

RESEARCH

Open Access



# Comprehensive characterization of putative genetic influences on plasma metabolome in a pediatric cohort

In-Hee Lee<sup>1</sup>, Matthew Ryan Smith<sup>2,3</sup>, Azam Yazdani<sup>4</sup>, Sumiti Sandhu<sup>1</sup>, Douglas I. Walker<sup>5</sup>, Kenneth D. Mandl<sup>1,6,7</sup>, Dean P. Jones<sup>2</sup> and Sek Won Kong<sup>1,7\*</sup>

## Abstract

**Background:** The human exposome is composed of diverse metabolites and small chemical compounds originated from endogenous and exogenous sources, respectively. Genetic and environmental factors influence metabolite levels, while the extent of genetic contributions across metabolic pathways is not yet known. Untargeted profiling of human metabolome using high-resolution mass spectrometry (HRMS) combined with genome-wide genotyping allows comprehensive identification of genetically influenced metabolites. As such previous studies of adults discovered and replicated genotype–metabotype associations. However, these associations have not been characterized in children.

**Results:** We conducted the largest genome by metabolome-wide association study to date of children ( $N = 441$ ) using 619,688 common genetic variants and 14,342 features measured by HRMS. Narrow-sense heritability ( $h^2$ ) estimates of plasma metabolite concentrations using genomic relatedness matrix restricted maximum likelihood (GREML) method showed a bimodal distribution with high  $h^2$  ( $> 0.8$ ) for 15.9% of features and low  $h^2$  ( $< 0.2$ ) for most of features (62.0%). The features with high  $h^2$  were enriched for amino acid and nucleic acid metabolism, while carbohydrate and lipid concentrations showed low  $h^2$ . For each feature, a metabolite quantitative trait loci (mQTL) analysis was performed to identify genetic variants that were potentially associated with plasma levels. Fifty-four associations among 29 features and 43 genetic variants were identified at a genome-wide significance threshold  $p < 3.5 \times 10^{-12}$  ( $= 5 \times 10^{-8}/14,342$  features). Previously reported associations such as *UGT1A1* and bilirubin; *PYROXD2* and methyl lysine; and *ACADS* and butyrylcarnitine were successfully replicated in our pediatric cohort. We found potential candidates for novel associations including *CSMD1* and a monoethyl alcohol triglyceride ( $m/z$  781.7483, retention time (RT) 89.3 s); *CALN1* and Tridecanol ( $m/z$  283.2741, RT 27.6). A gene-level enrichment analysis using MAGMA revealed highly interconnected modules for dADP biosynthesis, sterol synthesis, and long-chain fatty acid transport in the gene-feature network.

**Conclusion:** Comprehensive profiling of plasma metabolome across age groups combined with genome-wide genotyping revealed a wide range of genetic influence on diverse chemical species and metabolic pathways. The developmental trajectory of a biological system is shaped by gene–environment interaction especially in early life.

\*Correspondence: [Sekwon.Kong@childrens.harvard.edu](mailto:Sekwon.Kong@childrens.harvard.edu)

<sup>1</sup> Computational Health Informatics Program, Boston Children's Hospital, 401 Park Drive, Boston, MA 02215, USA

Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Therefore, continuous efforts on generating metabolomics data in diverse human tissue types across age groups are required to understand gene–environment interaction toward healthy aging trajectories.

## Background

Metabolites are indicators and effectors of biological processes that are controlled by genetic and environmental factors; thus, metabolite levels reflect homeostatic and pathological status [1]. Furthermore, metabolites are the regulators of epigenetic modification, gene expression, and protein activity [2]. Comprehensive profiling of metabolites in an individual—a metabolome—can provide a snapshot of the host genetic makeup and its interaction with diet and environmental exposures. Genetic variants affect metabolite levels by regulating gene expression and/or changing protein function in metabolic pathways. A single base change in DNA sequence can have a strong impact on metabolite concentrations in patients with an inborn error of metabolism (IEM); examining perturbed metabolic pathways helps elucidate the molecular pathophysiology of human diseases and therapeutic targets [3]. Furthermore, metabolite levels can be used as endophenotype that mediates genetic risks for common diseases [4] and predict inter-individual differences in drug response that are associated with pharmacogenetic variations [5].

Endogenous metabolite levels are highly heritable [6, 7]; genetically influenced metabolomes (GIMs) can be discerned with targeted or untargeted metabolomic profiling combined with genome-wide genotyping [8–11]. Using a genome-wide association study (GWAS) framework, a recent meta-analysis confirmed the reported associations between single nucleotide variants (SNVs) and metabolomes from independent cohort studies of adults [12]. The human metabolome is likely to contain hundreds of thousands of chemicals [13]; however, the breadth of chemical space coverage was limited up to few hundreds in previous studies. Therefore, the full extent of GIMs in the human metabolome is not yet known.

Untargeted high-resolution metabolomics (HRM) platforms enable quantitative measurements for tens of thousands of features with mass-to-charge ratios ( $m/z$ ) with retention times (in seconds; RT) from endogenous and exogenous origins in biospecimens [3, 14]. A liquid chromatography high-resolution mass spectrometry (LC-HRMS) platform combined with genome-wide genotyping can provide a comprehensive snapshot of GIMs. With this platform, we previously evaluated the coverage of chemical space and constructed a global correlation map of the human plasma metabolome,

measuring metabolites produced by the gut microbiome and xenobiotics and finding that many metabolites were associated with demographic characteristics [15].

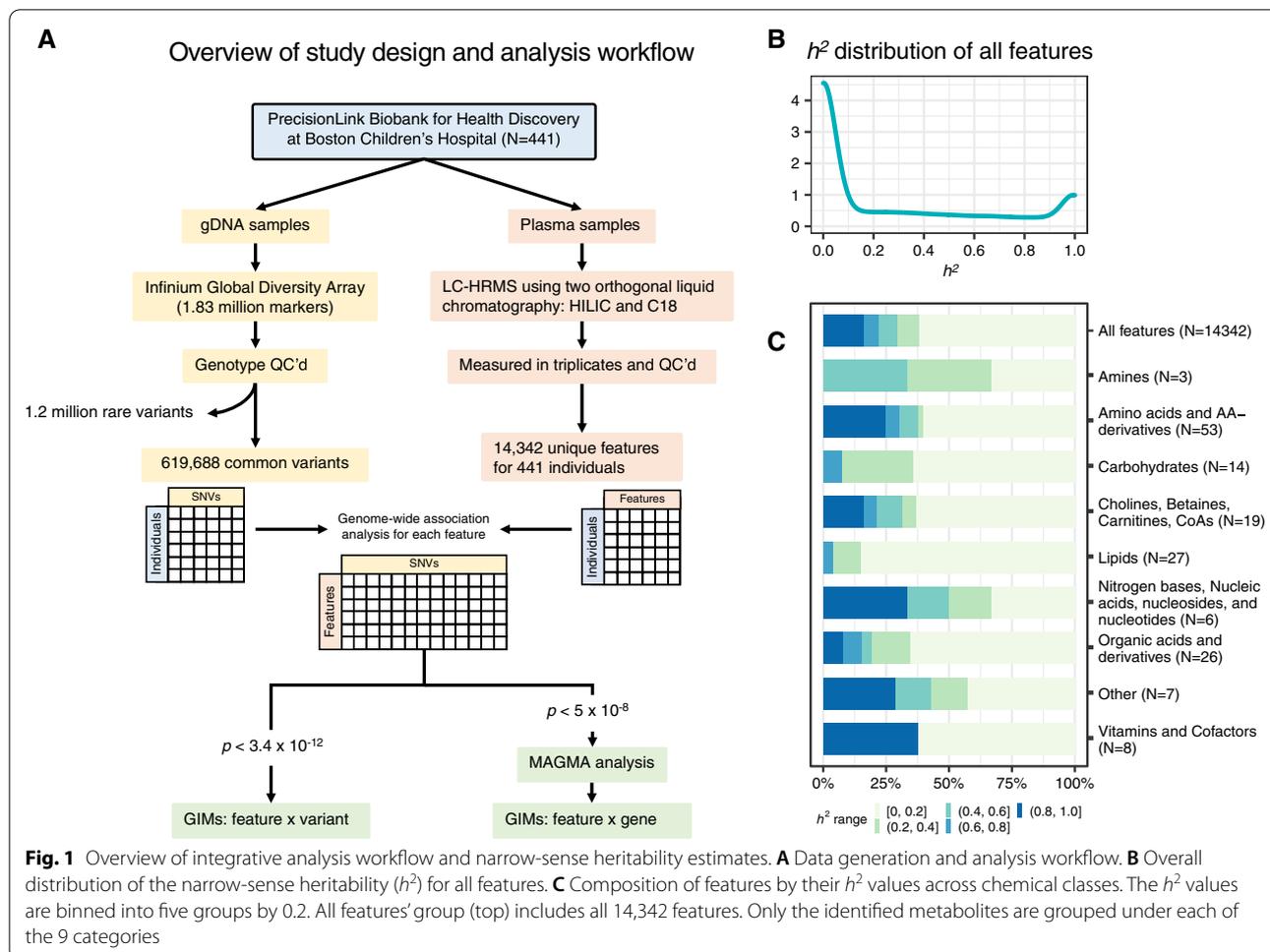
In the current study, we deployed an untargeted LC-HRMS platform to analyze plasma samples collected from a pediatric cohort ( $N=441$ ) for which common genetic variants were characterized using genome-wide genotyping microarray. By interrogating 14,342 features and 619,688 common genetic variants, we first estimated narrow-sense heritability for all features. Further we performed a genome by metabolome-wide association study (GxMWAS) to replicate previously reported GIMs in our pediatric cohort and to discover novel GIM candidates after controlling for age, gender, and global genetic ancestry. Our results clearly revealed the extent of genetic contribution to metabolome across age groups for a wide range of chemical species.

## Results

### Chemical space coverage and the features associated with demographic factors

Samples, data generation, and analysis workflow are depicted in Fig. 1A. All LC-HRMS analysis was performed in triplicate using a dual column chromatography scheme that included hydrophilic interaction liquid chromatography (HILIC) and reversed phase liquid chromatography (RPLC; C18) columns to maximize the chemical space coverage. A total of 14,342 features were quantitatively measured with accurate  $m/z$  and RT for HILIC and C18 columns ( $N=8739$  and 5603, respectively). All features were subjected for metabolite annotation using xMSannotator [16], and high or medium confidence annotations at a 5-ppm mass tolerance window were used to reduce incorrect annotations (see Methods). A subset of features was experimentally identified using LC-MS/MS with authentic standards ( $N=97$  for HILIC and 69 for C18, Additional file 2: Table S1) [17]. The demographic characteristics of our cohort are summarized in Additional file 3: Table S2. To prioritize metabolites that were associated with demographic variables, we used a generalized linear model while controlling for batch effect and global genetic ancestry using the top ten eigenvectors—hereafter referred to as principal components (PCs) 1–10—from EIGENSTRAT analysis of 619,688 common variants (see Additional file 1: Fig. S1 for population stratification with the first two PCs).

A total of 338 features were associated with age and gender (false discovery rate (FDR) $<0.05$ ,  $N=120$  and



218, respectively), of which 19 were among the identified metabolites using reference standards ( $N=9$  and 10, respectively; Additional file 4: Table S3). Among the age-correlated features, urate and creatinine are known to be correlated with age [18, 19] and we found novel associations including cystine, trimethyllysine, quinic acid, butyrobetaine, arachidonic acid, and retinol. Between females and males, amino acid derivatives (hydroxyproline, hydroxylysine, dimethylarginine), carnitine, cholines ((lysoPC(18:0) and lysoPC(18:1)), microbial product (valerobetaine), and a product of urea cycle (fumaric acid) were significantly different among the identified metabolites.

As of the impact of global genetic ancestry on individual's metabolotype, PC1 that distinguishes African descents from non-African descents was correlated with 141 features including 8 identified metabolites: amino acid derivatives (citrulline, indoleacetate, and kynurenine), carbohydrate (arabinose), vitamins (thiamine (B1), nicotinamide (B3)), and cholines (lysoPE(18:0), lysoPE(20:3)) (FDR < 0.05) (Additional file 4: Table S3). Except for PC4

and PC7, the other PCs were not correlated with any of the identified metabolites. PC4 was correlated with subgroups of European descent and PC7 distinguished two individuals of European descents. Nonetheless, the features that were significantly correlated with PCs may reflect combined effect of genetic and environmental factors, such as diet and lifestyle. We did not observe any bias in  $m/z$  and RT for the significant features associated with demographic variables—age, gender, and global genetic ancestry (Additional file 1: Figs. S2–S4).

Next, we performed pathway enrichment analysis using Mummichog [20] with statistical scores obtained from univariate analyses as described above. Pathways were selected for adjusted  $p$  value < 0.01 with  $\geq 5$  identified metabolites or uniquely annotated features (i.e., level 2 according to Schymanski et al. [21], see Methods) overlapping for each pathway (Additional file 5: Table S4). Glutamate metabolism pathway including five identified metabolites—glutamine, glutamate, carbamoyl phosphate, 2-oxoglutarate, and *N*-methylglycine—was significant for age (adjusted  $p$  value 0.00495).

Tricarboxylic acid (TCA) cycle pathway was enriched with differentially detected metabolites between males and females. Ascorbate and aldarate metabolism and pentose phosphate pathways were significant for PC1 and PC4.

#### Narrow-sense heritability estimation of feature levels

To capture the genetic influence on the variance of metabolite levels across individuals, we estimated a narrow-sense heritability ( $h^2$ ) of feature levels using genomic relatedness matrix (GRM) restricted maximum likelihood (GREML) method implemented in genome-wide complex trait analysis (GCTA) [22]. We found a wide range of  $h^2$  estimates with a bimodal distribution having 15.9% of features with high  $h^2$  ( $>0.8$ ) and 62.0% with low  $h^2$  ( $<0.2$ ) (Fig. 1B). Next, we checked  $h^2$  distributions in each chemical species with the identified metabolites. Carbohydrates and lipids had low  $h^2$  overall with 7.1% and 3.7% had  $h^2$  between 0.6–0.8, respectively. No feature was high  $h^2$  ( $>0.8$ ) for these chemical classes. In comparison, large proportions of amino acids and derivatives and nucleic acids

had high  $h^2$  ( $>0.8$ ) (24.5% and 33.3%, respectively) (Fig. 1C).

#### Genome by metabolome-wide association analysis

We calculated age and gender corrected feature intensities and included the top 10 PCs as covariates to perform GxMWAS (see Methods) discovering 54 associations among 29 features and 43 common genetic variants at the threshold of  $p < 3.5 \times 10^{-12}$  ( $= 5 \times 10^{-8}/14,342$  features) (Table 1 for selected associations; Fig. 2 for associations involving identified metabolites with  $p < 5 \times 10^{-8}$ ; Additional file 6: Table S5 for full list of associations at genome-wide significance level,  $p < 5 \times 10^{-8}$ ; Additional file 7: Table S6 for full list of high and medium confidence annotations at 5-ppm mass tolerance). On average, a genetic variant was associated with  $1.3 \pm 0.79$  features (range 1–4) and a feature was associated with the median of one variant (range 1–5). Most variants associated with feature levels were intronic (25 of 43, 58.1%) and three (7.0%) were in protein coding exons.

A previously reported GIM between *PYROXD2* and Ne,Ne dimethyllysine ( $m/z$  175.1442, RT 104.5) was successfully replicated in our pediatric cohort, which was indeed the strongest association ( $p = 6.2 \times 10^{-21}$ ) in our

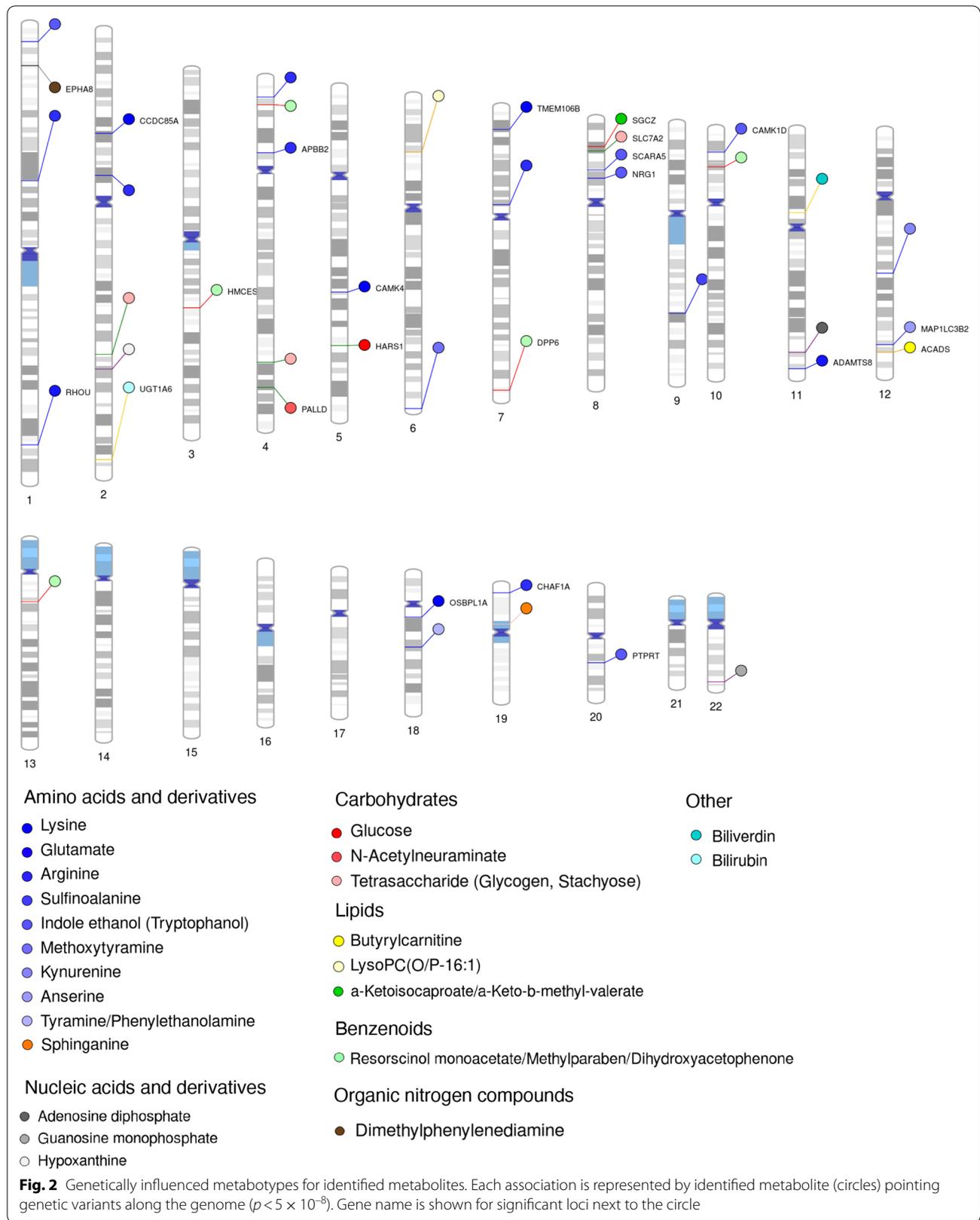
**Table 1** Genetically influenced metabolites. Significant genetic variant–metabolite associations at a genome-wide significance of  $p < 3.5 \times 10^{-12}$  ( $= 5 \times 10^{-8}/14,342$  features) are shown.

Feature ( $m/z$ , RT)	Lead SNP position*	EA, OA	EAF	Beta	Standard error	$p$ value	Candidate genes
$m/z$ 394.8915, RT 68.1	chr9:72,902,395	G, A	0.14	− 0.49	0.063	4.94E−14	<i>ALDH1A1</i>
$m/z$ 640.3195, RT 278.2	chr20:38,052,517	G, T	0.08	− 1.23	0.167	7.35E−13	<i>RPRD1B</i>
	chr10:66,637,441	A, G	0.07	− 0.69	0.090	1.56E−13	<i>CTNNA3</i>
$m/z$ 709.0644, RT 190.3	chr8:47,661,411	A, G	0.07	0.93	0.128	1.80E−12	<i>SPIDR</i>
$m/z$ 101.5811, RT 100.6	chr10:98,377,943	T, C	0.46	− 0.61	0.065	4.82E−19	<i>PYROXD2**</i>
$m/z$ 161.1285, RT 106.1 (N6-Methyl-L-lysine)	chr10:98,377,943	T, C	0.46	− 0.49	0.061	6.98E−15	<i>PYROXD2**</i>
$m/z$ 162.1321, RT 106.9 (N6-Methyl-L-lysine)	chr10:98,377,943	T, C	0.46	− 0.54	0.064	3.32E−16	<i>PYROXD2**</i>
$m/z$ 175.1442, RT 104.5 (Ne,Ne dimethyllysine)	chr10:98,377,943	T, C	0.46	− 0.55	0.056	6.20E−21	<i>PYROXD2**</i>
$m/z$ 188.9574, RT 44.2	chr12:66,599,541	C, T	0.06	− 1.22	0.164	5.69E−13	<i>GRIP1</i>
	chr15:88,215,350	A, C	0.15	− 0.84	0.116	2.85E−12	<i>NTRK3</i>
	chr21:14,901,801	G, C	0.13	− 0.85	0.117	1.94E−12	<i>NRIP1</i>
$m/z$ 220.1777, RT 24	chr16:7,273,260	C, T	0.08	− 1.03	0.137	3.45E−13	<i>RBFOX1</i>
$m/z$ 269.0023, RT 125.5 (fenson)	chr4:40,241,127	G, C	0.06	0.47	0.061	9.20E−14	<i>RHOH</i>
$m/z$ 283.2741, RT 27.6	chr7:12,354,019	G, A	0.12	0.27	0.037	1.55E−12	<i>VWDE</i>
	chr7:72,362,534	A, G	0.07	0.35	0.047	3.02E−13	<i>CALN1</i>
$m/z$ 300.2167, RT 37.1	chr2:210,210,185	C, T	0.31	0.68	0.077	5.62E−17	<i>ACADL**</i>
$m/z$ 781.7483, RT 89.3	chr8:3,674,391	C, T	0.11	0.73	0.093	2.29E−14	<i>CSMD1</i>

The features with identified metabolites or high/medium confidence annotations are listed with chemical names. EA, effect allele; OA, other allele; EAF, effect allele frequency;  $m/z$ , mass-to-charge ratio; and RT, retention time in seconds

\*Represented as chromosome:position based on the human reference genome GRCh38

\*\*Previously reported genetically influenced metabolotypes in adults



results (Fig. 3A). This association has been independently replicated by the studies using urine, plasma, and cerebrospinal fluid (CSF) samples [4, 23–25]. The *ACADL* gene encodes acyl-CoA dehydrogenase long chain (ACADL) that is a subunit of the four enzymes involved in the initial step of mitochondrial beta-oxidation of straight-chain fatty acid. One missense and two intronic variants were significantly associated with a feature ( $m/z$  300.2167, RT 37.1) that was annotated as menthol propylene glycol carbonate (HMDB identifier: HMDB0039785) at 5-ppm tolerance (i.e., level 5 identification by criteria of Schymanski et al. [21]) (Fig. 3B). The association between nonanoyl carnitine and *ACADL* ( $p=2.3 \times 10^{-9}$ ) did not pass our stringent statistical significance threshold, while this association has been previously reported.

Lipid species had low  $h^2$  overall; however, we found novel GIM candidates for the features that uniquely matched to lipids based on the level 5 identification criteria of Schymanski et al. [21]. Among novel GIMs discovered in our cohort, Tridecanol ( $m/z$  283.2741, RT 27.6) level was associated with intronic SNVs in *CALN1* and *VWDE* ( $p$  values  $3.02 \times 10^{-13}$  and  $1.55 \times 10^{-12}$ , respectively) (Fig. 3C, D). A monostearyl alcohol triglyceride ( $m/z$  781.7483, RT 89.3) was associated with an intronic SNV (rs2624100) in the *CSMD1* (CUB and Sushi Multiple Domains 1) gene ( $p=2.29 \times 10^{-14}$ ).

Some feature levels were associated with multiple genomic loci. For instance, a feature ( $m/z$  188.9574, RT 44.2) was associated with variants in three genes—*GRIP1*, *NTRK3*, and *NRIP1*—on chromosomes 12, 15, and 21 ( $p$  values  $5.69 \times 10^{-13}$ ,  $2.85 \times 10^{-12}$  and  $1.94 \times 10^{-12}$ , respectively). The *GRIP1* (glutamate receptor-interacting protein 1) gene on chromosome 12q14.3 is involved in synapse formation [26]. The *NTRK3* (neurotrophic receptor tyrosine kinase 3) gene on chromosome 15q25.3 encodes a receptor tyrosine kinase that binds to its ligand neurotrophin-3 and plays a role in nervous system development. AF127577.4 is a long non-coding RNA on chromosome 21q11.2 and 5'-end overlaps with the *NRIP1* gene. Nuclear receptor-interacting protein 1 (NRIP1) is a nuclear protein that interacts with the hormone-dependent nuclear receptors and expressed in neuronal and glial cells [27].

For the identified metabolites, we found several GIM candidates including those previously reported at a less

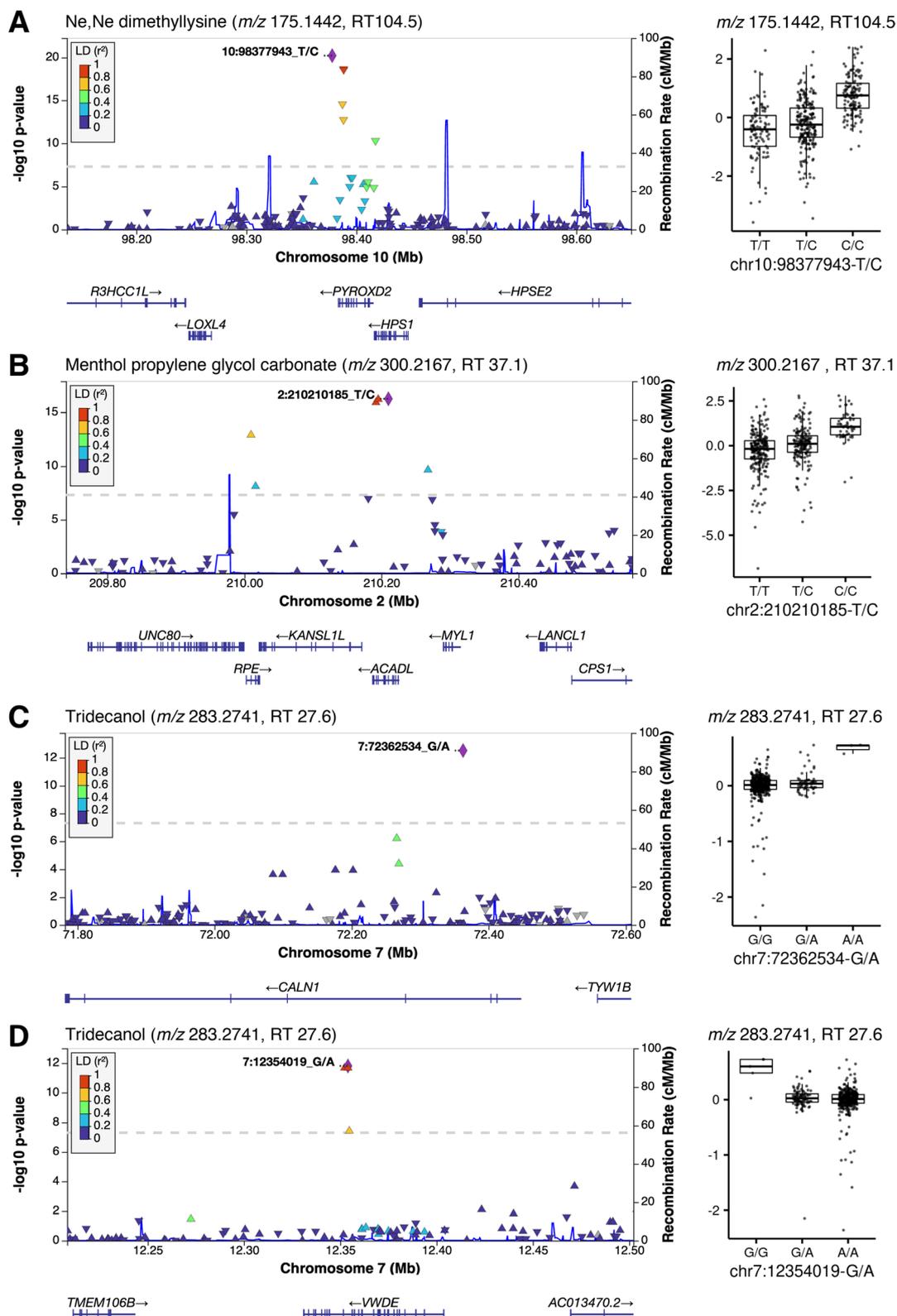
stringent threshold of  $p < 5 \times 10^{-8}$ . For amino acids and its derivatives, we found GIM candidates for arginine, glutamate, lysine, and sulfinoalanine. Arginine and lysine levels were associated with multiple genetic loci in different genes. Six loci including intronic variants in the *APBB2* and *CHAF1A* genes were significantly associated with plasma arginine level. Lysine level was associated with four loci including intronic variants in the *CCDC85A*, *CAMK4*, *TMEM106B*, and *OSBPL1A* genes. Glutamate is the most abundant excitatory neurotransmitter in CNS, and its plasma level was significantly associated with genetic variants in the *ADAMTS8* and *RHOA* genes. For carbohydrates, glucose level was significant for an intronic variant (rs11954514) in the *HARS1* (Histidyl-tRNA synthetase 1) gene that is a disease-causing gene for Usher syndrome type 3b (MIM ID: 614504) [28]. An intronic variant in the *SLC7A2* gene was associated with plasma glycogen level. The *SLC7A2* gene encodes a cationic amino acid transporter and is implicated in arginine metabolism. Slc7a2 knockout mice had 20% higher blood glucose compared to wild-type mice [29]. *N*-Acetylneuraminic acid (Neu5Ac) is a sialic acid found in cell membrane. In neuronal cells, Neu5Ac residues are found in membrane bound glycoproteins, i.e., gangliosides. Neu5Ac interacts with bacterial and viral pathogens in diverse cell types. Multiple SNVs in the *PALLD* gene were associated with plasma Neu5Ac levels. Palladin, encoded by the *PALLD* gene, is a cytoskeletal protein found in actin filaments. Among lipid species, butyrylcarnitine-*ACADS* association was notable, which has been reported in independent studies. All associations significant at a threshold of  $p < 5 \times 10^{-8}$ , along with  $m/z$ , RT and annotation, are listed in Additional file 6: Table S5. The full list of annotations with high or medium confidence by xMSannotator is shown in Additional file 7: Table S6.

### Gene-level enrichment with the variants associated with features

Next, we performed a gene-level enrichment analysis for each feature with its GWAS summary statistics using Multi-marker Analysis of GenoMic Annotation (MAGMA) [30]. A total of 572 genes were enriched with variants associated with 217 features at FDR < 0.01 (Additional file 8: Table S7). The most significant association between gene and feature was found for the *SKIDA1* gene

(See figure on next page.)

**Fig. 3** Regional plots for significant genotype–metabotype associations. The genomic coordinates (x-axis) are based on GRCh38. The variant with the strongest  $p$  value is highlighted with its coordinate, reference, and variant alleles. The boxplot (shown next to the regional plot) shows the distribution of normalized feature intensity by genotype for the strongest variant. Feature annotation is from xMSannotator with high and medium confidence or from HMDB with exact mass at 5-ppm tolerance (level 5 annotation according to Schymanski et al.) **A** significant loci for Ne,Ne dimethyllysine ( $m/z$  175.1442, RT 104.5) around *PYROXD2* on chromosome 10. **B** loci for Menthol propylene glycol carbonate ( $m/z$  300.2167, RT 37.1) on *ACADL* on chromosome 2. **C, D** shows the two separate loci on chromosome 7 for Tridecanol ( $m/z$  283.2741, RT 27.6), located in two genes *CALN1* and *VWDE*, respectively



**Fig. 3** (See legend on previous page.)

and fenson ( $m/z$  269.0023, RT 125.5) ( $FDR$   $6.64 \times 10^{-8}$ ). Among the identified metabolites, the *HLA-C* gene was associated with arginine level ( $FDR$  0.0084). Bilirubin was significantly associated with UDP-glucuronosyltransferase (UDP1A) isoforms, which has been replicated by independent studies using different metabolomics platforms [4, 7, 31, 32]. Of the ten features associated with *UGT1A* isoforms, six were identified and/or annotated as bilirubin, while four features were not. Correlation structure of these features showed that the unmatched features could represent other chemicals than bilirubin (Additional file 1: Fig. S5) as *UGT1As* are the enzymes of the glucuronidation pathway processing small lipophilic molecules such as steroids, bilirubin, hormones, and drugs into water-soluble and excretable metabolites.

We created a network of gene-feature associations with MAGMA results to check the interconnectivity (Fig. 4A and Additional file 1: Fig. S6). The average number of neighbors was 2.0 and a one-to-one association was found for 67 gene-feature pairs. The largest subgraph had 13 features and 72 genes. We checked enriched gene ontology terms for the genes in each subgraph with 6 or more genes. Six subgraphs were enriched with one or more of the Gene Ontology (GO) biological pathway terms (hypergeometric test,  $FDR < 0.05$ ). A subgraph with 4 features and 19 genes was enriched with the genes involved in purine metabolism such as dADP (deoxyadenosine diphosphate) biosynthetic process (hypergeometric test,  $FDR$  0.018, Fig. 4B). A feature ( $m/z$  467.256, RT 278.1) was associated with the variants in 12 genes that were enriched in nucleosome assembly (hypergeometric test,  $FDR$  0.013, Fig. 4C). Seven genes associated with a feature ( $m/z$  522.734, RT 43.2) were enriched for plasma membrane long-chain fatty acid transport and ketone body biosynthetic process (hypergeometric test,  $FDR$  0.036 and 0.036, respectively, Fig. 4D).

#### Revealing the underlying network modules with genetically influenced metabolotypes

Focusing on the 29 GIMs with  $p < 3.5 \times 10^{-12}$ , we identified a GIM-causal network at the type I error rate of 5%. Briefly, causal networks are based on conditional (in)dependency established in the principles of Mendelian

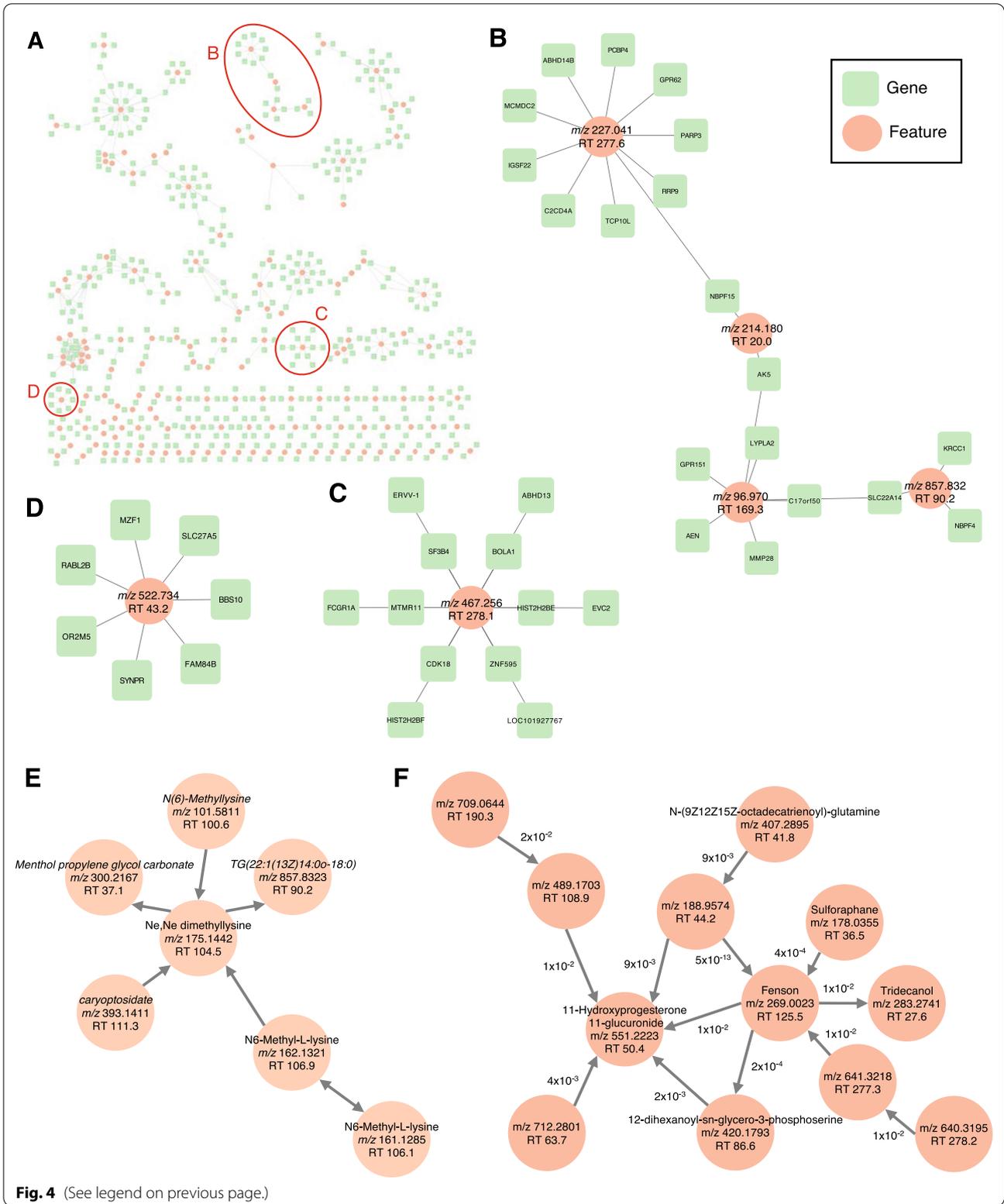
randomization [33]. We found two disjoint modules of interconnected GIMs. The modules comprised 7 and 12 GIMs with directed connections pointing a prediction target in each module (Fig. 4E, F). A prediction target captures the effect from multiple other GIMs in the module, so its concentration levels can be representative of the module [33]. In the module with seven features (Fig. 4E), menthol propylene glycol carbonate ( $m/z$  300.2167, RT 37.1) showed a significant connectivity/dependency ( $p = 2 \times 10^{-4}$ ) with Ne,Ne dimethyllysine ( $m/z$  175.1442, RT 104.5). In the module with 12 features, the connection between an unannotated feature (hilic\_1914;  $m/z$  188.9574, RT 44.2) and fenson ( $m/z$  269.0023, RT 125.5) was one of the most significant connectivities ( $p = 5 \times 10^{-13}$ ) (Fig. 4F).

#### Discussion

In the human body, metabolites have diverse biological functions such as the regulation of epigenome, transcription, translation, protein function, and signal transduction. Further, metabolites are indicators of gene-environment interactions. Previous studies reported genotype-metabotype associations to highlight the genetic contribution to metabolite concentrations; however, the extent of GIMs in the human exposome has not yet been known partly due to the limited coverages of chemical space in previous studies. In the current study, we used an untargeted metabolomics platform that provided a snapshot of thousands of metabolites. Age is one of the key demographic factors contributing the development of exposome. In our previous study, age was correlated with 40.4% of metabolites measured in plasma samples [15]. Interestingly, age-correlated features were enriched for xenobiotics but depleted for nucleic acids and its derivatives. In the current study, we aimed to estimate the contribution of genetic factors to plasma concentrations of diverse chemical species for generally healthy individuals from 5 months to 60 years of age, while most of our study cohort (82.3%) were 20 years old or younger. The genetic contribution to metabolite levels varied across chemical classes. Narrow-sense heritability ( $h^2$ ) was small ( $< 0.2$ ) for most features suggesting that environment factors might contribute more substantially to the human

(See figure on next page.)

**Fig. 4** Gene-feature network. The gene-feature network (A) with its modules enriched with the Gene Ontology biological pathway terms (B–D) and causal networks composed of 29 GIMs based on conditional (in)dependency augmented with principles of Mendelian randomization (E–F). **A** Gene-feature network is constructed using MAGMA results (false discovery rate  $< 0.01$ ). One-to-one associations are found for 67 gene-feature pairs; however, the other genes and features are interconnected to form modules. **B** A module with four features and the genes enriched with dADP (deoxyadenosine diphosphate) biosynthetic process. **C** Twelve genes associated with a feature ( $m/z$  467.256, RT 278.1) are over-represented for nucleosome assembly. **D** Another feature ( $m/z$  522.734, RT 43.2) is associated with the seven genes functioning plasma membrane long-chain fatty acid transport. **E, F** Two modules identified from the causal network analysis of 29 GIMs. Annotated metabolites in each module provide information about the unannotated metabolites in the same module specially if they are highly connected (e.g., Ne,Ne dimethyllysine ( $m/z$  175.144, RT 104.5) in (E). In the module in (F), a strong dependency ( $p = 5 \times 10^{-13}$ ) between an unannotated feature ( $m/z$  188.9574, RT 44.2) and fenson ( $m/z$  269.0023, RT 125.5) may provide annotation information for the feature



**Fig. 4** (See legend on previous page.)

exposome than genetic factors, overall. For instance, less than 10% of carbohydrates and lipids—likely linked with diet and lifestyle—had  $h^2 > 0.6$  and none of these species had  $h^2 > 0.8$ . In contrast, the genetic influence seemed to be larger for some of amino acids and nucleic acids.

We successfully replicated the previous findings (e.g., *UGT1A1* and bilirubin; *PYROXD2* and methyllysines; and *ACADS* and butyrylcarnitine) that were described in adults. The strongest associations were found between lysine derivatives (i.e., methyllysines) and the genetic variants in *PYROXD2*. This association has been replicated in multiple tissue types. Of note, Panyard and colleagues used 672 CSF samples to perform GxMWAS for 378 metabolites of which the most significant association was found between N6-methyllysine and rs2147896 in *PYROXD2* [4]. The same association was also discovered in our analysis of plasma metabolome. Indeed, ten out of top 16 GxMWAS results from the CSF study were also replicated in our study. These results suggest that, for IEMs caused by rare genetic variants with large effect, surrogate tissues such as blood can be used to understand the molecular pathophysiology for which the primary affected tissues are difficult to biopsy.

Environmental contribution to plasma levels of carbohydrates and lipids was larger than genetic factors; however, we found strong GIM candidates for these chemical species as well. Tridecanol ( $m/z$  283.2741, RT 27.6) and intronic SNVs in the *CALNI* and *VWDE* genes were significantly associated. The *CALNI* (calneuron 1) is a candidate gene for schizophrenia and intelligence, which was discovered in independent GWASs. The function of *VWDE* (von Willebrand factor D and EGF domains) gene product is not yet known while previous GWASs discovered the risk alleles for frontotemporal dementia, depression, and coronary artery diseases in this gene.

A monostearyl alcohol triglyceride ( $m/z$  781.7483, RT 89.3) level was associated with the *CSMD1* gene that encodes a transmembrane protein CSMD1. CSMD1 is an inhibitor of the complement component 3 (C3) convertases that produce C3b and C3a-desArg. Of note, the active C3 fragment, C3a-desArg has insulin-like effects and is involved in triglyceride metabolism [34]. C3 knockout mice demonstrate reduced body weight and fat mass [35]. *CSMD1* is reported as a candidate gene for schizophrenia [36].

For carbohydrates, we found a novel association between plasma glucose level and an intronic variant in the *HARS1* gene encoding HARSL. Loss of function mutations in *HARS1* cause peripheral neuropathies that is likely caused by reduced translation efficiency [37, 38]. In patients with diabetes mellitus, peripheral neuropathy is a common complication observed in 6–51% [39].

Reported GIMs are mostly one-to-one relationships between a metabolite level and a genetic variant (or multiple variants in linkage disequilibrium); however, we found subgraphs with interconnected genes and features in our network analysis. Modularity of gene–metabolite network showed the complexity of genetic contributions to metabolic pathways. Using a causal inferential network analysis method, we extracted direct and indirect contributions of genetic factors to plasma metabolite levels. Of note, a triglyceride and methyl lysine showed a significant connectivity, while the role of *PYROXD2* in triglyceride metabolism is not yet known.

Our study had some limitations. Firstly, our sample size ( $N=441$ ) was not sufficient to replicate some of previous findings at the stringent statistical threshold of  $p < 3.5 \times 10^{-12}$  ( $= 5 \times 10^{-8}/14,342$  features). For instance, the *ACADS* gene encodes short-chain acyl-CoA dehydrogenase (SCAD) and genetic variants in this gene are associated with mitochondrial fatty acid oxidation function. More than 55 mutations in the *ACADS* gene were reported in patients with SCAD deficiency (SCADD) with increased plasma concentration of butyrylcarnitine [40]. In our analysis, butyrylcarnitine level was associated with eight SNPs in upstream, coding, intronic variants of the *ACADS* gene at  $p < 5 \times 10^{-8}$ ; however, none of these loci passed the threshold  $p < 3.5 \times 10^{-12}$ . Secondly, annotation of features measured by LC-HRMS was not complete. A total of 14,342 features were identified with the combination of unique  $m/z$ , RT, and peak intensity in our plasma samples. To reduce false annotations, we highlighted significant associations among 891 features with high confidence annotations and 2431 features with medium confidence annotations according to xMSannotator in addition to 166 identified metabolites confirmed with authentic standards. xMSannotator integrates correlation structure of measured features in a dataset with multistage clustering and refinement using diverse chemical annotation databases (e.g., KEGG, HMDB, T3DB, and Lipid Maps) to reduce false annotations. Therefore, some interesting features with significant associations require further annotation and identification of chemical compounds. For instance, a feature ( $m/z$  283.2741, RT 27.6) that was associated with the *CALNI* and *VWDE* genes was not annotated by xMSannotator with high or medium confidence nor identified with authentic standard. This feature was uniquely mapped to Tridecanol (HMDB identifier: HMDB0013316) with exact mass ( $m/z$ ) at 1.02 ppm tolerance; however, further investigation is required to confirm structure. Lastly, children enrolled for the current study included patients with rare genetic disorder (e.g., cystic fibrosis) and common diseases (e.g., epilepsy, anemia, and diabetes mellitus).

Medication history at the time of blood draw for metabolomics profiling was not used.

## Conclusions

In summary, we performed the most comprehensive analysis to date of the plasma metabolome in a pediatric cohort. Our unbiased profiling methods revealed a wide range of genetic contributions to metabolites for different chemical species as well as complex gene–metabolite associations. The developmental trajectory of a biological system is shaped by gene–environment interactions especially in early life course. Environmental exposures of endogenous and exogenous origins modulate health and aging trajectory of an individual, while genetic factors modify the environmental effect. Continuous efforts on the chemical identification of significant features in HRMS experiments and generating paired genomic and metabolomic data in diverse human tissue types across age groups are required to understand gene–environment interaction toward healthy aging trajectories.

## Methods

### Subjects

Individuals were enrolled in the PrecisionLink Biobank for Health Discovery at Boston Children's Hospital (BCH) from January 2016 to November 2019 [41]. The participants are enrolled throughout the hospital, across diverse clinical settings. Informed consent is obtained from all participants enrolling in the Biobank and provides permission to: (1) access electronic health record (EHR) data for research; (2) collect and use of residual specimens produced as by-products of routine care; and (3) share de-identified data and specimens outside of the institution. We collected 441 plasma samples from 230 females and 211 males with mean ages 15.7 and 14.3 years old, respectively (ranges from 4.8 months to 60.1 years) (Additional file 3: Table S2). The International Classification of Diseases (versions 9 and 10) and SNOMED CT codes were collected for participants from the BCH Cerner EHR database. To comply with the Health Insurance Portability and Accountability Act rules for protected health information, medical record identifiers and personal information were removed from the EHR extracts and universal unique identifiers (UIDs) were assigned to everyone. All analyses were performed with UIDs, age at blood collection, gender information, and sample identifiers for plasma and DNA samples, which were provided by the BCH Biobank. The study was reviewed and approved by the BCH Institutional Review Board.

### Plasma and genomic DNA samples

Genomic DNA (gDNA) and plasma samples were obtained from the PrecisionLink Biobank at Boston Children's Hospital (BCH). Participants are given the opportunity to also consent to collection of a 4 mL blood sample for research use, from which DNA and plasma aliquots are obtained. In conjunction with other scheduled clinical laboratories, the whole blood is collected from participants in EDTA treated tubes. When received in the Biobank Core Lab, the blood is centrifuged at  $2000 \times g$  for 10 min at room temperature. Plasma is then aliquoted into 0.5 mL microcentrifuge tubes and stored at  $-80^\circ\text{C}$  in the Biobank Core Lab facility until requested. gDNA is extracted from the whole blood using Genra Puregene Extraction Kit (Qiagen Sciences Inc, Germantown, MD) or Chemagic B5k Extraction Kits (PerkinElmer, Waltham, MA) resulting in two 0.225 mL aliquots. DNA samples are stored at  $-80^\circ\text{C}$  until requested for research use at which point they undergo normalization and QC. The PrecisionLink Biobank initiative is approved by the BCH Institutional Review Board (protocol number—P00000159).

### High-resolution metabolomics profiling of plasma samples

Plasma samples were thawed and aliquoted at BCH Biobank for shipping to Emory University in dry ice package. We randomized plasma samples to each batch to balance age and sex between batches of HRMS profiling. Plasma samples were extracted by treating 50  $\mu\text{L}$  aliquots with acetonitrile containing 14 stable isotope internal standards with 100  $\mu\text{L}$  to precipitate proteins. Samples were then equilibrated on ice for 30 min and centrifuged for 10 min at 13,400 rpm at  $4^\circ\text{C}$ . The supernatant was transferred to autosampler vials and kept in a refrigerated autosampler until analysis. Each extract was analyzed in triplicate using a dual column chromatography scheme that includes hydrophilic interaction liquid chromatography (HILIC; XBridge BEH Amide XP HILIC column; Waters, Waltham, MA;  $50 \times 2.1$  mm,  $2.5 \mu\text{m}$ ) and reversed phase liquid chromatography (RPLC; C18 column; Higgins Analytical, Mountain View, CA;  $50 \times 2.1$  mm,  $2.6 \mu\text{m}$ ). The chromatography was coupled with HRMS in positive (HILIC) and negative electrospray ionization (ESI) modes (RPLC) that enabled an increased coverage of the plasma metabolome. Mass spectral data were collected with a 5-min mobile phase gradient on a Thermo Q-Exactive HF high-resolution mass spectrometer (Thermo Fisher, San Diego, CA) set to collect data from  $m/z$  of 85 to 1275 at a resolution of 120,000 [17]. Raw data were converted to mzXML using ProteoWizard, and data preprocessing, which included peak detection, noise filtering, peak quantification and alignment, averaging signals of triplicates, peak-matching, and

batch effect correction, was completed using apLCMS and xMSanalyzer [42, 43]. Raw and processed metabolomic datasets are deposited in the Metabolomics Workbench [44] (Study\_ID: ST002331, Project\_ID: PR001495, <https://doi.org/10.21228/M8GM6Q>).

#### Metabolite identification and annotation

All features were annotated using xMSannotator [16], which utilizes a multistage clustering algorithm in order to provide confidence scores for annotated metabolites. Subsequent identities of features were compared to our confirmed library of identified metabolites [17] which utilized co-elution relative to authentic standards and ion dissociation mass spectrometry (level 1 identification by criteria of Schymanski et al. [21]) with a tolerance of 5 ppm and 30 s. The remaining annotations with high or medium confidence provided by xMSannotator stages 4 and 5 possessed a M – H/M + H adduct, detected in the negative/positive mode, respectively, and were made using the KEGG (Kyoto Encyclopedia of Genes and Genomes) [45]; HMDB (Human Metabolome Database) [46]; T3DB (Toxin and Toxin Target Database) [47]; and Lipid Maps [48] databases at 5 ppm tolerance [17]. For the significant features without high or medium confidence annotations, we added annotations that had exact mass at a 5-ppm tolerance in the HMDB database (level 5 identification by the criteria of Schymanski et al. [21]).

#### Genome-wide genotyping

Illumina Global Diversity Array (GDA) was used to genotype 1.83 million variants. This platform is used for the All of US program and includes 0.61 million common variants and 0.68 million rare variants covering diverse race and ethnic groups. More importantly, 0.53 million known clinically implicated variants are included such that GDA can be used for monitoring genetic risks for various common diseases, pharmacogenomics, and frequently mutated genes in rare disorders and cancer. Frozen gDNA samples were thawed for aliquoting and shipping to the Partners Center for Personalized Genomic Medicine (PCPGM) for genotyping using GDAs. All samples were quantitated using picogreen to assess the concentration of double-stranded DNA. QC of the microarrays was carried out by inspecting the Controls Dashboard within GenomeStudio analysis software (Illumina, San Diego, CA). These controls monitor internal spike-in probes at various points of the process and allow the QC of sample-dependent and sample-independent processes. After validating input of 300 ng to each assay, gDNA was amplified using a whole genome amplification process. After fragmentation of the DNA, the sample is hybridized to 50-mer probes attached to the Infinium BeadChip, stopping one base before the

interrogated base. Single base extension was then carried out to incorporate a labeled nucleotide. Dual color (Cy3 and Cy5) staining allowed the nucleotide to be detected by the iSCAN reader (Illumina, San Diego, CA) and was converted to genotype during analysis with GenomeStudio.

#### Statistical analysis

Features with high coefficient of variation across triplicate measurements and detected in less than 80% of the samples were removed prior to statistical analysis. Batch effects were corrected using ComBat [49]. The peak intensity values were log-2 transformed and adjusted for age, sex, and batches of LC-HRMS profiling by taking the residual values from generalized linear regression with the covariates. For each feature, individuals that were outside of three standard deviations from residual mean value were excluded to reduce potential spurious associations with rare variants.

Out of genotype data from 453 individuals and 1.82 million variants, we used variants and samples that passed all the following exclusion criteria: (1) variants or samples with missing rate less than 2%, (2) bi-allelic variants in autosome with minor allele frequency of 5% or greater, (3) variants passing Hardy–Weinberg equilibrium test ( $p$  value threshold of  $10^{-6}$ ), (4) samples without excess heterozygosity (within 3 standard deviations from average across samples), (5) unrelated samples with pairwise King-robust estimator less than 0.177. The final genotype data used in GxMWAS consisted of 441 samples and 619,688 variants in autosomes (including 1036 indels and 17 tri-allelic SNPs each split into 2 bi-allelic SNPs).

The narrow-sense heritability ( $h^2$ ) of each feature level was estimated using genomic related matrix (GRM) restricted maximum likelihood (GREML) implemented in genome-wide complex trait analysis (GCTA) [22]. For each feature, we fitted a generalized linear model (`--glm`) [50] implemented in PLINK 2.00a2.3LM [51] with the dosages of minor alleles as independent variables (additive genetic model) and top 10 PCs as covariates to account for population structure. The gene-level enrichment analysis was done for each feature with its summary statistics using MAGMA version 1.0 [30]. All subsequent statistical analysis was performed using R statistical language (version 4.1.2; R Foundation for Statistical Computing, Vienna, Austria).

#### Causal network analysis of significant GIMs

We systematically integrated the 29 GIMs and their genetic determinants to identify the GIM-causal network at level 0.05 using the G-DAG algorithm [33]. The causal Bayesian network is augmented with principles of Mendelian randomization (MR). The MR approach is an

instrumental variable (IV) technique to identify causal relationships. The assumptions are.

1. IVs are associated with GIMs.
2. IVs are exogenous variables, not affected by metabolites.
3. IVs do not have pleiotropic effect.

We satisfied the assumption 1 by using the genetic determinants of GIMs that are strongly associated with GIMs. The assumption 2 is satisfied since the genetic variations affect metabolites unidirectionally. The assumption 3 is assessed using conditional independence test embedded in the G-DAG algorithm [33].

The GIM-causal network identified from this approach is represented as a directed graph, where the direction represents the direction of effect.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40246-022-00440-w>.

**Additional file 1. Fig. S1.** Population stratification with the first two principal components for all subjects in the study. **Fig. S2.** Distribution of m/z and RT for features associated with age. **Fig. S3.** Distribution of m/z and RT for features associated with sex. **Fig. S4.** Distribution of m/z and RT for features associated with PC1. **Fig. S5.** Correlation matrix among ten features associated with *UGT1A* and isoforms. **Fig. S6.** Visualization of gene-feature association identified using MAGMA.

**Additional file 2. Table S1.** List of features experimentally identified with authentic standards.

**Additional file 3. Table S2.** Demographic characteristics of the pediatric cohort in the study.

**Additional file 4. Table S3.** Features associated with demographic features (age, sex, and genetic ancestry). Only experimentally identified metabolites are listed.

**Additional file 5. Table S4.** Pathways enriched among features associated with demographic features (age, sex, and PC1) by Mummichog.

**Additional file 6. Table S5.** List of all genome-wide significant variant-feature associations.

**Additional file 7. Table S6.** Feature annotations by xMSannotator with high or medium confidence and with 5-ppm mass tolerance.

**Additional file 8. Table S7.** List of all significant gene-feature associations by MAGMA.

## Acknowledgements

Authors acknowledge material and data support from the PrecisionLink Biobank for Health Discovery at Boston Children's Hospital. We thank Mass General Brigham Biobank Genomics Core for providing technical expertise and sample processing using the Illumina Infinium platform. We want to acknowledge Vilinh Tran for her supports with the technical expertise in mass spectrometry.

## Author contributions

DPJ, KDM, and SWK contributed to study design. DPJ, DIW, and MRS were involved in performing LC-HRMS measurements, quality control, and annotation. I-HL and SWK contributed to analysis of genotype data. MRS, I-HL, SS, AY, and SWK were involved in analysis and management of metabolomics data. I-HL, MRS, AY, KDM, and SWK wrote the paper. All authors read and approved the final manuscript.

## Funding

This study was supported by funding from National Institute of Health (NIMH R01MH107205, ORIP R24OD024622, NHLBI R01HL150401, and NCATS U01TR002623).

## Availability of data and materials

The metabolomic dataset is available in the Metabolomics Workbench (Study\_ID: ST002331, Project\_ID: PR001495, <https://doi.org/10.21228/M8GM6Q>).

## Declarations

### Ethics approval and consent to participate

The study was reviewed and approved by the BCH Institutional Review Board.

### Consent for publication

Not applicable.

### Competing interests

The authors declare no competing financial interests.

## Author details

<sup>1</sup>Computational Health Informatics Program, Boston Children's Hospital, 401 Park Drive, Boston, MA 02215, USA. <sup>2</sup>Division of Pulmonary, Allergy, and Critical Care Medicine, Department of Medicine, Emory University, Atlanta, GA 30602, USA. <sup>3</sup>Atlanta Department of Veterans Affairs Medical Center, Decatur, GA 30033, USA. <sup>4</sup>Center of Perioperative Genetics and Genomics, Department of Anesthesiology, Perioperative and Pain Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA. <sup>5</sup>Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA. <sup>6</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA. <sup>7</sup>Department of Pediatrics, Harvard Medical School, Boston, MA 02115, USA.

Received: 16 August 2022 Accepted: 22 November 2022

Published online: 08 December 2022

## References

1. Vermeulen R, Schymanski EL, Barabasi AL, Miller GW. The exposome and health: where chemistry meets biology. *Science*. 2020;367(6476):392–6.
2. Rinschen MM, Ivanisevic J, Giera M, Siuzdak G. Identification of bioactive metabolites using activity metabolomics. *Nat Rev Mol Cell Biol*. 2019;20(6):353–67.
3. Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat Rev Mol Cell Biol*. 2016;17(7):451–9.
4. Panyard DJ, Kim KM, Darst BF, Deming YK, Zhong X, Wu Y, Kang H, Carlsson CM, Johnson SC, Asthana S, et al. Cerebrospinal fluid metabolomics identifies 19 brain-related phenotype associations. *Commun Biol*. 2021;4(1):63.
5. Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine. *Nat Rev Drug Discov*. 2016;15(7):473–84.
6. Long T, Hicks M, Yu HC, Biggs WH, Kirkness EF, Menni C, Zierer J, Small KS, Mangino M, Messier H, et al. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat Genet*. 2017;49(4):568–78.
7. Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, Huang J, Arnold M, Erte I, Forgetta V, Yang TP, et al. An atlas of genetic influences on human blood metabolites. *Nat Genet*. 2014;46(6):543–50.
8. Yet I, Menni C, Shin SY, Mangino M, Soranzo N, Adams J, Suhre K, Spector TD, Kastenmuller G, Bell JT. Genetic influences on metabolite levels: a comparison across metabolomic platforms. *PLoS ONE*. 2016;11(4):e0153672.
9. Kastenmuller G, Raffler J, Gieger C, Suhre K. Genetics of human metabolism: an update. *Hum Mol Genet*. 2015;24(R1):R93–101.
10. Suhre K, Wallaschofski H, Raffler J, Friedrich N, Haring R, Michael K, Wasner C, Krebs A, Kronenberg F, Chang D, et al. A genome-wide association study of metabolic traits in human urine. *Nat Genet*. 2011;43(6):565–9.

11. Illig T, Gieger C, Zhai G, Romisch-Margl W, Wang-Sattler R, Prehn C, Altmajer E, Kastenmuller G, Kato BS, Mewes HW, et al. A genome-wide perspective of genetic variation in human metabolism. *Nat Genet.* 2010;42(2):137–41.
12. Lotta LA, Pietzner M, Stewart ID, Wittemans LBL, Li C, Bonelli R, Raffler J, Biggs EK, Oliver-Williams C, Auyeung VPW, et al. A cross-platform approach identifies genetic regulators of human metabolism and health. *Nat Genet.* 2021;53(1):54–64.
13. Jones DP. Sequencing the exposome: A call to action. *Toxicol Rep.* 2016;3:29–45.
14. Uppal K, Walker DI, Liu K, Li S, Go YM, Jones DP. Computational metabolomics: a framework for the million metabolome. *Chem Res Toxicol.* 2016;29(12):1956–75.
15. Kong SW, Hernandez-Ferrer C. Assessment of coverage for endogenous metabolites and exogenous chemical compounds using an untargeted metabolomics platform. *Pac Symp Biocomput.* 2020;25:587–98.
16. Uppal K, Walker DI, Jones DP. xMSannotator: an R package for network-based annotation of high-resolution metabolomics data. *Anal Chem.* 2017;89(2):1063–7.
17. Liu KH, Nellis M, Uppal K, Ma C, Tran V, Liang Y, Walker DI, Jones DP. Reference standardization for quantification and harmonization of large-scale metabolomics. *Anal Chem.* 2020;92(13):8836–44.
18. Stiburkova B, Bleyer AJ. Changes in serum urate and urate excretion with age. *Adv Chronic Kidney Dis.* 2012;19(6):372–6.
19. Krzyzanski W, Smits A, Van Den Anker J, Allegaert K. Population model of serum creatinine as time-dependent covariate in neonates. *AAPS J.* 2021;23(4):86.
20. Li S, Park Y, Duraisingham S, Strobel FH, Khan N, Soltow QA, Jones DP, Pulendran B. Predicting network activity from high throughput metabolomics. *PLoS Comput Biol.* 2013;9(7):e1003123.
21. Schymanski EL, Jeon J, Gulde R, Fenner K, Ruff M, Singer HP, Hollender J. Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol.* 2014;48(4):2097–8.
22. Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM. Concepts, estimation and interpretation of SNP-based heritability. *Nat Genet.* 2017;49(9):1304–10.
23. Rueedi R, Ledda M, Nicholls AW, Salek RM, Marques-Vidal P, Morya E, Sameshima K, Montoliu I, Da Silva L, Collino S, et al. Genome-wide association study of metabolic traits reveals novel gene-metabolite-disease links. *PLoS Genet.* 2014;10(2):e1004132.
24. Nicholson G, Rantalainen M, Li JV, Maher AD, Malmodin D, Ahmadi KR, Faber JH, Barrett A, Min JL, Rayner NW, et al. A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. *PLoS Genet.* 2011;7(9):e1002270.
25. Raffler J, Friedrich N, Arnold M, Kacprowski T, Rueedi R, Altmajer E, Bergmann S, Budde K, Gieger C, Homuth G, et al. Genome-wide association study with targeted and non-targeted NMR metabolomics identifies 15 novel loci of urinary human metabolic individuality. *PLoS Genet.* 2015;11(9):e1005487.
26. Pfennig S, Foss F, Bissen D, Harde E, Treeck JC, Segarra M, Acker-Palmer A. GRIP1 binds to ApoER2 and EphrinB2 to induce activity-dependent AMPA receptor insertion at the synapse. *Cell Rep.* 2017;21(1):84–96.
27. Feng X, Krogh KA, Wu CY, Lin YW, Tsai HC, Thayer SA, Wei LN. Receptor-interacting protein 140 attenuates endoplasmic reticulum stress in neurons and protects against cell death. *Nat Commun.* 2014;5:4487.
28. Puffenberger EG, Jinks RN, Sougnez C, Cibulskis K, Willert RA, Achilly NP, Cassidy RP, Fiorentini CJ, Heiken KF, Lawrence JJ, et al. Genetic mapping and exome sequencing identify variants associated with five novel diseases. *PLoS ONE.* 2012;7(1):e28936.
29. Spears E, Shou M, Siv WA, Dai C, Chen W, Powers AC, Dean D. 198-OR: role for the cationic amino acid transporter Slc7a2 in alpha-cell proliferation and islet hormone secretion. *Diabetes.* 2019;68(Supplement 1):198-OR.
30. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol.* 2015;11(4):e1004219.
31. Schlosser P, Li Y, Sekula P, Raffler J, Grundner-Culemann F, Pietzner M, Cheng Y, Wuttke M, Steinbrenner I, Schultheiss UT, et al. Genetic studies of urinary metabolites illuminate mechanisms of detoxification and excretion in humans. *Nat Genet.* 2020;52(2):167–76.
32. Suhre K, Shin SY, Petersen AK, Mohney RP, Meredith D, Wagele B, Altmajer E, CardioGram, Deloukas P, Erdmann J, et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature.* 2011;477(7362):54–60.
33. Yazdani A, Yazdani A, Samiei A, Boerwinkle E. Identification, analysis, and interpretation of a human serum metabolomics causal network in an observational study. *J Biomed Inform.* 2016;63:337–43.
34. Barbu A, Hamad OA, Lind L, Ekdahl KN, Nilsson B. The role of complement factor C3 in lipid metabolism. *Mol Immunol.* 2015;67(1):101–7.
35. Murray I, Havel PJ, Sniderman AD, Cianflone K. Reduced body weight, adipose tissue, and leptin levels despite increased energy intake in female mice lacking acylation-stimulating protein. *Endocrinology.* 2000;141(3):1041–9.
36. Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature.* 2014;511(7510):421–7.
37. Meyer-Schuman R, Antonellis A. Evidence for a dominant-negative mechanism in HARS1-mediated peripheral neuropathy. *FEBS J.* 2021;288(1):91–4.
38. Safka Brozkova D, Deconinck T, Griffin LB, Ferbert A, Haberlova J, Mazanec R, Lassuthova P, Roth C, Pilunthanakul T, Rautenstrauss B, et al. Loss of function mutations in HARS cause a spectrum of inherited peripheral neuropathies. *Brain.* 2015;138(Pt 8):2161–72.
39. Hicks CW, Selvin E. Epidemiology of peripheral neuropathy and lower extremity disease in diabetes. *Curr Diab Rep.* 2019;19(10):86.
40. Kim SH, Park HD, Sohn YB, Park SW, Cho SY, Ji S, Kim SJ, Choi EW, Kim CH, Ko AR, et al. Mutations of ACADS gene associated with short-chain acyl-coenzyme A dehydrogenase deficiency. *Ann Clin Lab Sci.* 2011;41(1):84–8.
41. Bourgeois FT, Avillach P, Kong SW, Heinz MM, Tran TA, Chakrabarty R, Bickel J, Sliz P, Borglund EM, Kornetsky S, et al. Development of the precision link Biobank at Boston Children's Hospital: challenges and opportunities. *J Pers Med.* 2017;7(4):21.
42. Yu T, Park Y, Johnson JM, Jones DP. apLCMS—adaptive processing of high-resolution LC/MS data. *Bioinformatics.* 2009;25(15):1930–6.
43. Uppal K, Soltow QA, Strobel FH, Pittard WS, Gernert KM, Yu T, Jones DP. xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. *BMC Bioinform.* 2013;14:15.
44. Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, Edison A, Fiehn O, Higashi R, Nair KS, et al. Metabolomics workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* 2016;44(D1):D463–470.
45. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000;28(1):27–30.
46. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, et al. HMDB: the Human Metabolome Database. *Nucleic Acids Res.* 2007;35(Database issue):D521–6.
47. Wishart D, Arndt D, Pon A, Sajed T, Guo AC, Djoumbou Y, Knox C, Wilson M, Liang Y, Grant J, et al. T3DB: the toxic exposome database. *Nucleic Acids Res.* 2015;43(Database issue):D928–34.
48. Fahy E, Sud M, Cotter D, Subramaniam S. LIPID MAPS online tools for lipid research. *Nucleic Acids Res.* 2007;35(Web Server issue):W606–612.
49. Liu Q, Walker D, Uppal K, Liu Z, Ma C, Tran V, Li S, Jones DP, Yu T. Addressing the batch effect issue for LC/MS metabolomics data in data preprocessing. *Sci Rep.* 2020;10(1):13856.
50. Hill A, Loh PR, Bharadwaj RB, Pons P, Shang J, Guinan E, Lakhani K, Kilty I, Jelinsky SA. Stepwise distributed open innovation contests for software development: acceleration of genome-wide association analysis. *Gigascience.* 2017;6(5):1–10.
51. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7.

## Publisher's Note

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