Dark matter in host-microbiome metabolomics: Tackling the unknowns—A review

B.Y. Loulou Peisl [a, b], Emma L. Schymanski [a, *], Paul Wilmes [b, *]

[a] Environmental Cheminformatics Group, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 7, Avenue des Hauts Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg
[b] Eco-Systems Biology Group, LCSB, University of Luxembourg, 7, Avenue des Hauts Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg

GRAPHICAL ABSTRACT

ARTICLE INFO

ABSTRACT

The "dark matter" in metabolomics (unknowns) represents an exciting frontier with significant potential for discovery in relation to biochemistry, yet it also presents one of the largest challenges to overcome. This focussed review takes a close look at the current state-of-the-art and future challenges in tackling the unknowns with specific focus on the human gut microbiome and host-microbe interactions. Metabolomics, like metabolism itself, is a very dynamic discipline, with many workflows and methods under development, both in terms of chemical analysis and post-analysis data processing. Here, we look at developments in the mutli-omic analyses and the use of mass spectrometry to investigate the exchange of metabolites between the host and the microbiome as well as the environment within the microbiome. A case study using HuMiX, a microfluidics-based human-microbial co-culture system that enables the co-culture of human and microbial cells under controlled conditions, is used to highlight opportunities and current limitations. Common definitions, approaches, databases and elucidation techniques from both the environmental and metabolomics fields are covered, with perspectives on how to merge these, as the boundaries blur between the fields. While reflecting on the number of unknowns remaining to be conquered in typical complex samples measured with mass spectrometry (often orders

of magnitude above the "knowns"), we provide an outlook on future perspectives and challenges in elucidating the relevant "dark matter".

1. Introduction

1.1. The human gut microbiome

Complex assemblages of microorganisms populate the human body and these microbiomes are emerging as key players in human health and disease [1]. The largest reservoir of microbial biomass is the gastrointestinal tract. The human gut microbiome is considered a central hub that integrates environmental inputs, such as diet and surroundings, with genetic and immune signals to affect the host's overall physiology, including metabolism [2]. The gut microbiome confers essential metabolic and other functions to human physiology including digestion of food components [3], synthesis of essential vitamins [4], stimulation and regulation of the immune system [5], out-competition of pathogens [6], removal of toxins and carcinogens [7], and support of intestinal function [8]. Many of these functions are interconnected as the gut microbiome contributes to overall human metabolism [9,10] and the microbial metabolites produced play essential roles in immunomodulatory processes [11]. In the context of the human immune system, there is a tight interconnection whereby the immune system may affect the gut microbiome and its metabolic capacity, and *vice versa* [12]. The gut microbiome also interfaces with other body systems via the circulatory, immune, endocrine and nervous systems. Changes to the microbial ecology of the gut may culminate in dysbiosis, a pathological imbalance in the gut microbiota with implied dysfunctions in the complex set of processes governing human health [13].

The advent of high-throughput sequencing and its application to the complex microbiota of the human gut has provided essential new insights into the structural diversity and functional potential of the gut microbiome. Essential attributes of the human gut microbiome uncovered through these studies include extensive genetic diversity [14,15], distinct community types [16,17], apparent functional stability despite variation in community structure [18], the influence of host genetics in shaping community structure [19], inter-individual variability [20] and apparent intra-individual stability [21,22], and the overall importance of extrinsic and intrinsic host factors in shaping community composition [23]. Furthermore, through the use of sequencing-based methods, largely applied in case-control study designs, dysbiosis has been implicated in the aetiology of numerous idiopathic conditions including irritable bowel syndrome [24], inflammatory bowel disease [25], liver cirrhosis [26], type 1 diabetes [27,28], type 2 diabetes [29], obesity [30,31], cardiovascular disease [32], colorectal cancer [33,34], rheumatoid arthritis [35] and most recently Parkinson's disease [36—38].

Metagenomic analysis involving random shotgun sequencing of community genomic DNA has revealed the genetic potential of the gut microbiota, especially in relation to metabolic transformations and disease [39]. Additionally, metatranscriptomic and meta-proteomic analyses have identified which genes are expressed by the microbiota under specific conditions [40]. However, as community-wide metabolism reflects the actual, cumulative phenotypes of the different populations which comprise the microbiome, (meta-)metabolomics is likely the most sensitive indicator for disease-linked processes. This in turn makes it well suited for identifying discriminant features based on which mechanistic hypotheses can be formulated linking, for example, dysbiotic microbiota to disease pathogenesis. Therefore, differences in microbial community structure reflective of dysbiosis have been linked to changes in microbial metabolism in the gut, e.g. alteration of microbial phosphatidylcholine metabolism in the context of atherosclerosis [32] or increased biosynthesis of branched chain amino acids in the context of insulin resistance [41]. Metabolic activity within the gut microbiome also impacts drug metabolism and efficacy [42,43]. A detailed knowledge of gut microbiome-mediated metabolic transformations is therefore essential to understand how the gut microbiome impacts human phenotypes, especially in relation to the panoply of diseases now associated with changes in the gut microbiome. In this context, metabolomic analyses of gut microbiome small molecule extracts have allowed the identification of disease-specific signatures (recently reviewed in Ref. [44]).

1.2. Metabolites and the microbiome

Given the pronounced metabolic activity of the gut microbiome, which includes multiple unique catabolic and anabolic reactions not catalyzed by human cells, microbial metabolism has to be considered as an integral part of human physiology. In general terms, the major known metabolite classes produced and transformed by gut microbiota with known effects on human physiology include organic acids (lactate, succinate, formate, etc.), short chain fatty acids (acetate, propionate, butyrate, etc.), lipids (ceramides, lysophosphatidylcholines, phosphatidylcholines, etc.), branched-chain fatty acids (valerate, isobutyrate, isovalerate, etc.), branched-chain amino acids (leucine, isoleucine, valine), vitamins (biotin, folate, niacin, etc.), bile acids (deoxycholic acid, lithocholic acid, etc.), and neurotransmitters (GABA, serotonin, etc.). Apart from these specific metabolite classes, the gut microbiome catalyzes a broad spectrum of different biotransformations (Fig. 1). Importantly, apart from the known microbiome-driven metabolic reactions, microbiome-derived metabolomic datasets are characterized by a significant fraction (>90%) of as yet uncharacterized metabolite features that do not have any match in public databases [45]. Many of these "unknowns" are highly likely to represent "missing links" in microbial metabolism and human-microbe molecular interactions [46]. The systematic study of the microbiome-conferred metabolome therefore requires extensive future study, not least because a detailed understanding of the functional microbiome represents an essential prerequisite for future rational interventions leveraging the gut microbiome to alter host phenotype. In this context, the unknowns represent an important focus for investigation.

1.3. Mass spectrometry: from metabolites to the environment

Mass spectrometry (MS) is often the analytical method of choice for discovery-based untargeted metabolomics analyses. Although a lot of excellent metabolomics is performed with nuclear magnetic resonance (NMR) techniques, it is not the focus of this article. The post-analysis identification workflows of MS-based approaches depend highly on the analytical set-up used, with many different databases and software tools now available. Separation techniques,
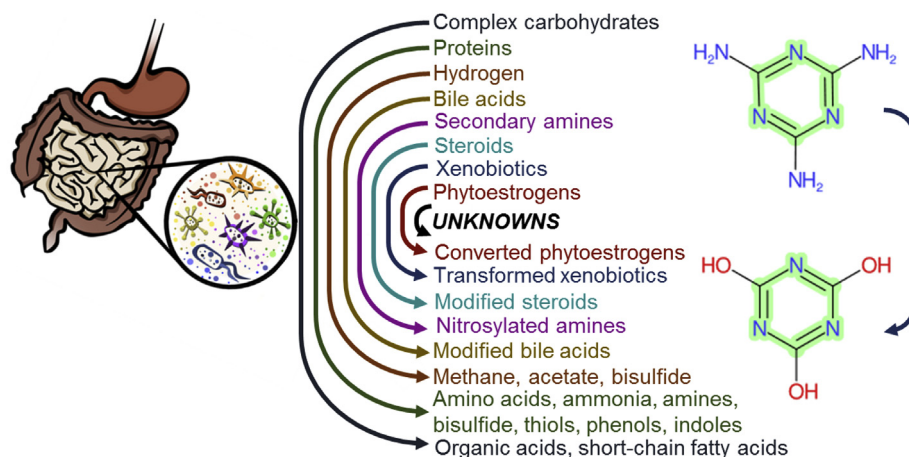
Fig. 1. Microbiome-driven biotransformations in the gut. The microbiome-mediated transformation of the xenobiotic melamine into cyanuric acid is shown as an example (from Ref. [7]).

most commonly gas chromatography (GC) and liquid chromatography (LC) are often, but not always, used upstream of the MS to detect more or less polar substances, respectively. Traditionally, GC-MS is coupled to electron ionization (EI; GC-EI-MS), fixed at an energy of 70 eV, which is a hard ionization technique that typically yields reproducible, fragment-rich spectra with unit mass accuracy. Comprehensive libraries of several hundred thousand GC-EI-MS spectra now exist (more details below) and thus "unknown identification" workflows have evolved around spectral database lookup. Due to the high fragmentation energy, the molecular ion is often either absent or of low abundance (estimates indicate this may occur in up to 40% of cases) [47] and this makes further elucidation challenging. While Computer-Assisted Structure Elucidation (CASE) by GC-EI-MS is possible, it is not yet suitable for routine application [48] and, thus, further elucidation efforts generally require manual interpretation. To expand the chemical coverage, metabolomics samples undergoing GC-MS are often derivatized to increase the volatility of substances containing polar functional groups, which creates additional challenges for structure elucidation of those substances that do not have any database match. Techniques that couple GC with higher resolution to obtain the exact mass of the parent ion and/or fragments are becoming more common; the data processing workflows for this analysis are thus more similar to those for LC-HR-MS described below.

For LC, the most common ionization techniques are electrospray ionization (ESI), atmospheric pressure chemical ionization (APCI) and atmospheric pressure photon-ionization (APPI) [49], which yield the exact mass of the parent ion(s) at high resolution. Fragment information is generated in a second stage (MS2), with ions in the first stage (MS1) isolated either according to a narrow mass-to-charge ratio (m/z) — data-dependent acquisition — or with a broad range (data-independent acquisition). Time of Flight (TOF) and Orbitraps are now commonly used for untargeted analysis [50], where the exact options for data acquisition depend highly on the instrument and vendor and are beyond the scope of the present review. The final result is a two-stage spectrum, where the MS1 typically provides information about the molecular ion and other adducts (precursor or parent information — which, for unknowns, must be grouped together to form "features" that correspond to one molecule) and the MS2 typically provides fragmentation information. Both stages then contribute to the information available for identification, such as the molecular formula determination and subsequent search for possible matching candidates (structures). The lack of standardization and reproducibility coupled to the lack

of substance coverage within spectral libraries [51,52] means that unknown identification workflows based on high-resolution MS (HR-MS) have evolved to be less spectral library dependent than GC-EI-MS (Fig. 5).

MS1 and MS2 information is used for elucidation of the unknown metabolites or small molecules in HR-MS experiments corresponding to the detected features (grouped MS1 signals). The terminology used to describe the concepts involved in the identification of unknowns is as diverse as the many approaches used across disciplines. Many terms depend on the scientific field and the study question. In metabolomics experiments, two main approaches are usually described, targeted and untargeted analysis. The terms metabolite and metabolite features are commonly used to describe what is detected by the mass spectrometer. The targeted method, also referred to as metabolic profiling, is in principle analogous to target screening. Here, one identifies and quantifies a list of known target substances (metabolites, small molecules) according to their mass spectra and retention times/indices [53]. In the untargeted or non-targeted approach, also described as metabolic fingerprinting, all detected metabolite features are considered during data analysis without necessarily any prior knowledge about their identity [53]. The identification of features detected with untargeted methods is also referred to as dereplication or annotation. While metabolites were traditionally considered to be of natural origin (originating from metabolic processes in the organism, where often only the elements C, H, N, O, P and S were considered relevant), the extension of metabolomic techniques into fields such as personalized medicine and exposomics has expanded the window of "metabolites" beyond the traditional definition to also include anthropogenic compounds (e.g. drugs, pharmaceuticals and their metabolites), i.e. expanding into the realm of small molecules in general.

1.3.1. Expanding the window of metabolomics

As the human gut is one of the main interfaces between the external environment (e.g. food, chemical exposure) and the body, studying the microbiome goes beyond the known metabolic networks and involves the "environment within". In the environmental sciences, all "small molecules" often need to be considered by default, although the focus is often on those of anthropogenic origin. As a result, the substance classes and properties considered are often far wider than those that would be under study in traditional metabolomics studies (although, as stated above, the fields are now moving together). All elements must be considered

and in particular popular approaches to restrict the domain to "expected" features, such as outlined by the "Seven Golden Rules" [54], may no longer apply as some environmental contaminants are deliberately created to have unique properties (e.g. poly-halogenated substances). Identification of features in the environmental chemistry literature is often split into three main approaches, i.e. target, suspect and non-target screening ([50,55] and Fig. 5). For target screening, a reference standard for each "target" or known compound should be available in-house and measured with the exact same settings as the samples to enable confirmation of the identity. At least two orthogonal pieces of analytical information should be used for confirmation, such as retention behavior on the chromatographic system as well as the mass and fragmentation information from the mass spectrometer. Often, full quantification is also performed, depending on the sample context. This can either be done separately (e.g. with a targeted analysis), combined (target screening and quantification with untargeted data acquisition, also termed post-target screening [50,56]) or in coupled systems [57]. Following target screening, HR-MS data allows suspect screening, which can be used to search for compounds that may be *expected* in the sample but are not available as reference standards in-house at the start of the experiment (reference standards are needed eventually for full confirmation of the suspect identity). Initiatives such as the NORMAN Suspect Exchange [58] and the chemical lists available in the US EPA CompTox Chemistry Dashboard [59] have enabled access to various "suspect lists" relevant to environmental investigations including surfactants, pesticides, pharmaceuticals and perfluorinated substances. As the identity of the compound is "suspected", the molecular formula is known and can be used to calculate the exact mass of the expected ion and the isotope pattern. This and other evidence such as adduct information and, where possible fragment data, can be used for a tentative identification. Exact confirmation still requires a reference standard, such that the aim of untargeted analysis would be to "upgrade" suspects to target substances, to allow more sensitive analysis and quantitation in future investigations. A parallel to metabolomics would be the use of metabolic networks to direct the identification of untargeted metabolites. Target and suspect screening are finally complemented by non-target screening. This applies to the remaining masses that are detected in samples, where neither prior information about the potential presence, nor reference standards are available. These masses must be grouped into "features", which require full elucidation. As for suspect screening, relevant non-targets, where possible, should become target compounds or "knowns" to allow easier detection in subsequent investigations.

### 1.4. Shining light on dark matter through mass spectrometry

While metabolism is highly dynamic and can potentially reveal information in realtime and high-throughput [60], metabolite identification and the biological interpretation still remain major bottlenecks in metabolomics [61]. This has resulted in the term "dark matter", i.e. what remains *unknown* in samples. Dark matter, however, has many contexts. The "dark matter" of the universe is the hypothetical matter that remains invisible because it does not interact with electromagnetic radiation, but is estimated to encompass 85% of the universe's total mass [62–64]. The "dark matter" of biology often refers to the biological information that cannot be produced or interpreted by current methods — for example, in the field of microbiology only 1–15% of all bacteria and archaea can be cultivated in the lab with current cultivation methods [65] and there is large demand for new methods that enable the investigation of these "unculturables" *in situ* [65]. Here, multi-omic analyses, including metabolomics, of microbial

consortia is essential [40]. In metabolomics specifically, Silva et al. [66] refer to chemical signatures that remain unannotated as "dark matter", stating that "in an untargeted metabolomics experiment only 1.8% of the spectra can be annotated". This is the definition of "dark matter" adopted here, expanding the term chemical signatures to correspond to metabolite (or more comprehensively small molecule) features.

Targeted methods are now realistically capable of detecting tens to hundreds of "known" compounds per sample, whether these be with unit- or high-resolution techniques (e.g. Refs. [46,61]). While some specialist analytical laboratories now have thousands of reference standards in house, analytical limitations still restrict per-sample target determination to several hundred targets to obtain sufficient sensitivity and fragmentation information for confirmation. Increased coverage of targets is obtained with several measurements if necessary. Untargeted methods, in contrast, enable a view of all detectable substances in one measurement, which is why they are so valuable for discovery-based investigations. However, overviews in both the metabolomics [60] and environmental fields [50] reveal that tens of thousands of masses, corresponding to several thousand features, are reported in most untargeted studies — such that there is in general an order of magnitude difference between "known" and "unknown" features — a formidable challenge for elucidation that requires careful detection and prioritization of relevant features. A detailed breakdown of Swiss wastewaters revealed that after extensive target and suspect screening, an average of over 10,000 non-target masses remained on average per ionization mode, which could be broken down into approximately 8000 features per mode. Approximately 1300 of these features in each mode had isotope and adduct information associated with them. The remaining single masses were of insufficient intensity to obtain such information [55]. Fuhrer and Zamboni looked at the number of features reported in non-comparable metabolomics studies (drastically different samples and analytical methods, although all involved chromatographic separation) and reported a similar range of ~4000 to 8000 features per sample [60]. The challenge for untargeted methods is thus to determine which and how many of these features are relevant for further identification. Mahieu and Patti [67] recently used systems-level annotation (including data reduction with isotopic labelled experiments to remove non-biological signals and extensive degeneracy filters) to reduce >25,000 features to 892 unique metabolites for the relatively simple case of *E. coli* — although it is not yet clear how many of these are "known" or tentatively annotated. While such detailed methods are not available to all sample contexts, especially environmental matrices, they reveal many interesting facets in the data that require further investigation and, likely, the adjustment of current data processing workflows (discussed further below). Analytically one could thus postulate that the relevant "dark matter" in untargeted studies would generally be of the order of several hundreds to thousands of features. Ideally, untargeted studies should be complemented by targeted analysis and/or suspect screening to identify the known or potential known analytes first.

There are numerous methods to tackle this dark matter, yet also inherent limitations. As outlined above, analytical limitations of the broad untargeted methods are such that insufficient information will be available to identify a large proportion of the unknowns beyond a tentative annotation based on exact mass only (see e.g. Ref. [68] and Fig. 2). Provisional annotations can be made (e.g. via "suspect screening"); in metabolomics often using resources such as KEGG [69,70] and HMDB [71,72]. However, several categories of "unknowns" exist in the "dark matter" and a quote by the former US Secretary of Defense, Donald Rumsfeld, relating to weapons of mass destruction in Iraq is increasingly used to describe this [73] by categorizing information into known knowns, known unknowns
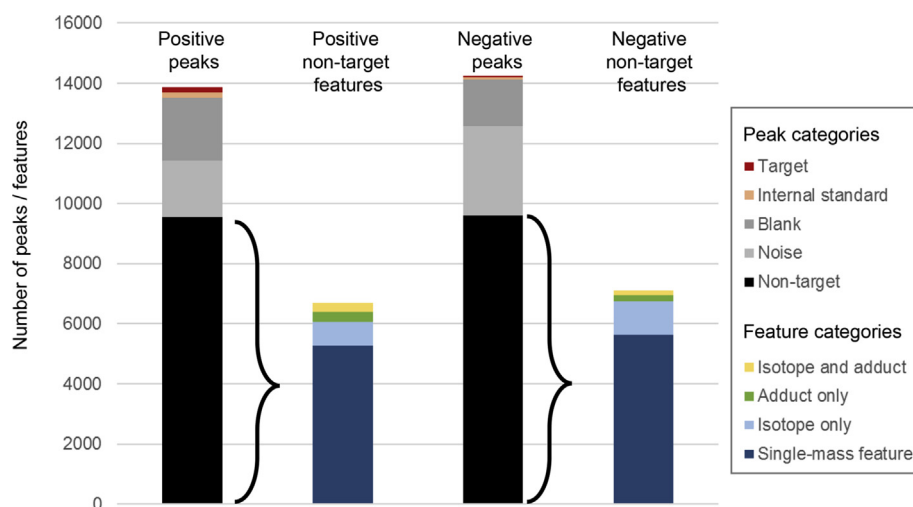
Fig. 2. Average breakdown of peaks (masses) and grouped features detected in 10 Swiss wastewater samples per mode following blank/noise removal (blue bars) and extensive target screening (red). The average 10,000 masses per mode could be grouped into approx. 7000 features per mode, of which only several hundred had either isotopic or adduct information (or both). Features remaining "black" are very challenging to elucidate/confirm. Details in Ref. [55]. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

and unknown unknowns [74]. This concept was introduced with the context of identification in mass spectrometry by Little et al. [74]. They defined a *known known* as a "compound suspected to be in the mixture, whose identity is to be confirmed by mass spectrometric analyses", a *known unknown* as "a compound that is unknown to the investigator, but is cited in the chemical literature or mass spectrometry reference databases" and an *unknown unknown* as "a compound that is not previously cited". Stein [75] then expanded this to form "Rumsfeld quadrants" by adding *unknown knowns* and redefining these according to the stricter context of library search and identification confidence. If a compound was expected in the samples and was identified by the library search it is a *known known*. *Unknown knowns* are unexpected by the analyst but identified by the library, whereas *known unknowns* are expected but not identified by the library. Finally, substances that are neither found nor expected are *unknown unknowns* [75]. Here, we slightly modify the concept presented by Little et al. [74], as it is closer to the definition of target, suspect and non-target, adding to this the "unknown known" as a substance that is known as a complex mixture, but not documented anywhere as an individual substance (and thus cannot be found with current computational methods; Table 1).

Given the current methodologies and feature denomination frameworks, we review the current state-of-the-art with respect to microbiome-linked unknown metabolites, methods to elucidate these and provide perspectives on the further developments required for enhanced identification success rates in the coming years.

2. Tackling the metabolic dark matter of the microbiome

2.1. Laboratory-based microbiome studies: HuMiX

In order to systematically resolve the complex landscape of human-microbial molecular interactions, experimental models that allow the probing and manipulation of metabolic interdependencies are essential. Widely used animal models, e.g. germ-free/gnotobiotic mice, exhibit important limitations resulting from differences in diet, divergent gut topology, and differences in the immune system, which render them sub-optimal with respect to testing the multitude of hypotheses which arise from multi-omic analyses of human microbiota [76]. Whereas experiments in human subjects may be possible for testing certain questions [77,78], such approaches are not possible in most cases due to ethical limitations. However, most importantly, human subjects do not offer the combinatorial power that is required to systematically probe and manipulate metabolism. In order to address these limitations, *in vitro* models have been developed based on cell culture technologies, which more or less recapitulate the *in vivo* conditions of the human gut [76,79]. Essential characteristics that need to be reflected in any good model for studying metabolic interactions include the ability to perfuse the device with dedicated culture media to allow the establishment of aerobic conditions for human cell culture and anaerobic conditions for gut bacteria, free exchange of small molecules between the individual cell contingents, easy access to the individual cell contingents following specific experimental regimes and compatibility with time-resolved metabolomic measurements.

Table 1
Categories for knowns and unknowns relating to dark matter in mass spectrometry. Modified definitions from [73—75].

Known known	Expected to occur
	Reference standard available
	Confirmed analytically
Known unknown	"Suspected" or unknown to investigator
	Documented in databases, literature
Unknown known	Known as a mixture or expert knowledge
	Undocumented as individual compound
Unknown unknown	Compound previously undocumented
	Full elucidation and confirmation required

18

In order to facilitate the study of human-microbial molecular interactions, HuMiX (human-microbial cross-talk), a microfluidics-based human-microbial co-culture system, was engineered to allow the culture of relevant human and microbial cell populations under conditions representative of those encountered in the gut [80]. In its most basic form, HuMiX consists of three parallel microfluidic channels and these are referred to as the microbial, epithelial and perfusion microchambers, thereby reflecting their primary purposes (see Fig. 3). The stacked and aligned channels follow a distinct spiral pattern to optimize the footprint and act as culture chambers with dedicated inlets and outlets, which allow for the inoculation of the relevant cell contingents and the precise control of the physiochemical conditions within each microchamber. The channels are separated from one another by semi-permeable polycarbonate membranes with pore sizes and coatings varying according to their particular function (see Fig. 3 and [80] for more detail). The model relies on an intricate experimental protocol to allow the establishment of co-cultures [80]. Briefly, following assembly and sterilization of the HuMiX device, human epithelial cells are cultured in the epithelial cell microchamber for 7 days until they have fully differentiated and formed a tight monolayer across the membrane. Following the establishment of an epithelial barrier, bacterial and/or immune cells are inoculated in the microbial and perfusion microchambers, respectively (see Fig. 3).

During co-culture, growth medium is typically perfused through the perfusion chamber via a peristaltic pump to recreate the peristaltic motions and intraluminal fluid flow present within the gut, thereby creating conditions that are representative of a healthy intact epithelial barrier. The environment within the HuMiX device is routinely monitored. Eluents can be collected from each microchamber, thereby allowing the continuous probing of the extracellular metabolome [80]. In addition, the modular architecture of the HuMiX device allows easy separation of the individual microchambers, which in turn ensure that the individual cell contingents can be probed following a given experimental regime [80].

The dimensions of the microchambers ensure that laminar flow profiles are provided, which, in turn, guarantees that all exchange between the cell contingents occurs via diffusion. Given the pore size cut-offs and biopolymers covering the membranes, the vast majority of the molecular exchange between the cell is via small molecules. The major strength of the HuMiX model is its ability to provide a reproducible and representative environment for both human and microbial cells. Thus, HuMiX limits the overall complexity typically encountered in the human gut to allow targeted investigations of host-microbe metabolic interactions in the absence of major matrix effects. As a result, HuMiX is intricately suited to systematically study human-microbial metabolic interactions as well as the impact of these interactions on cellular physiology. Additionally, the relative simplicity of the sample matrix (compared with more complex metabolomics and environmental matrices) enhances the ability for successful elucidation of detectable unknowns.

2.2. Metabolic dark matter in HuMiX experiments

In a previous investigation [80], HuMiX was employed to investigate changes in the metabolic profiles of human epithelial gut cells and bacteria under different co-culture conditions. Evaluation of biomolecular extracts in conjunction with other multi-omic datasets revealed that the metabolomics data reflected differing metabolic processes under the different co-cultivation scenarios, demonstrating the potential for HuMiX to allow systematic probing of both human and microbial metabolism. Although the original study design was not intended to investigate unknown metabolites, many significant unknowns were detected under the various co-culture scenarios, highlighting the panoply of so far uncharacterized small molecules that are either involved in or affected by host-microbe interactions. The following gives a brief overview:

In scenario A, a human epithelial gut cell line (Caco-2), was co-cultivated in HuMiX together with the facultative anaerobic commensal bacterium *Lactobacillus rhamnosus* GG (LGG). Mono-cultured Caco-2 cells served as a control. The intracellular metabolites of the human cells were then analyzed and compared between the co-culture and its control. Different metabolic profiles showed the effect of the co-culture on the metabolism of the human cells. Scenario B investigated the influence on the bacterial metabolism, where intracellular metabolites from LGG co-cultured with Caco-2 were compared to those of mono-cultured LGG. Again, changes in the metabolite profile could be observed. To furthermore demonstrate that HuMiX could also be employed for strictly anaerobic bacteria and potentially for more complex microbial communities, Caco-2 cells were co-cultivated together with LGG and the strictly anaerobic bacterium *Bacteroides caccae* (Scenario C). Analogous to the first scenario, the intracellular metabolites of the human cells were compared to a mono-cultured control of Caco-2. Here, the comparison of the metabolite profiles showed the most striking differences.
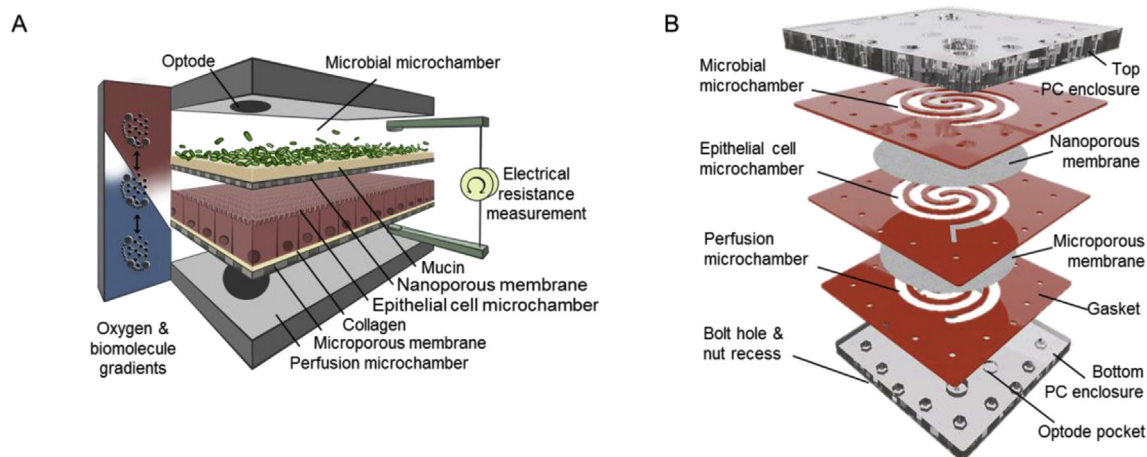

Fig. 3. HuMiX set-up. Extracted from Shah et al., 2016 [80]: (A) Principle of HuMiX. (B) Expanded view. PC: polycarbonate.

Following the co-cultivations described above, extracted polar metabolite fractions were analyzed in detail as these most clearly reflect metabolic interactions and responses on both the human as well as the microbial side. A double derivatization was carried out, firstly methoxyamine hydrochloride, then N-methyl-N-(trimethylsilyl)trifluoroacetamide (MSTFA). GC-EI-MS analysis was performed with m/z 70 to 800 and an alkane mix for retention index (RI) calibration. The chromatograms were processed using MetaboliteDetector [81] and an in-house library for derivatized metabolites. For prioritization, the relative intensities of each detected metabolite feature were compared between the biological replicates and the sample groups (co-culture versus control) and ordered according to their p-value. A summary of the top 100 metabolites for each scenario is given in Fig. 4, showing the proportion of known and unknown metabolites above and below a significance of p = .05. There were 44, 84 and 70 unknowns in scenarios A, B and C; shown in Fig. 4 A, B and C respectively − in other words nearly two thirds of these metabolites. Of these 198 unknowns, 106 were significantly different compared with the respective controls (p < .05) [80]. While it is possible that these numbers contain duplicates (see discussion below), there are no cases that are clearly duplicate on the basis of fragmentation and RI information.

This study was chosen to demonstrate the potential of HuMiX for metabolomics and as a practical example with a high number of relevant unknowns. Although it was not originally designed to investigate the unknowns, it nicely illustrates certain, current limitations. The double derivatization (trimethylsilylation (TMS), methoximation), which is necessary for increased volatility, makes it difficult to find metabolite features/spectra in other databases (such as NIST, despite the presence of some derivatized metabolites in there). It can also lead to duplicates, as different derivatized forms may relate to the same metabolite. Further, derivatization also changes the retention time, limiting the use of retention index prediction (as most predicted values e.g. in NIST correspond with underivatized substances). In addition, with GC-EI-MS alone it is not possible to acquire accurate mass data and as the molecular ion is often either low abundance or not visible at all, it is often difficult to calculate a molecular formula for further structure elucidation. While some formula prediction approaches are implemented e.g. within MOLGEN-MS [82] and NIST, neither of these are suited to higher-throughput methods needed for discovery-based screening. Without coupling to high accuracy MS, it often remains difficult to align unknowns detected in EI-MS with unknowns detected in HR-MS measurements. However, as mentioned above, the use of a system such as HuMiX offers exciting possibilities to elucidate these unknowns in future experiments. The design minimizes matrix effects, which can be a severe limitation in more complex samples, while showing high reproducibility between biological replicates. The system also allows the investigation of intra- and extracellular metabolites separately. Future HuMiX experiments will further probe the interactions between human cells and a complex, yet reduced synthetic microbiota to investigate and elucidate the production and exchange of intra- and extracellular metabolites of both human and microbial cells. Combined with additional HR-MS approaches and computational analyses, HuMiX shows great potential to tackle the unknowns in host-microbiome metabolomics.

3. Approaches for identifying dark matter

As outlined above, the features associated with dark matter far outnumber the "known knowns" (targets) and it still remains challenging to reconcile EI-MS features with HR-MS features. However, there are a plethora of methods and workflows aimed at supporting the identification of unknowns in mass spectrometry and many articles cover various aspects of these more comprehensively than possible here, e.g. Refs. [47,51,83–87]. Instead, the following material outlines the basic approaches with selected examples and a specific focus on very new approaches, especially those relevant to identification of unknowns in the microbiome. Fig. 5 includes a very simplified workflow for the major approaches, consistent with the terminology used in this article.

Microbiome studies are still relatively new and thus mostly at the "discovery" stage. Using the definitions above, most studies so far restrict themselves to "known knowns" (targets; reference standards in house) and "known unknowns" (discovered by suspect or non-target screening). Full elucidation of "unknown unknowns" (undocumented substances) is still extremely time consuming and beyond the scope of large screening studies, except for carefully selected features of high interest (see discussions in prioritization below). Approaches to deal with the "unknown knowns", i.e. complex mixtures such as homologues that are frequently observed in all samples, are still in their infancy and this remains a large challenge in tackling the dark matter with the current workflows [88].
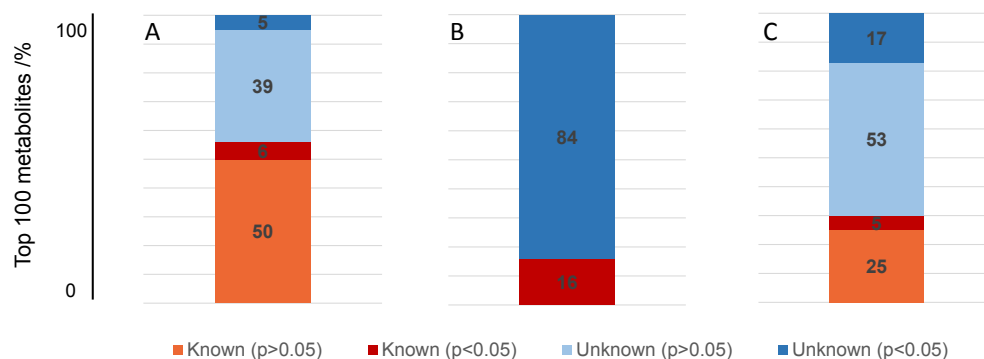
Fig. 4. Known and unknown features according to the level of significance of differences under different co-culture scenarios (Comparison of human-microbial co-cultures to their respective mono-culture controls). Red and blue indicate knowns and unknowns, respectively, with lighter shading indicating p > .05 and darker shading p < .05, respectively. (A) Intracellular metabolites of human cells (Caco-2) after co-culture with LGG. (B) Intracellular metabolites of LGG after co-culture with human cells. (C) Intracellular metabolites of human cells (Caco-2) after co-culture with LGG and *B. caccae*. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)
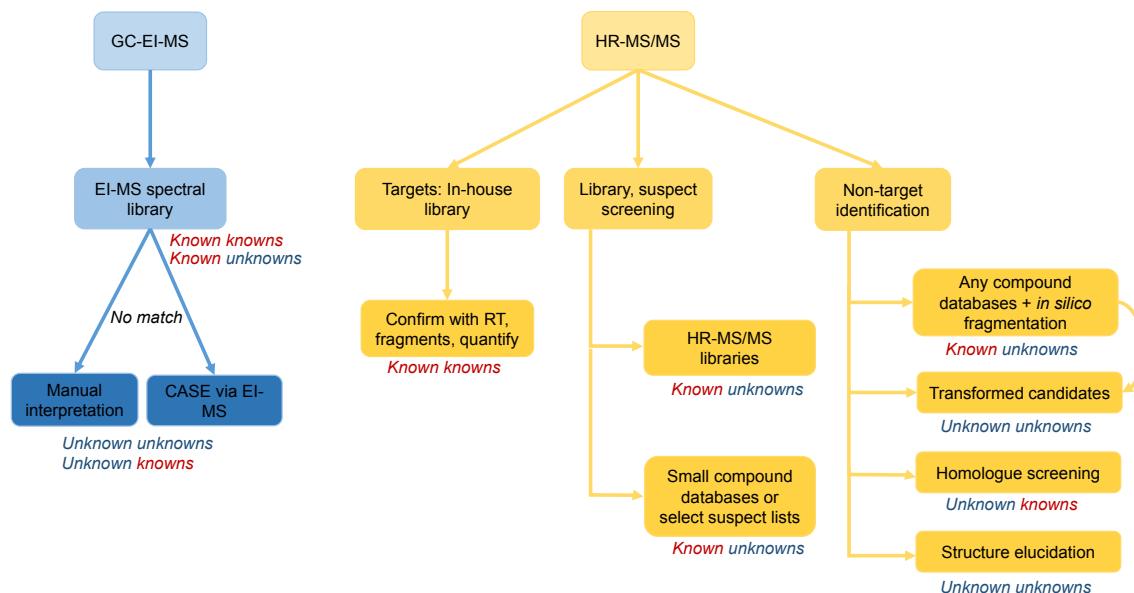
Fig. 5. Simplified workflows for unknown identification in GC-EI-MS (left) and HR-MS/MS (right).

3.1. "Known unknowns": resources

Information sources to find "known unknowns" for typical computational workflows include compound databases (containing chemical information, but not necessarily spectral information) and spectral libraries (containing chemical information and their associated spectra; here MS). Fig. 6 below contains an overview of major databases and spectral libraries.

Despite the millions of structures contained in PubChem and ChemSpider, and the hundreds of thousands of spectra in EI-MS libraries, many unknowns often still remain. For resources such as MS libraries it is often clear when no spectra match (indicating

an "unknown" that is beyond the library), while for large compound databases, often tens to thousands of candidate structures may match a given mass and, thus, candidate selection is necessary (see discussion below).

Importantly, non-target screening and unknown identification should not be performed in isolation, but are far more successful with supporting target and suspect screening to find as many "known unknowns" as possible. Full elucidation without prior information is an extremely time-consuming task and gathering as much information from the sample, experiment and other features present is critical to success. Target analysis and/or suspect screening allows comparatively efficient (tentative) identification
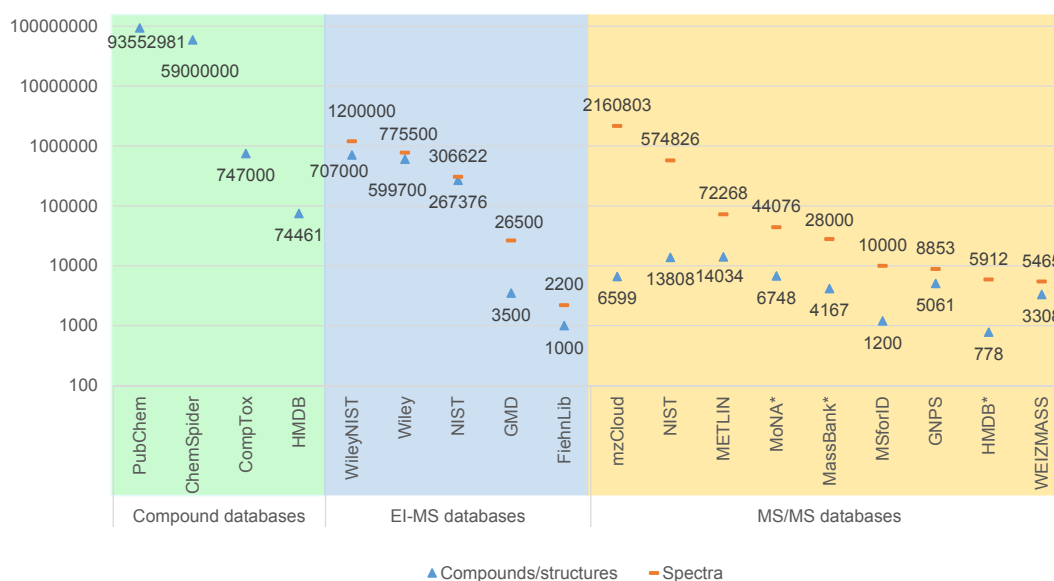
Fig. 6. An overview of selected compound databases (PubChem [89,90], ChemSpider [91,92], the CompTox Chemistry Dashboard [59,93] and HMDB [71,72,94]) and EI-MS databases [95–99] and soft ionization MS/MS databases. *indicates that EI-MS and/or predicted spectra were excluded from the numbers. Information for the MS/MS databases collated from the individual resources (mzCloud [100], NIST [97], METLIN [101,102], MoNA [103], MassBank [104,105], MSforID [106], GNPS [107], HMDB compounds with MS/MS only [71,72] and WEIZMASS [108] and [51]).

of expected features. With EI-MS, this step is often performed directly with large reference spectral libraries, either coupled with NIST and/or Wiley (see Fig. 6) or with smaller specialised collections specific to the experimental context, such as GOLM [98] or local libraries, such as described above for the HuMiX experiments. While this approach can also be used for the softer ionisation methods, the libraries are not as extensive (see Fig. 6) and the success rate at annotating tentative identities to detected peaks is still quite low. Although some libraries (e.g. mzCloud) now contain more spectra than the NIST EI-MS library, the number of compounds behind the spectra is several orders of magnitude lower. The spectral numbers are generally much larger than compound numbers in these libraries as many spectra are recorded with varying collision energies to overcome the lack of reproducibility and standardization (Fig. 6 for the breakdown).

For soft ionization measurements, alternative workflows to discover targets and suspects have evolved, either via peak picking and mass matching within peak lists, or via extracted ion chromatograms (see e.g. Ref. [50] for a more detailed delineation reviewing the various approaches used during a collaborative trial in the environmental community). Recently, target and untargeted analysis have been coupled together in microbiome studies, from analysis through to data processing and although this set-up still requires optimization, it is an important first step towards streamlining the whole process [57]. One key improvement needed to further improve "suspect screening" (looking for expected compounds without the reference standard in-house) via a reduction in the false positives is to use fragment information such that not only the exact mass, but also specific fragments can be used to increase the confidence in the match. For instance, Melnik et al. [57] used one fragment per target for confirmation, which is more than many but insufficient to delineate closely-related structures and ideally more should be used in workflows. Kaufmann et al. recently used predicted fragments (see below) to reduce false positives in residue screening of complex matrices [109]. The fragment information (with multiple fragments from measured spectra) was an essential quality control criterion in a recent pilot study to evaluate the potential of an Early Warning System (NormaNEWS) for Contaminants of Emerging Concern (CECs) to enable rapid determination of CEC distribution throughout Europe using retrospective screening techniques [58,59,110]. As a result, a list of top 3 fragments merged from various collision energies of all spectra per compound uploaded to the European MassBank [105] has recently been added to the public domain via the CompTox Chemistry Dashboard [59,93] and the NORMAN Suspect Exchange [58]; the latter list contains also a more extensive breakdown (per collision energy and instrument). While this list is just a prototype, expanding this to other collections will help increase the confidence in target and suspect screening greatly. The full mass spectra are still needed to capture intensity variations for more comprehensive confirmation in the workflows (e.g. MassBank, GNPS). Fragment matching serves to provide a quick tentative match to avoid too many false positives.

3.2. "Known unknowns": candidate selection without library spectra

The fragmentation information described so far was obtained on measured data. *In silico* fragmenters can be used to predict fragments in the absence of spectral information to establish whether potential candidate(s) could produce the observed (measured) fragments. Many *in silico* approaches exist, including combinatorial fragmenters (e.g. MetFrag, MAGMa) [111,112], machine learning approaches (CFM-ID, CSI:FingerID) [113,114], rule-based fragmenters (e.g. MS-FINDER) [115], all of which participated in the

2016 Critical Assessment of Small Molecule Identification (CASMI) contest [86]. Other approaches also exist, for instance some MS libraries (e.g. METLIN) and compound databases (e.g. HMDB) are now saving predicted spectra — whereby spectra are predicted by *in silico* fragmenters such as the ones mentioned above. Generally speaking, these *in silico* approaches depend on compound databases as a source of candidates. Metabolomics studies often limit themselves to KEGG (to find those candidates known to be associated with metabolism, usually only a few candidates per mass or formula) or extend themselves to PubChem (hundreds to thousands of candidates per mass/formula). Environmental studies tended rather to default straight to ChemSpider or PubChem as KEGG was too limited. The advent of the CompTox Dashboard as a smaller, environmentally relevant resource will help reduce candidate lists in these investigations [116]. An alternative to compound databases is structure generation, where all candidates with a given formula (and substructure restrictions) are enumerated. Very few studies so far with HR-MS use structure generation to source candidates and these are restricted to transformation studies with very detailed parent (and thus substructure) information (e.g. Ref. [117]). In contrast, structure generation coupled with substructure information via classifiers and rule-based candidate fragmentation and ranking has yielded a CASE system for EI-MS that can support elucidation in some cases (perspectives summarized in Ref. [48]). Such CASE systems with structure generation for EI-MS or even HR-MS are not yet quick or successful enough for large non-target screening studies, but rather may form a part of detailed elucidation on prioritized unknowns.

One outcome of the recent CASMI contest [86] and Ruttkies et al. [111] was the effect of the metadata on the success rate of tentatively identifying "known unknowns" amongst unknowns. For the datasets evaluated (all reference standards), the number of correct structures correctly ranked in first place rose from ~6 to 30% to over 70% with metadata, primarily driven by the number of references. The number of references is obviously highly relevant for environmental sciences, where well-known substances are more likely to occur in the environment (i.e. the references could be postulated to scale with the use). The relevance of the reference information to metabolomics studies is still a subject of debate in the community. However, as the name "suspect screening" implies, the reference information is merely supporting information to help find potentially interesting masses; these should be confirmed with analytical evidence before being considered anything other than tentative.

Metadata such as reference information will potentially greatly assist in finding the "known unknowns", but metabolomics *per se* involves metabolites and various transformation processes (including abiotic) also occur in the environment. Thus, another source of candidates becomes predicted metabolites or transformation products (TPs) from given starting products. Predictive systems such as enviPath [118,119], BioTransformer [120] or CATALOGIC [121] take a given structure as input and output transformed structures according to defined rules, certainties and also reaction steps desired. Combinatorial explosion is a limitation if too many steps are considered. The MINEs database [122,123] contains predicted metabolites/TPs of smaller compound databases such as KEGG [69], calculated using "promiscuous enzymes" theory, stored for access in methods such as the *in silico* fragmenters mentioned above. An alternative is to use metabolic logic to link potentially related masses in before/after samples [124], which is reasonably similar in concept to the principles behind the molecular networking approach [107,125]. Realistically, several approaches are still needed to fully explore the complexity of microbiome samples, as this contains both genetic, transcriptional, enzymatic, regulatory and not least metabolic signals and processes as well as the inputs (e.g. food, anthropogenic substances).

3.3. Dark matter and identification challenges

3.3.1. "Unknown knowns"

Homologue series (substances related by discrete building blocks, such as CH_2, CH_2CH_2O, CF_2) are prevalent in both metabolomics and environmental samples. Classic examples include lipids and surfactants. However, homologue screening (e.g. Refs. [126,127]) reveals that many more series exist than can be annotated, while several commonly observed mass differences cannot yet be annotated in HR-MS environmental studies (e.g. Ref. [55]). A huge challenge for assessing and annotating the "dark matter" will be to make progress on the identification of at least one member of such series, as there are often dozens of peaks and features (if not more) related to each series. There are detailed resources for lipids such as LipidMaps [128] and excellent computational approaches coupling prediction of lipid structures and the spectra [129]. However, the so-called Unknown and Variable Composition, Complex Reaction Products and Biological Materials (UVCBs) are often poorly characterized (and poorly defined) homologues and can be released into the environment in huge amounts. Thousands of high exposure UVCBs exist [59] and developing computational methods to deal with this complexity will be critical to cross-annotating the dark matter.

3.3.2. Feature grouping: in-source fragments and unexpected adducts

As mentioned above, one of the greatest challenges in untargeted data processing is reconciling the detected masses into relevant features for further identification efforts. Common efforts such as CAMERA [130,131] and the R package nontarget [55,132] (used to create Fig. 2) use rule-based approaches to screen for known adducts and isotopes. Few approaches consider the large role of in-source fragments in a systematic way, especially for certain compound classes, as these are not documented as well as potential adducts. Approaches such as RAMClust help explore this in a data-driven way [133]. If not considered carefully, some compound classes such as carboxylic acids may produce in-source fragments more intense than the precursor peak in the first stage MS1 (full scan) and, thus, identification/annotation may be performed on the wrong ion. A wider investigation into the recognition of in-source fragments in the MS1 and dealing with the consequences during feature detection is greatly needed. An important step will be the inclusion of these spectra in libraries. Preliminary results on data extracted previously for MassBank indicate that in-source fragments are far more common (and far more intense) than long suspected by the community [134] and the latest NIST MS library (NIST2017) now includes in-source fragments and also the MS/MS spectra corresponding to these fragments [135], which will greatly facilitate the recognition of mis-annotations. Failure to recognize in-source fragments results in mis-identifications if e.g. the non-target mass is a fragment and not the parent ion, and they are a potential significant contribution to the dark matter. Mahieu and Patti also showed very recently that analyte multimers, situational adducts and interactions with background ions can all contribute greatly to peak degeneracy [67], representing more layers of complexity to consider when consolidating the detected masses into actual features and assigning the correct "parent mass" for further identification efforts. Integrating these aspects into untargeted data analysis workflows is a challenge that the community needs to overcome in the next few years.

3.3.3. Prioritization

Finally, choosing the relevant "known unknowns" to confirm or the "unknown unknowns" for further elucidation efforts using untargeted approaches still, in reality, involves "cherry picking" or going for the low hanging fruits. The sample amount from discovery-based studies is rarely sufficient to allow for additional elucidation such as NMR techniques, although impressive couplings with LC are now enabling elucidation efforts where it was previously not possible. Efforts such as isotope labeling help systematically select those features that are metabolized in metabolomics studies, but not in environmental studies. Assigning annotations or tentative identifications to peaks in untargeted studies with even the largest compound databases still leaves many unknowns and focus is often on identifying the most intense, those with the most promising annotation/metadata or those with clear isotope and adduct signals as well as fragment information. Successful elucidation/confirmation of substances without an isotope pattern is rare (see the remaining "single masses" in Fig. 2), as the intensity is often insufficient for fragments and thus any form of structural information. Additionally, substances of lower intensity often experience matrix interferences. An exception are the homologues, where low intensity peaks in the series can still be cross-annotated successfully when some higher intensity peaks in the series are available, assuming sufficient additional evidence such as chromatographic peak shape is present.

4. Future perspectives and challenges

As outlined above and highlighted in Section 2.2, many challenges exist on the way to elucidating the "dark matter" of the microbiome, yet much progress in terms of available methodologies has been made. One challenge is to reconcile metabolite features from previous studies performed with (derivatized) GC-EI-MS with features detected in HR-MS measurements, especially where derivatization potentially complicates the overlap. Although GC and LC have different compound domains, it is clear that some substance classes will be detected with both techniques. A recent Collaborative Non-target Screening Trial run in Europe on a common water sample revealed 5.4% overlap between the substances reported by all participants with GC versus LC-based methods (target and suspect/library screening) [50]. This trial was one of the first collaborative efforts where results were merged between GC-EI-MS, GC-EI-MS coupled with chemical ionization (CI), GCxGC (two dimensional gas chromatography) techniques and finally GC-APCI-MS. While no institute in this trial had derivatized substances, the challenges nominated by researchers still apply — including the lack of streamlined workflows for GCxGC or for combining EI-MS with exact masses from CI. The identification approaches used for the GC-APCI-MS data were far more similar to the LC-HRMS approaches described above, as the EI-MS databases do not apply to APCI spectra. It was only possible to estimate the 5.4% overlap between LC and GC on the "knowns" in this trial. However, when derivatization is in the mix it is much more challenging to reconcile the unknown features (EI-MS spectrum plus retention index; mostly derivatized) with supporting measurements in LC-HRMS unless there is a coupling of GC with CI or APCI-MS to act as a "bridge" between the data. Thus, it remains difficult to estimate how much of the dark matter may overlap between LC and GC in microbiome-host interactions. Judging by the user difficulties during the trial and progress since then, one of the challenges over the next few years will be to establish workflows to assist in the merging of data and peak lists across chromatographic and ionization types in order to answer this question. This will become gradually easier as high-resolution GC methods become more accessible.

In silico methods to support the tentative identification of unknowns have come a long way in a few years, yet mass spectrometry is also inherently limited. The "dark matter" likely consists of

many signals resulting from complex isomeric mixtures that will not realistically be distinguishable with MS; nor will full elucidation be possible with MS-based methods alone. One of the challenges for MS-based workflows will be to find new ways of dealing with this, to move away from a single-structure-based annotation/identification towards annotation with generic structures, where some degree of structural uncertainty can be displayed. While this is often possibly pictorially to some extent, cheminformatics database exchange formats are not all yet able to support this (e.g. InChIs and InChIKeys are only defined for discrete structures, while multiple forms of "generic" or "extended" SMILES already exist but are not standardized). As for TP prediction, combinatorial explosion will prevent the generation of all possible isomers for even relatively simple isomer sets, such that "representative structures" (e.g. Refs. [88,136]) or "related structures" (e.g. Refs. [116,137,138]) may be a way forward for MS-based workflows to deal with these issues. As stated above, the correct recognition of potential in-source fragment peaks and other analytical phenomena (e.g. unexpected adducts) is another improvement that will assist in correctly identifying the "true" precursor peaks.

Another advanced technology to investigate unknown metabolites that should also be mentioned is single-cell MS. For example high-density microarrays for MS allow for high-throughput metabolite analysis of single cells with a detection limit of 100 amol to 10 fmol [139,140]. These approaches enable the detailed and minute description of metabolic differences between individual cells based for example on cell age or cell cycle stage [139].

With increased data and knowledge sharing between the environmental and metabolomics communities, as well as the multi-omics layers, the perspectives for "illuminating the dark matter" [66] are looking bright and despite the challenges that still lie ahead, good progress is being made to address some of the critical issues currently holding up the workflows. In this context, it will be particularly important to integrate metabolomics data with other omics data. As has already been demonstrated, the inclusion of metadata dramatically increases the number of metabolite features which can be identified (see detailed discussion above). In particular, the generation of concomitant metagenomic, metatranscriptomic and metaproteomic data on human microbiome samples will allow proper contextualization of metabolomic data in terms of active metabolic pathways and, thus, predictions of expected metabolites. Furthermore, by performing integrated multiomic analyses on longitudinally collected samples in combination with advanced time-series analysis methods, one will be able to associate the expression of proteins of unknown function to specific metabolite features which in turn will allow causal inferences regarding enzymatic activity. Such approaches relating proteomic to metabolomic data have proven fruitful for the elucidation of unknown metabolic traits in communities of low diversity [141,142]. However, given rapid advances in sequencing as well as proteomic and metabolomic methodologies especially in relation to enhanced depth of coverage of biomolecular pools, such approaches are now poised to find application in the more complex communities of the human gut microbiome.

Although the integration of multiple other omes will allow more comprehensive exploitation of the metabolomics data, really understanding the function and origin of the unknowns will require extensive work. Here, a prioritized list of unknown metabolites which are for example commonly seen across the human population could be established. Such a list would then provide the basis for detailed biochemical characterization not least in terms of linking them to possible proteins of unknown function which are abundantly present in metaproteomic datasets [143].

References

[1] K.J. Pflughoeft, J. Versalovic, Human microbiome in health and disease, Annu. Rev. Pathol. 7 (2012) 99–122.

[2] C.A. Thaiss, N. Zmora, M. Levy, E. Elinav, The microbiome and innate immunity, Nature 535 (2016) 65–74.

[3] W.H. Karasov, C. Martínez del Rio, E. Caviedes-Vidal, Ecological physiology of diet and digestive systems, Annu. Rev. Physiol. 73 (2011) 69–93.

[4] J.G. LeBlanc, C. Milani, G.S. de Giori, F. Sesma, D. van Sinderen, M. Ventura, Bacteria as vitamin suppliers to their host: a gut microbiota perspective, Curr. Opin. Biotechnol. 24 (2013) 160–168.

[5] L.V. Hooper, D.R. Littman, A.J. Macpherson, Interactions between the microbiota and the immune system, Science 336 (2012) 1268–1273.

[6] N. Kamada, S.-U. Seo, G.Y. Chen, G. Nunez, Role of the gut microbiota in immunity and inflammatory disease, Nat. Rev. Immunol. 13 (2013) 321–335.

[7] S.P. Claus, H. Guillou, S. Ellero-Simatos, The gut microbiota: a major player in the toxicity of environmental pollutants? Npj Biofilms Microbiomes 2 (2016) 16003.

[8] M.T. Abreu, Toll-like receptor signalling in the intestinal epithelium: how bacterial recognition shapes intestinal function, Nat. Rev. Immunol. 10 (2010) 131–144.

[9] M. Li, B. Wang, M. Zhang, M. Rantalainen, S. Wang, H. Zhou, Y. Zhang, J. Shen, X. Pang, M. Zhang, H. Wei, Y. Chen, H. Lu, J. Zuo, M. Su, Y. Qiu, W. Jia, C. Xiao, L.M. Smith, S. Yang, E. Holmes, H. Tang, G. Zhao, J.K. Nicholson, L. Li, L. Zhao, Symbiotic gut microbes modulate human metabolic phenotypes, Proc. Natl. Acad. Sci. 105 (2008) 2117–2122.

[10] J.L. Sonnenburg, F. Backhed, Diet-microbiota interactions as moderators of human metabolism, Nature 535 (2016) 56–64.

[11] Y. Furusawa, Y. Obata, S. Fukuda, T.A. Endo, G. Nakato, D. Takahashi, Y. Nakanishi, C. Uetake, K. Kato, T. Kato, M. Takahashi, N.N. Fukuda, S. Murakami, E. Miyauchi, S. Hino, K. Atarashi, S. Onawa, Y. Fujimura, T. Lockett, J.M. Clarke, D.L. Topping, M. Tomita, S. Hori, O. Ohara, T. Morita, H. Koseki, J. Kikuchi, K. Honda, K. Hase, H. Ohno, Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells, Nature 504 (2013) 446–450.

[12] Y. Shi, L. Mu, An expanding stage for commensal microbes in host immune regulation, Cell. Mol. Immunol. 14 (4) (2017) 339–348.

[13] S.W. Olesen, E.J. Alm, Dysbiosis is not an answer, N. Microbiol. 1 (2016) 16228.

[14] J. Qin, R. Li, J. Raes, M. Arumugam, K.S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D.R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H.B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Dore, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, S.D. Ehrlich, J. Wang, A human gut microbial gene catalogue established by metagenomic sequencing, Nature 464 (2010) 59–65.

[15] J. Li, H. Jia, X. Cai, H. Zhong, Q. Feng, S. Sunagawa, M. Arumugam, J.R. Kultima, E. Prifti, T. Nielsen, A.S. Juncker, C. Manichanh, B. Chen, W. Zhang, F. Levenez, J. Wang, X. Xu, L. Xiao, S. Liang, D. Zhang, Z. Zhang, W. Chen, H. Zhao, J.Y. Al-Aama, S. Edris, H. Yang, J. Wang, T. Hansen, H.B. Nielsen, S. Brunak, K. Kristiansen, F. Guarner, O. Pedersen, J. Dore, S.D. Ehrlich, M. Consortium, P. Bork, J. Wang, An integrated catalog of reference genes in the human gut microbiome, Nat Biotech 32 (2014) 834–841.

[16] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D.R. Mende, G.R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H.B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E.G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W.M. de Vos, S. Brunak, J. Dore, M. Consortium, J. Weissenbach, S.D. Ehrlich, P. Bork, Enterotypes of the human gut microbiome, Nature 473 (2011) 174–180.

[17] G. Falony, M. Joossens, S. Vieira-Silva, J. Wang, Y. Darzi, K. Faust, A. Kurilshikov, M.J. Bonder, M. Valles-Colomer, D. Vandeputte, R.Y. Tito, S. Chaffron, L. Rymenans, C. Verspecht, L. De Sutter, G. Lima-Mendez, K. D'hoe, K. Jonckheere, D. Homola, R. Garcia, E.F. Tigchelaar, L. Eeckhaudt,

J. Fu, L. Henckaerts, A. Zhernakova, C. Wijmenga, J. Raes, Population-level analysis of gut microbiome variation, Science 352 (2016) 560–564.

[18] C. Huttenhower, D. Gevers, R. Knight, S. Abubucker, J.H. Badger, A.T. Chinwalla, H.H. Creasy, A.M. Earl, M.G. FitzGerald, R.S. Fulton, M.G. Giglio, K. Hallsworth-Pepin, E.A. Lobos, R. Madupu, V. Magrini, J.C. Martin, M. Mitreva, D.M. Muzny, E.J. Sodergren, J. Versalovic, A.M. Wollam, K.C. Worley, J.R. Wortman, S.K. Young, Q. Zeng, K.M. Aagaard, O.O. Abolude, E. Allen-Vercoe, E.J. Alm, L. Alvarado, G.L. Andersen, S. Anderson, E. Appelbaum, H.M. Arachchi, G. Armitage, C.A. Arze, T. Ayvaz, C.C. Baker, L. Begg, T. Belachew, V. Bhonagiri, M. Bihan, M.J. Blaser, T. Bloom, V. Bonazzi, J. Paul Brooks, G.A. Buck, C.J. Buhay, D.A. Busam, J.L. Campbell, S.R. Canon, B.L. Cantarel, P.S.G. Chain, I.-M.A. Chen, L. Chen, S. Chhibba, K. Chu, D.M. Ciulla, J.C. Clemente, S.W. Clifton, S. Conlan, J. Crabtree, M.A. Cutting, N.J. Davidovics, C.C. Davis, T.Z. DeSantis, C. Deal, K.D. Delehaunty, F.E. Dewhirst, E. Deych, Y. Ding, D.J. Dooling, S.P. Dugan, W. Michael Dunne, A. Scott Durkin, R.C. Edgar, R.L. Erlich, C.N. Farmer, R.M. Farrell, K. Faust, M. Feldgarden, V.M. Felix, S. Fisher, A.A. Fodor, L.J. Forney, L. Foster, V. Di Francesco, J. Friedman, D.C. Friedrich, C.C. Fronick, L.L. Fulton, H. Gao, N. Garcia, G. Giannoukos, C. Giblin, M.Y. Giovanni, J.M. Goldberg, J. Goll, A. Gonzalez, A. Griggs, S. Gujja, S. Kinder Haake, B.J. Haas, H.A. Hamilton, E.L. Harris, T.A. Hepburn, B. Herter, D.E. Hoffmann, M.E. Holder, C. Howarth, K.H. Huang, S.M. Huse, J. Izard, J.K. Jansson, H. Jiang, C. Jordan, V. Joshi, J.A. Katancik, W.A. Keitel, S.T. Kelley, C. Kells, N.B. King, D. Knights, H.H. Kong, O. Koren, S. Koren, K.C. Kota, C.L. Kovar, N.C. Kyrpides, P.S. La Rosa, S.L. Lee, K.P. Lemon, N. Lennon, C.M. Lewis, L. Lewis, R.E. Ley, K. Li, K. Liolios, B. Liu, Y. Liu, C.-C. Lo, C.A. Lozupone, R. Dwayne Lunsford, T. Madden, A.A. Mahurkar, P.J. Mannon, E.R. Mardis, V.M. Markowitz, K. Mavromatis, J.M. McCorrison, D. McDonald, J. McEwen, A.L. McGuire, P. McInnes, T. Mehta, K.A. Mihindukulasuriya, J.R. Miller, P.J. Minx, I. Newsham, C. Nusbaum, M. O'Laughlin, J. Orvis, I. Pagani, K. Palaniappan, S.M. Patel, M. Pearson, J. Peterson, M. Podar, C. Pohl, K.S. Pollard, M. Pop, M.E. Priest, L.M. Proctor, X. Qin, J. Raes, J. Ravel, J.G. Reid, M. Rho, R. Rhodes, K.P. Riehle, M.C. Rivera, B. Rodriguez-Mueller, Y.-H. Rogers, M.C. Ross, C. Russ, R.K. Sanka, P. Sankar, J. Fah Sathirapongsasuti, J.A. Schloss, P.D. Schloss, T.M. Schmidt, M. Scholz, L. Schriml, A.M. Schubert, N. Segata, J.A. Segre, W.D. Shannon, R.R. Sharp, T.J. Sharpton, N. Shenoy, N.U. Sheth, G.A. Simone, I. Singh, C.S. Smillie, J.D. Sobel, D.D. Sommer, P. Spicer, G.G. Sutton, S.M. Sykes, D.G. Tabbaa, M. Thiagarajan, C.M. Tomlinson, M. Torralba, T.J. Treangen, R.M. Truty, T.A. Vishnivetskaya, J. Walker, L. Wang, Z. Wang, D. V Ward, W. Warren, M.A. Watson, C. Wellington, K.A. Wetterstrand, J.R. White, K. Wilczek-Boney, Y. Wu, K.M. Wylie, T. Wylie, C. Yandava, L. Ye, Structure, function and diversity of the healthy human microbiome, Nature 486 (2012) 207–214.

[19] J.K. Goodrich, J.L. Waters, A.C. Poole, J.L. Sutter, O. Koren, R. Blekhman, M. Beaumont, W. Van Treuren, R. Knight, J.T. Bell, T.D. Spector, A.G. Clark, R.E. Ley, Human genetics shape the gut microbiome, Cell 159 (2014) 789–799.

[20] T. Yatsunenko, F.E. Rey, M.J. Manary, I. Trehan, M.G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R.N. Baldassano, A.P. Anokhin, A.C. Heath, B. Warner, J. Reeder, J. Kuczynski, J.G. Caporaso, C.A. Lozupone, C. Lauber, J.C. Clemente, D. Knights, R. Knight, J.I. Gordon, Human gut microbiome viewed across age and geography, Nature 486 (2012) 222–227.

[21] J.J. Faith, J.L. Guruge, M. Charbonneau, S. Subramanian, H. Seedorf, A.L. Goodman, J.C. Clemente, R. Knight, A.C. Heath, R.L. Leibel, M. Rosenbaum, J.I. Gordon, The long-term stability of the human gut microbiota, Science 341 (2013), 1237439.

[22] S. Schloissnig, M. Arumugam, S. Sunagawa, M. Mitreva, J. Tap, A. Zhu, A. Waller, D.R. Mende, J.R. Kultima, J. Martin, K. Kota, S.R. Sunyaev, G.M. Weinstock, P. Bork, Genomic variation landscape of the human gut microbiome, Nature 493 (2013) 45–50.

[23] A. Zhernakova, A. Kurilshikov, M.J. Bonder, E.F. Tigchelaar, M. Schirmer, T. Vatanen, Z. Mujagic, A.V. Vila, G. Falony, S. Vieira-Silva, J. Wang, F. Imhann, E. Brandsma, S.A. Jankipersadsing, M. Joossens, M.C. Cenit, P. Deelen, M.A. Swertz, R.K. Weersma, E.J.M. Feskens, M.G. Netea, D. Gevers, D. Jonkers, L. Franke, Y.S. Aulchenko, C. Huttenhower, J. Raes, M.H. Hofker, R.J. Xavier, C. Wijmenga, J. Fu, Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity, Science 352 (2016) 565–569.

[24] C. Manichanh, A. Eck, E. Varela, J. Roca, J.C. Clemente, A. González, D. Knights, R. Knight, S. Estrella, C. Hernandez, D. Guyonnet, A. Accarino, J. Santos, J.-R. Malagelada, F. Guarner, F. Azpiroz, Anal gas evacuation and colonic microbiota in patients with flatulence: effect of diet, Gut 63 (2014) 401–408.

[25] D.N. Frank, A.L. St Amand, R.A. Feldman, E.C. Boedeker, N. Harpaz, N.R. Pace, Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases, Proc. Natl. Acad. Sci. 104 (2007) 13780–13785.

[26] N. Qin, F. Yang, A. Li, E. Prifti, Y. Chen, L. Shao, J. Guo, E. Le Chatelier, J. Yao, L. Wu, J. Zhou, S. Ni, L. Liu, N. Pons, J.-M. Batto, S.P. Kennedy, P. Leonard, C. Yuan, W. Ding, Y. Chen, X. Hu, B. Zheng, G. Qian, W. Xu, S.D. Ehrlich, S. Zheng, L. Li, Alterations of the human gut microbiome in liver cirrhosis, Nature 513 (2014) 59–64.

[27] C.T. Brown, A.G. Davis-Richardson, A. Giongo, K.A. Gano, D.B. Crabb, N. Mukherjee, G. Casella, J.C. Drew, J. Ilonen, M. Knip, H. Hyöty, R. Veijola, T. Simell, O. Simell, J. Neu, C.H. Wasserfall, D. Schatz, M.A. Atkinson, E.W. Triplett, Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes, PLos One 6

[28] A. Giongo, K.A. Gano, D.B. Crabb, N. Mukherjee, L.L. Novelo, G. Casella, J.C. Drew, J. Ilonen, M. Knip, H. Hyoty, R. Veijola, T. Simell, O. Simell, J. Neu, C.H. Wasserfall, D. Schatz, M.A. Atkinson, E.W. Triplett, Toward defining the autoimmune microbiome for type 1 diabetes, ISME J. 5 (2011) 82–91.

[29] J. Qin, Y. Li, Z. Cai, S. Li, J. Zhu, F. Zhang, S. Liang, W. Zhang, Y. Guan, D. Shen, Y. Peng, D. Zhang, Z. Jie, W. Wu, Y. Qin, W. Xue, J. Li, L. Han, D. Lu, P. Wu, Y. Dai, X. Sun, Z. Li, A. Tang, S. Zhong, X. Li, W. Chen, R. Xu, M. Wang, Q. Feng, M. Gong, J. Yu, Y. Zhang, M. Zhang, T. Hansen, G. Sanchez, J. Raes, G. Falony, S. Okuda, M. Almeida, E. LeChatelier, P. Renault, N. Pons, J.-M. Batto, Z. Zhang, H. Chen, R. Yang, W. Zheng, S. Li, H. Yang, J. Wang, S.D. Ehrlich, R. Nielsen, O. Pedersen, K. Kristiansen, J. Wang, A metagenome-wide association study of gut microbiota in type 2 diabetes, Nature 490 (2012) 55–60.

[30] P.J. Turnbaugh, R.E. Ley, M.A. Mahowald, V. Magrini, E.R. Mardis, J.I. Gordon, An obesity-associated gut microbiome with increased capacity for energy harvest, Nature 444 (2006) 1027–1131.

[31] R.E. Ley, P.J. Turnbaugh, S. Klein, J.I. Gordon, Microbial ecology: human gut microbes associated with obesity, Nature 444 (2006) 1022–1023.

[32] Z. Wang, E. Klipfell, B.J. Bennett, R. Koeth, B.S. Levison, B. DuGar, A.E. Feldstein, E.B. Britt, X. Fu, Y.-M. Chung, Y. Wu, P. Schauer, J.D. Smith, H. Allayee, W.H.W. Tang, J.A. DiDonato, A.J. Lusis, S.L. Hazen, Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease, Nature 472 (2011) 57–63.

[33] A.D. Kostic, D. Gevers, C.S. Pedamallu, M. Michaud, F. Duke, A.M. Earl, A.I. Ojesina, J. Jung, A.J. Bass, J. Tabernero, J. Baselga, C. Liu, R.A. Shivdasani, S. Ogino, B.W. Birren, C. Huttenhower, W.S. Garrett, M. Meyerson, Genomic analysis identifies association of Fusobacterium with colorectal carcinoma, Genome Res. 22 (2012) 292–298.

[34] J.R. Marchesi, B.E. Dutilh, N. Hall, W.H.M. Peters, R. Roelofs, A. Boleij, H. Tjalsma, Towards the human colorectal cancer microbiome, PLos One 6 (2011), e20447.

[35] J.U. Scher, A. Sczesnak, R.S. Longman, N. Segata, C. Ubeda, C. Bielski, T. Rostron, V. Cerundolo, E.G. Pamer, S.B. Abramson, C. Huttenhower, D.R. Littman, Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis, Elife 2 (2013), e01202.

[36] F. Scheperjans, V. Aho, P.A.B. Pereira, K. Koskinen, L. Paulin, E. Pekkonen, E. Haapaniemi, S. Kaakkola, J. Eerola-Rautio, M. Pohja, E. Kinnunen, K. Murros, P. Auvinen, Gut microbiota are related to Parkinson's disease and clinical phenotype, Mov. Disord. 30 (2015) 350–358.

[37] T.R. Sampson, J.W. Debelius, T. Thron, S. Janssen, G.G. Shastri, Z.E. Ilhan, C. Challis, C.E. Schretter, S. Rocha, V. Gradinaru, M.-F. Chesselet, A. Keshavarzian, K.M. Shannon, R. Krajmalnik-Brown, P. Wittung-Stafshede, R. Knight, S.K. Mazmanian, Gut microbiota regulate motor deficits and neuroinflammation in a model of Parkinson's disease, Cell 167 (2017) 1469–1480 e12.

[38] A. Heintz-Buschart, U. Pandey, T. Wicke, F. Sixel-Döring, A. Janzen, E. Sittig-Wiegand, C. Trenkwalder, W.H. Oertel, B. Mollenhauer, P. Wilmes, The nasal and gut microbiome in Parkinson's disease and idiopathic rapid eye movement sleep behavior disorder, Mov. Disord. (2017), https://doi.org/10.1002/mds.27105.

[39] C. Noecker, C.P. McNally, A. Eng, E. Borenstein, High-resolution characterization of the human microbiome, Transl. Res. 179 (2017) 7–23.

[40] E.E.L. Muller, E. Glaab, P. May, N. Vlassis, P. Wilmes, Condensing the omics fog of microbial communities, Trends Microbiol. 21 (2013) 325–333.

[41] H.K. Pedersen, V. Gudmundsdottir, H.B. Nielsen, T. Hyotylainen, T. Nielsen, B.A.H. Jensen, K. Forslund, F. Hildebrand, E. Prifti, G. Falony, E. Le Chatelier, F. Levenez, J. Doré, I. Mattila, D.R. Plichta, P. Pöhö, L.I. Hellgren, M. Arumugam, S. Sunagawa, S. Vieira-Silva, T. Jørgensen, J.B. Holm, K. Trošt, MetaHIT Consortium, K. Kristiansen, S. Brix, J. Raes, J. Wang, T. Hansen, P. Bork, S. Brunak, M. Oresic, S.D. Ehrlich, O. Pedersen, Human gut microbes impact host serum metabolome and insulin sensitivity, Nature 535 (2016) 376–381.

[42] T.A. Clayton, D. Baker, J.C. Lindon, J.R. Everett, J.K. Nicholson, Pharmacometabonomic identification of a significant host-microbiome metabolic interaction affecting human drug metabolism, Proc. Natl. Acad. Sci. U. S. A 106 (2009) 14728–14733.

[43] H.J. Haiser, D.B. Gootenberg, K. Chatman, G. Sirasani, E.P. Balskus, P.J. Turnbaugh, Predicting and manipulating cardiac drug inactivation by the human gut bacterium Eggerthella lenta, Science 341 (2013) 295–298.

[44] P. Vernocchi, F. Del Chierico, L. Putignani, Gut microbiota profiling: metabolomics based approach to unravel compounds affecting human health, Front. Microbiol. 7 (2016) 1144.

[45] H. Roume, E. Muller, T. Cordes, J. Renaut, K. Hiller, P. Wilmes, A biomolecular isolation framework for eco-systems biology, ISME J. 7 (2013) 110–121.

[46] P.C. Dorrestein, S.K. Mazmanian, R. Knight, Finding the missing links among metabolites, microbes, and the host, Immunity 40 (2014) 824–832.

[47] F. Hufsky, K. Scheubert, S. Böcker, Computational mass spectrometry for small-molecule fragmentation, TrAC - Trends Anal. Chem. 53 (2014) 41–48.

[48] A. Kerber, R. Laue, M. Meringer, C. Rücker, E. Schymanski, Mathematical Chemistry and Chemoinformatics : Structure Generation, Elucidation, and Quantitative Structure-property Relationships, Walter de Gruyter GmbH, Berlin, Germany, 2014, ISBN 978-3-11-030007-9.

[49] J.J. Pitt, Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry, Clin. Biochem. Rev. 30 (2009) 19–34.

[50] E.L. Schymanski, H.P. Singer, J. Slobodnik, I.M. Ipolyi, P. Oswald, M. Krauss,

T. Schulze, P. Haglund, T. Letzel, S. Grosse, N.S. Thomaidis, A. Bletsou, C. Zwiener, M. Ibáñez, T. Portolés, R. De Boer, M.J. Reid, M. Onghena, U. Kunkel, W. Schulz, A. Guillon, N. Noyon, G. Leroy, P. Bados, S. Bogialli, D. Stipaničev, P. Rostkowski, J. Hollender, Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis, Anal. Bioanal. Chem. 407 (2015) 6237–6255.

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

[52] M. Bedair, L.W. Sumner, Current and emerging mass-spectrometry technologies for metabolomics, TrAC - Trends Anal. Chem. 27 (2008) 238–250.

[53] O. Fiehn, Metabolomics - the link between genotypes and phenotypes, Plant Mol. Biol. 48 (2002) 155–171.

[54] T. Kind, O. Fiehn, J. Bruand, P. Baldi, M. Nicklaus, P. Stothard, Z. Chang, J. Woolsey, R. Edgar, S. Federhen, L. Geer, W. Helmberg, Y. Kapustin, D. Kenton, O. Khovayko, D. Lipman, T. Madden, D. Maglott, J. Ostell, K. Pruitt, G. Schuler, L. Schriml, E. Sequeira, S. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T. Suzek, R. Tatusov, T. Tatusova, L. Wagner, E. Yaschenko, Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry, BMC Bioinforma. 2007 81 (40) (2007) 323–324.

[55] E.L. Schymanski, H.P. Singer, P. Longrée, M. Loos, M. Ruff, M.A. Stravs, C. Ripollés Vidal, J. Hollender, Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry, Environ. Sci. Technol. 48 (2014) 1811–1818.

[56] R. Bade, N.I. Rousis, L. Bijlsma, E. Gracia-Lor, S. Castiglioni, J.V. Sancho, F. Hernandez, Screening of pharmaceuticals and illicit drugs in wastewater and surface waters of Spain and Italy by high resolution mass spectrometry using UHPLC-QTOF MS and LC-LTQ-Orbitrap MS, Anal. Bioanal. Chem. 407 (2015) 8979–8988.

[57] A.V. Melnik, R.R. da Silva, E.R. Hyde, A.A. Aksenov, F. Vargas, A. Bouslimani, I. Protsyuk, A.K. Jarmusch, A. Tripathi, T. Alexandrov, R. Knight, P.C. Dorrestein, Coupling targeted and untargeted mass spectrometry for metabolome-microbiome-wide association studies of human fecal samples, Anal. Chem. 89 (2017) 7549–7559.

[58] The Norman Early Warning System (NormaNEWS)|NORMAN, 2012. http://www.norman-network.com/?q=node/244 (accessed September 12, 2017).

[59] Chemistry Dashboard|Home, 2017. https://comptox.epa.gov/dashboard (accessed September 5, 2017).

[60] T. Fuhrer, N. Zamboni, High-throughput discovery metabolomics, Curr. Opin. Biotechnol. 31 (2015) 73–78.

[61] D.C. Sévin, A. Kuehne, N. Zamboni, U. Sauer, Biological insights through nontargeted metabolomics, Curr. Opin. Biotechnol. 34 (2015) 1–8.

[62] Dark Matter|CERN, 2017. http://home.cern/about/physics/dark-matter (accessed August 23, 2017).

[63] 3 Things We Know and Don't Know about Dark Matter, 2016. https://newscenter.lbl.gov/2016/05/24/3-knowns-3-unknowns-dark-matter/ (accessed August 23, 2017).

[64] Dark Energy, Dark Matter|Science Mission Directorate, 2017. https://science.nasa.gov/astrophysics/focus-areas/what-is-dark-energy/ (accessed August 23, 2017).

[65] C. Lok, Mining the microbial dark matter, Nature 522 (2015) 270–273.

[66] R.R. da Silva, P.C. Dorrestein, R.A. Quinn, Illuminating the dark matter in metabolomics, Proc. Natl. Acad. Sci. U. S. A 112 (2015) 12549–12550.

[67] N.G. Mahieu, G.J. Patti, Systems-level annotation of a metabolomics data set reduces 25 000 features to fewer than 1000 unique metabolites, Anal. Chem. 89 (19) (2017) 10397–10406.

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

[69] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa, KEGG: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 27 (1999) 29–34.

[70] KEGG, Kyoto Encyclopedia of Genes and Genomes, KEGG, 1995. http://www.genome.jp/kegg/ (accessed September 12, 2017).

[71] D.S. Wishart, T. Jewison, A.C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, A. Scalbert, HMDB 3.0-the human metabolome database in 2013, Nucleic Acids Res. 41 (2013) D801–D807.

[72] Human Metabolome Database, 2006. http://www.hmdb.ca/ (accessed September 12, 2017).

[73] D.H. Rumsfeld, Defense.gov News Transcript: DoD News Briefing - Secretary Rumsfeld and Gen. Myers, 2002. http://www.defense.gov/transcripts/transcript.aspx?transcriptid=2636 (accessed September 29, 2017).

[74] J.L. Little, C. Cleven, S.D. Brown, Identification of Known Unknowns Utilizing Accurate Mass Data and Chemical Absracts Service Databases, 2011, pp. 1–12.

[75] S. Stein, Mass Spectral Reference Libraries: an Ever-expanding Resource for Chemical Identification, vol. 84, 2012, pp. 7274–7282, 17.

[76] J. Fritz, M. Desai, P. Shah, J. Schneider, P. Wilmes, From meta-omics to causality: experimental models for human microbiome research, Microbiome 1 (2013) 14.

[77] S. Di Caro, H. Tao, A. Grillo, C. Elia, G. Gasbarrini, A.R. Sepulveda, A. Gasbarrini, Effects of lactobacillus GG on genes expression pattern in small bowel mucosa, Dig. Liver Dis. 37 (2017) 320–329.

[78] P. van Baarlen, F. Troost, C. van der Meer, G. Hooiveld, M. Boekschoten, R.J.M. Brummer, M. Kleerebezem, Human mucosal in vivo transcriptome responses to three lactobacilli indicate how probiotics may modulate human cellular pathways, Proc. Natl. Acad. Sci. Unit. States Am. 108 (2011) 4562–4569.

[79] J.Z.H. von Martels, M. Sadaghian Sadabad, A.R. Bourgonje, T. Blokzijl, G. Dijkstra, K.N. Faber, H.J.M. Harmsen, The role of gut microbiota in health and disease: in vitro modeling of host-microbe interactions at the aerobe-anaerobe interphase of the human gut, Anaerobe 44 (2017) 3–12.

[80] P. Shah, J.V. Fritz, E. Glaab, M.S. Desai, K. Greenhalgh, A. Frachet, M. Niegowska, M. Estes, C. Jäger, C. Seguin-Devaux, F. Zenhausern, P. Wilmes, A microfluidics-based in vitro model of the gastrointestinal human–microbe interface, Nat. Commun. 7 (2016) 11535.

[81] K. Hiller, J. Hangebrauk, C. Jäger, J. Spura, K. Schreiber, D. Schomburg, Metabolite detector: comprehensive analysis tool for targeted and non-targeted GC/MS based metabolome analysis, Anal. Chem. 81 (2009) 3429–3439.

[82] M. Meringer, E.L. Schymanski, Small molecule identification with MOLGEN and mass spectrometry, Metabolites 3 (2013) 440–462.

[83] T. Kind, O. Fiehn, Advances in structure elucidation of small molecules using mass spectrometry, Bioanal. Rev. 2 (2010) 23–60.

[84] A.A. Aksenov, R. da Silva, R. Knight, N.P. Lopes, P.C. Dorrestein, Global chemical analysis of biology by mass spectrometry, Nat. Rev. Chem. 1 (2017) 54.

[85] K. Mansouri, A. Abdelaziz, A. Rybacka, A. Roncaglioni, A. Tropsha, A. Varnek, A. Zakharov, A. Worth, A.M. Richard, C.M. Grulke, D. Trisciuzzi, D. Fourches, D. Horvath, E. Benfenati, E. Muratov, E.B. Wedebye, F. Grisoni, G.F. Mangiatordi, G.M. Incisivo, H. Hong, H.W. Ng, I.V. Tetko, I. Balabin, J. Kancherla, J. Shen, J. Burton, M. Nicklaus, M. Cassotti, N.G. Nikolov, O. Nicolotti, P.L. Andersson, Q. Zang, R. Politi, R.D. Beger, R. Todeschini, R. Huang, S. Farag, S.A. Rosenberg, S. Slavov, X. Hu, R.S. Judson, CERAPP: collaborative estrogen receptor activity prediction project, Environ. Health Perspect. 124 (2016) 1023–1033.

[86] E.L. Schymanski, C. Ruttkies, M. Krauss, C. Brouard, T. Kind, K. Dührkop, F. Allen, A. Vaniya, D. Verdegem, S. Böcker, J. Rousu, H. Shen, H. Tsugawa, T. Sajed, O. Fiehn, B. Ghesquière, S. Neumann, Critical assessment of small molecule identification 2016: automated methods, J. Cheminform 9 (2017) 1–21.

[87] D. Dias, O. Jones, D. Beale, B. Boughton, D. Benheim, K. Kouremenos, J.-L. Wolfender, D. Wishart, Current and future perspectives on the structural identification of small molecules in biological systems, Metabolites 6 (2016) 46.

[88] E.L. Schymanski, A.J. Williams, Open science for identifying "known unknown" chemicals, Environ. Sci. Technol. 51 (2017) 5357–5359.

[89] The PubChem Project, 2004. https://pubchem.ncbi.nlm.nih.gov/ (accessed September 5, 2017).

[90] S. Kim, P.A. Thiessen, E.E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B.A. Shoemaker, J. Wang, B. Yu, J. Zhang, S.H. Bryant, Pub chem substance and compound databases, Nucleic Acids Res. 44 (2016) D1202–D1213.

[91] H.E. Pence, A. Williams, ChemSpider: an online chemical information resource, J. Chem. Educ. 87 (2010) 1123–1124.

[92] ChemSpider|Search and Share Chemistry, 2015. http://www.chemspider.com/ (accessed September 5, 2017).

[93] A.J. Williams, C.M. Grulke, J. Edwards, A.D. McEachran, K. Mansouri, N.C. Baker, G. Patlewicz, I. Shah, J.F. Wambaugh, R.S. Judson, A.M. Richard, The CompTox Chemistry Dashboard: a community data resource for environmental chemistry, J. Cheminform 9 (2017) 61.

[94] D.S. Wishart, D. Tzur, C. Knox, R. Eisner, A.C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M.A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D.D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G.E. Duggan, G.D. MacInnis, A.M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B.D. Sykes, H.J. Vogel, L. Querengesser, HMDB: the human metabolome database, Nucleic Acids Res. 35 (2007) D521–D526.

[95] Wiley, Wiley Registry 11th Edition/NIST 2014 Mass Spectral Library (DVD), Wiley, NIST, 2016. http://eu.wiley.com/WileyCDA/WileyTitle/productCd-1119284228.html (accessed September 5, 2017).

[96] H.M. Oberacher, Wiley Registry of Tandem Mass Spectral Data, John Wiley, 2012. http://eu.wiley.com/WileyCDA/WileyTitle/productCd-1118037448.html (accessed September 29, 2017).

[97] NIST Standard Reference Database 1A V17|NIST, 2014. https://www.nist.gov/srd/nist-standard-reference-database-1a-. v17 (accessed September 5, 2017).

[98] GMD - the Golm Metabolome Database, 2008. http://gmd.mpimp-golm.mpg.de/ (accessed September 5, 2017).

[99] T. Kind, G. Wohlgemuth, D.Y. Lee, Y. Lu, M. Palazoglu, S. Shahbaz, O. Fiehn, FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry, Anal. Chem. 81 (2009) 10038–10048.

[100] mzCloud − Advanced Mass Spectral Database, 2013. https://www.mzcloud.org/ (accessed September 5, 2017).

[101] C.A. Smith, G. O'Maille, E.J. Want, C. Qin, S.A. Trauger, T.R. Brandon, D.E. Custodio, R. Abagyan, G. Siuzdak, METLIN: a metabolite mass spectral database, Ther. Drug Monit. 27 (2005) 747–751. http://www.ncbi.nlm.nih.

gov/pubmed/16404815 (accessed September 5, 2017).

[102] METLIN, 2017. https://metlin.scripps.edu/landing_page.php?pgcontent=mainPage (accessed September 5, 2017).

[103] MassBank of North America, 2017. http://mona.fiehnlab.ucdavis.edu/ (accessed September 5, 2017).

[104] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M.Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, T. Nishioka, MassBank: a public repository for sharing mass spectral data for life sciences, J. Mass Spectrom. 45 (2010) 703–714.

[105] MassBank|European MassBank (NORMAN MassBank) Mass Spectral Database, 2006. https://massbank.eu/MassBank/ (accessed September 5, 2017).

[106] MS for ID, 2009. https://msforid.com/about/ (accessed September 5, 2017).

[107] M. Wang, J.J. Carver, V.V. Phelan, L.M. Sanchez, N. Garg, Y. Peng, D.D. Nguyen, J. Watrous, C.A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V Melnik, M.J. Meehan, W.-T. Liu, M. Crüsemann, P.D. Boudreau, E. Esquenazi, M. Sandoval-Calderón, R.D. Kersten, L.A. Pace, R.A. Quinn, K.R. Duncan, C.-C. Hsu, D.J. Floros, R.G. Gavilan, K. Kleigrewe, T. Northen, R.J. Dutton, D. Parrot, E.E. Carlson, B. Aigle, C.F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B.T. Murphy, L. Gerwick, C.-C. Liaw, Y.-L. Yang, H.-U. Humpf, M. Maansson, R.A. Keyzers, A.C. Sims, A.R. Johnson, A.M. Sidebottom, B.E. Sedio, A. Klitgaard, C.B. Larson, C.A.B.P.D. Torres-Mendoza, D.J. Gonzalez, D.B. Silva, L.M. Marques, D.P. Demarque, E. Pociute, E.C. O'Neill, E. Briand, E.J.N. Helfrich, E.A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J.J. Kharbush, Y. Zeng, J.A. Vorholt, K.L. Kurita, P. Charusanti, K.L. McPhail, K.F. Nielsen, L. Vuong, M. Elfeki, M.F. Traxler, N. Engene, N. Koyama, O.B. Vining, R. Baric, R.R. Silva, S.J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P.G. Williams, J. Dai, R. Neupane, J. Gurr, A.M.C. Rodríguez, A. Lamsa, C. Zhang, K. Dorrestein, B.M. Duggan, J. Almaliti, P.-M. Allard, P. Phapale, L.-F. Nothias, T. Alexandrov, M. Litaudon, J.-L. Wolfender, J.E. Kyle, T.O. Metz, T. Peryea, D.-T. Nguyen, D. VanLeer, P. Shinn, A. Jadhav, R. Müller, K.M. Waters, W. Shi, X. Liu, L. Zhang, R. Knight, P.R. Jensen, B.O. Palsson, K. Pogliano, R.G. Linington, M. Gutiérrez, N.P. Lopes, W.H. Gerwick, B.S. Moore, P.C. Dorrestein, N. Bandeira, Sharing and community curation of mass spectrometry data with global natural products social molecular networking, Nat. Biotechnol. 34 (2016) 828–837.

[108] N. Shahaf, I. Rogachev, U. Heinig, S. Meir, S. Malitsky, M. Battat, H. Wyner, S. Zheng, R. Wehrens, A. Aharoni, The WEIZMASS spectral library for high-confidence metabolite identification, Nat. Commun. 7 (2016) 12423.

[109] A. Kaufmann, P. Butcher, K. Maden, S. Walker, M. Widmer, Using in silico fragmentation to improve routine residue screening in complex matrices, J. Am. Soc. Mass Spectrom. (2017) 1–11.

[110] N. Alygizakis, S. Samanipour, J. Hollender, M. Ibáñez, S. Kaserzon, V. Kokkali, J. A. van Leerdam, J. Mueller, M. Pijnappels, M. Reid Malcolm, E.L. Schymanski, J. Slobodnik, N. Thomaidis, K. Thomas, Exploring The Potential Of A Global Emerging Contaminant Early Warning Network Through The Use Of Restrospective Suspect Screening With High-Resolution Mass Spectrometry, (submitted, 2017).

[111] C. Ruttkies, E.L. Schymanski, S. Wolf, J. Hollender, S. Neumann, MetFrag relaunched: incorporating strategies beyond in silico fragmentation, J. Cheminform 8 (2016) 1–16.

[112] MAGMa, 2017. https://www.emetabolomics.org/ (accessed September 29, 2017).

[113] K. Dührkop, H. Shen, M. Meusel, J. Rousu, S. Böcker, Searching molecular structure databases with tandem mass spectra using CSI: FingerID, Proc. Natl. Acad. Sci. U. S. A 112 (2015) 12580–12585.

[114] CFM-ID, 2017. http://cfmid.wishartlab.com/ (accessed September 29, 2017).

[115] MS-FINDER, 2017. http://prime.psc.riken.jp/Metabolomics_Software/MS-FINDER/index.html (accessed September 12, 2017).

[116] A.D. McEachran, J.R. Sobus, A.J. Williams, Identifying known unknowns using the US EPA's comptox chemistry dashboard, Anal. Bioanal. Chem. 409 (2017) 1729–1735.

[117] S. Huntscha, T.B. Hofstetter, E.L. Schymanski, S. Spahr, J. Hollender, Biotransformation of benzotriazoles: insights from transformation product identification and compound-specific isotope analysis, Environ. Sci. Technol. 48 (2014) 4435–4443.

[118] J. Wicker, T. Lorsbach, M. Gütlein, E. Schmid, D. Latino, S. Kramer, K. Fenner, enviPath – the environmental contaminant biotransformation pathway resource, Nucleic Acids Res. 44 (2016) D502–D508, https://doi.org/10.1093/nar/gkv1229.

[119] enviPath, 2017. https://envipath.org/ (accessed September 12, 2017).

[120] Z. Shi, Learning to Predict the Sites of Metabolism and Metabolic Endpoints, 2016, https://doi.org/10.7939/R3G44HW2X.

[121] S. Dimitrov, T. Pavlov, N. Dimitrova, D. Georgieva, D. Nedelcheva, A. Kesova, R. Vasilev, O. Mekenyan, Simulation of chemical metabolism for fate and hazard assessment. II CATALOGIC simulation of abiotic and microbial degradation, SAR QSAR Environ. Res. 22 (2011) 719–755, https://doi.org/10.1080/1062936X.2011.623322.

[122] MINE, 2014. http://minedatabase.mcs.anl.gov/#/home (accessed September 12, 2017).

[123] J.G. Jeffryes, R.L. Colastani, M. Elbadawi-Sidhu, T. Kind, T.D. Niehaus, L.J. Broadbelt, A.D. Hanson, O. Fiehn, K.E.J. Tyo, C.S. Henry, MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics, J. Cheminform 7 (2015) 44.

[124] J.E. Schollée, E.L. Schymanski, S.E. Avak, M. Loos, J. Hollender, Prioritizing unknown transformation products from biologically-treated wastewater using high-resolution mass spectrometry, multivariate statistics, and metabolic logic, Anal. Chem. 87 (2015) 12121–12129.

[125] J. Watrous, P. Roach, T. Alexandrov, B.S. Heath, J.Y. Yang, R.D. Kersten, M. van der Voort, K. Pogliano, H. Gross, J.M. Raaijmakers, B.S. Moore, J. Laskin, N. Bandeira, P.C. Dorrestein, Mass spectral molecular networking of living microbial colonies, Proc. Natl. Acad. Sci. U. S. A 109 (2012) E1743–E1752.

[126] M. Loos, H. Singer, Nontargeted homologue series extraction from hyphenated high resolution mass spectrometry data, J. Cheminform 9 (2017).

[127] Homologous Series Detection Eawag, 2016. http://www.envihomolog.eawag.ch/ (accessed September 12, 2017).

[128] Lipid Maps, 2017. http://www.lipidmaps.org/ (accessed September 29, 2017).

[129] H. Tsugawa, K. Ikeda, W. Tanaka, Y. Senoo, M. Arita, M. Arita, Comprehensive identification of sphingolipid species by in silico retention time and tandem mass spectral library, J. Cheminform 9 (2017).

[130] CAMERA, 2003. https://bioconductor.org/packages/release/bioc/html/CAMERA.html (accessed December 5, 2017).

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

[132] CRAN R Project, 2017. https://cran.r-project.org/web/packages/nontarget/index.html (accessed December 5, 2017).

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

[134] B. Lauper, Exploration of In-source Fragmentation of Environmental Standards and Technology, July 10, 2017. Masters Project Thesis, 75 pages, Dübendorf, Switzerland.

[135] X. Yang, P. Neta, S.E. Stein, Extending a tandem mass spectral library to include MS2 spectra of fragment ions produced in-source and MSn spectra, J. Am. Soc. Mass Spectrom. 28 (2017) 2280–2287.

[136] S.D. Dimitrov, D.G. Georgieva, T.S. Pavlov, Y.H. Karakolev, P.G. Karamertzanis, M. Rasenberg, O.G. Mekenyan, UVCB substances: methodology for structural description and application to fate and hazard assessment, Environ. Toxicol. Chem. 34 (2015) 2450–2462.

[137] CompTox-alkylbenzenesulfonate, Linear, 2017. https://comptox.epa.gov/dashboard/dsstoxdb/results?search=DTXSID3020041 (accessed September 29, 2017).

[138] CompTox-C3-C15 Sulfophenyl Carboxylates, 2017. https://comptox.epa.gov/dashboard/dsstoxdb/results?search=DTXSID90891722 (accessed September 29, 2017).

[139] K. Rosenthal, V. Oehling, C. Dusny, A. Schmid, Beyond the bulk: disclosing the life of single microbial cells, FEMS Microbiol. Rev. (2017) 751–780.

[140] R. Zenobi, Single-cell metabolomics: analytical and biological perspectives, Science 342 (2013) 1243259.

[141] P. Wilmes, B.P. Bowen, B.C. Thomas, R.S. Mueller, V.J. Denef, N.C. VerBerkmoes, R.L. Hettich, T.R. Northen, J.F. Banfield, Metabolome-proteome differentiation coupled to microbial divergence, MBio 1 (2010), e00246–10.

[142] C.R. Fischer, P. Wilmes, B.P. Bowen, T.R. Northen, J.F. Banfield, Deuterium-exchange metabolomics identifies N-methyl lyso phosphatidylethanolamines as abundant lipids in acidophilic mixed microbial communities, Metabolomics 8 (2012) 566–578.

[143] A. Heintz-Buschart, P. Wilmes, Human gut microbiome: function matters, Trends Microbiol. (2017), https://doi.org/10.1016/j.tim.2017.11.002.

B.Y. Loulou Peisl is a PhD candidate at the Luxembourg Centre for Systems Biomedicine, University of Luxembourg. Her topic is "Elucidation of previously uncharacterized metabolites involved in human-microbial cross-talk". Her main interests are the human gut microbiome and its implications on health and disease, metabolomics and parasitology. She has co-authored one publication. Her research is supported by a doctoral fellowship from the Luxembourg National Research Fund (AFR PhD Grant 11609531).

Emma Schymanski recently joined the Luxembourg Centre for Systems Biomedicine, University of Luxembourg to establish the Environmental Cheminformatics group. She has over 40 peer-reviewed publications, book chapters and a book. Her research combines cheminformatics and computational mass spectrometry approaches to elucidate the unknowns in complex samples, primarily with non-target screening. An advocate for open science, she is involved in and organizes several activities to improve the exchange of data, information and ideas between scientists. https://wwwen.uni.lu/lcsb/research/environmental_cheminformatics

Paul Wilmes is Associate Professor for Systems Ecology at the Luxembourg Centre for Systems Biomedicine of the University of Luxembourg. He has published more than 70 peer-reviewed publications. His main research interest is on unravelling fundamental ecological relationships within and between microbial populations *in situ*. His group has pioneered the development of appropriate wet- and dry-lab methodologies for carrying out systematic molecular measurements of microbial consortia. These approaches are allowing unprecedented study of microbiome-host molecular interactions. In this context, his group has also recently developed a microfluidics-based *in vitro* model of the human-microbial gastrointestinal interface called HuMiX. http://wilmeslab.uni.lu