Human Molecular Genetics, 2020, Vol. 29, No. 23 3830–3844

doi: 10.1093/hmg/ddaa257
Advance Access Publication Date: 5 December 2020
Association Studies Article

ASSOCIATION STUDIES ARTICLE

Urinary metabolite quantitative trait loci in children and their interaction with dietary factors

Beatriz Calvo-Serra[1,2,3], Léa Maitre[1,2,3], Chung-Ho E. Lau[4], Alexandros P. Siskos[4,5], Kristine B. Gützkow[6], Sandra Andrušaitytė[7], Maribel Casas[1,2,3], Solène Cadiou[8], Leda Chatzi[9], Juan R. González[1,2,3], Regina Grazuleviciene[7], Rosemary McEachan[10], Rémy Slama[8], Marina Vafeiadi[11], John Wright[10], Murieann Coen[4,12], Martine Vrijheid[1,2,3], Hector C. Keun[4,5], Geòrgia Escaramís[13,3,‡] and Mariona Bustamante[1,2,3,‡,*,†]

[1]ISGlobal, Barcelona 08003, Spain, [2]Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain, [3]CIBER Epidemiología y Salud Pública (CIBERESP), Madrid 28029, Spain, [4]Division of Systems Medicine, Department of Metabolism, Digestion and Reproduction, Imperial College London, London SW7 2AZ, UK, [5]Cancer Metabolism and Systems Toxicology Group, Division of Cancer, Department of Surgery and Cancer, Imperial College London, Hammersmith Hospital Campus, London W12 0NN, UK, [6]Department of Environmental Health, Norwegian Institute of Public Health, Oslo 0213, Norway, [7]Department of Environmental Science, Vytautas Magnus University, Kaunas 44248, Lithuania, [8]Team of Environmental Epidemiology, IAB, Institute for Advanced Biosciences, Inserm, CNRS, CHU-Grenoble-Alpes, University Grenoble-Alpes, Grenoble 38000, France, [9]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles 90033, USA, [10]Bradford Institute for Health Research, Bradford BD9 6RJ, UK, [11]Department of Social Medicine, Faculty of Medicine, University of Crete, Heraklion 71003, Greece, [12]Oncology Safety, Clinical Pharmacology and Safety Sciences, R&D, AstraZeneca, Cambridge CB2 0RE, UK and [13]Departament de Biomedicina, Institut de Neurociències, Universitat de Barcelona (UB), Barcelona 08036, Spain

*To whom correspondence should be addressed. Tel: +34 933160167; Fax: +34 932147320; Email: mariona.bustamante@isglobal.org

Abstract

Human metabolism is influenced by genetic and environmental factors. Previous studies have identified over 23 loci associated with more than 26 urine metabolites levels in adults, which are known as urinary metabolite quantitative trait loci (metabQTLs). The aim of the present study is the identification for the first time of urinary metabQTLs in children and their interaction with dietary patterns. Association between genome-wide genotyping data and 44 urine metabolite levels measured by proton nuclear magnetic resonance spectroscopy was tested in 996 children from the Human Early Life Exposome project. Twelve statistically significant urine metabQTLs were identified, involving 11 unique loci and 10 different metabolites. Comparison with previous findings in adults revealed that six metabQTLs were already known, and one had been described in serum and three were involved the same locus as other reported metabQTLs but had different urinary metabolites. The remaining two metabQTLs represent novel urine metabolite-locus associations, which are reported for the

[†]Mariona Bustamante, http://orcid.org/0000-0003-0127-2860
[‡]These authors contributed equally to this work.
Received: August 18, 2020. **Revised:** November 26, 2020. **Accepted:** November 30, 2020

first time in this study [single nucleotide polymorphism (SNP) rs12575496 for taurine, and the missense SNP rs2274870 for 3-hydroxyisobutyrate]. Moreover, it was found that urinary taurine levels were affected by the combined action of genetic variation and dietary patterns of meat intake as well as by the interaction of this SNP with beverage intake dietary patterns. Overall, we identified 12 urinary metabQTLs in children, including two novel associations. While a substantial part of the identified loci affected urinary metabolite levels both in children and in adults, the metabQTL for taurine seemed to be specific to children and interacted with dietary patterns.

Introduction

Metabolite levels are considered a complex trait since their variation in the human body is influenced by various factors, specifically environmental conditions (nutrition, exposure to xenobiotics, smoking and alcohol consumption, physical activity, etc.) and genetic variants (1,2,3).

The study of inborn errors of metabolism has allowed the identification and functional characterization of many enzymes and other proteins involved in human metabolism (4,5). Recent technological advances in metabolomics and genetics have allowed to systematically explore the less conspicuous influences of more common and less deleterious genetic variants on human metabolism, known as metabolite quantitative trait loci (metabQTL), by conducting genome-wide association studies (GWAS) (6,7,8,9,10). In fact, these studies have suggested that inborn errors of metabolism are only extreme cases of a wide spectrum of genetic variation in human metabolism (1,2,3). The study of metabQTLs might help to understand how genes control metabolic pathways in the human body (1) and facilitate causal inference in metabolite–disease associations, by applying Mendelian randomization analyses (11). Furthermore, it can shed light on the interplay between genetic variation and dietary patterns and its effect on the metabolite levels and disease.

During recent years, more than 150 loci have been linked to serum metabolite levels (6,7,9) and 23 loci to more than 26 urine metabolite levels in adults (8,9,10). Nevertheless, to the best of our knowledge, there is very little information on whether these metabQTLs also affect the metabolism in early life. Indeed, in other traits, such as body mass index (BMI), it has been shown that the effects of some of the loci change across the life course (12–17). Similarly, lactase persistence and non-persistence alleles represent another example of genetic variants in which the effect—that is, lactase activity—is different in infancy than after the weaning phase (18).

In this study, we aimed to identify urinary metabQTLs in children. Using data from the Human Early Life Exposome (HELIX) project involving 996 children of European ancestry (19), we identified 12 urinary metabQTLs. Subsequently, we performed fine-mapping and functional annotation of these metabQTLs, compared them with findings in adults and tested their interaction with dietary factors.

Results

Description of study participants and study workflow

For this study, 996 European ancestry children presenting both urinary metabolite levels and genome-wide genotyping data were selected from the HELIX project (Supplementary Material, Fig. S1) (19). Children were similarly distributed among the cohorts (Table 1), with similar distributions of obese, overweight and male children between cohorts (Supplementary Material, Table S1). Around half of the children were males and mean age was 7.9 years old. A noticeable percentage of the children presented overweight (15.06%) or obesity (5.82%), and more than 90% of them had urine pooled samples available, which provide a better average of the daily metabolic fluctuations, including dietary intake, than spot urine samples (20).

Urinary metabolite concentrations were similarly distributed across the six cohorts (Supplementary Material, Table S2).

Table 1. Descriptive of the study population

Total, *N*	996
Cohort, *N* (%)	
BiB	88 (8.84)
EDEN	136 (13.65)
KANC	197 (19.78)
MoBa	200 (20.08)
Rhea	187 (18.78)
SAB	188 (18.88)
Sex, *N* (%)	
Male	545 (54.72)
Female	451 (45.28)
Age (years), mean (SD)	7.93 (1.57)
Obesity status[a], *N* (%)	
Thin	6 (6.12)
Normal	714 (71.69)
Overweight	150 (15.06)
Obesity	58 (5.82)
NA	13 (1.31)
Mother education level, *N* (%)	
Low	111 (11.14)
Middle	339 (34.04)
High	514 (51.61)
NA	7 (0.70)
Meat intake[b], *N* (%)	
Tertile 1	336 (33.73)
Tertile 2	286 (28.71)
Tertile 3	358 (35.94)
Dairy products intake[b], *N* (%)	
Tertile 1	354 (35.54)
Tertile 2	329 (33.03)
Tertile 3	297 (29.82)
Fish intake[b], *N* (%)	
Tertile 1	398 (39.96)
Tertile 2	293 (29.42)
Tertile 3	289 (29.01)
Beverages consumption[b], *N* (%)	
Tertile 1	395 (39.66)
Tertile 2	350 (35.14)
Tertile 3	235 (23.59)
Urine sample type, *N* (%)	
Night	32 (3.21)
Morning	31 (3.11)
Combined pool	933 (93.67)

[a]z-score of the BMI classification according to the World Health Organization.
[b]For all 4 food groups, 16 individuals had missing information (16.01%).
SAB: INMA Sabadell subcohort.

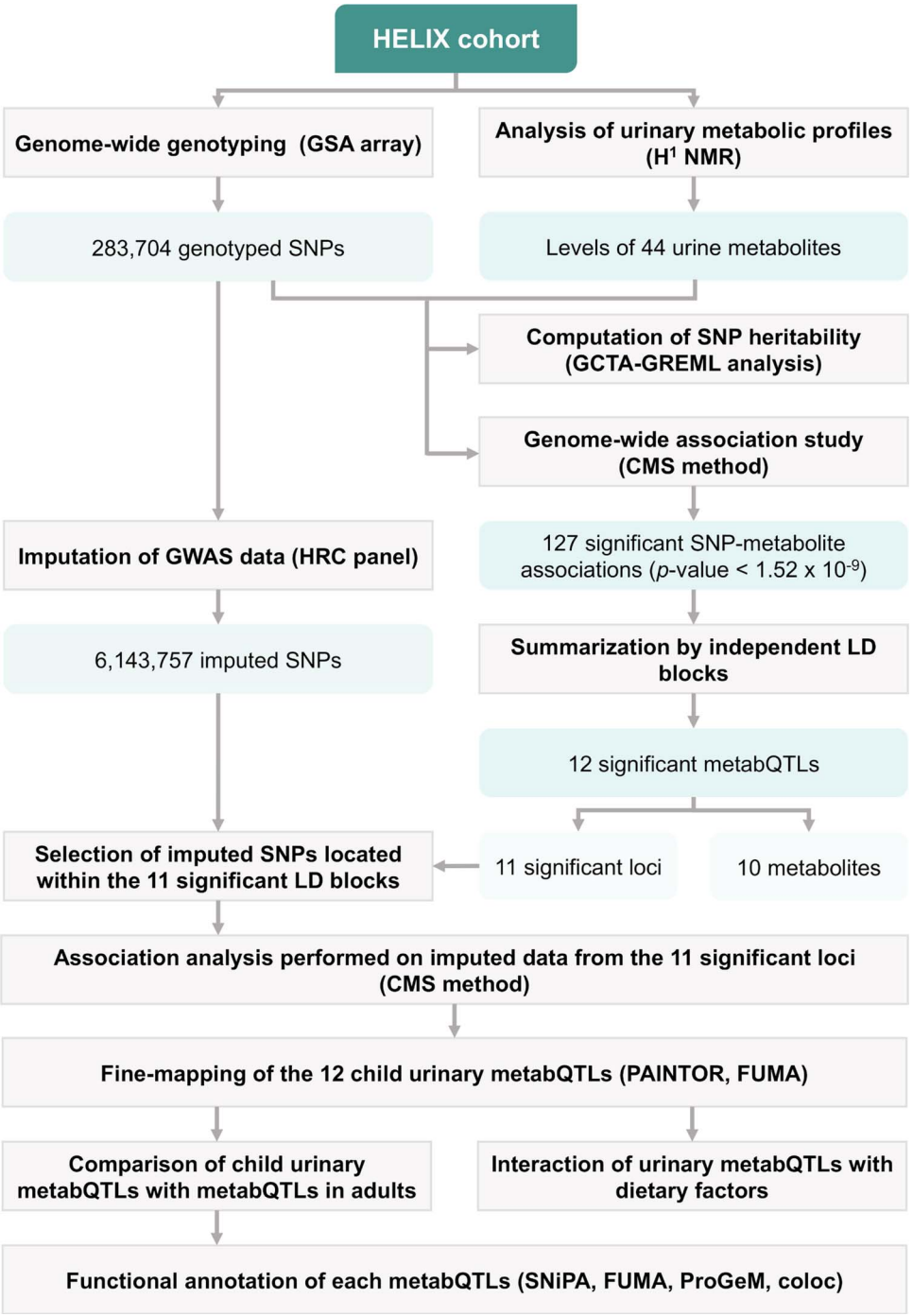

Figure 1. Workflow of the study. Genome-wide genotyped genetic data of 996 HELIX children were analyzed to obtain the SNP heritability of 44 urinary metabolites by GCTA-GREML (62,63) analysis as well as to identify metabQTLs using the CMS method. Results of the genome-wide association study were summarized per independent LD blocks, based on a recombination map (64), in order to obtain the loci displaying a statistically significant association. This led to the identification of 12 metabQTLs, involving 11 different loci and 10 metabolites. Next, we conducted association analysis using HRC imputed genetic data in these 11 loci, followed by fine-mapping by the Bayesian-based method PAINTOR v3.0 (23). Then, comparison of our findings with results from other studies in adults was done. SNiPA (25), FUMA (26) and ProGeM (27) were used to further narrow down the lead SNPs as well as to functionally annotate the lead SNP identified for each metabQTL. Finally, gene-by-environment interactions were tested for candidate metabolites.

Moreover, no clustering of the cohorts can be observed in a principal components analysis plot from the urinary and serum metabolome (Supplementary Material, Fig. S2).

Figure 1 shows the workflow followed in the study, which includes calculation of the single nucleotide polymorphism (SNP)-heritability (h^2) of urinary metabolites, identification of child urinary metabQTLs, fine-mapping and functional annotation, comparison with metabQTLs in adults and testing their interaction with dietary factors.

SNP-h^2 of urinary metabolites in children

SNP-h^2, which is indicative of the proportion of phenotypic variance due to the additive effects of a set of SNPs, was

computed using the Genome-wide complex trait analysis (GCTA) with the Genome-based restricted maximum likelihood (GREML) method for the 44 urinary metabolites considering the 283 704 genotyped SNPs. The values of phenotypic and genotypic variation for each trait can be observed in Supplementary Material, Table S3. For 28 of the metabolites, SNP-h^2 was statistically significant (*P*-value < 0.05) and ranged from 0.36 to 1 with mean [standard deviation (SD)] of 0.65 (0.23). Four metabolites showed a SNP-h^2 of 1 (hippurate, 4-deoxyerythronic acid, leucine and trimethylamine).

Identification of 12 urinary metabQTLs in children

In order to identify urinary metabQTLs, the Covariates for Multiphenotype Studies (CMS) method was used (21). The CMS algorithm uses as covariates collected correlated phenotypes that vary with the outcome but not with the genotype because of shared risk factors; this decreases phenotypic variance independently of the genotype and, hence, increases the statistical power (21). In our study, serum and urinary metabolite levels served as possible covariates for the CMS algorithm. Moreover, models were adjusted for child's age, sex, the first 20 GWAS principal components (PCs) and urine sampling type.

After multiple-testing correction (corrected *P*-value threshold $< 1.52 \times 10^{-9}$), 127 statistically significant SNP–metabolite associations were obtained (Supplementary Material, Table S4). They involved 11 unique loci and 10 different metabolites in a total of 12 metabQTLs. In the statistical models, CMS selected between 24 and 30 covariates (median = 28). On average, 60% of the included covariates were serum metabolite levels. The increase in statistical power achieved by CMS is visible in the quantile–quantile (Q–Q) plot (Supplementary Material, Fig. S3). It allowed the identification of two extra metabQTLs (with glycine and 3-hydroxyisovalerate) which would have not been identified using standard linear regression models (Supplementary Material, Fig. S4).

Fine-mapping of child urinary metabQTLs

To gain insight into the 12 metabQTLs, we ran CMS again in these loci, but now using genetic data imputed with the Haplotype Reference Consortium (HRC) panel (22). The lead SNP in these 12 metabQTLs can be seen in Table 2, and additional significant SNPs in each locus are shown in Supplementary Material, Table S5. The SNP-h^2 of all 10 metabolites displaying a statistically significant association with a locus was significant (nominal *P*-value < 0.05). The percentage of phenotypic variance of each metabolite explained by the lead SNP ranged from 2.49 to 10.26%, except for the lead SNP at the locus associated with trimethylamine, which explained up to 48.23% of the phenotypic variance (Table 2). Fine-mapping with the Probabilistic Annotation INTegratOR (PAINTOR) framework identified minimal credible sets that included between 2 and 13 SNPs (Table 2) (23). The list of these SNPs can be seen in Supplementary Material, Table S6.

Comparison of child urinary metabQTLs with adult metabQTLs

The 12 locus–metabolite associations identified in children were compared with previous studies in adults (Supplementary Material, Table S7). Six out of the 12 metabQTLs had already been reported in urine in adult individuals (Table 2; Supplementary Material, Table S8 in green) (8,24). All of them had consistent direction of the effect. It should be noted that we compared loci for 3-hydroxybutyrate + 3-aminoisobutyrate with loci for 3-aminoisobutyrate alone as we could not isolate the NMR signals of these two metabolites.

The other six urinary metabQTLs were novel SNP–metabolite associations. However, four of the loci involving these six metabQTLs were known loci related to other urinary metabolites or to the same metabolite but in other biological matrices. In particular, the locus associated with urinary 3-hydroxyisovalerate levels in our study had been previously described for 3-hydroxyisovaleric acid-carnitine, an intermediate of 3-hydroxyisovalerate, in the serum of adult subjects (Table 2; Supplementary Material, Table S8 in orange) (7). Furthermore, two loci associated with N-acetylneuraminic acid and one of them associated also with 3-aminoisobutyrate had previously been related to N-acetylaspartate in adults, but in the opposite direction (Table 2; Supplementary Material, Table S8 in yellow) (8). The last two metabQTLs, involving the urine metabolites 3-hydroxyisobutyrate (chromosome 9, lead SNP rs2274870) and taurine (chromosome 11, lead SNP rs12575496), represented novel loci described for the first time in this study (Table 2; Supplementary Material, Table S8 in red). For these novel metabQTLs, we performed fixed-effects based meta-analyses by cohort, and while the metabQTL for 3-hydroxyisobutyrate was consistent across the six cohorts ($I^2 = 0\%$), the metabQTL for taurine showed some heterogeneity ($I^2 = 88\%$), yet all cohorts, except SAB, presented the same direction of the effect (Supplementary Material, Fig. S5, Table S9).

We then examined whether the seven additional metabQTLs described in adults (8), and for which the associated metabolite had been measured in HELIX, were at least nominally significant in HELIX children (Supplementary Material, Table S10). One of them could not be tested as we could not find a proxy SNP, and, among the remaining, four of them were nominally replicated in children. They involved two additional loci for glycine, one for 3-hydroxyisovalerate and one for tyrosine.

Functional annotation of child urinary metabQTLs

In order to increase our insight into the biological mechanisms underlying the metabQTLs, identification of candidate causal genes was carried out by combining information from several tools (Supplementary Material, Table S11): Single Nucleotide Polymorphism Annotator (SNiPA) (25), Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) (26), Prioritization of candidate causal Genes at Molecular QTLs (ProGeM) (27) and Colocalisation Tests of Two Genetic Traits (coloc) (28) for co-localization with kidney cortex expression quantitative trait loci (eQTLs) (GTEX v8), the tissue most likely involved in regulating urinary metabolite levels.

For the six known urine metabQTLs, our strategy identified the same potential top causal genes as described in (8) through manual curation of the literature (Supplementary Material, Table S12). In addition, two of the metabQTLs were also identified as eQTLs linked to their potential causal gene in several tissues (Table 2; Supplementary Material, Tables S11 and S12).

The candidate gene for the locus related to N-acetylneuraminic acid and 3-aminoisobutyrate was *N-Acetyltransferase 8* (*NAT8*) gene, also described by (8). However, our strategy also identified the *Centrosome and basal body associated protein* (*ALMS1*) gene with the same score. The other locus for N-acetylneuraminic acid contained as potential causal gene *Alpha 1–3-N-acetyl-galactosaminyltransferase* and *Alpha 1–3-galactosyltransferase* (*ABO*), as described before (8). The potential

Table 2. Summarized results of the 12 urinary metabQTLs identified using imputed genetic data

Phenotype	Region	Chr.	Pos.	SNP	A1	A2	MAF	Standard beta	Standard P-value	CMS beta	CMS P-value	CMS r^2	N. covs.	% Var. explained	SNP-h^2	N. SNPs credible set/N. SNPs locus	Potential causal gene	eQTL of the potential causal gene
Glycine	610	5	150693906	rs12655635[a]	T	C	0.162	−0.300	5.40E-07	−0.266	3.18E-11	0.563	29	2.49	0.591059*	7/7	SLC36A2	No
Lysine	1601	19	33346341	rs12459052	T	C	0.471	0.395	2.23E-18	0.363	3.76E-18	0.216	29	7.4	0.341748*	2/318	SLC7A9	Yes
Trimethylamine	1072	10	100151305	rs942814	G	A	0.363	−1.025	9.22E-144	−0.972	3.34E-146	0.187	28	48.28	1*	2/228	PYROXD2	No
2-Hydroxyisobutyrate	1254	12	122297738	rs2596144	T	C	0.139	−0.659	1.32E-24	−0.632	3.36E-34	0.373	24	10.01	0.38161*	6/88	*HPD*	No
3-Hydroxybutyrate + 3-aminoisobutyrate	1180	12	351467	rs11613331[a]	G	A	0.466	−0.322	2.38E-12	−0.340	5.13E-17	0.271	30	4.82	0.760986*	5/10	SLC6A13	Yes
3-Hydroxybutyrate + 3-aminoisobutyrate	544	5	35039486	rs37370	C	T	0.100	0.829	3.19E-25	0.782	4.29E-28	0.264	28	10.26	0.760986*	2/62	AGXT2	No
3-Hydroxyisovalerate	389	3	182757084	rs2270967	G	A	0.307	0.299	2.20E-09	0.296	7.31E-13	0.353	30	2.88	0.416787*	13/31	MCCC1	Yes
N-Acetylneuraminic acid	1006	9	136146597	rs550057	T	C	0.269	0.345	4.42E-12	0.341	2.00E-14	0.254	30	4.7	0.854645*	2/12	ABO	Yes
N-Acetylneuraminic acid	181	2	73900900	rs10206899	C	T	0.219	−0.495	4.23E-20	−0.507	8.66E-25	0.208	27	8.13	0.854645*	5/471	ALMS1 or NAT8	Yes/Yes
3-Aminoisobutyrate	181	2	73610116	rs17008991	A	G	0.091	−0.487	1.56E-10	−0.490	2.09E-12	0.220	28	4.03	0.614055*	2/192	ALMS1 or NAT8	Yes/Yes
Taurine	1140	11	80936097	rs12575496	T	C	0.054	−0.729	1.15E-12	−0.643	1.54E-11	0.202	30	4.96	0.330066*	8/24	-	NA
3-Hydroxyisobutyrate	988	9	107515214	rs2274870*	G	A	0.365	0.295	1.63E-10	0.277	7.85E-14	0.381	30	4.02	0.523611*	2/12	NIPSNAP3A	Yes

These 12 loci had significant genome-wide *P*-values of association (*P*-value < 1.52E-9). SNPs coordinates are reported on human reference genome GRCh37/hg19. Each locus–metabolite association is classified into four groups (colors) depending on the existence of previous publications describing the association in adults. Known urine locus-metabolite associations are colored as green ($N = 6$, 1st–6th rows); known locus–metabolite associations described only in serum are colored in orange ($N = 1$, 7th row); known urine locus, but associated to other urine metabolites, are colored in yellow ($N = 3$, 8th–10th rows) and novel locus–metabolite associations are colored in red ($N = 2$, 11th and 12th rows). Phenotype, urine metabolite associated to the locus; Region, number of the independent genome block in which the locus is located; Chr., chromosome; Pos., position; SNP, SNP selected as lead; A1, minor and effect allele; A2, non-effect allele; MAF, minor allele frequency; Standard beta, beta coefficient for the SNP from the standard linear regression model; Standard *P*-value, *P*-value for the lead SNP in the standard linear regression model; CMS beta, beta coefficient from the CMS model (including correlated covariates into the model); CMS *P*-value, *P*-value for the lead SNP in the CMS model; CMS r2, phenotype variance that can be explained only by the correlated covariates added to the model by CMS; N. covs., number of correlated covariates added by CMS; % Var. expl., percentage of phenotypic variance explained by the SNP; SNP-h^2, SNP-heritability of the metabolite; N. SNPs credible set, number of SNPs selected for the credible set versus the number of statistically significant SNPs in that locus; Potential causal gene, gene selected as causal after running FUMA, ProGeM and SNiPA and considering the gene annotation performed in other studies (8); eQTL of the potential causal gene, whether the SNP has been reported to be an eQTL of the causal gene in another study.

[a]Same SNP is identified as the lead SNP in the analysis performed without imputed genetic data.

*Nominal *P*-value of the SNP-heritability computed by GREML-GCTA is *P*-value < 0.05.

causal gene for the loci related to 3-hydroxyisovalerate was *Methylcrotonoyl-CoA carboxylase 1* (*MCCC1*), as proposed elsewhere (8). In all four cases, the metabQTLs were also identified as eQTLs of their potential causal genes in several tissues (Table 2; Supplementary Material, Tables S11 and S12).

Finally, regarding the two novel urinary metabQTLs, we identified *Nipsnap homolog 3A* (*NIPSNAP3A*) as the potential causal gene for the 3-hydroxyisobutyrate metabQTL (Fig. 2A). The SNP identified as lead for this metabQTL is a missense variant (p.Arg100Gln) in *NIPSNAP3A* and also has been reported to be an eQTL for the same gene in several tissues (Table 2; Supplementary Material, Tables S11 and S12). Furthermore, the metabQTL locus has been associated to the response to cytidine analogs (*gemcitabine*) trait (29) according to the NHGRI-EBI GWAS Catalog (Supplementary Material, Table S13), and visual inspection of stacked locus zoom plots suggest that they could be the same signal (Supplementary Material, Fig. S6).

On the contrary, we could not find any causal gene for the metabQTL associated with taurine (Table 2; Supplementary Material, Tables S11 and S12). Moreover, the lead SNP was neither an eQTL nor a SNP associated with a trait according to the NHGRI-EBI GWAS Catalog (Supplementary Material, Table S13). The regional association plot for taurine shows that the lead rs12575496 SNP is in an intergenic region (Fig. 2B): the closest upstream gene (at 0.46 Mb) is the *Long intergenic non-protein coding RNA 2720* (*LINC02720*) and the closest downstream gene (at 0.65 Mb) is the *MIR4300 Host gene* (*MIR4300HG*). Both of them are long non-coding RNAs with unknown functions.

Although some of the metabQTLs were in linkage disequilibrium (LD) with eQTLs described in several tissues, the co-localization analysis could not prove that the signal for the metabQTL was also involved in the regulation of the expression of nearby genes in kidney cortex (Supplementary Material, Table S14).

Interaction of urinary metabQTLs with dietary factors in children

Interactions between lead SNPs, representing five different metabQTLs and dietary factors were studied. These five metabQTLs, which involved four different metabolites (taurine, trimethylamine, lysine and 3-hydroxybutyrate + 3-aminoisobutyrate), were selected for analysis because a previous association between the metabolite and dietary patterns had been described in the literature (30–34) (see Material and Methods). The studied dietary factors were meat, fish, dairy products and beverage intake frequencies, which were treated in tertiles (Table 1). The different cohorts presented slightly different intake frequencies (Supplementary Material, Table S1).

Two gene-by-environment interactions (GxE) were found to be statistically significant after Bonferroni multiple-testing correction (Fig. 3). They involved SNP rs12575496 at chromosome 11, which interacted with meat (*P*-value = 2.50×10^{-4}) and beverage (*P*-value = 2.84×10^{-4}) frequency intake with respect to taurine levels. Children with the CC genotype had similar taurine levels regardless of their meat and beverage frequency intake, while children with the alternate T allele had increasing levels of taurine with increasing meat and beverage frequency intake. Neither of the two GxE interactions involving rs12575496, urinary taurine levels and meat or beverage dietary patterns replicated for serum taurine levels (*P*-value interaction = 0.117 and 0.2003, respectively) (Supplementary Material, Table S15). In fact, urinary and serum taurine levels did not correlate (Spearman's rank correlation coefficient (ρ) = −0.0113). The fixed-effects based meta-analyses by cohort showed heterogeneous results across cohorts ($I^2 > 50\%$), especially for the interaction between the taurine metabQTL and meat dietary patterns (Supplementary Material, Fig. S7, Table S15). Only the interaction term between the metabQTL and the third tertile of beverage intake remained significant (*P*-value = 0.006).

Discussion

The GWAS of 44 urine metabolite levels measured in 996 children from the HELIX project resulted in the identification of 12 metabQTLs, involving 11 unique loci and 10 different metabolites. Overall, six of them were already known urinary metabQTLs in adults, three involved known loci but associated with different urinary metabolites, and one locus had been described in serum in adults and was also nominally significant in urine, and finally, the last two represented novel urinary metabQTLs, described for the first time in HELIX children.

The six urinary metabQTLs already known in adults involve the following metabolites: glycine, lysine, trimethylamine, 2-hydroxyisobutyrate and the sum of 3-hydroxybutyrate plus 3-aminoisobutyrate. Fine mapping allowed us to reduce the number of SNPs in the loci to a maximum of seven. The genes annotated to these loci following our strategy coincided with the genes manually curated by (8): *Alanine—glyoxylate aminotransferase 2* (*AGXT2*), *4-Hydroxyphenylpyruvate dioxygenase* (*HPD*), *Pyridine nucleotide-disulfide oxidoreductase domain 2* (*PYROXD2*) and three members of the solute carrier family (*SLC6A13*, *SLC7A9* and *SLC36A2*) (Supplementary Material, Table S12).

The metabQTL identified for 3-hydroxyisovalerate had already been described in serum for 3-hydroxyisovaleric acid-carnitine, an intermediate of 3-hydroxyisovalerate (7). The proposed causal gene for the serum metabQTL was *MCCC1*, which is in agreement with the causal gene proposed by us. This gene is responsible for the first reaction in the synthesis of 3-hydroxyisovalerate by converting 3-methylcrotonyl-CoA to 3-methylglutaconyl-CoA in the mitochondria (35).

Three metabQTLs involved two known urinary loci but associated with different metabolites: one at chromosome 9 associated with N-acetylneuraminic acid, and the other at chromosome 2 also associated with N-acetylneuraminic acid as well as with 3-aminoisobutyrate. As far as we know, we are the first to study genetic variation related to N-acetylneuraminic acid, and, thus, no comparison is possible. Differently from the current study, no associations between 3-aminoisobutyrate and SNPs at this locus were described in (8). In contrast, associations between the same two loci and N-acetylaspartate were found in (8), but in the opposite direction. However, an association displaying a negative effect—as in our results—between the locus on chromosome 2 and N-acetylornithine in serum was identified in another study (7). Thus, it seems that this locus participates in the regulation of several metabolites. We identified two candidate causal genes for this locus: *N-acetyltransferase* 8 (*NAT8*) and *ALMS1*. Given the biologically meaningful link between the function of the *NAT8* gene product and the associated metabolites (acetylated compounds), *NAT8* seems a plausible candidate gene. In addition, SNPs in the region act as eQTLs for *NAT8* in several tissues and have been related to kidney function (36). At the same time, *ALMS1* has been implicated in a number of kidney health disorders (37). Regarding the metabQTL in chromosome 9 associated with N-acetylneuraminic acid, we and Raffler *et al.* (8) attributed the causal gene to the *ABO*. The lead SNP in the locus is

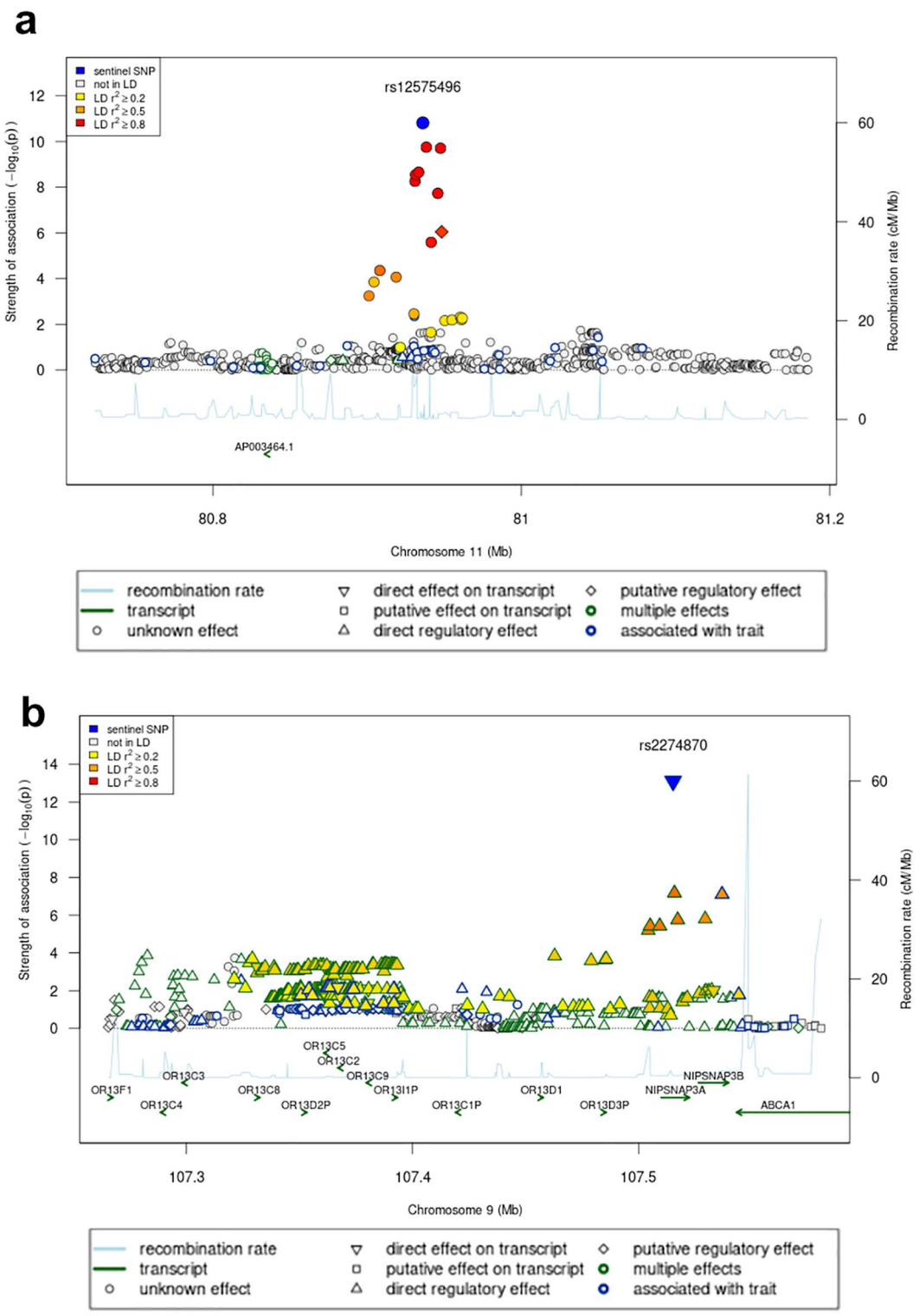

Figure 2. Regional plot of the two novel urinary metabQTLs identified in HELIX children. (**A**) Regional plot of the metabQTL associated to 3-hydroxyisobutyrate. (**B**) Regional plot of the metabQTL associated to taurine. Both regional plots represent association results from the imputed data. The *x*-axis represents the genomic position (in Mb). The *y*-axis displays the strength of the association [represented as $-\log_{10}$(*P*-value)]. The SNPs' symbols represent the functional annotations, while the colors show the pairwise LD correlations to the sentinel variant. Furthermore, the plot shows the estimated recombination rate, genes and regulatory elements. All functional annotation displayed in the plots was obtained by the SNiPA annotation browser (25).

an eQTL for the *ABO* gene in whole blood, pancreas, skeletal muscle and adipose visceral tissue. This gene codifies for a glycosyltransferase, which determines the blood group of an individual by modifying the oligosaccharides on red blood cell surface glycoproteins (38). Although these glycoproteins are not modified by the addition of N-acetylneuraminic

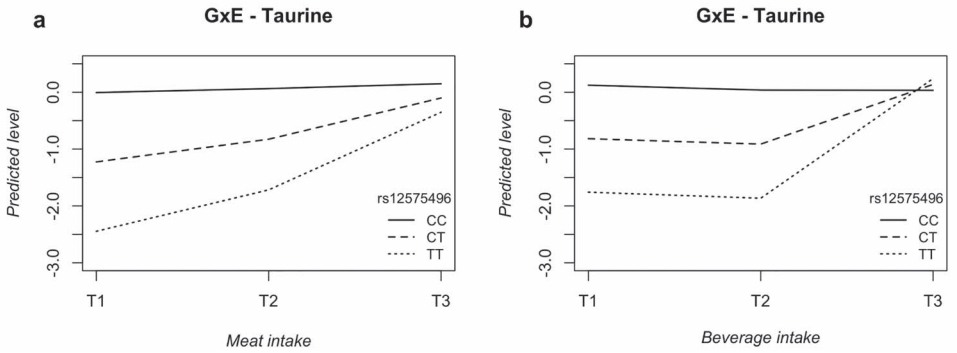

Figure 3. Graph representing the gene-by-environment (GxE) interactions for taurine. The horizontal axis represents the different values for (**A**) meat or (**B**) beverages intake frequency (from low to high intake frequency): first tertile (T1), second tertile (T2) and third tertile (T3). The vertical axis shows the predicted taurine urinary level. In each graph, the corresponding SNP and its genotypes are displayed, being, in both cases, 'T' the effect allele. Both graphs show non-parallel lines, implying a different SNP effect depending on the environmental condition: (A) meat intake and (B) beverage intake.

acid, the *ABO* gene product does modify these proteins by addition of other sialic acids, such as N-acetylglucosamine or N-acetylgalactosamine. All sialic acids are formed from N-acetylneuraminic acid (39). Therefore, by modifying the levels of other sialic acids through addition on the glycoproteins of red blood cells, the levels of their precursor (N-acetylneuraminic acid) could also be affected. Alternatively, the proton nuclear magnetic resonance spectroscopy (^{1}H NMR) peak assigned to N-acetylneuraminic acid could also include other metabolites with N-acetyl groups, such as N-acetylglucosamine and N-acetylgalactosamine.

We identified, for the first time, two novel urinary metabQTLs in HELIX children: at chromosome 9 for 3-hydroxyisobutyrate and at chromosome 11 for taurine. The fixed-effects based meta-analysis of both metabQTLs showed consistent results across cohorts. 3-hydroxyisobutyrate was not measured in the largest study to date in adult population (8), and thus the locus at chromosome 9 could still be relevant in adults. 3-hydroxyisobutyrate is a ketone body, reabsorbed in the kidneys through vesicles by a Na^+-gradient-dependent system, especially when there is a low availability of glucose (40). The lead SNP (rs2274870) is a missense variant (p.Arg100Gln) at the *NIPSNAP3A* gene. Also, it is in LD with eQTLs for *NIPSNAP3A* in several tissues, but we could not prove co-localization in the kidney cortex. *NIPSNAP3A* encodes for a protein expressed in several tissues, including the kidneys, that participates in the vesicular transport, and thus it could be involved in the reabsorption of 3-hydroxyisobutyrate (41). Furthermore, visual inspection of the region suggests that this metabQTL could also be involved in the response to *gemcitabine*, a chemotherapy drug. An effect of *NIPSNAP3A* on drug pharmacokinetics could be responsible for this finding, rather than 3-hydroxyisobutyrate being causal of chemoresistance to *gemcitabine*, since no link has been reported between this metabolite and the drug.

The metabQTL for taurine seemed to be specific of HELIX children because, although the metabolite was measured in adults (8), no signal was observed at this locus. Taurine (2-aminoethanesulfonic acid) is the most abundant, semi-essential, sulfur-containing amino acid. It is not incorporated into proteins; rather taurine acts as an intracellular osmolyte as well as a neurotransmitter, and it has many functional properties, including antioxidant and neuroprotectant effects (42). Moreover, it has several potentially beneficial effects to prevent the metabolic syndrome: it reduces triglycerides, improves insulin resistance, lowers cholesterol and reduces blood pressure by acting on the renin-angiotensin-aldosterone system, among others (33,43). Our functional annotation strategy did not provide any candidate causal gene for the taurine metabQTL. The credible set contained eight SNPs, presenting the lead one (rs12575496) a low minor allele frequency (MAF) (~0.05) and located in a gene desert region (44). Besides the two closest genes (at ~0.5 Mb), which are non-coding genes with unknown function, the lead SNP is located at 1.68 Mb from the *Prolylcarboxypeptidase* (*PRCP*) gene. *PRCP* codes for an enzyme that cleaves C-terminal amino acids linked to proline in peptides, such as angiotensin II, III and des-Arg9-bradykinin, resulting in a decrease in blood pressure (45).

Diet is the main source of taurine, although smaller amounts are also synthesized endogenously in the liver. Its levels in blood are regulated through urinary excretion. Two of the principal diet sources of taurine are meat, specially turkey and chicken, and energy drinks, where taurine is added as a supplement (46). We found that both meat and beverage intake frequency interacted with SNP rs12575496 to determine the urinary taurine levels. Although the link between meat and taurine is clear, meat being an external source of taurine, the possible link between beverage consumption and taurine remains more elusive. Previous findings have reported an interplay between taurine, which acts as a renal osmolyte, and high sugar intake-induced hypertension due to renin-angiotensin system dysregulation (33). The intake of beverages, characterized by high sugar levels, may alter blood pressure, leading to changes in renal osmolytes excretion, including taurine. In our data, the r^2 between meat and beverages intake was 0.006, which could be indicative that the two interactions are independent. In carriers of the T allele, dietary intake of meat and beverages correlated with urinary levels, while in carriers of the C allele, urinary levels were independent of dietary intake. This suggests that carriers of the C allele have the ability to adjust taurine renal excretion to environmental conditions (food intake), whereas carriers of the T allele seem to lack this food intake-dependent taurine renal excretion control. However, this proposed mechanism does not seem to apply in serum since this GxE interaction was not found for serum taurine levels. Nevertheless, replication in other datasets of the taurine urinary metabQTL and its interaction with diet would be needed, especially given the fact that there was substantial heterogeneity among cohorts and the P-values of the meta-analyses of the interaction terms were mostly not statistically significant. No other interactions between metabQTLs and dietary factors were found after multiple-testing correction.

Although our SNP-h^2 results were in the range of estimations obtained in other studies in adults (3,6), they must be interpreted with caution due to the relatively small sample size, which causes the SNP-h^2 estimate to substantially fluctuate, often hitting ≈0 or ≈1 boundary values. Therefore, these values of 0 and 1 do not represent accurate SNP-h^2 estimations but are rather an indicator of low or high SNP-h^2, respectively. However, the estimated SNP-h^2 of 1 for hippurate does not seem to be in line with findings in another study, where a low familial proportion of biological variance was reported for hippurate in urine (47). Regarding the 10 metabolites for which we detected significant metabQTLs, they presented an average SNP-h^2 of 0.58, higher than the average SNP-h^2 estimated for the other 34 metabolites, which was 0.46. In addition, the average inter-individual variability calculated from the urine samples collected during 1 week in a subset of 20 HELIX children (20) was also higher for these 10 metabolites (40.71%) compared with the average of the others (22.69%). In order to quantify the extent to which the identified metabolic loci capture metabolite variance, the estimated SNP-h^2 was compared with the metabolite variance explained by the lead SNP. For all 10 metabolites for which metabQTLs were identified, SNP-h^2 was higher than the metabolite variance explained by the lead SNP, which ranged between 2.49 and 10.26%. This is indicative that, apart from the lead SNPs, there are other SNPs (in the same locus or others) which are contributing to the variance of these metabolite levels. However, for trimethylamine, the estimated SNP-h^2 is likely to be highly influenced by the lead SNP (rs942814), which explains up to 48.28% of the metabolic variance.

Our study has many strengths. First, the use of the CMS method resulted in an increase in statistical power, thus, evidencing the benefit of the use of shared variance between variables for the identification of metabQTLs. Second, to the best of our knowledge, this is the first study in which genetic influences on urine metabolites have been identified in children and one of the few studies—either in adults or children—focusing on urine metabolites. Moreover, pooled urine samples, which best control for intra-individual variability (20), were available for the majority of individuals. Previous analysis of phenotypic variability in some of these samples reported an excellent analytical reproducibility and precision (20). Furthermore, identified metabQTLs were fine-mapped to credible sets covering from 2 to 13 SNPs, and 11 of them were functionally annotated with plausible relationships between the metabolite and the gene function. Finally, GxE were tested for candidate metabolites.

There are, however, some shortcomings in this project. First, the final sample size ($N = 996$) was relatively small. For this reason, a minimum MAF of 5% was selected, since associations with SNPs with a lower MAF could have led to false positive associations. Second, although the analyses of serum and urine metabolite levels in the cohorts were harmonized and samples were fully randomized, the results regarding GxE interactions could be influenced by the differences in dietary patterns among cohorts as well as by other potential confounding factors present in the cohorts. Third, owing to the computational burden present in the CMS method, neither genome-wide imputed genotype data nor urine metabolite ratios were considered in the present study. Imputed genotype data were only evaluated for statistically significant metabQTLs as part of the fine mapping approach. Fourth, the comparison between urinary metabQTLs in adults and children was restricted to common metabolites measured in both populations, subjected to potential limitations of metabolite annotation and limited by our statistical power. Indeed, we could nominally (P-value < 0.01) replicate four metabQTLs previously described in (8). Finally, the sample size was smaller than the one recommended for GCTA-GREML analysis, causing SNP-h^2 estimations to vary within a great range.

In summary, this is the first study identifying urinary metabQTLs in children. We have detected 12 metabQTLs, involving 11 unique loci and 10 different metabolites. Comparison with previous findings highlighted a substantial fraction of SNP–metabolite associations common to children and adults, while the association with taurine seemed to be child-specific or at least not previously reported in adults. Moreover, urinary taurine levels resulted from the combined effect of genetic variation and dietary patterns.

Materials and Methods

The HELIX project

HELIX is a collaborative research project across six established and ongoing longitudinal population-based birth cohort studies in six European countries (19,46): the Born in Bradford (BiB) study in the UK (48), the Étude des DÉterminants prÉ et postnataux du dÉveloppement et de la santÉ de l'ENfant (EDEN) study in France (49), the INfancia y Medio Ambiente (INMA) cohort in Spain (50), the Kaunas cohort (KANC) in Lithuania (51), the Norwegian Mother, Father and Child Cohort Study (MoBa) (52) and the Mother–Child Cohort study in Crete (Rhea) in Greece (53). For 1301 children, exposure data, 'omics' molecular profiles and child health outcomes were measured at 6–11 years of age. Research has been carried out according to the international and national guidelines and regulations (including the declaration of Helsinki) (20). All six HELIX cohorts had been in existence for some years and had undergone the required evaluation and obtained all the required permissions for their cohort recruitment and follow-up visits. Participants were asked to sign an informed consent form for specific HELIX work, including clinical examination and biospecimen collection (19). Each cohort also confirmed that relevant informed consent and approval were in place for the secondary use of data from pre-existing data. The work in HELIX was covered by new ethics approvals in each country. The final dataset included 996 European ancestry children presenting both serum and urine metabolite levels as well as genotyping data (Supplementary Material, Fig. S1). Workflow of the study is show in Fig. 1.

Data processing

Data processing, as well as the rest of the analyses, were performed using R ('The R Project for Statistical Computing') software environment (v3.5.0) (54), unless stated otherwise.

Genotype data, quality control and imputation

Peripheral blood was collected in EDTA tubes, centrifuged and the buffy coat was saved for DNA extraction. DNA was extracted using the Chemagen kit (Perkin Elmer) at the CeGen-ISCIII (Universidad de Santiago de Compostela). Genome-wide genotyping of 1397 HELIX samples was performed using the Infinium Global Screening Array (GSA) MD version 1 (Illumina), which includes 692 367 variants, at the Human Genomics Facility (HuGe-F), Erasmus MC (www.glimdna.org). Genotype calling was carried out using the GenTrain2.0 algorithm based on a custom cluster file implemented in the GenomeStudio software. SNPs coordinates were reported on human reference genome GRCh37/hg19 and on + strand. DNA extraction and genome-wide genotyping were performed using the same harmonized protocol in all cohorts.

Quality control (QC) was performed with the PLINK program following standard recommendations (54,55). Sample QC consisted of filtering by <97% sample call rate (N = 43), sex discordance (N = 8), heterozygosity (N = 0), relatedness—including potential DNA contamination—(N = 10) and duplicate samples (N = 19). Next, the Peddy program was used to predict ancestry from GWAS data (57). To do so, 6642 genetic variants which are highly polymorphic among populations were extracted from HELIX and compared with the 1000G project data (58). Ancestry predicted from GWAS was, then, contrasted with the ancestry recorded in the questionnaires, which led to the identification and exclusion of 12 samples with discordant ancestry. Overall, 93 samples (6.7%) were filtered out, and 1304 samples remained after sample QC. Regarding genetic variant QC, 408 217 SNPs were filtered out because they either had a variant call rate below 95% (N = 4046), were not in the Hardy–Weinberg equilibrium (HWE) (*P*-value $< 1 \times 10^{-06}$) (N = 913), displayed MAFs < 5% (N = 388 822) or were not aligned to the reference genomes (N = 14 436). Finally, SNPs located in the mitochondrial DNA and in the non-canonical pseudo-autosomal region (PAR) of chromosome Y were removed (N = 446). As a result, 283 704 SNPs remained in the dataset.

Imputation of the GWAS data was performed with the Imputation Michigan server (59) using the HRC cosmopolitan panel, version r1.1 2016 (22). Before imputation, PLINK GWAS data were converted into VCF format, and variants were aligned with the reference genome. The phasing of the haplotypes was done with Eagle v2.4 (60) and the imputation with minimac4 (61), both implemented in the code by the Imputation Michigan server. Chromosome X was imputed, including PAR and non-PAR regions. In total, we retrieved 40 405 505 variants after imputation. Then, we applied the following QC criteria to the imputed dataset: imputation accuracy (*r2*) > 0.9, MAF > 1%, HWE *P*-value $> 1 \times 10^{-06}$; and genotype probabilities were converted to genotypes using the best guest approach. The final post-imputation QCed dataset consisted of 1304 samples and 6 143 757 variants (PLINK format, Genome build: GRCh37/hg19, + strand).

Metabolomic data and QC

Although this project focused on the identification of urinary metabQTLs, serum metabolite levels were considered as covariates in the statistical models (see in the following). Urine and serum samples were collected and processed according to identical pre-defined standardized protocols across all six cohorts (30). As for urine samples, two samples, representing last night-time and first morning voids, were collected on the evening and morning before follow-up visit and were subsequently pooled to generate a more representative sample of the last 24 h for metabolomic analysis (*n* = 933). Either the night-time void (*n* = 32) or morning void (*n* = 31) sample was analyzed in cases where a pooled sample was missing (Table 1). Serum was obtained from the peripheral blood collected during the follow-up visit, with a mean postprandial interval of 3.65 h.

Urinary metabolic profiles were analyzed by ^{1}H NMR on a 14.1 Tesla (600 MHz ^{1}H) NMR spectrometer at Imperial College London, following a non-targeted approach. Samples from all cohorts and batches were fully randomized. A total of 44 metabolites belonging to 22 metabolic classes were annotated: quantification was achieved for 24 metabolites, and the remaining 20 metabolites were semi-quantified (Supplementary Material, Table S16) (20,30). The mean of the coefficients of variations (CV) across the 44 NMR detected urinary metabolites was 11% [SD = 5.7%]. Individual CV and type of evidence used for metabolite assignment is indicated in Supplementary Material, Table S16 (30). Urine metabolite levels were normalized with the median fold change method and one-half of the minimum value was used as an offset. Then data were log2 transformed. A summary of urinary metabolite concentrations across the six cohorts can be found in Supplementary Material, Table S2.

The levels of the serum metabolites were analyzed by a targeted approach using the Absolute*IDQ* p180 kit (BIOCRATES Life Sciences AG) following the manufacturer's protocol using LC–MS/MS and Agilent HPLC 1100 liquid chromatography coupled to a SCIEX QTRAP 6500 triple quadrupole mass spectrometer at Imperial College London (20,30,62). Samples from all cohorts and batches were fully randomized. A total of 177 serum metabolites were quantified (Supplementary Material, Table S17). Serum metabolite concentrations were $\log_2$ transformed prior to statistical analyses (30).

Dietary data

Data on the food intake frequency of 44 food items from 11 main food groups were collected through a short food frequency questionnaire, and the average number of times per week that each food item was consumed was recorded. Frequencies were, then, classified into tertiles. Only data from the following four main food groups, which are known to be associated to at least one metabolite with an identified metabQTL, were considered in this project: meat (processed meat, poultry and red meat), fish (canned fish, oily fish, white fish and seafood), dairy products (yogurt, cheese, milk and dairy desserts) and beverages (sugar-sweetened soft and fizzy drinks and artificially sweetened, low-sugar or sugar-free soda) (30). Dietary data were available for 980 individuals.

Calculation of SNP-h^2 of urinary metabolites

SNP-h^2 for the 44 metabolites was computed using the genome-wide complex trait analysis (GCTA) program, which estimates the additive contribution to a trait's heritability of a particular subset of SNPs (GREML analysis) (63). First, estimates of genetic sharing between individuals, represented as a genetic relationship matrix (GRM), from only the genome-wide autosomal common SNPs (MAF > 5%) were computed by GCTA (63). Then, GCTA's restricted maximum likelihood (REML) (64) was used to estimate the variance explained by the GRM (random-effects) adjusted for selected fixed-effects (age, sex and urine sampling type) in a linear mixed-effects model. The resulting SNP heritability from the REML analysis represented the proportion of phenotypic variation (i.e. metabolite levels) explained by the genotypic variation (i.e. set of SNPs).

Identification of child urinary metabQTLs

Data from the six cohorts were combined in a single dataset to perform the analyses for the identification of metabQTLs. The reasons for this were: (i) the well-established and harmonized protocol across cohorts; (ii) the absence of batch effect between cohorts in the serum and urine metabolome data (Supplementary Material, Fig. S4) and (iii) the control of potential population substratification by adjusting for the first 20 GWAS PCs.

For the identification of metabQTLs, we applied the CMS method (21). It consists of, when testing for association between a genotype and a phenotype (outcome), including as covariates other collected correlated phenotypes that vary with the outcome because of shared risk factors. The core of the method

is an approach to select these covariates that are correlated with the phenotype but not with the genotype, thereby decreasing the phenotypic variance independently of the genotype (e.g. phenotypic variance due to environmental factors), hence, increasing the power. This approach achieves a gain in power equivalent to that resulting from a 2- to 3-fold increase in sample size (21). In this study, which aimed to identify urinary metabQTLs, both urine and serum metabolite levels served as possible covariates of the model in which a urine metabolite was the outcome variable. Therefore, serum metabolite levels were only included in the model as covariates since no serum metabQTLs were studied.

The CMS pipeline is a three-step code: (i) pre-processing and pre-filtering of the data, which pre-selects covariates (metabolites levels expressed in log2) for each studied phenotype and standardizes all variables (i.e. metabolite levels and variables adjusting the model) to a standard normal distribution (mean = 0 and SD = 1); (ii) CMS running for each SNP—metabolite combination and (iii) summarization of results by approximately independent LD blocks (64) and creation of tables and plots (65). The statistical analysis performed by CMS consisted of linear regression models under the assumption of an additive genetic model after selecting—if any—the correlated covariates to be included in the model. Models were adjusted for age, sex, 20 first GWAS PCs calculated in European ancestry children—accounting for a 31.2% of the GWAS variability—and urine sampling type (night only, morning only or pooled sample). Categorical variables, such as sex and urine sampling type, were treated as dummy variables and were subsequently standardized in the first step of the CMS pipeline as well as the rest of variables. SNPs located in the non-PAR region of chromosome X were codified as 0–2 in males and as 0–1–2 in females. Instead, SNPs in the PAR region were codified as 0–1–2 in both males and females.

In a first screening, CMS was applied on the genome-wide genotyped SNPs. Significant associations were determined by dividing the standard genome-wide significance threshold of 5×10^{-8} by the virtual number of independent tests across the real number of tests performed computed by taking into account the degree of correlation between urinary metabolites levels, named as the effective number of tests (ENT), adapted from elsewhere (67,68). A P-value of 1.52×10^{-9} was obtained. For the metabQTLs identified through analysis of the genotyped data, we repeated the CMS analysis with genotyped plus HRC imputed SNPs (MAF > 0.05) located within the 11 independent LD blocks in which the 11 statistically significant metabQTLs were found. For each statistically significant locus–metabolite association identified in this second screening, a lead SNP was identified (see in the following). The proportion of the phenotype's variance explained by the lead SNP was computed for each metabolite according to a method used by Teslovich *et al.* (69).

Moreover, we performed standard linear regression analyses by cohort (adjusted for sex, age, urine sample type and 20 first GWAS) of the two novel metabQTLs identified in order to address potential heterogeneity across cohorts. CMS was not used for this analysis since, for these two metabQTLs, results provided by the CMS algorithm were very similar to those obtained without adjusting for the covariates selected by CMS (Table 2). Results by cohort were combined through fixed-effects-based meta-analysis and represented in forest plots with the meta R package (70).

Fine-mapping of child urinary metabQTLs

Fine-mapping was performed with the Bayesian-based method PAINTOR v3.0 (23). PAINTOR used the genotype–phenotype association results from the second screening (genotype plus HRC imputed data) to calculate, for each SNP, its posterior inclusion probability (PIP). PIP allows to estimate the minimum set of SNPs, known as 'credible set', that contains all causal SNPs with a probability of alpha. In this study, we used an alpha of 0.1 (confidence interval of 90%). Next, the selection of the lead SNP for each statistically significant locus was carried out. To do so, apart from the credible set obtained by PAINTOR, the FUMA (26) web-based platform was ran using default parameters except for the LD threshold used to define 'independent' SNP(s) within each locus, which was set to $r^2 < 0.4$. Results from these two resources were combined into a score table in which each proposed lead SNP obtained a score based on: (i) if it was present in the credible set obtained by PAINTOR; (ii) if FUMA had selected it as a lead SNP; (iii) if FUMA had selected it as an independent significant SNP and (iv) if an association between this SNP and levels of a metabolite (either serum or urinary) was described in the NHGRI-EBI GWAS Catalog (71) (Supplementary Material, Table S11).

Comparison of child urinary metabQTLs with metabQTLs in adults

MetabQTLs identified in our study in children were compared with results from other GWAS studies of urinary metabolites in adults (Supplementary Material, Table S7). To do so, the summary statistics of all significant SNP–metabolite associations from these studies were converged together to create a reference dataset for comparison. The Metabolomics GWAS Server (7,72) and the NHGRI-EBI GWAS Catalog (71) (study GCST003119 (8) downloaded on 08/06/2019) were used to download most of these results, while others were manually added from the published paper. For those significant metabQTLs not previously described in urine, we searched whether they had been described in other biological matrices, such as serum. For this, we queried the Metabolomics GWAS Server (7,72) and the NHGRI-EBI GWAS Catalog (71).

Functional annotation of child urinary metabQTLs

Mapping SNP(s) for each statistically significant locus to candidate genes that are most likely causally linked to the observed changes in the metabolite levels was carried out using three different functional annotation tools: SNiPA (25), FUMA (26) and ProGeM (27). As for ProGeM, the size of the LD region was the only parameter modified to 500 kb upstream and 500 kb downstream of the SNP. These tools, despite not performing statistical co-localization analysis, carry out extensive searches across multiple eQTL data sources to identify if any of the queried SNPs, or SNPs in LD, have been identified as eQTLs in diverse tissues (with the exception of kidneys). To complement them, we performed a co-localization analysis between all metabQTLs and eQTLs for kidney cortex described in GTEx v8, using the coloc R package with default parameters (28). This package tests five hypotheses [H_0: no association with either trait; H_1: association with metabolite levels (metabQTL) only; H_2: association with gene expression (eQTL) only; H_3: association with metabolite levels and gene expression, two distinct causal variants and H_4: association with metabolite levels and gene expression, one shared causal variant] and provides a posterior probability at the SNP level for each one. Summary statistics and MAF of all common SNPs in our association analysis (using imputed SNPs) and GTEx data for kidney cortex located within 1 Mb centered at the lead SNP of each metabQTL were considered for the analysis.

Selection of the candidate gene(s) affected by the lead SNP and presumably causally linked to the observed changes in the metabolite levels was done similarly to the selection of the lead SNP. The outputs of SNiPA, FUMA and ProGeM were analyzed and a summary table was built in which each gene proposed by either of the three programs received a score based on: (i) how many programs identified the gene as candidate gene; (ii) if it was the nearest gene; (iii) if the lead SNP was an eQTL linked to the gene; (iv) if the lead SNP was located in a coding region of the gene and (v) if an association between the lead SNP and the gene was described in the NHGRI-EBI GWAS Catalog (71) (Supplementary Material, Table S11). This classification was manually curated based on the UCSC Genome Browser (73).

Furthermore, as performed by FUMA, each lead SNP was searched in the NHGRI-EBI GWAS Catalog to find previously reported associations between the SNP and a variety of phenotypes (Supplementary Material, Table S13). The stacked locus zoom plot representing 3-hydroxyisobutyrate metabQTL and an association with response to *gemcitabine* in the same locus was obtained by the 'LocusCompareR' R package.

Interaction of urinary metabQTLs with dietary factors

For statistically significant metabQTLs, where the metabolite had been related to a dietary factor in the literature or in the Exposome-Explorer database (31), GxE interactions were tested. In the case of taurine, the Exposome-Explorer indicated an association with protein intake (31), thus, we tested the interactions between the lead SNP of the taurine metabQTL and the different sources of proteins: SNP × meat, SNP × dairy and SNP × fish intake. Furthermore, an association between sugar-sweetened beverage intake and taurine urinary levels is described in the literature (32), so the SNP × beverages consumption interaction was also studied. Trimethylamine is produced by the action of gut microbiota using dietary precursors such as choline, betaine or l-carnitine, which are present in meat, fish and daily products, among others, thus we tested SNP × meat, SNP × dairy and SNP × fish intake. Shellfish intake is associated with lysine urine levels (34), thus, for lysine, we tested SNP × fish intake since shellfish intake is included in the fish intake variable. Finally, in the case of the trait combining 3-hydroxybutyrate and 3-aminoisobutyrate levels, an interaction between the lead SNP and beverage consumption was computed (30). For the rest of metabolites with metabQTLs, no associations between their levels and food intake could be found in the Exposome-Exporer database.

GxE were tested for the lead SNP of each metabQTLs identified by the fine-mapping of the association results of the second CMS analysis in which genotyped plus HRC imputed SNPs were tested. Standard linear regression models were used under the assumption of an additive genetic factor interacting with a dietary factor. Models were adjusted for age, sex, urine sampling type and the 20 first PCs. Present analysis included 980 individuals with available dietary pattern data. To find statistically significant interactions with the lead SNP, an analysis of variance (ANOVA) was carried out to compare the model with the interactions against the null model (Supplementary Material, Table S15). Moreover, we analyzed the two statistically significant GxE interactions, involving taurine and meat or beverages intake, by cohort. Results by cohort were combined through fixed-effects based meta-analysis and were represented in forest plots.

Supplementary Material

Supplementary Material is available at *HMG* online.

Authors' Contributions

M.B. and G.E. conceptualized the study. B.C.-S. performed all the statistical analysis. M.Vr. coordinated and obtained main funding of the HELIX project. M.B., G.E. and K.B.G. obtained genome-wide genotyping data. L.M., C.-H.E.L. and A.P.S. obtained metabolomics data under the supervision of M.Co. and H.C.K. K.B.G., S.A., M.Ca., S.C., L.Ch., J.R.G., R.G., R.Mc., R.S., M.Va., J.W. and M.Vr. obtained data and biological samples from the cohorts. B.C.-S., G.E. and M.B. wrote the first version of the manuscript. All authors critically read and approved the manuscript.

Acknowledgements

We would like to thank all the children and their families for their generous contribution. MoBa are grateful to all the families in Norway who took part in this ongoing cohort study. The authors would also like to thank Ingvild Essén, Jorunn Evandt for thorough field work, Heidi Marie Nordheim for biological sample management and the MoBa–administrative unit.

Conflicts of interest. None declared.

Funding

European Community's Seventh Framework Programme (FP7/2007-206); [HELIX project (308333)]; the H2020-EU.3.1.2.—Preventing Disease Programme [ATHLETE project (874583)]. The genotyping was supported by Instituto de Salud Carlos III (PI17/01225) and co-funded by the European Union (ERDF, 'A way to make Europe') and the Centro Nacional de Genotipado-CEGEN (PRB2-ISCIII). BiB received core infrastructure funding from the Wellcome Trust (WT101597MA) and a joint grant from the UK Medical Research Council and Economic and Social Science Research Council (MR/N024397/1). INMA data collections were supported by grants from the Instituto de Salud Carlos III, Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública (CIBERESP) and the Generalitat de Catalunya-CIRIT. KANC was funded by the grant of the Lithuanian Agency for Science Innovation and Technology (6-04-2014_31V-66). The Norwegian Mother, Father and Child Cohort Study is supported by the Norwegian Ministry of Health and Care Services and the Ministry of Education and Research. The Rhea project was financially supported by European projects (EU FP6-2003-Food-3-NewGeneris, EU FP6. STREP Hiwate, EU FP7 ENV.2007.1.2.2.2. Project No 211250 Escape, EU FP7-2008-ENV-1.2.1.4 Envirogenomarkers, EU FP7-HEALTH-2009-single stage CHICOS, EU FP7 ENV.2008.1.2.1.6. Proposal No 226285 ENRIECO, EU-FP7-HEALTH-2012 Proposal No 308333 HELIX) and the Greek Ministry of Health (Program of Prevention of obesity and neurodevelopmental disorders in preschool children, in Heraklion district, Crete, Greece: 2011-2014; 'Rhea Plus': Primary Prevention Program of Environmental Risk Factors for Reproductive Health, and Child Health: 2012-15). We acknowledge support from the Spanish Ministry of Science and Innovation through the Centro de Excelencia Severo Ochoa 2019-2023 Program (CEX2018-000806-S) and support from the Generalitat de Catalunya through the CERCA Program. M.Ca. holds a Miguel Servet fellowship (CP16/00128) funded by Instituto de Salud Carlos III and cofunded by the European Social Fund Investing in your future.

Availability of Data, Material and Code

Summarized results (SNP, coordinates, alleles, effect, standard error, *P*-value) are available at the HELIX webpage:

https://helixomics.isglobal.org/. The main code can be found at: https://github.com/beacalvo/metabQTLs.

Ethics Statement

Prior to the start of HELIX, all six cohorts on which HELIX is based had been in existence for some years, had undergone the required evaluation by national ethics committees and had obtained all the required permissions for their cohort recruitment and follow-up visits. At enrolment in the HELIX subcohort and panel studies, participants were asked to sign an informed consent form for the specific HELIX work, including clinical examination and biospecimen collection and analysis. An Ethics Task Force was established to support the HELIX project on ethical issues, for advice on the project's ethical compliance, identification and alerting to changes in legislation where applicable. Specific procedures are in place within HELIX to safeguard the privacy of study subjects and confidentiality of data. First, any reported study results pertain to analyses of aggregate data; no variables or combination of variables that can identify an individual will be associated with any published or unpublished report of this study. Primary databases with personal information (such as geocodes, dates, questionnaires or health outcomes) have been stored on separate computers with personal identifiers removed. Subjects are identified by a unique study number, linking all basic data required for the study. The master key file linking the study numbers with personal identifiers is maintained in each cohort. For the dataset analysis, all information that enables identification of an individual (dates, geocodes, etc.) is removed before distribution of datasets to the researchers. All data exchanges will adhere to the most up-to-date EU and national data protection regulations.

References

1. Kastenmüller, G., Raffler, J. and Gieger, C. (2015) Genetics of human metabolism: an update. *Hum. Mol. Genet.*, **24**, R93–R101.
2. Dumas, M.E. (2012) Metabolome 2.0: quantitative genetics and network biology of metabolic phenotypes. *Mol. BioSyst.*, **8**, 2494–2502.
3. Dharuri, H., Demirkan, A., van Klinken, J.B., Mook-Kanamori, D.O., van Duijn, C.M., t Hoen, P.A. and Willems van Dijk, K. (2014) Genetics of the human metabolome, what is next? *Biochim. Biophys. Acta-Mol. Basis Dis.*, **1842**, 1923–1931.
4. Parsons, L.N. (2014) *Inborn Errors of MetaboLism*, 1st edn. Elsevier Inc., Amsterdam, Netherlands, Vol. **73**.
5. Suhre, K., Raffler, J. and Kastenmüller, G. (2016) Biochemical insights from population studies with genetics and metabolomics. *Arch. Biochem. Biophys.*, **589**, 168–176.
6. Kettunen, J., Tukiainen, T., Sarin, A.P., Ortega-Alonso, A., Tikkanen, E., Lyytikäinen, L.P., Kangas, A.J., Soininen, P., Würtz, P., Silander, K. *et al.* (2012) Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet.*, **44**, 269–276.
7. Shin, S.Y., Fauman, E.B., Petersen, A.K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I., Forgetta, V., Yang, T.P. *et al.* (2014) An atlas of genetic influences on human blood metabolites. *Nat. Genet.*, **46**, 543–550.
8. Raffler, J., Friedrich, N., Arnold, M., Kacprowski, T., Rueedi, R., Altmaier, E., Bergmann, S., Budde, K., Gieger, C., Homuth, G. *et al.* (2015) Genome-wide association study with targeted and non-targeted NMR metabolomics identifies 15 novel loci of urinary human metabolic individuality. *PLoS Genet.*, **11**, e1005487.
9. Nicholson, G., Rantalainen, M., Li, J.V., Maher, A.D., Malmodin, D., Ahmadi, K.R., Faber, J.H., Barrett, A., Min, J.L., Rayner, N.W. *et al.* (2011) A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. *PLoS Genet.*, **7**, e1002270.
10. Rueedi, R., Ledda, M., Nicholls, A.W., Salek, R.M., Marques-Vidal, P., Morya, E., Sameshima, K., Montoliu, I., Da Silva, L., Collino, S. *et al.* (2014) Genome-wide association study of metabolic traits reveals novel gene-metabolite-disease links. *PLoS Genet.*, **10**, e1004132.
11. Ference, B.A., Ray, K.K., Catapano, A.L., Ference, T.B., Burgess, S., Neff, D.R., Oliver-Williams, C., Wood, A.M., Butterworth, A.S., Di Angelantonio, E. *et al.* (2019) Mendelian randomization study of ACLY and cardiovascular disease. *N. Engl. J. Med.*, **380**, 1033–1042.
12. Haworth, C.M.A., Carnell, S., Meaburn, E.L., Davis, O.S., Plomin, R. and Wardle, J. (2008) Increasing heritability of BMI and stronger associations with the FTO gene over childhood. *Obesity*, **16**, 2663–2668.
13. Graff, M., Ngwa, J.S., Workalemahu, T., Homuth, G., Schipf, S., Teumer, A., Völzke, H., Wallaschofski, H., Abecasis, G.R., Edward, L. *et al.* (2013) Genome-wide analysis of BMI in adolescents and young adults reveals additional insight into the effects of genetic loci over the life course. *Hum. Mol. Genet.*, **22**, 3597–3607.
14. Winkler, T.W., Justice, A.E., Graff, M., Barata, L., Feitosa, M.F., Chu, S., Czajkowski, J., Esko, T., Fall, T., Kilpeläinen, T.O. *et al.* (2015) The influence of age and sex on genetic associations with adult body size and shape: a large-scale genome-wide interaction study. *PLoS Genet.*, **11**, e1005378.
15. Elks, C.E., den Hoed, M., Zhao, J.H., Sharp, S.J., Wareham, N.J., Loos, R.J. and Ong, K.K. (2012) Variability in the heritability of body mass index: a systematic review and meta-regression. *Front. Endocrinol. (Lausanne)*, **3**, 29.
16. Couto-Alves, A., De Silva, N., Karhunen, V., Sovio, U., Das, S., Taal, H.R., Warrington, N.M., Lewin, A.M., Kaakinen, M., Cousminer, D. *et al.* (2019) GWAS on longitudinal growth traits reveals different genetic factors influencing infant, child, and adult BMI. *Sci. Adv.*, **5**, eaaw3095.
17. Helgeland, Ø., Vaudel, M., Juliusson, P.B., Holmen, O.L., Juodakis, J., Bacelis, J., Jacobsson, B., Lindekleiv, H., Hveem, K., Lie, R.T. *et al.* (2019) Genome-wide association study reveals a dynamic role of common genetic variation in infant and early childhood growth. *Nat. Commun.*, **10**, 4448.
18. Swallow, D.M. (2003) Genetics of lactase persistence and lactose intolerance. *Annu. Rev. Genet.*, **37**, 197–219.
19. Maitre, L., de Bont, J., Casas, M., Robinson, O., Aasvang, G.M., Agier, L., Andrušaitytė, S., Ballester, F., Basagaña, X., Borràs, E. *et al.* (2018) Human early life exposome (HELIX) study: a European population-based exposome cohort. *BMJ Open*, **8**, e021311.
20. Maitre, L., Lau, C.E., Vizcaino, E., Robinson, O., Casas, M., Siskos, A.P., Want, E.J., Athersuch, T., Slama, R., Vrijheid, M. *et al.* (2017) Assessment of metabolic phenotypic variability in children's urine using 1H NMR spectroscopy. *Sci. Rep.*, **7**, 46082.
21. Aschard, H., Guillemot, V., Vilhjalmsson, B., Patel, C.J., Skurnik, D., Ye, C.J., Wolpin, B., Kraft, P. and Zaitlen, N. (2017) Covariate selection for association screening in multiphenotype genetic studies. *Nat. Genet.*, **49**, 1789–1795.
22. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C.,

Danecek, P., Sharp, K. *et al.* (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.*, **48**, 1279–1283.
23. Kichaev, G., Yang, W.Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P. and Pasaniuc, B. (2014) Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.*, **10**, e1004722.
24. Suhre, K., Wallaschofski, H., Raffler, J., Friedrich, N., Haring, R., Michael, K., Wasner, C., Krebs, A., Kronenberg, F., Chang, D. *et al.* (2011) A genome-wide association study of metabolic traits in human urine. *Nat. Genet.*, **43**, 565–569.
25. Arnold, M., Raffler, J., Pfeufer, A., Suhre, K. and Kastenmüller, G. (2015) SNiPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics*, **31**, 1334–1336.
26. Watanabe, K., Taskesen, E., Van Bochoven, A. and Posthuma, D. (2017) Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.*, **8**, 1826.
27. Stacey, D., Fauman, E.B., Ziemek, D., Sun, B.B., Harshfield, E.L., Wood, A.M., Butterworth, A.S., Suhre, K. and Paul, D.S. (2019) ProGeM: A framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res.*, **47**, e3.
28. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C. and Plagnol, V. (2014) Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.*, **10**, e1004383.
29. Li, L., Fridley, B.L., Kalari, K., Niu, N., Jenkins, G., Batzler, A., Abo, R.P., Schaid, D. and Wang, L. (2014) Discovery of genetic biomarkers contributing to variation in drug response of cytidine analogues using human lymphoblastoid cell lines. *BMC Genomics*, **15**, 93.
30. Lau, C.E., Siskos, A.P., Maitre, L., Robinson, O., Athersuch, T.J., Want, E.J., Urquiza, J., Casas, M., Vafeiadi, M., Roumeliotaki, T. *et al.* (2018) Determinants of the urinary and serum metabolome in children from six European populations. *BMC Med.*, **16**, 202.
31. Neveu, V., Moussy, A., Rouaix, H., Wedekind, R., Pon, A., Knox, C., Wishart, D.S. and Scalbert, A. (2017) Exposome-explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors. *Nucleic Acids Res.*, **45**, D979–D984.
32. Gibbons, H., McNulty, B.A., Nugent, A.P., Walton, J., Flynn, A., Gibney, M.J. and Brennan, L. (2015) A metabolomics approach to the identification of biomarkers of sugar-sweetened beverage intake. *Am. J. Clin. Nutr.*, **101**, 471–477.
33. Chesney, R.W., Han, X. and Patters, A.B. (2010) Taurine and the renal system. *J. Biomed. Sci.*, **17**, S4.
34. Playdon, M.C., Sampson, J.N., Cross, A.J., Sinha, R., Guertin, K.A., Moy, K.A., Rothman, N., Irwin, M.L., Mayne, S.T., Stolzenberg-Solomon, R. *et al.* (2016) Comparing metabolite profiles of habitual diet in serum and urine. *Am. J. Clin. Nutr.*, **104**, 776–789.
35. Mock, D.M., Stratton, S.L., Horvath, T.D., Bogusiewicz, A., Matthews, N.I., Henrich, C.L., Dawson, A.M., Spencer, H.J., Owen, S.N., Boysen, G. *et al.* (2011) Urinary excretion of 3-hydroxyisovaleric acid and 3-hydroxyisovaleryl carnitine increases in response to a leucine challenge in marginally biotin-deficient humans. *J. Nutr.*, **141**, 1925–1930.
36. Köttgen, A., Pattaro, C., Böger, C.A., Fuchsberger, C., Olden, M., Glazer, N.L., Parsa, A., Gao, X., Yang, Q., Smith, A.V. *et al.* (2010) Multiple new genetic loci associated with kidney function and chronic kidney disease: the CKDGen consortium. *Nat Genet.*, **42**, 376–384.
37. Köttgen, A., Glazer, N.L., Dehghan, A., Hwang, S.J., Katz, R., Li, M., Yang, Q., Gudnason, V., Launer, L.J., Harris, T.B. *et al.* (2010) New loci associated with kidney function and chronic kidney disease. *Nat. Genet.*, **42**, 376–384.
38. Hosoi, E. (2008) Biological and clinicel aspects of *ABO* blood group system. Biological and clinicel aspects of *ABO* blood group system. *J. Med. Investig.*, **55**, 174–182.
39. Schauer, R. and Kamerling, J.P. (2018) Exploration of the sialic acid world. *Adv. Carbohydr. Chem. Biochem.*, **75**, 1–213.
40. Garcia, M.L., Benavides, J. and Valdivieso, F. (1980) Ketone body transport in renal brush border membrane vesicles. *Biochim. Biophys. Acta-Biomembr.*, **600**, 922–930.
41. Buechler, C., Bodzioch, M., Bared, S.M., Sigruener, A., Boettcher, A., Lapicka-Bodzioch, K., Aslanidis, C., Duong, C.Q., Grandl, M. *et al.* (2004) Expression pattern and raft association of NIPSNAP3 and NIPSNAP4, highly homologous proteins encoded by genes in close proximity to the ATP-binding cassette transporter A1. *Genomics*, **83**, 1116–1124.
42. Ripps, H. and Shen, W. (2012) Review: taurine: A 'very essential' amino acid. *Mol. Vis.*, **18**, 2673–2686.
43. Chen, W., Guo, J., Zhang, Y. and Zhang, J. (2016) The beneficial effects of taurine in preventing metabolic syndrome. *Food Funct.*, **7**, 1849–1863.
44. Smigielski, E.M., Sirotkin, K., Ward, M. and Sherry, S.T. (2000) dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.*, **28**, 352–355.
45. Tan, F., Morris, P.W., Skidgel, R.A. and Erdös, E.G. (1993) Sequencing and cloning of human prolylcarboxypeptidase (angiotensinase C). Similarity to both serine carboxypeptidase and prolylendopeptidase families. *J. Biol. Chem.*, **268**, 16631–16638.
46. Wójcik, O.P., Koenig, K.L., Zeleniuch-Jacquotte, A., Costa, M. and Chen, Y. (2010) The potential protective effects of taurine on coronary heart disease. *Atherosclerosis*, **208**, 19–25.
47. Nicholson, G., Rantalainen, M., Maher, A.D., Li, J.V., Malmodin, D., Ahmadi, K.R., Faber, J.H., Hallgrímsdóttir, I.B., Barrett, A., Toft, H. *et al.* (2011) Human metabolic profiles are stably controlled by genetic and environmental variation. *Mol. Syst. Biol.*, **7**, 525.
48. Wright, J., Small, N., Raynor, P., Tuffnell, D., Bhopal, R., Cameron, N., Fairley, L., Lawlor, D.A., Parslow, R., Petherick, E.S. *et al.* (2013) Cohort profile: the born in Bradford multi-ethnic family cohort study. *Int. J. Epidemiol.*, **42**, 978–991.
49. Heude, B., Forhan, A., Slama, R., Douhaud, L., Bedel, S., Saurel-Cubizolles, M.J., Hankard, R., Thiebaugeorges, O., De Agostini, M., Annesi-Maesano, I. *et al.* (2016) Cohort profile: the EDEN mother-child cohort on the prenatal and early postnatal determinants of child health and development. *Int. J. Epidemiol.*, **45**, 353–363.
50. Guxens, M., Ballester, F., Espada, M., Fernández, M.F., Grimalt, J.O., Ibarluzea, J., Olea, N., Rebagliato, M., Tardón, A., Torrent, M. *et al.* (2012) Cohort profile: the INMA-INfancia y Medio Ambiente-(environment and childhood) project. *Int. J. Epidemiol.*, **41**, 930–940.
51. Grazuleviciene, R., Danileviciute, A., Nadisauskiene, R. and Vencloviene, J. (2009) Maternal smoking, GSTM1 and GSTT1 polymorphism and susceptibility to adverse pregnancy outcomes. *Int. J. Environ. Res. Public Health*, **6**, 1282–1297.
52. Magnus, P., Birke, C., Vejrup, K., Haugan, A., Alsaker, E., Daltveit, A.K., Handal, M., Haugen, M., Høiseth, G., Knudsen, G.P. *et al.* (2016) Cohort profile update: the Norwegian mother and child cohort study (MoBa). *Int. J. Epidemiol.*, **45**, 382–388.

53. Chatzi, L., Leventakou, V., Vafeiadi, M., Koutra, K., Roumeliotaki, T., Chalkiadaki, G., Karachaliou, M., Daraki, V., Kyriklaki, A., Kampouri, M. *et al.* (2017) Cohort profile: the mother-child cohort in Crete, Greece (Rhea study). *Int. J. Epidemiol.*, **46**, 1392–1393k.
54. R. Core Team (2017) *R: A Language and Environment for Statistical Computing*.
55. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
56. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M. and Lee, J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, **4**, 7.
57. Pedersen, B.S. and Quinlan, A.R. (2017) Who's who? Detecting and resolving sample anomalies in human DNA sequencing studies with peddy. *Am. J. Hum. Genet.*, **100**, 406–413.
58. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A. and Abecasis, G.R. (2015) A global reference for human genetic variation. A global reference for human genetic variation. *Nature*, **526**, 68–74.
59. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M. *et al.* (2016) Next-generation genotype imputation service and methods. *Nat. Genet.*, **48**, 1284–1287.
60. Loh, P.R., Danecek, P., Palamara, P.F., Fuchsberger, C., A Reshef, Y., Finucane, H.K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R. *et al.* (2016) Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet.*, **48**, 1443–1448.
61. Fuchsberger, C., Abecasis, G.R. and Hinds, D.A. (2015) Minimac2: faster genotype imputation. *Bioinformatics*, **31**, 782–784.
62. Siskos, A.P., Jain, P., Römisch-Margl, W., Bennett, M., Achaintre, D., Asad, Y., Marney, L., Richardson, L., Koulman, A., Griffin, J.L. *et al.* (2017) Interlaboratory reproducibility of a targeted metabolomics platform for analysis of human serum and plasma. *Anal. Chem.*, **89**, 656–665.
63. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G. *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, **42**, 565–569.
64. Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, **88**, 76–82.
65. Berisa, T. and Pickrell, J.K. (2016) Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, **32**, 283–285.
66. Gallois, A., Mefford, J., Ko, A., Vaysse, A., Julienne, H., Ala-Korpela, M., Laakso, M., Zaitlen, N., Pajukanta, P. and Aschard, H. (2019) A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. *Nat. Commun.*, **10**, 4788.
67. Castagné, R., Boulangé, C.L., Karaman, I., Campanella, G., Santos Ferreira, D.L., Kaluarachchi, M.R., Lehne, B., Moayyeri, A., Lewis, M.R., Spagou, K. *et al.* (2017) Improving visualization and interpretation of metabolome-wide association studies: an application in a population-based cohort using untargeted 1H NMR metabolic profiling. *J. Proteome Res.*, **16**, 3623–3633.
68. Li, M.X., Yeung, J.M., Cherny, S.S. and Sham, P.C. (2012) Evaluating the effective numbers of independent tests and significant *P*-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.*, **131**, 747–756.
69. Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J. *et al.* (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**, 707–713.
70. Balduzzi, S., Rücker, G. and Schwarzer, G. (2019) How to perform a meta-analysis with R: a practical tutorial. *Evid. Based Ment. Health*, **22**, 153–160.
71. Buniello, A., MacArthur, J., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
72. Suhre, K., Shin, S.Y., Petersen, A.K., Mohney, R.P., Meredith, D., Wägele, B., Altmaier, E., CARDIoGRAM, Deloukas, P., Erdmann, J. *et al.* (2011) Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, **477**, 54–62.
73. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
74. Köttgen, A., Raffler, J., Sekula, P. and Kastenmüller, G. (2018) Genome-wide association studies of metabolite concentrations (mGWAS): relevance for nephrology. *Semin. Nephrol.*, **38**, 151–174.
75. Montoliu, I., Genick, U., Ledda, M., Collino, S., Martin, F.P., le Coutre, J. and Rezzi, S. (2013) Current status on genome-metabolome-wide associations: an opportunity in nutrition research. *Genes Nutr.*, **8**, 19–27.