet composition GWAS summary statistics, and we thank 23andMe, Inc., for sharing physical activity GWAS summary statistics. A full list of acknowledgements is provided in **Supplementary Information 13.**

Author Contributions

DJB, CCC, PDK, and SFWM designed and oversaw the study. RdV proposed the phenotype construction. SFWM was the lead analyst, responsible for GWAS, quality control, meta-analysis, summarizing the overlap across the results of the various GWAS, heritability analyses, genetic correlation analysis, phenotypic correlation analyses, out-of-sample prediction, and all bioinformatics analyses. RKL assisted with GWAS in UKB, and RKL and AO assisted with cohort-level quality control. CL performed the replication analyses, and was supervised by PT and DJB. CAPB performed the impG imputation of DietGen summary statistics. SFWM prepared the majority of figures with assistance from RKL; PB and CL also prepared some figures. JJL provided helpful advice and feedback on various aspects of the study design. All authors contributed to and critically reviewed the manuscript. DJB, CCC, PDK and SFWM made especially major contributions to the writing and editing.

Competing interests

Pauline M Emmett was funded by Nestlé Nutrition. The authors declare no other competing interests.

Additional information

Supplementary Information is available for this paper at [URL]. Correspondence and requests for materials should be addressed to SFWM (s.f.w.meddens@vu.nl), PDK (p.d.koellinger@vu.nl), CCC (carsonc@niddk.nih.gov), or DJB (daniel.benjamin@gmail.com).

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures and tables

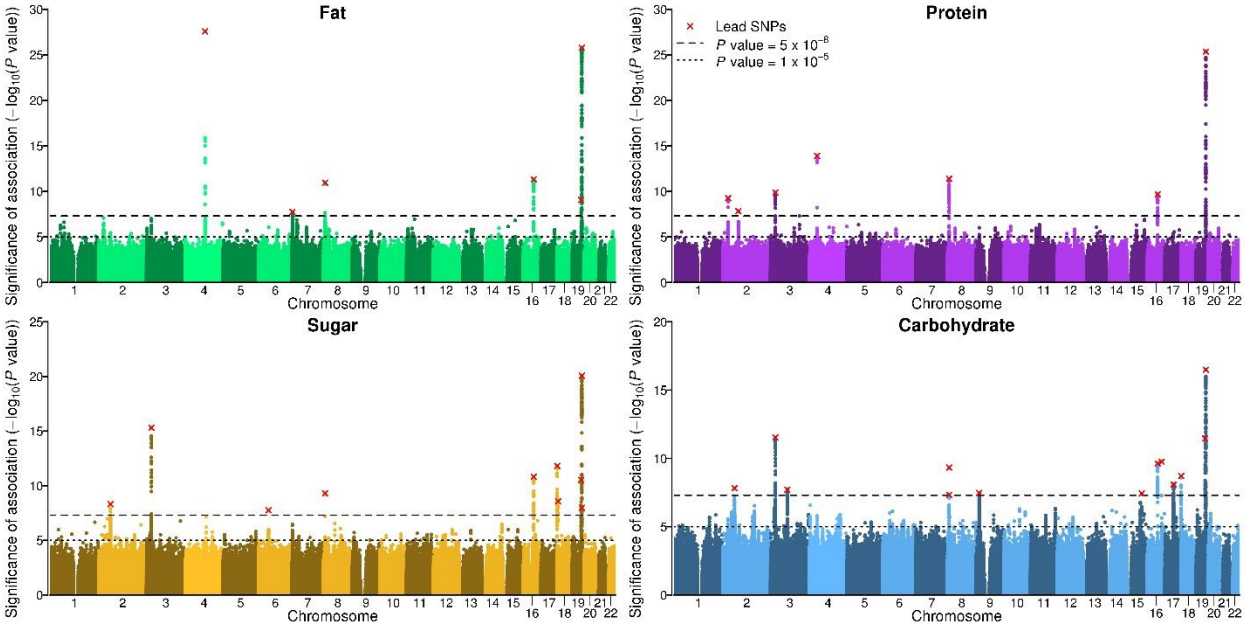


Figure 1 Manhattan plots | The *x*-axis is SNP chromosomal position; the *y*-axis is the SNP *P* value on a $-\log_{10}$ scale; the horizontal dashed line marks the threshold for genome-wide ($P = 5 \times 10^{-8}$) and suggestive ($P = 1 \times 10^{-5}$) significance; and each approximately independent (pairwise $r^2 < 0.1$) genome-wide significant association ("lead SNP") is marked by a red cross.

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

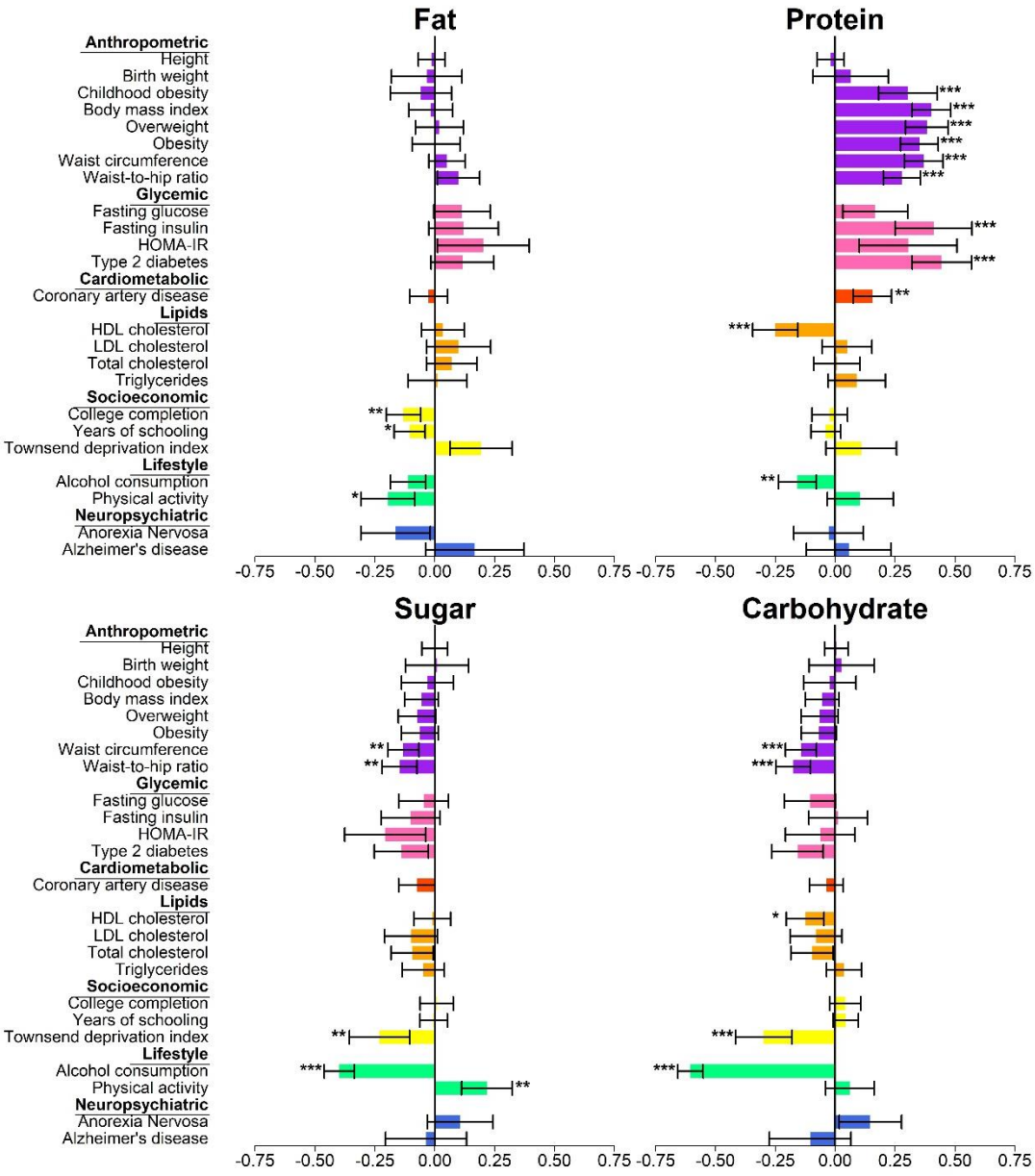


Figure 2 Genetic correlations | Genetic correlations were estimated with bivariate LD Score (LDSC) regression. Error bars show 95% confidence intervals, while asterisks denote Bonferroni-corrected *P* value thresholds (* $P_{Bonferroni} < 0.05$, ** < 0.01, *** < 0.001), corrected for 24 traits. The colours represent the different functional domains.

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

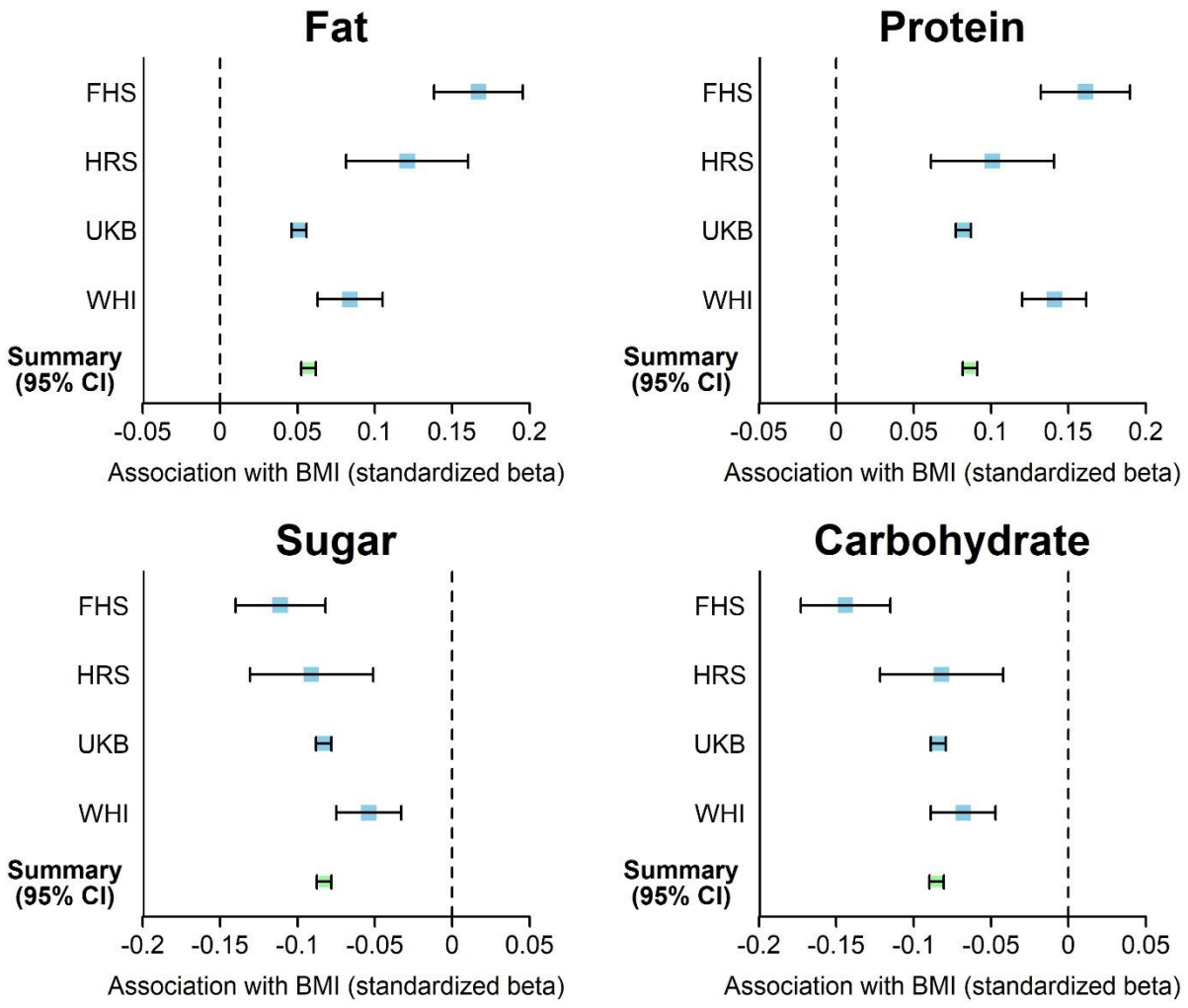


Figure 3 Phenotypic associations with Body Mass Index | Forest plots depicting the phenotypic associations between diet composition and Body Mass Index (BMI) in four independent cohorts, in terms of standardized betas (with errors bars indicating 95% confidence intervals). These standardized regression coefficients were obtained from a regression of BMI on the focal macronutrient and several covariates (sex, age, educational attainment, and household income). FHS = Framingham Heart Study (N = 4,413), HRS = Health and Retirement Study (N = 2,394), UKB = UK Biobank (N = 158,046), WHI = Women's Health Initiative (N = 8,628). The summary estimate was based on a fixed-effects, inverse-variance weighted meta-analysis of all four cohorts.

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

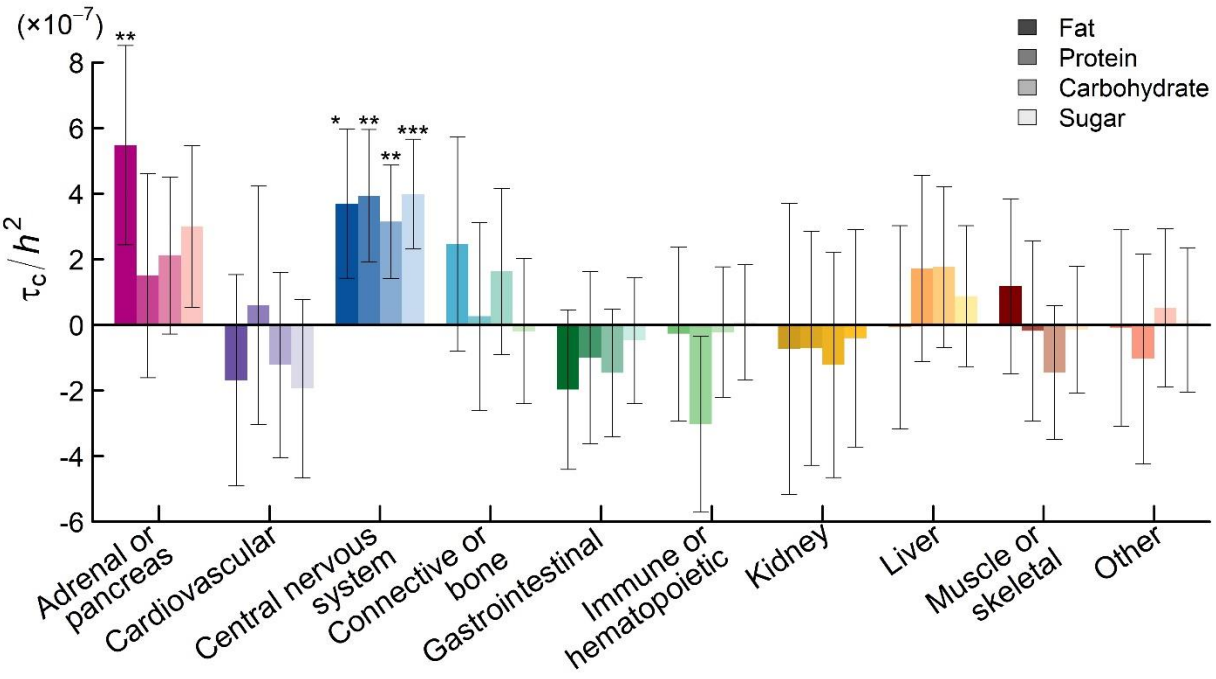


Figure 4 LD Score partitioning of heritability – Tissues | Functional partitioning of the heritability of diet composition phenotypes with stratified LD Score regression, where tissues were ascertained by Finucane et al. on the basis of chromatin data. The panel shows the partial regression coefficient (τ_C) from the stratified regression, divided by the LD Score heritability of the diet composition phenotype (h^2). Each estimate of τ_C comes from a separate stratified LD Score regression, where we also controlled for the 52 functional annotation categories in the "baseline" model. Error bars represent 95% confidence intervals. The phenotypes are ordered from left to right (FAT, PROTEIN, SUGAR, and CARBOHYDRATE), from darker to lighter shades. Asterisks (*) denote significant deviation from zero after Bonferroni correction for 10 tissues: * $P < \frac{0.05}{10}$, ** $P < \frac{0.01}{10}$, *** $P < \frac{0.001}{10}$.

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

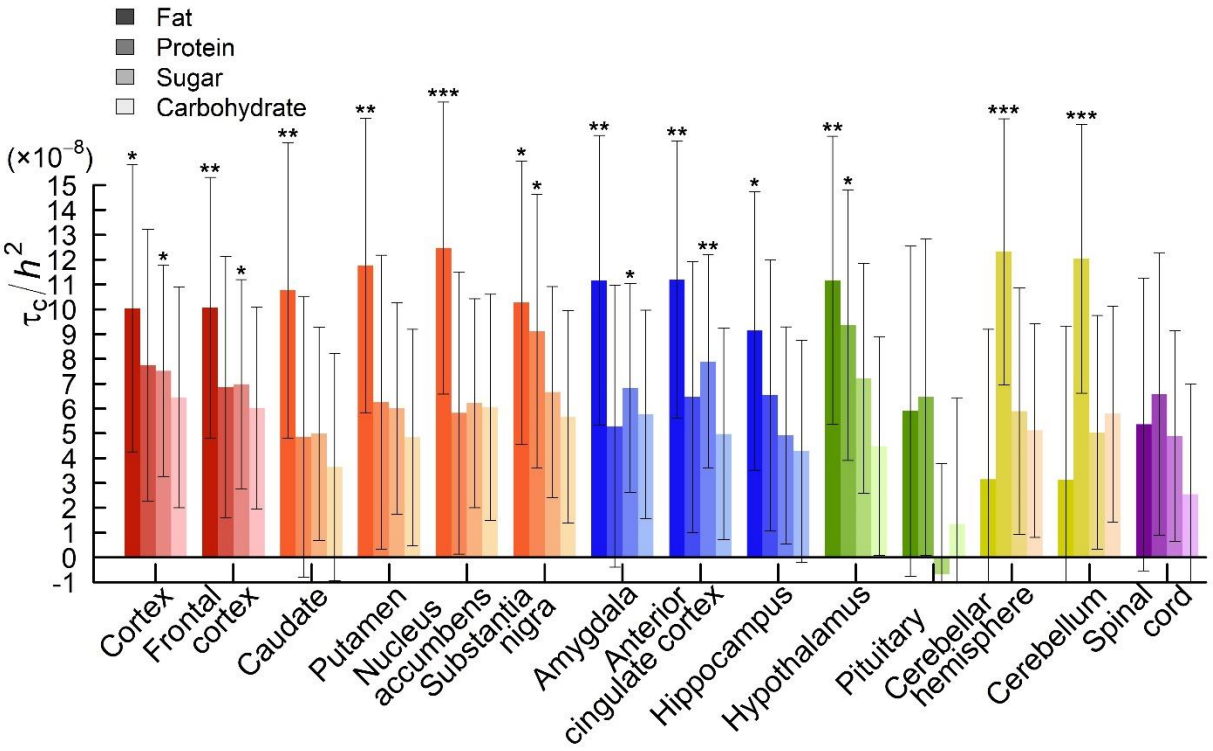


Figure 5 LD Score partitioning of heritability – Brain regions | Functional partitioning of the heritability of diet composition phenotypes with stratified LD Score regression, where tissues were ascertained by Finucane et al. on the basis of sets of specifically-expressed genes in GTEx data ("LDSC-SEG"). The sets of specifically-expressed genes in these analyses compared the focal tissue to other bodily tissues. The panel shows the partial regression coefficient (τ_C) from the stratified regression, divided by the LD Score heritability of the diet composition phenotype (h^2) to facilitate comparison between traits. Each estimate of τ_C comes from a separate stratified LD Score regression, where we also controlled for the 52 functional annotation categories in the "baseline" model. Error bars represent 95% confidence intervals. Asterisks (*) denote significant deviation from zero after Bonferroni correction for 53 tissues; * $P < \frac{0.05}{53}$, ** $P < \frac{0.01}{53}$, *** $P < \frac{0.001}{53}$. Each group of colored bars represents an anatomical region (ordered from left to right: red – cortex, orange – basal ganglia, blue – limbic system, green – hypothalamus-pituitary, yellow – cerebellum, and purple – spinal cord).

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Top hit in locus for	SNPID	CHR	BP	Effect allele	Beta	*P* value	Nearest gene
PROTEIN	rs780094	2	27,741,237	t	0.018	5.58E-10	*GCKR*
SUGAR	rs12713415	2	60,205,134	c	-0.019	4.88E-09	*AC007100.1*
CARBOHYDRATE	rs10206338	2	60,209,981	a	-0.016	1.52E-08	*AC007100.1*
PROTEIN	rs445551	2	79,697,982	a	0.019	1.49E-08	*CTNNA2*
CARBOHYDRATE	rs10510554	3	25,099,776	t	0.019	2.94E-12	*AC133680.1*
PROTEIN	rs1603978	3	25,108,236	a	0.019	1.35E-10	*AC092422.1*
SUGAR	rs7619139	3	25,110,415	a	-0.024	4.98E-16	*AC092422.1*
CARBOHYDRATE	rs10433500	3	85,546,798	a	0.016	1.96E-08	*CADM2*
PROTEIN	rs13146907	4	39,425,248	a	-0.022	1.24E-14	*KLB*
FAT	rs1229984	4	100,239,319	t	0.098	2.64E-28	*ADH1B*
SUGAR	rs13202107	6	51,395,463	a	-0.020	1.77E-08	*SNORD66*
FAT	rs57193069	7	1,862,417	a	-0.016	1.80E-08	*MAD1L1*
CARBOHYDRATE	rs7012637	8	9,173,209	a	0.017	4.68E-10	*AC022784.6*
FAT	rs7012814	8	9,173,358	a	-0.019	1.12E-11	*AC022784.6*
SUGAR	rs7012814	8	9,173,358	a	0.019	4.99E-10	*AC022784.6*
CARBOHYDRATE	rs9987289	8	9,183,358	a	-0.026	4.64E-08	*AC022784.6*
PROTEIN	rs1461729	8	9,187,242	a	0.032	4.09E-12	*AC022784.6*
CARBOHYDRATE	rs10962121	9	15,702,704	t	-0.015	3.40E-08	*CCDC171*
CARBOHYDRATE	rs2472297	15	75,027,880	t	-0.018	3.73E-08	*CYP1A1*
PROTEIN	rs55872725	16	53,809,123	t	0.018	2.09E-10	*FTO*
SUGAR	rs9972653	16	53,814,363	t	-0.020	1.53E-11	*FTO*
FAT	rs9927317	16	53,820,996	c	-0.024	4.77E-12	*FTO*
CARBOHYDRATE	rs7190396	16	53,822,502	t	0.018	2.39E-10	*FTO*
CARBOHYDRATE	rs1104608	16	73,912,588	c	0.018	1.74E-10	*AC087565.1*
CARBOHYDRATE	rs36123991	17	44,359,663	t	0.021	8.24E-09	*ARL17B*
SUGAR	rs8097672	18	1,839,601	a	0.030	1.54E-12	*AP005230.1*
CARBOHYDRATE	rs8097672	18	1,839,601	a	0.023	1.95E-09	*AP005230.1*
SUGAR	rs341228	18	6,395,336	t	0.019	2.72E-09	*L3MBTL4*
FAT	rs429358	19	45,411,941	t	0.024	8.65E-10	*APOE*
SUGAR	rs429358	19	45,411,941	t	-0.028	2.97E-11	*APOE*
CARBOHYDRATE	rs429358	19	45,411,941	t	-0.027	3.49E-12	*APOE*
FAT	rs33988101	19	49,218,111	t	-0.029	1.66E-26	*MAMSTR*
SUGAR	rs838144	19	49,250,239	t	-0.028	8.53E-21	*IZUMO1*
CARBOHYDRATE	rs838144	19	49,250,239	t	-0.023	3.26E-17	*IZUMO1*
PROTEIN	rs838133	19	49,259,529	a	-0.032	4.52E-26	*FGF21*
SUGAR	rs62132802	19	49,270,872	t	-0.020	1.07E-08	*FGF21*

Table 1 Diet composition lead SNPs | GWAS summary statistics of the 36 diet composition lead SNPs (i.e., the top hit in the locus for each phenotype). A total of 21 of these lead SNPs are approximately independent. **Supplementary Table 5.1** reports the effect alleles and summary statistics across all four phenotypes for each individual lead SNP. MAF = minor allele frequency (weighted average across cohorts). Beta = semi-standardized (i.e., increase in phenotypic standard

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

deviations per effect allele). All *P* values are genomic-controlled (corrected for LDSC intercept). All genomic coordinates are in GRCh37.

	FAT	PROTEIN	Sugar	CARBOHYDRATE
FAT	--	--	--	--
PROTEIN	-0.019 (0.068)	--	--	--
SUGAR	-0.513 (0.040)***	-0.307 (0.057)***	--	--
CARBOHYDRATE	-0.607 (0.032)***	-0.226 (0.048)***	0.728 (0.020)***	--

Table 2 Genetic correlations between macronutrients | Genetic correlation analysis results obtained from bivariate LD Score regression (with block jackknife standard errors in brackets). Only HapMap3 SNPs were used in this analysis. The results show the genetic correlations among the four phenotypes calculated using the summary statistics from the combined meta-analyses. *** Denotes *P* value < 0.001 for the null hypothesis of zero genetic correlation. All estimates also differed from 1 and -1 with *P* < 0.001.

Authors

S Fleur W Meddens[1,2,3]*, Ronald de Vlaming[1], Peter Bowers[4], Casper AP Burik[1], Richard Karlsson Linnér[1,2,3], Chanwook Lee[4], Aysu Okbay[1], Patrick Turley[5,6,7], Cornelius A Rietveld[8,9,3], Mark Alan Fontana[10,11], Mohsen Ghanbari[9,12], Fumiaki Imamura[13], George McMahon[14], Peter J van der Most[15], Trudy Voortman[9], Kaitlin H Wade[14], Emma L Anderson[14], Kim VE Braun[9], Pauline M Emmett[16], Tonũ Esko[17], Juan R Gonzalez[18,19,20], Jessica C Kiefte-de Jong[9,21], Jian'a Luan[13], Claudia Langenberg[13], Taulant Muka[9], Susan Ring[14], Fernando Rivadeneira[22], Josje D Schoufour[9], Harold Snieder[15], Frank JA van Rooij[9], Bruce HR Wolffenbuttel[23], *23andMe Research Team, EPIC-InterAct Consortium, Lifelines Cohort Study*, George Davey Smith[14], Oscar H Franco[9], Nita G Forouhi[13], M Arfan Ikram[9], Andre G Uitterlinden[22], Jana V van Vliet-Ostaptchouk[23], Nick J Wareham[13], David Cesarini[24], K Paige Harden[25], James J Lee[26], Daniel J Benjamin[27,7,28]*, Carson C Chow[29]*, Philipp D Koellinger[1]*

* Corresponding authors

[1] School of Business and Economics, VU University Amsterdam, De Boelelaan 1105, 1081 HV, Amsterdam, The Netherlands

bioRxiv preprint doi: https://doi.org/10.1101/383406; this version posted August 2, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[2] Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, VU University Amsterdam, De Boelelaan 1085, 1081 HV, Rotterdam, The Netherlands
[3] Erasmus University Rotterdam Institute for Behavior and Biology, Erasmus School of Economics, Erasmus University Rotterdam, Burgemeester Oudlaan 50, Rotterdam, 3062 PA, Rotterdam, The Netherlands
[4] Department of Economics, Harvard University, Cambridge, Massachusetts 02138, USA
[5] Analytical and Translational Genetics Unit, Massachusetts General Hospital, Richard B. Simches Research building, 185 Cambridge St, CPZN-6818, Boston, MA 02114, USA
[6] Stanley Center for Psychiatric Genomics, The Broad Institute at Harvard and MIT, 75 Ames St, Cambridge, MA 02142, USA
[7] Behavioral and Health Genomics Center, Center for Economic and Social Research, University of Southern California, 635 Downey Way, Los Angeles, CA 90089, USA
[8] Department of Applied Economics, Erasmus School of Economics, Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3062 PA, Rotterdam, The Netherlands
[9] Department of Epidemiology, Erasmus MC, University Medical Center, Wytemaweg 80, 3015 GE, Rotterdam, The Netherlands
[10] Center for the Advancement of Value in Musculoskeletal Care, Hospital for Special Surgery, 535 East 70th Street, New York, NY 10021, USA
[11] Department of Healthcare Policy and Research, Weill Cornell Medical College, Cornell University, 402 East 67th Street, New York, NY 10065, USA
[12] Department of Genetics, School of Medicine, Mashhad University of Medical Sciences, Azadi Square, University Campus, 9177948564, Mashhad, Iran
[13] MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus Cambridge, CB2 0QQ Cambridge, United Kingdom
[14] Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Oakfield House, Oakfield Grove, BS8 2BN, Bristol, United Kingdom
[15] Department of Epidemiology, University of Groningen, University Medical Center Groningen, Hanzeplein 1, 9713 GZ Groningen, The Netherlands
[16] Population Health Sciences, Bristol Medical School, University of Bristol, Oakfield House, Oakfield Grove, BS8 2BN, Bristol, United Kingdom
[17] Estonian Genome Center, University of Tartu, Tartu, 51010, Estonia
[18] Barcelona Institute for Global Health (ISGlobal), Barcelona, 8003, Spain
[19] Universitat Pompeu Fabra (UPF), Barcelona, 8003, Spain
[20] CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, 280229, Spain
[21] Leiden University College, Anna van Buerenplein 301, 2595 DG, Den Haag, The Netherlands
[22] Department of Internal Medicine, Erasmus MC University Medical Center, Wytemaweg 80, Rotterdam, 3015 GE, the Netherlands
[23] Department of Endocrinology, University of Groningen, University Medical Center Groningen, Hanzeplein 1, 9713 GZ, Groningen, The Netherlands
[24] Department of Economics, New York University, 19 W. 4th Street, New York, NY 10012, USA
[25] Department of Psychology, University of Texas at Austin, 108 E. Dean Keeton Stop #A8000, Austin, TX 78704, USA
[26] Department of Psychology, University of Minnesota Twin Cities, 75 East River Parkway, Minneapolis, MN 55455, USA
[27] National Bureau of Economic Research, 1050 Massachusetts Ave, Cambridge, MA 02138, USA
[28] Department of Economics, University of Southern California, 635 Downey Way, Los Angeles, CA 90089, USA
[29] Laboratory of Biological Modeling, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA