




# Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics

Zijuan Lai<sup>1,2,9</sup>, Hiroshi Tsugawa<sup>3,4,9</sup> , Gert Wohlgemuth<sup>1</sup>, Sajjan Mehta<sup>1</sup> , Matthew Mueller<sup>1</sup>, Yuxuan Zheng<sup>2</sup>, Atsushi Ogiwara<sup>5</sup>, John Meissen<sup>1</sup>, Megan Showalter<sup>1</sup>, Kohei Takeuchi<sup>6</sup>, Tobias Kind<sup>1</sup>, Peter Beal<sup>2</sup>, Masanori Arita<sup>3,7</sup> & Oliver Fiehn<sup>1,8</sup> 

**Novel metabolites distinct from canonical pathways can be identified through the integration of three cheminformatics tools: BinVestigate, which queries the BinBase gas chromatography–mass spectrometry (GC-MS) metabolome database to match unknowns with biological metadata across over 110,000 samples; MS-DIAL 2.0, a software tool for chromatographic deconvolution of high-resolution GC-MS or liquid chromatography–mass spectrometry (LC-MS); and MS-FINDER 2.0, a structure-elucidation program that uses a combination of 14 metabolome databases in addition to an enzyme promiscuity library. We showcase our workflow by annotating *N*-methyl-uridine monophosphate (UMP), lysomonogalactosyl-monopalmitin, *N*-methylalanine, and two propofol derivatives.**

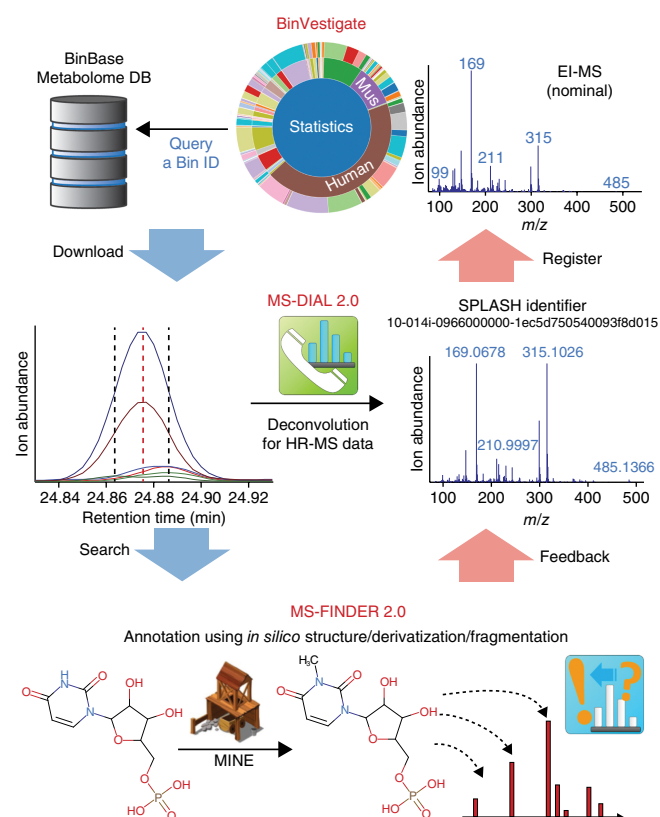
Many more unknown peaks than identified compounds are detected in untargeted metabolomics experiments, in large part because publicly available mass spectra libraries are still very small in comparison to the chemical sphere of more than 68 million known compounds<sup>1</sup>. Although GC-MS spectra have been collected systematically and in a standardized manner for more than 30 years in the NIST and Wiley libraries, which now encompass over 267,000 unique compounds, only about 40% of the reliably detectable peaks can be identified in a typical metabolomic profile. This ‘dark matter of metabolomics’<sup>2</sup> can be explained by (a) a lack of knowledge of enzymatic transformations<sup>3</sup>, including substrate promiscuity<sup>4</sup>; (b) metabolic damage by spontaneous

reactions or enzyme errors<sup>5</sup>; (c) signatures of exogenous compounds, for example, from environmental sources<sup>6</sup>; (d) the combined metabolic impact of a community of species, such as gut microbiota<sup>7</sup>; and (e) the formation of chemical artifacts during analytical protocols<sup>8</sup>. Recently, the new term ‘epimetabolite’<sup>9</sup> was suggested to encompass metabolites modified by enzymatic transformations that gain physiological functions in a biological system, similar to proteins affected by post-translational modifications. A strategy to identify epimetabolites aims at reducing the number of important (functional) unknowns by investigating multiple studies simultaneously, including cross-species analyses<sup>10</sup>. Once the origin, relevance, and specificity of these unknowns have been determined, accurate mass spectrometry and cheminformatics tools can be used to annotate and validate chemical structures.

We here present a unified method for functional and structural annotation of unknown epimetabolites (**Fig. 1**). BinBase is our large GC-MS-based untargeted metabolomics database encompassing 1,561 studies with 114,795 samples for various species, organs, matrices, and experimental conditions that have been acquired over the past 13 years<sup>11</sup>. In BinBase, 9,563 unique metabolites have been discovered so far; 1,020 of these have been identified through comparison to mass spectral libraries of authentic standards<sup>12</sup>, and 256 represent known chemical artifacts. To query biological metadata for each metabolite, we developed BinVestigate (<http://binvestigate.fiehnlab.ucdavis.edu/>), a tool that provides open access information about abundance, frequency, species, and organ origin. Our MS-DIAL 2.0 tool (<http://prime.psc.riken.jp/>) can be used to obtain deconvoluted spectra from high-resolution GC-MS data as a prerequisite for compound identification. MS-DIAL was previously developed for LC-MS data processing<sup>13</sup> but now enables processing of both LC-MS/MS and GC-MS data. Finally, unknowns can be annotated by their elemental formulas and *in silico* mass spectral fragmentation with MS-FINDER 2.0 (<http://prime.psc.riken.jp/>)<sup>14</sup>. MS-FINDER integrates structures and formulas for 224,622 known metabolites and now also includes 643,307 hypothetical compounds from the enzyme promiscuity database MINE-DB (<http://minedatabase.mcs.anl.gov/>)<sup>15</sup>. Notably, MS-DIAL 2.0 links mass spectra directly to BinVestigate and MS-FINDER 2.0; these tools are also available as stand-alone software. Here we illustrate five successful examples for this strategy, ranging from the discovery of new methylation products in mammalian and microbial cells to the identification of plant-specific metabolites and transformations of exposome compounds. The performance of the three programs, including false discovery rates in BinBase, is discussed in the Online Methods.

<sup>1</sup>West Coast Metabolomics Center, UC Davis, Davis, California, USA. <sup>2</sup>Department of Chemistry, UC Davis, Davis, California, USA. <sup>3</sup>RIKEN Center for Sustainable Resource Science, Yokohama, Japan. <sup>4</sup>RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. <sup>5</sup>Reifys Inc., Tokyo, Japan. <sup>6</sup>Perfume Development Research Laboratory, Kao Corporation, Tokyo, Japan. <sup>7</sup>National Institute of Genetics, Mishima, Japan. <sup>8</sup>Department of Biochemistry, King Abdulaziz University, Jeddah, Saudi Arabia. <sup>9</sup>These authors contributed equally to this work. Correspondence should be addressed to M.A. ([arita@nig.ac.jp](mailto:arita@nig.ac.jp)) or O.F. ([ofiehn@ucdavis.edu](mailto:ofiehn@ucdavis.edu)).

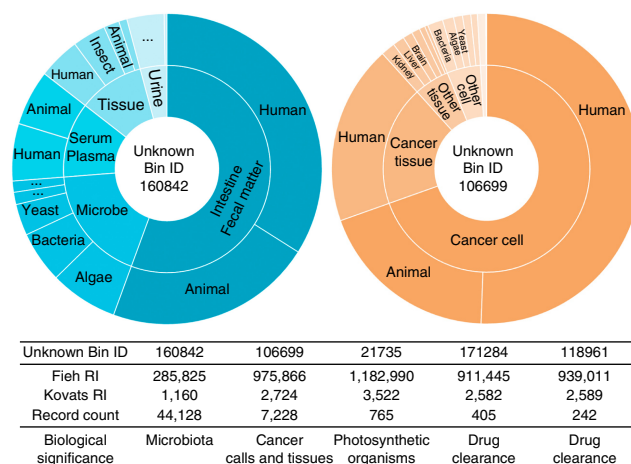
RECEIVED 31 AUGUST; ACCEPTED 26 OCTOBER; PUBLISHED ONLINE 27 NOVEMBER 2017; DOI:10.1038/NMETH.4512



**Figure 1** | A workflow for the functional and structural identification of unknown metabolites. Top, BinVestigate searches for metabolomics study metadata and (nominal) electron ionization (EI)-MS spectra in BinBase, with results shown as sunburst diagrams to illustrate the biological origin (species, organ(s), cell type(s)) of unknown compounds. Middle, MS-DIAL 2.0 carries out spectral deconvolution of unknown compounds from GC-HR-MS or LC-HR-MS/MS data for identification. Bottom, MS-FINDER 2.0 interprets spectra from GC-electrospray ionization (ESI)-MS and LC-ESI-MS/MS to annotate unknown compounds in combination with the enzyme promiscuity structure database (MINE). The tools are fully integrated via MS-DIAL. Each tool is also available as standalone program.

BinVestigate can be used to query unknowns from metabolomics studies and to prioritize and select targeted unknowns for structural identification on the basis of their cross-study specificity and relevance (Online Methods). In constructing BinBase, we followed West Coast Metabolomics Center quality control standards to keep absolute signal intensities within two-fold deviation from the mean and to avoid detector saturation, thus making intensities comparable across species and sample types. Besides the frequency of detection in specific organs and species, BinVestigate uses average signal intensities to highlight relevance across studies.

As an example, we detected an unknown BinBase metabolite (BB160842) in 44,128 samples (Fig. 2), 90% of which were from microbial, fecal, or plasma studies. Its signal intensity was five to ten times higher in human or animal fecal matter compared with that in microbial cells, and up to 20-fold higher compared with that in body fluids or tissues. This suggests a compound of microbial origin that is excreted into human plasma. Another metabolite, BB106699, was detected in 7,228 samples, most abundantly in diverse cancer cell lines and cancer tissues (Fig. 2). It had up to 100-fold higher signal intensity in myeloma cancer cell

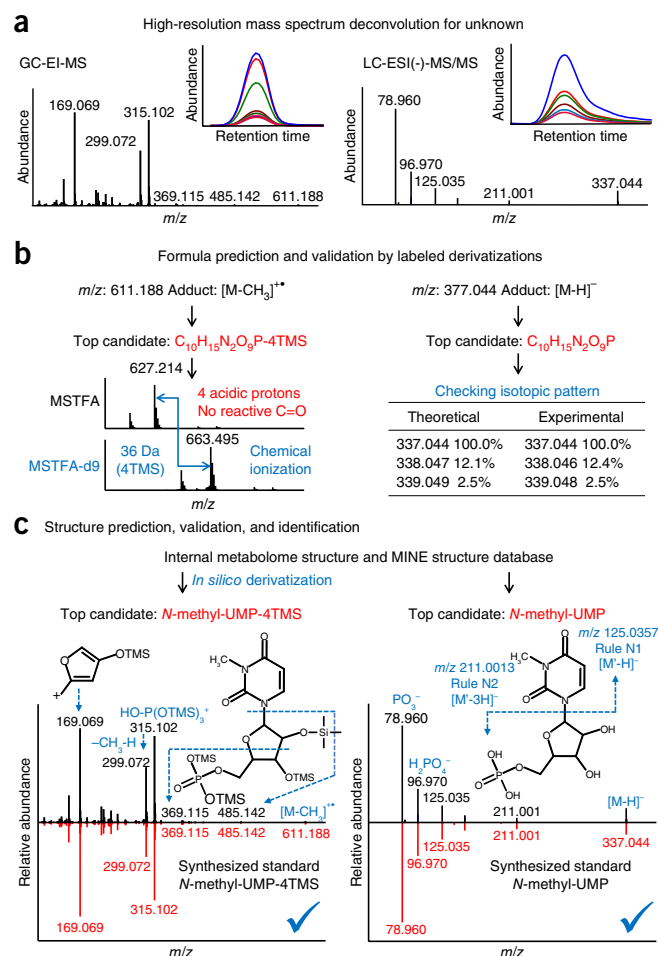


**Figure 2** | Metabolomic meta-analysis for origin exploration by BinVestigate. Unknown BinBase IDs 160842 and 106699 were queried in over 114,000 samples to determine cross-study specificity and biological relevance. In the sunburst diagrams, the area of the circular sector for each organ (inner cycle) or species (outer cycle) was mathematically determined on the basis of the average signal intensity of the unknown compound when present in such an origin. The BinBase ID, Fiehn retention index (RI), Kovats RI, number of annotation records, and biological significance for the five unknowns discussed in this work are summarized in the table at the bottom of the figure.

lines compared with that in other cell types such as mouse kidney cells. This compound was never found in fecal matter or bacterial samples, which supports the notion that it might be exclusive to eukaryotes and have a specific role in cancer. Similarly, BB21735 (Supplementary Fig. 1) was found exclusively in 765 samples of algae, marine bacteria and cyanobacteria, and plants, but never in human or animal samples, which suggests a dedicated role in the biochemistry of photosynthetic organisms. Finally, BB171284 and BB118961 were found in only two clinical cohort studies in plasma and urine from 405 and 242 samples, respectively. As both studies involved pharmaceutical treatments, these compounds seemed to be important for phase 2 drug clearance.

To identify unknown compounds with our workflow, we examined high-resolution (HR) accurate mass GC-MS data with different ionization techniques, and used LC-MS/MS for validation. Unlike LC-MS/MS metabolomics, GC-MS-based analyses lead to extensive fragmentation right at the source of ionization, even under soft chemical ionization. We therefore developed a new version of our data processing software MS-DIAL 2.0 to handle GC-MS data, including spectral deconvolution and compound identification (Supplementary Figs. 2 and 3, Supplementary Data Set 1, and Online Methods). Multiple MS data types (low-resolution MS or HR-MS, as well as GC-MS, LC-MS, and LC-MS/MS) of any major vendor or open data format (Supplementary Fig. 4) are supported. With MS-DIAL 2.0, deconvoluted spectra from GC-HR-MS and LC-HR-MS/MS are extracted from representative biological samples that contain the unknown metabolites.

In our workflow, MS-FINDER 2.0 is used next for structure elucidation of unknown HR-MS spectra, as demonstrated here for BB106699 (Fig. 3, Supplementary Figs. 5 and 6, and Online Methods). First, the molecular adduct ion is identified to pursue the chemical formula of the unknown compound. As is often



**Figure 3** | The identification of *N*-methyl-UMP by MS-DIAL 2.0 and MS-FINDER 2.0. (a) Spectral deconvolution: fragment ions and molecular adduct ions of BinBase ID 106699 were deconvoluted and confirmed via MS-DIAL 2.0. GC-HR-MS analytics were used for structure elucidation (left), then LC-MS/MS was applied as an additional evidence line to validate the discovery (right). (b) Formula prediction and validation:  $C_{10}H_{15}N_2O_9P$  was scored and ranked first in MS-FINDER 2.0 on the basis of mass errors, isotope ratio errors, and subformula assignments. Using GC-MS, we obtained chemical ionization data with different derivatization methods (MSTFA versus MSTFA-d9) to verify the formula and obtain the number of acidic protons; with LC-MS, between theoretical values and experimental values, the mass errors were only 1 mDa, and the isotopic ratio errors were within 1%. (c) Structure prediction, validation, and identification. Structure candidates were retrieved from MINE DB and the unified metabolome structure database in MS-FINDER, and *in silico*-fragmented on the basis of hydrogen-rearrangement rules, bond-dissociation energy, and a comprehensive fragmentation rule library (including GC-EI-MS and LC-ESI-MS/MS). *N*-methyl-UMP was ranked as the most likely structure in MS-FINDER 2.0 with computational assigned substructures. The mass spectra and retention times in GC-MS (left) and LC-MS/MS (right) were matched between BinBase ID 106699 in a cancer cell sample, and validated with a chemically synthesized *N*-methyl-UMP standard.

observed for trimethylsilylated (TMS) metabolites in GC-MS, the molecular ion ( $[M]^{++}$ ) was absent from the hard electron ionization spectrum. We found that its initial methyl cleavage fragment ion ( $[M-CH_3]^+$ ) was present in very low abundance, which impeded the calculation of the elemental formula for this

unknown compound (Fig. 3a). Softer methane chemical ionization GC-HR-MS yielded a pattern of additional, highly characteristic molecular adduct ions. For BB106699, the molecular mass was calculated as 626.212 Da on the basis of the alignment of ions at  $m/z$  611.118 ( $[M-CH_3]^+$ ),  $m/z$  627.214 ( $[M+H]^+$ ),  $m/z$  655.244 ( $[M+C_2H_5]^+$ ), and  $m/z$  667.245 ( $[M+C_3H_5]^+$ ). MS-FINDER 2.0 calculated  $C_{10}H_{15}N_2O_9P$  as the most probable elemental formula by using an optimized algorithm for TMS derivatives that applies both heuristic and chemical rules<sup>14</sup> (Fig. 3b). The formula was further validated by comparison of regular TMS derivatization to stable-isotope-labeled d9-trimethylsilyl, which directly yielded the number of TMS groups, or the number of acidic protons in the unknown molecule (Fig. 3b). The fact that GC-MS-based metabolomics requires derivatization allowed us to omit isomeric structures of  $C_{10}H_{15}N_2O_9P$  with fewer than four acidic protons. Moreover, a comparison of the chemical derivatization results on the basis of methoximation versus ethoximation showed that the unknown compound had no aldehyde or ketone functional groups<sup>16</sup>. In subsequent LC-HR-MS/MS analysis, we validated the formula with both accurate mass and natural isotope abundance information (Fig. 3b). In an analogous manner, we confirmed the formulas  $C_4H_9NO_2$  for BB160842 with two acidic protons,  $C_{25}H_{48}O_9$  for BB21735 with five acidic protons, and  $C_{18}H_{26}O_8$  for the isomers BB171284 and BB118961 with five acidic protons (Supplementary Table 1).

Using the elemental formulas, we retrieved all potential isomer structures from databases for these five unknown BinBase metabolites. For the retrieval of structure candidates, MS-FINDER 2.0 searches a combination of 14 metabolomics databases comprising 47,311 formulas and 224,622 unique known structures (Online Methods). However, enzyme promiscuity and a general lack of knowledge about enzyme reactions may be why many unknown compounds remain unidentified. Therefore, MS-FINDER 2.0 also incorporates all MINE-DB structures<sup>15</sup>, a collection of 643,307 virtual metabolites that are predicted on the basis of generalized enzymatic transformations as applied to KEGG pathway metabolites.

For BB106699,  $C_{10}H_{15}N_2O_9P$  yielded six isomers from the 14 metabolomic databases in MS-FINDER 2.0, and 33 isomers in MINE-DB. We then ranked all structures by matching the experimental spectra to predicted spectra for all isomers, considering chemical substructures recognized from the mass spectra, as well as biochemical likelihood. For the annotation of chemical substructures from GC-MS spectra, MS-FINDER 2.0 exploits 228 true positive fragmentation patterns from 80 reports published over the past 50 years<sup>17</sup>. These rules confirmed the presence of hydroxyl groups and a phosphate moiety in BB106699, a secondary amine and a carboxylic acid in BB160842, glycosylation of BB21735, and glucuronidation for BB171284 and BB118961 (Supplementary Table 1).

The unknown compound BB106699 was finally identified as *N*-methyl-UMP, a MINE predicted metabolite that has never been reported in biological samples. It ranked as the most likely structure in MS-FINDER 2.0 with all fragment ions rationalized by substructure annotations (Fig. 3c). We validated the identification of *N*-methyl-UMP by synthesizing an authentic standard (Online Methods) and comparing retention times and mass spectra generated by both GC-MS and LC-MS/MS to alternative *O*-methyl-UMP isomers (Supplementary Fig. 7). In the same manner, we annotated BB160842 as *N*-methylaniline



(Supplementary Fig. 8), BB21735 as lyso-monogalactosylmonopalmitin (Supplementary Fig. 9), BB171284 as 4-hydroxypropofol-1-glucuronide (Supplementary Fig. 10), and BB118961 as 4-hydroxypropofol-4-glucuronide (Supplementary Fig. 11).

Open access metabolomics repositories such as the NIH MetabolomicsWorkbench<sup>18</sup> and EBI MetaboLights<sup>19</sup> are important for enabling the comparison of metabolomics results to identified compounds. However, for comparisons of unknown metabolites across different biological studies, it is critical to standardize data acquisition methods and data processing parameters; currently only GC-MS is standardized. Our strategy of using MS-DIAL 2.0 with BinVestigate and MS-FINDER 2.0 outperformed alternative software-based approaches to deconvolution and compound identification for untargeted metabolomics analysis (Supplementary Tables 1 and 2).

Our approach revealed, for example, that *N*-methyl-UMP was highly upregulated in cancer cells and cancer tissues compared with its levels in any other cell type or tissue. Recently, methylation of small molecules has been shown to directly regulate cellular progression in stem cells<sup>20</sup>, thus raising the possibility of related mechanisms in cancer cells or the use of methylated metabolites as cancer biomarkers. More broadly, it has been shown quite regularly that small chemical alterations of metabolites may remove these compounds from primary biochemistry pathways, and that such modified metabolites (i.e., epimetabolites) subsequently gain regulatory functions, such as with oxylipins. For the discovery of epimetabolites, the integration of BinVestigate, MS-DIAL, and MS-FINDER provides a systematic strategy to make use of the complete set of mass spectral information and biochemical metadata to successfully find and rank the most likely chemical structures.

## METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the [online version of the paper](#).

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

## ACKNOWLEDGMENTS

This work was supported by the US National Science Foundation (NSF)–Japan Science and Technology Agency (JST) Strategic International Collaborative Research Program (SICORP) for Japan–United States metabolomics. We appreciate funding from the US National Science Foundation (projects MCB 113944 and MCB 1611846 to O.F.), the US National Institutes of Health (U24 DK097154 to O.F.),

and AMED–Core Research for Evolutionary Science and Technology (AMED-CREST) and JSPS KAKENHI (grants 15K01812, 15H05897, 15H05898, 17H03621 to M.A.).

## AUTHOR CONTRIBUTIONS

Z.L., H.T., M.A., and O.F. designed the research. G.W. and S.M. developed the BinVestigate program. H.T. developed the MS-DIAL 2.0 and MS-FINDER 2.0 programs. Z.L. performed the sample preparation, instrumental analysis, and data processing for unknown-compound identification. M.S. contributed biological and LC-MS studies for the identification of *N*-methyl-UMP. T.K. trained Z.L. in cheminformatics and contributed to validation of MS-FINDER. M.M. wrote the front end for BinVestigate. Z.L. and H.T. performed performance validation and program comparison for MS-DIAL 2.0 and MS-FINDER 2.0. Y.Z. and P.B. synthesized the *N*-methyl-UMP standard compound. A.O. improved the raw data file reader in ABF conversion. J.M., K.T., and O.F. contributed to the identification of lyso-MGMP and propofol derivatives. Z.L., H.T., M.A., and O.F. thoroughly discussed this project and wrote the manuscript.

## COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the [online version of the paper](#).

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

- Kim, S. *et al. Nucleic Acids Res.* **44**, 1202–1213 (2016).
- da Silva, R.R., Dorrestein, P.C. & Quinn, R.A. *Proc. Natl. Acad. Sci. USA* **112**, 12549–12550 (2015).
- Hanson, A.D., Pribat, A., Waller, J.C. & de Crécy-Lagard, V. *Biochem. J.* **425**, 1–11 (2009).
- Khersonsky, O. & Tawfik, D.S. *Annu. Rev. Biochem.* **79**, 471–505 (2010).
- Linster, C.L., Van Schaftingen, E. & Hanson, A.D. *Nat. Chem. Biol.* **9**, 72–80 (2013).
- Rappaport, S.M., Barupal, D.K., Wishart, D., Vineis, P. & Scalbert, A. *Environ. Health Perspect.* **122**, 769–774 (2014).
- Wikoff, W.R. *et al. Proc. Natl. Acad. Sci. USA* **106**, 3698–3703 (2009).
- Kumari, S., Stevens, D., Kind, T., Denkert, C. & Fiehn, O. *Anal. Chem.* **83**, 5895–5902 (2011).
- Showalter, M.R., Cajka, T. & Fiehn, O. *Curr. Opin. Chem. Biol.* **36**, 70–76 (2017).
- Patti, G.J. *et al. Metabolomics* **10**, 737–743 (2014).
- Fiehn, O., Wohlgemuth, G. & Scholz, M. In *Data Integration in the Life Sciences* (eds. Ludäscher, B. & Raschid, L.) 224–239 (Springer-Verlag, 2005).
- Kind, T. *et al. Anal. Chem.* **81**, 10038–10048 (2009).
- Tsugawa, H. *et al. Nat. Methods* **12**, 523–526 (2015).
- Tsugawa, H. *et al. Anal. Chem.* **88**, 7946–7958 (2016).
- Jeffries, J.G. *et al. J. Cheminform.* **7**, 44 (2015).
- Fiehn, O. *Trends Analyt. Chem.* **27**, 261–269 (2008).
- Lai, Z. & Fiehn, O. *Mass Spectrom. Rev.* <http://dx.doi.org/10.1002/mas.21518> (2016).
- Sud, M. *et al. Nucleic Acids Res.* **44**, D463–D470 (2016).
- Haug, K. *et al. Nucleic Acids Res.* **41**, D781–D786 (2013).
- Sperber, H. *et al. Nat. Cell Biol.* **17**, 1523–1535 (2015).

## ONLINE METHODS

**BinBase.** BinBase is a large GC-TOF-MS-based metabolomics database encompassing 1,561 studies with 114,795 samples for various species, organs, matrices, and experimental conditions. Because of the physics of GC-MS, analysis is restricted to thermostable small molecules up to 650 Da in size, even if derivatization by trimethylsilylation is used to reduce boiling points. Molecules profiled by trimethylsilylation–GC-MS-based metabolomics include amino acids; di- and tripeptides; hydroxyl acids; organic phosphates; fatty acids; alcohols; sugar acids; mono-, di-, and trisaccharides, including sugar acids and sugar alcohols; aromatic acids; nucleosides and mononucleotides (but not di- or trinucleotides); sterols; polyamines; and a large variety of miscellaneous compounds.

BinBase uses a retention index and mass spectral quality filtering system based on GC-TOF-based mass spectral deconvolution results as input<sup>21</sup> to store and report unique metabolite signals that are detected in metabolomic studies. Through the connected MiniX system<sup>22</sup>, all studies in BinBase are associated with metadata such as species, organs, cell types, and treatments. The BinBase algorithm has been published previously<sup>11,23</sup> and used over the past 13 years. It relies on mass spectral deconvolution of GC-TOF-MS data by the Leco ChromaTOF software and uses a multi-tiered filter system with different settings to annotate deconvoluted instrument peak spectra as unique database entries ('bins'). For typical studies on mammalian plasma with about 50–60 samples, about 1,000 peaks would be detected by ChromaTOF in at least one chromatogram at signal-to-noise ratios of >5. BinBase removes low-abundant, inconsistent, and noisy peaks that cannot be assigned to existing bins in BinBase and for which the spectral quality is too low to allow the generation of a new bin in BinBase, resulting in data sets that typically report 400–500 peaks for mammalian plasma samples. Compound identifications in BinBase are managed by the administrator with spectral libraries and retention index information from the Fiehnlib libraries<sup>12</sup> and NIST mass spectra. In a typical final BinBase report such as on mammalian plasma, about 30–40% of the reported bins are noted as identified metabolites—that is, about 150 compounds, including database identifiers such as KEGG, PubChem, and InChI keys.

**BinVestigate.** BinVestigate is an open access query tool (<http://binvestigate.fiehnlab.ucdavis.edu>) that can be used to obtain information on known/unknown compounds present in BinBase. BinVestigate uses data from trimethylsilyl-derivatized GC-MS-based metabolomics with respect to the frequency, intensity and origin of such metabolites. Unknowns can be queried in two ways in BinVestigate: (i) users can obtain result data from the West Coast Metabolomics Center (WCMC) or download public WCMC data from the free NIH database (<http://www.metabolomicsworkbench.org>)<sup>18</sup>; or (ii) users can match electron ionization (EI)-MS spectra obtained from their own GC-MS data sets against BinBase within narrow Kovats retention index windows to gain a Bin ID for cross-study analysis. BinVestigate result data are downloaded as CSV files and are represented by sunburst diagrams. Some information, such as cell line genotypes and specific treatments, is currently withheld to maintain the confidentiality of study specifics related to WCMC user data. For that reason, the WCMC uploads public data with more specific biological details to the NIH Metabolomics Workbench.

BinVestigate uses MongoDB for data storage and retrieval. The database is accessible and extendable through the use of its REST services and the RSQL query language. To populate the MongoDB database, we used a Spring-based integration workflow to associate the study design information from our in-house study design database MiniX with metabolomic information. The metabolome data were contributed by the in-house data processing system BinBase, which is based in PostgreSQL. Metabolite abundance data are normalized to the intensity of the sum of the internal standards (fatty acid methyl esters) to level the absolute differences between analyses over time. For comparison of abundances in the BinVestigate sunburst diagrams, it is important to note that a mere twofold difference in normalized abundance between organs or species should be ignored because such values reflect average intensities across biological studies, and thus are very much dependent on the conditions of such biological studies (for example, mutants, stress conditions, and other factors that may greatly influence metabolite abundance). In contrast, a 10-fold or 50-fold difference in relative intensity certainly indicates a high likelihood of different metabolite concentrations across different organs or species.

For users who query their own GC-MS mass spectra, Kovats retention indices are computed inside the integration workflow to enable access to BinVestigate for the general metabolomic community. Mass spectral similarity scores are calculated by the composite measure of the NIST algorithm. BinVestigate uses Java and Scala programming languages for data processing, and JavaScript for the graphic user interface. Query results are available as JSON-based documents or XLS-compatible CSV files. D3JS is used for data visualization.

To test the broad usability of BinVestigate, we tested one statistically significant unknown that was published by a European group investigating human cytomegalovirus infection<sup>24</sup>. In a similarity search with mass spectrum and retention index, the unknown U1804 (ref. 24) matched our BinBase unknown ID 8270 (**Supplementary Fig. 12**).

**BinBase false discovery rate testing.** We tested the data processing accuracy of BinBase by determining false positive, true positive, false negative and true negative spectra annotation rates for the five unknown biomarkers highlighted in this research report. We used the following equations:

$$\text{FDR} = \text{FP}/(\text{TP} + \text{FP}) \quad (1)$$

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{FP} + \text{FN} + \text{TP} + \text{TN}) \quad (2)$$

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN}) \quad (3)$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \quad (4)$$

where FDR is the false discovery rate, FP is the number of false positives, TP is the number of true positives, TN is the number of true negatives, and FN is the number of false negatives. For BB106699 (*N*-methyl-UMP), BinBase stored a total of 324,471 experimental mass spectra within the retention index search range. We identified a total of 7,363 true positive spectra for *N*-methyl-UMP. To determine false positives, we chose spectra with an ion abundance ratio of  $m/z$  (352/315) > 0.3. We chose this ratio because UMP elutes within the retention index window

(about 2 s later than *N*-methyl-UMP) and shares most fragment ions with *N*-methyl-UMP, except for  $m/z$  352, which is abundant in UMP but absent from *N*-methyl-UMP (Supplementary Fig. 13). With this criterion, we found three false positive spectra, at an FDR of 0.04%. The number of false negatives should be a lot higher than the number of false positives because BinBase was designed to assign ‘known peaks’ in a conservative way. Yet very low-abundant spectra and very complex chromatograms may lead to mass spectral deconvolution errors, and thus to false negative peak reports. False negatives were defined as spectra in the retention index search range that were not annotated as *N*-methyl-UMP but had  $m/z$  ion ratios of 169/315 between 10:1 and 1:1 and 169/299 between 10:1 and 1:1 (i.e., 1.0 to 10.0) and  $m/z$  169 > 30% base peak intensity. With these criteria, 1,472 spectra were possible false negatives, yielding an overall sensitivity of 83.3%, specificity of 100%, and predictive accuracy of 99.4%. We used similar detailed analyses for the other four BinBase spectra (Supplementary Table 3).

A close investigation of histograms and raw chromatograms for BB160842, however, showed coelution of *N*-methylalanine and the isomer 2-aminobutyric acid for many sample chromatograms. For false positives we used the criterion  $m/z$  218 > 2% base peak intensity, which represents an ion that is a typical fragment for  $\alpha$ -amino acids<sup>17</sup>. Unfortunately, *N*-methylalanine also shows low-abundant  $m/z$  218 ions, which are much less abundant than in 2-aminobutyric acid. 2-aminobutyric acid also shows most fragments that occur in *N*-methylalanine (base peak  $m/z$  130,  $m/z$  100,  $m/z$  114,  $m/z$  147, and  $m/z$  204). *N*-methylalanine presents low-abundant diagnostic ions ( $m/z$  144,  $m/z$  142, and  $m/z$  175) that are even less abundant and present in different ratios in pure 2-aminobutyric acid. Using these diagnostic ions, we validated the detection of pure *N*-methylalanine in biological samples, as well as the detection of pure 2-aminobutyric acid in other samples. In most chromatograms, however, total peak intensities were too small to allow deconvolution of the quantities of both coeluting compounds owing to the low abundance of the diagnostic ions. Thus, BB118961 should be regarded as reflecting a mixture of both compounds in BinVestigate.

**MS-DIAL 2.0.** MS-DIAL 2.0 is designed as a universal program for MS data processing that supports any mass spectrometry approach, including GC-MS, GC-MS/MS, LC-MS, and LC-MS/MS. It is vendor independent and supports data conversion from file formats of many instrument manufacturers, namely, Agilent, Bruker, Leco, Sciex, Shimadzu, Thermo, and Waters. This software also supports any data acquisition method, from nominal or accurate mass analysis to data-dependent or data-independent MS/MS. It runs with a user-friendly graphical user interface on a Windows system (.NET Framework 4.0 or later with at least 4 GB RAM memory). MS-DIAL 2.0 and its source code are freely downloadable at the PRIME website ([http://prime.psc.riken.jp/Metabolomics\\_Software/MS-DIAL/](http://prime.psc.riken.jp/Metabolomics_Software/MS-DIAL/)).

A summarized workflow for processing high-resolution GC-MS data is shown in Supplementary Figure 2, with three primary metabolites shown as examples: glycerol, phosphate, and leucine. The peak maxima of these metabolites coelute within 1.02 s with 3-s peak widths. MS-DIAL 2.0 spots all  $m/z$  peaks and determines peak spot properties (Supplementary Fig. 2a), and then constructs peak groups on the basis of local maxima of the second

Gaussian filtered array of sharpness values (Supplementary Fig. 2b). The most important part is the subsequent chromatogram deconvolution to assign  $m/z$  spots and fractions of shared  $m/z$  intensities to specific peak groups (Supplementary Fig. 2c). The deconvolution follows a least-squares regression model based on unique ions, similar to the original MS-DIAL algorithm<sup>13</sup> implemented for data-independent MS/MS chromatogram deconvolution. The program substantially improved the spectral similarities of all coeluting metabolites in the example data, greatly increasing the number of positively identified metabolites. For compound identification, a total of 15,302 GC-MS spectra and 21,770 LC-MS/MS spectra are currently available in the internal mass spectral database in MS-DIAL 2.0.

**Raw data handling and MS-DIAL scalability.** The data stream, including file formats and converters, for MS-DIAL is summarized in Supplementary Figure 4. MS-DIAL can import mzML, netCDF, and Analysis Base Framework (ABF) formats, but ABF format is recommended for rapid data retrieval and for efficient data access. The ABF file converter is freely available online (<http://www.reifycs.com/AbfConverter/index.html>). ABF file conversion and compatibility with MS-DIAL have been validated for open-access formats such as mzML and netCDF, as well as vendor formats from Agilent Technologies (.D), Bruker Daltonics (.D), Leco (.netCDF), Sciex (.WIFF), Shimadzu (.LCD), Thermo Fisher Scientific (.RAW), and Waters (.RAW).

**Peak detection.** Profile data are centroided in MS-DIAL 2.0 before peak detection. First, data points are smoothed with a linearly weighted smoothing average used as the default setting. Noise is defined by ion amplitude and first and second derivatives. Peak start and end retention times are first approximated on the basis of noise levels. Then the local minima within adjacent 5-point windows are explored to determine optimal peak start and end times by forward and back tracing. To avoid defining peak starts and ends too far away from the peak maxima, users can define an average peak width (APW) parameter. APW is used as a clamp for peak width definition within a maximum of  $\pm 2$  APW. MS-DIAL 2.0 involves background subtraction for filtering of chemical noise (Supplementary Fig. 14). Once the initial peak detection program is finished, the unsmoothed raw chromatogram is retrieved as a control. Peaks are excluded if the ion abundance of one neighbor point from the peak top is zero, because the smoothing algorithm may construct signal resembling actual peaks even for chromatographic noise. A secondary filter is used to exclude baseline noise arising from a sequence of ‘peak-like’ spike noise that may occur when too many peaks are detected in the initial peak-picking algorithm. Here the amplitude of spike noise is defined as the difference between two adjacent scan points. The current filter will exclude peaks if four spike noise signals are programmatically detected within  $\pm 5$  APWs of a peak top.

**MS1Dec deconvolution.** MS-DIAL 2.0 deconvolution, termed MS1Dec, starts with MS1 fragment ions. The peak-spotting program is first executed over the entire retention time and  $m/z$  range. In MS-DIAL, the detected  $m/z$ -retention time features are termed peak spots. The peak quality value is defined for each spot through comparison of ideal slope values that evaluate the peak smoothness (i.e., whether the peak contains any spike noise within its peak



width). There are three quality levels: high if the ideal slope value is higher than 0.999, middle if the value is between 0.9 and 0.999, and low if it is less than 0.9. The peak sharpness value evaluates peak symmetry in combination with absolute intensity. Ideal slope and peak sharpness values were defined according to previous work<sup>13</sup>.

Subsequently, all peak spots that have identical peak widths and peak top retention times are combined into single arrays. For each array, peak sharpness values are summed and a second Gaussian derivative filter is applied to construct 'peak groups'. The smoother is defined by a default sigma value of 0.5 to join  $m/z$  peak maxima even in cases of small derivations. In practice, MS-DIAL 2.0 requires at least two scan differences in the peak tops of coeluting metabolites to be distinguished because local maxima of the smoothed sharpness arrays are recognized as peak maxima and ignore neighboring peaks.

The main purpose of deconvolution is to estimate the peak abundance of  $m/z$  traces that are shared by two or more coeluting metabolites. This is achieved through the definition of model peak  $m/z$  traces in the retention time region of each peak group. High-quality  $m/z$  traces are used to construct model peak fitting for each  $m/z$  trace by least-squares regression. Middle-quality  $m/z$  traces will be used if there are no high-quality traces in a focused peak group. Peak groups that consist of only low-quality traces are recognized as 'not detected'. For formation of the model peak for a peak group, the peak intensities that are above 90% of their base peaks are summed for modeling to increase the model accuracy for low ion signals. Peak maxima, peak start points, and peak end points are determined through tracing of the local maximum, left local minimum, and right local minimum, respectively.

For the deconvolution of local model peaks  $M_t(n)$ , coeluting adjacent model peaks are considered if peak start and end points of the two (or more) model peaks overlap. For practical reasons, the current program considers up to four coeluting metabolites ( $M_{t-2}(n)$  and  $M_{t-1}(n)$  for the left side of the target compound;  $M_{t+1}(n)$  and  $M_{t+2}(n)$  for the right side of the target compound) for the chromatogram deconvolution of targeted compound  $M_t(n)$ . Therefore, the raw  $m/z$  trace  $C(n)$  will be decomposed to the model peaks as follows:

$$C(n) = aM_{t-2}(n) + bM_{t-1}(n) + cM_t(n) + dM_{t+1}(n) + eM_{t+2}(n) + fn + g$$

**Retention time and full mass spectral similarity.** Retention time and mass spectral similarities are used for compound identification and peak alignment in data processing. To determine retention time (or index) similarity, we use a Gaussian function under the assumption that the potential retention time drifts between sets of chromatograms will follow a Gaussian distribution. MS-DIAL 2.0 uses a combined value as 'full mass spectral similarity' with a weight factor of 2:2:1 for dot product, reverse dot product, and matched fragment ratio. To calculate the overall similarities of chromatogram alignments, the software sums values of retention time and mass spectral similarity.

**Compound identification.** Deconvoluted spectra are matched against mass spectral libraries that are imported in NIST MSP format. Library match hits are ranked against experimental data on the basis of the total retention time (or index) and mass spectral similarity across all samples that are processed in a batch. Users can define cutoff thresholds for both parameters. MS-DIAL 2.0

supports two retention indices: Kovats RI, based on alkanes, and Fiehn RI, based on fatty acid methyl esters.

**MS-DIAL aligner.** The alignment algorithm for detecting peak groups across all samples of a data processing batch was optimized for GC-MS data. The aligner runs in the following procedures: (a) creation of a reference table; (b) fitting of each sample peak table to the reference peak table; (c) filtering of aligned peaks; and (d) missing-value interpolation. For LC-MS/MS data, the MS-DIAL aligner focuses on MS1 precursor ions. For GC-MS data, the MS-DIAL aligner determines the unique ion used for peak quantification, termed the 'quant mass'. The  $m/z$  of the highest ion abundance in a high-quality trace is defined as the quant mass in the program, and the  $m/z$  from the middle-quality trace will be used if no high-quality trace is present. A user-defined sample will serve for the creation of a starting reference table of all deconvoluted peak groups. Additional peak groups from further samples are inserted if the total retention and spectral similarity between the sample peak groups is lower than a user-defined cutoff compared with the existing peak groups in the reference table. This insertion routine is repeated for all peaks of all samples. The final table is used as the reference peak table for peak alignment. Each sample peak table is assigned into the reference peak table as the following criterion:

$$\text{Score} = a * \text{RT similarity} + b * \text{MS similarity}$$

where RT is retention time, and MS is mass spectra. The coefficients  $a$  and  $b$  can be set by the user. After all peaks of all samples have been fitted to the reference peak table, the alignment peak table including retention time, quant mass, and intensity is constructed with each row referred to as an alignment spot. The representative quant mass for each aligned spot is defined by the consideration of ion abundances and frequencies of quant mass among samples and is used for peak height and peak area determination. The average retention time (or index) and average quant mass (for accurate mass data) are calculated. A 'fill percentage' with respect to the positive detected sample number in a peak group is obtained. The results of compound identification obtained by matching the reference database against the experimental peak with the best retention time and mass spectral similarity are stored. The corresponding spectrum for each aligned peak group is retrieved from the imported samples for each identified sample peak, or the spectrum with the highest total ion intensities for unidentified peaks.

The interpolation program for missing values is executed as follows: (1) for the quantification mass of each aligned peak group, maximum and minimum retention times (or indices) are recorded with maximum and minimum peak widths; (2) for gap filling, local maxima and minima signal intensities are determined for the quant mass within the retention time (index) window of (minimum RT, maximum RT)  $\cap$  (quant mass average – sigma mass, quant mass average + sigma mass); (3) peak height is defined by the amplitude difference of local maxima and minima, and the corresponding estimated peak area is calculated on the basis of the average peak width and actual peak maxima.

**MS-FINDER 2.0.** MS-FINDER 2.0 is a universal program for structure elucidation of unknown mass spectra in LC-MS/MS<sup>14</sup> and, now, GC-MS. Although vendors provide specific GC-MS/MS instruments, in practice we have observed a high degree of

in-source fragmentation in GC-MS spectra (both hard electron ionization and soft chemical ionization), making the distinction between GC-MS and GC-MS/MS spectra unnecessary. MS-FINDER 2.0 is compatible with the Windows system (.NET Framework 4.0 or later with at least 8 GB RAM memory). It is freely downloadable at the PRIME website ([http://prime.psc.riken.jp/Metabolomics\\_Software/MS-FINDER/](http://prime.psc.riken.jp/Metabolomics_Software/MS-FINDER/)). MS-FINDER 2.0 uses a conventional spectral database search function based on dot product, reverse dot product, and matched fragment ratio. More important, this program features a computational mass spectral fragmentation (*in silico* fragmenter search) for structure annotation. Here the technical details of an *in silico* fragmenter search for full GC-MS spectra of TMS compounds is described. Additionally, the molecular adduct ions, often the  $[M]^{+}$  or  $[M-CH_3]^{+}$  radical ions, have to be manually determined by peak alignments in MS-DIAL 2.0.

**Molecular formula generator.** Structure elucidation in MS-FINDER 2.0 begins with formula prediction. Full GC-MS spectra are mostly ionized compounds with odd electron (radical) ions, denoted by  $+•$  (Supplementary Fig. 5a). This program supports 11 elements for formula generation: carbon (C), hydrogen (H), oxygen (O), nitrogen (N), sulfur (S), phosphorus (P), fluorine (F), chlorine (Cl), bromine (Br), iodine (I), and silicon (Si). CHONSPSi atoms were used for the program evaluation presented in this paper.

Elemental formulas are computationally generated with valence rules and elemental ratio checks. Next, the number of TMS and methoxy (MeOX) moieties are simulated as follows. Here we use the formula  $C_{22}H_{47}N_2O_9PSi_4$  as an example (Supplementary Fig. 5b). MS-FINDER 2.0 recognizes the origin of all Si elements as TMS moieties. Therefore,  $C_{22}H_{47}N_2O_9PSi_4$  is converted to  $C_{10}H_{15}N_2O_9P$ , with each  $C_3H_8Si$  subtracted as one trimethylsilyl from the derivatized formula. The number of MeOX moieties is simulated by the number of N atoms. All simulated candidates are used as the results of formula generation, and hence  $C_{10}H_{15}N_2O_9P$  (0 MeOX),  $C_9H_{12}NO_9P$  (1 MeOX), and  $C_8H_9O_9P$  (2 MeOX) are obtained from the original  $C_{10}H_{15}N_2O_9P$  where  $CH_3N$  is deleted per one N atom.

**Molecular formula ranking.** Formulas are ranked on the basis of the sum of five diagnostic scores, specifically, mass error, isotopic ratio error, formula assignment to fragment ions, neutral loss matching, and presence of the formula in the unified metabolome structure database. Our unified metabolome structure database is an integrated database for retrieving biologically reported formulas (90,227 unique formulas in total) in 15 repositories: BMDB, ChEBI, DrugBank, ECMDB, FooDB, HMDB, KNApSACk, PlantCyc, PubChem (Biomolecules), SMPDB, T3DB, UNPD, YMDB, STOFF, and MINE. The virtual enzyme expansion database MINE is evaluated separately from other repositories. If a formula candidate is present in one of the 14 databases (except for MINE), the evaluation score is 0.5; otherwise it is 0. To this value, the number of databases that include this formula, standardized by 0.5, is added. After this, 0.2 is added if the formula is also present in the MINE database. The formula database is stored in an EFD file of the MS-FINDER folder in ASCII file format.

**Searching of structure candidates and *in silico* derivatization.** Currently, MS-FINDER 2.0 has three options for the retrieval

of structural isomers: the unified metabolome database of 14 repositories with 224,622 unique structures, the MINE database with 643,307 unique structures, and the PubChem REST service for approximately 70 million structures (Supplementary Fig. 5c). Each repository in the combined metabolome database can also be selected by itself. The combined metabolome database and MINE database are stored in ESD and MSD files in the MS-FINDER folder in ASCII file format.

After the structural data for a given formula are retrieved, the structure is computationally derivatized on the basis of the simulated trimethylsilyl and MeOX numbers via the following procedures: (1) the acidic protons attached with ONSP heteroatoms are recognized as the reactive protons amenable to trimethylsilyl derivatization; (2) the carbonyl groups as ketones or aldehydes are recognized as reactive  $C=O$  for MeOX derivatization unless further ONSP heteroatoms are attached like carboxylic acids; (3) candidate structures are excluded if the number of acidic protons and carbonyl groups is less than the number of simulated trimethylsilyl and MeOX; (4) trimethylsilyl derivatization is prioritized by  $OH > COOH > NH_2 > SH > NHR$ ; (5) MeOX derivatization is prioritized by  $R_1(C=O)R_2 > R(C=O)H$ ; and (6) identical functional groups are derivatized with the same priority, that is, the order of derivatization is determined by the order of atomic numbering.

**Ranking of structure candidates.** Hydrogen rearrangement rules (rules P1–P5 and N1–N4 for positive and negative ion mode, respectively) have been established to interpret mass spectra of LC-MS/MS with collision-induced dissociation (CID) as previously reported<sup>14</sup>. Structure candidates are ranked by the integrated score of hydrogen rearrangement rules, fragment linkage and bond dissociation energy. Now the updated rule-based mass spectral fragmentation library also includes structure elucidation for GC-MS spectra. A total of 533 fragment ions are rationalized by  $m/z$ , formula, and SMILES code and stored in an EIF file of the MS-FINDER folder in ASCII file format. MS-FINDER 2.0 uses this rule-based fragmentation library for substructure assignments because GC-MS spectra produce intense fragmentation schemes that include electron shifts, hydrogen rearrangements, homolytic or heterolytic bond cleavages, and intramolecular rearrangements, rather than stabilizing fragments by aromatization. This program excludes aromatic fragment ions unless the aromatic substructures are detected in the original molecules. The likelihood of a fragment ion with assigned substructure is evaluated by a Gaussian function based on experimental mass errors. Advanced likelihood on the basis of molecular fingerprints in combination with similarity calculation methods, such as Jaccard, will be used in the next version.

For fragment ions without substructure assignment by the rule-based library, *in silico* spectral annotation is carried out through simulation of an  $\alpha$ -cleavage process for up to two bonds with a consideration of  $\pm 2$  hydrogen rearrangements. To specify the appropriate fragments within mass tolerance, computational ions are assigned to observed ions according to the following priorities: (1) fragments from the first cleavage that differ by up to two hydrogens from a neutralized substructure; (2) fragments from the second cleavage with assigned precursors in a higher  $m/z$  area; and (3) fragments of minimum mass errors. This scoring system is identical to that for LC-CID-MS/MS spectra except the penalty of hydrogen rearrangement rules: at current,



the hydrogen rearrangement rules are always considered as true for GC-MS spectra.

**Performance validation of MS-DIAL 2.0.** The scalability and functionality of MS-DIAL 2.0 for GC-MS data processing were validated with six raw data files from five major MS vendors (**Supplementary Fig. 3** and **Supplementary Data Set 1**). All raw data files (except Bruker and Thermo) are available at the PRIME website ([http://prime.psc.riken.jp/?action=drop\\_index](http://prime.psc.riken.jp/?action=drop_index)) with the following data sources:

- (1) LECO GC-TOF(MS): The biological sample was *Euglena gracilis*; analysis procedures were done according to the “LECO GC-TOF MS” protocol from a previous report<sup>25</sup>.
- (2) Agilent GC-Q(MS): The biological sample was NIST standard human plasma; analysis procedures were done according to the “Agilent GC-Quadrupole MS” protocol from a previous report<sup>25</sup>.
- (3) Agilent GC-QTOF(MS): The biological sample was *Chlamydomonas reinhardtii*; analysis procedures are described below in the “Reagents and sample preparation” section of this paper.
- (4) Shimadzu GC-Q(MS): The raw data were obtained from previously reported data<sup>26</sup>.
- (5) Bruker GC-Q(MS): The raw data were kindly provided by Bruker Daltonics.
- (6) Thermo GC-QE(MS): The raw data were kindly provided by Thermo Fisher Scientific.

The MS-DIAL 2.0 data processing procedures were followed by the software tutorial ([http://prime.psc.riken.jp/Metabolomics\\_Software/MS-DIAL/](http://prime.psc.riken.jp/Metabolomics_Software/MS-DIAL/)). The analyzing parameters and MS libraries can be downloaded at [http://prime.psc.riken.jp/?action=drop\\_index](http://prime.psc.riken.jp/?action=drop_index). The actual processing times for nominal and accurate mass GC-MS data were 20–30 s and 1–2 min, respectively. The identification results were manually confirmed by the MS-DIAL 2.0 graphical user interface.

**Performance validation of MS-FINDER 2.0.** The performance of MS-FINDER 2.0 for structure elucidation was tested against the accurate mass GC-EI-MS spectra of 441 TMS compounds. The sample preparation and analytical conditions are described below. For each compound, the molecular mass and the trimethylsilyl and MeOX number of the derivatized form were determined by manual investigation, and the formula, SMILES and InChIKey of the nonderivatized form were generated by ChemAxon MolConverter and Calculator (<http://www.chemaxon.com>).

The mass tolerance, relative abundance cutoff, and isotopic ratio tolerance were set to 0.01 Da, 1%, and 20%, respectively. Filtering of “LEWIS and SENIOR check” and “common range for element ratio check” was activated. The targeted atoms were set as C, H, O, N, S, and P with the option of “TMS-MeOX derivatized compound.” Tree depth was set to 2, and the “use of fragmentation library for electron ionization” option was applied. For batch process, the top 100 formula candidates were transferred to the structure searching procedure. Unified metabolome databases including BMDB, ChEBI, DrugBank, ECDDB, FooDB, HMDB, KNApSack, PlantCyc, PubChem (Biomolecules), SMPDB, T3DB, UNPD, YMDB, and STOFF were selected for structure searching. We retrieved PubChem ‘Biomolecules’ from the ~70 million compounds in PubChem by restricting the search to “Biomolecular and

interaction pathway,” then to “Biosystems and pathways.” Currently, 12,400 compounds are retrievable from PubChem in this way.

MS-FINDER 2.0 was tested by three structure resource sets. The first set was the internal metabolome database with 14 repositories as mentioned above, termed FINDMetDB. To increase the search space, we also included the MINE and PubChem databases for the second and third sets, which were termed FINDMetDB+MINE and FINDMetDB+PubChem, respectively. Therefore, the numbers of total unique structures for the software accuracy test were 13,869, 92,628, and 280,245 in FINDMetDB, FINDMetDB+MINE, and FINDMetDB+PubChem, respectively. The performance test results of MS-FINDER 2.0 and of ‘random sampling method’ are shown in **Supplementary Figure 6**. The logP and natural product likeness values were calculated by ChemAxon Calculator (<http://www.chemaxon.com>) and Natural Product Likeness Calculator (<https://sourceforge.net/projects/np-likeness/>). With a mass tolerance of 10 mDa for spectral annotation of CHNOSP elements, the probability of finding the correct structure for the top hit or among the top 3, top 5, or top 10 hits was 49.2%, 72.1%, 82.1%, or 91.8%, respectively.

**Software comparison for MS-DIAL 2.0 and MS-FINDER 2.0.** We compared the results obtained by MS-DIAL 2.0 (version 2.52) and MS-FINDER 2.0 (version 2.10) against those from other, alternative programs. For the performance of GC-MS chromatogram deconvolution, MS-DIAL 2.0, AMDIS<sup>27</sup>, AnalyzerPro, and ChromaTOF were tested with identical raw data ([http://prime.psc.riken.jp/?action=drop\\_index](http://prime.psc.riken.jp/?action=drop_index)). Deconvoluted mass spectra were exported from the data processing software and imported to the NIST MS Search program (<http://chemdata.nist.gov/mass-spc/ms-search/>) to obtain match scores for compound annotation. On the basis of the results of nine primary metabolites from four coeluting peak groups, MS-DIAL 2.0 outperformed AMDIS, AnalyzerPro, and ChromaTOF for most individual spectral similarity matches and for the average match scores (**Supplementary Table 2**).

For the functionality of *in silico* GC-MS mass spectral annotation, MS-FINDER 2.0, CFM-ID<sup>28</sup>, MetFrag<sup>29</sup>, Molecular Structure Correlator (MSC), and Mass Frontier were evaluated with the mass spectra ([http://prime.psc.riken.jp/?action=drop\\_index](http://prime.psc.riken.jp/?action=drop_index)) of our five BinBase unknowns highlighted in this paper with same structure candidate lists that were downloaded from PubChem by formula query, computationally derivatized and generated in ChemAxon Instant JChem (<http://www.chemaxon.com>) (**Supplementary Table 1**). For calculation, scoring, and ranking of 25–59 different isomers, MS-FINDER 2.0 required 2–12 s of processing time, similar to the MetFrag and MSC programs, whereas CFM-ID and Mass Frontier needed considerably more time (5–108 min). More important, MS-FINDER 2.0 was the only software that could confidently identify the five unknown compounds presented here as top hits; other programs had an average ranking of 3.8 (CFM-ID) to 9.6 (Mass Frontier).

The computer condition for data processing was an Intel (R) Core (TM) i7 with 4 GHz CPU and 8 GB RAM with the Windows 7 system. The settings used for MS-DIAL 2.0 and MS-FINDER 2.0 were the default values as mentioned above. The parameters of other programs were as follows:

AMDIS: The software application was downloaded from <http://chemdata.nist.gov/>. The match factor penalty level was ‘Very Strong’. The scan direction was ‘Low to High’. The adjacent peak

subtraction was 'None'. Resolution, sensitivity, and shape requirement were all 'Medium'.

**AnalyzerPro:** The software application was purchased from Spectral Works. The data processing settings used were the vendor-suggested default parameters.

**ChromaTOF:** The software application was purchased from LECO Corporation. The data processing settings used were the vendor-suggested default parameters.

**CFM-ID:** The web application was used in <http://cfmid.wishartlab.com/>. The spectra type was 'EI'. The number of results, mass tolerance, and scoring function were 20, 0.01 Da, and dot product, respectively.

**MetFrag:** The web application was used in <http://msbi.ipb-halle.de/MetFrag/>. The process mode was '[M]'. MZABS and MZPPM were 0.01 Da and 20, respectively.

**MSC:** The software application was purchased from Agilent Technologies. The data processing settings used were the vendor-suggested default parameters.

**Mass Frontier:** The software application was purchased from Thermo Fisher Scientific. The data processing settings used were the vendor-suggested default parameters.

**Reagents and sample preparation.** The following reagents and authentic standard compounds were obtained from the named suppliers: water, isopropanol, and acetonitrile (Fisher Optima); pyridine (Acros Organics); C8–C30 fatty acid methyl esters, methoxyamine hydrochloride, ethoxyamine hydrochloride, *N*-methyl-*N*-(trimethylsilyl)-trifluoroacetamide (MSTFA), *N*-methyl-*N*-(trimethyl-d9-silyl)-trifluoroacetamide (MSTFA-d9), ammonium formate, formic acid, and *N*-methyl-L-alanine (Sigma-Aldrich); 2'-*O*-methyluridine-5'-triphosphate, 3'-*O*-methyluridine-5'-triphosphate, 5-methyluridine-5'-triphosphate (TriLink BioTechnologies); 4-hydroxypropofol-1-*O*-β-D-glucuronide, and 4-hydroxypropofol-4-*O*-β-D-glucuronide (Toronto Research Chemicals).

All data used here were taken from previous studies<sup>30–32</sup>. Briefly, all metabolite extraction procedures were kept on ice, and the quantities for sample aliquots were 25 μL for blood plasma, 5 × 10<sup>6</sup> for cells, 5 mg for tissues, and 2 mL for algae cultures. Metabolites were extracted with 1,000 μL of degassed acetonitrile: isopropanol:water (3:3:2, v/v/v) and then homogenized, centrifuged, decanted, and evaporated. Extracts were cleaned with 500 μL of degassed acetonitrile:water (1:1, v/v) to remove triglycerides and membrane lipids, and evaporated again. For GC-MS analysis, internal standard C8–C30 fatty acid methyl esters were added to determine the retention index. The dried samples were derivatized with 10 μL of methoxyamine hydrochloride (or ethoxyamine hydrochloride) in pyridine and subsequently by 90 μL of MSTFA (or MSTFA-d9) for trimethylsilylation of acidic protons. For LC-MS analysis, the extracted samples were resuspended in 50 μL of acetonitrile:water (4:1, v/v) and applied to the instrument.

**Analytical conditions.** All data used here were taken from previous studies; please see refs. 30–32 for representative examples. Briefly, for GC-MS analysis, we used an Agilent 7890A GC system (Agilent Technologies) and an Agilent 7200 accurate mass Q-TOF mass spectrometer (Agilent Technologies) with the transfer line temperature maintained at 290 °C. Chromatography was done on an Rxi-5Sil MS column (30 × 0.25 mm, 0.25 μm; Restek) with helium (99.999%; Airgas) at a constant flow of 1 mL/min. The GC temperature

program was set as follows: initial temperature of 60 °C with a hold time of 1 min, a temperature ramp of 10 °C/min to 325 °C, and a final hold time of 9.5 min at 325 °C. The injection volume was 1 μL in splitless mode at 250 °C. Mass spectra were acquired from *m/z* 50 to *m/z* 800 at a 5-Hz scan rate and 750-V detector voltage in both electron ionization (EI) mode and chemical ionization (CI) mode. Other data acquisition parameters were as follows: EI ion source temperature, 230 °C; EI electron energy, 70 eV; CI ion source temperature, 300 °C; CI electron energy, 135 eV; CI gas flow rate, 20%; CI gas, methane (99.999%; Airgas).

For LC-MS analysis, the initial separation was achieved on an Agilent 1290 infinity LC system (Agilent Technologies) with an Acquity UPLC BEH amide column (150 × 2.1 mm, 1.7 μm; Waters). The mobile phases consisted of (A) 10 mM ammonium formate and 0.125% formic acid in water and (B) acetonitrile:water (95:5, v/v) with 10 mM ammonium formate and 0.125% formic acid. The gradient was 0 min, 100% B; 2 min, 100% B; 7.7 min, 70% B; 9.5 min, 40% B; 10.3 min, 30% B; 12.8 min, 100% B; 16.8 min, 100% B. Sample volumes of 2 μL and 5 μL were used for the injection in ESI (+) and ESI (–), respectively, with a flow rate of 0.4 mL/min. The autosampler temperature was 4 °C, and the column temperature was 45 °C. The mass spectrometer was equipped with an Agilent 6530 accurate mass Q-TOF system (Agilent Technologies). MS and MS/MS data were collected at a 4-Hz scan rate and *m/z* 50–800 mass range. Collision energy was applied at 20 eV. Mass calibration was maintained at a constant infusion of reference ions at *m/z* 121.0509, *m/z* 922.0098 for positive mode and *m/z* 119.0363, *m/z* 966.0007 for negative mode.

**Synthetic procedures.** Glassware was oven-dried at 100 °C overnight before the reaction. All reagents were purchased from commercial sources (Sigma-Aldrich or Fisher Scientific) and were used without further purification unless noted otherwise. Reactions were carried out under an atmosphere of dry argon. Liquid reagents were introduced by disposable syringes. Thin layer chromatography was done with EMD silica gel 60, F254 precoated thin layer chromatography plates. Short- and long-wave visualization were carried out with a Mineralight multiband ultraviolet lamp at 254 and 365 nm, respectively. Flash column chromatography was done with Merck silica gel (Sorbent Technologies; 60–200 mesh). Purification of nucleotide monophosphate was done on a column of Sephadex DEAE-A25. The resin was swollen in 1 M NaHCO<sub>3</sub> at 4 °C for 1 d and washed with deionized water before use, unless noted otherwise. The fractions containing nucleotide monophosphate were identified by a Beckman DU-7400 UV-Vis scanning spectrophotometer and Applied Biosystems QTrap mass spectrometer.

**N<sup>3</sup>-methyluridine synthesis.** N<sup>3</sup>-methyluridine was synthesized as previously described<sup>33,34</sup> with minor modifications. In brief, uridine (1.504 g, 6.16 mmol) and K<sub>2</sub>CO<sub>3</sub> (1.704 g, 12.33 mmol) were added to a mixture of DMF (7.5 mL) and acetone (7.5 mL). Methyl iodide (383 μL, 6.16 mmol) was added dropwise to the suspension. The system was then refluxed for 5 h. The solvent was removed *in vacuo*. The residue was purified by chromatography on a flash silica column. Elution was done with 5–10% (v/v) MeOH in CH<sub>2</sub>Cl<sub>2</sub>. Fractions containing the product were dried *in vacuo*. Product was recrystallized in MeOH, which yielded N<sup>3</sup>-methyluridine as a white crystal.

***N*<sup>3</sup>-methyluridine 5'-monophosphate synthesis.** *N*<sup>3</sup>-methyluridine 5'-monophosphate was synthesized as previously described<sup>35</sup>, with minor modifications. *N*<sup>3</sup>-methyluridine and proton sponge were dried overnight in a vacuum oven. We dissolved nucleoside (180 mg, 0.69 mmol) in freshly distilled trimethyl phosphate (8 mL) by heating the solution, and then cooled the solution to −15 °C. Dry proton sponge (444 mg, 2.07 mmol) was then added to the solution, and the solution was stirred at −15 °C for 20 min. Distilled phosphorus oxychloride (64 µL, 0.69 mmol) was added dropwise under argon with a micro-syringe. The reaction solution was then stirred at −15 °C. After 2 h at −15 °C, a solution of 1 M triethylammonium bicarbonate (30 mL, pH 8) was added. The clear solution was stirred at room temperature for 45 min and then freeze-dried. The crude mixture was dissolved in water and purified on a Sephadex DEAE-A25 column with a linear gradient of 0.01–0.10 M triethylammonium bicarbonate buffer. Fractions containing *N*<sup>3</sup>-methyluridine 5'-monophosphate were identified by UV spectrophotometer and mass spectrometry. Combined fractions were evaporated under reduced pressure, yielding *N*<sup>3</sup>-methyluridine 5'-monophosphate as a white solid.

**Life Sciences Reporting Summary.** Further information on experimental design is available in the **Life Sciences Reporting Summary**.

**Data availability.** The data that support the findings of this study are available publicly through BinVestigate for summary investigation. Raw data files (including LC-HR-MS/MS and GC-HR-MS data) are available from the corresponding author on request. Raw data and result data for low-resolution GC-TOF-MS are available from the public NIH MetabolomicsWorkbench website.

21. Styczynski, M.P. *et al. Anal. Chem.* **79**, 966–973 (2007).
22. Scholz, M. & Fiehn, O. In *Pacific Symposium on Biocomputing* 169–180 (World Scientific, 2007).
23. Fiehn, O. *et al. Plant J.* **53**, 691–704 (2008).
24. Fattuoni, C. *et al. Clin. Chim. Acta* **460**, 23–32 (2016).
25. Fiehn, O. *Curr. Protoc. Mol. Biol.* **114**, 30.4.1–30.4.32 (2016).
26. Tsugawa, H. *et al. J. Biosci. Bioeng.* **112**, 292–298 (2011).
27. Stein, S.E. *J. Am. Soc. Mass Spectrom.* **10**, 770–781 (1999).
28. Allen, F., Pon, A., Greiner, R. & Wishart, D. *Anal. Chem.* **88**, 7689–7697 (2016).
29. Ruttkies, C., Strehmel, N., Scheel, D. & Neumann, S. *Rapid Commun. Mass Spectrom.* **29**, 1521–1529 (2015).
30. Budczies, J. *et al. BMC Genomics* **13**, 334 (2012).
31. Lee, D.Y., Park, J.J., Barupal, D.K. & Fiehn, O. *Mol. Cell. Proteomics* **11**, 973–988 (2012).
32. Hartman, A.L. *et al. Proc. Natl. Acad. Sci. USA* **106**, 17187–17192 (2009).
33. Flosadóttir, H.D., Jónsson, H., Sigurdsson, S.T. & Ingólfsson, O. *Phys. Chem. Chem. Phys.* **13**, 15283–15290 (2011).
34. Yamamoto, I., Kimura, T., Tateoka, Y., Watanabe, K. & Ho, I.K. *J. Med. Chem.* **30**, 2227–2231 (1987).
35. El-Tayeb, A., Qi, A. & Müller, C.E. *J. Med. Chem.* **49**, 7076–7087 (2006).



## Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see [Reporting Life Sciences Research](#). For further information on Nature Research policies, including our [data availability policy](#), see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### ► Experimental design

#### 1. Sample size

Describe how sample size was determined.

We do not describe a particular biological cohort study.

#### 2. Data exclusions

Describe any data exclusions.

No data were excluded.

#### 3. Replication

Describe whether the experimental findings were reliably reproduced.

We provide False Discovery Rate calculations in the supplement information on the question how often BinBase would report false positive or false negative results.

#### 4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

We do not describe a particular biological cohort study.

#### 5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

This is a database paper where all samples are blinded to the investigators.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

#### 6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

- |                                     |  |
|-------------------------------------|--|
| n/a                                 | Confirmed  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> The <u>exact sample size</u> ( <i>n</i> ) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A statement indicating how many times each experiment was replicated   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as an adjustment for multiple comparisons  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The test results (e.g. <i>P</i> values) given as exact values whenever possible and with confidence intervals noted  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A clear description of statistics including <u>central tendency</u> (e.g. median, mean) and <u>variation</u> (e.g. standard deviation, interquartile range)   |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clearly defined error bars  |

See the web collection on [statistics for biologists](#) for further resources and guidance.

## ► Software

Policy information about [availability of computer code](#)

### 7. Software

Describe the software used to analyze the data in this study.

MS-FINDER 2.0, MS-DIAL 2.0 and BinVestigate

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* [guidance for providing algorithms and software for publication](#) provides further information on this topic.

## ► Materials and reagents

Policy information about [availability of materials](#)

### 8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

all software source code and access are freely available

### 9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

n/a

### 10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

n/a

b. Describe the method of cell line authentication used.

n/a

c. Report whether the cell lines were tested for mycoplasma contamination.

n/a

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by [ICLAC](#), provide a scientific rationale for their use.

n/a

## ► Animals and human research participants

Policy information about [studies involving animals](#); when reporting animal research, follow the [ARRIVE guidelines](#)

### 11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

This is a database study without use of particular samples.

Policy information about [studies involving human research participants](#)

### 12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

This is a database study without use of particular samples.