Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics

Romanas Chaleckis[1,2], Isabel Meister[1,2], Pei Zhang[1,2] and Craig E Wheelock[1,2]

Accurate annotation is vital for data interpretation; however, metabolite identification is a major bottleneck in untargeted metabolomics. Although community guidelines for metabolite identification were published over a decade ago, adaptation of the recommended standards has been limited. The complexity of LC–MS data due to combinations of various chromatographic and mass spectrometric acquisition methods has resulted in the advent of diverse workflows, which often involve non-standardized manual curation. Herein, we review the parameters involved in metabolite reporting and provide a workflow to estimate the level of confidence in reported metabolite annotation. The future of metabolite identification will be heavily based upon the use of metabolome data repositories and associated data analysis tools, which will enable data to be shared, re-analyzed and re-annotated in an automated fashion.

Addresses
[1] Gunma University Initiative for Advanced Research (GIAR), Gunma University, Gunma, Japan
[2] Division of Physiological Chemistry II, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden

Corresponding author: Wheelock, Craig E (craig.wheelock@ki.se)

Current Opinion in Biotechnology 2019, **55**:44–50

This review comes from a themed issue on **Analytical biotechnology**

Edited by **Saulius Klimasauskas** and **Linas Mazutis**

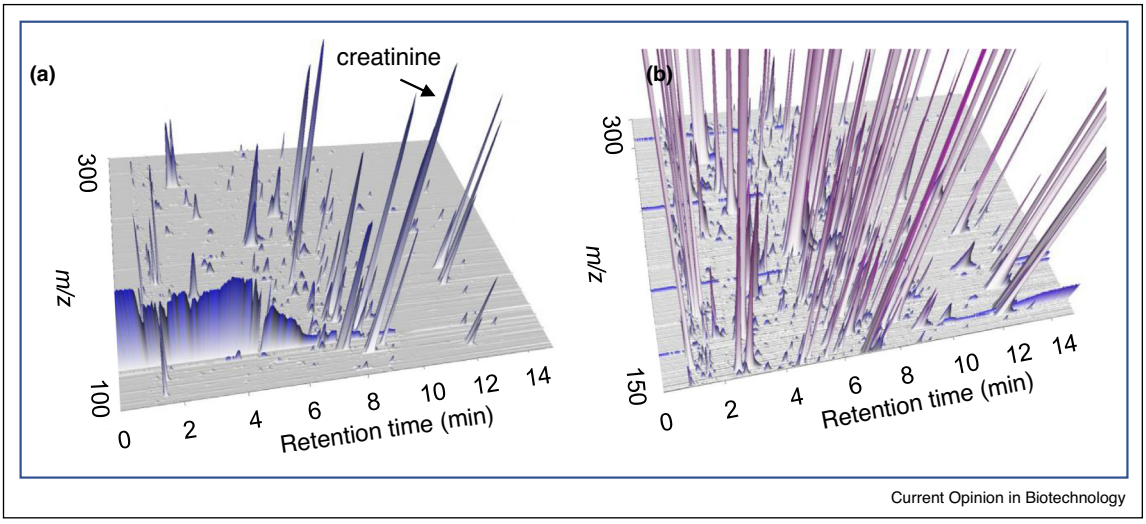
https://doi.org/10.1016/j.copbio.2018.07.010

Metabolomics and sample complexity

The metabolome refers to the complement of small molecules (usually <1500 Da [1]) present in cells, tissues and body fluids. Hundreds to thousands of well-known primary metabolites can be found across species due to shared core metabolic pathways among organisms (e.g., glycolysis, TCA cycle) [2]. However, there is a multitude of secondary metabolites that are not directly involved in the development, growth or reproduction of the organism, which display a high degree of structural variability. Many of these metabolites are produced by plants and microbes as means of communication and warfare [3]. Upon ingestion by other organisms, these metabolites are further modified resulting in a concomitant increase in the structural diversity [4]. Accordingly, the number and structural variety of metabolites constitutes a significant analytical challenge in terms of detection and annotation.

Liquid chromatography (LC) coupled to mass spectrometry (MS) has become an established technique for metabolomics studies due to high sensitivity [5–7]. Targeted approaches can measure and potentially quantify specific classes or groups of metabolites, while untargeted approaches aim to acquire as much metabolic information as possible. Although the accuracy, resolution and speed of the instrumentation for LC–MS metabolomics have improved over the last decade, covering the complexity of the metabolome remains a major challenge. An individual sample can contain thousands to tens of thousands of metabolites with diverse chemical structures of varying concentration [8]. For example, >20 000 features can be detected in a single metabolomics run; however, the number of reported metabolites is usually at least an order of magnitude lower [9••]. The experimental design, including sample extraction and choice of the LC and MS methods, inherently favors the detection of certain categories of metabolites. In addition, many of the signals detected from the LC–MS system are artifacts from sample collection and handling as well as adducts, complexes and fragments of metabolites [10]. High abundant signals are often traced back to concentrated metabolites such as glucose (4 mM in blood), creatinine (10 mM in urine (Figure 1)), salts (e.g., 140 mM sodium in blood), chemicals introduced during sampling (e.g., EDTA in plasma collection at high mM) or system contaminants (e.g., plasticizers). Individuals can evidence variable levels of specific compounds and their associated metabolites due to diet (e.g., anserine, trimethylamine-*N*-oxide), lifestyle (e.g., cotinine, caffeine), disease (e.g., 1,5-anhydroglucitol), and/or therapeutic status (e.g., paracetamol) [5,11–15]. This complexity is further confounded by variations in the individual microbiome composition. Finally, usually more than two thirds of the detected peaks are of low abundance, a fraction of which are potentially significant, but not yet reported metabolites (Figure 2). The proportion of the peaks in the different abundance categories depends very much on the sample, its preparation protocol, LC–MS method as well as data processing cut-offs and settings. Our estimation is on the optimistic side, with reported numbers being lower (e.g.,

Figure 1



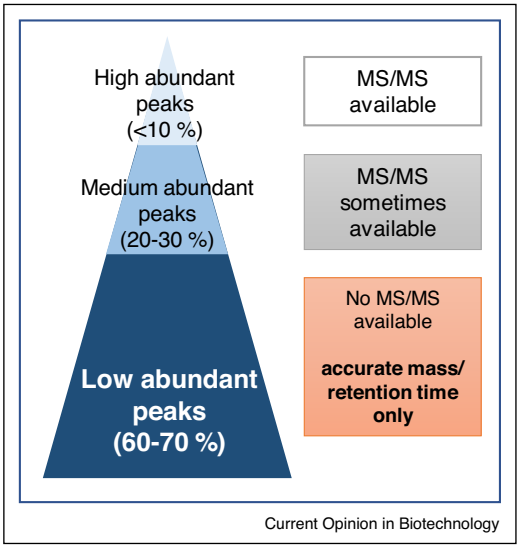
3D plots of urine LC–MS metabolomics data measured on a HILIC column in positive ionization mode as described in Naz *et al*. 2017 [32]. **(a)** Overview of the chromatographic data (RT 0–15 min, 100–300 *m/z*). The large peak indicated by the arrow is creatinine. Note the noise signals and busy regions in the chromatography. **(b)** 10-Fold zoom-in (RT 0–15 min, 150–300 *m/z*) of panel A. Plots were generated using MZmine 2 software [54]. RT: retention time.

the high, medium, and low abundant peaks are ~1, 7 and 92% in the respective categories [16]).

Metabolite annotation: importance and current standards

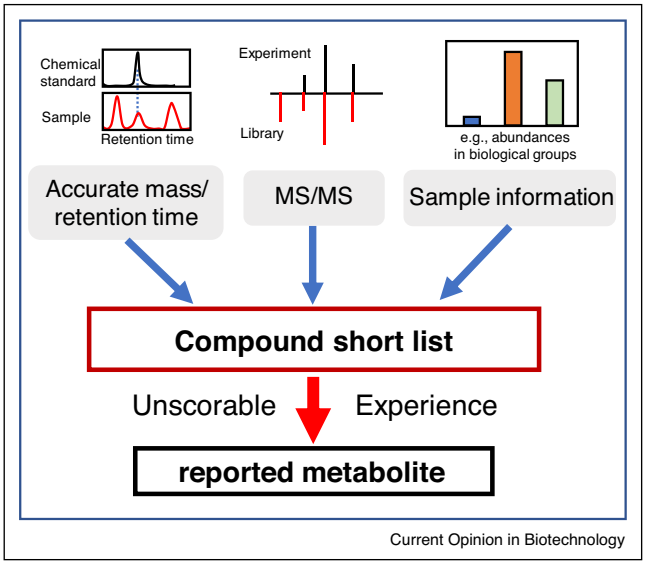
In order to convert LC–MS data into biological information, metabolites need to be annotated and the number of data processing tools is continuously growing [17•,18]. Usually a single compound identity is chosen over several other valid possibilities for enrichment and pathway analyses (e.g., CHEMRICH [19], MetaboAnalyst [20]) or integration with other omics data sets. A correct annotation therefore is pivotal. For LC–MS-based metabolomics, several criteria are used for metabolite identification, including AM (accurate mass), RT (retention time), MS/MS (fragmentation pattern), and information on the study design (Figure 3). In order to assess the confidence of an annotation, the Metabolomics community proposed defined metrics [18,21,22]. Briefly, level 0 is the strongest level of annotation and includes stereochemistry discrimination, followed by level 1 that requires the use of a chemical standard and at least two orthogonal techniques (e.g., AM and RT). Level 2 is confirmation by a class-specific standard, and level 3 by one parameter (e.g., AM), while level 4 is the feature-level without annotation. However, the existing standards are not rigorously adhered to [23] and the current level 1 description is insufficient to reflect the full range of annotation efforts beyond its minimal requirements (e.g., AMRT and MS/MS and ion ratio). Depending on the field, there have been efforts for more explicit reporting standards (e.g., plant metabolomics [24]), especially including the handling of MS/MS information for compound identification. At the metabolomics community-level; however, a coordinated update is still pending, adding confusion in an already highly technical and demanding research field. The field would benefit from increased requirements from publishers and funding agencies to follow the recommended reporting standards as well as to make the data available in public repositories.

Figure 2



Abundance distribution of the peak intensities and availability of MS/MS spectra in metabolome data.

Figure 3



A common annotation pipeline. Different pieces of information are used to compile a short list of candidate compounds. The final metabolite identities are reported following inspection and curation of the data.

Why are AM and RT not sufficient in untargeted metabolomics?

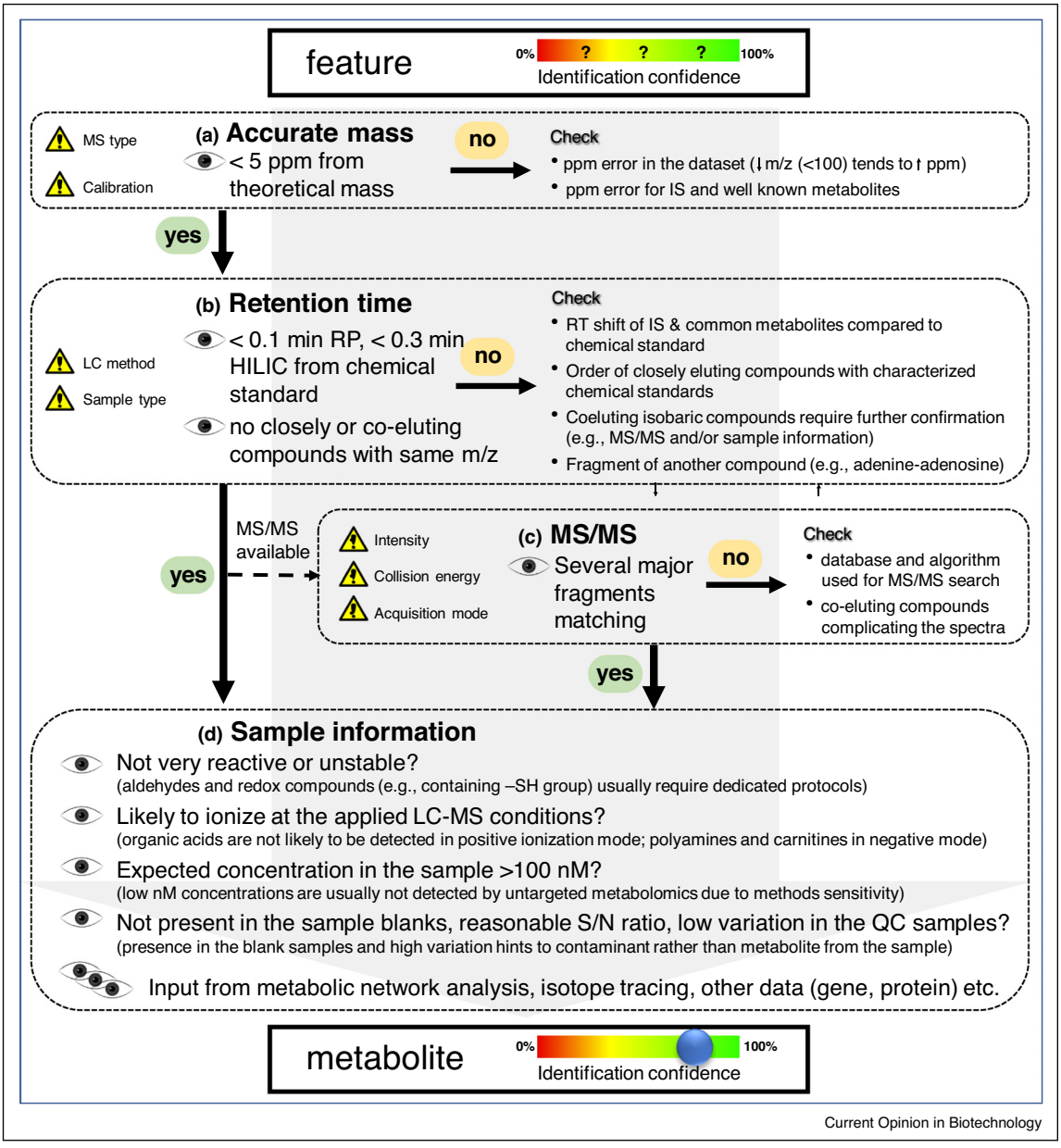
Current high-resolution Time-of-Flight and Orbitrap-based mass spectrometers have low to sub-ppm error [25], enabling high precision in estimating AM from which the elemental composition can be determined. AM is usually the core component for compound annotation (Figure 4a). However, even a sub-ppm accuracy alone is insufficient for unambiguous formula generation [26] and already a relatively simple formula such as glutamine ($C_5H_{10}N_2O_3$) can have over one million theoretical structures [27]. Secondly, for some compounds, the molecular ion that is generally used for formula generation and structure elucidation is minor, with an in-source fragment being more prominent [28•]. In the case of untargeted metabolomics, including possible in-source fragmentation and adducts dramatically expands the number of potential candidates. Finally, the routine method for formula generation, usually with the elements C, H, O, N, P and S, eventually Cl, Na and K, might be insufficient for annotating an unknown. When sampling at the population level, exogenous and rare compounds can reach μM concentrations in some samples and become detectable by metabolomics. For example, selenium-containing metabolites, such as selenoneine, have been measured in human blood due to consumption of Se-containing fish [29]. Given the diversity of secondary metabolism, especially when accounting for widely unknown microbial contributions, there are often several plausible metabolite candidates.

Annotation at level 1 confidence can be achieved if RT is included as an orthogonal parameter to MS, which requires the use of chemical standards. In order to address this need, an increasing number of commercial chemical libraries have become available (IROA, MetaSci, Merck). This has significantly reduced the resource-consuming endeavor of constructing in-house AMRT libraries. In the case where confirmation by a chemical standard is not feasible, metabolite annotation can be supported by RT predictions based upon modelling [30] and on 'projections' to similar LC methods [31•]. Among other limitations, modelling-based RT predictions usually need an already characterized chromatographic system and RT projections only work well for similar LC methods. The major challenge regarding the use of RT for untargeted metabolomics annotation is the high number of closely eluting compounds. In fact, maximizing the number of compounds that can be monitored within a single analytical run inevitably results in a crowded chromatogram (Figure 1). The problem is that RTs in some LC–MS-based methods (e.g., HILIC — hydrophilic interaction liquid chromatography) are not stable enough to provide sufficient resolution for closely eluting compounds; even minor differences in chromatographic conditions (e.g., pH, matrix effects) can result in RT shifts and affect compound elution order. Consequently, using AM and RT alone for compound annotation does not always unequivocally identify a single candidate [32,33]. As a matter of fact, the evaluation of the RT parameter for compound annotation involves not only reporting the deviation to the RT of chemical standards, but also monitoring closely and potentially co-eluting same *m/z* from multiple compounds in the sample (Figure 4b).

MS/MS strengths and limitations

By contrast to AM alone, the MS/MS fragments originating from the same molecule can be used to elucidate the chemical formula and structure and therefore contribute to the discrimination of closely eluting compounds. In data dependent targeted MS/MS approaches, ions are isolated in a narrow (1–4 Da) window before fragmentation, while data independent acquisition techniques gather MS/MS spectral information over a broader or full mass range [34]. For data independent acquisition, specific tools (e.g., MS-DIAL [35], MetDIA [36]) are needed to deconvolute the MS/MS spectra. In addition, in-source fragmentation spectra from MS1 can be obtained by CAMERA or RAMclust [37,38] providing additional annotation support. Compound identification using MS/MS spectra has been greatly facilitated by the growing number of spectral databases, algorithm improvements and user-friendly software [18,39]. For example, software such as MS-FINDER enables the user to easily interrogate several databases [40]. However, multiple challenges remain for performing a MS/MS search in spectral databases. First, each molecule has an optimal fragmentation energy, which might not be the one applied in the data

Figure 4



Step-by-step workflow to assess the confidence of compound identification. Confidence in metabolite identification is established by fulfillment of the indicated criteria. **(a)** Accurate mass is the starting point for compound identification. Higher ppm errors suggest technical issues or false annotation. **(b)** Retention times are highly dependent on the LC method used. Better chromatographic separation provides increased confidence in the annotation. **(c)** MS/MS spectra are a powerful way to verify the identity of a compound. The comparison depends on many factors, therefore in practice an overlap of several major fragments between library and experiment can be used as minimum requirement. **(d)** Sample information can provide important information for strengthening the confidence in compound annotation. Metabolite identification criteria are indicated by the symbol of the eye, and the caution symbol indicates experimental parameters that should be considered when annotating metabolites. *Abbreviations*: MS, mass spectrometry; LC, liquid chromatography; IS, internal standard (e.g., isotopically labeled amino acid); RP, reversed phase; HILIC, hydrophilic interaction liquid chromatography; S/N, signal to noise ratio; QC, quality control.

acquisition. In other words, the further the experimental fragmentation energy from the optimum, the less informative the spectra. Moreover, small molecules (<150 Da) often generate few and rather generic fragments making the interpretation difficult. Finally, the majority of the features in the LC–MS data are of low abundance (Figure 3), posing a challenge to obtain a meaningful MS/MS spectrum. In addition, acquiring MS/MS data reduces the time available for the MS1 scans thereby decreasing the overall method sensitivity [41]. For smaller scale studies

with abundant sample material, LC–MS measurements can be repeated several times with different LC–MS methods, but for large-scale studies methods with simultaneous MS/MS spectra acquisition are preferable [32,41]. Alternatively, only MS1 data are gathered in study samples to increase the sensitivity and in addition MS/MS data are only obtained from pooled quality control samples. In this case, the MS/MS spectra are averaged over the entire study and do not necessarily reflect the MS/MS spectra in individual samples. Therefore, in order to obtain usable MS/MS spectra for identifications, compound separation is important as well as striking a balance between sensitivity and specificity.

Another major issue in MS/MS annotation is the lack of reporting standards for spectral matching combined with the complexity of parameters to report. First, a scoring for the quality of the spectral match between experimental and database fragments should be provided. A scoring might already be provided within each identification software, but it varies according to the scoring formulas specific to each tool. For example, the number of matched fragments versus the match in fragment intensities might be weighted differently, hence complicating the evaluation of the spectral match quality. Also, the extent of coverage used for spectral matching should be reported. For a simple AM database match, smaller customized databases simplify the annotation by reducing the choices and thereby reducing false discovery rates [42] — at the cost of missing less common compounds [43]. Therefore, the choice of database to interrogate the data influences the list of potential candidates for annotation and only recently FDR (false discovery rate) estimation methods have been developed [44•]. Finally, the use of real-data and/or *in silico* databases for spectral matching will also have an impact on the annotation confidence. In fact, new MS/MS identification tools are regularly benchmarked by the CASMI challenge [45], with the combination of real-data repositories and *in silico* fragmentation algorithms currently providing the best results [46]. However, reporting the source and type of spectra that were used for annotating each metabolite is another level of complexity. In practice, to evaluate the compound annotation, the number of several major MS/MS fragments matching with library spectra is important (Figure 4c); in the case of co-eluting compounds, there is increased confidence if a fragment specific for a particular compound is found.

Many aspects of metabolomics data processing are understood and multiple tools have been developed [17•,18]. Although the standard scenarios are served well with current software, many of the more uncommon applications require makeshift solutions and manual curation. For compound annotation, it is especially challenging to incorporate additional information due to study-to-study variation in sample preparation and LC–MS techniques (Figure 4d). In consequence, the quality of the curations depends on the background and the experience of the curator (Figure 2). In practice, however, assessing the accuracy of annotation is a daunting task. There is a need to provide clear parameters for annotation confidence for each reported compound. In Figure 4, we provide a step-by-step workflow to assess the parameters used in the annotation process. These parameters can be extracted from the raw data and should be made readily available to the research community for all published studies.

Annotation, the way forward

It remains a challenge to annotate novel compounds and low abundant compounds with no usable MS/MS data available in addition to AM and RT. Therefore, developing targeted extraction protocols for particular metabolite classes together with specific depletion of highly abundant compounds (e.g., molecular imprinting [47]) could be a promising approach. In addition, the gap in unknown compounds might be closing as more enzymatic functions are understood. Of note, it is estimated that >600 and >2000 protein enzymatic functions remain to be characterized in budding yeast and humans, respectively [48]. Therefore, integrating the knowledge of metabolic networks with database searches [49] should provide more powerful identification tools. Databases used for compound identification keep growing. For example, the Human Metabolome Database (HMDB) currently contains >100 000 compounds [50] and PubChem consists of 100 million compounds. Periodic re-annotation cycles as new compounds are discovered might result in additional information, provided that the raw data is available to the community.

The continued improvements and automation of data analysis workflows are expected to result in essentially all data (re)processing being performed in the cloud. A clear example of this future is the application of XCMS online with support of artificial intelligence [41,51••,52]. Open spectral repositories with downloadable spectra and straightforward submission system (e.g., MassBank and MoNa (MassBank of North America)) will play a key role in facilitating such community efforts. For example, experimental MS/MS data together with RT for thousands of compounds using different chromatography are readily available for download at MoNA. In LC–MS, integrating multiple studies analyzed with similar methods should facilitate annotation, such as BinBase for GC–MS data [53••]. Therefore, raw data sharing opportunities (e.g., Metabolights, Metabolomics Workbench) combined with automated processing and easy to report parameters should assist in reporting the annotation confidence level. In the coming years, we expect the metabolite annotation to become less of a burden to researchers due to improvements in software tools and the growth of databases with the contribution of the whole research community.

Acknowledgements

We thank the Wheelock lab members and Tada Ipputa (National Institute of Genetics, Mishima, Japan) for the discussion and suggestions. We acknowledge support from the Gunma University Initiative for Advanced Research (GIAR), the Swedish Heart Lung Foundation (HLF 20170734, HLF 20170603), and the Swedish Research Council (2016-02798). This work was supported in part by The Environment Research and Technology Development Fund (ERTDF) (Grant No 5-1752). IM was supported by Japan Society for the Promotion of Science (JSPS) postdoctoral fellowship (P17774).

References and recommended reading

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

- • of special interest
- •• of outstanding interest

1. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S *et al.*: **HMDB: the Human Metabolome Database**. *Nucleic Acids Res* 2007, **35**:D521-6.

2. DeBerardinis RJ, Thompson CB: **Cellular metabolism and disease: what do metabolic outliers teach us?** *Cell* 2012, **148**:1132-1144.

3. Demain AL, Fang A: *The Natural Functions of Secondary Metabolites*. Berlin, Heidelberg: Springer; 2000 http://dx.doi.org/10.1007/3-540-44964-7_1 http://link.springer.com/10.1007/3-540-44964-7_1.

4. Scalbert A, Brennan L, Manach C, Andres-Lacueva C, Dragsted LO, Draper J, Rappaport SM, van der Hooft JJJ, Wishart DS: **The food metabolome: a window over dietary exposure**. *Am J Clin Nutr* 2014, **99**:1286-1308.

5. Dunn WB, Lin W, Broadhurst D, Begley P, Brown M, Zelena E, Vaughan AA, Halsall A, Harding N, Knowles JD *et al.*: **Molecular phenotyping of a UK population: defining the human serum metabolome**. *Metabolomics* 2015, **11**:9-26.

6. Guo L, Milburn MV, Ryals JA, Lonergan SC, Mitchell MW, Wulff JE, Alexander DC, Evans AM, Bridgewater B, Miller L *et al.*: **Plasma metabolomic profiles enhance precision medicine for volunteers of normal health**. *Proc Natl Acad Sci U S A* 2015, **112**:E4901-E4910.

7. Tadaka S, Saigusa D, Motoike IN, Inoue J, Aoki Y, Shirota M, Koshiba S, Yamamoto M, Kinoshita K: **jMorp: Japanese multi omics reference panel**. *Nucleic Acids Res* 2018, **46**:D551-D557.

8. Cajka T, Fiehn O: **Toward merging untargeted and targeted methods in mass spectrometry-based metabolomics and lipidomics**. *Anal Chem* 2016, **88**:524-545.

9. •• Mahieu NG, Patti GJ: **Systems-level annotation of a metabolomics data set reduces 25 000 features to fewer than 1000 unique metabolites**. *Anal Chem* 2017, **89**:10397-10406.
Using 13C-based credentialing in*E. coli*, the authors demonstrate that the number of unique metabolites is at least an order of magnitude smaller than that of the detected features. An online database was setup to share the results.

10. Xu Y-F, Lu W, Rabinowitz JD: **Avoiding misannotation of in-source fragmentation products as cellular metabolites in liquid chromatography–mass spectrometry-based metabolomics**. *Anal Chem* 2015, **87**:2273-2281.

11. Psychogios N, Hau DD, Peng J, Guo AC, Mandal R, Bouatra S, Sinelnikov I, Krishnamurthy R, Eisner R, Gautam B *et al.*: **The human serum metabolome**. *PLoS One* 2011, **6**:e16957.

12. Bouatra S, Aziat F, Mandal R, Guo AC, Wilson MR, Knox C, Bjorndahl TC, Krishnamurthy R, Saleem F, Liu P *et al.*: **The human urine metabolome**. *PLoS One* 2013, **8**:e73076.

13. Halama A, Kulinski M, Kader SA, Satheesh NJ, Abou-Samra AB, Suhre K, Mohammad RM: **Measurement of 1,5-anhydroglucitol in blood and saliva: from non-targeted metabolomics to biochemical assay**. *J Transl Med* 2016, **14**:140.

14. Cheung W, Keski-Rahkonen P, Assi N, Ferrari P, Freisling H, Rinaldi S, Slimani N, Zamora-Ros R, Rundle M, Frost G *et al.*: **A metabolomic study of biomarkers of meat and fish intake**. *Am J Clin Nutr* 2017, **105**:600-608.

15. Huang Y, Hui Q, Walker DI, Uppal K, Goldberg J, Jones DP, Vaccarino V, Sun YV: **Untargeted metabolomics reveals multiple metabolites influencing smoking-related DNA methylation**. *Epigenomics* 2018, **1**:116.

16. Chaleckis R, Ebe M, Pluskal T, Murakami I, Kondoh H, Yanagida M: **Unexpected similarities between the Schizosaccharomyces and human blood metabolomes, and novel human metabolites**. *Mol Biosyst* 2014, **10**:2538-2551.

17. • Misra BB: **New tools and resources in metabolomics: 2016–2017**. *Electrophoresis* 2018, **39**:909-923.
One of the most recent reviews on the tools and resources available for metabolomics. An extensive list of tools with their estimated ease of use is provided.

18. Blaenović I, Kind T, Ji J, Fiehn O: **Software tools and approaches for compound identification of LC–MS/MS data in metabolomics**. *Metabolites* 2018, **8**:31.

19. Barupal DK, Fiehn O: **Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets**. *Sci Rep* 2017, **7**:14567.

20. Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, Wishart DS, Xia J: **MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis**. *Nucleic Acids Res* 2018, **37**:W652.

21. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, Fan TW-M, Fiehn O, Goodacre R, Griffin JL *et al.*: **Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI)**. *Metabolomics* 2007, **3**:211-221.

22. Creek DJ, Dunn WB, Fiehn O, Griffin JL, Hall RD, Lei Z, Mistrik R, Neumann S, Schymanski EL, Sumner LW *et al.*: **Metabolite identification: are you sure? And how do your peers gauge your confidence?**. *Metabolomics* 2014, **10**:350-353.

23. Spicer RA, Salek R, Steinbeck C: **Compliance with minimum information guidelines in public metabolomics repositories**. *Sci Data* 2017, **4**:170137.

24. Fernie AR, Aharoni A, Willmitzer L, Stitt M, Tohge T, Kopka J, Carroll AJ, Saito K, Fraser PD, DeLuca V: **Recommendations for reporting metabolite data**. *Plant Cell* 2011, **23**:2477-2482.

25. Eichhorn P, Pérez S, Barceló D: *Time-of-Flight Mass Spectrometry Versus Orbitrap-Based Mass Spectrometry for the Screening and Identification of Drugs and Metabolites*. Elsevier; 2012.

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

27. Kerber A, Laue R, Meringer M, Rucker C: **Molecules in silico: the generation of structural formulae and its applications**. *J Comp Chem Jpn* 2004, **3**:85-96.

28. • Domingo-Almenara X, Montenegro-Burke JR, Benton HP, Siuzdak G: **Annotation: a computational solution for streamlining metabolomics analysis**. *Anal Chem* 2018, **90**:480-489.
This review details current computational annotation strategies for LC–MS data. The challenges of in-source fragmentation and overlap of MS/MS between similar compounds is well-illustrated.

29. Yamashita M, Yamashita Y, Ando T, Wakamiya J, Akiba S: **Identification and determination of selenoneine, 2-selenyl-N alpha, N alpha, N alpha-trimethyl-L-histidine, as the major organic selenium in blood cells in a fish-eating population on remote Japanese Islands**. *Biol Trace Elem Res* 2013, **156**:36-44.

30. Hall LM, Hill DW, Bugden K, Cawley S, Hall LH, Chen M-H, Grant DF: **Development of a reverse phase HPLC retention index model for nontargeted metabolomics using synthetic compounds**. *J Chem Inf Model* 2018, **58**:591-604.

31. • Stanstrup J, Neumann S, Vrhovsek U: **PredRet: prediction of retention time by direct mapping between multiple chromatographic systems**. *Anal Chem* 2015, **87**:9421-9428.

PredRet is a user-friendly tool for mapping retention times between similar chromatographic systems. The tool is integrated in an online database and in addition to the retention time prediction tool it allows to share retention times with other users.

32. Naz S, Gallart-Ayala H, Reinke SN, Mathon C, Blankley R, Chaleckis R, Wheelock CE: **Development of a liquid chromatography–high resolution mass spectrometry metabolomics method with high specificity for metabolite identification using all ion fragmentation acquisition**. *Anal Chem* 2017, **89**:7933-7942.

33. Chaleckis R, Naz S, Meister I, Wheelock CE: **LC–MS-based metabolomics of biofluids using all-ion fragmentation (AIF) acquisition**. *Methods Mol Biol* 2018, **1730**:45-58.

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

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

36. Li H, Cai Y, Guo Y, Chen F, Zhu Z-J: **MetDIA: targeted metabolite extraction of multiplexed MS/MS spectra generated by data-independent acquisition**. *Anal Chem* 2016, **88**:8757-8764.

37. Kuhl C, Tautenhahn R, Böttcher C, Larson TR, Neumann S: **CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets**. *Anal Chem* 2011, **84**:283-289.

38. Broeckling CD, Afsar FA, Neumann S, Ben-Hur A, Prenni JE: **RAMClust: a novel feature clustering method enables spectral-matching-based annotation for metabolomics data**. *Anal Chem* 2014, **86**:6812-6817.

39. Kind T, Tsugawa H, Cajka T, Ma Y, Lai Z, Mehta SS, Wohlgemuth G, Barupal DK, Showalter MR, Arita M *et al.*: **Identification of small molecules using accurate mass MS/MS search**. *Mass Spectrom Rev* 2017, **56**:1121.

40. Tsugawa H: **Advances in computational metabolomics and databases deepen the understanding of metabolisms**. *Curr Opin Biotechnol* 2018, **54**:10-17.

41. Benton HP, Ivanisevic J, Mahieu NG, Kurczy ME, Johnson CH, Franco L, Rinehart D, Valentine E, Gowda H, Ubhi BK *et al.*: **Autonomous metabolomics for rapid metabolite identification in global profiling**. *Anal Chem* 2015, **87**:884-891.

42. Matsuda F, Shinbo Y, Oikawa A, Hirai MY, Fiehn O, Kanaya S, Saito K: **Assessment of metabolome annotation quality: a method for evaluating the false discovery rate of elemental composition searches**. *PLoS One* 2009, **4**:e7490.

43. Böcker S: **Searching molecular structure databases using tandem MS data: are we there yet?** *Curr Opin Chem Biol* 2017, **36**:1-6.

44. • Scheubert K, Hufsky F, Petras D, Wang M, Nothias L-F, Dührkop K, Bandeira N, Dorrestein PC, Böcker S: **Significance estimation for large scale metabolomics annotations by spectral matching**. *Nat Commun* 2017, **8**:1494.

Describes methods for FDR estimation adjusted for each project.

45. Schymanski EL, Ruttkies C, Krauss M, Brouard C, Kind T, Dührkop K, Allen F, Vaniya A, Verdegem D, Böcker S *et al.*: **Critical assessment of small molecule identification 2016: automated methods**. *J Cheminformatics* 2017, **9**:22.

46. Blaenović I, Kind T, Torbainović H, Obrenović S, Mehta SS, Tsugawa H, Wermuth T, Schauer N, Jahn M, Biedendieck R *et al.*: **Comprehensive comparison of in silico MS/MS fragmentation tools of the CASMI contest: database boosting is needed to achieve 93% accuracy**. *J Cheminformatics* 2017, **9**:32.

47. Gao B, Li Y, Zhang Z: **Preparation and recognition performance of creatinine-imprinted material prepared with novel surface-imprinting technique**. *J Chromatogr B Analyt Technol Biomed Life Sci* 2010, **878**:2077-2086.

48. Ellens KW, Christian N, Singh C, Satagopam VP, May P, Linster CL: **Confronting the catalytic dark matter encoded by sequenced genomes**. *Nucleic Acids Res* 2017, **45**:11495-11514.

49. Alden N, Krishnan S, Porokhin V, Raju R, McElearney K, Gilbert A, Lee K: **Biologically consistent annotation of metabolomics data**. *Anal Chem* 2017 http://dx.doi.org/10.1021/acs.analchem.7b02162.

50. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, Sajed T, Johnson D, Li C, Karu N *et al.*: **HMDB 4.0: the human metabolome database for 2018**. *Nucleic Acids Res* 2017 http://dx.doi.org/10.1093/nar/gkx1089.

51. •• Forsberg EM, Huan T, Rinehart D, Benton HP, Warth B, Hilmers B, Siuzdak G: **Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online**. *Nat Protoc* 2018, **13**:633-651.

The authors introduce the XCMS online platform for metabolomics data processing and interpretation. In addition, the platform supports the integration of multi-omics data.

52. Warth B, Spangler S, Fang M, Johnson CH, Forsberg EM, Granados A, Martin RL, Domingo-Almenara X, Huan T, Rinehart D *et al.*: **Exposome-scale investigations guided by global metabolomics, pathway analysis, and cognitive computing**. *Anal Chem* 2017, **89**:11505-11513.

53. •• Lai Z, Tsugawa H, Wohlgemuth G, Mehta S, Mueller M, Zheng Y, Ogiwara A, Meissen J, Showalter M, Takeuchi K *et al.*: **Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics**. *Nat Methods* 2018, **15**:53-56.

This review highlights the utility of large data repositories for annotating metabolites. The ability to incorporate various types of study information provides suggestions for annotation.

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