How close are we to complete annotation of metabolomes?
Viant, Mark; Kurland, Irwin J.; Jones, Martin R.; Dunn, Warwick

DOI:
10.1016/j.cbpa.2017.01.001

License:
Creative Commons: Attribution (CC BY)

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Viant, MR, Kurland, IJ, Jones, MR & Dunn, WB 2017, 'How close are we to complete annotation of metabolomes?' Current Opinion in Chemical Biology, vol. 36, pp. 64-69. DOI: 10.1016/j.cbpa.2017.01.001

Link to publication on Research at Birmingham portal
How close are we to complete annotation of metabolomes?
Mark R Viant1,2, Irwin J Kurland3, Martin R Jones1 and Warwick B Dunn1,2

The metabolome describes the full complement of the tens to hundreds of thousands of low molecular weight metabolites present within a biological system. Identification of the metabolome is critical for discovering the maximum amount of biochemical knowledge from metabolomics datasets. Yet no exhaustive experimental characterisation of any organismal metabolome has been reported to date, dramatically contrasting with the genome sequencing of thousands of plants, animals and microbes. Here, we review the status of metabolome annotation and describe advances in the analytical methodologies being applied. In part through new international coordination, we conclude that we are now entering a new era of metabolome annotation.

Addresses
1 School of Biosciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK
2 Phenome Centre Birmingham, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK
3 Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461, USA

Corresponding author: Viant, Mark R (m.viant@bham.ac.uk)

Introduction
Metabolomics is the multidisciplinary field of research concerned with the study of metabolomes, the complement of naturally-occurring and exogenous (e.g. environmental pollutants), low-molecular-weight (typically <1500 Da) metabolites present within biological systems [1]. Comprising of precursors, intermediates and products of biochemical pathways, metabolites constitute some of the terminal products of higher cellular processes and collectively provide a ‘fingerprint’ of the complex interplay between genome and environment. From an analytical perspective, both the measurement and identification of whole metabolomes presents a considerable challenge, not least due to the vast structural heterogeneity of metabolites, their large number (e.g. an estimated 200 000 structurally-distinct secondary metabolites across the plant kingdom [2]) and their wide concentration ranges (estimated to span 12 orders of magnitude [3]). As a point of clarity, the formal definitions of metabolite annotation and identification, as developed by the Chemical Analysis Working Group of the Metabolomics Standards Initiative (MSI) [4], are shown in Table 1. The categorical scoring system defines ‘identification’ as the most rigorous (level 1) while ‘annotation’ does not require such exhaustive analytical validation (levels 2 and 3). Currently there are no completed lists of experimentally-derived metabolites that describe the metabolome of any model organism, not even as putative annotations.

Meaningful biological inferences may only be drawn from metabolomics datasets where peaks can be structurally identified as named metabolites, that is it is only when we are empowered to move beyond discussing unidentified peaks to rigorously identified metabolites that we can fully engage in describing metabolic pathways and integrate metabolism with other levels of biological hierarchy. For over a decade, molecular identification has remained the principal technical bottleneck in metabolomics [5,6]. Hence, for metabolomics to deliver its full potential in fields from medicine to ecology, innovations in analytical workflows are urgently required. Yet based on the literature, it is readily apparent that the core metabolomics workflow has changed little over the past 15 years, typically comprising of sampling, measurement of metabolites by mass spectrometry (MS) and/or nuclear magnetic resonance (NMR) spectroscopy, data processing and statistical analyses, with a view to discovering peaks of biological importance [7,8]. Those peaks are typically searched against databases, providing limited putative annotation. Rarely do investigators undertake the challenging and time consuming step of identifying peaks that are not present in databases [9], using methods that are common to natural products chemistry such as fractionation, high resolution accurate mass MS, and 1D and 2D NMR for structure determination. Typically, a significant proportion of detected peaks are not annotated or identified, dependent on the analytical platform used and sample type. Hence it is appropriate to conclude that all experimental metabolomics studies to date would have generated additional biological insights were metabolite identification a more tractable process, that in turn may have allowed for more complete metabolome network.

Current Opinion in Chemical Biology 2017, 36:64–69
Edited by Georg Pohnert and Frank C Schroeder
derivation. Metabolite identification remains a colossal challenge and a step change is needed. Here, we review the status of metabolome annotation, introduce the important role of model organisms, and describe the analytical methodologies being applied.

Can model organisms help metabolome annotation?
A critical question is how such a transformative change will occur in metabolomics, to address this more than decade long problem. We believe a combination of approaches is required, including new analytical strategies, computational algorithms and database resources, and also a concerted effort by the metabolomics community to solve this bottleneck. This latter point has recently been recognised with the formation, in 2015, of a scientific task group of the International Metabolomics Society to progress the characterisation of metabolomes by initially focusing on a few model organisms [10*]. The value of model organisms across biology and medicine is huge [11]. While seemingly disparate, research into bacteria, yeast, insects, worms, fish, rodents and plants has shown that the core biochemical operating principles have been conserved across all living organisms. Hence findings derived from non-mammalian model animals, for example can shed light on biological processes in humans (Table 2).

The Model Organism Metabolomes (MOM) task group’s philosophy is to leverage upon the critical mass of research activity and knowledge that exists for model organisms, that is to encourage the community to focus their metabolite identification efforts on systems we know the most about already (i.e. have species-specific metabolite databases [12–14]), that have sequenced genomes (hence can create genome-wide metabolic reconstructions to predict metabolism; [15]), and that when the metabolomes are successfully identified this knowledge will be of greatest value to the community [10*]. The two primary aims of the MOM task group are to integrate disparate model organism-focused research groups into an interactive community, and to share, discuss and develop the analytical and bioinformatics strategies to progress the identification of model organism metabolomes, resulting in best practice documents (Figure 1). Ultimately, this task group has set a grand challenge: to identify and map all ‘system’ metabolites onto metabolic pathways, to develop quantitative metabolic models for model organisms, and to relate organism metabolic pathways within the context of evolutionary metabolomics, that is phylogenetic metabolomics [10*]. Efforts have begun to optimise analytical methods for metabolome identification, for example in Escherichia coli [16], Saccharomyces cerevisiae [17*] and Caenorhabditis elegans [18], as well as to mine the literature for existing knowledge, for example in S. cerevisiae [19]. An atlas of tissue-specific metabolomes has also been initiated for Drosophila melanogaster, including both polar and lipophilic metabolites [20].

### Table 1

<table>
<thead>
<tr>
<th>Level of confidence</th>
<th>Description</th>
<th>Requisite analytical data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>‘Identified metabolites’</td>
<td>Two orthogonal analytical techniques applied to the analysis of both the metabolite of interest and to a chemical reference standard of suspected structural equivalence, with all analyses performed under identical analytical conditions within the same laboratory; Examples of appropriate orthogonal data: accurate mass via MS with retention time; accurate mass MS and fragmentation data or isotopic pattern; 2D NMR spectra; full 'H and/or 13C NMR spectra and so on</td>
</tr>
<tr>
<td>Level 2</td>
<td>‘Putatively annotated compounds’</td>
<td>As for levels 3 and 4, including spectral (NMR and/or MS) similarity with public or commercial libraries</td>
</tr>
<tr>
<td>Level 3</td>
<td>‘Putatively characterised compound classes’</td>
<td>As for level 4, plus spectral and/or physicochemical properties consistent with a particular class of organic compounds</td>
</tr>
<tr>
<td>Level 4</td>
<td>‘Unknown’</td>
<td>A discernible spectral signal (NMR, MS or other) that can be reproducibly detected and quantified</td>
</tr>
</tbody>
</table>

Extending our analytical strategies to progress metabolome identification
With the ambition to more deeply characterise the complete metabolomes of model organisms, what recent developments in analytical chemistry have been applied? Unlike for genomics, where disruptive technologies are relatively common [21], the analytical methods used in metabolomics have changed relatively little over the last decade. What has occurred recently is a considerable growth of targeted methods for studying swaths of metabolism, likely driven by the very frustration of limited peak annotation in non-targeted metabolomics, as discussed above. For example, a number of targeted LC–MS/MS assays have been developed to profile from a few tens to a couple of hundred metabolites in rice [22,23,24*] and mammalian samples [25,26]. While benefitting from yielding metabolic data that is identified and often quantitative, all of these studies only scrape the surface of the thousands of metabolites estimated to comprise a metabolome. Hence, to an extent, this shift to targeted assays is a distraction (except for cases where
the targets are already known) from solving the real challenge of metabolite identification and understanding biology in greater detail utilising non-targeted metabolomics.

So what is the current status of non-targeted metabolomics for fully characterising metabolomes? Both gas chromatography (GC) and liquid chromatography (LC) methods continue to be developed, including multidimensional chromatography and the application of multiple columns for the separation of different classes of metabolites. For example, a ‘broad spectrum’ GC–MS method has been developed to measure non-volatile

<table>
<thead>
<tr>
<th>Latin name</th>
<th>Common name</th>
<th>Database</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escherichia coli</td>
<td>–</td>
<td>E. coli Metabolome Database (ECMDB)</td>
<td>3755 small molecules from textbooks, scientific journals, metabolic reconstructions and electronic databases</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>Yeast</td>
<td>Yeast Metabolome Database (YMDB)</td>
<td>16 042 small molecules from textbooks, scientific journals, metabolic reconstructions and electronic databases</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>Nematode</td>
<td>Small Molecule Identifiers (SMIDs)</td>
<td>ca. 180 experimentally identified metabolites</td>
</tr>
<tr>
<td>Daphnia magna</td>
<td>Water flea</td>
<td>None currently</td>
<td>–</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>Fruit fly</td>
<td>None currently</td>
<td>–</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>Zebrafish</td>
<td>None currently</td>
<td>–</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>Mouse</td>
<td>Mouse Multiple tissue Metabolome DataBase (MMMDB)</td>
<td>Contains ca. 200 known metabolites per tissue and many unknown peaks</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>Thale cress</td>
<td>AraCyc</td>
<td>Metabolic pathway reconstruction and experimental data, contains 2802 compounds</td>
</tr>
<tr>
<td>Medicago truncatula</td>
<td>Barrel medic</td>
<td>MedicCyc 2.0</td>
<td>Metabolic pathway reconstruction, contains &gt;400 pathways with related genes, enzymes and metabolites</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>Rice</td>
<td>OryzaCyc 4.0</td>
<td>Metabolic pathway reconstruction, contains 2487 compounds</td>
</tr>
<tr>
<td>Solanum lycopersicum</td>
<td>Tomato</td>
<td>TomatoCyc 2.0</td>
<td>Metabolic pathway reconstruction, contains 2550 compounds</td>
</tr>
</tbody>
</table>

Table 2
Species-specific metabolome databases available for the model organisms prioritised for deeper investigation by the Model Organism Metabolomes task group [10]

Figure 1
Schematic workflow for the deep experimental characterisation of the metabolomes of model organisms. The creation of such knowledge in open-access metabolome databases would greatly accelerate the study of metabolism in an evolutionary context, that is *phylometabolomics*. 

Current Opinion in Chemical Biology 2017, 36:64–69  www.sciencedirect.com
metabolites in tropical fruits [27]; a total of 92 peaks were detected of which the authors identified only 45. Utilising a comprehensive GCxGC approach, coupled with head-space solid phase microextraction (HS SPME) and a time of flight (ToF) MS, Alves et al. reported the identification of 257 volatile metabolites from S. cerevisiae distributed over more than a dozen chemical families [28]. For a recent review of advanced multi-dimensional separations in mass spectrometry, see Ref. [29]. The inherent requirements of GC–MS for the thermal stability and volatility of the analytes, or derivatives thereof, means that, alone, this technique is unable to facilitate comprehensive metabolome annotation. Instead, the majority of non-targeted metabolomics studies continue to employ LC, and there is an increasing trend towards the application of several column types in a given study; for example Tufi et al. not only used two LC columns (C18 and HILIC) but also GC–MS to study a freshwater snail Lymnaea stagnalis [30]. This more comprehensive analytical approach was applied specifically to obtain a broader picture of the hydrophilic and lipophilic metabolome.

The application of multiple analytical platforms, as applied to L. stagnalis [30] is an emerging trend in metabolomics. Geier et al. [31] applied three different platforms to analyse C. elegans, including 1D 1H NMR spectroscopy, GC/MS and UPLC–MS. The deeper integration of NMR and MS data in automated metabolite identification pipelines is an emerging topic [32]. An even broader range of metabolites were measured in 31 varieties of rice using HS SPME GC–MS, primary polar metabolites by GC-ToF-MS, both polar and semi-polar compounds by 1H NMR and direct infusion MS, and multi-elemental analysis using ICP-MS [33]. While more time intensive and costly, deeper characterisation of organismal metabolomes currently requires such a multi-platform strategy. Fortunately, as long as the metabolomic knowledge is captured in relevant open access databases, such as MetaboLights [34], then this strategy only needs to be conducted once. A related project to deeply annotate the metabolome of the NIH model species Daphnia magna (waterflea) is underway in the primary authors’ laboratory, applying multiple extraction methods, LC–MS/MS and MS² methods, GC-Orbitrap MS, and 1D and 2D NMR spectroscopy. Progress has also been reported in the integration of several platforms to enable metabolite identification by UHPLC–SPE–NMR–MS [35]. In addition, approaches such as ion mobility mass spectrometry [36] and ultrahigh resolution mass spectrometry [37] hold considerable promise for contributing to metabolome annotation projects.

Another methodology that has considerable potential for aiding metabolite identification is stable isotope labelling, for example to probe the sulfur metabolome of Arabidopsis [38,39]. Isotopic ratio outlier analysis (IROA) is another isotope labelling technology that is designed to generate specific 13C isotopic patterns in metabolites for both high resolution LC–MS and GC–MS [17,40–43]. Unlike other stable isotope labelling methods, rather than utilising natural abundance and 98–99% enrichment for the control and experimental populations, respectively [44–48], IROA uses an enrichment level of 95% and 5% 13C. This leads to more observable isotopic peaks in the mass spectra in predictable and diagnostic patterns. Recent studies have demonstrated the promise of IROA for metabolic phenotyping in model organisms, including for prototrophic S. cerevisiae [17,49] and C. elegans [45]; the latter was grown in liquid culture with 13C-labeled E. coli that was first grown in M9 minimal media on either 95% or 5% 13C glucose, creating labelled C. elegans. These 95% 13C and 5% 13C glucose labelling experiments, when extracted and combined, show distinctive IROA patterns: 12C-derived molecules, 13C-derived molecules, artifacts (lack IROA patterns) and derivatives of exogenously applied compounds. Only metabolites of biological origin will have mirrored 13C and 13C metabolite peaks at the same retention time. Furthermore, the abundance of the heavy isotopologues in the 5% 13C samples (M + 1, M + 2, etc. the 13C envelope) or light isotopologues in the 95% 13C samples (M – 1, M – 2, etc. the 13C envelope), follows the binomial distribution for 13C in metabolite products based on the initial substrate enrichment. The mass difference between the 12C monoisotopic peak and the 13C monoisotopic peak indicates the number of carbons in the metabolite. Uniquely, the accurate mass IROA–GC/MS protocol developed, using both chemical ionization (CI) and electron ionization (EI), extends the information acquired from the isotopic peak patterns for molecular formulae generation. The process has been formulated as an algorithm, in which the numbers of carbons, methoximations and silylations are used as search constraints, and an accurate mass CI IROA library with retention times based on the Fiehn protocol has been published [17]. The combination of CI and EI IROA protocols affords a metabolite identification procedure that can identify co-eluting metabolites [17]. In summary, non-targeted stable isotope metabolite profiling using IROA reduces the complexity for global stable isotope metabolite identification [50], and can extend metabolome analysis by identifying ‘known unknowns’ with an IROA mirror pattern, and generating the number of carbons in the unknown metabolite.

Conclusions
The hugely beneficial impact of the Human Genome Project on 21st century science is undeniable [51,52]. No such large-scale experimental characterisations of organism metabolomes have been reported and many of the studies published to date describe only a fraction of the estimated size of a metabolome. That said, efforts are now underway from text mining to novel experimental approaches that offer to accelerate this process, and coordination of some of these activities is being achieved
through the Metabolomics Society's task group. Activity in metabolome identification is therefore expected to increase over the next couple of years with significant returns on this investment within 5–10 years. Looking further ahead, challenges will include developing analytical strategies to quantify several thousand known metabolites (simultaneously) as well as the spatial localisation of these compounds, for example using MS imaging [53,54]. Ultimately, a better understanding of the parts list is going to facilitate growth of several fields, including phylometabolomics, the study of organism metabolic pathways in the context of evolution.

Acknowledgements
This work was in part supported by the UK Natural Environment Research Council (NE/J017442/1), Thermo Fisher Scientific (CASE award) and by an Einstein-Mount Sinai Diabetes Research Center Grant (P60DK020541).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest


10. Edison AS, Hall RD, Junot C, Karp PD, Kurland IJ, Mistrik R, Reed LK, Saito K, Salek RM, Steinbeck C et al.: The time is right to focus on model organism metabolomes. Metabolites 2016, 6. Describes the launch of the first internationally coordinated action to accelerate research into the deep characterisation of model organism metabolomes. A team of scientists have laid down the challenge “to identify and map all ‘system’ metabolites onto metabolic pathways, to develop quantitative metabolic models for model organisms, and to relate organism metabolic pathways within the context of evolutionary metabolomics.”


First report using IROA technology in combination with accurate mass GC/TOF-MS, used to examine the S. cerevisiae metabolome. An accurate mass Cl IROA library containing 126 metabolites with retention times based on the Fiehn protocol was established. The combination of Cl and EI IROA protocols identifies co-eluting metabolites and differentiates metabolite spectra from artifacts, and “known unknown” metabolites were easily separated, with the number of carbons identified from the IROA patterns.


24. Schafer M, Bruttig C, Baldwin IT, Kallenbach M: High-throughput quantification of more than 100 primary- and secondary-metabolites, and phytohormones by a single solid-phase extraction based sample preparation with analysis by UHPLC−HESI−MS/MS. Plant Methods 2016, 12.

Excellent demonstration of the quantitative capabilities of targeted metabolite analysis, here assigning more than 100 primary and secondary metabolites simultaneously following a single extraction of plant material. Such an approach is highly recommended when the metabolite ‘targets’ are already known, and hence non-targeted metabolomics is not required.

brain and serum using hydrophilic interaction ultra-
high-performance liquid chromatography–tandem mass


31. Good example of the application of several separation techniques—both GC and LC—with the latter using two different column chemistries (C18 and HILIC). This approach was used to increased the breadth of experimental measurement of the snail metabolome.


33. Good example of the application of multiple analytical platforms to enable a deeper characterisation of this important model organism’s metabolome. Multiple tissue extraction strategies were examined and an 80% methanol solution was determined to be the best.


An exemplary paper, demonstrating through hierarchical cluster analysis, both the complementarity and distinct merits of multiple analytical platforms in the study of rice variant metabolome constituents. Important data is interrogated in a way that aligns well with the vision of the MOM task group, i.e. phylometabolomic comparisons, for a crop species of fundamental societal importance.


