Title

Structure Annotation of All Mass Spectra in Untargeted Metabolomics

Journal

Analytical Chemistry, 91(3)

ISSN

0003-2700

Authors

Blaženović, Ivana

Kind, Tobias

Sa, Michael R

et al.

Publication Date

2019-02-05

DOI

10.1021/acs.analchem.8b04698

Peer reviewed

HHS Public Access

Author manuscript

Anal Chem. Author manuscript; available in PMC 2024 September 26.

Published in final edited form as:

Anal Chem. 2019 February 05; 91(3): 2155–2162. doi:10.1021/acs.analchem.8b04698.

Structure Annotation of All Mass Spectra in Untargeted Metabolomics

Ivana Blaženović[†], **Tobias Kind**[†], **Michael R. Sa**[†], **Jian Ji**[‡], **Arpana Vaniya**[†], **Benjamin Wancewicz**[†], **Bryan S. Roberts**[†], **Hrvoje Torbašinović**[§], **Tack Lee**[‖], **Sajjan S. Mehta**[†], **Megan R. Showalter**[†], **Hosook Song**[‖], **Jessica Kwok**[†], **Dieter Jahn**[⊥,#], **Jayoung Kim**[▽,○,◆,¶], **Oliver Fiehn**[*,†]

[†]West Coast Metabolomics Center, University of California, Davis, Davis, California 95616, United States

[‡]School of Food Science, State Key Laboratory of Food Science and Technology, Jiangnan University, Wuxi, Jiangsu 330047, China

[§]Inovatus Ltd., Zagreb 10000, Croatia

[‖]Department of Urology, Inha University College of Medicine, Incheon 22212, South Korea

[⊥]Institute of Microbiology, Technische UniversitaẗBraunschweig, Braunschweig 38106, Germany

[#]Braunschweig Integrated Centre of Systems Biology (BRICS), Technische UniversitaẗBraunschweig, Braunschweig 38106, Germany

[▽]Departments of Surgery and Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States

[○]Department of Medicine, University of California Los Angeles, Los Angeles, California 90095, United States

[◆]Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, United States

[¶]Department of Urology, Ga Cheon University College of Medicine, Incheon 22212, South Korea

[*]Corresponding Author: Corresponding author address: NIH West Coast Metabolomics Center, UC Davis Genome Center, Room 1313, 451 Health Sci Drive, Davis, CA 95616.

Author Contributions

I.B., T.K., J.J., K.J., D.J., and O.F. designed the study. Data processing and compound ID were performed by I.B. Data analysis, figures, and tables were produced by I.B. and O.F. T.K. processed data using CSI:FingerID, B.R. and B.W. assisted in sample preparation, data preprocessing, and validation, J.K. and O.F. edited the manuscript, H.T. assisted with Java support, A.V. performed IDs using *in silico* fragmentation tools, M.S. provided assistance with the HILIC library, and K.J., T.L., and H.S. selected the cohort and provided the IC urine samples. The manuscript was written through the support of all authors. All authors have given approval to the final version of the manuscript.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.8b04698.

Table S1, settings used for LC-MS/MS data processing of polar metabolites and biogenic amines (XLSX)

Table S2, multilevel compound annotations (XLSX)

Table S3, HILIC library annotations of all 43 IC patients (XLSX)

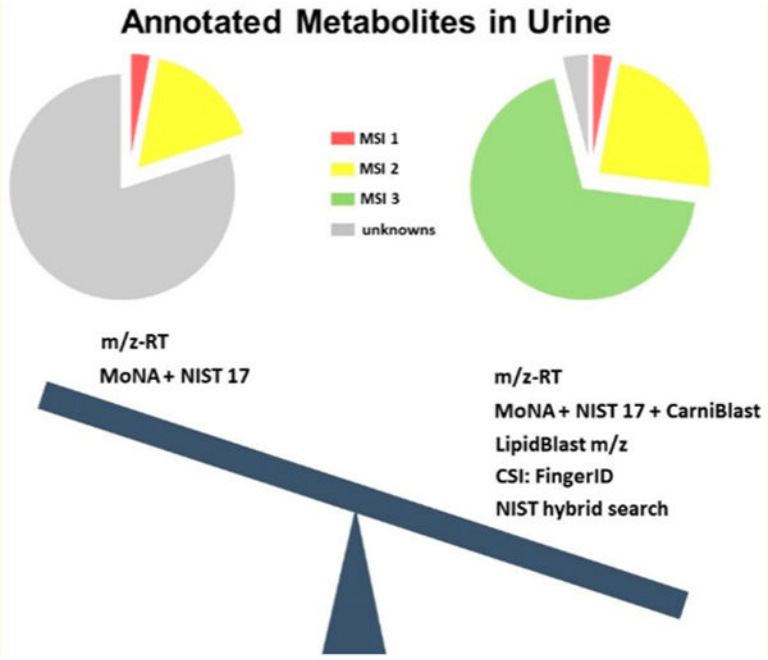
Table S4, structural classification of all spectra from 43 IC patients (XLSX)

The authors declare no competing financial interest.

Abstract

Urine metabolites are used in many clinical and biomedical studies but usually only for a few classic compounds. Metabolomics detects vastly more metabolic signals that may be used to precisely define the health status of individuals. However, many compounds remain unidentified, hampering biochemical conclusions. Here, we annotate all metabolites detected by two untargeted metabolomic assays, hydrophilic interaction chromatography (HILIC)-Q Exactive HF mass spectrometry and charged surface hybrid (CSH)-Q Exactive HF mass spectrometry. Over 9,000 unique metabolite signals were detected, of which 42% triggered MS/MS fragmentations in data-dependent mode. On the highest Metabolomics Standards Initiative (MSI) confidence level 1, we identified 175 compounds using authentic standards with precursor mass, retention time, and MS/MS matching. An additional 578 compounds were annotated by precursor accurate mass and MS/MS matching alone, MSI level 2, including a novel library specifically geared at acylcarnitines (CarniBlast). The rest of the metabolome is usually left unannotated. To fill this gap, we used the *in silico* fragmentation tool CSI:FingerID and the new NIST hybrid search to annotate all further compounds (MSI level 3). Testing the top-ranked metabolites in CSI:Finger ID annotations yielded 40% accuracy when applied to the MSI level 1 identified compounds. We classified all MSI level 3 annotations by the NIST hybrid search using the ClassyFire ontology into 21 superclasses that were further distinguished into 184 chemical classes. ClassyFire annotations showed that the previously unannotated urine metabolome consists of 28% derivatives of organic acids, 16% heterocyclics, and 16% lipids as major classes.

Graphical Abstract



Metabolomics is used as one of the major -omics tools to tackle the complex area of personalized medicine and health.[1] Target analysis of metabolites is an integral part of clinical laboratories worldwide. Conversely, untargeted metabolomics provides

comprehensive insights into complex metabolomes and allows for discovery of novel biomarkers and generating new metabolic hypothesis. Yet, untargeted metabolomics is challenged by very low identification rates.[2,3] Since there is no single platform capable of capturing the entire metabolome of urine, we have employed two chromatographic platforms that are highly suited for untargeted metabolome analysis: hydrophilic interaction chromatography (HILIC; for polar metabolite profiling) and charged-surface hybrid chromatography (CSH, for lipidomics profiling). While lipids are usually low abundant in (aqueous) urine samples, recent technological advancements of high-resolution mass spectrometry (MS) have largely improved the comprehensive lipid profiling of cells, tissues, and biofluids, including urine. Lipids can serve as important biomarkers even in urine samples, for example for prostate cancer[4] or segmental glomerulosis.[5] Combined, metabolomics and lipidomics reveal biologically active metabolites in urine and provide a diagnostic chemical signature of human metabolic phenotypes. The urinary metabolome is associated with urological diseases, including bladder dysfunctions such as interstitial cystitis/bladder pain syndrome (IC).[6–8] IC is characterized by chronic bladder and/or pelvic pain, as well as nocturia and an increase in urinary frequency and urgency.[9–11] The work presented here investigated how many urine metabolites from IC patients could be identified, as defined by the Metabolomics Standards Initiative (MSI),[12] using freely available comprehensive metabolite annotation tools, novel databases, and libraries that were developed and used here for the first time.[13]

EXPERIMENTAL SECTION

Extraction.

Subjects, urine specimen collection, and clinical and pathological features of subjects were described in a previous paper from our laboratory.[14] Deidentified urine samples were stored at −80 °C until further analysis. Urinary lipids were extracted with methanol and methyl *tert*-butyl ether both containing a cocktail of lipid standards.[15] Water was subsequently added for phase separation. This extraction protocol extracts all main lipid classes in urine with high recoveries, specifically phosphatidylcholines (PC), sphingomyelins (SM), phosphatidylethanolamines (PE), lysophospha-tidylcholines (LPC), ceramides (Cer), cholesteryl esters (CholE), and triacylglycerols (TG).[16] Lipid standards were purchased from Avanti Polar lipids (Alabaster, USA). After concentrating extracts to complete dryness, samples were reconstituted prior to LC-MS analysis as published before.[17] Polar metabolites were retrieved by using the polar phase of the lipid extraction procedure. Samples were dried in a centrivap prior to a cleanup step of 50% acetonitrile and dried again. Samples were reconstituted for HILIC-MS analysis in an 80:20 acetonitrile:water solution containing internal standards from Sigma and CDN Isotopes.

Instrumentation.

All measurements were carried out on a Thermo Q Exactive instrument. For lipidomics measurements, 1 μL of diluted samples was separated on a Waters Acquity UPLC CSH C18 column (100 × 2.1 mm; 1.7 μm) coupled to an Acquity UPLC CSH C18 VanGuard precolumn (5 × 2.1 mm; 1.7 μm). The column was maintained at 65 °C with a flow rate of 0.6 mL/min. The positive ionization mobile phases consisted of (A)

acetonitrile:water (60:40, v/v) with ammonium formate (10 mM) and formic acid (0.1%) and (B) 2-propanol:acetonitrile (90:10, v/v) with ammonium formate (10 mM) and formic acid (0.1%). The negative ionization mobile phases consisted of (A) acetonitrile:water (60:40, v/v) with ammonium formate (10 mM) and (B) 2-propanol:acetonitrile (90:10, v/v) with ammonium formate (10 mM). The separation was conducted under the following gradient: 0 min 15% B; 0–2 min 30% B; 2–2.5 min 48% B; 2.5–11 min 82% B; 11–11.5 min 99% B; 11.5–12 min 99% B; 12–12.1 min 15% B; 12.1–15 min 15% B. The Q Exactive MS instrument was operated using positive mode electrospray ionization using the following parameters: Mass range, 120−1200 *m*/*z*; Sheath gas flow rate, 60; Aux gas flow rate, 25; Sweep gas flow rate, 2; Spray Voltage (kV) 3.6; Capillary temp, 300 °C; S-lens RF level, 50; Aux gas heater temp, 370 °C. Full MS parameters: Resolution, 60,000; AGC target, 1e6; Maximum IT, 100 ms; Spectrum data type, Centroid. Data dependent MS2 parameters: Resolution, 15,000; AGC target, 1e5; Maximum IT, 50 ms; Loop count, 4; TopN, 4; Isolation Window, 1.0 *m*/*z*; Fixed First Mass, 70.0 *m*/*z*; (N)CE/stepped (N)CE, 20, 30, 40; Spectrum data type, Centroid.

For profiling polar compounds and biogenic amines, HILIC-Q Exactive MS/MS data acquisition was performed. One μL of diluted samples was separated on a Waters Acquity UPLC BEH Amide column (150 × 2.1 mm; 1.7 μm) coupled to an Acquity UPLC BEH Amide VanGuard precolumn (5 × 2.1 mm; 1.7 μm). The column was maintained at 45 °C with a flow rate of 0.4 mL/min. The mobile phases consisted of (A) water with ammonium formate (10 mM) and formic acid (0.125%) and (B) acetonitrile:water (95:5, v/v) with ammonium formate (10 mM) and formic acid (0.125%). The separation was conducted under the following gradient: 0 min 100% B; 0–2 min 100% B; 2–7.7 min 70% B; 7.7–9.5 min 40% B; 9.5–10.25 min 30% B; 10.25–12.75 min 100% B; 12.75–17 min 100% B.

The Q Exactive MS instrument was operated using positive mode electrospray ionization (ESI HILIC) with the following parameters: Mass range, 60–900 *m*/*z*; Sheath gas flow rate, 60; Aux gas flow rate, 25; Sweep gas flow rate, 2; Spray Voltage (kV) 3.6; Capillary temp, 300 °C; S-lens RF level, 50; Aux gas heater temp, 370 °C. Full MS parameters: Microscans, 1; Resolution, 60,000; AGC target, 1e6; Maximum IT, 100 ms; Number of scans, 1; Spectrum data type, Centroid. Data dependent MS2 parameters: Microscans, 1; Resolution, 15,000; AGC target, 1e5; Maximum IT, 50 ms; Loop count, 4; MSX count, 1; TopN, 4; Isolation Window, 1.0 *m*/*z*; Isolation offset 0.0 *m*/*z*; (N)CE/stepped (N)CE, 20, 30, 40; Spectrum data type, Centroid.

Data Processing and Compound Identification.

The LC-MS/MS data was analyzed by MS-DIAL software.[18] Detailed parameter settings are listed in Supplemental Table 1 (HILIC and lipidomics data processing settings). Data tables containing accurate masses, retention times, and peak heights were exported, and further analysis was performed in R and Metabox.[19] Automated annotation of metabolites was performed separately for polar metabolites and lipids. Table S1 lists libraries, methods, and software used for each platform.

Metabolite annotations were achieved using a combination of different tools. On MSI level 1, we developed and used a novel HILIC-MS/MS library of 1,102 authentic standards

including retention time, precursor mass, and MS/MS spectra. All spectra, retention times, and chromatography conditions are freely available at MassBank of North America (http://massbank.us). Search windows were used as follows: 0.1 min RT tolerance (for the alignment of peaks), 0.0001 Da tolerance for the precursor masses, and 0.05 Da tolerance for the MS/MS spectral matching. Similarly, we used lipid retention times and MS/MS spectra for lipidomics identifications.[15] On MSI level 2, we annotated compounds that did not trigger MS/MS fragmentations in data dependent mode but that were still identified based on accurate mass and retention time using the HILIC-MS/MS library in addition to manually curated lipid retention times. Moreover, MSI level 2 annotations were also based on accurate mass and MS/MS annotations for spectra for which no authentic retention time library was available, such as the NIST17, HMDB,[20] GNPS,[21] the new CarniBlast library, and the LipidBlast libraries.[17,22,23] For MSI level 3 annotations, we used CSI:FingerID,[24] the NIST-Hybrid Search,[25] and LipidBlast accurate mass search services.

RESULTS AND DISCUSSION

MSI Level 1 Annotations.

The number of precursors that triggered MS/MS fragmentations was sample dependent. Table S2 contains all 3,894 merged spectra for all samples that were aligned and processed by MS-DIAL software which were subsequently used for MSI level 1 and 2 annotations (Table 1). Compound identifications with the highest level of confidence (MSI level 1) were achieved using libraries of authentic standards. All library spectra and retention times were acquired under identical conditions as the experimental urine spectra. Specifically, a new HILIC-Q Exactive MS/MS library was established using 1,102 authentic compounds measured in positive mode. Data and metadata for this library can be downloaded from MassBank of North America (Fiehn HILIC). By matching experimental urine spectra against library retention times (RT), accurate precursor masses (m/z), and MS/MS spectra, overall 175 compounds were identified at MSI level 1. Specifically, we identified 72 lipids in CSH-Q Exactive MS/MS as members of 7 lipid classes and 103 hydrophilic compounds using HILIC-Q Exactive MS/MS as amino acids, biogenic amines, and other polar compound classes (Table S2). Detailed settings and cutoffs are listed in Table S1.

MSI Level 2 Annotations.

Retention-time based libraries of authentic standards are necessarily smaller than the complement of available MS/MS spectra in public or licensed mass spectral libraries. Therefore, it is a common practice in metabolomics research to perform mass spectral similarity searches of experimental to library MS/MS spectra to increase the annotation rate. While metabolite MS/MS fragmentations are independent of chromatography conditions, spectra often show differences due to slightly different fragmentation parameters or different mass spectrometers used. In addition, many metabolites show only a few characteristic fragment ions, rendering the use of classic spectral similarity searches unreliable. To retain high confidence, we combined accurate precursor mass and MS/MS searches for all over queries, using 750 dot-product score (HILIC-MS/MS) and 400 reverse dot-product (CSH-MS/MS) as lower threshold below which no further correct match hits were expected.

Subsequently, each spectrum was manually inspected to verify spectral matches and retention time matches, where available.

MassBank of North America (MoNA; http://massbank.us) currently contains over 260,000 mass spectra from 15 individual mass spectral repositories such as MassBank, MassBank EU, GNPS, ReSpect, LipidBlast, MetaboBase, and HMDB, covering more than 80,000 compounds. We combined MoNA spectra with MS/MS data from the NIST17 library, the largest available licensed repository with over 550,000 experimental spectra from 13,808 chemical compounds. We merged all spectra from both resources into one.msp within MS-DIAL software for mass spectral similarity matching. In total, this approach yielded 480 identified compounds on MSI level 2.

While investigating and validating MS/MS spectra, we observed many spectra that appeared similar to 17 acylcarnitines annotated by using LipidBlast or MoNA spectra. Acylcarnitines in urine serve as biomarkers for bladder cancer,[26] diabetic nephropathy,[27] obesity,[28] and human kidney cancer.[29] The identification of acylcarnitines has to be performed either using authentic reference compounds or with reference library spectra.[30] However, only a few tandem mass spectra of acylcarnitines exist in commercial (NIST) and open mass spectral libraries (MassBank,[31] METLIN,[32] Respect DB[33]), covering less than 50 acylcarnitine structures. Conversely, when using a structure similarity search in the CAS SciFinder literature database, we found 453 acylcarnitine-like structures of which only 62 were commercially available. Such finding indicated a high chemical diversity of acylcarnitines that could not possibly be closed by purchasing more chemical compounds. To overcome this gap and identify all urinary acylcarnitines, we developed an *in silico* tandem mass spectral library of acylcarnitines using structure templates[23] similar to our previous LipidBlast[22] and FAHFA predicted MS/MS libraries.[34] Here, we constructed the CarniBlast library of 2,400 acylcarnitine species covering a wide range of saturated, unsaturated, -hydroxyl, -keto, -dicarboxylic, and oxidized acyl chain substituted acylcarnitines. We matched all experimental MS/MS spectra from both polar and lipidomics profiling against the new *in silico* database of acylcarnitines. After removing duplicates and manually validating each candidate spectrum, we identified 67 novel acylcarnitines through the CarniBlast library in addition to the 17 acylcarnitines obtained by LipidBlast and MoNA. Detailed settings and cutoffs are listed in Table S1.

All urinary metabolomic data were acquired by data-dependent MS/MS. Yet, we used retention-time based MS/MS libraries such as the new HILIC-Q Exactive MS/MS repository. For lipids, we have recently shown[35] that compound annotations can be based on accurate precursor mass and retention time alone with high confidence. Using these two orthogonal parameters (m/z and RT), six further acylcarnitines were annotated at MSI level 2. In the same manner, we assigned 201 further compounds that were too low abundant to trigger MS/MS fragmentations but that were covered in our m/z and RT libraries that were acquired under the same experimental conditions.

In combination, we identified 578 metabolites at MSI level 2 confidence (39 lipids, 85 acylcarnitines, and 454 hydrophilic compounds). Metadata such as dot product, reverse dot

product scores, number of matching ions, and MS-DIAL calculated MS2 similarity have been taken into account. Detailed results are listed in Table S2.

MSI Level 3 Annotations.

When combining compound annotations on MSI level 1 and 2, the 753 compound annotations only covered 19.3% of the acquired urinary MS/MS spectra (Figure 1). While this number of annotated compounds is already significantly higher than most other studies on urine metabolomics,[36,37] it is worrisome that more than 80% of all MS/MS spectra remain unannotated in metabolomics screens. Many biologists will focus their attention only on identified compounds and not even perform statistical assessments on complete metabolome data, including unknowns. Yet, it appears very likely that the fraction of more than 80% unknowns might include very important biomarkers or signatures of diseases, food patterns, exposome compounds, or other significant chemicals. We therefore used three tools to investigate this dark matter of metabolomics[38] closer: (a) accurate mass search, (b) structure elucidation tools, and (c) mass spectral library hybrid search. We used the most exhaustive MS/MS files for MSI level 3 annotations, using raw MS/MS spectra from individual IC patients for each stepped collision energy to enable best-possible annotations. Spectra were exported as either mgf or msp files and used in the different software programs. For HILIC-MS/MS analyses, we used between 5,192 and 6,447 MS/MS spectra; for CSH-MS/MS lipidomics analyses, the number of spectra ranged from 5,705 to 7,050 MS/MS spectra per sample (Table 1 and Table S2). The number of raw spectra is higher than in the MS-DIAL processed file because MS-DIAL merged the stepped collision energies during data acquisition.

First, we used precursor mass lookups. In general, simple mass lookups yield many false discoveries due to a plethora of isomers and isobars at a given accurate mass level. This problem is especially pronounced in HILIC-MS for which hardly any constraint can be applied with respect to the number of possible chemicals. Yet, for lipidomics assays, lipids can already be assigned to specific lipid classes with some level of confidence based on m/z and retention time (in relation to MSI level 1 and 2 annotated lipids within the same study). Additional structure information on such lipid class annotations and their acyl chains cannot be made using accurate mass alone. We extracted m/z precursor information from the MS-DIAL output in positive ionization mode and used the m/z lookup macro within LipidBlast v49 to assign additional lipids. We used a 5 mDa mass tolerance for lipid assignments based on the mass accuracy of the Q Exactive instrument. This way, an additional 96 lipid annotations were obtained for the lipidomics data set in positive ionization mode.

Second, we used cheminformatics tools to annotate accurate mass HILIC-MS/MS spectra to likely chemical structures. A range of software tools has been published such as MS-FINDER,[39] Sirius, CFM-ID,[40] and others. Here, we used two programs, Sirius 4.0 with CSI:FingerID interface[24] and the new NIST17 hybrid-search.[25] Sirius/CSI:FingerID scored highly during the latest CASMI structure identification challenges[41,42] but has never been applied to urinary metabolomics. The NIST17 hybrid-search software was released after the CASMI challenges but offers advantages by greatly expanding the utility of existing mass spectral libraries. For Sirius/CSI:Finger ID, MS/MS spectra were exported as a MGF

file from each raw file using the MSConvert program. The largest MGF file contained 6,447 MS/MS spectra. Formulas were assigned at 10 ppm search windows, retaining the 10 best formula candidates. Using Sirius, spectra were processed within 2 min on a 16 CPU machine, assigning formulas to a total of 6,184 features (96%). Subsequently, CSI:FingerID performed spectral fingerprint matching via a Web service to annotate isomer structures. Within 5 min processing time, 728 MS/MS spectra were assigned to chemical structures in the biodatabase, a filtered version of Pubchem containing over 270,000 structures of biological interest, and 1,557 MS/MS spectra were assigned to chemicals in the much larger PubChem database. Scored results of all isomeric structures were exported as CSV files. For structures returned by biodatabase searches, CSI:Finger ID yielded up to 130 results per MS/MS scan and up to 10,000 structure candidates per MS/MS spectrum in PubChem queries. Time-consuming manual investigations have to be performed to select the most likely structures. To test CSI:Finger ID accuracy, we selected 103 MS/MS spectra from the urinary HILIC-Q Exactive MS/MS data set that were unambiguously assigned by authentic standards and tested these spectra within 5 ppm mass accuracy and a biodatabase structure query. Using our publicly available HILIC library (see above), 41 compounds (40%) were correctly annotated by CSI:Finger ID as top hit, and 54 MS/MS spectra (52%) were correctly assigned within the top 3 isomer candidates. Detailed results are given in Table S3. CSI:FingerID is not optimized for use in lipidomics MS/MS spectra.

Third, we used the novel NIST17 hybrid search[25] that combines mass spectral library-based scoring with calculating fragment and precursor mass shifts for chemical modifications of library structures. Such a tool mimics the experience of well-trained chemists[43] because known biochemical modifications such as methylations or acetylations produce epimetabolites that are removed from their classic functions in canonical metabolic pathways.[44] Four examples of how the NIST17 hybrid search works are given in Figure 2 for head-to-tail MS/MS spectral comparisons of methyladenosine/adenosine, phosphothreonine/threonine, hydroxyarginine/arginine, and acetylmethionine/methionine. Spectra of modified metabolites show distinct shifts in precursor masses and fragments when comparing to nonmodified library spectra. Yet, the NIST17 hybrid search correctly associated the modified spectra with their best scoring related library spectra.

We exported all lipidomics and HILIC-MS/MS spectra from 43 interstitial cystitis patients from MS-DIAL to the NIST pepSearch software and used the NIST17 hybrid search function. The software supports batch processing, enables users to include or exclude specific MS/MS spectra, and yields quick overviews of the complement of chemical structures in mass spectral profiling studies. Results are given in Table S2. Within 10 min processing time, 95% of all spectra were assigned with structure annotations and compound names, including a set of confidence scores such as forward and reverse dot products and a probability score. Hybrid search annotations must be treated with caution as they do not represent an identification but rather a nearest known neighbor to the unknown spectrum. While in many cases NIST17 hybrid searches give correct results (Figure 2), overall results highlight a high probability of chemical class annotations (MSI level 3) rather than exact structures. We therefore used these results for exactly this purpose, classifying the thousands of patient urinary MS/MS spectra to chemical classes. For this purpose, we implemented a batch search version of the automated chemical hierarchy

classification system ClassyFire[45] (https://cfb.fiehnlab.ucdavis.edu). ClassyFire requires International Chemical Identifiers (InChI keys) as input.[46] Today, InChI keys are a standard tool in all chemical and biochemical databases to assign and compare chemical structures with machine-readable, unique keys. The ClassyFire Batch search utilizes the ClassyFire API to look up provided InChIKeys and, if no match is found, to query its nonstereo form. It yields a tabular CSV version of the results. We used the online tool Chemical Translation Services (CTS)[47] to convert chemical names from the NIST17 hybrid pepSearch results to InChI keys. This conversion reduced the compound list by ~1% because some NIST17 hybrid search chemical names are not yet included in PubChem or the other 200 chemical databases that support the CTS tool. Hence, of the average number of 5,250 MS/MS spectra found per patient in lipidomics and HILIC-MS/MS, about 95% of all spectra were now annotated by exact chemical structures or by chemical classes (MSI level 3, Figure 1). Results of classifications are organized into Kingdom, Superclass, Class, Subclass, and two parent levels. Detailed results are given in Table S4, with varying chemical classes present in urines of different patients. An average MSI level 3 classification is given in Figure 3 using the superclass and subclass classifications as defined by ClassyFire. Roughly one-third of the urinary metabolome was classified as chemicals containing aromatic rings or heterocycles, one-third was classified as compounds containing ketones, alcohols, or acids, while the remaining one-third consisted of lipids, phenylpropanoids, and nitrogenous- or sulfur containing organics. As expected, organic phosphates, organometallic, or other compound classes comprised less than 1% of the urinary metabolome.

CONCLUSION

Unlike proteomic MS/MS spectra assignments, the field of metabolomics currently lacks generally accepted and validated automated calculations of compound identification confidence levels with false-discovery rate assessments. As remedy, structure annotation in untargeted MS/MS metabolomics reports must be annotated with MSI confidence levels to detail which metabolites can be trusted and used for metabolic pathway annotations (MSI level 1 and 2), especially if annotated spectra use accurate mass information and manual curation. While the majority of acquired MS/MS spectra cannot be annotated with certainty to specific chemicals, Sirius/CSI:Finger ID and NIST17 hybrid search results yield many structure assignments that are worthy to be validated by acquiring spectra from corresponding authentic chemicals. In addition, MSI level 3 chemical classes can be ordered by ClassyFire and used for chemical class enrichment statistics,[48] for example, in biomarker discovery studies. Moreover, MSI level 3 classifications may yield differences in urinary chemicals that detail differences in subjects due to diet and chemical exposure in epidemiology studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

The authors acknowledge support from National Institutes of Health grant U2C ES030158 (to O.F.) and the following grants to J.K.: NIH U01 DK103260, R01 DK100974, U01 DP006079, NIH NCATS UCLA CTSI UL1

TR000124, Department of Defense grant W81XWH-15-1-0415, IMAG-INE NO IC research grant, the Steven Spielberg Discovery Fund in Prostate Cancer Research Career Development Award, and the U.S.-Egypt Science and Technology Joint Fund. The funders had no role in the design, data collection, and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

(1). Jacob M; Lopata AL; Dasouki M; Abdel Rahman AM Mass Spectrom. Rev 2017, DOI: 10.1002/mas.21548.

(2). Bloszies CS; Fiehn O Current Opinion in Toxicology 2018, 8, 87–92.

(3). Blazenovic I; Kind T; Ji J; Fiehn O Metabolites 2018, 8, 31. [PubMed: 29748461]

(4). Yang JS; Lee JC; Byeon SK; Rha KH; Moon MH Anal. Chem 2017, 89, 2488–2496. [PubMed: 28192938]

(5). Erkan E; Zhao X; Setchell K; Devarajan P Pediatr. Nephrol 2016, 31, 581–588. [PubMed: 26537928]

(6). Antunes-Lopes T; Cruz CD; Cruz F; Sievert KD Current opinion in urology 2014, 24, 352–357. [PubMed: 24841379]

(7). Kuo HC Int. J. Urol 2014, 21 (S1), 34–41. [PubMed: 24807491]

(8). Pedroza-Diaz J; Rothlisberger S Biochemia medica 2015, 25, 22–35. [PubMed: 25672464]

(9). Clemens JQ; Clauw DJ; Kreder K; Krieger JN; Kusek JW; Lai HH; Rodriguez L; Williams DA; Hou X; Stephens A; Landis JR; MAPP Research Network. J. Urol 2015, 193, 1554–1558. [PubMed: 25463989]

(10). Naliboff BD; Stephens AJ; Afari N; Lai H; Krieger JN; Hong B; Lutgendorf S; Strachan E; Williams D; MAPP Research Network. Urology 2015, 85, 1319–1327. [PubMed: 26099876]

(11). Hanno PM; Burks DA; Clemens JQ; Dmochowski RR; Erickson D; Fitzgerald MP; Forrest JB; Gordon B; Gray M; Mayer RD; Newman D; Nyberg L Jr.; Payne CK; Wesselmann U; Faraday MM J. Urol 2011, 185, 2162–2170. [PubMed: 21497847]

(12). Schymanski EL; Jeon J; Gulde R; Fenner K; Ruff M; Singer HP; Hollender J Environ. Sci. Technol 2014, 48, 2097–2098. [PubMed: 24476540]

(13). Kind T; Tsugawa H; Cajka T; Ma Y; Lai Z; Mehta SS; Wohlgemuth G; Barupal DK; Showalter MR; Arita M; Fiehn O Mass Spectrom. Rev 2018, 37, 513–532. [PubMed: 28436590]

(14). Wen H; Lee T; You S; Park SH; Song H; Eilber KS; Anger JT; Freeman MR; Park S; Kim JJ Proteome Res 2015, 14, 541–548.

(15). Cajka T; Fiehn O: LC–MS-Based Lipidomics and Automated Identification of Lipids Using the LipidBlast In-Silico MS/MS Library. In Lipidomics: Methods and Protocols; Bhattacharya SK, Ed.; Springer New York: New York, NY, 2017; pp 149–170, DOI: 10.1007/978-1-4939-6996-8_14.

(16). Matyash V; Liebisch G; Kurzchalia TV; Shevchenko A; Schwudke D J. Lipid Res 2008, 49, 1137–1146. [PubMed: 18281723]

(17). Cajka T; Smilowitz JT; Fiehn O Anal. Chem 2017, 89, 12360–12368. [PubMed: 29064229]

(18). Tsugawa H; Cajka T; Kind T; Ma Y; Higgins B; Ikeda K; Kanazawa M; VanderGheynst J; Fiehn O; Arita M Nat. Methods 2015, 12, 523–526. [PubMed: 25938372]

(19). Wanichthanarak K; Fan S; Grapov D; Barupal DK; Fiehn O PLoS One 2017, 12, No. e0171046. [PubMed: 28141874]

(20). Wishart DS; Feunang YD; Marcu A; Guo AC; Liang K; Vazquez-Fresno R; Sajed T; Johnson D; Li C; Karu N; Sayeeda Z; Lo E; Assempour N; Berjanskii M; Singhal S; Arndt D; Liang Y; Badran H; Grant J; Serra-Cayuela A; Liu Y; Mandal R; Neveu V; Pon A; Knox C; Wilson M; Manach C; Scalbert A Nucleic Acids Res 2018, 46, D608–D617. [PubMed: 29140435]

(21). Wang M; Carver JJ; Phelan VV; Sanchez LM; Garg N; Peng Y; Nguyen DD; Watrous J; Kapono CA; Luzzatto-Knaan T; Porto C; Bouslimani A; Melnik AV; Meehan MJ; Liu WT; Crusemann M; Boudreau PD; Esquenazi E; Sandoval-Calderon M; Kersten RD; Pace LA; Quinn RA; Duncan KR; Hsu CC; Floros DJ; Gavilan RG; Kleigrewe K; Northen T; Dutton RJ; Parrot D; Carlson EE; Aigle B; Michelsen CF; Jelsbak L; Sohlenkamp C; Pevzner P; Edlund A; McLean J; Piel J; Murphy BT; Gerwick L; Liaw CC; Yang YL; Humpf HU; Maansson M; Keyzers

RA; Sims AC; Johnson AR; Sidebottom AM; Sedio BE; Klitgaard A; Larson CB; Boya P CA; Torres-Mendoza D; Gonzalez DJ; Silva DB; Marques LM; Demarque DP; Pociute E; O'Neill EC; Briand E; Helfrich EJN; Granatosky EA; Glukhov E; Ryffel F; Houson H; Mohimani H; Kharbush JJ; Zeng Y; Vorholt JA; Kurita KL; Charusanti P; McPhail KL; Nielsen KF; Vuong L; Elfeki M; Traxler MF; Engene N; Koyama N; Vining OB; Baric R; Silva RR; Mascuch SJ; Tomasi S; Jenkins S; Macherla V; Hoffman T; Agarwal V; Williams PG; Dai J; Neupane R; Gurr J; Rodriguez AMC; Lamsa A; Zhang C; Dorrestein K; Duggan BM; Almaliti J; Allard PM; Phapale P Nat. Biotechnol 2016, 34, 828–837. [PubMed: 27504778]

(22). Kind T; Liu KH; Lee DY; DeFelice B; Meissen JK; Fiehn O Nat. Methods 2013, 10, 755–758. [PubMed: 23817071]

(23). Kind T; Okazaki Y; Saito K; Fiehn O Anal. Chem 2014, 86, 11024–11027. [PubMed: 25340521]

(24). Duhrkop K; Shen H; Meusel M; Rousu J; Bocker S Proc. Natl. Acad. Sci. U. S. A 2015, 112, 12580–12585. [PubMed: 26392543]

(25). Burke MC; Mirokhin YA; Tchekhovskoi DV; Markey SP; Heidbrink Thompson J; Larkin C; Stein SE J. Proteome Res 2017, 16, 1924–1935. [PubMed: 28367633]

(26). Kim WT; Yun SJ; Yan C; Jeong P; Kim YH; Lee IS; Kang HW; Park S; Moon SK; Choi YH; Choi YD; Kim IY; Kim J; Kim WJ Yonsei Med. J 2016, 57, 865–871. [PubMed: 27189278]

(27). Mirzoyan K; Klavins K; Koal T; Gillet M; Marsal D; Denis C; Klein J; Bascands JL; Schanstra JP; Saulnier-Blache JS Biochem. Biophys. Res. Commun 2017, 487, 109–115. [PubMed: 28396151]

(28). Rauschert S; Uhl O; Koletzko B; Hellmuth C Ann. Nutr. Metab 2014, 64, 314–324. [PubMed: 25300275]

(29). Ganti S; Taylor SL; Kim K; Hoppel CL; Guo L; Yang J; Evans C; Weiss RH Int. J. Cancer 2012, 130, 2791–2800. [PubMed: 21732340]

(30). Solberg HE; Bremer J Biochim. Biophys. Acta, Gen. Subj 1970, 222, 372–380.

(31). Horai H; Arita M; Kanaya S; Nihei Y; Ikeda T; Suwa K; Ojima Y; Tanaka K; Tanaka S; Aoshima K; Oda Y; Kakazu Y; Kusano M; Tohge T; Matsuda F; Sawada Y; Hirai MY; Nakanishi H; Ikeda K; Akimoto N; Maoka T; Takahashi H; Ara T; Sakurai N; Suzuki H; Shibata D; Neumann S; Iida T; Tanaka K; Funatsu K; Matsuura F; Soga T; Taguchi R; Saito K; Nishioka TJ Mass Spectrom 2010, 45, 703–714.

(32). Zhu Z-J; Schultz AW; Wang J; Johnson CH; Yannone SM; Patti GJ; Siuzdak G Nat. Protoc 2013, 8, 451–460. [PubMed: 23391889]

(33). Sawada Y; Nakabayashi R; Yamada Y; Suzuki M; Sato M; Sakata A; Akiyama K; Sakurai T; Matsuda F; Aoki T; Hirai MY; Saito K Phytochemistry 2012, 82, 38–45. [PubMed: 22867903]

(34). Ma Y; Kind T; Vaniya A; Gennity I; Fahrmann JF; Fiehn OJ Cheminf 2015, 7, 53.

(35). Blazenovic I; Shen T; Mehta SS; Kind T; Ji J; Piparo M; Cacciola F; Mondello L; Fiehn O Anal. Chem 2018, 90, 10758–10764. [PubMed: 30096227]

(36). Bouatra S; Aziat F; Mandal R; Guo AC; Wilson MR; Knox C; Bjorndahl TC; Krishnamurthy R; Saleem F; Liu P; Dame ZT; Poelzer J; Huynh J; Yallou FS; Psychogios N; Dong E; Bogumil R; Roehring C; Wishart DS PLoS One 2013, 8, No. e73076. [PubMed: 24023812]

(37). Zaghlool SB; Mook-Kanamori DO; Kader S; Stephan N; Halama A; Engelke R; Sarwath H; Al-Dous EK; Mohamoud YA; Roemisch-Margl W; Adamski J; Kastenmuller G; Friedrich N; Visconti A; Tsai PC; Spector T; Bell JT; Falchi M; Wahl A; Waldenberger M; Peters A; Gieger C; Pezer M; Lauc G; Graumann J; Malek JA; Suhre K Hum. Mol. Genet 2018, 27, 1106–1121. [PubMed: 29325019]

(38). da Silva RR; Dorrestein PC; Quinn RA Proc. Natl. Acad. Sci. U. S. A 2015, 112, 12549–12550. [PubMed: 26430243]

(39). Tsugawa H; Kind T; Nakabayashi R; Yukihira D; Tanaka W; Cajka T; Saito K; Fiehn O; Arita M Anal. Chem 2016, 88, 7946–7958. [PubMed: 27419259]

(40). Allen F; Pon A; Wilson M; Greiner R; Wishart D Nucleic Acids Res 2014, 42, W94–99. [PubMed: 24895432]

(41). Schymanski EL; Ruttkies C; Krauss M; Brouard C; Kind T; Duhrkop K; Allen F; Vaniya A; Verdegem D; Bocker S; Rousu J; Shen H; Tsugawa H; Sajed T; Fiehn O; Ghesquiere B; Neumann S J. Cheminf 2017, 9, 22.

(42). Blazenovic I; Kind T; Torbasinovic H; Obrenovic S; Mehta SS; Tsugawa H; Wermuth T; Schauer N; Jahn M; Biedendieck R; Jahn D; Fiehn O J. Cheminf 2017, 9, 32.

(43). Nikolic D; Macias C; Lankin DC; van Breemen RB Rapid Commun. Mass Spectrom 2017, 31, 1385–1395. [PubMed: 28558170]

(44). Showalter MR; Cajka T; Fiehn O Curr. Opin. Chem. Biol 2017, 36, 70–76. [PubMed: 28213207]

(45). Djoumbou Feunang Y; Eisner R; Knox C; Chepelev L; Hastings J; Owen G; Fahy E; Steinbeck C; Subramanian S; Bolton E; Greiner R; Wishart DS J. Cheminf 2016, 8, 61.

(46). Heller SR; McNaught A; Pletnev I; Stein S; Tchekhovskoi D J. Cheminf 2015, 7, 23.

(47). Wohlgemuth G; Haldiya PK; Willighagen E; Kind T; Fiehn O Bioinformatics 2010, 26, 2647–2648. [PubMed: 20829444]

(48). Barupal DK; Fiehn O Sci. Rep 2017, 7, 14567. [PubMed: 29109515]

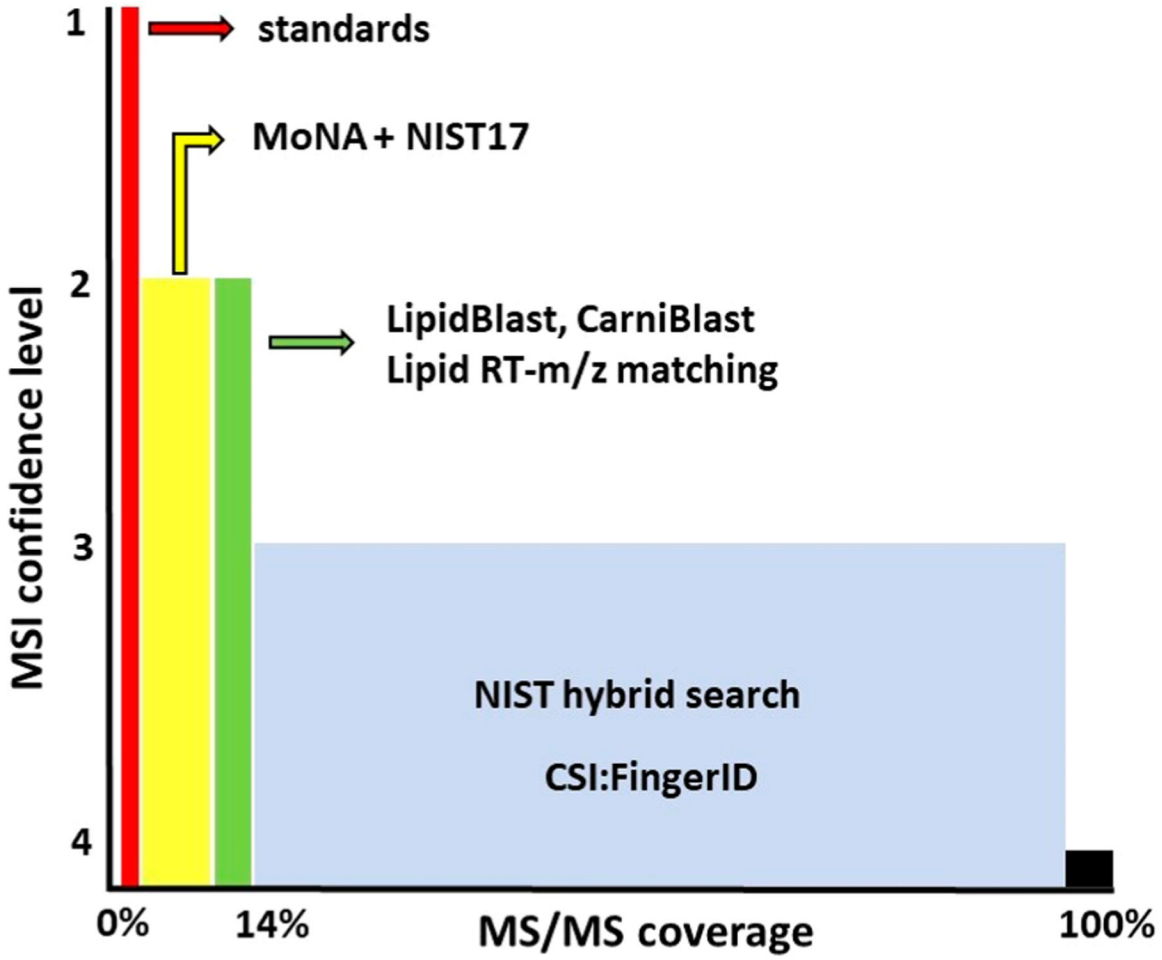


Figure 1.
Categorized overview of the complete annotation of MS/MS spectra of human urine metabolomes based on MSI level 1, 2, and 3 confidence scores.

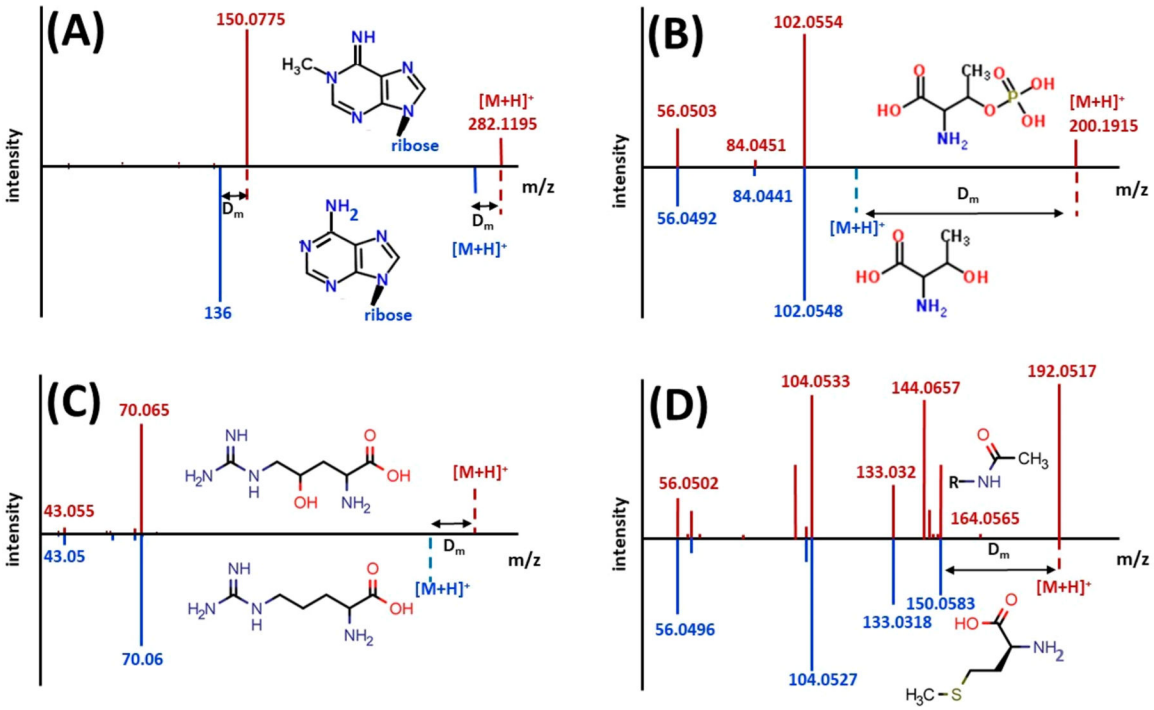


Figure 2.

Head-to-tail comparison of MS/MS spectra of distinct shifts in spectra of modified versions of canonical metabolites. (A) methylation: 1-methyladenosine to adenosine, (B) phosphorylation: phosphothreonine to threonine, (C) hydroxylation: hydroxyarginine to arginine, (D): acetylation: *N*-acetylmethionine to methionine.

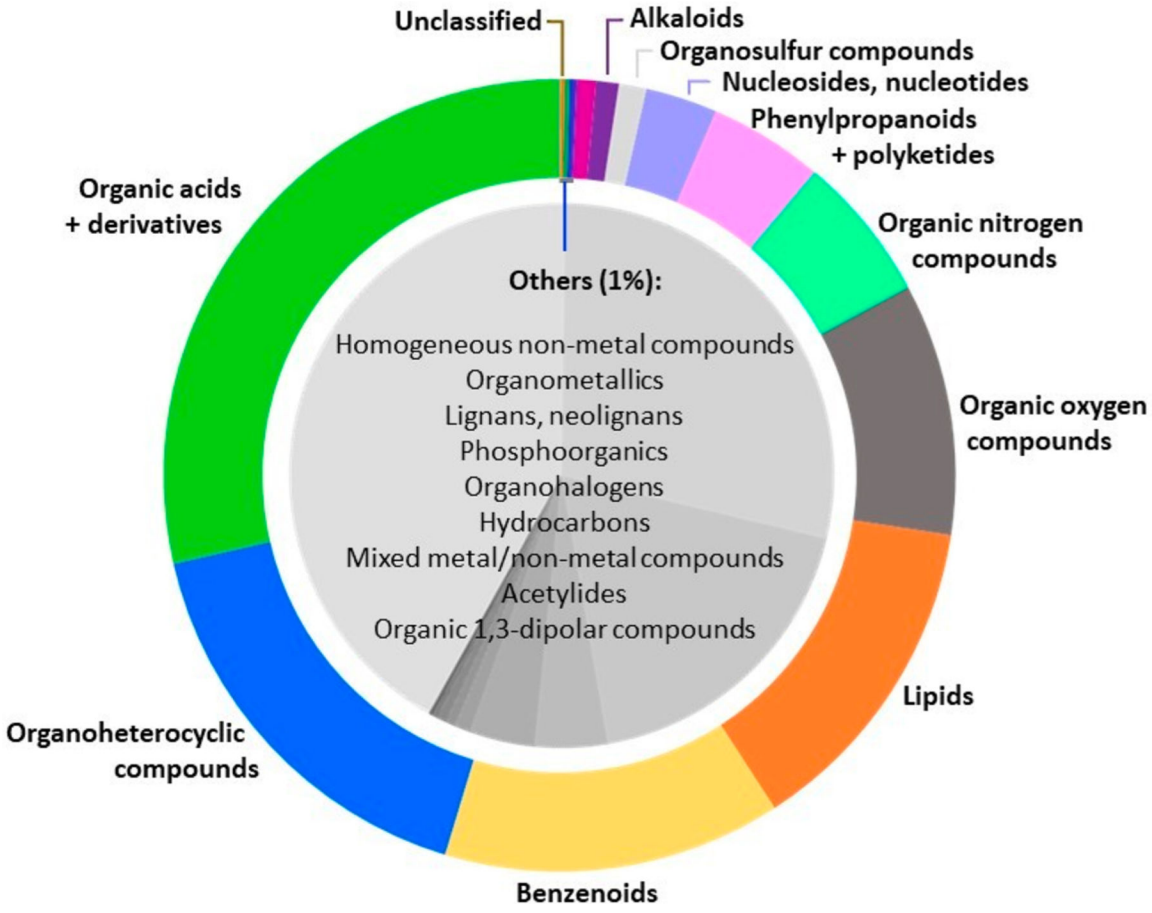


Figure 3.

Structural categorization of compounds present in urine samples of 43 subjects diagnosed with interstitial cystitis. Chemicals are structured according to the “Superclass level” of the ClassyFire classification system.

Table 1.

Results of Comprehensive Annotation of Urinary Metabolomics and Lipidomics MS/MS Spectra

chromatography and databases	type of matching	MSI level of annotation	no. of annotations
HILIC	precursor *m*/*z*, RT experimental library MS/MS	MSI level 1	103
lipidomics	*m*/*z*, RT, experimental + *in silico* library MS/MS		72
HILIC: MoNA+NIST17	precursor *m*/*z*, experimental library MS/MS	MSI level 2	440
HILIC	precursor *m*/*z* and RT		13
lipidomics: CarniBlast	*m*/*z*, *in silico* library MS/MS		18
HILIC: CarniBlast	precursor *m*/*z*, *in silico* library MS/MS		107
lipidomics: mzRT lookup	precursor *m*/*z* with RT curation	MSI level 3	96
HILIC and lipidomics: NIST17 hybrid search	MS/MS (hybrid and experimental library)		6,447
HILIC: Sirius/CSI:FingerID	precursor *m*/*z* and *in silico* predicted MS/MS		728