HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2015 December 01.

Published in final edited form as:

Nat Methods. 2015 June ; 12(6): 523–526. doi:10.1038/nmeth.3393.

MS-DIAL: Data Independent MS/MS Deconvolution for Comprehensive Metabolome Analysis

Hiroshi Tsugawa[1,2], **Tomas Cajka**[3], **Tobias Kind**[3], **Yan Ma**[3], **Brendan Higgins**[4], **Kazutaka Ikeda**[5,6], **Mitsuhiro Kanazawa**[7], **Jean VanderGheynst**[4], **Oliver Fiehn**[3,8], and **Masanori Arita**[1,9]

[1]RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa, Japan

[2]Department of Biotechnology, Graduate School of Engineering, Osaka University, Suita, Osaka, Japan

[3]Genome Center, University of California Davis, Davis, California, USA

[4]Department of Biological and Agricultural Engineering, University of California Davis, Davis, California, USA

[5]RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa, Japan

[6]Japan Science and Technology Agency, Kawaguchi, Saitama, Japan

[7]Reifycs Inc., Minato-ku, Tokyo, Japan

[8]Department of Biochemistry, Faculty of Sciences, King Abdulaziz University, Jeddah, Saudi-Arabia

[9]National Institute of Genetics, Mishima, Shizuoka, Japan

Abstract

Data-independent acquisition (DIA) in liquid chromatography tandem mass spectrometry (LC-MS/MS) provides more comprehensive untargeted acquisition of molecular data. Here we provide an open-source software pipeline, MS-DIAL, to demonstrate how DIA improves simultaneous identification and quantification of small molecules by mass spectral deconvolution. For reversed phase LC-MS/MS, our program with an enriched LipidBlast library identified total 1,023 lipid compounds from nine algal strains to highlight their chemotaxonomic relationships.

CORRESPONDING AUTHORS. Oliver Fiehn (ofiehn@ucdavis.edu), Masanori Arita (arita@nig.ac.jp).

Author contributions

H.T., O.F., and M.A. designed the research. H.T. developed the MS-DIAL program. H.T. and T.C. analyzed the samples. T.C. and Y.M contributed to the improvement of MS-DIAL program. H.T., T.K., and Y.M. performed the lipid annotations for the retention time prediction. H.T., T.K., and K.I. improved and optimized the LipidBlast library. H.T., B.H., and J.V. prepared algal samples. M.K. developed the ABF file and the converter for this project. H.T., O.F. and M.A. thoroughly discussed about this project and wrote the manuscript. T.C., T.K., B.H., and J.V. also contributed to the manuscript.

Completing Financial Interests Statement

The authors declare no competing financial interests.

Precursor- or data-independent MS/MS acquisition (DIA) methods in liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) have been gathering attentions for untargeted analysis of biomolecules[1,2]. In contrast to traditional data-dependent MS/MS acquisitions, DIA methods can obtain all fragment ions for all precursors simultaneously, thereby increasing the coverage of observable molecules and reducing false negative identifications. The problem is, however, the contamination of MS/MS spectra due to its wider isolation window (10–25 Da or more) for precursor ion selections. Moreover, the DIA process dissociates the link between precursors and their fragment ions, compromising the molecular identification process.

In proteomics, OpenSWATH software has partly addressed these problems[2]. After extracting product ion chromatograms for the corresponding precursor range, chromatogram peaks are grouped, scored and statistically assessed by false discovery rate (FDR) in the mProphet algorithm[3]. Unfortunately, this approach is not directly applicable to metabolomics. While spectral similarity in shotgun proteomics is probabilistically estimated by presence or absence of peak groups, compound annotations in metabolomics rely on overall match scores of experimental to library spectra. In addition, no FDR calculation schemes by validated decoy techniques exist in metabolomics[4]. Therefore, DIA MS/MS spectra must be purified from fragment ions of co-eluting compounds and noise ions for metabolomic annotations to achieve high overall library matching scores.

The solution is mathematical deconvolution of fragment ions to extract original spectra and to re-associate the precursor-fragment links. A deconvolution approach is also reported by Nikolskiy et al.[5], but their program, decoMS2, requires two different collision energies, low (usually 0V) and high, in each precursor range to solve the mathematical equations. Interestingly, automatic mass spectral deconvolution and identification systems are routine today in gas chromatography coupled to mass spectrometry (GC-MS)[6,7]. DIA-type mass fragmentation schemes are the norm in hard electron-ionization GC-MS in contrast to soft electrospray-ionization LC-MS/MS. Analogous to these successful GC-MS data processing systems, we have developed Mass Spectrometry – Data Independent AnaLysis software (MS-DIAL) that implements a new deconvolution algorithm for DIA data sets. It is a data-processing pipeline for untargeted metabolomics applicable to either data independent or precursor-dependent MS/MS fragmentation methods.

The raw vendor-format data or the common mzML data are first converted into ‘Analysis Base File’ (ABF) format for rapid data retrieval[8] (Fig. 1a). Then, precursor ion peaks are efficiently spotted (hereafter *peak spotting*) by exploring two continuous data axes: retention-time and accurate mass. Each spot represents a detected peak (Fig. 1b), and our MS^{2}Dec algorithm is applied to each spot to deconvolute spectra in the respective precursor ion range. The MS^{2}Dec algorithm first extracts the product spectra for each precursor peak on all MS/MS chromatograms (the raw chromatograms are shown in regular lines in Fig. 1c) to recover the precursor-product links as a result of deconvolution. The deconvolution itself is based on its established GC-MS counterpart[6] with substantial modifications based on accurate mass information instead of nominal masses, enabling analyses of large-scale DIA data sets. It uses the least square optimization to extract ‘model peaks’ (See **Online Methods**) in MS/MS chromatograms (the reconstructed model chromatograms are shown in

thick lines in Fig. 1c). Finally, the pure MS/MS spectrum is determined by the peak heights of reconstructed chromatograms, thus removing the background noise and extracting the spectrum out of co-eluted metabolites (Fig. 1c). Compound identification is performed by means of retention time, mass accuracy, isotope ratio, and MS/MS similarity matching in combination with libraries from publicly available databases (e.g. MassBank[9] and LipidBlast[10]) (Fig. 1d). The MS-DIAL also implements additional functions required for untargeted metabolomics such as peak alignment, filtering and missing value interpolation (**Online Methods**).

MS-DIAL is available on Windows (.NET Framework 4.0 or later; RAM: 4.0 GB or more). The program is downloadable at the PRIMe (http://prime.psc.riken.jp/) website. It supports mzML and major MS vendor formats including Agilent Technologies (.D), AB Sciex (.Wiff), Thermo Fisher Scientific (.RAW), Bruker Daltonics (.D), and Waters (.RAW). The program is intended for large-scale analyses such as cohort studies: it accesses raw data sequentially and keeps only their peak information on memory. The actual processing time for an average 600 MB per assay file in our study was less than 1.2 min with an Intel Core i7-4700MQ CPU (2.4 GHz) with 8 GB RAM on Windows 8.1.

We here used sequential window acquisition of all theoretical mass spectra (SWATH) acquisition as the DIA approach and compared to the traditional data-dependent acquisition (DDA) for validation. The first showcase of our MS^2Dec deconvolution is a human plasma sample with hydrophilic interaction chromatography (Fig. 2 and **Online Methods**). Two metabolites, metoclopramide and norcocaine, exhibited only 1.8 s difference in their elution-time (at 2.95 and 2.98 min, respectively) and fell within the same 25-Da window of the SWATH acquisition. While the unique MS and abundance of metoclopramide could be marginally confirmed in the raw MS/MS spectrum (similarity score 0.72), the spectrum of norcocaine was thoroughly masked under the peaks from metoclopramide (similarity score 0.48) if no mass spectral deconvolution was applied. In this study, the similarity score was calculated by dot-product scoring method (See **Online Methods**). The MS-DIAL program extracted the pure MS/MS spectrum of norcocaine (similarity score 0.80), although contamination of higher mass peaks (e.g. *m/z* 227) was not completely suppressed. The similarity score of metoclopramide was also improved to 0.86 by deconvolution. More examples for the other metabolites are available in Supplementary Fig. 1.

The main showcase is the lipidomic analysis of nine algal species using the LipidBlast library[10]. Prior to the analysis, the library was thoroughly extended to cover major plant and algal lipids such as monogalactosyl, digalactosyl, and sulfoquinovosyl diacylglycerols (MGDG, DGDG, and SQDG) and diacylglyceryl trimethyl homoserine (DGTS) (Supplementary Table 1 and **Online Methods**). Moreover, to improve identification accuracies, we predicted the retention times for all molecules in LipidBlast specifically for our chromatography method by partial least squares regression (PLS-R)[11] on their PaDEL[12] molecular descriptors (**Online Methods**). Predicted retention times exhibited a standard deviation of 0.14 min when compared to retention times of lipid standards, which was almost equivalent to the regressed standard deviation of the actually measured dataset (Fig. 3a and Supplementary Data 1).

We first tested the overall effect of using MS/MS deconvolution on spectral accuracy for lipid profiling at 10 ms accumulation time. Indeed, spectral similarity scores were substantially improved by mass spectral deconvolution in comparison to the raw centroid spectra using 21-Da isolation window, approaching the quality of 1-Da isolation window spectra in targeted acquisitions (DDA) (Fig. 3b). Importantly, the SWATH acquisition with MS-DIAL covered a larger number of phospho- and glycolipids in both positive and negative ionization modes compared to the DDA mode (Fig. 3c and Supplementary Table 2). The only exception was SQDG lipids, whose identification scores worsened due to the low abundance of its characteristic peak (*m/z* 225) (Supplementary Fig. 2). When we re-analyzed the same sample of *Chlamydomonas reinhardtii* under 30 ms accumulation time and a 65-Da isolation window, the number of identified lipids was notably increased not only for SQDG but also for all other lipid classes (Fig. 3c and Supplementary Table 3). Even for this wide isolation window, the deconvoluted MS/MS spectra kept > 90% similarities against the targeted spectra except for SQDG (14:0/16:0) (Supplementary Fig. 2, 3). This result implies that a wider SWATH window appears preferable for lipid profiling in negative mode, while the precursor-isolation windows and accumulation times may need further optimizations. Overall a total of 1,023 lipids were identified, of which the SWATH acquisition covered > 90% (Fig. 3c) and yielded 310 additional lipids that were not detected using the data-dependent MS/MS acquisition (Supplementary Data 2,3).

We conducted hierarchical clustering analysis (HCA) on the lipidomic profile of all 1,023 distinct lipid molecules in nine species to define their overall similarities (**Online Methods**). The clustering result was in full concordance with the commonly accepted phylogenetic tree (Fig. 3d): the nine investigated species were found to clearly cluster, distinguishing the five plantae, three chromista, and one protozoa. Plantae species contained fatty acids of mainly 16 or 18 carbons, whereas protozoa and chromista were comprised of very long-chain (VLC) fatty acids (> 18 carbons). Among plantae, *Chlamydomonas* and *Dunaliella* (chlorophyceae) contained DGTS whereas the two *Chlorella* species (trebouxiophyceae) and UTEX 2341 did not. VLC polyunsaturated fatty acids (PUFAs) of 20 carbons or more such as eicosapentanoic acid or docosahexanoic acid were mostly identified in *Nannochloropsis oculata* and *Euglena gracilis*. In addition, the total quantity of DGTS and PA were highly characteristic to these species (Supplementary Fig. 4, 5). Note that we here used the culture collection ID for the green algae strain UTEX 2341, since its identity is controversial between *Chlorella minutissima* (original classification) and *Nannochloropsis* species[13,14]. Interestingly, our chemotaxonomy suggested that UTEX 2341 is most probably a *Chlorella* rather than *Nannochloropsis*. Moreover, our method led to the detection of lipids that had not been identified previously such as 18:5 PUFA in *C. reinhardtii*, *N. oculata*, and *Pleurochrysis carterae*, odd chain lipids in all nine algal species, and DGTS lipids in *E. gracilis*, *D. salina*, and *N. oculata* (Supplementary Fig. 6). All major lipid classes that have been previously reported in multiple publications[13–16] were all identified in our single experiment.

In summary, MS-DIAL resolves entangled MS/MS spectra in SWATH acquisition by a two-step process: precursor-peak spotting and MS/MS-level deconvolution. With this software, data-independent MS/MS acquisitions can provide high efficacy and accuracy for

metabolome coverage. In this sense, MS-DIAL presents a major step forward to solve the bottleneck in metabolomics: compound identification and annotation[17]. It is important to note that unlike any other software to date, MS-DIAL combines information from four sources: accurate mass, isotope ratios, retention time prediction and MS/MS fragment matching, exceeding the two orthogonal parameters required by the Metabolomics Standards Initiative[18]. A more rationalized confidence score for each parameter setting and their combination will need to be explored in detail for a variety of matrices[19].

Since the MS/MS information potentially includes all detectable ions, the DIA results allow analyses *a posteriori* and alleviate the otherwise incurring cost to re-analyze the same samples with different precursor-selection. Although we purposely focused on compound identifications instead of quantitative aspects, MS-DIAL software substantially supports normalization methods, as required for specific needs in quantitative projects. In addition, MS-DIAL can be used with other DIA methods such as All-ions MS/MS, MSc^2, and All Ion Fragmentation[20]. Accuracy of deconvolution results, however, will depend on data acquisition scan speed and parameter settings such as scan width, overall sensitivity and data accumulation types. While SWATH data acquisition fits well with our deconvolution method, MS-DIAL also supports flexible precursor-mass windows from low to high *m/z*. It will also benefit the proteomics community, where true mass spectral deconvolution has not been commonly utilized for peptide identifications.

Online Methods

Peak detection

Smoothing—The peak detection algorithm starts with a smoothing method with respect to retention time and accurate mass. The MS-DIAL program utilizes the linearly weighted smoothing average[21] that is simple and robust (Supplementary Note Equation 1). The software also supports several other smoothing methods: moving average[21], Savitzky-Golay[21] and binomial filter[22].

Peak Detection—The basic concept of peak detection algorithm consists of differential calculus and noise estimations (Supplementary Fig. 7). After the smoothing for retention time (against extracted ion chromatogram) or accurate mass (against mass chromatogram), peak detection is performed[8]. To evaluate noise, the program determines three threshold values automatically: (1) the maximum amplitude differences between two adjacent points, (2) the maxima of the first derivatives, and (3) the maxima of second derivatives in a chromatogram. The derivatives are calculated by five-point approximations (Supplementary Note Equation 2 and 3). From only the values below 5% of each maximum, medians of amplitude differences, first derivatives, and second derivatives are computed as the threshold values for peak detection and are hereafter called AF (amplitude filter), FF (first-order derivative filter), and SF (second-order derivatives filter), respectively. When a computed median is near zero, 0.0001 is used instead.

The left edge of peaks is recognized when the amplitude and the first-order derivative both exceed AF and FF in two adjacent points. In order to locate the edge more accurately, the local minimum of the adjacent 5-point window is explored by back-tracing from the

detected start position. The peak top is recognized when the sign of the first-order derivative changes and the second-order derivative is less than SF. The right edge is recognized by the same criteria as the left.

Peak spotting

The term, 'peak spotting', is derived from the visualization method in the MS-DIAL software and refers to peak detection based on retention time and MS1 data axes. The base peak chromatogram is formed for each mass slice of 0.1 *m/z* with a step size of 0.05 *m/z* (default), allowing all data points to belong to two adjacent slices (Supplementary Fig. 8a). Each data point of the base peak chromatogram has its scan number, retention time, base peak *m/z*, and base peak intensity. The peak detection algorithm as described above is applied to the base-peak chromatogram and detected peak-tops are shown as 'spots'. Two spots of the same retention time and close *m/z* value in adjacent bins are merged by comparing their peak heights (Supplementary Fig. 8b). Although useful algorithms for automated noise- and background reduction have been known[23], we chose to exclude unwanted peaks simply by means of a user-defined exclusion mass list.

Centroiding spectra

When the profile mode data is analyzed in the MS-DIAL program, the spectral centroiding is simply performed (Supplementary Note Equation 4): After the same peak detection algorithm described above is performed, the ions in the user-defined region between the peak's left and right edges are accumulated.

MS^{2}Dec deconvolution

The MS^{2}Dec procedure is applied to all spots that are detected in the peak spotting method. It consists of (1) centroiding of MS/MS peaks that are consistent to each spot, (2) extraction of the MS/MS chromatogram for each centroided spot, (3) smoothing and baseline correction, (4) model peak extraction, and (5) model peak fitting for each MS/MS chromatogram by means of the least square method.

In contrast to the GC-MS approach[6], we process high-resolution mass (instead of nominal mass) for the de-convolution. Specifically, we extract MS/MS chromatograms of the centroid MS/MS spectrum corresponding to the precursor ion detected in the spotting process, i.e., the precursor ion of the DIA MS/MS spectrum. For each spot, all MS/MS chromatograms within the following retention time (RT) range are extracted:

$$\mathrm{RTrange} \in (\mathrm{peak\ top\ retention\ time} - 1.5 \times \mathrm{peak\ width}, \mathrm{peak\ top\ retention\ time} + 1.5 \times \mathrm{peak\ width})$$

The peak width is the region between peak left- and right edges of the focused peak spot.

After smoothing, each MS/MS chromatogram is subjected to the following baseline correction (Supplementary Fig. 9).

Step 1. Each chromatogram is separated into a user-defined 'segment' value.

Step 2. Local minimum within a user-defined 'band width' value is extracted and stored.

Step 3. The median value of minimal points is computed for each segment, and points more than the median are discarded.

Step 4. Baseline is determined by connecting the remaining minimal points.

Then, the peak detection algorithm is applied to each baseline-corrected MS/MS chromatogram. For each detected peak, two scores, *ideal slope* (Supplementary Note Equation 5–7) and *sharpness* values (Supplementary Note Equation 8–10), are calculated. For these values please refer to the previous work[24].

Examples of our least square method are shown in Supplementary Data 4. In our software program, ideal slope score of > 0.95 is necessary to be considered 'a model peak'. For model peak candidates, their sharpness scores and the scan number are stored, and the second Gaussian derivative filter (Supplementary Note Equation 11) is fit to their array of sharpness values. The data point region selected for each model peak candidate is described using a model peak chromatogram $M(n)$ which has the baseline corrected chromatogram information from peak left- to peak right edge. The least square method for the deconvolution is performed as follows.

$$C(n)=aM_1(n)+bM_2(n)+cM_3(n)+dn+e \quad \text{(eq. 1)}$$

The original chromatogram $C(n)$ is decomposed into three base vectors $M_1(n)$, $M_2(n)$, and $M_3(n)$. One vector, $M_2(n)$, corresponds to the model peak from the focused peak spot sided by two adjacent peaks on both sides, $M_1(n)$ and $M_3(n)$. Note that the purpose of deconvolution is to determine $M_k(n)$ (k = 1, 2, 3) and coefficients *a, b, c, d,* and *e*. So far, $M_k(n)$ is determined as described above. Here, the coefficients are calculated as $b = X^{-1}Y$ as shown below.

$$\begin{pmatrix} a \\ b \\ c \\ d \\ e \end{pmatrix} = \begin{pmatrix} \|M_1(n)\|^2 & (M_1(n),M_2(n)) & (M_1(n),M_3(n)) & (M_1(n),n) & (M_1(n),1) \\ (M_2(n),M_1(n)) & \|M_2(n)\|^2 & (M_2(n),M_3(n)) & (M_2(n),n) & (M_2(n),1) \\ (M_3(n),M_1(n)) & (M_3(n),M_2(n)) & \|M_3(n)\|^2 & (M_3(n),n) & (M_3(n),1) \\ (n,M_1(n)) & (n,M_2(n)) & (n,M_3(n)) & \|n\|^2 & (n,1) \\ (1,M_1(n)) & (1,M_2(n)) & (1,M_3(n)) & (1,n) & \|1\|^2 \end{pmatrix}^{-1} \begin{pmatrix} (M_1(n),C(n)) \\ (M_2(n),C(n)) \\ (M_3(n),C(n)) \\ (n,C(n)) \\ (1,C(n)) \end{pmatrix} \quad \text{(eq. 2)}$$

If the siding model peak $M_1(n)$ or $M_3(n)$ is not found within the extracted retention time region, the Equation 3, 4, and 5 will be changed to the corresponding matrix: $X(3 \times 3)$, $b(3 \times 1)$, and $Y(3 \times 1)$ are used when both $M_1(n)$ and $M_3(n)$ are not found, and $X(4 \times 4)$, $b(4 \times 1)$, and $Y(4 \times 1)$ are used when either $M_1(n)$ or $M_3(n)$ is not found. In the special case when $M_2(n)$ is not found (this case would be possible when all MS/MS chromatograms are impure), an *ad hoc* model peak is inserted (instead of a *null* value) as follows. When the peak spotting algorithm is performed in RT vs. MS1 axis, model peaks of ideal slope value 1 are stored as peak candidates. Then, one model peak which has the median sharpness value within the candidates is used as the *ad hoc* model peak. Although inserting a *null* value is

another option, we hypothesize that a compound was missed in the detected spot in RT vs. MS1.

Compound identification

The software program utilizes the NIST MS format (NIST MSP ASCII) file for the reference library. Four criteria, (1) retention time, (2) accurate mass, (3) isotope ratio, and (4) MS/MS spectrum information, are used for peak identification. Each score gives the standardized range from 0 to 1, meaning no similarity and a perfect match, respectively. The subscript '*act.*' and '*lib.*' of each equation describe the measurement value and the theoretical value, respectively.

Accurate mass and Retention time similairties—

$$Accurate\ mass(MS1)\ or\ RT\ similarities = exp\left\{-0.5 \times \left(\frac{Experimental\ value - Theoretical\ value}{\delta}\right)^2\right\} \quad \text{(eq. 3)}$$

The background hypothesis of the equations for accurate mass and retention time similarities is that the differences between experimental and theoretical values follow the Gaussian distribution. The standard deviation δ (user-defined) is also used as the search tolerance. If retention time information is not included in the MSP file of metabolites, the similarity value of retention time is not calculated.

Isotope ratio—If metabolite information in the MSP file includes the molecular formula, the theoretical isotopic distribution is calculated from [M+0] to [M+5] by means of binomial and McLaurin expansion. An example for C_2H_6O is as follows.

$$({}^{12}\mathrm{C}+{}^{13}\mathrm{C})^2({}^{1}\mathrm{H}+{}^{2}\mathrm{H})^6({}^{16}\mathrm{O}+{}^{17}\mathrm{O}+{}^{18}\mathrm{O}) = ({}^{12}\mathrm{C}_2{}^{1}\mathrm{H}_6{}^{16}\mathrm{O})\left(1+\frac{{}^{13}\mathrm{C}}{{}^{12}\mathrm{C}}\right)^2\left(1+\frac{{}^{2}\mathrm{H}}{{}^{1}\mathrm{H}}\right)^6\left(1+\frac{{}^{17}\mathrm{O}}{{}^{16}\mathrm{O}}+\frac{{}^{18}\mathrm{O}}{{}^{16}\mathrm{O}}\right) \quad \text{(eq. 4)}$$

Here, the letter such as ^{12}C shows the natural abundance of each element. The contents except for the molecular mass [M+0] ($^{12}C_2{}^{1}H_6{}^{16}O$) is expanded. Note that each coefficient value of expanded elements indicates the relative, i.e., isotope abundances with respect to the molecular ion ($^{12}C_2{}^{1}H_6{}^{16}O$). Then, the relative abundances are compared between theoretical values and actual values. The intensity of [M+0] is normalized to 1. The similarity value of the isotope ratio is calculated as follows.

$$Isotope\ ratio\ similarity = 1 - \sum |r_{act.i} - r_{lib.i}| \mathrm{with} r_i = \frac{I_{M+i}}{I_M},\ 1 \le i \le 5 \quad \text{(eq. 5)}$$

The I_M and I_{M+i} show the intensity of the molecular ion and the isotope peak, respectively.

Spectral similarity—For the MS/MS spectral similarity, the MS-DIAL program utilizes the combined values of dot-product, reverse dot-product, and the matched fragments ratio with the reference product ions. The weight among the dot-product, reverse dot-product, and the matched fragments ratio is 1:1:1 in the current MS-DIAL software setting. The

amplitude of mass spectrum is normalized so that the highest amplitude of the product value becomes 1. The abundance *A* of each *m/z* is the integrated value within the user-defined MS^2 tolerance. The dot product calculation in the MS-DIAL program are performed as follows.

$$dot\ product = \frac{(\sum wA_{act.} wA_{lib.})^2}{\sum wA_{act.}{}^2 \sum wA_{lib.}{}^2} \text{with} w = 1/\left(1 + \frac{A}{\sum A - 0.5}\right) \quad \text{(eq. 6)}$$

$\sum A$ is the sum of relative abundances. The reverse dot product is also calculated in the same way (Supplementary Note Equation 12). The coefficient *w* is the weight value in order to reduce the effect of high abundance intensities. In the dot-product calculation, the half abundance of the measured spectrum is used if the corresponding mass-peak does not exist in a library spectrum. Un-wanted peaks derived from isotopic ions or back ground noise may decrease the dot-product score. In the reverse dot-product, the spectrum in the reference library is used to calculate the score. The ion abundance of the reference spectrum is halved when the pairing mass-peak does not exist in the query spectrum.

Total similarity—The four scores are utilized for compound identification.

$$Total\ score = \frac{MS/MS similarity + MS1 similarity + RT similarity + 0.5 \times isotope\ ratio\ similarity}{3.5} \times 100 \quad \text{(eq. 7)}$$

If retention time information is not available for compound identification, the total score is calculated as Supplementary Note Equation 13. If the formula information is not available, the total score is calculated as Supplementary Note Equation 14. If both retention time and formula information are not available, the total score is calculated as Supplementary Note Equation 15. The compound with highest total score (above the user-defined threshold) is assigned to each focused peak. When the MS/MS spectrum is not obtained for data dependent MS/MS acquisition, the MS/MS similarity is recognized as zero and the denominator described above is decremented by 1.

Peak alignment, filtering and missing value interpolation

The algorithm of peak alignment in MS-DIAL is based on the idea of Joint Aligner implemented in MZmine[26]. It consists of four major steps: (1) making a reference table, (2) fitting each sample peak table to the reference peak table, (3) filtering aligned peaks, and (4) interpolating missing values. The summary of MS-DIAL peak alignment algorithm is described in Supplementary Fig. 10.

Making a reference peak table—As shown in Supplementary Fig. 10a, the reference peak table consisting of retention time (RT) and *m/z* is created as follows.

Step 1. A user-defined ‘reference file’ which is one of the aligned samples is used as the basis of reference peak table.

Step 2. Information of each sample peak table is inserted to the reference peak table (Supplementary Fig. 10b). The condition is as follows.

$$\text{if} |RT_{sam.} - RT_{ref.}| > \delta_{RT} \cup |Mass_{sam.} - Mass_{ref.}| > \delta_{Mass} \; then \; insert \; to \; peak \; table \quad \text{(eq. 8)}$$

δ_{RT} and δ_{Mass} are user-defined tolerance values for RT (δ_{RT}) and MS1 accurate mass (δ_{Mass}), respectively.

Step 3. Repeat the function for all peaks from all samples.

The reference peak table is utilized in order to associate each peak in each sample.

Fitting each sample peak table to the reference peak table—Each peak in the sample data is associated with the reference peak list using the following criterion.

$$Score = a \times exp\left\{-0.5 \times \left(\frac{RT_{sam.} - RT_{ref.}}{\delta_{RT}}\right)^2\right\} + b \times exp\left\{-0.5 \times \left(\frac{Mass_{sam.} - Mass_{ref.}}{\delta_{Mass}}\right)^2\right\} \quad \text{(eq. 9)}$$

The coefficient is user-defined RT factor (*a*) and MS1 accurate mass factor (*b*), respectively. δ_{RT} and δ_{Mass} are the same as the above criteria to construct the reference peak table. Finally, aligned peak table including alignment ID, average RT, average *m/z*, and intensities of all samples is generated.

Filtering aligned peaks—MS-DIAL provides the simple filter in order to exclude unwanted alignment ID (Supplementary Fig. 10c). Three step filtering is applied for each alignment ID.

Step 1. If all peak intensities of samples in a row are missing or undetected, the alignment information is removed.

Step 2. If the percentage of filled peaks in an alignment ID is less than the user-defined peak count filter (default 0%), the information is removed.

Step 3. This is optional but if all quality control (QC) samples are not filled in an alignment ID, the information is removed.

Interpolating missing values—In each alignment ID, the intensity information of all samples is not always filled. As shown in Supplementary Fig. 10d, such missing values after the above process are interpolated in MS-DIAL as follows.

Step 1. The average retention time and average *m/z* of 'filled' peaks are calculated.

Step 2. A local maximum from the following range is stored for the missing value.

$$(RT_{average} - \delta_{RT}, RT_{average} + \delta_{RT}) \cap (Mass_{average} - \delta_{Mass}, Mass_{average} + \delta_{Mass}) \quad \text{(eq. 10)}$$

Databases

The MassBank revision 173, ReSpect updated in 2012/9/25, and LipidBlast version 3 were downloaded. The spectrum data were converted to the NIST MSP format. For the hydrophilic metabolite identification, the NIST 12 MS/MS library was also utilized in

addition to MassBank and ReSpect libraries. For the algal lipid identification, fatty acid 16:2, 16:3, 16:4, and 16:5 spectrum information were added to LipidBlast library. The position of double bonds was determined according to previous reports[16]. Moreover, the adduct ions and the MS/MS spectral information of formic acid were added to PC, lysoPC, MGDG, and DGDG for the lipid identification in negative ion mode analysis. In order to determine the ion abundances for each lipid class the heuristic model was constructed from the data sets of DDA MS/MS as described above. The MSP format libraries (MassBank, ReSpect, and LipidBlast), and the LipidBlast excel macro file are downloadable under http://prime.psc.riken.jp/.

Retention time prediction for lipids

The SDF files of all lipids in LipidBlast were constructed as follows. The SDF files of PC, lysoPC, PE, lysoPE, PG, PI, PS, and PA were downloaded from LIPID MAPS[26]. The SDF files for the other lipid classes were created from SMILES code written in LipidBlast by ChemAxon JChem 6.3.0 molconvert (http://www.chemaxon.com), totaling 117,343 SDF files. They also included plasmenyl PC, PE, sphingomyelin, and cholesterol ester as lipid classes although these lipids were not the focus for algal lipid identifications. The PaDEL descriptor software was utilized to calculate 1D and 2D molecular descriptors and PubChem fingerprints from the SDF files[12]. Their exact masses were also generated by ChemAxon JChem molconvert. Then, redundant and uniform variables were excluded, and a total of 464 compound descriptors were used as predictor variables in the regression analysis. The in-house retention time information of 254 lipids was used for model development. Since the number of predictor variables (compound descriptors) were considerably higher than the number of data samples (the number of training set: 254), partial least square regression (PLS-R) was utilized in order to construct the retention time prediction model[11]. The program of PLS-R was written in Visual Basic for Application and the source code can be downloaded at http://prime.psc.riken.jp/. A seven-fold cross validation was used to calculate the predictive residual sum of squares (PRESS) and Q^2 value. The final model included six latent variables based on the PRESS and Q^2 value and the retention time information from the training samples. In this study, retention time information of newly identified 1,808 lipids from nine algal species was used for validating that accurate precursor ion masses and MS/MS spectra were also confirmed by retention time matching.

MS-DIAL software and data processing parameters

MS-DIAL is available in Windows OS (.NET Framework 4.0 or later; RAM: 4.0 GB or more). Its source code was written in the C# language with the windows presentation foundation (WPF) to develop the graphical user interface. The main source code such as peak detection, peak spotting, and MS^{2}Dec algorithm is downloadable at http://prime.psc.riken.jp/. The data processing parameter of MS-DIAL utilized in this study are described in Supplementary Table 4.

Chemotaxonomic tree by lipid descriptors and hierarchical clustering analysis

All lipids were annotated with subsequent manual verification of MS/MS spectral matching for compound identifications (Supplementary Table 1). A total of 1,808 (SWATH) and

1,521 (DDA) lipids (Supplementary Table 2) were first integrated disregarding the acyl chain positions (*sn1, sn2, sn3*), double bond positions, and stereoisomers (*E, Z*). For example, TAG(16:0/16:1/16:2), TAG(16:0/16:2/16:1), TAG(16:2/16:1/16:0), TAG(16:2/16:0/16:1), TAG(16:1/16:0/16:2), and TAG(16:1/16:2/16:0) were considered the same lipid. Likewise, lysoPC 16:1(7Z) and lysoPC 16:1(9E) were regarded as the same. For the remaining 1,023 lipids, presence or absence in each of nine species was represented as a binary data matrix of size 1023×9 (Supplementary Data 2).

Hierarchical clustering analysis was performed using the R statistical language (http://www.R-project.org) and the package 'amap' (http://CRAN.R-project.org/package=amap). The distance was calculated by 'correlation' in the package. The linkage was performed by 'average'. We cited the previous report[27] as the standard taxonomic tree.

Biospecimen and Algae strains

A single human plasma sample was obtained from the Cleveland Clinic from the GeneBank study[28]. The cultivation procedure of *Chlamydomonas reinhardtii* followed our previous report[29]. The *C. reinhardtii* CC125 strain was streaked out from cryopreserved stock and cultivated in 75 mL TAP medium in 125 mL shake flasks at 25 °C under constant illumination with cool-white fluorescent bulbs at a fluence rate of 70 μmol m^{-2} s^{-1} and with continuous stirring (100 rpm). Four independent cultures were used for this study. The starter culture was harvested at late log-phase and 1 mL cell suspensions were then shifted to 75 mL of fresh TAP medium in 125 mL shake flasks. At 0.2–0.6 OD_{680} during the late-log phase, 1 mL cell suspensions were injected into 1 mL of –80 °C cold quenching solution composed of 70% methanol in water, centrifuged at 12,000 g for 2 min, and pellets were lyophilized and stored at −80 °C until further analysis. The same quenching procedure was used for all algae strains.

UTEX 2341 (originally classified as *Chlorella minutissima*), *Chlorella sorokiniana* (UTEX 2805), and *Chlorella variabilis* (ATCC NC64A) were plated on ATCC #5 agar[30] and colonies were selected for inoculation into liquid cultures. All three *Chlorella* strains were cultivated simultaneously in 250 ml hybridization tubes with four independent cultures per strain. Hybridization tubes were filled with 200 ml media and maintained in a 28 °C water bath. Aeration was supplied at 125 ml per minute with 2% CO_2 mixed with air (v/v). Reactors were illuminated horizontally (10,000 lux) by T5 growth lamps operating on a 16:8 light/dark cycle and cultures were mixed by stir bar operating at ~150 rpm. UTEX 2341 was cultivated in N8-NH_4 medium[31], *C. sorokiniana* in N8 medium[32], and *C. variabilis* in N8-NH_4 medium supplemented with 20 mg/L yeast extract. Culture samples (1 ml) were quenched for lipidomics analysis during the late log growth stage.

The cultures of *Euglena gracilis* (UTEX B367), *Cricosphaera carterae* (UTEX LB1014), *Nannochloropsis oculata* (UTEX LB2164), *Dunaliella salina* (UTEX LB200), and *Pavlova lutheri* (UTEX LB1293) were purchased from the UTEX culture collection of algae[33]. Three technical replicates for each strain were prepared from quenched samples.

Reagent and Sample preparation

Water, isopropanol, and acetonitrile were purchased from Fisher Optima. Methanol was purchased from J.T. Baker. Ammonium formate, formic acid and methyl tert-butyl ester (MTBE) were purchased from Sigma Aldrich. Authentic standard compounds were purchased from Avanti Polar Lipids Inc., CDN Isotopes, Cayman Chemical, and Sigma Aldrich.

For hydrophilic interaction chromatography-MS/MS analysis of pharmaceutical agents present in a human plasma sample, all procedures for the metabolite extraction were kept on ice. 30 μL of human plasma was added to 1000 μL cold mix-solvent (acetonitrile/ isopropanol/water, 3:3:2, v/v/v) on ice, then vortexed for 10 s, and shaken for 5 min at 4 °C using the Orbital Mixing Chilling/Heating Plate (Torrey Pines Scientific Instruments). After 2 min centrifugation at 14,000 rcf, 300 μL of the supernatant was transferred to a new 1.5 mL Eppendorf tube and evaporated to dryness in a Labconco Centrivap cold trap concentrator. The dried sample was re-suspended with 60 μL (80% acetonitrile in water) including 0.038 μg/mL choline-D_9, 0.050 μg/mL TMAO-D_9, 0.020 μg/mL betaine-D_9, 10.0 μg/mL glutamine-D_5, and 1.48 μg/mL arginine-$^{15}N_2$ and centrifuged for 5 min at 16,000 rcf. The 50 μL aliquot was transferred to a glass amber vial (National Scientific) with a micro-insert (Supelco).

For lipid profiling, all samples for the metabolite extraction were kept on ice and performed as described previously[34]. 225 μL of MeOH including 1.64 μg/mL PE (17:0/17:0), 6.55 μg/mL PG (17:0/17:0), 1.10 μg/mL PC (17:0/0:0), 0.24 μg/mL sphingosine (d17:1), 0.55 μg/mL ceramide (d18:1/17:0), 0.44 μg/mL SM (d18:1/17:0), 54.5 μg/mL palmitic acid-D3, 0.44 μg/mL PC (12:0/13:0), 22.7 μg/mL cholesterol-D7, 0.27 μg/mL TAG (17:0/17:1/17:0), 2.18 μg/mL DAG (12:0/12:0/0:0), 13.1 μg/mL DAG (18:1/2:0/0:0), 4.36 μg/mL MAG (17:0/0:0/0:0), and 0.55 μg/mL PE (17:1/0:0) were added to each dried algae on ice and vortexed for 10 seconds. Then, the MTBE including 21.8 μg/mL cholesteryl ester (22:1) was added on ice and vortexed for 10 seconds. After shaking for 6 min at 4 °C in the orbital mixer, 188 μL water was added and vortexed for 20 s. After centrifugation for 2 min at 14,000 rcf, 350 μL of the supernatant was transferred to a new 1.5 mL Eppendorf tube and evaporated to dryness in the Labconco Centrivap cold trap concentrator. The dried sample was re-suspended in 108.6 μL MeOH:toluene 90:10 (v/v) with CUDA (12-[[(cyclohexylamino)carbonyl]amino]-dodecanoic acid, 50 ng/mL). After vortexing for 20 s, each sample was sonicated for 5 min at room temperature. After centrifugation for 2 min at 16,000 rcf, 50 μL of the supernatant was transferred to a glass amber vial with micro-insert. The *C. reinhardtii*, *C. sorokiniana*, and *C. variabilis* samples were diluted by adding 50 μL of MeOH:toluene 90:10 (v/v). Moreover, the *E. gracilis* sample was diluted by adding 200 μL of MeOH:toluene 90:10 (v/v).

Analytical conditions

The liquid chromatography system consisted of an Agilent 1290 system (Agilent Technologies Inc.) with a pump (G4220A), a column oven (G1316C), and an autosampler (G4226A). For hydrophilic metabolite analysis, mobile phase A was 10 mM ammonium formate with 0.125 % formic acid in water; mobile phase B was 95:5 acetonitrile:water (v/v)

with 10 mM ammonium formate and 0.125 % formic acid. An Acquity UPLC BEH Amide column (150×2.1 mm; 1.7 μm) coupled to a VanGuard BEH Amide pre-column (5×2.1 mm; 1.7 μm) (Waters; Milford, MA, USA) was used. 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. The column flow rate was 0.4 mL/min, autosampler temperature was 4 °C, injection volume was 2 μL, and column temperature was 45 °C. For lipid analysis, mobile phase A was 60:40 acetonitrile:water (v/v) with 10 mM ammonium formate and 0.1% formic acid; mobile phase B was 90:10 isopropanol:acetonitrile (v/v) with 10 mM ammonium formate and 0.1% formic acid.

The lipidomic LC method utilized an Acquity UPLC charged-surface hybrid (CSH) C18 column (100×2.1 mm; 1.7 μm) coupled to an Acquity CSH C18 VanGuard pre-column (5×2.1 mm; 1.7 μm) (Waters; Milford, MA, USA). The gradient was 0 min, 15% B; 2 min, 30% B; 2.5 min, 48% B; 11 min, 82% B, 11.5 min, 99% B; 12 min, 99% B; 12.1 min, 15% B; 15 min, 15% B. The column flow rate was 0.6 mL/min, autosampler temperature was 4 °C, injection volume was 3 μL in positive mode and 5 μL in negative mode, and column temperature was 65 °C.

Mass spectrometry was performed on an AB Sciex TripleTOF 5600+ system (Q-TOF) equipped with a DuoSpray ion source. All analyses were performed at the high sensitivity mode for both TOF MS and product ion scan. The mass calibration was automatically performed every 10 injections using an APCI positive/negative calibration solution via a calibration delivery system (CDS). For hydrophilic interaction chromatography analysis, SWATH (sequential window acquisition of all theoretical mass spectra) acquisition with positive ion mode was used as the data independent acquisition system. The SWATH parameters were MS^1 accumulation time, 50 ms; MS^2 accumulation time, 30 ms; collision energy, 45 V; collision energy spread, 15 V; cycle time, 640 ms; Q1 window, 25 Da; mass range, *m/z* 50–500. The other parameters were curtain gas, 35; ion source gas 1, 50; ion source gas 2, 50; temperature, 300 °C; ion spray voltage floating, 4.5 kV; declustering potential, 100 V; RF transmission, *m/z* 40: 33%, *m/z* 120: 33%, and *m/z* 390: 34%. For lipid analysis, six different methods were used; DDA (data-dependent acquisition) with positive ion mode, DDA with negative ion mode, SWATH acquisition (Q1 window, 21 Da) with positive ion mode, SWATH acquisition (Q1 window, 21 Da) with negative ion mode, SWATH acquisition (Q1 window, 65 Da) with positive ion mode, and SWATH acquisition (Q1 window, 65 Da). The common parameters in both SWATH/DDA and positive/negative ion mode were collision energy, 45 V; collision energy spread, 15 V; mass range, *m/z* 100–1250; curtain gas, 35; ion source gas 1, 60; ion source gas 2, 60; temperature, 350 °C; declustering potential, 80 V; RF transmission, *m/z* 80: 50%, *m/z* 200: 50%. The ion spray voltage floating of positive/negative ion mode were +5.5/–4.5 kV, respectively. The DDA parameters in both positive and negative ion modes were MS^1 accumulation time, 100 ms; MS^2 accumulation time, 50 ms; cycle time, 650 ms; dependent product ion scan number, 10; intensity threshold, 500; exclusion time of precursor ion, 5 s; mass tolerance, 20 mDa; ignore peaks, within 6 Da; dynamic background subtraction, TRUE. The SWATH parameters of 21/65 Da Q1 window were MS^1 accumulation time, 100/50 ms; MS^2 accumulation time, 10/30 ms; cycle time, 731/640 ms; Q1 window, 21/65 Da.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the NSF-JST Strategic International Collaborative Research Program (SICORP) for JP-US metabolomics. We also thank the Lipid MAPS consortium for providing us the lipid SDF files; ChemAxon for a free research license for the Marvin and JChem cheminformatics tools; Z. Tietel and N. Nguyen (UC Davis) for assisting with the sample preparation of algal species; T. Bamba, Y. Izumi, and T. Yamada (Osaka University) for suggestions and discussion of lipid annotation; D. Yukihira (Kyushu University) for discussion of retention time prediction; and A. Ogiwara (Reyfics Inc.) for development of the ABF file and for suggestions and discussion of MS-DIAL development. H.T. was also supported by Grant-in-Aid for Young Scientists (B) 25871136. This study was also supported by the US National Science Foundation (NSF MCB 113944), National Institutes of Health (NIH) (Grants P20 HL113452 and U24 DK097154), the Japan Science and Technology Agency (JST)-Core Research for Evolutionary Science and Technology (JST-CREST), and Database Integration Coordination Program by National Bioscience Database Center.

References

1. Zhu X, Chen Y, Subramanian R. Anal. Chem. 2014; 86:1202–1209. [PubMed: 24383719]
2. Röst HL, et al. Nat. Biotechnol. 2014; 32:219–223. [PubMed: 24727770]
3. Reiter L, et al. Nat. Methods. 2011; 8:430–435. [PubMed: 21423193]
4. Tsugawa H, et al. Anal. Chem. 2013; 85:5191–5199. [PubMed: 23581547]
5. Nikolskiy I, et al. Anal. Chem. 2013; 85:7713–7719. [PubMed: 23829391]
6. Stein SE. J. Am. Soc. Mass. Spectrum. 1999; 10:770–781.
7. Fiehn O, Wohlgemuth G, Scholz M. Proc. Lect. Note Bioinform. 2005; 3615:224–239.
8. Tsugawa H, Kanazawa M, Ogiwara A, Arita M. Bioinformatics. 2014; 30:2379–2380. [PubMed: 24753485]
9. Horai H, et al. J. Mass Spectrom. 2010; 45:703–714. [PubMed: 20623627]
10. Kind T, et al. Nature Methods. 2013; 10:755–758. [PubMed: 23817071]
11. Wold S, Sjostrom M, Eriksson L. Chemometr. Intell. Lab. 2001; 58:109–130.
12. Yap CW. J. Comput. Chem. 2011; 32:1466–1474. [PubMed: 21425294]
13. Gladu PK, Patterson GW, Wikfors GH, Smith BC. Journal of Phycology. 1995; 31:774–777.
14. Kind T, et al. J. Chromatogr. A. 2012; 1244:139–147. [PubMed: 22608776]
15. Haigh WG, et al. Biochim. Biopys. Acta. 1996; 19:183–190.
16. Giroud C, Gerber A, Eichenberger W. Plant Cell Physiol. 1988; 29:587–595.
17. Schymanski EL, Neumann S. Metabolites. 2013; 3:412–439. [PubMed: 24957999]
18. Sumner LW, et al. Metabolomics. 2007; 3:211–221. [PubMed: 24039616]
19. Creek DJ, et al. Metabolomics. 2014; 10:350–353.
20. Egertson JD, et al. Nat. Methods. 2013; 10:744–746. [PubMed: 23793237]

Methods-only reference

21. Savitzky A, Golay MJE. Anal. Chem. 1964; 36:1627–1639.
22. Lommen A. Anal. Chem. 2009; 81:3079–3086. [PubMed: 19301908]
23. Windig W, Phalp JM, Payne AW. Anal. Chem. 1996; 68:3602–3606.
24. Hiller K, et al. Anal. Chem. 2009; 81:3429–3439. [PubMed: 19358599]
25. Katajamaa M, Miettinen J, Oresic M. Bioinformatics. 2006; 22:634–636. [PubMed: 16403790]
26. Sud M, et al. Nucleic Acids Research. 2006; 35:527–532.
27. Cavalier-Smith T. Biological Reviews. 2007; 73:203–266. [PubMed: 9809012]
28. Wang Z, et al. Nature. 2011; 472:57–63. [PubMed: 21475195]

29. Lee DY, Park J-J, Barupal DK, Fiehn O. Mol. Cell. Proteomics. 2012; 11:973–988. [PubMed: 22787274]
30. ATCC. ATCC Medium 5: Sporulation Agar. 2013 http://www.atcc.org.
31. Higgins B, VanderGheynst J. PloS one. 2014; 9:e96807. [PubMed: 24805253]
32. Tanadul OU, Vandergheynst JS, Beckles DM, Powell AL, Labavitch JM. Biotechnol Bioeng. 2014; 111:1323–1331. [PubMed: 24474069]
33. Brand, JJ.; Andersen, RA.; Nobles, DR, Jr. Applied Phycology and Biotechnology. Second Edition. Oxford, UK: John Wiley & Sons, Ltd; 2013.
34. Matyash V, Liebisch G, Kurzchalia TV, Shevchenko A, Schwudke D. J. Lipid Res. 2008; 49:1137–1146. [PubMed: 18281723]

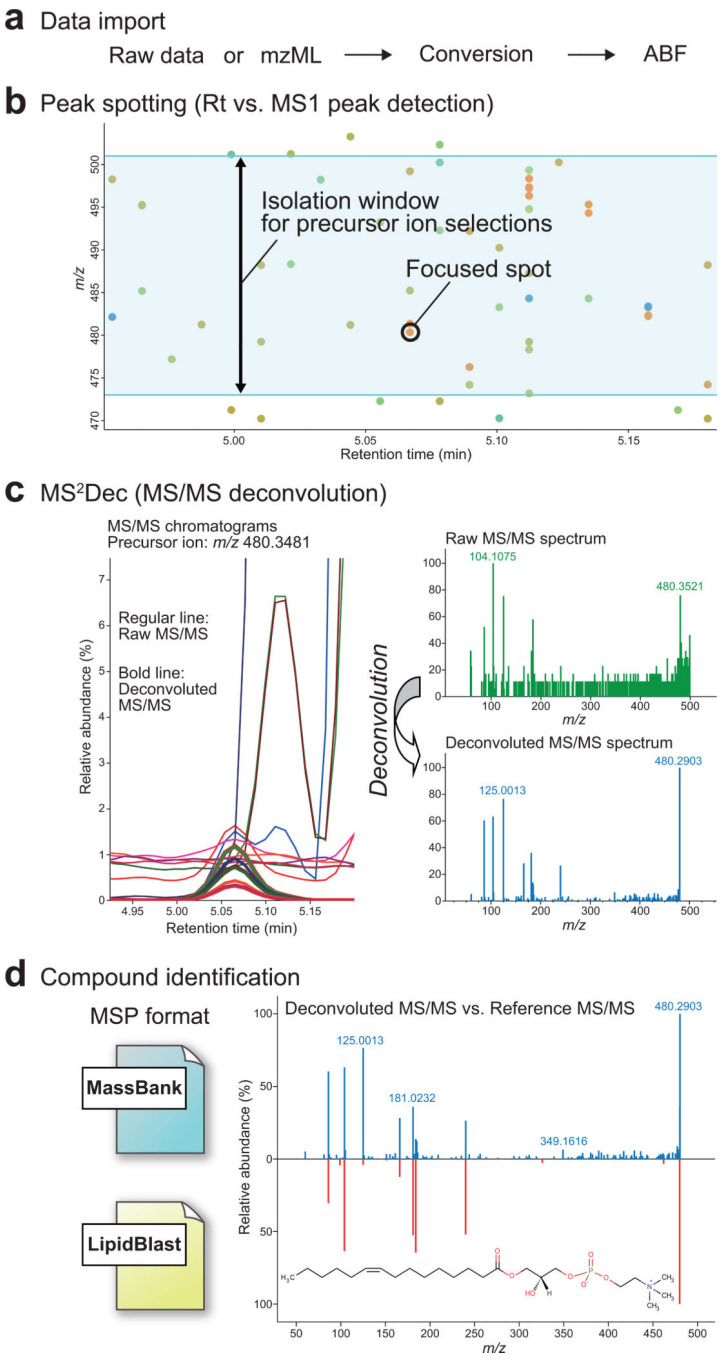

Figure 1. Main workflow of MS-DIAL program

(**a**) MS vendor format or mzML is converted to ABF binary format for rapid data retrieval. (**b**) Peak spotting (two-dimensional peak detection, see main text) is performed to determine precursor ions for MS/MS spectra. The detected precursor ions are described as spots. The blue color range describes the isolation window of precursor ions. The focused spot is also depicted as the following procedures. (**c**) The MS^{2}Dec deconvolution process includes chromatogram extractions (drawn by regular line on the left panel), model peak constructions (drawn by bold line on the left panel), and mass spectrum reconstructions

(right panel). (**d**) The MSP format is utilized for matching experimental mass spectra against mass spectral libraries such as MassBank or LipidBlast. The compound identification is performed by the weighted similarity score of retention time, accurate mass, isotope ratio, and MS/MS spectra.

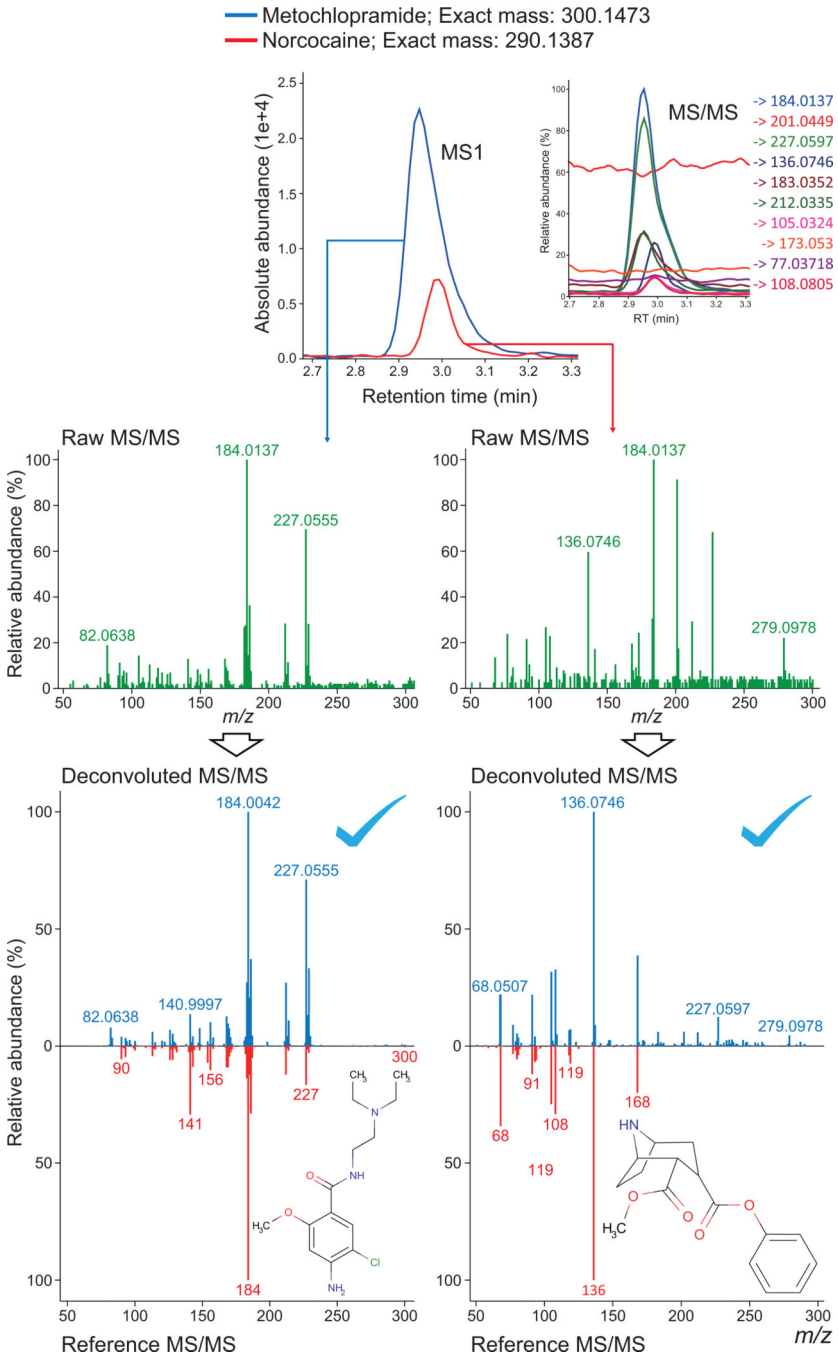

Figure 2. A deconvolution example with respect to SWATH acquisition with HILIC positive ion mode

Two pharmaceutical agents, metoclopramide and norcocaine, were detected in untargeted metabolomics screens and co-eluted within 1.8-s peak top difference. The MS/MS ion traces with respect to these two metabolites are also shown in the top-right panel of precursor ion traces. The middle panels show raw MS/MS spectra of metoclopramide (left) and norcocaine (right), respectively. The spectrum of metoclopramide dominates and masks that of norcocaine, making its detection highly difficult. The bottom panels show the

deconvoluted MS/MS spectrum and spectra matching results of metoclopramide (left) and norcocaine (right) yielding dot-product scores of 0.80 and 0.86, respectively.

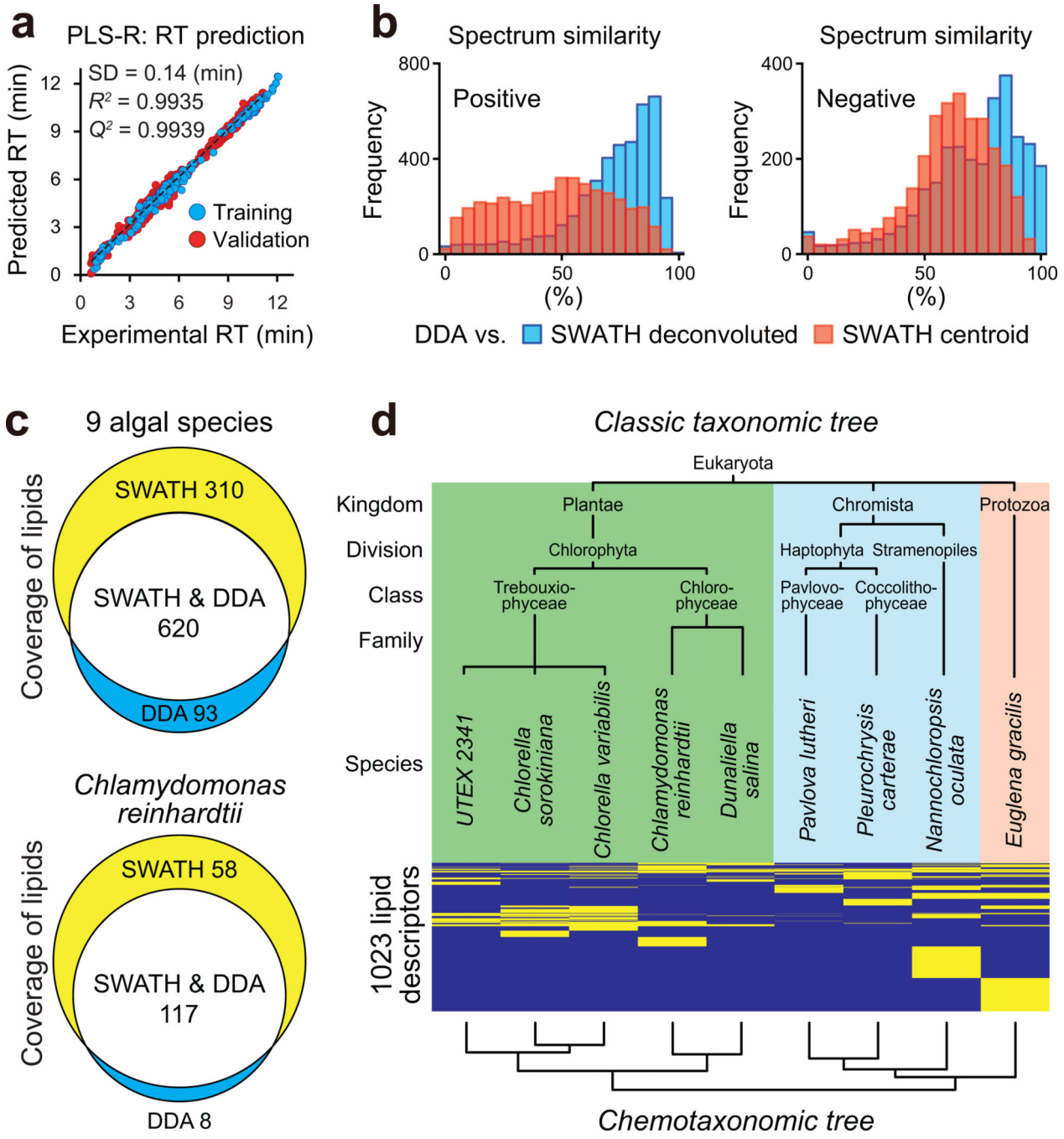

Figure 3. System validation for lipid profiling, lipid coverage and chemotaxonomic relationship of nine algal species

(**a**) The experimental and predicted retention times of 254 (training) and 1,808 (validation) lipids were plotted along *X*-axis and *Y*-axis, respectively. Prediction was performed by using PLS-R on 464 properties from the PaDEL-descriptor suite. The R square, Q square, and standard deviation of the validation set were 0.9935, 0.9939, and 0.14 min, respectively. (**b**) Comparison of mass spectra in positive (left) and negative (right) ion modes in commonly identified lipids between SWATH (data-independent) and the traditional data-dependent

(DDA) methods. The blue histogram shows the spectra similarity between the deconvoluted- and DDA spectra. The red histogram shows the similarity between the centroid (non-deconvoluted)- and DDA spectra. (**c**) Venn diagram of lipid coverages between SWATH and DDA methods. The top panel is the result of nine algal species at 10 ms (SWATH) and 50 ms (DDA) accumulation times in both ionization modes for product ion scanning. The bottom panel is the result of *Chlamydomonas reinhardtii* that the accumulation time of SWATH was 10 ms and 30 ms in positive- and negative ion modes, respectively. (**d**) Hierarchical clustering analysis for nine algal species and 1,023 binary variables. The top and bottom trees are from the classical taxonomies and chemotaxonomies, respectively. The yellow and blue colors between these trees show ‘included’ and ‘not included’ in each algae. UTEX 2341 is currently annotated as *Chlorella minutissima*. *Chlamydomonas reinhardtii* and *Dunaliella salina* are distinguished in Family levels as Chlamydomonadaceae and Dunaliellaceae, respectively.