nature
human behaviour

ARTICLES
https://doi.org/10.1038/s41562-019-0805-1

GWAS of 165,084 Japanese individuals identified nine loci associated with dietary habits

Nana Matoba[1,2], Masato Akiyama[1,3], Kazuyoshi Ishigaki[1], Masahiro Kanai[1,4], Atsushi Takahashi[1,5], Yukihide Momozawa[6], Shiro Ikegawa[7], Masashi Ikeda[8], Nakao Iwata[8], Makoto Hirata[9], Koichi Matsuda[10], Yoshinori Murakami[11], Michiaki Kubo[12], Yoichiro Kamatani[1,13]* and Yukinori Okada[14,15,16]*

Dietary habits are important factors in our lifestyle, and confer both susceptibility to and protection from a variety of human diseases. We performed genome-wide association studies for 13 dietary habits including consumption of alcohol (ever versus never drinkers and drinks per week), beverages (coffee, green tea and milk) and foods (yoghurt, cheese, natto, tofu, fish, small whole fish, vegetables and meat) in Japanese individuals ($n = 58{,}610$–$165{,}084$) collected by BioBank Japan, the nationwide hospital-based genome cohort. Significant associations were found in nine genetic loci (*MCL1-ENSA*, *GCKR*, *AGR3-AHR*, *ADH1B*, *ALDH1B1*, *ALDH1A1*, *ALDH2*, *CYP1A2-CSK* and *ADORA2A-AS1*) for 13 dietary traits ($P < 3.8 \times 10^{-9}$). Of these, ten associations between five loci and eight traits were new findings. Furthermore, a phenome-wide association study revealed that five of the dietary trait-associated loci have pleiotropic effects on multiple human complex diseases and clinical measurements. Our findings provide new insight into the genetics of habitual consumption.

Every individual has a different lifestyle including habitual dietary consumption, which is one of the keys to a healthy life. Considering that dietary habits confer both susceptibility to, and protection from, a variety of human diseases, elucidation of the latent factors affecting the individual's dietary habits should have beneficial clinical impacts. Variation in dietary patterns among individuals could be driven by several factors such as cultural lifestyle, environment and ethnicity. Previous studies have indicated the contribution of population-specific genetic factors to alcohol dependence in close relationship with population-specific natural selection signatures[1–3], which should warrant further genetic studies investigating the link with dietary habits in a broad spectrum of worldwide populations. In addition to alcohol consumption[4–11], recent large-scale, genome-wide association studies (GWAS) have successfully identified the genetic factors that influence habitual dietary consumption of a variety of foods and drinks, including coffee[11–16], fish[17] and various nutrients[18]. These results implicate the contribution of genetic heritability to wider ranges of dietary habits to a far greater extent than previously expected, warranting further genetic studies based on cohorts with extensive phenotypic records of dietary habits.

The BioBank Japan (BBJ) Project (https://biobankjp.org/english/index.html), a nationwide hospital-based genome cohort in Japan, has collected data on various types of dietary habits as well as the affected status of 47 complex diseases and clinical measurements[19]. The dietary items include traditional Japanese beverages and foods, such as green tea and fermented soybean (natto). Such dietary habit information could be a useful resource in investigating the link with susceptibility to, and outcome of, human diseases. For example, an inverse association was reported between mortality due to cardiovascular diseases and green tea consumption[20]. Epidemiological studies indicated that greater consumption of natto was associated with reduced risk of arterial stiffness[21]. However, whether genetic factors contribute to dietary habits among individuals remains unclear, except in the case of certain well-investigated traits such as alcohol and coffee. BBJ collects dietary habits using a food-frequency questionnaire (FFQ), which was developed as a tool to measure dietary consumption in epidemiological studies. For example, the Japan Public Health Centre-based Prospective Study[22] validated a self-administrated FFQ for the identification of potential causal evidence of the influence of dietary habits on disease distribution.

In the current study, to comprehensively investigate genetic factors contributing to dietary habits we conducted large-scale GWAS on 13 dietary habits of drink and food consumption with 165,084 Japanese subjects enroled in BBJ. To further elucidate the genetic architecture of habitual consumption, we comprehensively conducted a cross-trait genetic correlation analysis among dietary habits, as well as a phenome-wide association study (PheWAS) using 45 human complex diseases and 58 clinical measurements.

[1]Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [2]Department of Genetics, UNC Neuroscience Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [3]Department of Ophthalmology, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan. [4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [5]Department of Genomic Medicine, Research Institute, National Cerebral and Cardiovascular Center, Suita, Japan. [6]Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [7]Laboratory for Bone and Joint Diseases, RIKEN Center for Integrative Medical Sciences, Tokyo, Japan. [8]Department of Psychiatry, Fujita Health University School of Medicine, Toyotake, Japan. [9]Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [10]Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan. [11]Division of Molecular Pathology, the Institute of Medical Sciences, The University of Tokyo, Tokyo, Japan. [12]RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [13]Laboratory of Complex Trait Genomics, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan. [14]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan. [15]Laboratory of Statistical Immunology, Immunology Frontier Research Center, Osaka University, Suita, Japan. [16]Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita, Japan. *e-mail: yoichiro.kamatani@riken.jp; yokada@sg.med.osaka-u.ac.jp

Results

Characteristics of the samples studied. The BBJ cohort[19] collected DNA and clinical information, including items related to participants' lifestyle such as dietary habits, which were obtained through interviews and reviews of medical records using a standardized questionnaire. Of the 171,979 Japanese individuals with clinical records and GWAS genotyped data that passed a quality check as described previously[23,24] (Supplementary Fig. 1), those with information available on dietary habits were included in the final analysis. In total, data from 165,738 individuals (89,784 males and 75,954 females) were analysed in this study. The mean body mass index (BMI) for males and females was 23.55 and 23.00 kg m^{-2}, respectively (see detailed characteristics in Supplementary Tables 1 and 2). All individuals were diagnosed with one of the targeted 45 diseases (Supplementary Fig. 2 and Supplementary Table 1). To minimize any potential effects of disease affection status, we used the affected status of individuals as covariates in the genetic association analysis (Supplementary Table 3). The numbers of individuals analysed for each of the dietary items were as follows: alcohol phenotypes (ever versus never drinker ($n=165{,}084$) and drinks per week ($n=58{,}610$)); three types of beverage (coffee ($n=152{,}634$), green tea ($n=152{,}653$) and milk ($n=152{,}965$)); and eight types of food (yoghurt ($n=152{,}097$), cheese ($n=152{,}714$), natto ($n=152{,}678$), tofu ($n=152{,}943$), fish ($n=153{,}048$), small whole fish ($n=152{,}277$), vegetables ($n=153{,}001$) and meat ($n=152{,}857$)). In total, 13 dietary habits from 12 questionnaire items administered to Japanese subjects were analysed (Supplementary Table 2).

GWAS of 13 dietary habits on >160,000 Japanese subjects. In the GWAS on the 13 dietary habits, we tested the associations of 5,961,480 and 147,348 imputed single-nucleotide polymorphisms (SNPs) in autosomes and the X chromosome, respectively, with minor allele frequency (MAF) >0.01 and imputation quality score >0.7. The genomic inflation factor (λ_{GC}) and the linkage disequilibrium score regression (LDSC)[25] intercepts of the GWAS summary statistics ranged from 1.05 to 1.25 (mean = 1.13) and 1.02 to 1.12 (mean = 1.02), respectively (Supplementary Table 2 and Supplementary Fig. 3).

In total, nine distinct genetic loci (>1 Mb apart) satisfying the significance threshold of multiple comparisons based on the number of the tested traits were identified (that is, $P=(5.0\times10^{-8})/13=3.8\times10^{-9}$; Fig. 1, Table 1 and Supplementary Figs. 4 and 5). Of these nine loci, ten associations in five regions have not been reported in previous GWAS on dietary habits: *ALDH1B1* at 9p13 and *ALDH1A1* at 9q21 with ever versus never drinkers, *MCL1-ENSA* at 1q21 and *ADORA2A-AS1* at 22q11 with coffee consumption, and *ALDH2* at 12q24 with the six phenotypes (consumption of green tea, milk, yoghurt, natto, tofu and fish). We replicated the 13 associations previously reported in European (EUR) or Japanese populations ($P<5.0\times10^{-8}$; Table 1 and Supplementary Table 4). For example, a lead SNP at *ABCG2* (rs75544042; $P=4.9\times10^{-8}$) was in strong linkage disequilibrium ($r^2=0.94$ and 1.00 in the EUR population and the East Asian (EAS) population, respectively, in

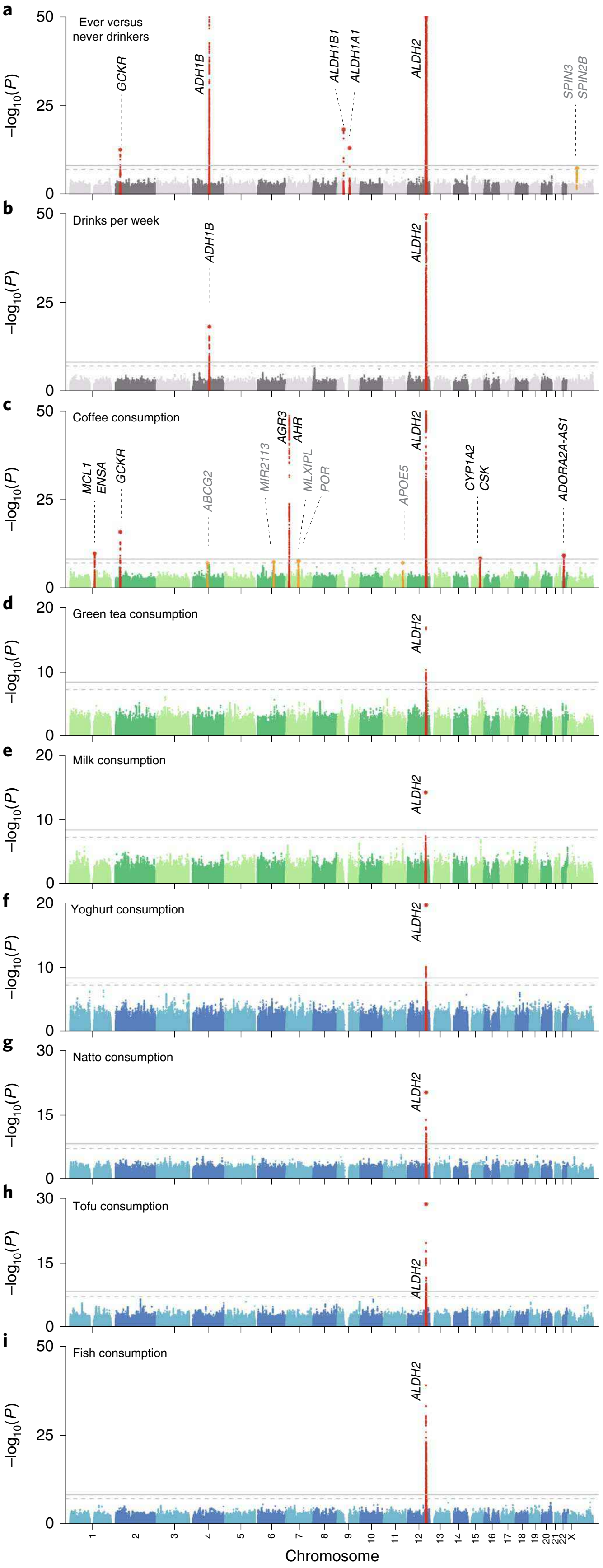

Fig. 1 | Manhattan plots for dietary habits from GWAS in a Japanese population. a–i, Manhattan plots from GWAS on dietary habits. Traits with association signals that satisfied the genome-wide significance threshold are indicated ($P<5.0\times10^{-8}$). Ever versus never drinkers ($n_{total}=165{,}064$) (**a**), drinks per week ($n_{total}=58{,}610$) (**b**) and consumption of coffee ($n_{total}=152{,}634$) (**c**), green tea ($n_{total}=152{,}653$) (**d**), milk ($n_{total}=152{,}965$) (**e**), yoghurt ($n_{total}=152{,}907$) (**f**), natto ($n_{total}=152{,}678$) (**g**), tofu ($n_{total}=152{,}943$) (**h**) and fish ($n_{total}=153{,}048$) (**i**). The *x* axis shows the position on each chromosome, while the *y* axis shows the $-\log_{10}(P)$ of SNPs. Solid and dotted grey lines indicate the study-wide ($P=3.8\times10^{-9}$) and genome-wide significance levels ($P=5.0\times10^{-8}$), respectively. Red and orange lines indicate regions satisfying the study-wide and genome-wide significance thresholds, respectively (see Supplementary Fig. 4 for results of remaining traits).

Table 1 | Associations of genomic loci with dietary habits in the Japanese population

Category	SNP	Position (hg19)	Allele	Alt. frequency			Nearby genes	β^a	s.e.m.	$P_{corrected}{}^b$	Refs. [c]	UKBB	
			Ref/alt.	GWAS	EAS	EUR						β	P
Significantly associated loci: $P < (5.0\times10^{-8})/13 = 3.8\times10^{-9}$													
Ever/never drinkers	rs1260326	2:27730940	T/C	0.44	0.52	0.59	*GCKR*	0.011	0.0015	1.5×10^{-13}	9	0.002	3.3×10^{-3}
Ever/never drinkers	rs1229984	4:100239319	T/C	0.22	0.30	0.97	*ADH1B*	0.038	0.0021	1.6×10^{-72}	8,30	0.022	4.6×10^{-18}
Ever/never drinkers	rs3043	9:38397355	G/C	0.69	0.70	0.76	*ALDH1B1*	−0.015	0.0016	3.3×10^{-19}	-	−0.001	0.37
Ever/never drinkers	rs8187929	9:75540504	T/A	0.03	0.05	0.00	*ALDH1A1*	0.031	0.0041	4.6×10^{-14}	-	NA	NA
Ever/never drinkers	rs671	12:112241766	G/A	0.25	0.17	0.00	*ALDH2*	−1.815	0.0121	$<1.0\times10^{-4,740}$	4,8,29	NA	NA
Drinks per week	rs1229984	4:100239319	T/C	0.23	0.30	0.97	*ADH1B*	0.070	0.0078	4.3×10^{-19}	8,9	−0.26	6.1×10^{-177}
Drinks per week	rs671	12:112241766	G/A	0.12	0.17	0.00	*ALDH2*	−0.430	0.0092	$<1.0\times10^{-450}$	4,8,29	NA	NA
Coffee	rs6681426	1:150586971	G/A	0.66	0.65	0.64	*MCL1, ENSA*	−0.078	0.0118	1.1×10^{-10}	-	−0.025	5.7×10^{-9}
Coffee	rs1260326	2:27730940	T/C	0.44	0.52	0.59	*GCKR*	0.096	0.0113	9.9×10^{-17}	14	0.040	6.2×10^{-21}
Coffee	rs4410790	7:17284577	T/C	0.37	0.41	0.62	*AGR3, AHR*	0.213	0.0119	8.1×10^{-68}	12,14,26,27	0.12	7.6×10^{-175}
Coffee	rs671	12:112241766	G/A	0.25	0.17	0.00	*ALDH2*	0.354	0.0131	3.8×10^{-153}	16,28	NA	NA
Coffee	rs58806801	15:75059546	G/A	0.22	0.24	0.07	*CYP1A2, CSK*	−0.084	0.0137	2.4×10^{-9}	12–14,26	−0.039	2.2×10^{-5}
Coffee	rs5760444	22:24878218	T/C	0.59	0.56	0.59	*ADORA2A-AS1*	−0.073	0.0114	3.7×10^{-10}	-	0.0056	0.18
Tea	rs671	12:112241766	G/A	0.25	0.17	0.00	*ALDH2*	0.090	0.0101	2.8×10^{-18}	-	NA	NA
Milk	rs671	12:112241766	G/A	0.25	0.17	0.00	*ALDH2*	0.113	0.0129	6.4×10^{-18}	-	NA	NA
Yoghurt	rs671	12:112241766	G/A	0.25	0.17	0.00	*ALDH2*	0.113	0.0117	6.0×10^{-21}	-	NA	NA
Natto	rs671	12:112241766	G/A	0.25	0.17	0.00	*ALDH2*	−0.114	0.0106	2.7×10^{-24}	-	NA	NA
Tofu	rs671	12:112241766	G/A	0.25	0.17	0.00	*ALDH2*	−0.105	0.0097	2.5×10^{-25}	-	NA	NA
Fish	rs671	12:112241766	G/A	0.25	0.17	0.00	*ALDH2*	−0.134	0.0087	6.5×10^{-50}	-	NA	NA
Suggested loci; $3.8\times10^{-9} < P < 5.0\times10^{-8}$													
Ever/never drinkers	rs150096	X:57105278	G/C	0.25	0.39	0.28	*SPIN3, SPIN2B*	−0.044	0.0078	2.0×10^{-8}	-	NA	NA
Coffee	rs75544042	4:89045331	A/G	0.70	0.71	0.91	*ABCG2*	0.069	0.0123	4.9×10^{-8}	14	NA	NA
Coffee	rs12189679	6:98333409	G/A	0.36	0.41	0.49	*MIR2113*	0.067	0.0117	2.5×10^{-8}	-	0.016	1.7×10^{-4}
Coffee	rs13234378	7:73026151	A/T	0.10	0.12	0.12	*MLXIPL*	0.108	0.0186	1.6×10^{-8}	14	0.063	2.2×10^{-24}
Coffee	rs3815455	7:75611756	C/T	0.41	0.37	0.30	*POR*	0.067	0.0114	1.1×10^{-8}	14	0.060	1.4×10^{-39}
Coffee	rs662799	11:116663707	G/A	0.65	0.71	0.92	*APOE5*	0.066	0.0117	4.2×10^{-8}	-	0.0065	0.46

NA, not available. Ref/alt, reference allele / alternative allele. [a] β indicates effect size for each SNP from BOLT-LMM, except for three SNPs: rs671, calculated using either glm or lm of R function for binominal traits or quantitative traits, respectively; and tw on the X chromosome. Meta-analysis results of male/female GWAS, $P<1.0\times10^{-300}$ were calculated using R with the Rmpfr package. [b] P values were corrected by applying genomic control. [c] Refs. indicates reference(s) where a corresponding locus was previously reported.

1000 Genome Project phase 1 (1KGP)), with rs14801012 reported by the Coffee and Caffeine Genetic Consortium[14]. Likewise, other loci associated with coffee consumption, including *GCKR*[11,14], *AGR3-AHR*[11,12,14,26,27] *CYP1A2-CSK*[11–14,26] and *ADORA2A-AS1* (ref. [11]), were also confirmed in the current study, suggesting that these loci may have biological effects on coffee consumption or caffeine metabolism shared across multiple ancestries. *ALDH2* (ref. [16,28]) for coffee consumption was also confirmed. The remaining replicated associations were *GCKR* (ref. [9]) and the well-known associations of *ALDH2* (ref. [4,5,8]) and *ADH1B*[8,9,29] with alcohol consumption. Furthermore, 13 additional loci showed at least nominal evidence of replication ($P < 1.0\times10^{-2}$), including rs11940694 at *KLB* with alcohol consumption ($P = 1.5\times10^{-4}$ for ever versus never drinkers, and $P = 4.6\times10^{-6}$ for drinks per week), and rs6265 near *BDNF* with coffee consumption ($P = 3.9\times10^{-6}$) (Table 1 and Supplementary Table 4). Stepwise conditional analysis further identified three additional loci associated with coffee consumption at 7p21 and 7q11 (Supplementary Table 5 and Supplementary Fig. 6).

We next compared allele frequency spectra and effect sizes of the associated SNPs across different ancestries. The MAF of the lead SNPs on the 16 independent associated loci ranged from 0.03 to 0.44 in Japanese (and a similar frequency range (0.05–0.48) in 1KGP EAS; Supplementary Fig. 7). Of these, two SNPs (rs8187929 and rs671) were monomorphic while rs1229883 was rare (MAF = 0.029) in the EUR population, but were more common in the EAS population (MAF = 0.046, 0.174 and 0.303 for rs8187929, rs671 and, rs1229984, respectively; Table 1 and Supplementary Fig. 7). We obtained dietary habit association results from UK Biobank (UKBB) through the Gene ATLAS database[30] (http://geneatlas.roslin.ed.ac.uk/; Table 1 and Supplementary Fig. 8) to reference the effects of our lead SNPs

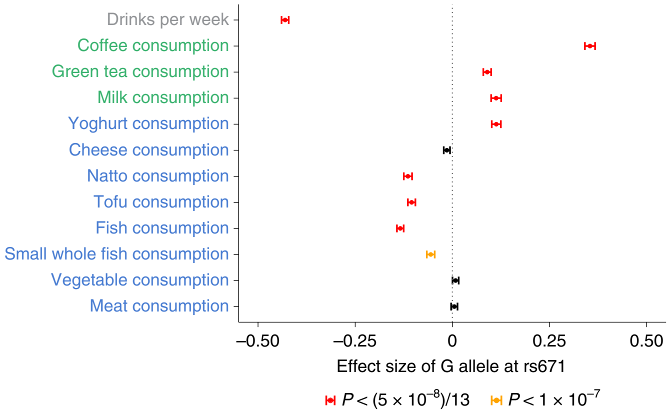

Fig. 2 | Effect of functional *ALDH2* variant (rs671) on dietary habits. Effect size estimates of the functional variant of *ALDH2*, rs671, on 12 phenotype categories. Circles and line widths represent the estimated relative risk and 95% confidence interval of RR, respectively, of each category. Red and orange denote statistical significance at $P < 5.0 \times 10^{-8}$ and $P < 1.0 \times 10^{-7}$, respectively.

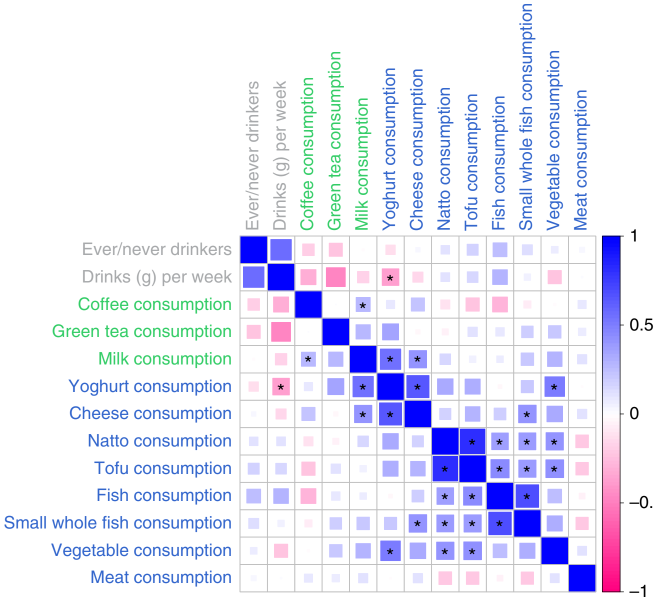

Fig. 3 | Genetic correlations among dietary habits. Full results of genetic correlations are listed in Supplementary Table 9. Blue and red represent positive and negative correlations, respectively. The size of the squares indicates significance level. Asterisks, FDR < 0.01.

in European individuals. In total, association results of 12 of the 16 lead SNPs were available in the UKBB results. We found that 9 of the 12 SNPs showed significant association ($P < 0.05/27 = 0.0019$) with the corresponding phenotypes in UKBB with the same directional effects, suggesting that these SNPs are involved in dietary traits beyond ethnicity. On the other hand, there was no association found between rs3043 and alcohol-drinking habits in UKBB ($P = 0.37$), while this SNP had similar MAF values in both populations (0.300 and 0.238 in EAS and EUR, respectively). While we did not observe an association of rs5760444 with coffee intake ($P = 0.18$), its association with tea intake was observed in UKBB ($P = 1.4 \times 10^{-37}$).

Specific SNPs considered to be related to dietary habits were also investigated (Supplementary Table 7). We observed associations with ever versus never drinkers and the SNPs responsible for phenylthiocarbamide bitter taste (for example, rs713598-G in *TAS2R38*, $\beta = -0.006$, $P = 2.2 \times 10^{-5}$), indicating that individuals who had never drunk alcohol are unlikely to prefer bitter taste. A further example of the association with nominal significance ($P < 0.05$) was tofu (rs10246939-C, $\beta = -0.018$, $P = 0.044$), which contains Nigari (magnesium chloride), suggesting the hypothesis that individuals who eat tofu are often likely to have the ability to taste bitter substances. Consumption of five items—fish, small whole fish, natto, tofu and vegetables—was found to be associated with rs307355-C (known to be sensitive to the *Umami* taste) at a nominal significance level ($P = 0.037$, 0.043, 0.034, 0.0088 and 0.013, respectively). Because the well-known lactose-intolerance SNPs at genes *MCM6* and *LCT* (for example, rs4988235 and rs182549) are monomorphic in the Japanese population, no association was observed in our GWAS.

Functional annotation of the lead SNPs (and variants in strong linkage disequilibrium ($r^2 > 0.8$)) was performed to investigate the biological effects of the identified associated variants (Supplementary Table 8). Six non-synonymous variants—rs1336900 on *HORMAD1*, rs2230061 on *CTSS*, rs1260326 on *GCKR*, rs1229984 on *ADH1B*, rs1057868 on *POR* and rs818787929 on *ALDH1A*—were identified. rs3043 in *ALDH1B* and rs1260326 in *GCKR* were also found to be overlapped ($r^2 > 0.8$ in EAS and EUR) with the *cis*-expression quantitative trait locus (*cis*-eQTL) lead SNP, according to the Gene-Tissue Expression (GTEx) Project database[31] (Supplementary Table 9).

Pleiotropic effects of loci *ALDH2* and *GCKR* on dietary habits. In this study, two loci with abundant pleiotropic effects among dietary habits were identified. One is the A*LDH2* region at 12q24, which is well known through its pleiotropic effects on various phenotypes including alcohol consumption, as represented by the functional Asian-specific non-synonymous variant of rs671 (p.Glu457Lys of *ALDH2*; $P < 1.0 \times 10^{-4,740}$ for ever versus never drinkers, and $P < 1.0 \times 10^{-450}$ for drinks per week in our study)[23,24,32,33]. An additional pleiotropic effect of rs671 across dietary habits on foods and beverages was also found (coffee, green tea, yoghurt, natto, tofu and fish): notably, the directional effects of rs671 on consumption were heterogeneous among traits (Fig. 2 and Supplementary Fig. 9). Comparisons of directional effects showed that the derived A allele of rs671 (Glu > Lys) had decreasing effects on the habitual consumption of alcohol, natto, tofu and fish, while having increasing effects on the habitual consumption of coffee, green tea, milk and yoghurt. Since the effects of rs671 on alcohol-related phenotypes were strong and alcohol drinking itself could affect other dietary habits, it is possible that the observed pleiotropy of rs671 could, partly and indirectly, reflect phenotypic correlations with alcohol drinking. We note that the rs671 genotype was initially filtered out as a significant deviation from Hardy–Weinberg equilibrium (HWE), which was not due to genotyping error (see details in Methods). In regard to another pleiotropic region of *GCKR* at 2p23, the non-synonymous variant rs1260326 (p.Leu446Pro) was associated with ever versus never drinkers and coffee consumption, with the same directional dosage effects on both traits. This SNP is also well known through having pleiotropic effects on many types of metabolism[24,34], such as triglycerides and total cholesterol.

Genetic differences in dietary habits between males and females. To characterize the sex-specific genetic architecture of dietary habits, sex-stratified GWAS on dietary habits were conducted, which identified additional loci with sex-specific associations with yoghurt consumption (*LINC00635* at 3q13 in males; Supplementary Figs. 3, 10 and 11).

The subsequent analysis using cross-sex genetic correlation found no evidence in the hypothesis that dietary habits genetically differed across sexes after multiple corrections (lowest $p_{(r_g)} =$

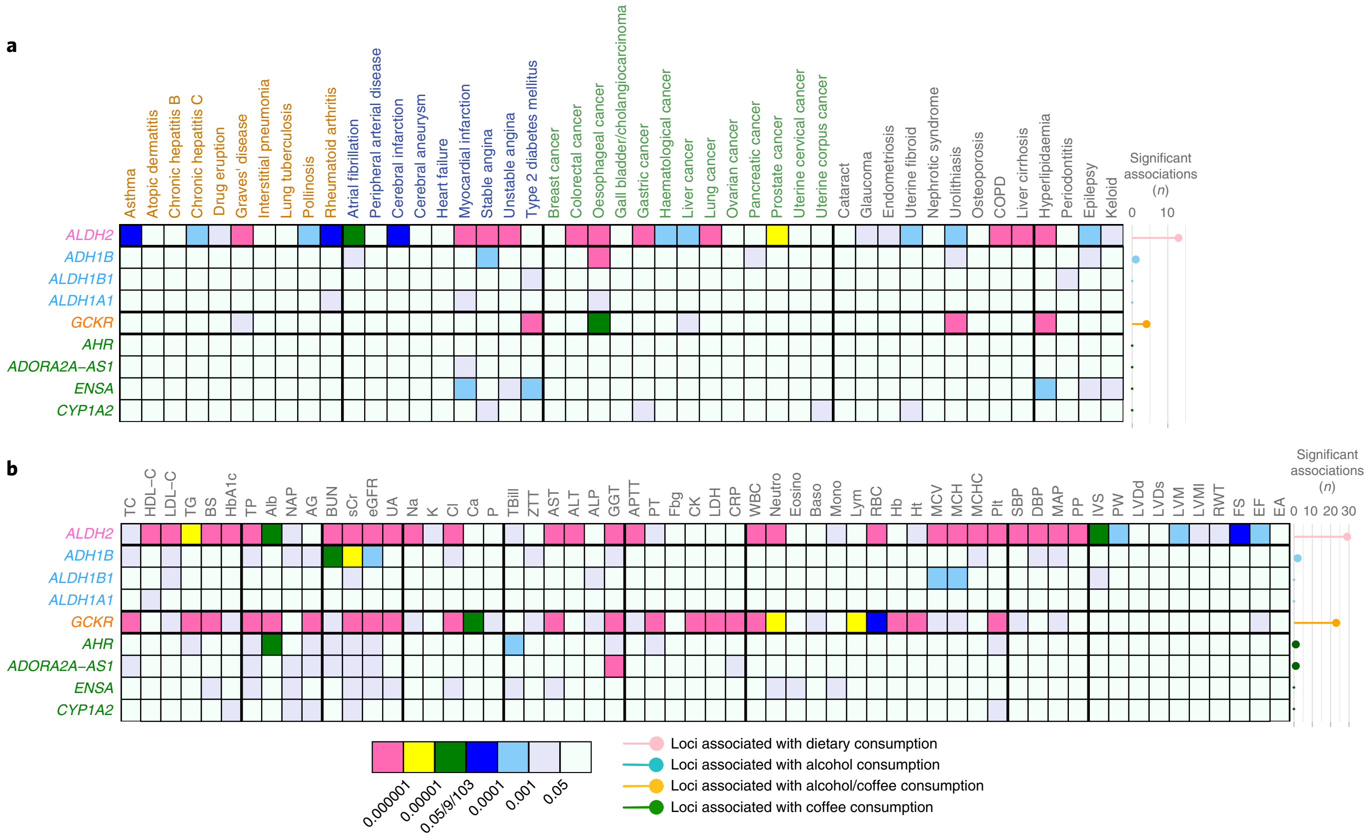

Fig. 4 | PheWAS matrix plot of significant SNPs associated with dietary habits. a,b, Phenome-wide associations between complex traits and loci associated with dietary habits. Complex diseases (**a**) and laboratory measurements (**b**) are plotted on the *x* axis. Colours indicate the *P* value for each genotype–phenotype test association; these were calculated by either logistic or linear regression adjusted by age and sex. The lollipop chart shows the number of association signals that met the significance threshold ($P<0.05/103/9$). Abbreviations and clinical phenotypes are described in Supplementary Table 10.

1) = 0.007; $r_g = 0.64$ (pairwise genetic corrections, r_g; Supplementary Table 2). While several genetic loci have sex-specific effects, the overall genetic components would influence the dietary habits between sexes similarly.

Genetic correlation among dietary habits. To identify genetic overlap, r_g values were estimated for the 13 dietary habits (in total 78 trait pairs) using cross-trait LDSC[35] (Fig. 3 and Supplementary Table 10). Of 78 trait pairs, 15 showed significant (false discovery rate (FDR) < 0.01) positive genetic correlations while drinks per week and yoghurt consumption had a negative genetic correlation ($r_g = -0.37$, $P = 6.0\times10^{-4}$). Interestingly, strong positive correlations were observed between foods made from the same ingredients (that is, tofu and natto from soybean ($r_g = 0.80$, $P = 9.5\times10^{-23}$) and cheese and yoghurt from milk ($r_g = 0.66$, $P = 9.3\times10^{-14}$)).

PheWAS of significant SNPs with dietary habits. Considering that dietary habits have epidemiological links to a variety of diseases, additional investigation was conducted to determine whether dietary habits share genetic architecture with human complex diseases. We carried out PheWAS to determine whether the SNPs associated with dietary habits were also associated with diseases or laboratory clinical measurements included in BBJ (103 clinical phenotypes for 143,658 Japanese individuals; Fig. 4 and Supplementary Tables 11 and 12)[24,34,36]. Among the 9 associated SNPs satisfying the study-wide significance with any of the dietary habits, 6 variants were associated with at least one of the 14 diseases or 49 laboratory measurements that satisfied Bonferroni correction ($P<0.05/9/103 = 5.39\times10^{-5}$). Three of these six SNPs were significantly associated with multiple phenotypes, including two previously described pleiotropic SNPs within dietary habits (rs671 at *ALDH2* and rs1260326 at *GCKR*).

Tissue specificity and biological pathways related to genetics of dietary habits. Translation of the association signals of genome-wide SNPs into upstream functional units contributes to the identification of biological mechanisms of complex human traits[24,37,38]. In this context, we first analysed cell-type specificity of the genetics of dietary habits by partitioning heritability across ten cell types (Fig. 5 and Supplementary Table 13). Two significant heritability enrichments were observed after multiple testing of cell type groups—liver with coffee consumption ($P = 0.0012$) and the central nervous system with yoghurt consumption ($P = 0.0007$). Although the PheWAS results showed no association between coffee consumption-related SNPs and liver-related diseases, the overall genetic architecture suggests that coffee consumption is related to the liver. This should be informative to further understanding of the biological effects of coffee consumption on chronic liver diseases such as hepatocellular carcinoma and liver cirrhosis, which have been shown to be negatively associated in epidemiological studies[39,40]. Previous studies reported that yoghurt consumers tended to have higher levels of education and healthier behaviours[41,42]; the findings of the present study partly agree that yoghurt consumption may be driven by behavioural motivation.

Pathway enrichment of the GWAS association signals of dietary habits was also evaluated in 1,077 pathways (Supplementary Fig. 12

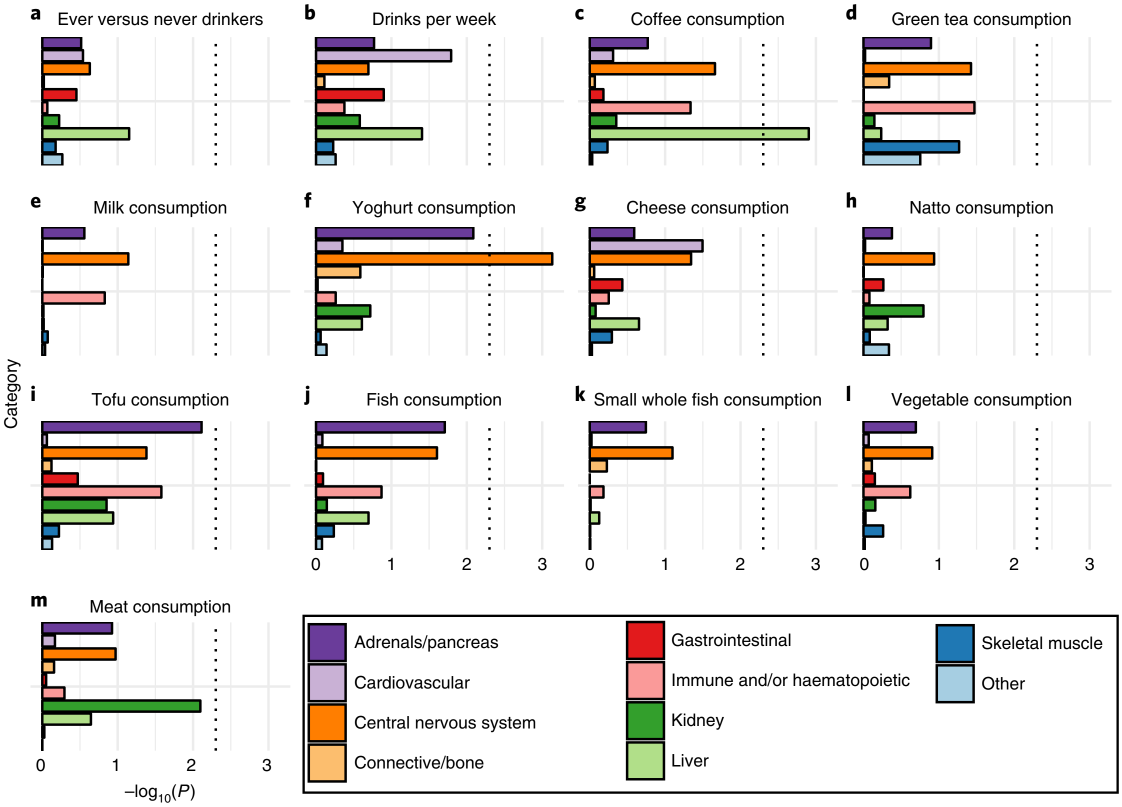

Fig. 5 | Cell-type-specific enrichment of dietary habits. Trait-relevant, cell-type-specific enrichment of the four dietary habits indicated by LDSC analysis. Vertical dashed lines indicate the significance threshold at $P=5.0\times10^{-3}$ after Bonferroni correction (0.05 per ten cell groups).

and Supplementary Table 14). None of the pathways indicated significant associations after application of multiple corrections ($P<0.05/1{,}077/13=3.5\times10^{-6}$). Notably, the most significantly enriched pathway was drug metabolism of cytochrome p450 (CYP) ($P=2.5\times10^{-5}$) for coffee consumption. CYP is an enzyme that is mainly expressed in the liver. Enrichment of this pathway was consistent with the results obtained in GWAS and tissue-specificity analysis. These findings support the role of gene *CYP2* in coffee consumption and caffeine metabolism, as previously reported[43–45].

Discussion

In this study, the genetic architecture of habitual consumption of foods and beverages in the Japanese population was investigated by conducting a large-scale, biobank-driven GWAS. This large sample size in a non-European population ($n_{\max}=165{,}084$) successfully identified ten new study-wide significant associations with dietary habits, including the previously well-studied behaviours of alcohol and coffee consumption. Genome-wide genetic correlation analysis of dietary habits showed that processed foods made from similar ingredients had positive genetic correlations (for example, between natto and tofu made from soybean). PheWAS results revealed potential associations between dietary habits and complex phenotypes.

ALDH2 and *ADH1B* are widely known as alcohol-related genes, and both are involved in alcohol metabolism. Although these genes, which encode alcohol dehydrogenase (ADH) and aldehyde dehydrogenase (ALDH), respectively, are often reported to have potential associations with alcohol metabolism by candidate gene approaches, no GWAS has yet identified any ADH- and ALDH-related genes other than *ALDH2* and *ADH1B*. In the present study, two new signals of genes *ALDH1B1* and *ALDH1A1* were found to be associated with ever versus never drinkers. The top SNP (rs8187929) found in the *ALDH1A1* region is a missense variant (p.Ile177Phe), which is common in 1KGP EAS but absent in the EUR population, highlighting the possibility of genetic heterogeneity between Europeans and East Asians. A previous study reported that this variant has been under natural selection pressure, in addition to the well-known functional SNPs rs1229984 in *ADH1B* and rs671 in *ALDH2* (ref. [2,3,46]). In *ALDH1B1*, the lead SNP in Japanese (rs3043) was located within the 3′-untranslated region (UTR) region while previous genetic associations reported in Europeans mostly indicated the missense variant of rs2228093 (p.Ala86Val) with non-drinking (or alcohol hypersensitivity), as well as higher total alcohol consumption[47–49]. These two SNPs were in low linkage disequilibrium ($r^2=0.29$ in EAS), and rs2228093 also indicated a significant association with ever versus never drinkers in our study ($P=1.3\times10^{-18}$). When we conditioned the association with rs3043, rs2228093 still demonstrated association just below the genome-wide significance level ($P=3.0\times10^{-7}$). These results suggested the existence of independent SNPs associated with alcohol consumption in *ALDH1B1*.

In this study, the largest numbers of susceptibility loci to date have been identified for coffee consumption ($n=8$). Although caffeine is an antagonist of the adenosine A1/A2a receptors ADORA1 and ADORA2A, no GWAS has reported an association of this gene with coffee consumption. Several studies of candidate genes/variants have reported the nominal association between the synonymous variant rs5751876 of *ADORA2A* and caffeine consumption[50]. In the present study, an association with *ADORA2A-AS1* adenosine A2a receptor anti-sense RNA1 was found. These two SNPs are in moderate-to-strong linkage disequilibrium in EAS, EUR and admixed African populations from 1KGP ($r^2=0.64$, 0.89 and 0.91, respectively), but the association at the genome-wide level was a new finding of the current study. We note that, after submission of our study, Zhong et al.[11] reported that rs2330783 in *SPECC1L-*

ADORA2A was associated with coffee consumption. Because the SNP was monomorphic in EAS from 1KGP, we were not able to test this association in our study.

There were several potential limitations to this study. First, data on the consumption of foods and beverages were obtained by FFQ, which consisted of four quantitative consumption dosage categories. The FFQ is among the most widely used dietary consumption surveys. Potential bias in the consumption dosage records inducted by the FFQ would not be directly linked to specific genetic markers, although there is no consensus on the potential effects of the FFQ on the genetic analysis of dietary habits. Second, some of our results may include potential confounders such as alcohol consumption and unmeasured environmental factors. Third, BBJ is a hospital-based cohort that includes individuals affected with any of the target diseases. Although analytical efforts to avoid such bias were applied, it would be difficult to avoid potential biases due to unmeasured factors. There is a possibility that differences in dietary habits between prediagnosed and diseased individuals could have affected the results, and that the dietary habit-associated loci identified were enriched for disease risk-associated loci per se. Hence, further studies of healthy populations may be needed for clarification. The fourth limitation is the lack of a replication study. Since the BBJ cohort is one of the largest genetic cohorts in Japan, it is difficult currently to obtain replication datasets of similar sample size and deep coverage of phenotypes to obtain sufficient statistical power. As a future step, further integration of GWAS and replication studies on dietary habits in a wider spectrum of phenotypes in a trans-ethnic manner, as well as cross-validation with epidemiological observations, is warranted to robustly elucidate the interactive landscape of dietary habits on human life and society.

In conclusion, this large-scale dietary habits GWAS on a Japanese population identified a total of ten new study-wide trait–locus associations, as well as genetic links among dietary habits and complex human diseases. The study results reveal new genetic architectures of dietary habits, which should contribute to further understanding of the link between biological factors and human diseases.

Methods

Study cohort. Baseline data, including sex, age, dietary habits and drinking history, were obtained from the BBJ Project[19], which was established in 2003 and has a registry of approximately 200,000 individuals[23,24,51]. All participants were diagnosed with any of 47 common diseases. In this study, we enroled individuals aged 20–89 years of age and diagnosed as having any of 45 diseases, including at least 100 affected individuals (Supplementary Tables 1 and 2). No statistical methods were used to predetermine sample sizes, but these are similar to those reported in previous publications[23,24,51]. All subjects provided written informed consent as approved by the ethical committee of RIKEN Yokohama Institute and the Institute of Medical Science, the University of Tokyo. This study was approved by the Ethical Committee of Osaka University Graduate School of Medicine.

Phenotypes. Our studies assessed all dietary habit phenotypes obtained by interviewing the participants in BBJ about the frequency of consumption of foods, beverages and alcohol using the standardized questionnaire. The items analysed were grouped into one of three categories: alcohol, other beverages and foods. Foods included both raw and processed foods: yoghurt, cheese, natto, tofu, fish, small whole fish, vegetables and meat. Beverages included coffee, green tea and milk. Green tea is the most widely consumed type of tea in Japan that does not involve the withering and oxidation of tea leaves, while other types of traditional tea could be included according to the questionnaire. For each food and beverage, the participants were asked to clarify the frequency of consumption on a four-point scale, where 1 = almost every day, 2 = 3–4 d per week, 3 = 1–2 d per week and 4 = rarely, based on normal frequency of consumption (see details in Supplementary note).

Based on their replies, the corresponding values were used for analysis as follows:

$$y(\text{outcome}) = \begin{cases} 7, \text{if individual consumes the item almost every day} \\ 3.5, \text{if individual consumes the item } \sim 3-4 \text{ times per week} \\ 1.5, \text{if individual consumes the item } \sim 1-2 \text{ times per week} \\ 0, \text{if individual consumes the item rarely} \end{cases}$$

In regard to alcohol consumption, two defined phenotypes were analysed: (1) ever versus never drinker and (2) number of drinks per week. Participants were asked about the type, volume (ml) and frequency of alcoholic drinks consumed (per week). The number of drinks consumed per week was calculated by multiplying the percentage of alcohol in the drinks by the volume and frequency. A pairwise correlation matrix of the phenotypes is shown in Supplementary Fig. 13.

Genotyping and imputation. Genotyping of ~200,000 individuals was performed using the Illumina HumanOmniExpress Exome-8 platform (v.1.0/v1.2) or both the Illumina HumanOmniExpress-12v1 and HumanExome-12 (v.1.0/v1.1) platforms, as described elsewhere[23,24,51] (see the detailed quality-control steps in Supplementary Fig. 1). Briefly, we excluded (1) individuals with sample call rate <0.98, (2) closely related individuals estimated by identity-by-state analysis through visual confirmation, (3) individuals with genotypic and phenotypic sex mismatch and (4) outliers from the East Asian cluster identified by PCA with three reference populations in the International HapMap Project: African, European American and East Asian. Through quality control of SNPs, those with call rate <0.99 or MAF <0.005 were excluded. In addition, SNPs with large allele frequencies across the SNP microarray platforms were excluded from the analysis. Genotyping data of autosomal chromosomes were phased using MACH (http://csg.sph.umich.edu//abecasis/MaCH/) and imputed using minimac (v.0.1.1) (https://genome.sph.umich.edu/wiki/Minimac), with a reference panel constructed from 275 unrelated EAS haplotypes from 1KGP phase1v3. Insertions, deletions and SNPs with MAF <0.01 or HWE P values (P_{HWE}) $< 1.0 \times 10^{-6}$ were excluded from the reference panel before imputation. Phasing and imputation of genotypes of the X chromosome were conducted separately for males and females.

Genome-wide association analyses. To identify associations among SNPs on autosomal chromosomes, the BOLT-LMM Bayesian mixed-model association method was performed, which implements linear mixed models (LMMs) with the covariates of age, age[2] and sex. Considering that BBJ is a hospital-based patient cohort, the status of the 45 target diseases affecting any of the individuals in GWAS was additionally included as a covariate (Supplementary Table 2). BOLT-LMM adjusts for cryptic population structures and relatedness among individuals: principal components were not included as covariates. We analysed SNPs on the X chromosome in the same manner separately for males and females using ProbABEL (https://github.com/GenABEL-Project/ProbABEL), including the same covariates as described above with the additional top ten principal components. SNPs with low imputation quality ($R_{\mathrm{sq}} < 0.7$) and MAF < 0.01 were excluded before conducting a fixed-effect meta-analysis using METAL (https://genome.sph.umich.edu/wiki/METAL). $P < 1.0 \times 10^{-300}$ was calculated with the Rmpfr package of R. LDSC[25] intercepts were used for genomic-control correction of the test statistics.

Pleiotropic association of *ALDH2* (rs671) and dietary habits. Due to significant deviations from HWE ($P_{\mathrm{HWE}} < 1.0 \times 10^{-6}$), the well-known alcohol-associated SNP of rs671 at *ALDH2* was initially filtered out from the SNP genotype data with the use of a SNP quality-control filter. To clarify whether an experimental error had occurred, we assessed the accuracy of rs671 genotypes by comparson with our in-house whole-genome sequencing (WGS) data ($n = 2{,}798$). Concordance of genotyping between the SNP microarray and the WGS data was 100%, indicating that the observed deviation of rs671 from HWE was not due to genotyping error, but rather to heterogeneity in allele frequency spectra among the domestic regions of Japan, caused by strong selection pressure[2]. Thus, to identify associations among the phenotype and rs671, genotypes were directly extracted from SNP genotyping data rather than from imputed dosage data. Then, linear and logistic regression analyses were performed in R with either the lm or glm function for binominal or quantitative traits, respectively. Age, age[2], sex and the top ten principal components and the disease affection status were used as covariates.

Look-up results from UKBB. Variant association results from UKBB were obtained from the Gene ATLAS database (http://geneatlas.roslin.ed.ac.uk/) by searching for the following key words: alcohol, coffee, tea, milk, yoghurt, natto, tofu and fish. The available traits in UKBB consisted of alcohol-drinker status, alcohol intake frequency, coffee intake, tea intake, non-oily fish intake and oily fish intake. Those SNPs that showed associations at the genome-wide significance level in the current study were selected ($P < 5.0 \times 10^{-8}$). We looked up the results for each of the SNPs with each corresponding trait from UKBB.

Investigation of SNPs related to taste ability. We assessed the variants of rs10246939, rs713598 and rs1726866 of *TSR2R38* related to bitter taste sensitivity, as well as those in strong linkage disequilibrium with them ($r^2 > 0.995$ each other in 1KGP EAS).

Functional annotation of identified SNPs. The pairwise measure of linkage disequilibrium (r^2) was calculated using PLINK (v.1.90b) (https://www.cog-genomics.org/plink/1.9/) with the EAS population. ANNOVAR (2016Feb01; http://annovar.openbioinformatics.org/en/latest/) software was used for the lead SNPs (and SNPs in linkage disequilibrium; $r^2 > 0.8$). *Cis*-eQTL effects of these SNPs were assessed using Genotype Tissue Expression Portal (GTEx) release v.7 (ref. [31])

(https://www.gtexportal.org/). If the lead SNP in GWAS and eQTL was in marked linkage disequilibrium ($r^2 > 0.8$) in both EAS and EUR populations, the SNP was considered to be overlapping. Genes with a significant eQTL in all 48 tissues (FDR < 0.01) were selected.

LDSC analysis. Linkage disequilibrium score regression[25] (https://github.com/bulik/ldsc/) was used to assess the level of inflation of GWAS after exclusion of polygenic effects, cross-trait genetic correlations and tissue-specific partitioned heritability enrichment. We adopted the HapMap3 SNPs, excluding the HLA region where the linkage disequilibrium pattern was complex[31], using the –merge-alleles flag of LDSC with precomputed linkage disequilibrium scores from 1KGP EAS downloaded from the LDSC software website (https://data.broadinstitute.org/alkesgroup/LDSCORE/). Cross-trait genetic correlations analysis among the current GWAS on dietary habits was performed (Supplementary Table 9). *P* values were corrected by FDR (Benjamini–Hochberg method) for the 78 trait pairs, and for pairs with FDR < 0.01 that were considered to be significantly correlated. Cross-sex genetic correlation analysis was also performed for each item, and differences were calculated by Wald test against $r_g = 1$.

Tissue-specific enrichment analysis was conducted as described previously[24,38]. Briefly, stratified LDSC was performed to partition heritability into ten cell type group-specific annotations, consisting of the adrenals/pancreas, central nervous system, cardiovascular, connective/bone, gastrointestinal, immune/haematopoietic, kidney, liver, skeletal muscle and other, which were obtained from the software website https://data.broadinstitute.org/alkesgroup/LDSCORE. As a reference for partitioning of heritability, linkage disequilibrium scores were generated using 1KGP phase3 (v.5) EAS haplotypes.

PheWAS. Logistic or linear regression analysis was used to estimate associations between dietary-related SNPs ($n = 10$) and BBJ phenotypes ($n = 103$; 45 diseases and 58 clinical measurements). The glm and lm functions implemented in R were used for the analyses. The significance threshold was set as $P < 4.9 \times 10^{-5}$ (0.05/9/103). Details of phenotype definitions are described elsewhere[25,36].

Pathway enrichment analysis. Gene prioritization and pathway analysis was carried out using the pathway-scoring algorithm PASCAL[37] (https://www2.unil.ch/cbg/index.php?title=Pascal) using GWAS summary statistics with default settings. We used the 275 independent EAS samples from 1KGP as a reference panel, with removal of low-frequency variants (MAF < 0.01). In total, 1,077 pathways containing KEGG (http://www.genome.jp/kegg/), REACTOME (https://reactome.org/) and BIOCARTA pathway-gene data provided in the software packages were analysed.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

GWAS summary statistics of the 13 dietary habits investigated are publicly available at the National Bioscience Database Centre (NBDC) Human Database (Research ID: hum0014) as open data with no access restrictions. GWAS genotype data were deposited at the NBDC Human Database (Research ID: hum0014).

Received: 8 October 2018; Accepted: 3 December 2019;
Published online: 20 January 2020

References

1. Johnson, K. E. & Voight, B. F. Patterns of shared signatures of recent positive selection across human populations. *Nat. Ecol. Evol.* **2**, 713–720 (2018).
2. Okada, Y. et al. Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nat. Commun.* **9**, 1631 (2018).
3. Okada, Y. eLD: entropy-based linkage disequilibrium index between multiallelic sites. *Hum. Genome Var.* **5**, 29 (2018).
4. Baik, I., Cho, N. H., Kim, S. H., Han, B.-G. & Shin, C. Genome-wide association studies identify genetic loci related to alcohol consumption in Korean men. *Am. J. Clin. Nutr.* **93**, 809–816 (2011).
5. Takeuchi, F. et al. Confirmation of *ALDH2* as a major locus of drinking behavior and of its variants regulating multiple metabolic phenotypes in a Japanese population. *Circ. J.* **75**, 911–918 (2011).
6. 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. USA* **113**, 14372–14377 (2016).
7. Yang, X. et al. Common variants at 12q24 are associated with drinking behavior in Han Chinese. *Am. J. Clin. Nutr.* **97**, 545–551 (2013).
8. Jorgenson, E. et al. Genetic contributors to variation in alcohol consumption vary by race/ethnicity in a large multi-ethnic genome-wide association study. *Mol. Psychiatry* **22**, 1359–1367 (2017).
9. Clarke, T. K. et al. Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK Biobank ($N = 112117$). *Mol. Psychiatry* **22**, 1376–1384 (2017).
10. Liu, M. et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).
11. Zhong, V. W. et al. A genome-wide association study of bitter and sweet beverage consumption. *Hum. Mol. Genet.* **28**, 2449–2457 (2019).
12. Sulem, P. et al. Sequence variants at *CYP1A1-CYP1A2* and *AHR* associate with coffee consumption. *Hum. Mol. Genet* **20**, 2071–2077 (2011).
13. Amin, N. et al. Genome-wide association analysis of coffee drinking suggests association with *CYP1A1*/*CYP1A2* and *NRCAM*. *Mol. Psychiatry* **17**, 1116–1129 (2012).
14. Coffee and Caffeine Genetics Consortiumet al. Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption. *Mol. Psychiatry* **20**, 647–656 (2015).
15. Pirastu, N. et al. Non-additive genome-wide association scan reveals a new gene associated with habitual coffee consumption. *Sci. Rep.* **6**, 31590 (2016).
16. Nakagawa-Senda, H. et al. A genome-wide association study in the japanese population identifies the 12q24 locus for habitual coffee consumption: the J-MICC study. *Sci. Rep.* **8**, 1493 (2018).
17. Mozaffarian, D. et al. Genome-wide association meta-analysis of fish and EPA+DHA consumption in 17 US and European cohorts. *PLoS One* **12**, e0186456 (2017).
18. Jiang, L., Penney, K. L., Giovannucci, E., Kraft, P. & Wilson, K. M. A genome-wide association study of energy intake and expenditure. *PLoS One* **13**, e0201555 (2018).
19. Nagai, A. et al. Overview of the BioBank Japan project: study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).
20. Nakachi, K., Matsuyama, S., Miyake, S., Suganuma, M. & Imai, K. Preventive effects of drinking green tea on cancer and cardiovascular disease: epidemiological evidence for multiple targeting prevention. *BioFactors* **13**, 49–54 (2000).
21. Uemura, H. et al. Inverse association between soy food consumption, especially fermented soy products intake and soy isoflavone, and arterial stiffness in Japanese men. *Sci. Rep.* **8**, 9667 (2018).
22. Tsugane, S. & Sawada, N. The JPHC study: design and some findings on the typical Japanese diet. *Jpn. J. Clin. Oncol.* **44**, 777–782 (2014).
23. Akiyama, M. et al. Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat. Genet.* **49**, 1458–1467 (2017).
24. Kanai, M. et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).
25. Bulik-Sullivan, B. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
26. 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**, e1002033 (2011).
27. Cornelis, M. C. et al. Genome-wide association study of caffeine metabolites provides new insights to caffeine metabolism and dietary caffeine-consumption behavior. *Hum. Mol. Genet.* **25**, ddw334 (2016).
28. Yin, G. et al. *ALDH2* polymorphism is associated with fasting blood glucose through alcohol consumption in Japanese men. *Nagoya J. Med. Sci.* **78**, 183–193 (2016).
29. Gelernter, J. et al. Genome-wide association study of alcohol dependence:significant findings in African- and European-Americans including novel risk loci. *Mol. Psychiatry* **19**, 41–49 (2014).
30. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat. Genet.* **50**, 1593–1599 (2018).
31. Lonsdale, J. et al. The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
32. Kim, Y. K. et al. Evaluation of pleiotropic effects among common genetic loci identified for cardio-metabolic traits in a Korean population. *Cardiovasc. Diabetol.* **15**, 20 (2016).
33. Sakaue, S. et al. Functional variants in *ADH1B* and *ALDH2* are non-additively associated with all-cause mortality in Japanese population. *Eur. J. Hum. Genet.* https://doi.org/10.1038/s41431-019-0518-y (2019).
34. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Researchet al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
35. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
36. Hirata, J. et al. Genetic and phenotypic landscape of the major histocompatibilty complex region in the Japanese population. *Nat. Genet.* **51**, 470–480 (2019).
37. Lamparter, D. et al. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Comput. Biol.* **12**, e1004714 (2016).
38. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

39. Kennedy, O. J. et al. Systematic review with meta-analysis: coffee consumption and the risk of cirrhosis. *Aliment. Pharmacol. Ther.* **47**, 562–5 (2016).
40. Inoue, M., Yoshimi, I., Sobue, T. & Tsugane, S. Influence of coffee drinking on subsequent risk of hepatocellular carcinoma: a prospective study in Japan. *J. Natl Cancer Inst.* **97**, 293–300 (2005).
41. Panahi, S., Fernandez, M. A., Marette, A. & Tremblay, A. Yogurt, diet quality and lifestyle factors. *Eur. J. Clin. Nutr.* **71**, 573–579 (2017).
42. D'Addezio, L., Mistura, L., Sette, S. & Turrini, A. Sociodemographic and lifestyle characteristics of yogurt consumers in Italy: results from the INRAN-SCAI 2005-06 survey. *Med. J. Nutr. Metab.* **8**, 119–129 (2015).
43. Kot, M. & Daniel, W. A. Effect of cytochrome P450 (CYP) inducers on caffeine metabolism in the rat. *Pharmacol. Rep.* **59**, 296–305 (2007).
44. Zanger, U. M. & Schwab, M. Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol. Ther.* **138**, 103–141 (2013).
45. Berthou, F. et al. Evidence for the involvement of several cytochromes P-450 in the first steps of caffeine metabolism by human liver microsomes. *Drug Metab. Dispos.* **19**, 561–567 (1991).
46. Wang, L.-X., Wen, S., Wang, C.-C., Zhou, B. & Li, H. Molecular adaption of alcohol metabolism to agriculture in East Asia. *Quat. Int.* **426**, 187–194 (2016).
47. Way, M. J., Ali, M. A., McQuillin, A. & Morgan, M. Y. Genetic variants in *ALDH1B1* and alcohol dependence risk in a British and Irish population: a bioinformatic and genetic study. *PLoS One* **12**, e0177009 (2017).
48. Linneberg, A. et al. Genetic determinants of both ethanol and acetaldehyde metabolism influence alcohol hypersensitivity and drinking behaviour among Scandinavians. *Clin. Exp. Allergy* **40**, 123–130 (2009).
49. Husemoen, L. L. N. et al. The association of *ADH* and *ALDH* gene variants with alcohol drinking habits and cardiovascular disease risk factors. *Alcohol. Clin. Exp. Res.* **32**, 1984–1991 (2008).
50. Cornelis, M. C., El-Sohemy, A. & Campos, H. Genetic polymorphism of the adenosine A2A receptor is associated with habitual caffeine consumption. *Am. J. Clin. Nutr.* **86**, 240–244 (2007).
51. Matoba, N. et al. GWAS of smoking behaviour in 165,436 Japanese people reveals seven new loci and shared genetic architecture. *Nat. Hum. Behav.* **3**, 471–477 (2019).

Acknowledgements

We are grateful to all participants enrolled in BBJ. We thank all the clinicians and organizations that contributed to the collection of samples and clinical information. This research was supported by the Tailor-Made Medical Treatment Programme (BBJ) of the Ministry of Education, Culture, Sports, Science and Technology and the Japan Agency for Medical Research and Development (AMED; grant nos. JP17km0305002, JP19km0405201 and JP19km045208), and by the Strategic Research Programme for Brain Sciences of AMED (no. JP19dm0107097). Y.O. was supported by the Japan Society for the Promotion of Science, KAKENHI (nos. 15H05911 and 19H01021), AMED (nos. JP19gm6010001, JP19ek0410041, JP19ek0109413 and JP19km0405211), Takeda Science Foundation and the Bioinformatics Initiative of Osaka University Graduate School of Medicine, Osaka University and Osaka University Centre of Medical Data Science, Advanced Clinical Epidemiology Investigator's Research Project. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

N.M., M.A., Y.K. and Y.O. contributed to study concept and design. M.H., K.M., Y. Murakami and M. Kubo collected and managed BBJ samples. Y. Momozawa and M. Kubo performed genotyping. N.M., M.A., K.I., M. Kanai and A.T. performed statistical analysis. S.I., M.I. and N.I. contributed to data acquisition. N.M., Y.K. and Y.O. wrote the manuscript. All authors reviewed and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-019-0805-1.

Correspondence and requests for materials should be addressed to Y.K. or Y.O.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Editor recognition statement Primary handling editor: Stavroula Kousta

nature research

Corresponding author(s): Yoichiro Kamatani
Yukinori Okada

Last updated by author(s): Nov 27, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We did not use any software for Data collection
Data analysis	We used publicly available software for the analyses; We used MACH, Minimac, ANNOVAR, BOLT-LMM, ProbABEL, LDSC, PASCAL and R. The software is described in the Methods section of the manuscript.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The National Bioscience Database Center (NBDC) Human Database (Research ID: hum0014) is publicly available without access restrictions.
GWAS genotype data from the participants was deposited at the NBDC Human Database (Research ID: hum0014).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size calculation was not carried out, but we used maximum samples in BioBank Japan, whose phenotype of each dietary habit was available.
Data exclusions	We excluded samples and SNPs based on the standard quality control procedure in GWAS as follows: Sample QC: call rate < 0.98, closely related individuals evaluated by identity-by- state or outlier of East Asian cluster identified by principal component analysis of all BBJ samples and the three reference populations: Africans, European Americans, and East Asians in the International HapMap Project. SNP QC before imputation: minor allele frequency (MAF) < 0.005 or SNP call rate < 0.99. For GWAS, SNPs met the imputation quality score (Rsq) ≥ 0.7 and MAF ≥ 0.01 were used. Detailed information on quality controls were described in method as well as Supplementary Fig.1.
Replication	We looked up UK BioBank association for identified loci in this study, and also checked our results for previously reported loci. Results were described in Table 1, Supplementary table 4 and 6, and Supplementary fig. 8. We did not split our sample into discovery and replication set.
Randomization	n/a
Blinding	n/a

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
□	☒	Human research participants
☒	□	Clinical data

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Our analyses were based on a maximum of 165,084 Japanese individuals from the BioBank Japan Project (total number was 165,738). Detailed information for each phenotype is shown in Supplementary Table 1 and 2.
Recruitment	We enrolled participants in BioBank Japan who have been recruited between June 2003 and March 2008 from 66 hospitals located throughout Japan. Participants have been diagnosed with any of 47 target diseases. The affected status for each disease was used as a covariate in association tests.
Ethics oversight	All the subjects provided written informed consent as approved by the ethical committee of RIKEN Yokohama Institute and the Institute of Medical Science, the University of Tokyo. This study was approved by the ethical committee of Osaka University Graduate School of Medicine.

Note that full information on the approval of the study protocol must also be provided in the manuscript.