HHS Public Access

Author manuscript

Pac Symp Biocomput. Author manuscript; available in PMC 2020 January 01.

Published in final edited form as:

Pac Symp Biocomput. 2020 ; 25: 587–598.

Assessment of coverage for endogenous metabolites and exogenous chemical compounds using an untargeted metabolomics platform

Sek Won Kong[1,2,*], **Carles Hernandez-Ferrer**[1,2]

[1]Computational Health Informatics Program, Boston Children's Hospital, 300 Longwood Avenue Boston, MA 02115, USA

[2]Department of Pediatrics, Harvard Medical School, Boston, MA 02115, USA

Abstract

Physiological status and pathological changes in an individual can be captured by metabolic state that reflects the influence of both genetic variants and environmental factors such as diet, lifestyle and gut microbiome. The totality of environmental exposure throughout lifetime – i.e., exposome – is difficult to measure with current technologies. However, targeted measurement of exogenous chemicals and untargeted profiling of endogenous metabolites have been widely used to discover biomarkers of pathophysiologic changes and to understand functional impacts of genetic variants. To investigate the coverage of chemical space and interindividual variation related to demographic and pathological conditions, we profiled 169 plasma samples using an untargeted metabolomics platform. On average, 1,009 metabolites were quantified in each individual (range 906 – 1,038) out of 1,244 total chemical compounds detected in our cohort. Of note, age was positively correlated with the total number of detected metabolites in both males and females. Using the robust Q_n estimator, we found metabolite outliers in each sample (mean 22, range from 7 to 86). A total of 50 metabolites were outliers in a patient with phenylketonuria including the ones known for phenylalanine pathway suggesting multiple metabolic pathways perturbed in this patient. The largest number of outliers (N=86) was found in a 5-year-old boy with alpha-1-antitrypsin deficiency who were waiting for liver transplantation due to cirrhosis. Xenobiotics including drugs, diets and environmental chemicals were significantly correlated with diverse endogenous metabolites and the use of antibiotics significantly changed gut microbial products detected in host circulation. Several challenges such as annotation of features, reference range and variance for each feature per age group and gender, and population scale reference datasets need to be addressed; however, untargeted metabolomics could be immediately deployed as a biomarker discovery platform and to evaluate the impact of genomic variants and exposures on metabolic pathways for some diseases.

*To whom correspondence should be addressed. sekwon.kong@childrens.harvard.edu.

Keywords

Metabolomics; Disease; Untargeted metabolomics; Mass spectrometry; Blood

1. Introduction

Genetic discoveries from genome-wide association studies (GWAS) and whole-exome and -genome sequencing (WES/WGS) discovered risk alleles for common diseases and pathogenic genetic variants in 10-52% of patients with rare genetic diseases[1]. WES gained its clinical utility[2]; however, understanding functional consequences of genetic variant in the context of disease phenotype is essential and yet remains as an outstanding challenge since generally healthy children also harbor tens of putative disease-associated genetic variants. A functional read out – e.g., gene expression profiling of affected tissue – could inform impacts of genetic variants[3]. Nonetheless, accessibility to affected tissue is often limited and especially challenging for developmental disorders.

Metabolites are direct read-outs of functional status of biological entities – i.e., cells, tissues, and organs – and also serve as a proxy for understanding their sources such as internal metabolic processes, gut microbiome, xenobiotics, dietary, and exogenous exposures [4]. Moreover, metabolites are active regulators of gene expression and protein activity[5]. A limited set of blood chemistry analytes is routinely used in clinical care, which provide crucial information regarding pathophysiology. Metabolomics aims to characterize all the small molecules in biological system using metabolomics platforms such as nuclear magnetic resonance (NMR) spectroscopy and chromatography coupled to mass spectrometry (MS)[6]. NMR reproducibly identifies chemical structure of unknown chemical features but is limited by its lower sensitivity and throughput compared to MS-based metabolomics. Therefore, untargeted metabolomics using a high-resolution MS is typically used for hypothesis-driven research studies and novel biomarker discovery[4].

Metabolomic profiling with blood and affected tissue could be more closely associated with phenotype compared to other omics profiles[7]. More importantly, perturbed metabolic pathways could suggest mechanistic insights into the pathophysiology of diseases[8]. Previous studies showed the analytical validity of MS-based metabolomics platforms and successfully demonstrated a utility in interpreting the impact of genetic variants for generally healthy individuals[9] or in discovering novel biomarkers for inborn errors of metabolism (IEMs)[10]. These studies approached an index case to find metabolite outliers compared to a background distribution constructed from generally healthy individuals for each metabolite. Here we investigated the extent of endogenous metabolites and exogenous chemical compounds that could be captured by untargeted metabolomics profiling of plasma samples from patients with diverse medical conditions to evaluate a potential of untargeted metabolomics profiling as a precision medicine platform.

2. Materials and Methods

2.1. Subjects

Individuals were enrolled in the Precision Link Health Discovery cohort at Boston Children's Hospital (BCH) from January 2016 to November 2017. Enrolled patients and their family members were consented in-person with permission to access electronic health records (EHRs), if available, for research and to share de-identified data and specimens within and outside of the institution[11]. We collected 169 plasma samples from 79 males and 90 females with mean ages 19.6 and 20.9 years old, respectively (ranges from 4.4 months to 59.7 years).

Data from patient databases at BCH were obtained using i2b2, which allows for queries of EHRs using International Classification of Diseases, Ninth Revision, Clinical Modification or Tenth Revision codes, Systematized Nomenclature of Medicine - Clinical Terms, and the dates when the codes were assigned to patients and demographic information. The queries of the institutional i2b2 database and analyses were performed and restricted to October to December 2018. For 123 patients of 169 enrolled individuals with plasma samples, we collected the prescription history of 1,194 drugs corresponding to 594,201 events in the i2b2 database. The study was approved by the Institutional Review Board of BCH.

2.2. Untargeted metabolomics profiling of plasma samples

Whole blood was collected in ethylenediaminetetraacetic acid (EDTA) treated lavender top tubes, from the Precision Link Biobank participants. EDTA tube was centrifuged at 2000 X G for 10 minutes at room temperature to obtain plasma. Plasma samples were then aliquoted 200uL/0.5ml microcentrifuge tubes and stored at −80C. These samples were shipped in a dry iced box to Metabolon (Research Triangle Park, NC) for untargeted metabolomics profiling. Sample handling, metabolomic profiling, quality control and data pre-processing is described in detail in the previous study[9]. In brief, proteins were precipitated with methanol under vigorous shaking for 2 min (Glen Mills GenoGrinder 2000, Glen Mills, Clifton, NJ) followed by centrifugation. The resulting extract was divided into four fractions:

- Two for analysis by two separate reverse phase (RP)/ultra-performance liquid chromatography (UPLC)-MS/MS methods with positive ion mode electrospray ionization (ESI).
- One for analysis by RP/UPLC-MS/MS with negative ion mode ESI.
- One for analysis by HILIC/UPLC-MS/MS with negative ion mode ESI.

To remove the organic solvent, samples were placed briefly on a TurboVap® (Zymark, Hopkinton MA). A Waters ACQUITY UPLC (Milford, MA) and a Thermo Scientific Q-Exactive high resolution/accurate mass spectrometer interfaced with a heated electrospray ionization (HESI-II) source and Orbitrap mass analyzer (Waltham, MA) operating at 35,000 mass resolution were utilized to analyze aliquots covering 70 – 1,000 mass-to-charge ratio (m/z). Raw data was extracted, peak-identified and quantified using area-under-the-curve using Metabolon's hardware and software. Deliverables from Metabolon included raw area counts, rescaled-to-median and imputed values, and sample volume normalized data with the

retention time/index (RI), m/z, chemical annotation according to Metabolon's proprietary database with public database identifiers including PubChem[12], the Human Metabolome Database (HMDB)[13] and Kyoto Encyclopedia of Genes and Genomes (KEGG)[14] if available.

2.3. Statistical analysis

We used a volume normalized and re-scaled – i.e., median equals to 1 for each metabolite – data generated by Metabolon software pipeline[9]. Missing values were imputed with minimum observed value for each metabolite. A complete data table including 1,244 metabolites for 169 individuals was used for further analysis. Overall, concentrations of both endogenous metabolites and exogenous chemicals showed log-normal distribution; however, some exogenous chemicals were detected only in a small number of samples and the distribution was skewed for some metabolites. The median absolute deviation (MAD) is a robust scale estimator that is widely used with the sample median; however, it is a symmetric estimator of dispersion and has a low efficiency for data with Gaussian distribution. To address these limitations of MAD in our analysis, z-scores were calculated from log-transformed values using Q_n estimator that is considered to be more robust for data with asymmetric distribution[15]. For each metabolite, we calculated Q_n estimator using the Q_n function implemented in the robustbase R library package.

To explore correlation structure of metabolome, a robust estimator of correlation was required since some of them (e.g., prescribed drugs) were measured in a small proportion of samples which could cause a bias with Pearson or Spearman correlation coefficients. Thus, we calculated rQ_n as described in Eq.(1) where u and v were calculated according to Eq.(2) with the sample medians, $\tilde{x}$ and $\tilde{y}$[16].

$$rQ_n = \begin{cases} \dfrac{Q_n^2(u) - Q_n^2(v)}{Q_n^2(u) + Q_n^2(v)} & \text{if } Q_n(x) \neq 0 \text{ and } Q_n(y) \neq 0 \\ NA & \text{if} Q_n(x) = 0 \text{ or } Q_n(y) = 0 \end{cases} \quad (1)$$

$$u = \frac{x - \tilde{x}}{\sqrt{2}Q_n(x)} + \frac{y - \tilde{y}}{\sqrt{2}Q_n(y)} \text{and } v = \frac{x - \tilde{x}}{\sqrt{2}Q_n(x)} - \frac{y - \tilde{y}}{\sqrt{2}Q_n(y)} \quad (2)$$

Statistical significance of pairwise correlation was estimated using a t-distribution with n-2 degrees of freedom, where t is the Fisher-transformed robust correlation coefficients. Multiple testing correction was performed by calculating false discovery rate (FDR) from distribution of p-values[17]. All analyses were performed in the R statistical software environment[18].

3. Results

3.1. Overview of untargeted metabolomics profiling

3.1.1. Chemical coverage—A total of 1,244 endogenous metabolites and exogenous chemical compounds – hereafter referred to as *features* in aggregate – were measured in 169

plasma samples. On average, 1,009 features per sample (ranges from 906 to 1,038) were measured above detection limits. The majority of features (i.e., 1,073 out of 1,244) were successfully quantified in more than 50% of individual samples; however, 105 out of 224 xenobiotics such as drugs and food metabolites were only detected in less than 20% of samples. There was no difference in the number of features detected between males and females (Welch's *t*-test, p-value 0.29); however, age was significantly correlated with the total number of detected features in both males and females (generalized linear model, p-value 6.78×10^{-12}). The total number of xenobiotics measured per sample was also correlated with age (p-value 5.54×10^{-9}) but not significantly different between males and females (p-value 0.922).

According to their chemical properties, each feature was assigned to one of nine super-classes (i.e., amino acids, carbohydrates, cofactor and vitamins, energy, lipids, nucleotides, partially characterized molecules, peptides, and xenobiotics) and unannotated molecules, and one of 112 subpathways (Figure 1A). Lipids (N = 423) and amino acids (N = 195) were the major classes of endogenous features quantified by the untargeted platform used in the current study. For xenobiotics, we could identify 244 chemical compounds from: food (N=54), tobacco (N=6), benzoate (N=22), xanthine (N=15), exogenous environmental chemicals (N=26), bacterial/fungal (N=1), and drug metabolites including analgesics (N=22), anti-inflammatory (N=5), antibacterial (N=14), antiviral (N=2), cardiovascular (N=10), gastrointestinal (N=4), metabolic (N=2), neurological (N=18), psychotropic (N=15), respiratory (N=5) and topical agents (N=3).

3.1.2. Global correlation structure of human plasma metabolome—To examine correlation structure of features, we created a network of 1,244 features (i.e., nodes) connected by edges of significant correlation for each pair. We used a robust estimator of correlation – i.e., rQ_n, and selected top-most significant correlations with correlation coefficient greater than 0.4. Using the 1,244 features and 17,659 significant correlations (false discovery rate (FDR) < 0.0001 and $|rQ_n| > 0.4$) as edges, we constructed a metabolomic network. A force directed layout – ForceAtlas2 – was used to spatialize the network[19]. Overall, features were clustered by super-pathways (Figure 1B). Unconnected nodes were mostly xenobiotics and their metabolites; however, some xenobiotics were significantly correlated with diverse super-classes of endogenous metabolites such as amino acids and lipids suggesting the impact of exogenous chemical compounds on different metabolic pathways. Interestingly, lipid species formed four distinct clusters: sphingomyelins, diacylglycerols, steroid metabolism and fatty acids. Amino acids were broadly connected with multiple super-classes including xenobiotics. Unannotated features – i.e., features with unique pair of m/z and retention time without matching information in multiple databases – formed clusters with different super-classes suggesting these unannotated features could be mapped to known super-classed based on correlational structure. Additional details on global network structure with chemical compound names and correlation structure of subnetworks are available at the supplementary website (https://tom.tch.harvard.edu/supples/metabolome).

3.2. Factors contributing interindividual variance in feature concentrations

3.2.1. Demographic variables—Except for sex hormones, we did not find features showing significantly different concentrations between males and females. An androgenic steroid, 5alpha-androstan-3alpha,17beta-diol monosulfate was significantly higher in males after controlling for the effect of age. Age was significantly correlated with 502 features (40.4% of 1,244 features, FDR < 0.05). The complete list of metabolites correlated with age and statistical scores are available at the supplementary website (https://tom.tch.harvard.edu/supples/metabolome). We checked whether age-correlated chemical compounds were more frequently observed for each of nine super-classes and unannotated chemicals. Xenobiotics were enriched with age-correlated chemical compounds (Fisher's exact test p-value 0.000027, odds ratio 2.49 with 95% confidence interval (CI) 1.595 – 3.954) and nucleotides were depleted for age-correlated metabolites (Fisher's exact test p-value 0.00096, odds ratio 0.27 with 95% CI 0.099 – 0.644). Interestingly, age-correlation could be nonlinear and only significant correlated in an age group (e.g., children vs. adults). For instance, creatinine was positive correlated with age in children then reached plateau in adults[20] (Figure 1C). Therefore, background distribution of metabolites should be constructed for each age group.

3.2.2. Use of antibiotics—Circulating metabolites of mammalian host are substantially affected by gut microbiota[21]. In the current study, 34 gut microbial products that are exclusively or mainly contributed by bacteria metabolism were detected (see Appendix). These microbial products were tightly correlated with aromatic amino acids and bile acids metabolism, and significantly correlated with 773 features (FDR < 0.01). From EHR, we identified medication history for 123 out of 169 individuals. We selected 68 individuals with active drug prescription or in a window of 14 days after finishing drug prescription and found 23 features matching the drugs prescribed in at least one patient. We captured 40.9% drug prescription ($N_{detected}$ & $N_{prescribed}$ = 67, $N_{prescribed}$ = 164) and identified 128 drug consumptions with no prescription. Two antibiotic drugs were detected in the matched prescribed drugs and used by nine patients. Thirty-four features matched with gut microbial products including p-cresol and 4-hydroxyphenylacetate that are tyrosine metabolic products of anaerobic *Clostridium difficile* and certain *Lactobacillus* strains. The concentration of three gut microbial products were significantly correlated with the prescription history of the two antibiotics: 3-indoxyl sulfate, indole propionate and p-cresol sulfate (logistic regression, FDR < 0.01, Figure 2A).

Low indoxyl sulfate level suggested the relevance of microbiota-derived indole and features thereof in mucosal integrity and protection from inflammation[22]. p-cresol and 4-hydroxyphenylacetate are tyrosine metabolic products of anaerobic *Clostridia*, and over-growth of this genera could be associated with gastrointestinal symptoms. Moreover, plasma levels of trimethylamine n-oxide, derived from dietary choline and carnitine through the action of gut microbiota, are associated with several cardiometabolic traits[23].

3.2.3. Impact of environmental chemical toxicants on blood metabolome—Per- and polyfluoroalkyl substances (PFAS) are a group of industrial chemicals including perfluorooctanoic acid (PFOA) and perfluorooctanesulfonic acid (PFOS), which are used in various industrial products including food containers and present in drinking water. PFOA is

a toxicant affecting multiple biological pathways and considered as non-genotoxic carcinogens. In our cohort, PFOA and PFOS were detected and quantified in 102 and 169 samples, respectively. Endogenous features were significantly correlated with PFOA and PFOS (N = 65 and 227, FDR < 0.05) with 52 features in common (Figure 2B).

3.3. Implication for medical conditions

Forty features were detected in less than three samples and 941 features (75.6% of all measured ones) were not normally distributed (Shapiro-Wilke test, p-value < 0.05). Thus, we used the Q_n estimator to calculate robust z-scores to detect outliers (i.e., |z-score| > 3) after excluding 52 features detected in less than three individuals. As a proof-of-concept, we checked outlier features in patients with IEMs and diabetes mellitus (DM).

3.3.1. Inborn errors of metabolism—Significantly higher levels of phenylalanine, phenyllactate, and phenylpyruvate were observed in a 40-year-old male with classical phenylketonuria (z-scores 9.39, 9.12 and 8.53, respectively). Interestingly, there were also significantly low concentration of alpha-ketoglutaramate (z-score - 7.67) potentially due to long-term use of Phe-restrictive diet throughout life. Additionally, 46 features were outliers in this patient suggesting the perturbation of phenylalanine pathway as well as the other metabolic pathways (Figure 3A).

Alpha-1-antitrypsin deficiency (A1AD) is an autosomal recessive disorder due to a mutation in *SERPINA1* and often presents respiratory symptoms and liver failure. An 8-year-old girl was diagnosed with A1AD, and her metabolomic profile showed perturbation of liver enzyme pathways including sterol, ceramide and bile acid metabolism. Vitamin A and its metabolites showed significantly low concentration compared to the others suggesting that vitamin A supplement would be required. We confirmed prescription history of multivitamins and the other cofactors in EHR for this patient.

3.3.2. Diabetes mellitus—Glucose and mannose concentrations were not consistently changed in the patients with DM; however, 1,5-anhydroglucitol (1,5-AG) was detected as an outlier in all patients with type I and II DM. For instance, metabolomic profile of a patient with type II DM showed significantly low concentration of 1,5-AG with higher glucose concentration (Figure 3B). When blood glucose levels exceed the renal glucose threshold, glucose is excreted to urine and re-absorption of 1,5-AG is inhibited resulting low 1,5-AG concentrations with hyperglycemic events[24]. *Two parental samples also showed low concentrations of 1,5-AG suggesting DM although medical records for these individuals were not available.*

4. Discussion

Using an untargeted metabolomics platform, we successfully profiled a broad range of internal and external exposures in plasma samples from a cohort comprising generally healthy and individuals with diverse pediatric disorders with a fraction of cost for measuring several clinical laboratory tests. Endogenous features such as lipids, amino acids and nucleic acids were consistently measured in both children and adults while the total number of detected features was correlated with age likely due to exposures to diverse exogenous

chemical compounds with aging. Internal metabolites correlated with exogenous chemical compounds (e.g., PFAS and PFOS) suggested potential metabolic pathways affected by such compounds. Moreover, the use of antibiotics was reflected in the concentration changes of gut microbial products.

The total number of chemical entities has not been reported in the human and no metabolomics platform can quantitatively measure the entirety of chemicals of endogenous and exogenous origins. Unbiased profiling of all chemical compounds present in the human tissues may not be possible in near future nor required to understand the impact of metabolomic changes due to underlying physiological changes and exposures, which require further investigation and theoretical/experimental model validations[25].

There are few challenges that needed to be addressed for clinical research use of metabolomics in the context of precision medicine. Firstly, the reliability of measurement should be established for accurate and reproducible results. The US Environmental Protection Agency initiated the Non-Targeted Analysis Collaborative Trial to evaluate untargeted metabolomics platforms[26]. A previous study showed a wide range of coefficient of variation from 0.96% to 119.1% for the features measured by the same metabolomics platform[10]. To gain broader applications, a systematic comparison of platforms would be crucial. Secondly, feature annotation is incomplete. High-resolution MS has a potential to characterize 10,000 – 30,000 features in a single run. However, only a fraction of these features could be annotated with known chemical properties in the current study (N = 988). Current computational annotation using m/z and retention time needs to be improved. The Human Metabolome Project provides a repository of features from various sources[13]; however, classification of features in terms of ontology and functional characteristics are challenging. After all, metabolome databases do not provide the same level of organized information compared to genomic sequence databases. Thirdly, population-scale reference datasets would be essential for determining reference ranges and interindividual variation in diverse population. Coordinated data sharing platforms such as the MetaboLights database[27] and Metabolomics Workbench[28] are highly required to facilitate the distribution of existing data, standards, protocols, and analytical tools. Lastly, tissue-wide metabolomics profiling could greatly advance our understanding of tissue-specific metabolomic characteristics and their implication in pathophysiology of human disease.

The proportion of liability explained by genetic variants is relatively small for both common and rare diseases. Moreover, allelic and locus heterogeneities are frequently observed[29]. If one of the goals of translational genomic medicine is to find right drug for right patient, genetic data alone cannot provide sufficient insights as to diagnostic and therapeutic planning for patients[30]. Functional genomic data such as transcriptomic, proteomic and metabolomic analysis of treatment-naïve and during the course of treatment would be required in addition to WES/WGS. An immediate application of metabolomics (i.e., metabotype) is to complement genotype for prioritizing, optimizing and monitoring treatment strategy for patients with IEMs; however, application of untargeted metabolomics could be broader. One potential use case could be to model metabolite concentration as endophenotype that is affected by polygenic risk and exogenous environmental exposure for common disease. Mendelian randomization studies using metabolite profile seek for causal

association of metabolic biomarkers in metabolic and cardiovascular diseases. Once the analytical validity of untargeted metabolomics platforms is established from population scale studies, further dissection of genetic and environmental contributions to common diseases would be possible.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work is supported by the grants R01MH017205 and U01TR002623 from the National Institutes of Health. Authors acknowledge material and data support from the Precision Link Biobank for Health Discovery at Boston Children's Hospital.

6.: Appendix

Table. The list of the 34 gut microbial products, exclusively or mainly contributed by bacteria metabolism, detected by the untargeted metabolomics platform used in the current study. Known metabolic pathways and the Human Metabolome Database (HMDB) identifiers (ID) are shown for each metabolite. Metabolites with a number (#), are compounds that are a structural isomer of another compound in the Metabolon spectral library.

Metabolite name	Bacterial pathway	HMDB ID
2-hydroxyhippurate	xenobiotic metabolism	HMDB00840
3-(3-hydroxyphenyl)propionate	aromatic amino acid metabolism	HMDB00375
3-(4-hydroxyphenyl)lactate	aromatic amino acid metabolism	HMDB00755
3-hydroxyhippurate	xenobiotic metabolism	HMDB06116
3-indoxyl sulfate	aromatic amino acid metabolism	HMDB00682
3-phenylpropionate	aromatic amino acid metabolism	HMDB00764
4-hydroxyhippurate	xenobiotic metabolism	HMDB13678
4-hydroxyphenylacetate	aromatic amino acid metabolism	HMDB00020
4-hydroxyphenylpyruvate	aromatic amino acid metabolism	HMDB00707
cholate	bile acid metabolism	HMDB00619
daidzein sulfate (1)	xenobiotic metabolism	
daidzein sulfate (2)	xenobiotic metabolism	
deoxycholate	bile acid metabolism	HMDB00626
genistein sulfate	xenobiotic metabolism	
glycocholenate sulfate	bile acid metabolism	
glycodeoxycholate 3-sulfate	bile acid metabolism	
glycolithocholate sulfate	bile acid metabolism	HMDB02639
glycoursodeoxycholate	bile acid metabolism	HMDB00708
hippurate	bile acid metabolism	HMDB00714
hyocholate	bile acid metabolism	HMDB00760
indoleacetate	aromatic amino acid metabolism	HMDB00197

Metabolite name	Bacterial pathway	HMDB ID
indoleacetylglutamine	aromatic amino acid metabolism	HMDB13240
indolelactate	aromatic amino acid metabolism	HMDB00671
indolepropionate	aromatic amino acid metabolism	HMDB02302
lithocholate sulfate (1)	bile acid metabolism	
methyl-4-hydroxybenzoate sulfate	xenobiotic metabolism	
propyl 4-hydroxybenzoate sulfate	xenobiotic metabolism	
p-cresol sulfate	aromatic amino acid metabolism	HMDB11635
phenol sulfate	aromatic amino acid metabolism	HMDB60015
phenylacetate	aromatic amino acid metabolism	HMDB00209
phenylacetylglutamine	aromatic amino acid metabolism	HMDB06344
phenyllactate	aromatic amino acid metabolism	HMDB00779
taurocholenate sulfate	bile acid metabolism	
taurolithocholate 3-sulfate	bile acid metabolism	HMDB02580
tauroursodeoxycholate	bile acid metabolism	HMDB00874
ursodeoxycholate	bile acid metabolism	HMDB00946

References

1. Adams DR, Eng CM. Next-Generation Sequencing to Diagnose Suspected Genetic Disorders. N Engl J Med. 2018;379(14): 1353–1362. [PubMed: 30281996]
2. Tan TY, Dillon OJ, Stark Z, et al. Diagnostic Impact and Cost-effectiveness of Whole-Exome Sequencing for Ambulant Children With Suspected Monogenic Conditions. JAMA Pediatr. 2017; 171(9):855–862. [PubMed: 28759686]
3. Deelen P, van Dam S, Herkert JC, et al. Improving the diagnostic yield of exome- sequencing by predicting gene-phenotype associations using large-scale gene expression analysis. Nat Commun. 2019; 10(1):2837. [PubMed: 31253775]
4. Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: beyond biomarkers and towards mechanisms. Nat Rev Mol Cell Biol. 2016;17(7):451–459. [PubMed: 26979502]
5. Rinschen MM, Ivanisevic J, Giera M, Siuzdak G. Identification of bioactive metabolites using activity metabolomics. Nat Rev Mol Cell Biol. 2019;20(6):353–367. [PubMed: 30814649]
6. Bingol K, Bruschweiler R. Two elephants in the room: new hybrid nuclear magnetic resonance and mass spectrometry approaches for metabolomics. Curr Opin Clin Nutr Metab Care. 2015;18(5): 471–477. [PubMed: 26154280]
7. Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nat Rev Genet. 2018;19(5): 299–310. [PubMed: 29479082]
8. Delaney NF, Sharma R, Tadvalkar L, Clish CB, Haller RG, Mootha VK. Metabolic profiles of exercise in patients with McArdle disease or mitochondrial myopathy. Proc Natl Acad Sci U S A. 2017;114(31):8402–8407. [PubMed: 28716914]
9. Guo L, Milburn MV, Ryals JA, et al. Plasma metabolomic profiles enhance precision medicine for volunteers of normal health. Proc Natl Acad Sci U S A. 2015;112(35):E4901–4910. [PubMed: 26283345]
10. Miller MJ, Kennedy AD, Eckhart AD, et al. Untargeted metabolomic analysis for the clinical screening of inborn errors of metabolism. J Inherit Metab Dis. 2015;38(6): 1029–1039. [PubMed: 25875217]
11. Bourgeois FT, Avillach P, Kong SW, et al. Development of the Precision Link Biobank at Boston Children's Hospital: Challenges and Opportunities. J Pers Med. 2017;7(4).
12. Kim S, Chen J, Cheng T, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 2019;47(D1):D1102–D1109. [PubMed: 30371825]

13. Wishart DS, Feunang YD, Marcu A, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res. 2018;46(D1):D608–D617. [PubMed: 29140435]
14. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353–D361. [PubMed: 27899662]
15. Rousseeuw PJ, Croux C. Alternatives to the Median Absolute Deviation. Journal of the American Statistical Association. 1993;88(424): 1273–1283.
16. Shevlyakov G, Smirnov P. Robust Estimation of the Correlation Coefficient: An Attempt of Survey2011;No. 40:147–156.
17. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. 1995.
18. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
19. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS One. 2014;9(6):e98679. [PubMed: 24914678]
20. Uemura O, Honda M, Matsuyama T, et al. Age, gender, and body length effects on reference serum creatinine levels determined by an enzymatic method in Japanese children: a multicenter study. Clin Exp Nephrol. 2011; 15(5):694–699. [PubMed: 21505953]
21. Wikoff WR, Anfora AT, Liu J, et al. Metabolomics analysis reveals large effects of gut micro flora on mammalian blood metabolites. Proc Natl Acad Sci U S A. 2009;106(10):3698–3703. [PubMed: 19234110]
22. Weber D, Oefner PJ, Hiergeist A, et al. Low urinary indoxyl sulfate levels early after transplantation reflect a disrupted microbiome and are associated with poor outcome. Blood. 2015; 126(14): 1723–1728. [PubMed: 26209659]
23. Zhu W, Gregory JC, Org E, et al. Gut Microbial Metabolite TMAO Enhances Platelet Hyperreactivity and Thrombosis Risk. Cell. 2016;165(1): 111–124. [PubMed: 26972052]
24. Buse JB, Freeman JL, Edelman SV, Jovanovic L, McGill JB. Serum 1,5-anhydroglucitol (GlycoMark): a short-term glycemic marker. Diabetes Technol Ther. 2003;5(3):355–363. [PubMed: 12828817]
25. Miller GW, Jones DP. The nature of nurture: refining the definition of the exposome. Toxicol Sci. 2014; 137(1): 1–2. [PubMed: 24213143]
26. Ulrich EM, Sobus JR, Grulke CM, et al. EPA's non-targeted analysis collaborative trial (ENTACT): genesis, design, and initial findings. Anal Bioanal Chem. 2019;411(4):853–866. [PubMed: 30519961]
27. Kale NS, Haug K, Conesa P, et al. MetaboLights: An Open-Access Database Repository for Metabolomics Data. Curr Protoc Bioinformatics. 2016;53:14 13 11–18.
28. Sud M, Fahy E, Cotter D, 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. [PubMed: 26467476]
29. McClellan J, King MC. Genetic heterogeneity in human disease. Cell. 2010; 141 (2):210–217. [PubMed: 20403315]
30. Beger RD, Dunn W, Schmidt MA, et al. Metabolomics enables precision medicine: "A White Paper, Community Perspective". Metabolomics. 2016; 12(10): 149. [PubMed: 27642271]

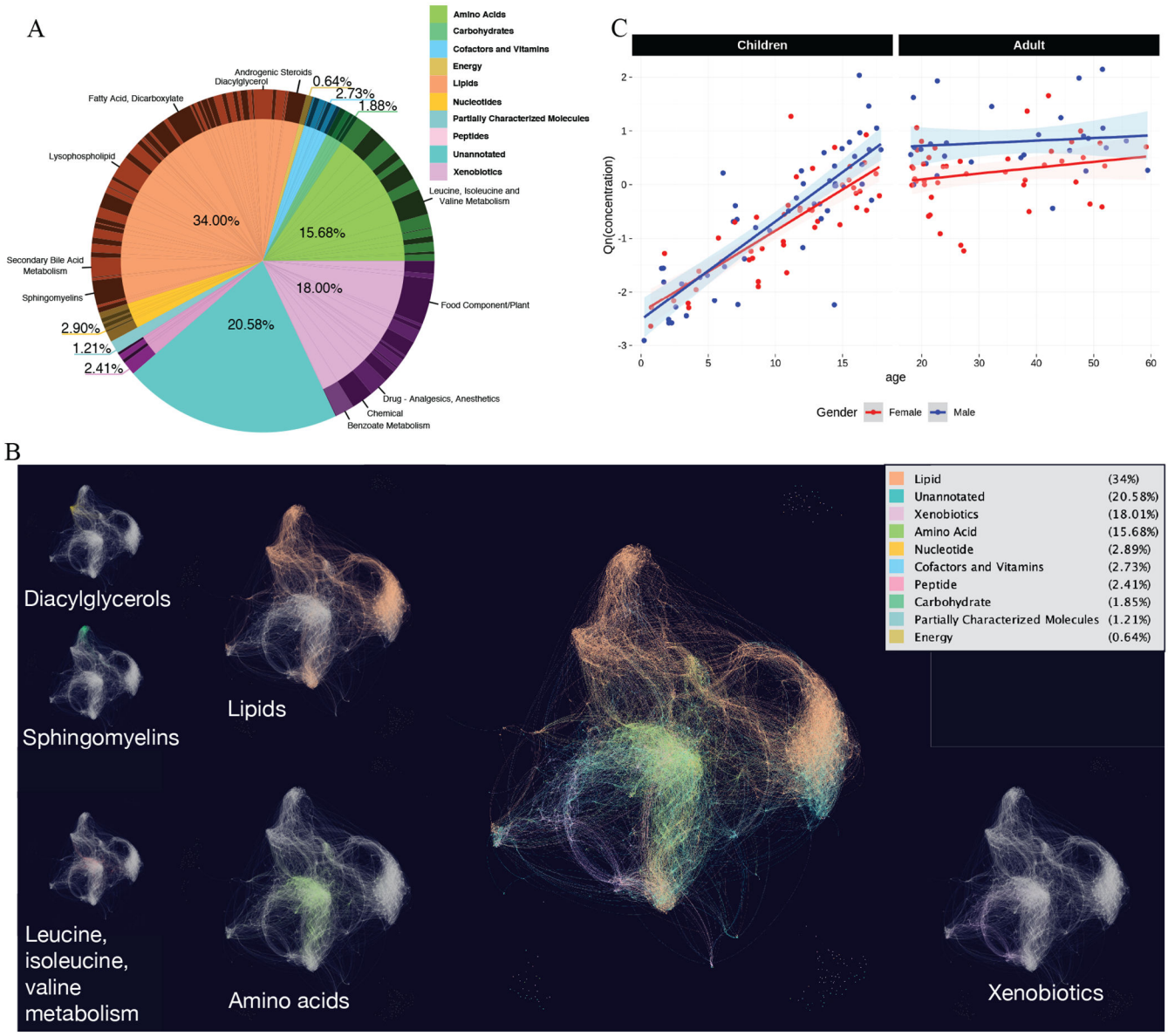

Fig. 1.

Chemical coverage and global correlation structure of 1,244 features measured by an untargeted metabolomics platform. (A) A significant proportion of measured features (N=256) are unannotated features for which chemical properties are not known although the features are consistently measured in multiple samples and showed correlations with known metabolites. Ten super-classes including lipids, amino acids, carbohydrates, vitamins, nucleotides, and xenobiotics are shown in the pie chart with subpathways in outer circle. (B) Correlation structure of metabolome. Lipids are clustered to multiple groups. Overall, amino acids, nucleotides, and carbohydrates are tightly correlated. Xenobiotics are associated with diverse endogenous metabolic pathways. (C) A total of 502 out of 1,244 features are significantly correlated with age (false discovery rate < 0.05) and correlation with age shows

a nonlinear relationship for some metabolites. For instance, creatinine concentration level is significantly correlated with age in children but not in adults.

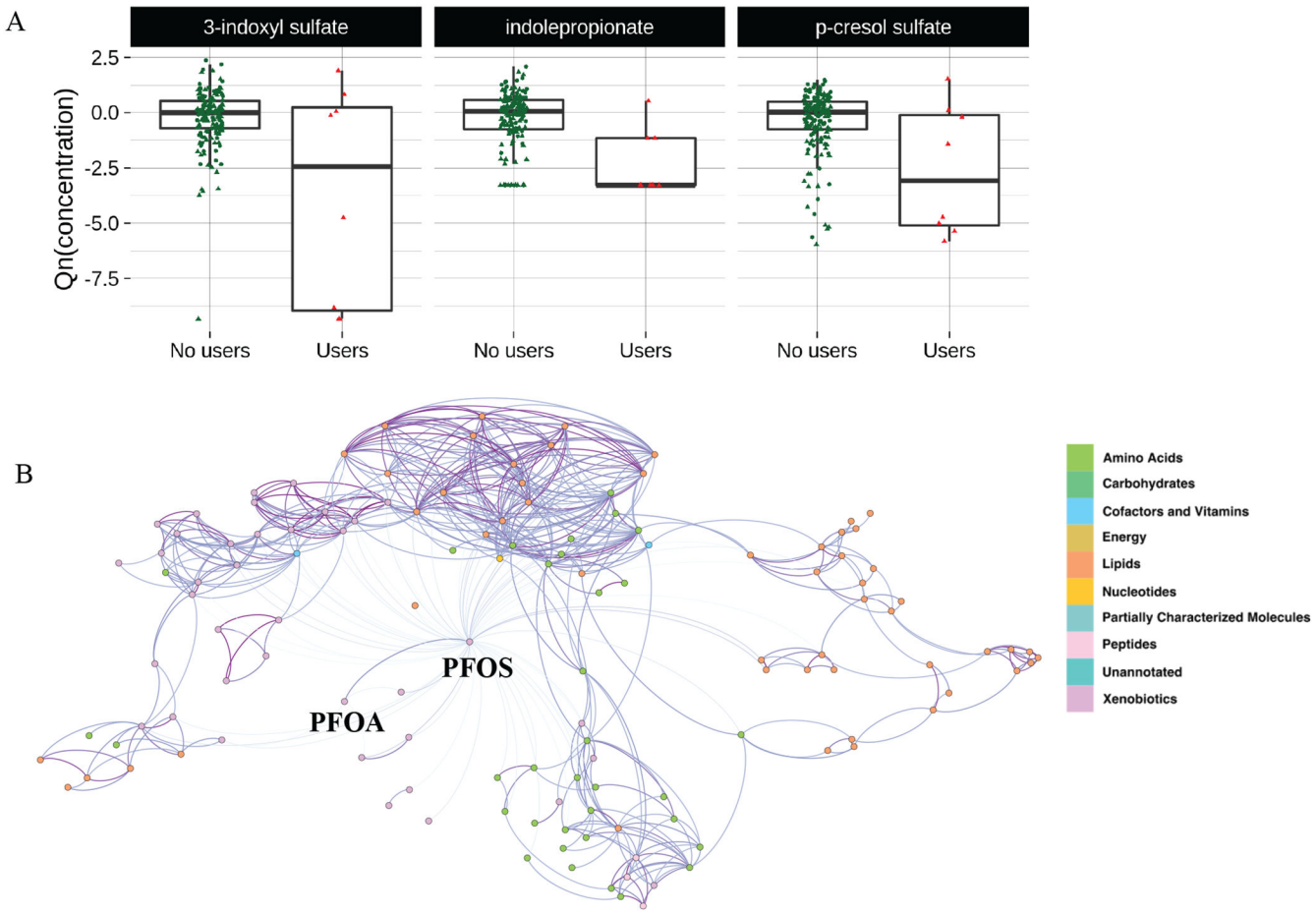

Fig. 2.

Impacts of xenobiotics and environmental chemicals on metabolome. (A) Three microbial products show significant differences in antibiotics users compared to non-users according to electronic health records. (B) A network of the metabolites significantly correlated with perfluorooctanesulfonic acid (PFOS) and perfluorooctanoic acid (PFOA). PFOS is strongly correlated with multiple metabolites (N=227) while PFOA is significantly correlated with PFOS and a few metabolites (N=65), suggesting different biological impacts of two chemical compounds of per- and polyfluoroalkyl substances (false discovery rate < 0.05). Only highly significant correlations (i.e., $|rQ_n| > 0.65$) are shown as edges.

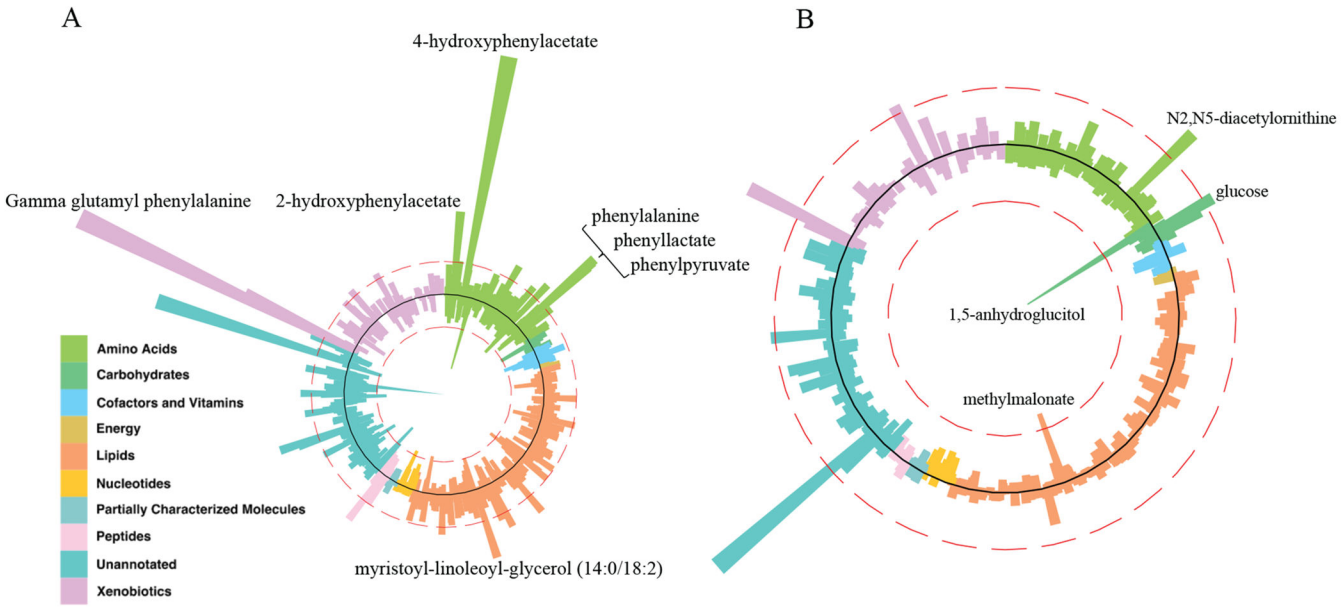

Fig. 3.

Metabolome-wide analysis of outlier features in a patient with phenylketonuria (A) and an individual with type II diabetes mellitus (B). Black solid line represents zero z-score for each metabolite (colored bars in radial). Inner and outer red dotted lines show −3 and 3 z-scores from the Q_n estimator for each feature.