Chemical Research in Toxicology®

Subscriber access provided by CORNELL UNIVERSITY LIBRARY

Review

Computational metabolomics: a framework for the million metabolome

Karan Uppal, Douglas I. Walker, Ken Liu, Shuzhao Li, Young-Mi Go, and Dean P. Jones

Chem. Res. Toxicol., **Just Accepted Manuscript** • DOI: 10.1021/acs.chemrestox.6b00179 • Publication Date (Web): 14 Sep 2016

Downloaded from http://pubs.acs.org on September 19, 2016

Just Accepted

"Just Accepted" manuscripts have been peer-reviewed and accepted for publication. They are posted online prior to technical editing, formatting for publication and author proofing. The American Chemical Society provides "Just Accepted" as a free service to the research community to expedite the dissemination of scientific material as soon as possible after acceptance. "Just Accepted" manuscripts appear in full in PDF format accompanied by an HTML abstract. "Just Accepted" manuscripts have been fully peer reviewed, but should not be considered the official version of record. They are accessible to all readers and citable by the Digital Object Identifier (DOI®). "Just Accepted" is an optional service offered to authors. Therefore, the "Just Accepted" Web site may not include all articles that will be published in the journal. After a manuscript is technically edited and formatted, it will be removed from the "Just Accepted" Web site and published as an ASAP article. Note that technical editing may introduce minor changes to the manuscript text and/or graphics which could affect content, and all legal disclaimers and ethical guidelines that apply to the journal pertain. ACS cannot be held responsible for errors or consequences arising from the use of information contained in these "Just Accepted" manuscripts.

ACS Publications

Chemical Research in Toxicology is published by the American Chemical Society.
1155 Sixteenth Street N.W., Washington, DC 20036
Published by American Chemical Society. Copyright © American Chemical Society.
However, no copyright claim is made to original U.S. Government works, or works produced by employees of any Commonwealth realm Crown government in the course of their duties.

Computational metabolomics: a framework for the million metabolome

Karan Uppal,[†] Douglas I. Walker,[†,‡,§] Ken Liu,[†] Shuzhao Li,[†,‡] Young-Mi Go[†], and Dean P. Jones[†,‡,*]

[†] Clinical Biomarkers Laboratory, Department of Medicine, Emory University, Atlanta, GA 30322

[‡] Hercules Exposome Research Center, Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA 30322

[§] Department of Civil and Environmental Engineering, Tufts University, Medford, MA 02155

*Correspondence:

Dean P. Jones, Ph.D.

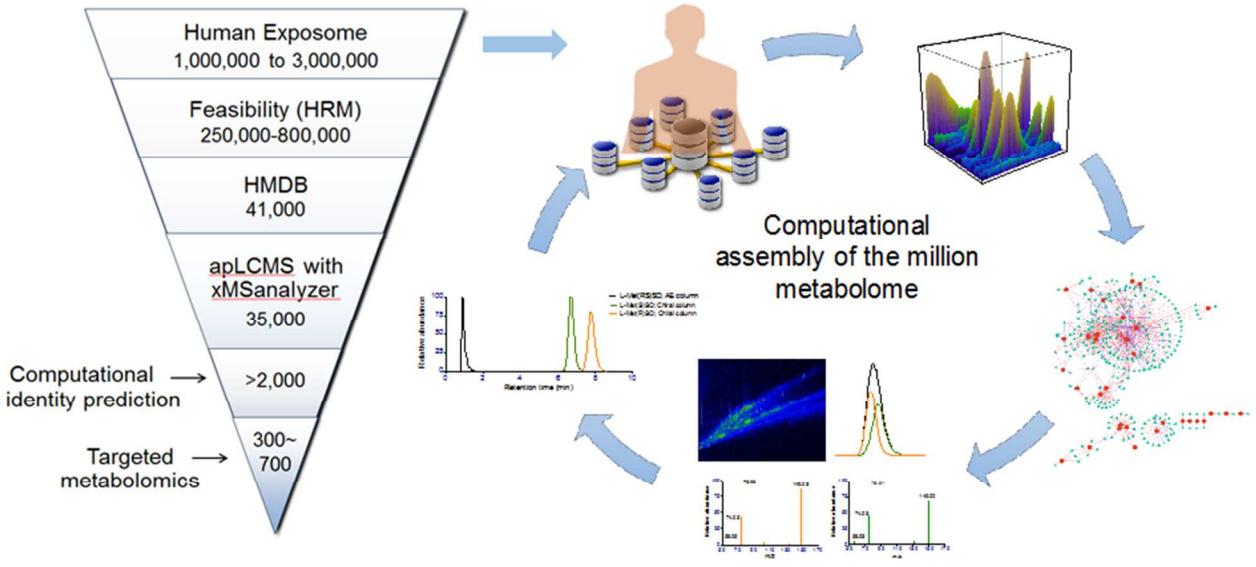
Department of Medicine, Pulmonary Division, Emory University, 205 Whitehead Biomedical Research Building, 615 Michael Street, Atlanta, GA 30322

Tel.: 404-727-5970, Fax: 404-712-2974.

Email: dpjones@emory.edu

CONTENTS

TOC graphic


Human Exposome
1,000,000 to 3,000,000
Feasibility (HRM)
250,000-800,000
HMDB
41,000
apLCMS with
xMSanalyzer
35,000
Computational
identity prediction
>2,000
Targeted
metabolomics
300~
700
Computational
assembly of the million
metabolome

ABSTRACT

"*Sola dosis facit venenum.*" These words of Paracelsus, "the dose makes the poison", can lead to a cavalier attitude concerning potential toxicities of the vast array of low abundance environmental chemicals to which humans are exposed. Exposome research teaches that 80-85% of human disease is linked to environmental exposures. The human exposome is estimated to include >400,000 environmental chemicals, most of which are uncharacterized with regard to human health. In fact, mass spectrometry measures >200,000 *m/z* features (ions) in microliter volumes derived from human samples; most are unidentified. This crystalizes a grand challenge for chemical research in toxicology: to develop reliable and affordable analytical methods to understand health impacts of the extensive human chemical experience. To this end, there appears to be no choice but to abandon the limitations of measuring one chemical at a time. The present article reviews progress in computational metabolomics to provide probability-based annotation linking ions to known chemicals and serve as a foundation for unambiguous designation of unidentified ions for toxicologic study. We review methods to characterize ions in terms of accurate mass *m/z*, chromatographic retention time, correlation of adduct, isotopic and fragment forms, association with metabolic pathways and measurement of collision-induced dissociation products, collision cross section and chirality. Such information can support a largely unambiguous system for documenting unidentified ions in environmental surveillance and human biomonitoring. Assembly of this data would provide a resource to characterize and understand health risks of the array of low-abundance chemicals to which humans are exposed.

1. Introduction: The "dark matter" of the human exposome

Rachel Carson's book, Silent Spring, published in 1962, awakened society to toxicological hazards from environmental exposures. As a result, procedures and regulatory policies to identify environmental hazards and risks of exposure were established to minimize health burden. The measures use technologies available decades ago and provide an affordable approach to minimize population risks from many hazardous chemicals. A consequence of this approach, however, is that most chemicals to which humans are exposed, the so-called "dark matter of the exposome", are largely uncharacterized and have minimal or no evaluation concerning toxicity.

Current analytical capabilities provide an opportunity to approach the problem differently, i.e., to develop universal exposure surveillance procedures [1, 2] in which health risks are associated with chemicals measured in populations using advanced biomonitoring methods. Such an approach sets new goals for mass spectrometry and analytical chemistry built upon recent explosive development of metabolomics capabilities. In this, environmental toxicologists have a critical role in guiding development of reliable and affordable methods for detailed human biomonitoring. Specifically, environmental chemicals are often present in human samples at three to four orders of magnitude lower abundance than intermediary metabolites. Thus, the environmental chemistry and toxicology challenge is to develop ways to scientifically study large numbers of un-identified, low abundance chemicals so that those associated with human disease can be isolated and identified.

The present review is focused on rapidly developing methods of computational metabolomics to address this challenge. Importantly for application to population surveillance and toxicology research concerning low abundance environmental chemicals, computational metabolomics uses a workflow that differs from more commonly used analytical methods which target analysis of known chemicals.[3, 4] At the most basic level, this difference involves distinguishing signal from noise, i.e., useful signal variation from non-useful signal variation. Analytical chemistry is biased toward assurance that a specific signal is reflective of a chemical of interest; for highly precise measurement, high qualitative and quantitative stringency requirements minimize error. In contrast, characterization of unidentified low abundance chemicals found in a small number of individuals cannot be achieved with the same rigor. The workflow for computational metabolomics is therefore biased toward inclusion of infrequent and less reliable signals. The expectation is that knowledge gain from the low abundance and uncharacterized signals will be cumulative, ultimately leading to understanding of health risks and sources of exposure and directing development of improved analytical methods for low abundance chemicals of concern. Hence, the present discussion addresses creation of a rigorous analytical chemistry data structure to facilitate systematic knowledge of the toxicology of currently uncharacterized low abundance chemicals found in humans. By necessity, such a goal will require use and integration of data from multiple analytical platforms and approaches; the headache created is a need to develop an unambiguous system to reliably designate tens of thousands of reproducible but unidentified mass spectral features so that the chemical toxicology research community

can pursue more specific aspects of thresholds and dose for those with adverse health impact.

2. Metabolomics for environmental biomonitoring

A rapid rise in metabolomics has occurred since 2000 (Figure 1) when chemometric methods were applied to nuclear magnetic resonance (NMR) spectroscopy to facilitate interpretation of complex spectra obtained from biological samples.[5, 6] Although the popularity of NMR is rapidly being supplanted by mass spectrometer-based methods (Figure 1), extension of computational methods to mass spectrometry is delivering another transformation in analytical chemistry, from analysis of one chemical at a time in a targeted manner to probability-based approaches to measure thousands in a single analysis. This transition is loosely discussed in terms of two categories of metabolomics research: targeted metabolomics and untargeted metabolomics. Targeted metabolomics focuses on a defined set of metabolites or pathways, while untargeted metabolomics aims to provide global profiling of small molecules in a biological system in an unbiased manner.[2, 3] High-resolution metabolomics (HRM) uses liquid chromatography (LC) or gas chromatography (GC) with high-resolution mass spectrometry and advanced data extraction algorithms to measure a broad spectrum of chemicals in biologic samples.[7, 8] In this, high-resolution mass spectrometry refers to instrumentation providing mass resolution of 30,000 [9] and includes Fourier-transform Ion-Cyclotron Resonance (FT-ICR) mass spectrometers, some Quadrupole-Time-of-Flight (Q-TOF) mass spectrometers, and specialized ion trap mass analyzers with an inner electrode that traps ions in an orbital motion (Orbitrap).[10, 11] FT-ICR and Orbitrap instruments are capable of higher mass resolution, e.g., 60,000 or more, and are also termed “ultra-high resolution” mass

spectrometers. HRM is noteworthy because application to plasma and urine provides a practical way to obtain detailed exposure and metabolic health information for precision medicine and also an affordable way to study cumulative life-long exposures in human exposome research. [1, 12]

Plasma and urine samples are commonly available during routine health examination, and HRM can be used with either to obtain information on environmental exposures, nutrient supply, central metabolic intermediates, metabolic wastes and hormonal signals.[1] In principle, such analyses can be used as an integrated measure of biologic responses, including effects of emotional stress, exercise and other health behaviors. [2, 3] Technological advancements and improved algorithms for HRM now enable reproducible detection of tens of thousands of metabolic features in biological samples. [13, 14] The number of chemicals represented is unknown, but ion dissociation of randomly selected features and correlation analyses of features suggest that the number of chemicals is also in the tens of thousands, most of which are unidentified.

Several studies have demonstrated the utility of HRM for human exposome research. In an untargeted metabolic profiling study, Go and Walker[15]detected, confirmed and quantified environmental chemicals present in plasma samples from 153 healthy humans. These included chemicals derived from food (caffeine, hippuric acid), insecticides (chlorobenzoic acid, chlorophenylacetic acid, pirimicarb, xylylcarb), herbicide (chlorsulfuron), tobacco (cotinine), flame retardants (triethylphosphate, triphenylphosphate, tris(2-chloropropyl)phosphate) and other commercial products (octylphenol, dibutylphthalate, dipropylphthalate, styrene, tetraethylene glycol). Less than half of these xenobiotics had previous publications reporting concentrations for

human plasma. Roca and Leon, [16] and Jamin and Bonvallot [17] obtained similar results showing ability to detect a large number of environmental chemical metabolites using untargeted chemical analysis.

In other applications, HRM was used to evaluate metabolites associated with polycyclic aromatic hydrocarbon (PAH) exposures in serum of military personnel.[3, 18] Correlations were observed for multiple metabolic products of naphthalene, pyrene, anthracene and benzo(a)pyrene [3] and also for metabolic pathways for linoleate, acyl carnitines, sphingolipids, methionine and cysteine.[18] Although some studies are available with concurrent air monitoring, measurements in blood and urine typically cannot discriminate sources of exposure. For instance, PAH may derive from air pollution, smoking or consumption of charbroiled foods. A study of 400 military personnel classified individuals as smokers or non-smokers based upon serum cotinine concentration.[19] This study found correlations of hydroxycotinine and naphthalene-1,2-diol with cotinine, as well as associations with many of the same pathways as correlated with PAH's. In principle, such analyses could be extended to examine exposures to dietary carcinogens, such as 2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine (PhIP) and 2-amino-3-methylimidazo[4,5-f]quinoline (IQ). Presently, however, this has not been evaluated for untargeted analyses using HRM.

An application of untargeted HRM to study of Parkinson's Disease (PD)[20] showed associations to chemical features with accurate mass match to polybrominated diphenyl ether (PBDE), tetrabromobisphenol A, octachlorostyrene and pentachloroethane. The chemical feature corresponding to PBDE was 1.5-fold higher in PD than controls, and a match to 2-amino-1,2-bis(p-chlorophenyl)ethanol was 1.5-fold higher in individuals with

rapid disease progression compared to slow progression. Untargeted HRM also detected 94 metabolic features associated with neovascular age-related macular degeneration (NVAMD) [21], including a match to *β*-2,3,4,5,6-pentachlorocyclohexanol (*β*-PCCH), a hydroxylated metabolite of *β*-lindane in insecticide formulations. Other correlated features matched the ^{37}Cl form of *β*-PCCH and other halogenated chemicals, suggesting a possible association for chemical exposures in NVAMD. Together, these studies show that contemporary HRM methods provide powerful approaches for biomonitoring of exogenous chemicals and studying their toxicities in epidemiological and laboratory research. The studies also emphasize that a large number of metabolic features in human samples are currently unidentified; whether these are natural or anthropogenic is unknown.

Detection of environmental chemicals within exposome research depends upon instrument sensitivity and response characteristics. The rapid advance in instrument quality is illustrated by a recent HRM study showing that chemicals could be quantified that differed by approximately eight orders of magnitude in absolute abundance.[15] Thus, even though the best platforms for cost, coverage and quantitative reproducibility are not well established, use of mass spectrometry protocols with simple protein removal, dual chromatography, dual electrospray ionization and triplicate analyses [3, 22] provides far more health-related chemical information than contemporary blood chemistry, NMR methods or other analytical procedures. This is not to imply that NMR spectroscopy and other methods have no place in health analyses, only that based on cost and extent of information provided, HRM offers advantages (discussed in Jones *et al*).[1, 19] NMR spectroscopy is particularly important, for instance, in structure elucidation of chemicals

and in non-invasive measurement of metabolites in vivo termed "magnetic resonance spectroscopy", performed using MRI instruments. The key conclusion is that in health and medicine, HRM provides a powerful approach to evaluate low abundance environmental exposures as well as nutrition, genetic factors impacting metabolism, adaptations to prior exposures, and disease development and progression.

3. The human metabolome

The human metabolome consists of *1)* endogenous metabolites with molecular weight < 2000 dalton,[23] including essential nutrients, amino acids, sugars, fatty acids, etc. and *2)* exogenous exposures including chemicals derived from food, drugs and pollution. [3, 8, 24] The exogenous influences originating from our diet, behavior, and lifestyles are cumulative throughout our lifespan and make up the human exposome.[25-27] Summation of these exposures indicates that humans are exposed to 1-3 million chemicals during their lifetime.[28] These exposures are important because they combine with genetic factors contributing to inter-individual metabolic variation and accounting for most human disease.[29-31] Multiple studies show that consumption of different diets contributes to inter-individual metabolic variations. [32-34]

We define "million metabolome" as an aggregate of endogenous metabolites and products of exogenously derived exposures measured in individuals across time and collectively among geographic populations. In this, one must note that the term is at least partially symbolic; there is no way to precisely estimate the number of metabolites in the human metabolome, but there is recognition that the number is probably greater than a million and that capability to measure one million is necessary to characterize human

exposures.[1] Only a fraction of the million metabolome is part of the core human metabolome [35] essential for life and preserved across populations. Most chemical experiences of individuals are highly variable, indicating that large populations may be needed to reach the million metabolome.

Although technological and computational advancements have facilitated detection of tens of thousands of ions, metabolite identification remains one of the biggest challenges of available analytical methods.[36-38] For the most powerful available methods, such as LCMS-based HRM, there is a need for development of advanced computational methods for systematic characterization of collected data. Figure 2 illustrates the gap between current metabolite detection abilities and likely size of the human exposome. Most targeted LCMS and GCMS methods detect 300-700 metabolites in biological samples, [39, 40] with NMR-based methods detecting less than half this range in biologic samples. Advanced computational methods facilitate detection of more than 35,000 ions from biological samples by LCMS [13]; feasibility studies show, however, that variation of data extraction parameters, e.g., increasing tolerance for coefficient of variation and % missing values, can allow detection of 250,000 to 800,000 mass-to-charge (*m/z*) features in human and animal samples.[1]

The ability to further analyze and characterize the human exposome in a fully automated or semi-automated manner requires development of a computational framework that can process different types of mass spectrometry data (MS, MS^2, MS^n, ion mobility MS, stereochemistry selective detection), provide predictions of metabolite identities, allow interpretation of metabolites using data-driven and knowledge-driven association methods, and combine orthogonal pieces of information to facilitate unambiguous

characterization and designation of both low-abundance and high-abundance metabolites. Collection of data over time in a cumulative database would further support a lifecycle framework for study of the individual exposures as a foundation to understand disease, predict individual risk, monitor progression and evaluate efficacy of interventions. [8] Databases are less expensive and more stable than stored blood samples and could be useful to monitor individual health changes as well as disease trends in populations. Thus, HRM could evolve into a central component of healthcare. Although potential benefits from understanding the human exposome are obvious, barriers exist due to the large number of exposures, variations in duration and intensity and costs associated with systematic study.

In the following sections, we review existing methods for MS with particular focus on LCMS, to develop a framework for the million metabolome. We include discussion of peak detection and alignment, quality assessment, metabolite annotation, network and pathway analysis, and metabolite identification, and propose inclusion of other measures to enhance metabolite identity prediction for both high-abundance and low-abundance metabolites.

4. Feature extraction, quality assessment, and data correction

Various tools have been developed to automate the process of peak detection, noise removal, intensity estimation, and feature alignment.[41-43] Figure 3 shows the typical steps involved in feature extraction, quality assessment, and data correction. None of these procedures can compensate for poor performance, chromatography instability or mass

spectrometry inaccuracy; consequently, quality control procedures are mandatory. Additionally, each sample constitutes a unique matrix so that replicate analyses, usually triplicate, are needed to verify analytical quality and assure reliable quantification.

4.1 Peak detection and alignment. The first step involves peak detection in individual files, and only features that meet the signal-to-noise threshold and/or peak shape criteria are kept for further analysis. For instance, apLCMS uses a three-step process for feature detection that involves grouping of data points based on *m/z* cutoff, splitting each group of *m/z* features based on the retention time dimension using kernel density estimation, and use of a run filter that takes into account the minimum length in the elution time dimension as well as proportion of time points in which the signal is detected to identify true peaks.[43] XCMS uses signal-to-noise and filtering criteria based on minimum number of peaks detected with minimum intensity>=I for removing features along with a density and wavelet transformation based method for peak detection.[42] Several other methods for noise removal and feature detection (centroid based, local maxima, recursive threshold, wavelet transform, and exact mass) in single files are implemented in MzMine2.[41] Other tools such as MetSign perform peak deconvolution using a two-stage process where the first derivative of the smoothed data is used to detect the dominant peaks and the second derivative is used to detect the hidden or low abundance peaks.[44] The performance of peak detection algorithms, especially for environmental exposures, can be improved by incorporating additional layers of information from biological and environmental databases, in-house databases of reliable peaks, and across multiple runs of the same sample. For instance, methods that utilize preexisting knowledge such as

information about known metabolites in the Human Metabolome Database (HMDB)[23] and pathway information in the Kyoto Encylopedia of Genes and Genomes (KEGG)[45] along with machine learning approaches can further enhance peak detection in biological samples.[46] Additionally, current algorithms perform feature detection individually within each LC/MS run and do not incorporate information across one or more technical replicates of a sample. In principle, combining information from multiple analyses prior to feature extraction could provide another means to reduce noise and improve feature extraction. Feature quality evaluation criteria such as signal-to-noise ratio and coefficient of variation remain an important subject to enhance confidence in low abundance or exogenous metabolites that could be present in only small number of samples and improve overall data quality prior to feature alignment.

After peak detection in individual LC/MS runs or profiles, alignment across all profiles is necessary to generate a combined feature set. Alignment is accomplished through *m/z* and retention time dewarping. The primary need is to correct the retention time dimension due to changes in pressure, column temperature, and column age over the course of an analytical run.[47] Most existing methods include a non-linear retention deviation estimation step, providing corrected retention times in individual profiles using the estimated deviation.[43, 47] Pairwise alignment is then completed by reference to the profile with maximum number of detected features, all other profiles are aligned with respect to the reference in a pairwise fashion using methods such as dynamic time warping, ObiWarp and kernel smoothing. [42, 43, 48]

A limitation to use of one sample as reference for aligning samples from multiple batches is that any distortions in retention time could affect the alignment results due to peak

mismatching. The aligned features are normally represented by median (or mean) *m/z* and retention time post-alignment. These estimates could be improved by following a hierarchical alignment procedure that first performs alignment of samples at a single sample level (across technical replicates), performs alignment within individual batches in the next step, and finally aligns all samples using the results from previous steps. Additionally, landmark peaks or use of "gold standard" metabolites as reference metabolites can improve retention time alignment and facilitate cross-laboratory comparisons. [14]

4.2 Parameter optimization. Parameter optimization is a crucial step in data extraction. Operational parameters of mass spectrometers differ, and fine-tuning of peak detection and alignment parameters is necessary for obtaining optimal results. [13, 36, 37, 49] xMSanalyzer is an R package that uses a scoring function based on number and quality of features determined based on coefficient of variation (CV) within technical replicates. xMSanalyzer is designed to work with apLCMS, XCMS, and other data extraction software. Another R package, IPO, is designed for optimizing peak picking, retention time correction, and peak grouping parameters in XCMS using replicate measures of a single sample and design of experiments framework.[49] In principle, there are a broad range of options for improvement with an important limitation being the computational time required for performing multiple extractions, integrating data and assessing quality of data with the different parameters.

4.3 Quality assessment and data correction. Web-based tools and R packages such as MetaboAnalyst, xMSanalyzer and MSPrep provide utilities for addressing quality assessment and correction. [13, 50, 51] The quality of individual features and samples can be evaluated based on CV within technical replicates, variability across pooled reference/quality control (QC) samples, percent missing values, signal-to-noise ratio, Principal Component Analysis to identify outliers and batch effects, and pairwise correlation within technical replicates to evaluate analytical reproducibility. [13, 36, 50] Various methods have been developed to address batch-effect problems. Dunn *et al*. [36] proposed QC-RLSC, a signal correction approach that fits a LOESS curve to the QC samples; the raw data for a feature is corrected relative to this interpolated correction curve.[36] Several R packages including sva and MSPrep offer batch-effect correction procedures.[51, 52] Methods for correcting batch effect include ComBat, which uses an empirical Bayes approach, and Surrogate Variable Analysis for removing batch-effects. [52, 53] Most data processing workflows for metabolomics have been developed for biomarker studies where analytical errors such as batch-effect errors could dramatically impact results and interpretation; however, it is challenging to address batch-related effects in exposome studies due to effect size considerations. Thus, one of the critical needs is improved batch correction procedure to address features with infrequent occurrence. Specifically, if there is an *m/z* feature present in only one sample in a batch and that feature is not present in corresponding pooled reference materials, then there are needs to be able to quantitatively compare that intensity to the same feature detected on another day.

Accurate mass measurement error is another source of error, and can occur due to temperature changes and improper instrument calibration.[36, 37] Mass accuracy plays a critical role during sample alignment and annotation. During the feature annotation process, measured *m/z* is compared to the theoretical *m/z* and only metabolites that are within the user-defined mass tolerance level are selected. The number of false positives can dramatically increase as the mass accuracy deteriorates. [37, 54] Internal standards and annotated features based on reference metabolites can be used for tracking mass accuracy and estimating mass measurement error. [55] Correction of mass errors can also improve alignment of datasets from multiple studies or batches.

To summarize, various approaches are available to enhance extraction of information on ions measured by mass spectrometry. These approaches provide quality assessment and data correction for general metabolomics use. The tools have been rigorously developed and provide an outstanding range of useful options. In terms of chemical detection, however, the limitations must be considered. By having a high stringency for signal to noise, one protects against identifying noise as signal. On the other hand, the high stringency is accompanied by dismissal of real signals as noise. Thus, to expand detection of low abundance chemicals, additional efforts need to focus on identification and reduction of noise signals. Also, improved batch correction procedures are needed to address features that are detected infrequently, such as unidentified environmental chemicals found in a small fraction of a population.

5. Data-driven clustering methods to identify sub-groups of related features

5.1 Correlation-based network and clustering analysis. An important advantage of computational metabolomics lies in the use of correlations among ion signals to aid in determination of chemical identity. Metabolites are interconnected by series of biochemical reactions, and this network of metabolites is organized in a hierarchical manner such that many small modules combine to form larger modules. [56, 57] Correlation-based network and modularity analysis is one approach to elucidate the association structure of metabolites. Although there are several mechanisms that could lead to correlations between metabolites, the association structure can be used to identify ions derived from the same metabolite, [58-60] identify biotransformations,[61] and detect associations between environmental exposures and endogenous metabolites.[15]

For high abundance unidentified chemicals, multiple spectral features arising from a single chemical provide valuable structural information to characterize a chemical. A network of ions where a pair of ions is linked if their correlation exceeds the significance threshold, e.g. $|r|>0.8$, can be generated to identify isotopes, adducts, and in-source fragments associated with a chemical (Figure 4). A similar approach can be used to identify biotransformations and other related metabolites. [60] Metabolome-wide association studies (MWAS) allow identification of associations between a specific target variable, e.g. cotinine levels in individuals, and metabolic profiles. [8, 62-64] In an MWAS, statistical tests are performed for association of a parameter (e.g., disease biomarker, chemical or other measured parameter) with each *m/z* feature to test for significance of association. Application of targeted MWAS using correlation-based criteria identified

choline-related metabolites and demonstrated similarity between correlation patterns of choline in different species (Figure 5). [64]

Correlation-based network analysis can also facilitate identification of in-source fragments. Gas-chromatography mass spectrometry with electron ionization sources results in a large number of characteristic spectra indicative of chemical functional groups and structure.[61, 65] Electrospray ionization can produce in-source fragmentation (e.g. loss of NH_3, H_2O, CHOOH, etc.) from electrical potentials or heat applied in the ion source. [66, 67] Because in-source fragments can mimic accurate masses of other common metabolites, computational methods that identify adducts, isotopes, and in-source fragments (based on clustering of highly correlated co-eluting ions) increases the ability to correctly assign chemical identities. An example is the in-source formation of pyroglutamate from glutamine or glutamate.[68] The identification of in-source fragments requires consideration of chromatographic conditions to separate possible co-eluting chemicals, as well as ion source conditions. When using soft ionization techniques, in-source fragmentation is only commonly observed for highly abundant metabolites, many low abundance chemicals will generate only a single detectable signal.[3, 18] To ensure detected, unannotated ions are unique chemicals, it is important to perform targeted MWAS to exclude the possibility of a signal originating from source fragments, adducts, and/or isotopes. To increase confidence of chemical identification, alternative detection methods with increased sensitivity for unknown chemicals and methods for defining unknown ions will be needed.

In addition to characterizing ions arising from known chemicals, MWAS using univariate and multivariate approaches can be used to generate hypotheses about biochemical roles of features with no database matches. This process uses targeted MWAS with validated metabolites or xMWAS, where "x" corresponds to other –omes (transcriptome, microbiome, genome, etc.). Krumsiek et al. used a systems-level approach where they combined genome-wide association analysis, knowledge-based pathway information, and metabolic networks to predict the identity of unknown metabolites.[69] Other studies have used integrative methods based on partial least squares regression (PLS) to determine correlations between the metabolome and the transcriptome, [70] proteome [71] and microbiome. [72] These methods combined with pathway and literature based information can provide alternative approaches for generating hypothesis about chemical identity, particularly for low abundance chemicals.

5.2 Retention time. Retention time is the time between sample injection and appearance of the maximum ion signal after chromatographic separation.[73] Chromatographic separation of complex mixtures is achieved by the differential rate of migration of chemicals through an analytical column. As chromatographic separation is dependent on column chemistry, choice of solvent, as well as physicochemical properties of a given chemical, the same chemical should have the same retention time (± few seconds) under the same chromatographic conditions over multiple injections. Application of kernel density estimation in the retention time dimension can be used for unsupervised grouping of features with similar chemical properties and assist in identifying adducts, isotopes, and in-source fragments when applied on distinct clusters of strongly correlated ions. [58]

5.3 Mass defect. Mass defect is the difference between the accurate mass and nominal mass of an ion and is a useful measure to facilitate isotope pattern reconstruction and identification of metabolite biotransformations.[74, 75] For high-resolution mass spectrometry, accurate mass information can be combined with mass defect filtering (MDF) techniques for finding isotopes, expected losses ($-H_2O$, $-2H_2O$, etc.), and biotransformations of known metabolites. [61, 75, 76] Furthermore, the MDF method can allow identification of features belonging to similar chemical classes, contain specific functional groups and homologous series.[77] In principle, additional use of mass defect for chemical identification could be derived from theoretical predictions based upon known elemental compositions of chemicals in ChemSpider.

6. Knowledge-driven methods for network and pathway analysis for metabolomics

Targeted metabolomics approaches often start with metabolite identification prior to pathway and network analysis. This is a valuable approach but can result in loss of information relative to chemicals without confirmed identity. In the present discussion, we consider pathway and network analysis prior to metabolite annotation and identification because computational metabolomics does not require a priori knowledge of *m/z* identity to obtain useful chemical information. Details of this alternate workflow are available. [3, 19] Importantly, this computational metabolomics approach enables use of otherwise uncharacterized mass spectral data.

Several thousands of metabolic reactions are collected in various databases,[78-81] which have been accumulated from biochemical research over many decades. The metabolic reactions are mostly interconnected by shared metabolites and are often organized into

pathways of dedicated functions. By mapping metabolites to these pathways, one can contextualize the data, greatly facilitating interpretation; however, the identity of metabolites in mass spectrometry data is often difficult to obtain and hinders the downstream pathway analysis.

A novel approach, named *mummichog*, was designed by Li *et al* [82] to rewrite the conventional metabolomic workflow. Since the computational prediction of metabolites from spectral peaks often results in multiple possibilities (see Section 7), a "null" distribution can be estimated by how these predicted metabolites from a metabolomics experiment map to all known metabolite reactions. Even though most are false annotations, the biological meaning in the data drives enrichment of metabolite subsets. The enrichment pattern of real metabolites compared to the null distribution is then tested statistically. Thus, *mummichog* can predict significant pathways and network modules directly from untargeted metabolomics data. To test prioritized hypotheses from *mummichog*, researchers can focus on validating only a handful of metabolites.

Mummichog has become a powerful tool to accelerate the rate of scientific discovery. [83-85] Multiple mechanistic studies have been supported by the *mummichog* approach, including T cell memory formation [86] and stress response in innate immune cells [87] Combined with common regression models and untargeted metabolomics, *mummichog* enables inclusion of metabolic pathway analysis in population studies. For example, using this combined approach, Hoffman *et al*. [88] identified metabolic pathways associated with age, sex, and genotype, including pathways involving the carnitine shuttle, glycerophospholipid metabolism, neurotransmitters and amino acid metabolism. Amino acid pathways, especially tyrosine metabolism, were also identified as associated with

nonalcoholic fatty liver disease using *mummichog* combined with statistical selection of relevant *m/z* ions. [89]

The HRM workflow has thus expanded from targeted analyses of a relatively small number of metabolites (300-700) supported by most metabolomics cores, to a much broader scope including thousands of metabolites from >20,000 ions. Most metabolic pathways are included, and the prioritization is agnostic, defined by the measured data. Approaches can be refined by using a highly stringent false discovery rate (e.g., $q < 0.05$) to select for metabolites most likely to represent real differences, or by using a raw *p*-value threshold of 0.05, expecting that any metabolite could represent a real difference. The former protects against type I statistical error while the latter protects against type II statistical error. Importantly, statistical tests for pathway enrichment using *mummichog* and a raw *p*-value threshold of 0.05 provide an effective compromise to protect against both type I and type II statistical error.

7. Metabolite annotation

"Annotation" is defined as "a note of explanation or comment" and should not be confused with "chemical identification". Chemical identification is ultimately required for mass spectral features of interest, but identification can be difficult and subject to different criteria for certainty.[90, 91] Importantly, metabolite identification is a major bottleneck in untargeted metabolomics.[36, 38] Most measured ions do not match known metabolites in databases using common adduct forms (Figure 6A). In high-resolution metabolomics analyses of human diseases, MWAS show that the accurate mass *m/z* for

more than half of the ions associated with human disease do not match any predicted ions for known chemicals in human metabolomic databases (Table 1). In recent years, several methods such as AStream, CAMERA, ProbMetab, and MetAssign have been developed for metabolite annotation [58, 92-94]. Most of these methods utilize *m/z*, retention time, adduct patterns, isotopes and correlation/clustering methods for metabolite annotation. AStream takes as input the processed feature table and uses correlation within *m/z* features, isotope patterns, retention time, and adduct patterns to annotate features using HMDB.[58] CAMERA uses a graph-clustering approach that incorporates correlation within raw signals, retention time, and adduct patterns for grouping ions derived from a single metabolite.[93] MetAssign and ProbMetab use Bayesian methods for assigning probabilities to annotations. [92, 94]

Additional information such as mass defect, modular network structure, pathway associations, elemental information and isotope ratios can improve confidence in identity prediction. [56, 82, 95] Methods utilizing multiple layers of information along with data-driven clusters (correlation-based, retention time, and mass defect as described in section 4) can further improve metabolite annotation and allow suspect screening for environmental exposures by assigning confidence levels to annotation. Ion dissociation analysis of metabolites annotated using the criteria described above shows that 80% of predicted identities are correct and overall >2,000 metabolites can be routinely annotated in human studies (Figure 6B). Various factors including *m/z* accuracy, selection of adducts, selection of database, consideration of isotopic forms, elemental and isotopic ratio checks influence the performance of annotation algorithms. Development of algorithms that use machine learning to predict retention time, adduct and isotope

probabilities, relative intensity, and various physical properties of previously validated metabolites, ionization modes, and columns could potentially improve performance of existing identity prediction methods. [61]

The capabilities for annotation and pathway mapping with these computational methods are truly advanced from just a few years ago, allowing simultaneous testing of most metabolic pathways for associations with any exposure, disease biomarker or measured health outcome. [2] Computational metabolomics has advanced analytical chemistry to a new level, one in which there is no longer a need to guess which pathways might be affected but rather to confidently interrogate most of the known metabolic pathways in a single step. At the same time, this accomplishment directs attention to the fact that known metabolites may represent less than half of the chemicals measured in a single experiment. Thus, as analytical chemistry moves beyond a one-chemical-at-a-time framework, the need for a better framework to address unidentified chemicals in the million metabolome becomes apparent.

8. Ion characterization and designation using knowledge-driven approaches

8.1 Metabolite identification. As discussed above, detected *m/z* ion and database matching is not sufficient for unambiguous identification. Multiple chemicals often exist for the same elemental formula, and positional isomers with very similar properties can pose a particular challenge for LC-MS and GC-MS identification. For high abundance ions likely to be metabolic intermediates, preferred metabolomics databases are HMDB and KEGG. However, in analysis of human samples, five features with accurate mass

identical to phenylalanine were observed. Searching ChemSpider [96] and METLIN [97] using the +H adduct of phenylalanine at a mass error threshold of ± 0.002 da identified 1,742 and 15 matches, respectively. Due to the presence of redundant database entries in Metlin and a large number of synthetic chemicals in ChemSpider, this example most likely over-estimates the number of unique chemicals that can be detected in a single human sample; however, it highlights the vastness of chemical space and difficulty in designating identities based on accurate mass alone. While rule-based annotation, retention time prediction and comparison to retention time index chemicals acting as landmarks improve confidence, complimentary information, such as retention time matching and molecular dissociation patterns relative to authentic standards are required to verify chemical identity. Ultimately, very rigorous standards are required for reliable assignment of correct configurations of very similar isomers.[90, 91] Different schemes have been proposed for ranking identification confidence of chemicals detected using high-resolution mass spectrometry, many of which rely upon comparison to reference spectra and molecular dissociation.[91, 98, 99] Specifically, the levels proposed by Schymanski *et al* [91] provide a clear framework for describing identification confidence of metabolites.

8.2 Ion dissociation. In untargeted metabolomics, the most common methodology for confirming the identity of detected chemicals is through comparison of the ion dissociation pattern (MS^2) obtained for a given precursor mass to reference standards or spectral databases. Ion dissociation is typically achieved through ion collision with inert gas and increased molecular vibrational energy, which disrupt covalent bonds and creates charged fragments that are then detected by a mass analyzer. For soft ionization

techniques, such as electrospray or chemical ionization, the precursor ion is typically the most abundant adduct (i.e. +H, +Na) and the detected fragments (referred to as MS^2 spectra) are consistent with loss of specific functional groups.[100] When trap-based mass filters are in use, extra levels of fragmentation can be achieved through fragmentation of ions obtained in the MS^2 spectrum (MS^n). The resulting fragment trees provide additional structural information and are useful for characterizing unknown molecules.[101] Thus, MS^2 and MS^n spectra are an intrinsic property of a molecule and represent an important dimension of ion definition in multi-vector space (Figure 7). Assembling the million metabolome will require computational approaches for processing, characterizing and utilizing MS^2 spectra.

8.3 Deconvolution of MS^2 spectra. Except for the most abundant chemicals, MS^2 software tools are required to generate clean spectra that accurately reflect fragments corresponding to a given precursor mass (MS^2 deconvolution). While fragments are detected using a high-mass accuracy analyzer, ion selection prior to fragmentation is typically achieved using unit resolution mass filters. To maximize the number of ions selected for fragmentation, a mass selection window of ± 1-2 *m/z* is often used, resulting in co-isolation of interfering ions that are also fragmented. Using the example described above, a theoretical isolation window of ± 1 *m/z* for generating spectra corresponding to the +H adduct of phenylalanine resulted in 28,485 matches in the ChemSpider database. Therefore, it can be expected as a rule, not an exception, that co-eluting compounds will be present during MS^2 analysis.

While many data pre-processing software packages can process MS^2 data, [34, 41, 102, 103] only a limited number provide deconvolution capable of generating sufficiently pure spectra of low abundance ions. To improve the quality of MS^2 data collected in biological samples, Smith *et al* [104] developed decoMS2 to remove interfering peaks and assign specific fragments to precursor ions detected in full scan mode. Deconvolution of MS^2 fragments is achieved with variable isolation windows to introduce variations in ions detected using full scan and MS^2 data; fragments are matched to ion peak shape by fitting a cubic spline. Application of decoMS2 to untargeted metabolomics provided improved detection of fragments and spectral matching scores; however, the need to use four separate scans for generating adequate data limits throughput and application in high-resolution instruments with slower scan speeds. Recently, sequential windowed acquisition of all theoretical fragments (SWATH) approaches have become available for deconvoluting fragments collected using large isolation windows and multiple scan events. [105-107] MS-DIAL is a standalone pre-processing software environment, and includes functions for full scan and MS^2 peak picking, alignment, deisotoping, MS^2 deconvolution and mass spectral searching.[108] Data can be processed from both data-dependent and data-independent scan events, with the latter useful for characterizing specific ions of interest and completing untargeted MS^2 analysis.

8.4 Clustering algorithm improves MS^2 deconvolution. Both of the software tools described above require differences in chromatographic retention time of precursor ions for accurate deconvolution from a single extracted ion chromatograms (EIC) data file. Due to the large number of chemical species that will need to be characterized for the

million metabolome, chromatographic resolution alone will be insufficient for collecting accurate MS^2. Algorithms incorporating biological variation naturally observed in human populations, as well as analytical variation, can be used to enhance detection of spectral fragments by employing full scan and MS^2 alignment followed by the correlation network approach described above. This functionality is available in RamClustR, which was developed by Broeckling *et al* [59] as an open source software package for clustering untargeted, multi-scan event high-resolution MS data. RamClustR provides a critical advance in processing MS^2 data. Through use of clustering in the intensity and time dimensions, spectral features from both full (isotopes, ionization fragments and adducts) and MS^2 (precursor ion fragments) scan data can be grouped based upon correlation and elution profile, providing an additional level of peak assignment not available when considering individual data files. In addition, RamClustR is compatible with a number of different input files, including XCMS objects and text based peak tables, enabling use with different data processing workflows. Currently, the clustering approach used in RamClustR is purely data-driven. Incorporating orthogonal information, such as isolation mass, mass defect and suspected chemical structures based upon full scan data will improve ability to generate sufficiently pure spectra for characterizing chemicals in the million metabolome. To date, no software packages offer all of these capabilities for MS^2 data. Additionally, one can envisage development of knowledgebase tools to further enhance speed and reliability for un-identified features.

8.5 Ongoing need for semiautomated and automated approaches. While advances have been made in algorithms providing extraction and deconvolution of MS^2 data in untargeted metabolomics, there is a pressing need for continued refinement of semi-

automated computational approaches. Of the many software tools currently available, none provide the throughput or capabilities required for the large-scale characterization of the million metabolome. Specifically, algorithms capable of accurately assigning fragments to precursor *m/z* ions with intensity values orders-of-magnitude lower than co-eluting metabolites must be developed. The resulting spectra will be required for uniquely defining chemical vectors in the million-metabolome space. Approaches for characterizing acquired spectra are discussed below.

8.6 Spectral databases. Databases containing both GC-MS and LC-based MS^2 spectral data are available, providing an important reference for identification and classification of features with MS^2 data. [23, 97, 109-116] Database chemical spectra are typically acquired using pure standards, although in some cases spectra acquired from authentic reference standards are complemented with *in silico* generated fragmentation patterns. [112] Matching is accomplished by calculating the similarity between the experimental fragmentation pattern and database spectra. The likelihood of a correct match is assessed with either a similarity or probabilistic score, which can be determined using a number of different calculation and weighting schemes that include information such as fragment masses, relative intensity, number of database chemicals with similar fragmentation patterns, neutral losses and precursor *m/z*. [38, 117] While MS^2 spectral matching provides greater confidence of identification than available from accurate mass matching alone, it is important to recognize that considerable overlap exists in fragmentation patterns due to the limited set of low energy pathways responsible for ion collision during dissociation [117]. Thus, spectral matching often results in false positives; complementary information

and analyst expertise is often needed when evaluating the correctness of a spectra match. Improved computational strategies to integrate complementary data and decrease reliance upon analyst expertise will be needed to improve reliability and throughput for environmental biomonitoring.

8.7 Collision Cross Section (CCS). Ion mobility spectrometry-mass spectrometry (IMS-MS) provides complementary structural information to improve confidence in chemical identification [118] and separation of isomers with the same atomic composition but different structures. In IMS-MS, movement of ions in the gas phase in an electric field is countered by collision of the ions with a buffer gas. Because separation is based on gas-phase mobility and not limited by constraints of solvent or stationary phase, IMS can separate species not easily separated by LC and GC. Uses have included analysis of lipids, [119] metabolites, [120] air pollutants [121] and pharmaceuticals,[122] suggesting a promising future for applications in environmental toxicology research.

The benefit of IMS for environmental exposure research is illustrated by separation of isobaric isoprene epoxy diols (IEPOX) in organic aerosol samples (Figure 8).[121] Organic aerosol species constitute a major fraction of airborne particles contributing to air pollution and impacting human health. The complex mixtures of organic aerosol species are difficult to resolve by commonly used LC-MS methods. In IMS-MS, ions with greater collision cross section (CCS) move more slowly and are separated from those moving more rapidly. IMS separation occurs over a millisecond time frame and is orthogonal to separation by LC or GC so that it adds resolving power. The resolving power is in the range of 20 to 200 for different instruments, but improvements are

ongoing. In the studies described in Figure 8, aerosol filters from different environmental monitoring locations were extracted and treated to convert isoprene epoxy diols (IEPOX) to hydroxysulfate esters. The results show that three different IEPOX isomers were sufficiently resolved to estimate combinations in the different samples. Note that the top bars indicate the uncertainty in drift time for each peak. The signal from the SOAS ambient filter, which does not align well with the other samples, may be due to the uncertainty in drift time. Such limitations can be resolved by improved instrumentation, additional study and improved computational methods. Thus, IMS-MS is expected to become a critical approach for resolution and identification of isobaric species in complex mixtures. Additionally, as discussed below, measured values for CCS, obtained from IMS-MS, could provide information to aid in unambiguous designation of unknown ions.

9. Unambiguous ion characterization and designation: current progress and future directions

For successful detection of a million metabolome, a new contextual construct is required to escape the limitations of studying only known chemicals with annotation databases. Rappaport emphasized that environmental chemicals are often 4-5 orders of magnitude lower abundance than endogenous metabolites.[123] Thus, under conditions where MS^2 is useful for confirmation of identity of endogenous metabolites, this creates challenges for ion dissociation studies of environmental chemicals. Additionally, MS^2 spectra for many environmental chemicals are not available in databases. Unlike accurate mass *m/z*, used for annotation and easily calculated by knowledge of the chemical formula and adduct

form, there are no computational tools available with sufficient throughput to provide accurate estimation of MS^2 spectra. As a result, considerable selection bias exists when using MS^2 databases for annotation or identification. Many of these databases were established for specific classes of chemicals, such as natural products, environmental chemicals, lipids and human metabolites. [124]

9.1 Multi-vector space. The ion characterization methods described above provide the basis for a new contextual construct to designate ions through definition in multi-vector space. Such an approach will require assembly of data for unidentified ions into recurrent spectral databases.[125] Robust measures will be needed to define accurate mass *m/z*, retention time, MS^2 spectra, collision cross section (CCS) and chirality, with each parameter providing a vector to uniquely define an ion within a multi-vector space of the million metabolome. In this framework, ion designation can be unambiguous even though chemical identity is unknown. Chemical identity (defined as 3-dimensional chemical structure) can be added as this becomes available, with priority established when dictated by relevance.

9.2 Accurate mass *m/z*. Mass analyzers are widely available to support measurement of *m/z* within 1 ppm. Such information can be particularly useful as a robust characteristic to describe unidentified ions. Mass resolution is important to assure that *m/z* reflects a single ion, and mass calibration is essential to assure the accuracy of the stated *m/z*.

9.3 Retention time. In LC, chemicals are separated based on partitioning between the stationary and mobile phase. Gradient-based chromatography methods manipulate mobile phase pH, aqueous or organic content over the course of a chromatographic run to elute chemicals from the column. Co-eluting chemicals generally possess similar properties, such as lipophilicity, hydrophobicity, ionic strength, and acid dissociation constant. [126] Therefore, the retention times of chemicals with known structures and physicochemical properties could serve as a reference to deduce qualitative physicochemical properties of an unknown metabolite.[127] For example, chemical retention in reversed phase chromatography is based on a chemical octanol/water partitioning coefficient, a measure of chemical lipophilicity. As lipophilicity increases, a chemical will have greater affinity for the stationary phase, resulting in greater retention times relative to other chemicals. By building a regression model (based upon the physicochemical properties and retention times of known index chemicals), the properties of an unknown chemical could be inferred based on absolute and relative retention time. Extending this concept to other modes of chromatography is also possible. In addition, if an unidentified *m/z* is detected with two orthogonal chromatographic separation techniques (i.e. reversed phase (C18) and HILIC, or anion exchange and C18), metabolite associations and retention time indexing can be cross-validated between multiple platforms.

9.4 Characterizing unknowns by MS^2. Numerous tools exist for characterizing MS^2 fragmentation patterns and predicting identification based upon spectral features. Interpretation is completed using a combination of different strategies and computational

approaches. When using spectral information to characterize detected unknowns, it is useful to classify the underlying methodology as either top-down (*in silico*) or bottom up (structural elucidation). Top-down approaches use theoretical models, often calibrated to experimentally collected MS^2 data, to predict fragmentation patterns based upon bond dissociation energies, ion physics, rearrangement and molecular functional groups. While there are currently no algorithms available providing high-accuracy MS^2 for all the different methods of dissociation, there are multiple heuristic methods that provide sufficient spectra that can be used for improving confidence in annotation and classifying characteristics of fragmented *m/z* ions. Many of the approaches combine *in silico* fragmentation with experimentally collected MS^2 data for fragment annotation and ranking likelihood of a correct identification.

With the recognition that current MS^2 spectra databases are insufficient for annotation of many of the chemical species detected during untargeted profiling, the availability of *in silico* approaches has increased.[128] MetFrag, [129] which was recently updated to improve fragmentation handling, increase computation speed, expand the number of available adduct forms, include suspect screening lists and incorporate retention time information, [130] is capable of providing predicted fragmentation patterns in both large (PubChem and ChemSpider), and specific databases (HMDB, KEGG and NORMAN). CFM-ID provides estimated fragmentation based upon a probabilistic, generative model. This provides functionalities for spectra prediction based upon ionization type and chemical IUPAC International Chemical Identifier (InChI), fragment peak annotation given chemical structure and spectra collected at low, medium and high energy settings, and

compound identification based upon comparison of predicted MS^2 spectra to experimentally collected data. [131, 132] Hybrid approaches have also been developed that leverage existing MS^2 spectral databases in combination with *in silico* fragmentation for reducing the number of suspected identifications. For example, MetFusion combines MS^2 spectral databases from MassBank and METLIN and *in silico* prediction using MetFrag, providing improved ranking of the correct chemical structure compared to that available from database matching or predicted fragmentation alone. [133] Improvement in identification accuracy is achieved through merging chemical similarity information, making possible the determination of *in silico* fragmentation accuracy when compared to experimentally generated MS^2 spectra. Computational approaches for creating fragmentation trees of *in silico* fragmentation patterns are also available, which are useful for further characterizing fragment structures and in structural elucidation. CSI:FingerID, which was developed to assist in annotating the "dark-matter of the metabolome",[134] uses a machine learning approach and reference chemical dataset to compute similarities between molecular fingerprints and fragmentation trees for predicting chemical structures based upon user provided MS^2 through MS^n. [135] The molecular fingerprints consist of 1,415 molecular properties, which are obtained from PubChem and Klekota–Roth fingerprints and used to identify possible matches based upon support vector machine (SVM) predictions.

There is a long history of using ion dissociation mass spectra for structural elucidation of organic molecules.[65, 136] As discussed above, spectral data are consistent with molecular structure; both fragments and neutral losses can be used to infer molecule properties. A

wide range of techniques and software are available for interpretation of features present in mass spectra.[61] For example, Mass Frontier contains libraries of fragmentation schemes from more than 100,000 individual reaction mechanisms, in addition to functionalities for analyzing and interpreting MS^2 and MS^n spectra. [137] Characterization of neutral losses can be used to identify functional groups, characterize transformation products, predict structure and determine chemical classification. To incorporate common neutral losses into computational spectrum interpretation, Ma *et al* [100] developed MS2Analyzer, which enables searching of MS^2 spectra based upon user-defined parameters, including neutral losses, *m/z* differences, product and precursor ions. The authors provide a list of 147 literature-reported neutral losses, which were used in validation studies of previously collected MS^2 data.

Molecular networking of MS^2 data[138] will also be important for development of the million metabolome. Molecular networking, which was originally developed as a dereplication strategy for natural product identification, uses a similarity network determined from spectrum relatedness to identify structurally similar chemicals. The resulting network can be used to identify chemicals sharing similar structural components and biotransformations. [139] Further development of molecular networking to include ion definition parameters in multi-vector space will be an important component of cross-laboratory and cross-platform assembly of the million metabolome.

9.5 Collision Cross Section. As indicated above, CCS provides an important complementary property to aid chemical identification. Similarly, for unambiguous

designation of unidentified ions, CCS provides a useful characteristic that is independent of MS^2, and retention time and partially independent of *m/z*. IMS technology is well developed and applications of IMS are extensive; recent introduction of IMS-MS from multiple manufacturers offers new opportunities for environmental chemical research (see Figure 8). Because separation is based upon different principles than GC or LC separation, IMS-MS can provide additional characterization not otherwise available. Several computational algorithms such as the trajectory method, exact hard sphere scattering method, and the projection approximation method have been developed for computationally predicting CCS values.[140-143] Recent studies have shown that combination of experimental CCS values obtained from IMS-MS, molecular modeling techniques, and theoretical CCS values obtained using MOBCAL or Sigma can aid in structural identification of drug metabolites, lipids, small molecules, and unknown structural isomers.[118, 144-146] Importantly, development of automated frameworks to use IMS-MS to determine CCS in combination with computational methods would greatly facilitate unambiguous ion designation for large numbers of unidentified ions.

9.6 Stereochemistry. Stereoisomers of biomolecules are well known[147] and, despite extensive study in chemistry, have been largely ignored in development of high-throughput metabolomics methods. When assembling the million metabolome, the ability to differentiate between forms will be an essential requirement due to enzyme selectivity and difference in biological effects.[148] For environmental chemicals, toxic interaction of chemicals with biological macromolecules also can be stereo-selective so that different stereoisomers can have different toxicity profiles. For instance, L- and D-

amino acids exist in human plasma at a ratio of >100:1. If the two co-elute by chromatography, the higher abundance form predominates so that changes in the abundance of the toxic, low-abundance form is not measured. This informs the broader challenge to environmental chemical analysis, i.e., a higher abundance form could mask the toxicity of a lower abundance stereoisomer. Consequently, separation and characterization of ion configuration is required for defining ions in multi-vector space.

Diastereomers, which are stereoisomers with different configurations of related stereocenters, will often exhibit different chromatographic behavior or molecular volumes. Therefore, it will often be possible to provide vector characterization for diastereoisomers based upon retention time relative to landmark ions and CCS. Standard chromatographic and mass spectrometer operating conditions do not provide sufficient selectivity for selective enantiomer detection, however, and additional analytical procedures will probably be required for defining chirality of many environmental chemicals.

The challenge is illustrated in Figure 9, where chiral chromatography was used to separate (R)- and (S)- forms of L-methionine-sulfoxide. Anion exchange (AE) chromatography provided insufficient separation of the two enantiomers but chiral chromatography detected both forms. Ion dissociation of the two enantiomers, performed at low resolution in the ion trap, showed identical fragmentation patterns, highlighting the need for chiral separation. Chiral chromatographic phases are available for LC, GC and capillary electrophoresis platforms for a wide range of applications. [149, 150]

Mass spectrometer operational parameters can also be altered to provide entantiomer discrimination. [151] Enantiomer selective chemical ionization is well developed,[152] but reagent gases must be selected for specific applications and this limits use in untargeted measures. In the study by Yao *et al*, [153] addition of chiral-selector chemicals to chromatographic mobile phase (including L- or D-*N*-tert-butoxycarbonylphenylalanine, L- or D-*N*-tert-butoxycarbonylproline, and L- or D-*N*-tert-butoxycarbonyl-O-benzylserine) enabled chiral recognition of 19 common amino acids due to enantiomer-specific disassociation efficiency of the diastereometric complex ions formed during ionization. Enantiomer selective detection is also possible through relative ion mobility in the presence of a chiral reagent drift gas.[154] Continued development of IMS- and selector-based measures for chirality is expected to vastly improve the ability to provide untargeted assessment of stereochemistry.

10. Conclusion and Perspective

Substantial advances in analytical chemistry have occurred through application of computational metabolomic methods to improve data extraction, reliability and interpretation of data from high-resolution mass spectral analyses of biological samples. The methods developed for computation metabolomics have built upon the important accomplishment of providing high confidence measures of 300-700 metabolites through targeted metabolomics; the expanded capabilities now enable moderate to high confidence measures of >2000 metabolites with representative metabolites in most metabolic pathways. Although advancements have been made, as illustrated in Figure 1, metabolomics is still in its early stages of development.

These knowledge-based approaches are limited by an inability to address the extensive range of chemicals to which humans are exposed. The greatest limitation lies in cost for targeted analyses, which cannot reasonably be expected to support measurement of tens of thousands of chemicals in large populations. Regulatory policies use risk assessment to minimize hazardous exposures and reduce the need for biomonitoring. As a result, most high production chemicals (30,000-40,000) are not monitored in the general population. Without known hazards, there is also little justification for studying the large diversity of natural chemicals to evaluate their health risks. Similarly, health risks of many chemicals originating from dietary, microbiome, therapeutic, commercial and environmental sources have not been evaluated, largely due to the costly and/or limited coverage of contemporary methods.

A second limitation of knowledge-based approaches lies in the mass spectral data libraries, which are highly biased and limited in coverage. HMDB contains 42,000 metabolites, which have accumulated at a rate of about 6,000 per year since inception. To obtain chemical identities of one million metabolites at this rate would take about 158 years. Consequently, there is a need for prioritization of ions for identification to maximally benefit society. A third limitation exists in the abundance of ions detected, many of which are too low to allow MS^2. Improved computational algorithms and noise reduction methods will be critical to address this challenge.

To address these limitations, we propose development of a multi-vector grid to designate unidentified and low abundance ions in terms of accurate mass *m/z*, indexed chromatographic retention time, intensity, MS^2, collision cross section, and chiral form. Development of a multi-vector ion definition grid will require a computational

framework that can merge information from multiple data sources and enhance the identification process for low and high abundance ions. Furthermore, hybrid network and pathway analysis approaches can be used to characterize unidentified ions by taking advantage of data-driven network structure, relationship of unidentified ions with other –omic measures and preexisting knowledge in pathway databases. Such a multi-dimensional system to characterize the million metabolome will facilitate chemical identification and improve understanding of environmental causes of human disease.

Conflict of Interest Disclosure

The authors declare no competing financial interest.

Funding information

The authors acknowledge support by NIH grants ES023485, ES019776, HL113451, OD018006, AG038746, ES025632, HL086773, California Breast Cancer Research Program 21UB-8002 and NIH contracts 1U2CES026560-01, HHSN272201200031C and HHSN27200009.

Abbreviations

CCS, Collision Cross Section; CV, Coefficient of Variation; EIC, Extracted Ion Chromatogram; GC, Gas Chromatography; HRM, High-resolution metabolomics; IMS-MS, Ion mobility spectrometry-mass spectrometry; InChI, International Chemical Identifier; LC, Liquid Chromatography; MDF, Mass Defect Filtering; MS, Mass

Spectrometry; MWAS, Metabolome-Wide Association Studies; m/z , mass-to-charge ratio; PLS, Partial Least Squares; QC, Quality control;

SVM, Support Vector Machine.

References

(1) Jones, D. P. (2016) Sequencing the exposome: A call to action. *Toxicology Rep. 3*, 29-45.

(2) Walker, D. I., Go, Y.-M., Liu, K., Pennell, K., and D. Jones, D. P. (2016) *Population Screening for Biological and Environmental Properties of the Human Metabolic Phenotype: Implications for Personalized Medicine*. Vol. 7, Elsevier.

(3) Walker, D. I., Mallon, T. M., Hopke, P. K., Uppal, K., Go, Y. M., Rohrbeck, P., Pennell, K. D., and Jones, D. P. (2016) Deployment-Associated Exposure Surveillance with High-Resolution Metabolomics. *J. Occup. Environ. Med. 58*, S12-21.

(4) Dennis, K. K., Marder, E., Balshaw, D. M., Cui, Y., Lynes, M. A., Patti, G. J., Rappaport, S. M., Shaughnessy, D. T., Vrijheid, M., and Barr, D. B. (2016) Biomonitoring in the Era of the Exposome. *Environ. Health Perspect.*

(5) Nicholson, J. K., Lindon, J. C., and Holmes, E. (1999) 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica 29*, 1181-1189.

(6) Lenz, E. M., Bright, J., Wilson, I. D., Morgan, S. R., and Nash, A. F. (2003) A 1H NMR-based metabonomic study of urine and plasma samples obtained from healthy human subjects. *J. Pharm. Biomed. Anal. 33*, 1103-1115.

(7) Johnson, J. M., Yu, T., Strobel, F. H., and Jones, D. P. (2010) A practical approach to detect unique metabolic patterns for personalized medicine. *The Analyst 135*, 2864-2870.

(8) Jones, D. P., Park, Y., and Ziegler, T. R. (2012) Nutritional metabolomics: progress in addressing complexity in diet and health. *Annu. Rev. Nutr. 32*, 183-202.

(9) Marshall, A. G., and Hendrickson, C. L. (2008) High-resolution mass spectrometers. *Annu. Rev. Anal. Chem. 1*, 579-599.

(10) Makarov, A. (2000) Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal. Chem. 72*, 1156-1162.

(11) Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M., and Graham Cooks, R. (2005) The Orbitrap: a new mass spectrometer. *J. Mass Spectrom. 40*, 430-443.

(12) Athersuch, T. (2016) Metabolome analyses in exposome studies: Profiling methods for a vast chemical space. *Arch. Biochem. Biophys. 589*, 177-186.

(13) Uppal, K., Soltow, Q. A., Strobel, F. H., Pittard, W. S., Gernert, K. M., Yu, T., and Jones, D. P. (2013) xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. *BMC Bioinformatics 14*, 15.

(14) Scalbert, A., Brennan, L., Fiehn, O., Hankemeier, T., Kristal, B. S., van Ommen, B., Pujos-Guillot, E., Verheij, E., Wishart, D., and Wopereis, S. (2009) Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics 5*, 435-458.

(15) Go, Y. M., Walker, D. I., Liang, Y., Uppal, K., Soltow, Q. A., Tran, V., Strobel, F., Quyyumi, A. A., Ziegler, T. R., Pennell, K. D., Miller, G. W., and Jones, D. P. (2015) Reference Standardization for Mass Spectrometry and High-resolution Metabolomics Applications to Exposome Research. *Toxicol. Sci. 148*, 531-543.

(16) Roca, M., Leon, N., Pastor, A., and Yusa, V. (2014) Comprehensive analytical strategy for biomonitoring of pesticides in urine by liquid chromatography-orbitrap high resolution masss pectrometry. *Journal of chromatography. A 1374*, 66-76.

(17) Jamin, E. L., Bonvallot, N., Tremblay-Franco, M., Cravedi, J. P., Chevrier, C., Cordier, S., and Debrauwer, L. (2014) Untargeted profiling of pesticide metabolites by LC-HRMS: an exposomics tool for human exposure evaluation. *Analytical and bioanalytical chemistry 406*, 1149-1161.

(18) Walker, D. I., Pennell, K., Uppal, K., Xia, X., Hopke, P., Utell, M., Phipps, R., Sime, P., Rohrbeck, P., Mallon, T., and Jones, D. P. (2016) Pilot Metabolome-Wide Association Study of Benzo(a)pyrene in Serum from Military Personnel. *J. Occup. Environ. Med. 58*, S44-52.

(19) Jones, D. P., Walker, D. I., Uppal, K., Rohrbeck, P., Malllon, T. M., and Go, Y. M. (2016) Metabolic Pathways and Networks Associated With Tobacco Use in Military Personnel. *J. Occup. Environ. Med. 58*, S111-116.

(20) Roede, J. R., Uppal, K., Park, Y., Lee, K., Tran, V., Walker, D., Strobel, F. H., Rhodes, S. L., Ritz, B., and Jones, D. P. (2013) Serum metabolomics of slow vs. rapid motor progression Parkinson's disease: a pilot study. *PLoS One 8*, e77629.

(21) Osborn, M. P., Park, Y., Parks, M. B., Burgess, L. G., Uppal, K., Lee, K., Jones, D. P., and Brantley, M. A., Jr. (2013) Metabolome-wide association study of neovascular age-related macular degeneration. *PLoS One 8*, e72737.

(22) Liu, K., Walker, D. I., Uppal, K., Tran, V., Rohrbeck, P., Mallon, T., and Jones, D. P. (2016) High-resolution metabolomics assessment of military personnel:

Evaluating analytical strategies for chemical detection. *J. Occup. Environ. Med. 58*, S53-61.

(23) Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R., and Scalbert, A. (2013) HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res. 41*, D801-807.

(24) Scalbert, A., Brennan, L., Manach, C., Andres-Lacueva, C., Dragsted, L. O., Draper, J., Rappaport, S. M., van der Hooft, J. J., and Wishart, D. S. (2014) The food metabolome: a window over dietary exposure. *Am. J. Clin. Nutr. 99*, 1286-1308.

(25) Miller, G. W., and Jones, D. P. (2014) The nature of nurture: refining the definition of the exposome. *Toxicol. Sci. 137*, 1-2.

(26) Wild, C. P. (2005) Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol. Biomarkers Prev. 14*, 1847-1850.

(27) Wild, C. P. (2012) The exposome: from concept to utility. *Int. J. Epidemiol. 41*, 24-32.

(28) Idle, J. R., and Gonzalez, F. J. (2007) Metabolomics. *Cell Metab. 6*, 348-351.

(29) Rappaport, S. M., and Smith, M. T. (2010) Epidemiology. Environment and disease risks. *Science 330*, 460-461.

(30) Breunig, J. S., Hackett, S. R., Rabinowitz, J. D., and Kruglyak, L. (2014) Genetic basis of metabolome variation in yeast. *PLoS Genet. 10*, e1004142.

(31) Draisma, H. H., Pool, R., Kobl, M., Jansen, R., Petersen, A. K., Vaarhorst, A. A., Yet, I., Haller, T., Demirkan, A., Esko, T., Zhu, G., Bohringer, S., Beekman, M., van Klinken, J. B., Romisch-Margl, W., Prehn, C., Adamski, J., de Craen, A. J., van Leeuwen, E. M., Amin, N., Dharuri, H., Westra, H. J., Franke, L., de Geus, E. J., Hottenga, J. J., Willemsen, G., Henders, A. K., Montgomery, G. W., Nyholt, D. R., Whitfield, J. B., Penninx, B. W., Spector, T. D., Metspalu, A., Eline Slagboom, P., van Dijk, K. W., t Hoen, P. A., Strauch, K., Martin, N. G., van Ommen, G. J., Illig, T., Bell, J. T., Mangino, M., Suhre, K., McCarthy, M. I., Gieger, C., Isaacs, A., van Duijn, C. M., and Boomsma, D. I. (2015) Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nat. Commun. 6*, 7208.

(32) Fardet, A., Llorach, R., Orsoni, A., Martin, J. F., Pujos-Guillot, E., Lapierre, C., and Scalbert, A. (2008) Metabolomics provide new insight on the metabolism of dietary phytochemicals in rats. *J. Nutr. 138*, 1282-1287.

(33) Rezzi, S., Ramadan, Z., Fay, L. B., and Kochhar, S. (2007) Nutritional metabonomics: applications and perspectives. *J. Proteome Res. 6*, 513-525.

(34) Edmands, W. M., Barupal, D. K., and Scalbert, A. (2015) MetMSLine: an automated and fully integrated pipeline for rapid processing of high-resolution LC-MS metabolomic datasets. *Bioinformatics 31*, 788-790.

(35) Park, Y. H., Lee, K., Soltow, Q. A., Strobel, F. H., Brigham, K. L., Parker, R. E., Wilson, M. E., Sutliff, R. L., Mansfield, K. G., Wachtman, L. M., Ziegler, T. R., and Jones, D. P. (2012) High-performance metabolic profiling of plasma from

seven mammalian species for simultaneous environmental chemical surveillance and bioeffect monitoring. *Toxicology 295*, 47-55.

(36) Dunn, W. B., Brown, M., Stephanie, A., Worton, K. D., Jones, R. L., Kell, D. B., and Heazell, A. E. P. (2011) The metabolome of human placental tissue: investigation of first trimester tissue and changes related to preeclampsia in late pregnancy. *Metabolomics 8*, 579-597.

(37) Johnson, C. H., Ivanisevic, J., Benton, H. P., and Siuzdak, G. (2015) Bioinformatics: the next frontier of metabolomics. *Anal. Chem. 87*, 147-156.

(38) Neumann, S., and Bocker, S. (2010) Computational mass spectrometry for metabolomics: identification of metabolites and small molecules. *Anal. Bioanal. Chem. 398*, 2779-2788.

(39) Sawada, Y., Akiyama, K., Sakata, A., Kuwahara, A., Otsuki, H., Sakurai, T., Saito, K., and Hirai, M. Y. (2009) Widely targeted metabolomics based on large-scale MS/MS data for elucidating metabolite accumulation patterns in plants. *Plant Cell Physiol. 50*, 37-47.

(40) Zhou, J., Liu, H., Liu, Y., Liu, J., Zhao, X., and Yin, Y. (2016) Development and Evaluation of a Parallel Reaction Monitoring Strategy for Large-Scale Targeted Metabolomics Quantification. *Anal. Chem. 88*, 4478-4486.

(41) Pluskal, T., Castillo, S., Villar-Briones, A., and Oresic, M. (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics 11*, 395.

(42) Tautenhahn, R., Bottcher, C., and Neumann, S. (2008) Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics 9*, 504.

(43) Yu, T., Park, Y., Johnson, J. M., and Jones, D. P. (2009) apLCMS--adaptive processing of high-resolution LC/MS data. *Bioinformatics 25*, 1930-1936.

(44) Wei, X., Shi, X., Kim, S., Zhang, L., Patrick, J. S., Binkley, J., McClain, C., and Zhang, X. (2012) Data preprocessing method for liquid chromatography-mass spectrometry based metabolomics. *Anal. Chem. 84*, 7963-7971.

(45) Kanehisa, M., and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res. 28*, 27-30.

(46) Yu, T., and Jones, D. P. (2014) Improving peak detection in high-resolution LC/MS metabolomics data using preexisting knowledge and machine learning approach. *Bioinformatics 30*, 2941-2948.

(47) Lange, E., Tautenhahn, R., Neumann, S., and Gropl, C. (2008) Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics 9*, 375.

(48) Mahieu, N. G., Spalding, J. L., and Patti, G. J. (2016) Warpgroup: increased precision of metabolomic data processing by consensus integration bound analysis. *Bioinformatics 32*, 268-275.

(49) Libiseller, G., Dvorzak, M., Kleb, U., Gander, E., Eisenberg, T., Madeo, F., Neumann, S., Trausinger, G., Sinner, F., Pieber, T., and Magnes, C. (2015) IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics 16*, 118.

(50) Xia, J., Mandal, R., Sinelnikov, I. V., Broadhurst, D., and Wishart, D. S. (2012) MetaboAnalyst 2.0--a comprehensive server for metabolomic data analysis. *Nucleic Acids Res. 40*, W127-133.

(51) Hughes, G., Cruickshank-Quinn, C., Reisdorph, R., Lutz, S., Petrache, I., Reisdorph, N., Bowler, R., and Kechris, K. (2014) MSPrep--summarization, normalization and diagnostics for processing of mass spectrometry-based metabolomic data. *Bioinformatics 30*, 133-134.

(52) Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics 28*, 882-883.

(53) Johnson, W. E., Li, C., and Rabinovic, A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics 8*, 118-127.

(54) Kind, T., and Fiehn, O. (2006) Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics 7*, 234.

(55) Shahaf, N., Franceschi, P., Arapitsas, P., Rogachev, I., Vrhovsek, U., and Wehrens, R. (2013) Constructing a mass measurement error surface to improve automatic annotations in liquid chromatography/mass spectrometry based metabolomics. *Rapid Commun. Mass Spectrom. 27*, 2425-2431.

(56) Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabasi, A. L. (2002) Hierarchical organization of modularity in metabolic networks. *Science 297*, 1551-1555.

(57) Steuer, R. (2006) Review: on the analysis and interpretation of correlations in metabolomic data. *Briefings in bioinformatics 7*, 151-158.

(58) Alonso, A., Julia, A., Beltran, A., Vinaixa, M., Diaz, M., Ibanez, L., Correig, X., and Marsal, S. (2011) AStream: an R package for annotating LC/MS metabolomic data. *Bioinformatics 27*, 1339-1340.

(59) Broeckling, C. D., Afsar, F. A., Neumann, S., Ben-Hur, A., and Prenni, J. E. (2014) RAMClust: a novel feature clustering method enables spectral-matching-based annotation for metabolomics data. *Anal. Chem. 86*, 6812-6817.

(60) Brown, M., Wedge, D. C., Goodacre, R., Kell, D. B., Baker, P. N., Kenny, L. C., Mamas, M. A., Neyses, L., and Dunn, W. B. (2011) Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets. *Bioinformatics 27*, 1108-1112.

(61) Kind, T., and Fiehn, O. (2010) Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal. Rev. 2*, 23-60.

(62) Holmes, E., Wilson, I. D., and Nicholson, J. K. (2008) Metabolic phenotyping in health and disease. *Cell 134*, 714-717.

(63) Nicholson, J. K., Holmes, E., and Elliott, P. (2008) The metabolome-wide association study: a new look at human disease risk factors. *J. Proteome Res. 7*, 3637-3638.

(64) Uppal, K., Soltow, Q. A., Promislow, D. E., Wachtman, L. M., Quyyumi, A. A., and Jones, D. P. (2015) MetabNet: An R Package for Metabolic Association Analysis of High-Resolution Metabolomics Data. *Front. Bioeng. Biotechnol. 3*, 87.

(65) McLafferty, F. W., and Tureček, F. E. (1993) *Interpretation of mass spectra*. University Science Books, Mill Valley.

(66) Kim, J. S., Monroe, M. E., Camp, D. G., 2nd, Smith, R. D., and Qian, W. J. (2013) In-source fragmentation and the sources of partially tryptic peptides in shotgun proteomics. *J. Proteome Res. 12*, 910-916.

(67) Xu, Y. F., Lu, W., and Rabinowitz, J. D. (2015) Avoiding misannotation of in-source fragmentation products as cellular metabolites in liquid chromatography-mass spectrometry-based metabolomics. *Anal. Chem. 87*, 2273-2281.

(68) Purwaha, P., Silva, L. P., Hawke, D. H., Weinstein, J. N., and Lorenzi, P. L. (2014) An artifact in LC-MS/MS measurement of glutamine and glutamic acid: in-source cyclization to pyroglutamic acid. *Anal. Chem. 86*, 5633-5637.

(69) Krumsiek, J., Suhre, K., Evans, A. M., Mitchell, M. W., Mohney, R. P., Milburn, M. V., Wagele, B., Romisch-Margl, W., Illig, T., Adamski, J., Gieger, C., Theis, F. J., and Kastenmuller, G. (2012) Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. *PLoS Genet. 8*, e1003005.

(70) Roede, J. R., Uppal, K., Park, Y., Tran, V., and Jones, D. P. (2014) Transcriptome–metabolome wide association study (TMWAS) of maneb and paraquat neurotoxicity reveals network level interactions in toxicologic mechanism *Toxicology Rep. 1*, 435-444.

(71) Go, Y. M., Roede, J. R., Orr, M., Liang, Y., and Jones, D. P. (2014) Integrated redox proteomics and metabolomics of mitochondria to identify mechanisms of cd toxicity. *Toxicol. Sci. 139*, 59-73.

(72) Cribbs, S. K., Uppal, K., Li, S., Jones, D. P., Huang, L., Tipton, L., Fitch, A., Greenblatt, R. M., Kingsley, L., Guidot, D. M., Ghedin, E., and Morris, A. (2016)

Correlation of the lung microbiota with metabolic profiles in bronchoalveolar lavage fluid in HIV infection. *Microbiome 4*, 3.

(73) Katajamaa, M., and Oresic, M. (2005) Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics 6*, 179.

(74) Zhang, H., Zhang, D., Ray, K., and Zhu, M. (2009) Mass defect filter technique and its applications to drug metabolite identification by high-resolution mass spectrometry. *J. Mass Spectrom. 44*, 999-1016.

(75) Sleno, L. (2012) The use of mass defect in modern mass spectrometry. *J. Mass Spectrom. 47*, 226-236.

(76) Breitling, R., Ritchie, S., Goodenowe, D., Stewart, M. L., and Barrett, M. P. (2006) prediction of metabolic networks using Fourier transform mass spectrometry data. *Metabolomics 2*, 155-164.

(77) Jobst, K. J., Shen, L., Reiner, E. J., Taguchi, V. Y., Helm, P. A., McCrindle, R., and Backus, S. (2013) The use of mass defect plots for the identification of (novel) halogenated contaminants in the environment. *Anal. Bioanal. Chem. 405*, 3289-3297.

(78) Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C. A., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., and Karp, P. D. (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res. 44*, D471-480.

(79) Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res. 42*, D199-205.

(80) Thiele, I., Swainston, N., Fleming, R. M., Hoppe, A., Sahoo, S., Aurich, M. K., Haraldsdottir, H., Mo, M. L., Rolfsson, O., Stobbe, M. D., Thorleifsson, S. G., Agren, R., Bolling, C., Bordel, S., Chavali, A. K., Dobson, P., Dunn, W. B., Endler, L., Hala, D., Hucka, M., Hull, D., Jameson, D., Jamshidi, N., Jonsson, J. J., Juty, N., Keating, S., Nookaew, I., Le Novere, N., Malys, N., Mazein, A., Papin, J. A., Price, N. D., Selkov, E., Sr., Sigurdsson, M. I., Simeonidis, E., Sonnenschein, N., Smallbone, K., Sorokin, A., van Beek, J. H., Weichart, D., Goryanin, I., Nielsen, J., Westerhoff, H. V., Kell, D. B., Mendes, P., and Palsson, B. O. (2013) A community-driven global reconstruction of human metabolism. *Nat. Biotechnol. 31*, 419-425.

(81) Herrgard, M. J., Swainston, N., Dobson, P., Dunn, W. B., Arga, K. Y., Arvas, M., Bluthgen, N., Borger, S., Costenoble, R., Heinemann, M., Hucka, M., Le Novere, N., Li, P., Liebermeister, W., Mo, M. L., Oliveira, A. P., Petranovic, D., Pettifer, S., Simeonidis, E., Smallbone, K., Spasic, I., Weichart, D., Brent, R., Broomhead, D. S., Westerhoff, H. V., Kirdar, B., Penttila, M., Klipp, E., Palsson, B. O., Sauer, U., Oliver, S. G., Mendes, P., Nielsen, J., and Kell, D. B. (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat. Biotechnol. 26*, 1155-1160.

(82) Li, S., Park, Y., Duraisingham, S., Strobel, F. H., Khan, N., Soltow, Q. A., Jones, D. P., and Pulendran, B. (2013) Predicting network activity from high throughput metabolomics. *PLoS Comput. Biol. 9*, e1003123.

(83) Cho, K., Mahieu, N. G., Johnson, S. L., and Patti, G. J. (2014) After the feature presentation: technologies bridging untargeted metabolomics and biology. *Curr. Opin. Biotechnol. 28*, 143-148.

(84) Li, S., Dunlop, A. L., Jones, D. P., and Corwin, E. J. (2016) High-Resolution Metabolomics: Review of the Field and Implications for Nursing Science and the Study of Preterm Birth. *Biol. Res. Nurs. 18*, 12-22.

(85) Li, S., Todor, A., and Luo, R. (2016) Blood transcriptomics and metabolomics for personalized medicine. *Comput. Struct. Biotechnol. J. 14*, 1-7.

(86) Xu, X., Araki, K., Li, S., Han, J. H., Ye, L., Tan, W. G., Konieczny, B. T., Bruinsma, M. W., Martinez, J., Pearce, E. L., Green, D. R., Jones, D. P., Virgin, H. W., and Ahmed, R. (2014) Autophagy is essential for effector CD8(+) T cell survival and memory formation. *Nat. Immunol. 15*, 1152-1161.

(87) Ravindran, R., Khan, N., Nakaya, H. I., Li, S., Loebbermann, J., Maddur, M. S., Park, Y., Jones, D. P., Chappert, P., Davoust, J., Weiss, D. S., Virgin, H. W., Ron, D., and Pulendran, B. (2014) Vaccine activation of the nutrient sensor GCN2 in dendritic cells enhances antigen presentation. *Science 343*, 313-317.

(88) Hoffman, J. M., Tran, V., Wachtman, L. M., Green, C. L., Jones, D. P., and Promislow, D. E. (2016) A longitudinal analysis of the effects of age on the blood plasma metabolome in the common marmoset, Callithrix jacchus. *Exp. Gerontol. 76*, 17-24.

(89) Jin, R., Banton, S., Tran, V. T., Konomi, J. V., Li, S., Jones, D. P., and Vos, M. B. (2016) Amino Acid Metabolism is Altered in Adolescents with Nonalcoholic Fatty Liver Disease-An Untargeted, High Resolution Metabolomics Study. *J. Pediatr. 172*, 14-19 e15.

(90) Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., Reily, M. D., Thaden, J. J., and Viant, M. R. (2007) Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics 3*, 211-221.

(91) Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P., and Hollender, J. (2014) Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ. Sci. Technol. 48*, 2097-2098.

(92) Daly, R., Rogers, S., Wandy, J., Jankevics, A., Burgess, K. E., and Breitling, R. (2014) MetAssign: probabilistic annotation of metabolites from LC-MS data using a Bayesian clustering approach. *Bioinformatics 30*, 2764-2771.

(93) Kuhl, C., Tautenhahn, R., Bottcher, C., Larson, T. R., and Neumann, S. (2012) CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem. 84*, 283-289.

(94) Silva, R. R., Jourdan, F., Salvanha, D. M., Letisse, F., Jamin, E. L., Guidetti-Gonzalez, S., Labate, C. A., and Vencio, R. Z. (2014) ProbMetab: an R package

for Bayesian probabilistic annotation of LC-MS-based metabolomics. *Bioinformatics 30*, 1336-1337.

(95) Kind, T., and Fiehn, O. (2007) Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics 8*, 105.

(96) Pence, H. E., and Williams, A. (2010) ChemSpider: An Online Chemical Information Resource. *J. Chem. Educ. 87*, 1123-1124.

(97) Smith, C. A., O'Maille, G., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., Custodio, D. E., Abagyan, R., and Siuzdak, G. (2005) METLIN: a metabolite mass spectral database. *Ther. Drug Monit. 27*, 747-751.

(98) Creek, D. J., Dunn, W. B., Fiehn, O., Griffin, J. L., Hall, R. D., Lei, Z., Mistrik, R., Neumann, S., Schymanski, E. L., Sumner, L. W., Trengove, R., and Wolfender, J.-L. (2014) Metabolite identification: are you sure? And how do your peers gauge your confidence? *Metabolomics 10*, 350-353.

(99) Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., Reily, M. D., Thaden, J. J., and Viant, M. R. (2007) Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics 3*, 211-221.

(100) Ma, Y., Kind, T., Yang, D., Leon, C., and Fiehn, O. (2014) MS2Analyzer: A software for small molecule substructure annotations from accurate tandem mass spectra. *Anal. Chem. 86*, 10724-10731.

(101) Vaniya, A., and Fiehn, O. (2015) Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics. *Trends in analytical chemistry : TRAC 69*, 52-61.

(102) Mayampurath, A. M., Jaitly, N., Purvine, S. O., Monroe, M. E., Auberry, K. J., Adkins, J. N., and Smith, R. D. (2008) DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. *Bioinformatics 24*, 1021-1023.

(103) Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem. 78*, 779-787.

(104) Nikolskiy, I., Mahieu, N. G., Chen, Y. J., Tautenhahn, R., and Patti, G. J. (2013) An untargeted metabolomic workflow to improve structural characterization of metabolites. *Anal. Chem. 85*, 7713-7719.

(105) Arnhard, K., Gottschall, A., Pitterl, F., and Oberacher, H. (2015) Applying 'Sequential Windowed Acquisition of All Theoretical Fragment Ion Mass Spectra' (SWATH) for systematic toxicological analysis with liquid chromatography-high-resolution tandem mass spectrometry. *Anal. Bioanal. Chem. 407*, 405-414.

(106) Peng, H., Chen, C., Saunders, D. M., Sun, J., Tang, S., Codling, G., Hecker, M., Wiseman, S., Jones, P. D., Li, A., Rockne, K. J., and Giesy, J. P. (2015) Untargeted Identification of Organo-Bromine Compounds in Lake Sediments by

Ultrahigh-Resolution Mass Spectrometry with the Data-Independent Precursor Isolation and Characteristic Fragment Method. *Anal. Chem. 87*, 10237-10246.

(107) Zhu, X., Chen, Y., and Subramanian, R. (2014) Comparison of information-dependent acquisition, SWATH, and MS(All) techniques in metabolite identification study employing ultrahigh-performance liquid chromatography-quadrupole time-of-flight mass spectrometry. *Anal. Chem. 86*, 1202-1209.

(108) Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., Kanazawa, M., VanderGheynst, J., Fiehn, O., and Arita, M. (2015) MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods 12*, 523-526.

(109) Bouslimani, A., Sanchez, L. M., Garg, N., and Dorrestein, P. C. (2014) Mass spectrometry of natural products: current, emerging and future technologies. *Nat. Prod. Rep. 31*, 718-729.

(110) Fahy, E., Sud, M., Cotter, D., and Subramaniam, S. (2007) LIPID MAPS online tools for lipid research. *Nucleic Acids Res. 35*, W606-612.

(111) 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, M. Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K., and Nishioka, T. (2010) MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom. 45*, 703-714.

(112) Kind, T., Liu, K. H., Lee do, Y., DeFelice, B., Meissen, J. K., and Fiehn, O. (2013) LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nat. Methods 10*, 755-758.

(113) Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmuller, E., Dormann, P., Weckwerth, W., Gibon, Y., Stitt, M., Willmitzer, L., Fernie, A. R., and Steinhauser, D. (2005) GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics 21*, 1635-1638.

(114) NIST. (2014) NIST/EPA/NIH Mass Spectral Library National Institute of Standards and Technology; US Secretary of Commerce.

(115) Oberacher, H., Whitley, G., and Berger, B. (2013) Evaluation of the sensitivity of the 'Wiley registry of tandem mass spectral data, MSforID' with MS/MS data of the 'NIST/NIH/EPA mass spectral library'. *J. Mass Spectrom. 48*, 487-496.

(116) Sawada, Y., Nakabayashi, R., Yamada, Y., Suzuki, M., Sato, M., Sakata, A., Akiyama, K., Sakurai, T., Matsuda, F., Aoki, T., Hirai, M. Y., and Saito, K. (2012) RIKEN tandem mass spectral database (ReSpect) for phytochemicals: a plant-specific MS/MS-based data resource and database. *Phytochemistry 82*, 38-45.

(117) Stein, S. (2012) Mass spectral reference libraries: an ever-expanding resource for chemical identification. *Anal. Chem. 84*, 7274-7282.

(118) Lapthorn, C., Pullen, F., and Chowdhry, B. Z. (2013) Ion mobility spectrometry-mass spectrometry (IMS-MS) of small molecules: separating and assigning structures to ions. *Mass Spectrom. Rev. 32*, 43-71.

(119) Groessl, M., Graf, S., and Knochenmuss, R. (2015) High resolution ion mobility-mass spectrometry for separation and identification of isomeric lipids. *The Analyst 140*, 6904-6911.

(120) Dwivedi, P., Wu, P., Klopsch, S. J., Puzon, G. J., Xun, L., and Hill, H. H. (2008) Metabolic profiling by ion mobility mass spectrometry (IMMS). *Metabolomics 4*, 63-80.

(121) Krechmer, J. E., Groessl, M., Zhang, X., Junninen, H., Massoli, P., Lambe, A. T., Kimmel, J. R., Cubison, M. J., Graf, S., Lin, Y.-H., Budisulistiorini, S. H., Zhang, H., Surrat, J. D., Knochenmus, R., Jayne, J. T., Worsnop, D. R., Jose-Luis Jimenez, J.-L., and Canagaratna, M. R. (2016) Ion Mobility Spectrometry-Mass Spectrometry (IMS-MS) for on- and off-line analysis of atmospheric gas and aerosol species. *Atmos. Meas. Tech. Discuss.*

(122) Kanu, A. B., Dwivedi, P., Tam, M., Matz, L., and Hill, H. H., Jr. (2008) Ion mobility-mass spectrometry. *J. Mass Spectrom. 43*, 1-22.

(123) Rappaport, S. M., Barupal, D. K., Wishart, D., Vineis, P., and Scalbert, A. (2014) The blood exposome and its role in discovering causes of disease. *Environ. Health Perspect. 122*, 769-774.

(124) Vinaixa, M., Schymanski, E. L., Neumann, S., Navarro, M., Salek, R. M., and Yanes, O. (2016) Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *TrAC, Trends Anal. Chem. 78*, 23-35.

(125) Mallard, W. G., Andriamaharavo, N. R., Mirokhin, Y. A., Halket, J. M., and Stein, S. E. (2014) Creation of libraries of recurring mass spectra from large data sets assisted by a dual-column workflow. *Anal. Chem. 86*, 10231-10238.

(126) Karger, B. L., Snyder, L. R., and Horvath, C. (1973) *Introduction to separation science*. Wiley-Interscience.

(127) Boswell, P. G., Schellenberg, J. R., Carr, P. W., Cohen, J. D., and Hegeman, A. D. (2011) Easy and accurate high-performance liquid chromatography retention prediction with different gradients, flow rates, and instruments by back-calculation of gradient and flow rate profiles. *J. Chromatogr. A 1218*, 6742-6749.

(128) Rathahao-Paris, E., Alves, S., Junot, C., and Tabet, J.-C. (2015) High resolution mass spectrometry for structural identification of metabolites in metabolomics. *Metabolomics 12*, 1-15.

(129) Wolf, S., Schmidt, S., Muller-Hannemann, M., and Neumann, S. (2010) In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics 11*, 148.

(130) Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J., and Neumann, S. (2016) MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J. Cheminform. 8*, 3.

(131) Allen, F., Greiner, R., and Wishart, D. (2015) Comparative fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics 11*, 98-110.

(132) Allen, F., Pon, A., Wilson, M., Greiner, R., and Wishart, D. (2014) CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res. 42*, W94-99.

(133) Gerlich, M., and Neumann, S. (2013) MetFusion: integration of compound identification strategies. *J. Mass Spectrom. 48*, 291-298.

(134) da Silva, R. R., Dorrestein, P. C., and Quinn, R. A. (2015) Illuminating the dark matter in metabolomics. *Proc. Natl. Acad. Sci. U. S. A. 112*, 12549-12550.

(135) Duhrkop, K., Shen, H., Meusel, M., Rousu, J., and Bocker, S. (2015) Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci. U. S. A. 112*, 12580-12585.

(136) Boyd, B., Basic, C., and Bethem, R. (2008) *Trace quantitative analysis by mass spectrometry*. John Wiley & Sons, Chichester in England.

(137) HighChem. (2015) MassFrontier v. 7. , Bratislava, HighChem Ltd.

(138) Yang, J. Y., Sanchez, L. M., Rath, C. M., Liu, X., Boudreau, P. D., Bruns, N., Glukhov, E., Wodtke, A., de Felicio, R., Fenner, A., Wong, W. R., Linington, R. G., Zhang, L., Debonsi, H. M., Gerwick, W. H., and Dorrestein, P. C. (2013) Molecular Networking as a Dereplication Strategy. *J. Nat. Prod. 76*, 1686-1699.

(139) Quinn, R. A., Phelan, V. V., Whiteson, K. L., Garg, N., Bailey, B. A., Lim, Y. W., Conrad, D. J., Dorrestein, P. C., and Rohwer, F. L. (2016) Microbial, host and xenobiotic diversity in the cystic fibrosis sputum metabolome. *ISME J. 10*, 1483-1498.

(140) Mesleh, M. F., Hunter, J. M., Shvarsburg, A. A., Schatz, G. C., and Jarrold, M. F. (1996) Structural Information from Ion Mobility Measurements: Effects of the Long-Range Potential. *J. Phys. Chem. 100*, 16082-16086.

(141) Shvartsburg, A. A., and Jarrold, M. F. (1996) An exact hard-spheres scattering model for the mobilities of polyatomic ions. *Chem. Phys. Lett. 261*, 86-91.

(142) D'Atri, V., Porrini, M., Rosu, F., and Gabelica, V. (2015) Linking molecular models with ion mobility experiments. Illustration with a rigid nucleic acid structure. *J. Mass Spectrom. 50*, 711-726.

(143) Shvarsburg, A. A., Schatz, G. C., and Jarrold, M. F. (1998) Mobilities of carbon cluster ions: critical importance of the molecular attractive potential. *J. Chem. Phys. 108*, 2416.

(144) Campuzano, I., Bush, M. F., Robinson, C. V., Beaumont, C., Richardson, K., Kim, H., and Kim, H. I. (2012) Structural characterization of drug-like compounds by ion mobility mass spectrometry: comparison of theoretical and experimentally derived nitrogen collision cross sections. *Anal. Chem. 84*, 1026-1033.

(145) Paglia, G., Kliman, M., Claude, E., Geromanos, S., and Astarita, G. (2015) Applications of ion-mobility mass spectrometry for lipid analysis. *Anal. Bioanal. Chem. 407*, 4995-5007.

(146) Reading, E., Munoz-Muriedas, J., Roberts, A. D., Dear, G. J., Robinson, C. V., and Beaumont, C. (2016) Elucidation of Drug Metabolite Structural Isomers Using Molecular Modeling Coupled with Ion Mobility Mass Spectrometry. *Anal. Chem. 88*, 2273-2280.

(147) Armstrong, D. W., Gasper, M., Lee, S. H., Zukowski, J., and Ercal, N. (1993) D-amino acid levels in human physiological fluids. *Chirality 5*, 375-378.

(148) McConathy, J., and Owens, M. J. (2003) Stereochemistry in Drug Action. *Prim. Care Companion J. Clin. Psychiatry 5*, 70-73.

(149) Stalcup, A. M. (2010) Chiral separations. *Annu. Rev. Anal. Chem. 3*, 341-363.

(150) Cavazzini, A., Pasti, L., Massi, A., Marchetti, N., and Dondi, F. (2011) Recent applications in chiral high performance liquid chromatography: a review. *Anal. Chim. Acta 706*, 205-222.

(151) Schug, K. A., and W. Lindner, W. (2005) Chiral molecular recognition for the detection and analysis of enantiomers by mass spectrometric methods. *J. Sep. Sci. 28*, 1932-1955.

(152) Filippi, A., Giardini, A., Piccirillo, S., and Speranza, M. (2000) Gas-phase enantioselectivity. *Int. J. Mass spectrom. 198*, 137-163.

(153) Yao, Z. P., Wan, T. S., Kwong, K. P., and Che, C. T. (2000) Chiral analysis by electrospray ionization mass spectrometry/mass spectrometry. 1. Chiral recognition of 19 common amino acids. *Anal. Chem. 72*, 5383-5393.

(154) Dwivedi, P., Wu, C., Matz, L. M., Clowers, B. H., Siems, W. F., and Hill, H. H., Jr. (2006) Gas-phase chiral separations by ion mobility spectrometry. *Anal. Chem. 78*, 8200-8206.

(155) Burgess, L. G., Uppal, K., Walker, D. I., Roberson, R. M., Tran, V., Parks, M. B., Wade, E. A., May, A. T., Umfress, A. C., Jarrell, K. L., Stanley, B. O., Kuchtey, J., Kuchtey, R. W., Jones, D. P., and Brantley, M. A., Jr. (2015) Metabolome-

Wide Association Study of Primary Open Angle Glaucoma. *Invest. Ophthalmol. Vis. Sci. 56*, 5020-5028.

(156) Cribbs, S. K., Park, Y., Guidot, D. M., Martin, G. S., Brown, L. A., Lennox, J., and Jones, D. P. (2014) Metabolomics of bronchoalveolar lavage differentiate healthy HIV-1-infected subjects from controls. *AIDS Res. Hum. Retroviruses 30*, 579-585.

(157) Frediani, J. K., Jones, D. P., Tukvadze, N., Uppal, K., Sanikidze, E., Kipiani, M., Tran, V. T., Hebbar, G., Walker, D. I., Kempker, R. R., Kurani, S. S., Colas, R. A., Dalli, J., Tangpricha, V., Serhan, C. N., Blumberg, H. M., and Ziegler, T. R. (2014) Plasma metabolomics in human pulmonary tuberculosis disease: a pilot study. *PLoS One 9*, e108854.

Biography

Karan Uppal, Ph.D., is an Assistant Professor in the Department of Medicine at Emory University. He received his BS in Biomedical Engineering from University of Iowa in 2007, MS in Bioinformatics from Georgia Institute of Technology in 2009, and Ph.D. in Bioinformatics with a minor in Predictive Analytics from Georgia Institute of Technology in 2015. His primary research focus is computational metabolomics, integrative omics, biomarker discovery, machine learning, and text mining. He has developed several tools and algorithms for metabolomics data processing, annotation, and network analysis. He is also working on identifying metabolic biomarkers of environmental exposures and diseases.

Douglas Walker received his BS in 2009 from the University of Massachusetts-Dartmouth in Civil and Environmental Engineering and is currently a Ph.D. candidate in the Department of Civil and Environmental Engineering at Tufts University. In 2013, he joined the Clinical Biomarkers Laboratory at Emory University, where he is currently employed. The primary focus of his research is to integrate measures of environmental exposure, health outcomes and high-resolution metabolomics. Application of this framework using advanced biostatistic/bioinformatic techniques provides a component for sequencing the human exposome and understanding the contribution of environmental exposures in disease pathophysiology.

Ken Liu has a BS in Chemistry from the Georgia Institute of Technology and is currently

a predoctoral student in the Molecular and Systems Pharmacology Department at Emory University. Ken joined the Clinical Biomarkers Laboratory in 2015 and is involved with projects using high-resolution metabolomics to improve understanding of clinical and experimental drug overdoses.

Shuzhao Li, Ph.D., is an Assistant Professor at Emory University School of Medicine. He develops computational methods for high-dimensional data, and applies those to systems immunology and precision medicine. His mummichog software brought genome-scale metabolic models into the field of high throughput metabolomics, and enabled pathway/network analysis for untargeted metabolomics. A goal of his ongoing work is to use multi-omics, multi-scale models to simulate the immune system and support clinical decisions.

Young-Mi Go, Ph.D., is Assistant Professor of Medicine in Emory University. She obtained Ph.D in Department of Pathology in University of Alabama at Birmingham and her study is focused on studying redox control mechanism and metabolic responses to environmental metals and stressors using cell and animal models. She serves Director of Experimental Metabolomics in Clinical Biomarkers Laboratory at Emory University.

Dean Jones, Ph.D., is Professor of Medicine and Director of the Clinical Biomarkers Laboratory at Emory University, Atlanta, GA. He has degrees in chemistry (BS, Univ Illinois, Urbana) and biochemistry (PhD, Oregon Health Sci Univ, Portland) and postdoctoral training in nutrition (Cornell) and molecular toxiciology (Karolinska

Institute, Stockholm). His research is supported by National Institute of Environmental Health Sciences and other research agencies. He has over 450 peer-reviewed research publications and reviews, largely focused on toxicologic mechanisms and human disease. His recent research has included use of ultra-high resolution mass spectrometry to develop methods to sequence the human exposome.

Table 1. Summary of disease-associated ions without matches in metabolomics databases. Results from human disease studies show that half of the ions significantly associated with disease do not match predicted ions of known metabolites in human metabolic databases.

	No. of significant ions	No. of unmatched ions	%, unmatched ions/significant ions	Reference
Glaucoma	41	12	29	[155]
AMD	40 (from 94)	26	65	[21]
HIV	20	7	35	[156]
Parkinson's disease	259	215	83	[20]
Tuberculosis	61	29	47	[157]
Average	**84.2**	**57.8**	**51.8**	

Figure legends

Figure 1. Increase in metabolomics publications over the last 15 years. Searches of PubMed for "metabolomics" or "metabonomics" with mass spectrometry (MS) or Nuclear Magnetic Resonance (NMR) spectroscopy showed that the pioneering applications of chemometrics to NMR analysis of biological samples resulted in a rapid increase in MS-based studies.

Figure 2. Gap between analytical need and current capabilities for metabolomics analysis of human samples. The human metabolome is estimated to contain 1-3 million chemicals. Most targeted liquid chromatography and gas chromatography based mas spectrometry methods detect 300-700 metabolites, underscoring the substantial need for improved methods to test for chemical exposures associated with human disease. Analytical coverage is improved by probability-based methods providing moderate to high confidence scores for annotations of more than 2000 metabolites. Advanced computational methods facilitate detection of more than 35,000 ions and feasibility studies show detection of 250,000 to 800,000 ions is possible.

Figure 3. High-resolution metabolomics data processing. Similar data processing procedures are used for peak picking and alignment. In xMSanalyzer, which is illustrated here, step one involves noise removal, peak detection, integration, and alignment at multiple parameter settings. In step two, feature and sample quality assessment is performed at each parameter combination. Next, an optimization procedure is performed

by merging and evaluating results from different parameter settings to improve data quality and detection coverage as data extraction using only one setting could give sub-optimal results. The merged results are then used for additional quality assessment and correction such as evaluation of internal standards and reference metabolites, mass calibration and batch-effect correction in step 4. Step 5 involves *m/z* based annotation of features using HMDB, KEGG, T3DB, and LipidMaps.

Figure 4. Correlation-based network analysis to identify related ions and metabolites. Data-driven network analysis can be used to identify modules/clusters of strongly associated ions. Some of these associations are a consequence of analytical correlations, such as multiple adducts formed from a single chemical, while other associations are a consequence of biological relationships. In the example shown here for the anesthetic ketamine, each sub-cluster shows strong associations between the primary form, adducts, isotopes, and ionization fragments derived from the same metabolite. Secondary correlations exist between biologically related metabolites, ketamine and its metabolites, norketamine and hydroxyketamine. Data from the studies of Jones et al. and Uppal *et al*.[19, 64]

Figure 5. Metabolome-wide association study for metabolite identification. Choline correlation in different species illustrates preservation of metabolic association structures; supporting metabolite annotation[64]. The network structures for humans and the common marmoset contain metabolites exhibiting similarly significant correlations with choline. Like correlations of adducts formed from a chemical during ionization, the existence of

network correlations of metabolites in biological systems provides a parameter for establishing confidence in identification, even for low abundance metabolites without quality MS/MS spectra. Figure is reproduced from Uppal et al. [64]

Figure 6. Computational identity prediction. A) Distribution of metabolic features in a human dataset with or without database matches in HMDB using common adduct forms showed that more than half of the ions reproducibly detected in human plasma did not have matches to known metabolites in HMDB. B) Evaluation of results for medium-to-high confidence matches from a healthy human dataset using a clustering approach based on correlation between ions across all samples, retention time, mass defect, adducts and isotopes pattern using MS/MS showed that 80% of matches are correct. Thus, methods are improving for identification of high abundance metabolites, with moderate to high confidence annotation for over 2000 chemical species. Despite the ability to characterize such a large number of metabolites, a much larger number of ions are without matches in databases, creating a major challenge for biological interpretation. Methods are needed to provide unambiguous designation of these ions to facilitate identification, especially for unidentified ions linked to human disease (e.g. see Table 1).

Figure 7. Ion definition in multi-vector space. Assembling the million metabolome will require an unambiguous system for defining detected, unidentified ions. In this framework, experimental measures including high-accuracy mass-to-charge ratio (*m/z)*, retention time relative to landmark chemicals, correlation structure, ion dissociation spectra, collision cross section (CCS) from ion mobility spectrometry and enantiomer

selective detection are combined to uniquely position an ion in chemical space. This is arbitrarily visualized here in terms of three-dimensional plots; expression could be made in terms of six or more one-dimensional vectors from a common origin. In this figure, three dimensions are designated in a way that leverages the capabilities of currently available analytical and computational approaches while enabling incorporation of future advances. The dimensions of Plot 1 on the left includes untargeted profiling on high-resolution, accurate mass (HRAM) mass spectrometers coupled with chromatographic separation prior to detection. The use of landmark chemicals provides retention time indices for relative elution and metabolic correlation structure, which is anchored against the accurate *m/z*. Plot 2 in the middle is largely defined by structural characteristics of the molecule, which are designated by ion dissociation of precursor *m/z* from Plot 1 and CCS. Plot 3 on the right is defined by relative quantification of enantiomers. Several chiral methods are available but will require development for automated use in ion characterization.

Figure 8. Separation of isobaric environmental chemicals by ion mobility spectroscopy (IMS)-mass spectroscopy. Organic aerosol species constitute a major fraction of airborne particles contributing to air pollution and adversely impacting health of humans and other species. The complex mixtures of organic aerosol species are difficult to resolve measure by conventional analytical methods, and little information is available concerning the levels or distribution of these chemicals in humans and other mammalian species. This figure from a recent application of IMS-MS to samples from the Southern Oxidant and Aerosol Study (SOAS) shows the utility of IMS-MS for this

challenging environmental issue. IMS-MS was performed for hydroxysulfate esters (HSE; $C_5H_{11}O_7S^-$) of isoprene epoxydiols (IEPOX) in four different aerosol filter samples. Dashed vertical lines designate signals for three different IMS peaks of isoprene epoxydiols (IEPOX) after conversion to respective hydroxysulfate esters. Different stereoisomers of IEPOX are formed by radical reactions from isoprene hydroxyhydroperoxide intermediates. The stereoisomers are sufficiently resolved to allow discrimination of the different species. The bars on the top denote the uncertainty in the drift time dimension for each peak and were determined from the standard error of the mean of a mobility calibration compound from its average drift time. Additional details are provided in the original publication (Figure 4) by Krechmer *et al*. [121] This figure was reproduced with permission granted by the original authors and Creative Commons Attribution 3.0 License.

Figure 9. Developmental need exists for enantiomer-selective designation. Many environmental chemicals exist as stereoisomers and this presents a challenge for chromatography and detection methods which do not resolve stereoisomers. Analytical data for S- and R-enantiomers of L-methionine sulfoxide illustrate the need for enantiomer-selective designation of ions. **A)** Anion exchange (AE) chromatography was unable to separate enantiomers of L-methionine-sulfoxide prior to detection, resulting in one peak representing the sum of the two enantiomers. Use of a chiral column that resulted in specific R- and S- interactions with the two enantiomers separate L-methionine(S)sulfoxide from L-methionine(R)sulfoxide, enabling quantification of each. **B-C)** Ion dissociation (MS^2) of the two enantiomers showed identical fragmentation

patterns and are indistinguishable when defined by accurate mass, retention time and MS^2 spectra. Thus, there is a need to develop methods to enable enantiomer-specific designation for ions in the million metabolome. Available analytical methods include chiral selectors, ion mobility with chiral gases and chromatographic separation using enantiomer specific retention mechanisms.

Figure 1

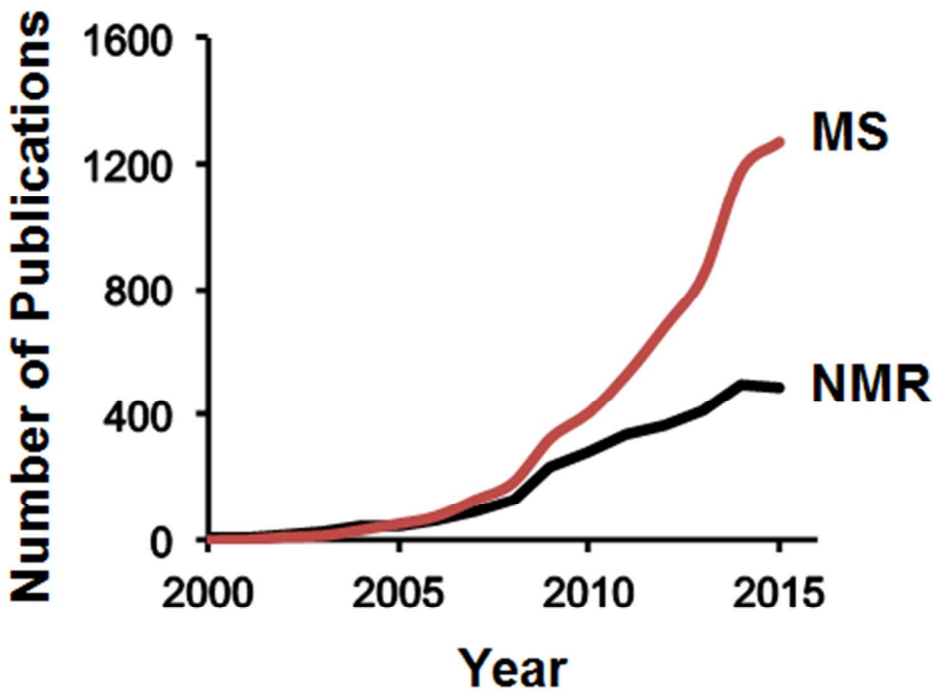

Number of Publications
1600
1200
800
400
0
2000
2005
2010
2015
Year
MS
NMR

Figure 2

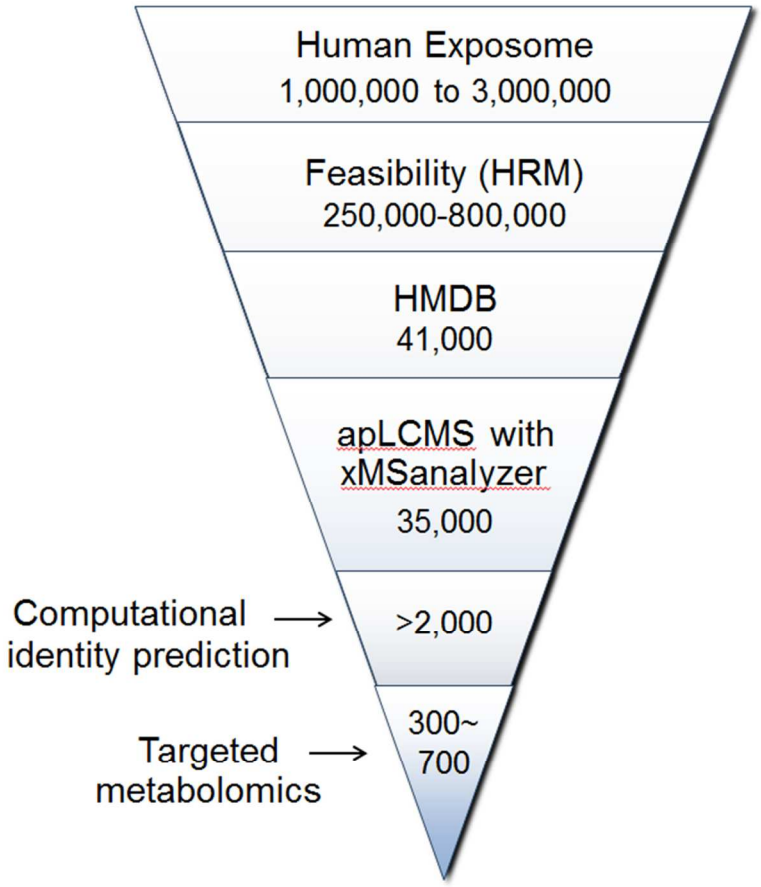

Human Exposome
1,000,000 to 3,000,000
Feasibility (HRM)
250,000-800,000
HMDB
41,000
apLCMS with
xMSanalyzer
35,000
Computational
identity prediction
>2,000
Targeted
metabolomics
300~
700

Figure 3

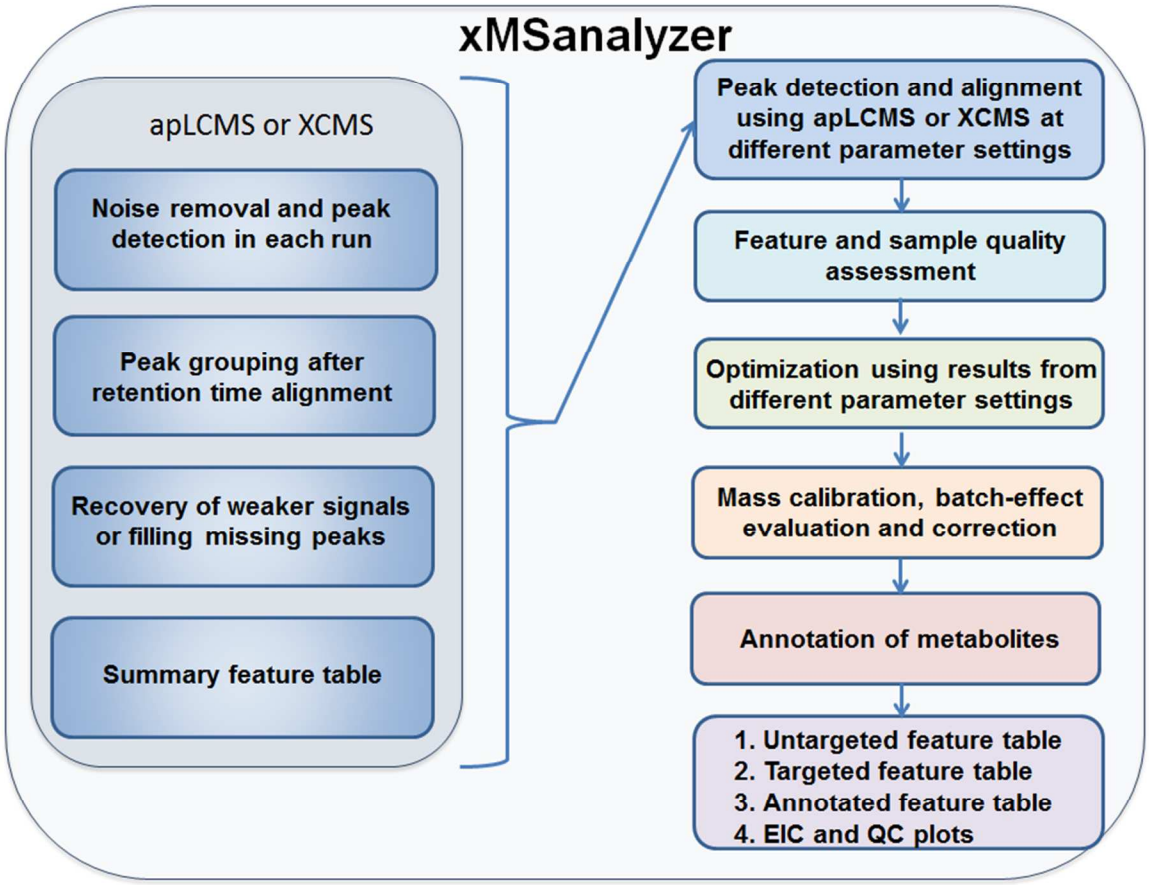

xMSanalyzer
apLCMS or XCMS
Noise removal and peak detection in each run
Peak grouping after retention time alignment
Recovery of weaker signals or filling missing peaks
Summary feature table
Peak detection and alignment using apLCMS or XCMS at different parameter settings
Feature and sample quality assessment
Optimization using results from different parameter settings
Mass calibration, batch-effect evaluation and correction
Annotation of metabolites
1. Untargeted feature table
2. Targeted feature table
3. Annotated feature table
4. EIC and QC plots

Figure 4

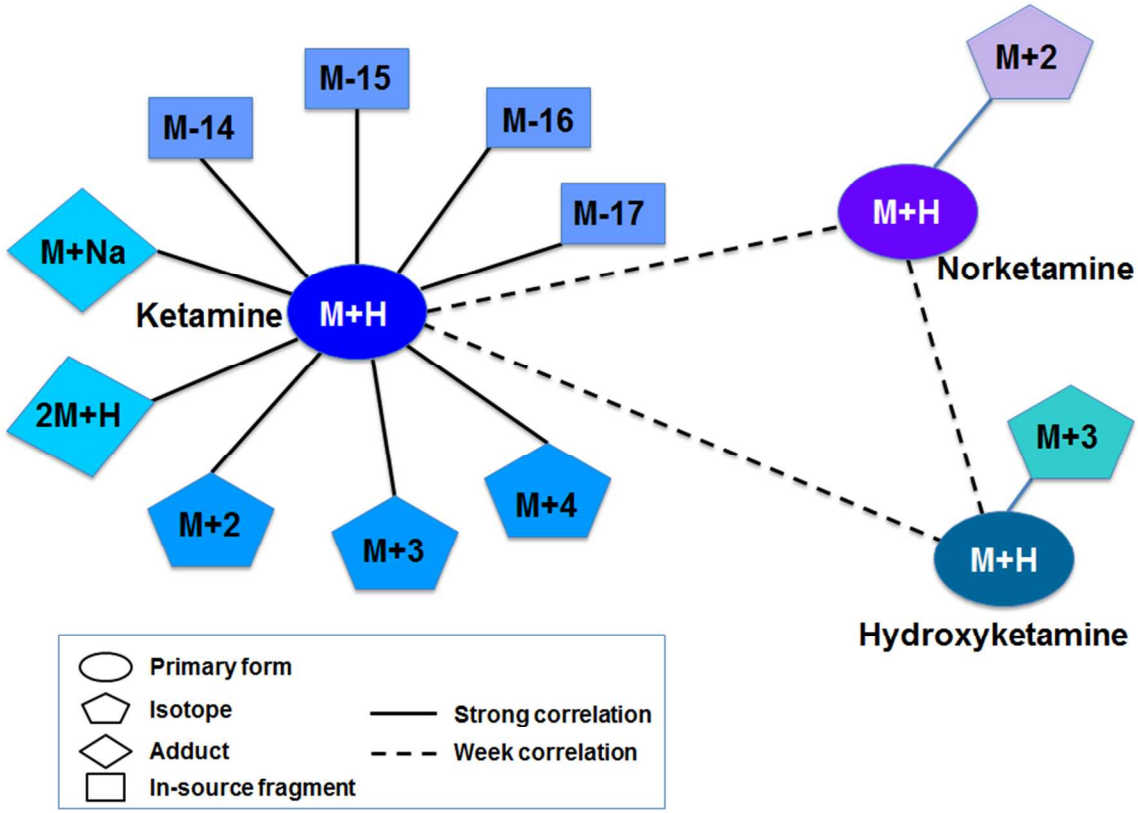

M-15
M-14
M-16
M-17
M+Na
Ketamine
M+H
2M+H
M+2
M+3
M+4
M+2
M+H
Norketamine
M+3
M+H
Hydroxyketamine
Primary form
Isotope
Adduct
In-source fragment
Strong correlation
Week correlation

Figure 5

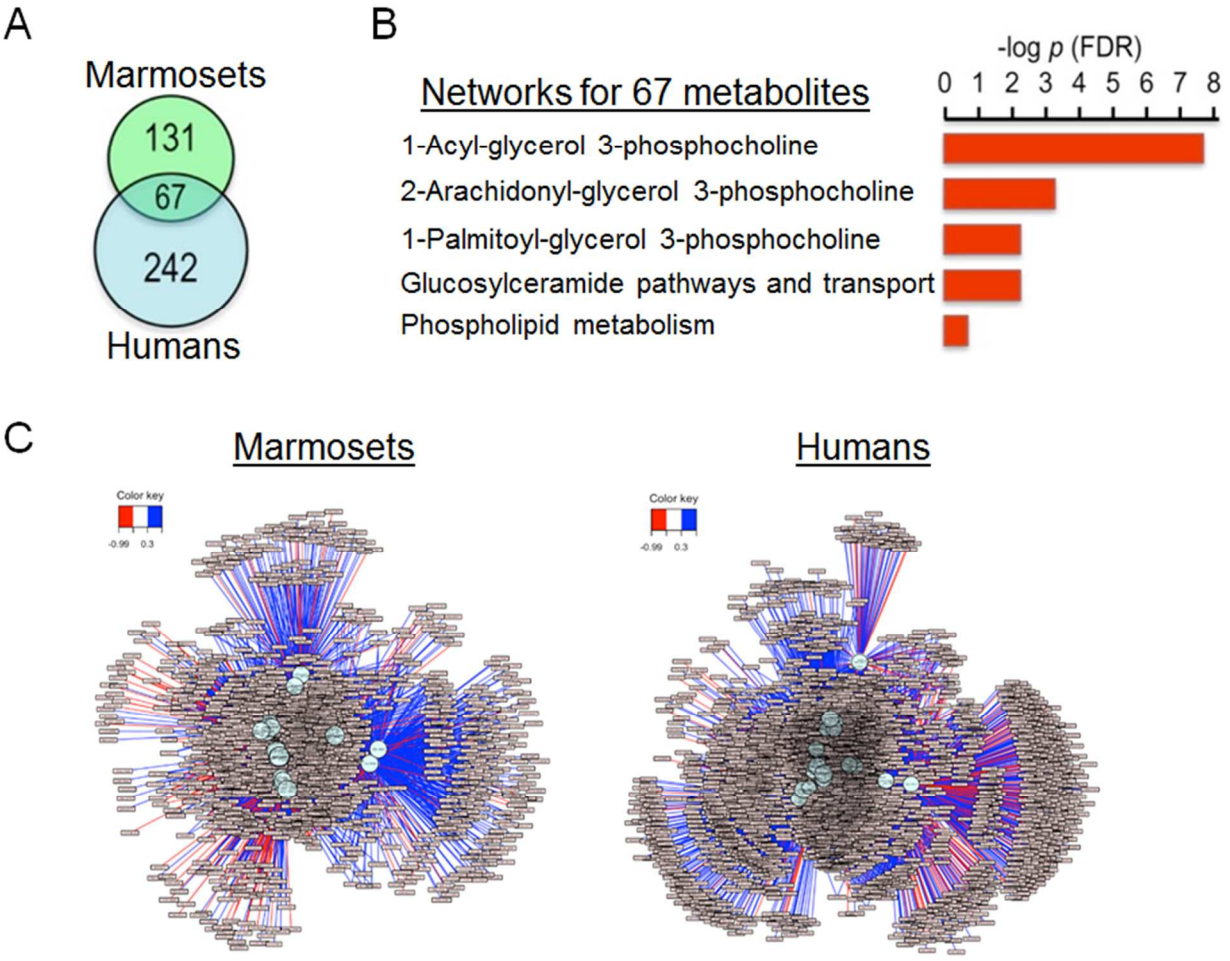

A
Marmosets
131
67
242
Humans
B
Networks for 67 metabolites
-log p (FDR)
0 1 2 3 4 5 6 7 8
1-Acyl-glycerol 3-phosphocholine
2-Arachidonyl-glycerol 3-phosphocholine
1-Palmitoyl-glycerol 3-phosphocholine
Glucosylceramide pathways and transport
Phospholipid metabolism
C
Marmosets
Humans
Color key
-0.99 0.3
Color key
-0.99 0.3

Figure 6

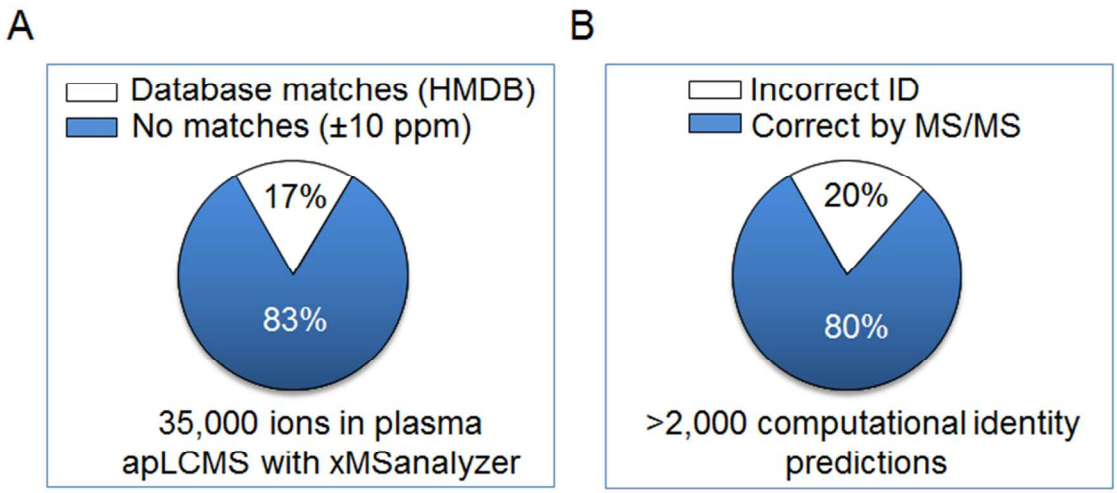

A
Database matches (HMDB)
No matches (±10 ppm)
17%
83%
35,000 ions in plasma
apLCMS with xMSanalyzer
B
Incorrect ID
Correct by MS/MS
20%
80%
>2,000 computational identity
predictions

Figure 7

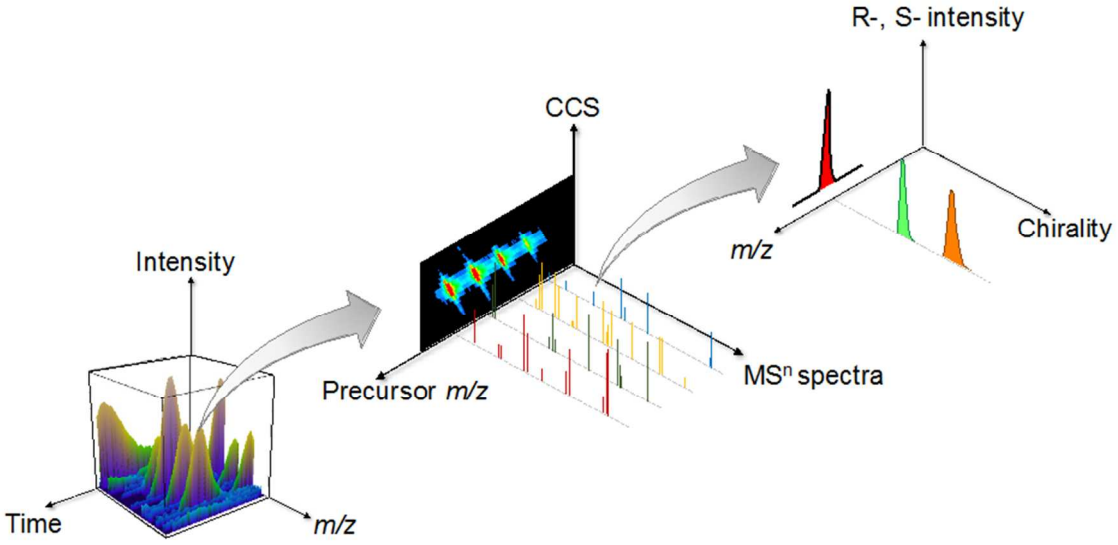

R-, S- intensity
CCS
Chirality
m/z
Intensity
MSⁿ spectra
Precursor m/z
Time
m/z

Figure 8

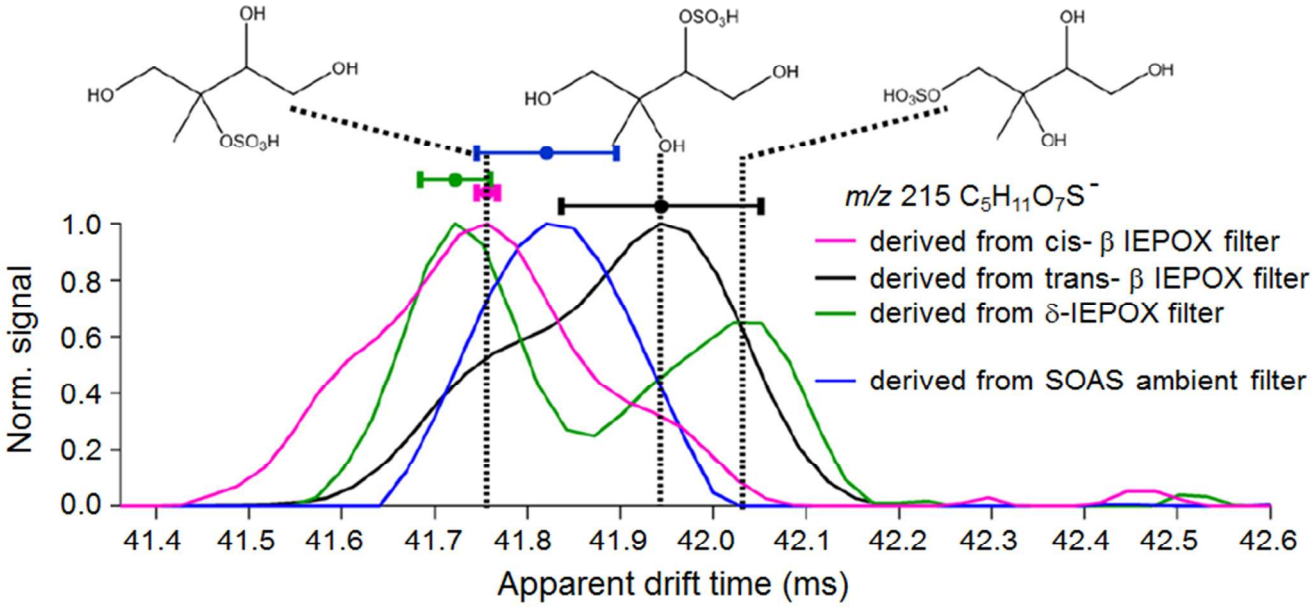
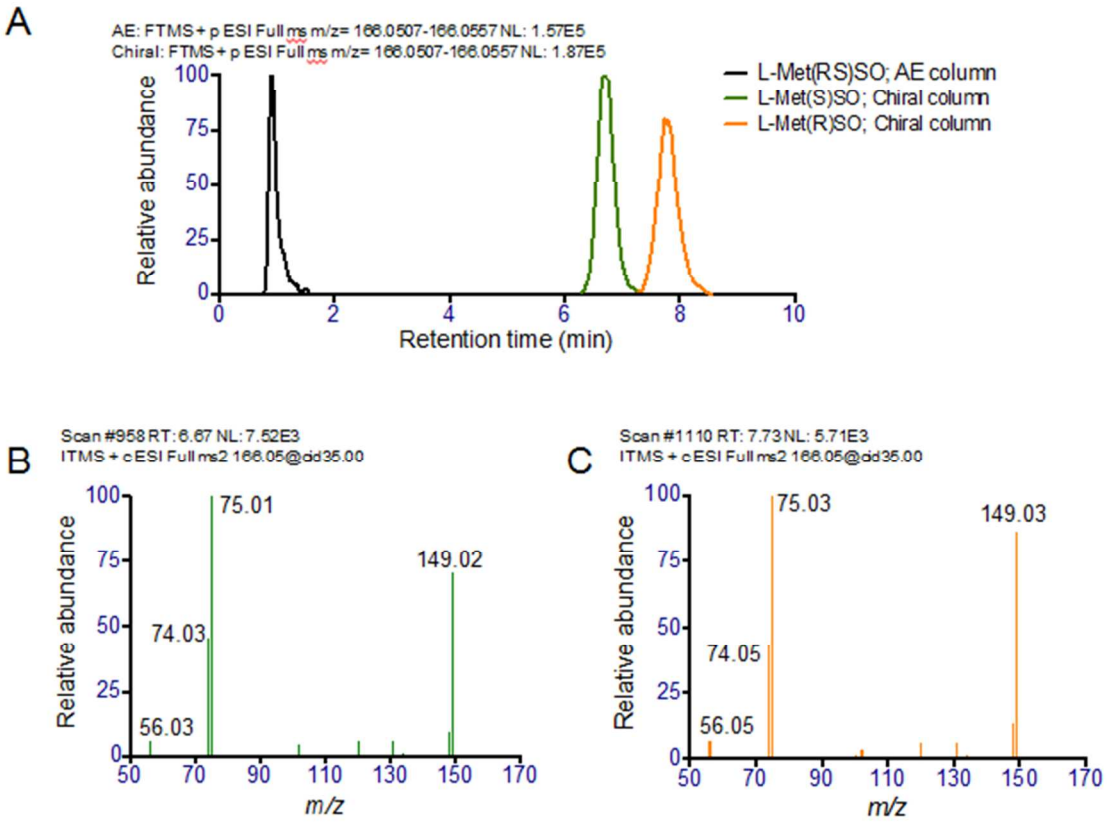

OH
HO
OH
OSO3H
OSO3H
HO
OH
OH
HO3SO
OH
OH
m/z 215 C5H11O7S-
derived from cis- β IEPOX filter
derived from trans- β IEPOX filter
derived from δ-IEPOX filter
derived from SOAS ambient filter
Norm. signal
1.0
0.8
0.6
0.4
0.2
0.0
41.4
41.5
41.6
41.7
41.8
41.9
42.0
42.1
42.2
42.3
42.4
42.5
42.6
Apparent drift time (ms)

Figure 9


A
AE: FTMS + p ESI Full ms m/z= 166.0507-166.0557 NL: 1.57E5
Chiral: FTMS + p ESI Full ms m/z= 166.0507-166.0557 NL: 1.87E5
L-Met(RS)SO; AE column
L-Met(S)SO; Chiral column
L-Met(R)SO; Chiral column
Relative abundance
Retention time (min)
B
Scan #958 RT: 6.67 NL: 7.52E3
ITMS + c ESI Full ms2 166.05@cid35.00
75.01
149.02
74.03
56.03
C
Scan #1110 RT: 7.73 NL: 5.71E3
ITMS + c ESI Full ms2 166.05@cid35.00
75.03
149.03
74.05
56.05
m/z

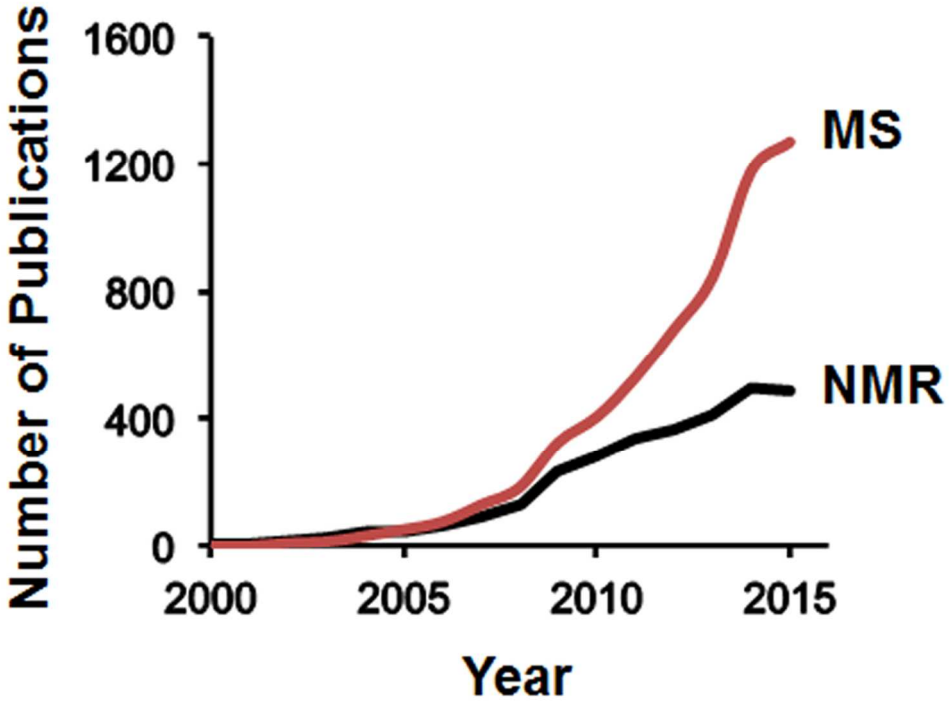


Figure 1

97x74mm (300 x 300 DPI)

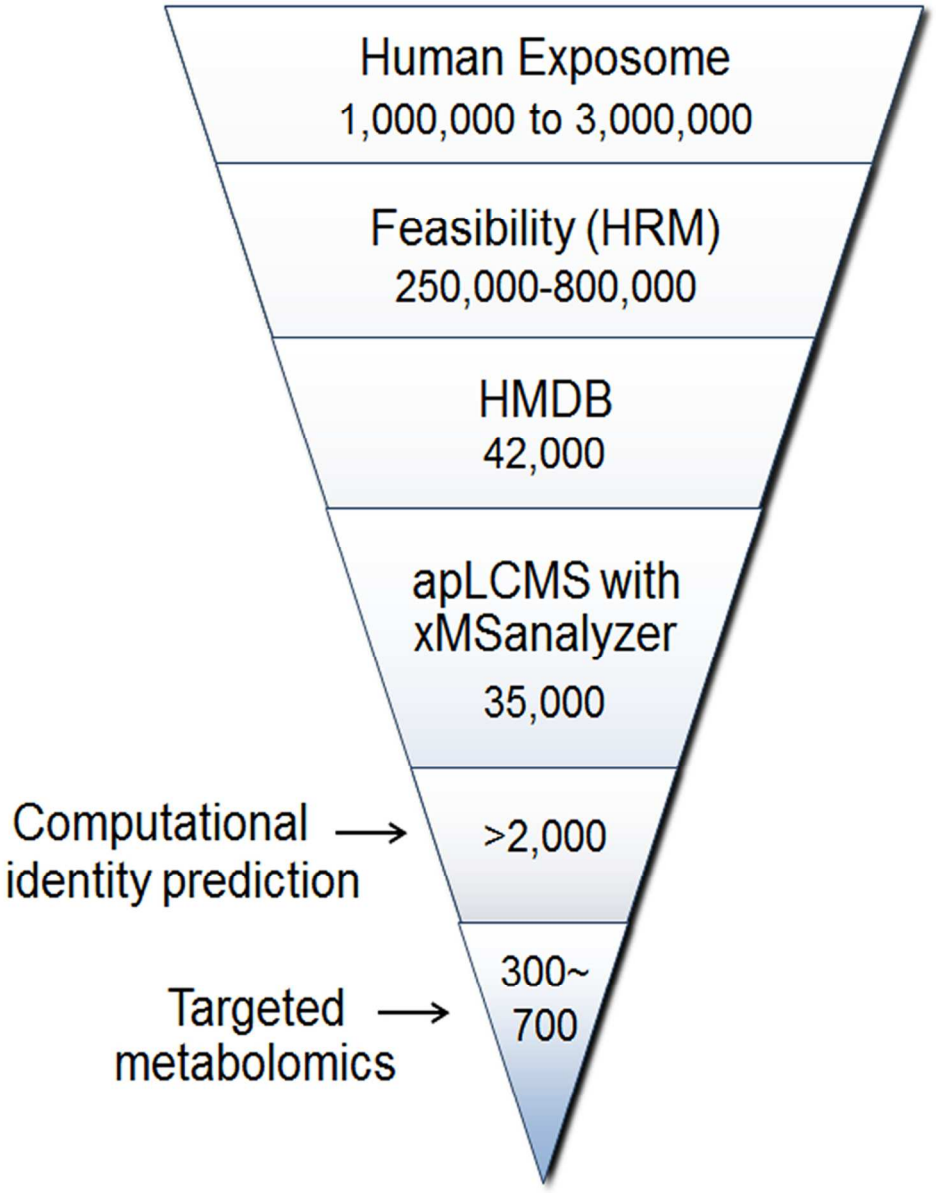


Figure 2

84x105mm (300 x 300 DPI)

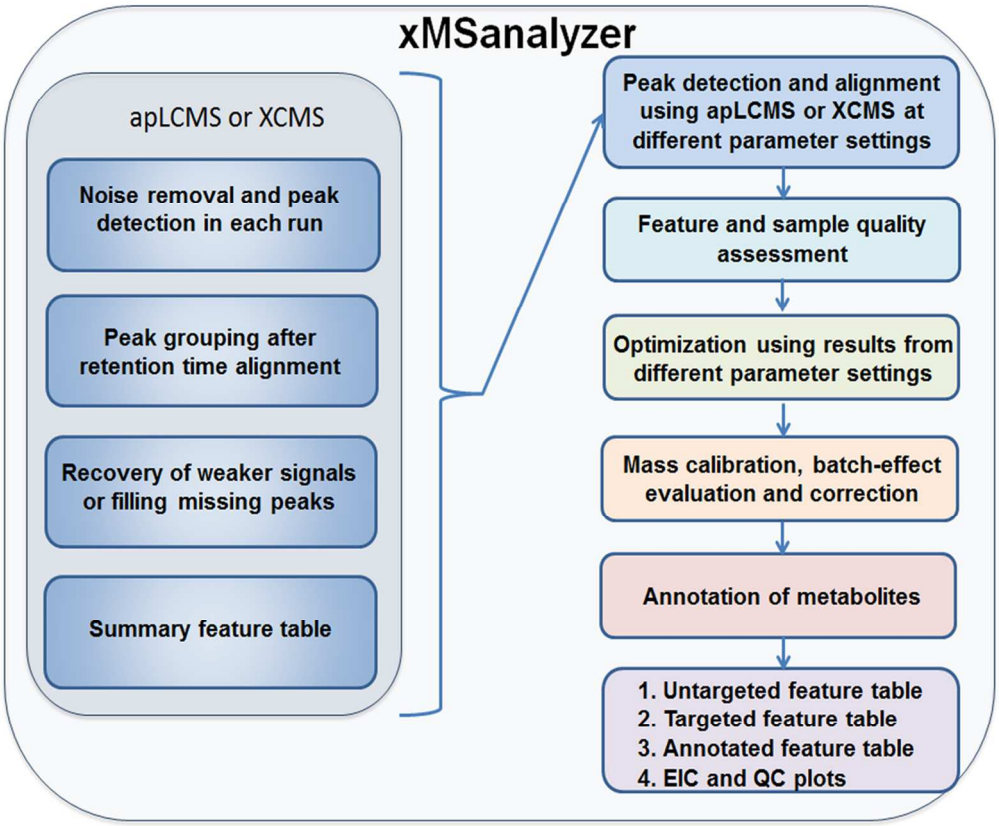


Figure 3

105x88mm (300 x 300 DPI)

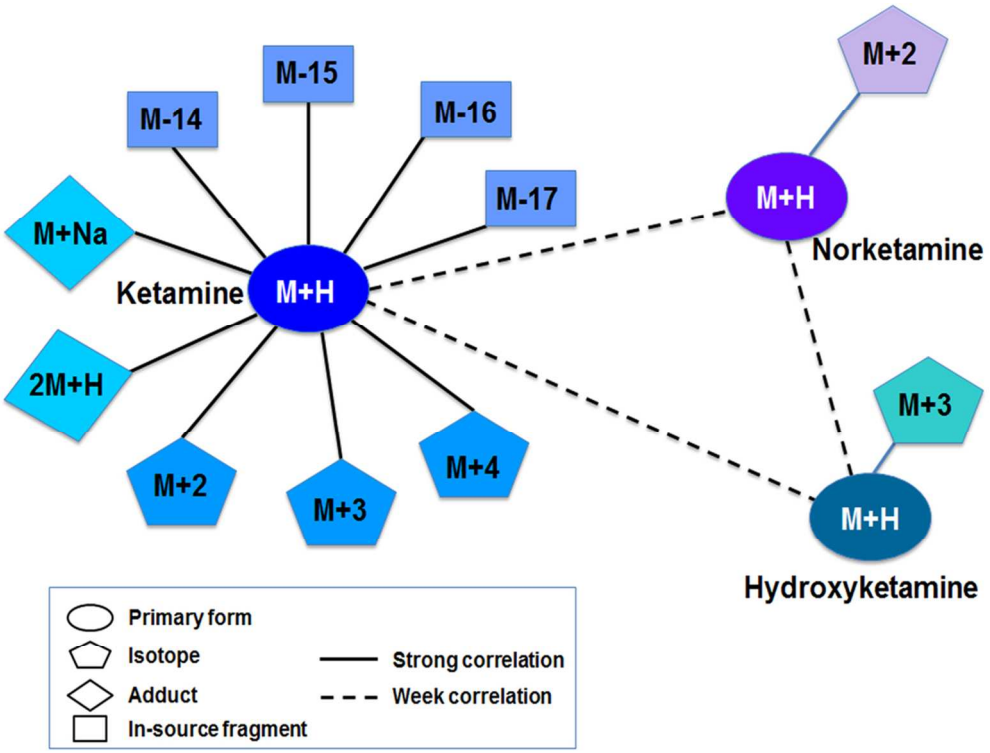


Figure 4

95x71mm (300 x 300 DPI)

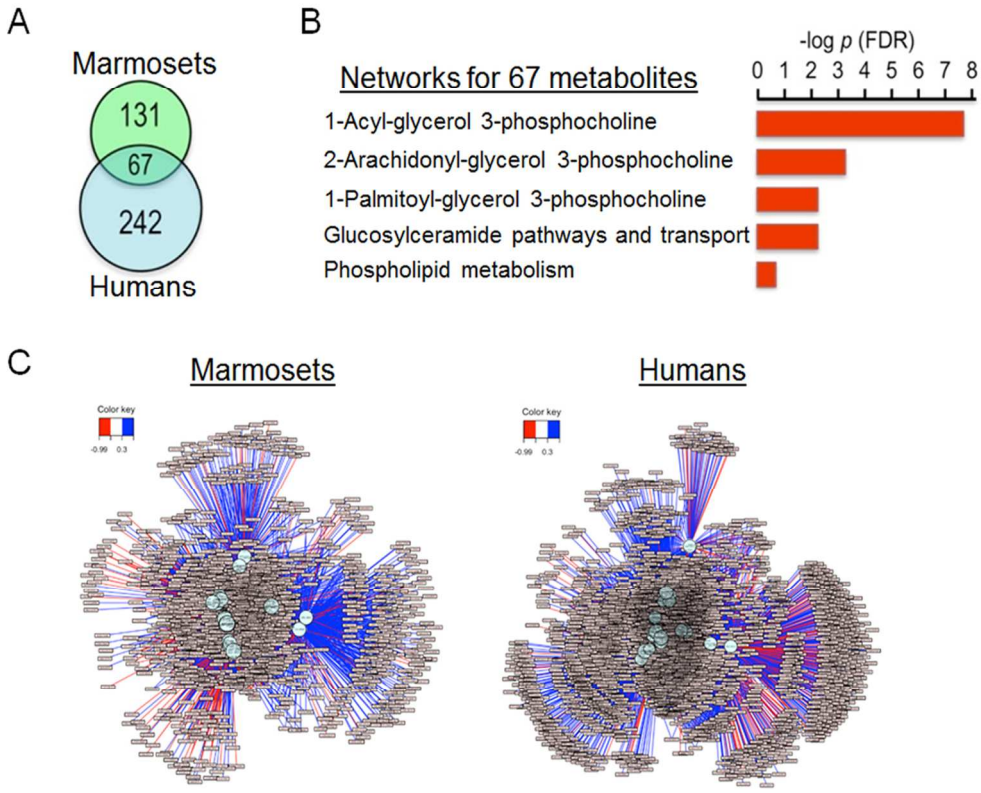


Figure 5

101x81mm (300 x 300 DPI)

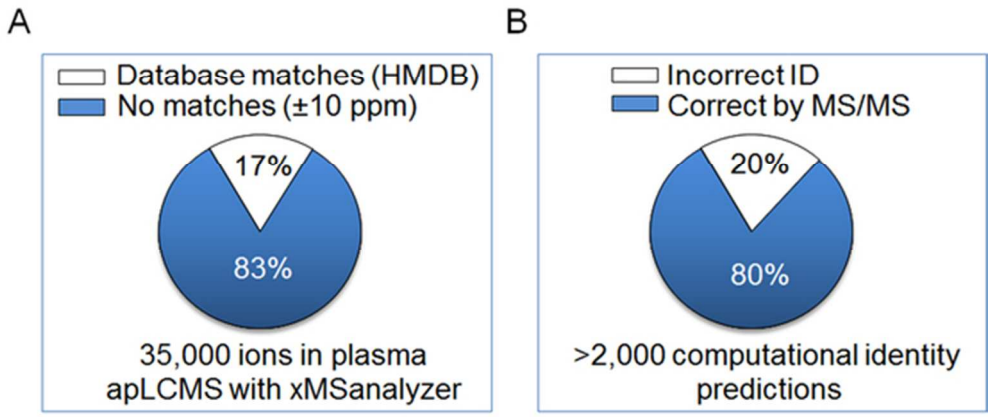
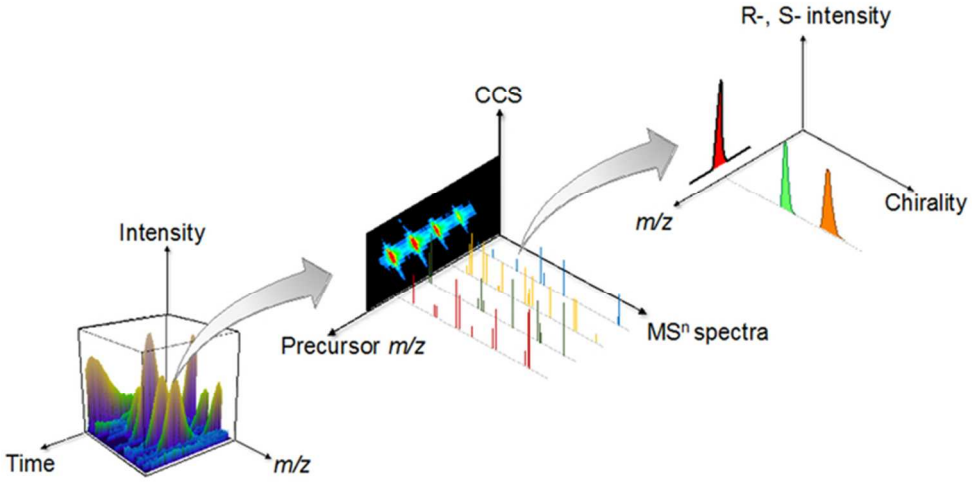
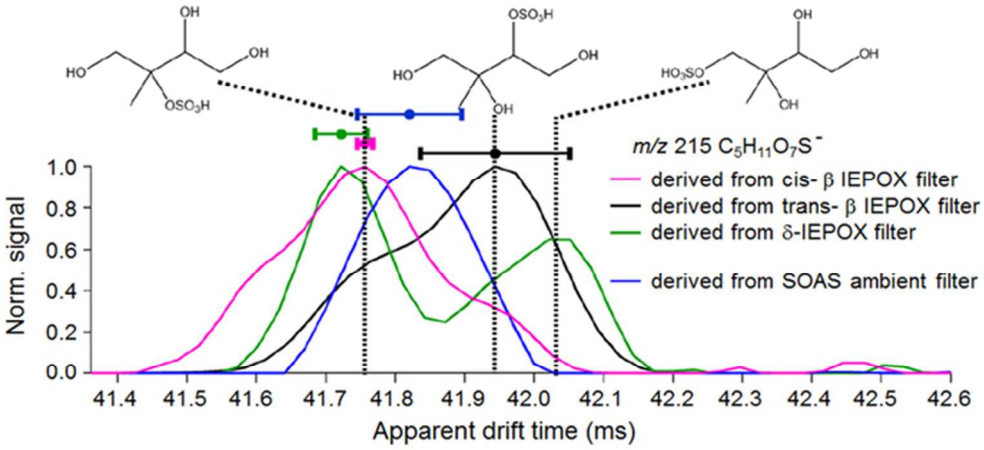


Figure 6

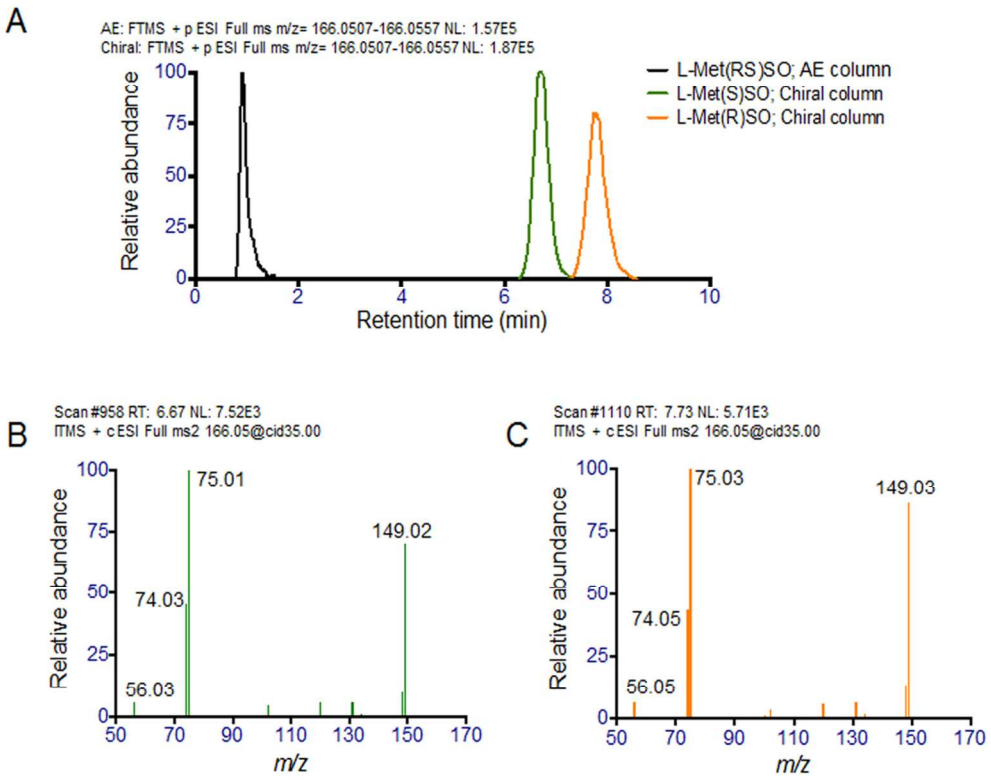
55x23mm (300 x 300 DPI)



63x31mm (300 x 300 DPI)



59x27mm (300 x 300 DPI)



101x81mm (300 x 300 DPI)