

MetDIA: Targeted Metabolite Extraction of Multiplexed MS/MS Spectra Generated by Data-Independent Acquisition

Hao Li, Yuping Cai, Yuan Guo, Fangfang Chen, and Zheng-Jiang Zhu

Anal. Chem., **Just Accepted Manuscript** • Publication Date (Web): 27 Jul 2016

Downloaded from <http://pubs.acs.org> on July 27, 2016

Just Accepted

"Just Accepted" manuscripts have been peer-reviewed and accepted for publication. They are posted online prior to technical editing, formatting for publication and author proofing. The American Chemical Society provides "Just Accepted" as a free service to the research community to expedite the dissemination of scientific material as soon as possible after acceptance. "Just Accepted" manuscripts appear in full in PDF format accompanied by an HTML abstract. "Just Accepted" manuscripts have been fully peer reviewed, but should not be considered the official version of record. They are accessible to all readers and citable by the Digital Object Identifier (DOI®). "Just Accepted" is an optional service offered to authors. Therefore, the "Just Accepted" Web site may not include all articles that will be published in the journal. After a manuscript is technically edited and formatted, it will be removed from the "Just Accepted" Web site and published as an ASAP article. Note that technical editing may introduce minor changes to the manuscript text and/or graphics which could affect content, and all legal disclaimers and ethical guidelines that apply to the journal pertain. ACS cannot be held responsible for errors or consequences arising from the use of information contained in these "Just Accepted" manuscripts.



MetDIA: Targeted Metabolite Extraction of Multiplexed MS/MS Spectra Generated by Data-Independent Acquisition

Hao Li, Yuping Cai, Yuan Guo, Fangfang Chen, and Zheng-Jiang Zhu*

Interdisciplinary Research Center on Biology and Chemistry, and Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai, 200032 P.R. China

Abstract

With recent advances in mass spectrometry, there is an increased interest in data-independent acquisition (DIA) techniques for metabolomics. With DIA technique, all metabolite ions are sequentially selected and isolated using a wide window to generate multiplexed MS/MS spectra. Therefore, DIA strategy enables a continuous and unbiased acquisition of all metabolites and increases the data dimensionality, but presents a challenge to data analysis due to the loss of the direct link between precursor ion and fragment ions. However, very few DIA data processing methods are developed for metabolomics application. Here, we developed a new DIA data analysis approach, namely MetDIA, for targeted extraction of metabolites from multiplexed MS/MS spectra generated using DIA technique. MetDIA approach considers each metabolite in the spectral library as an analysis target. Ion chromatographs for each metabolite (both precursor ion and fragment ions) and MS² spectra are readily detected, extracted, and scored for metabolite identification, referred as metabolite-centric identification. A minimum metabolite-centric identification score responsible for 1% false positive rate of identification is determined as 0.8 using fully ¹³C labeled biological extracts. Finally, the comparisons of our MetDIA method with data-dependent acquisition (DDA) method demonstrated that MetDIA could significantly detect more metabolites in biological samples, and is more accurate and sensitive for metabolite identifications. The MetDIA program and the metabolite spectral library is freely available on the internet.

Keywords

Data independent acquisition; metabolomics; targeted extraction; multiplexed spectra; metabolite-centric identification

Introduction

Liquid chromatography tandem mass spectrometry (LC-MS/MS)-based metabolomics is a powerful technology that enables high-throughput metabolic profiling of biological samples.¹⁻⁴ Data-dependent acquisition (DDA) approach is the most common strategy for the metabolite identification,⁵ that metabolite ions in a full scan spectrum (MS^1) are sequentially selected and isolated to generate MS/MS spectra (MS^2). Then metabolite structure is elucidated through MS/MS spectral similarity matching to the standard metabolite spectral library (such as METLIN,⁶ MassBank,⁷ and HMDB⁸), commonly referred as a spectrum-centric approach. However, DDA technology suffers from several limitations.^{9,10} For example, not all precursor ions are able to be readily isolated and fragmented in one experiment, and random selection of precursor ions cannot ensure reliable quality of MS/MS spectra.⁹ The selected precursor ions may also be derived from adducted ions and/or in-source fragmentation (such as $[M+Na]^+$ and $[M-H_2O+H]^+$) instead of molecular ions, thereby increasing the identification of false positives.¹¹

Recently, an alternative workflow to DDA, namely, data-independent acquisition (DIA) approach has been developed (such as MS^E ,¹² PAcIFIC,¹³ and SWATH¹⁴) for proteomics¹⁵⁻¹⁷ and metabolomics.¹⁸ Unlike DDA approach that isolates the precursor ions one by one at selected time points, DIA approach cycles through the whole mass range in segments of predetermined isolation windows, at each segment producing one multiplexed MS^2 spectrum derived from multiple precursor ions.^{13,18} Therefore, DIA strategy enables a continuous and unbiased acquisition of both MS^1 and MS^2 ions in time and ion intensity,¹⁹ and increases the dimensionality of data relative to DDA approach, in which MS^2 spectra are recorded only at selected time points. In addition, the large isolation window in DIA increases transmission efficiency and abundance of the fragment ions,¹⁴ thereby improving the MS^2 spectrum sensitivity. However, compared to DDA approach, the loss of the direct link between a precursor ion and their fragment ions in multiplexed MS/MS spectra makes the subsequent data analysis nontrivial. In proteomics, several data analysis methods and programs, such as OpenSWATH,²⁰ DIA-Umpire,¹⁷ and Skyline²¹ have been developed to address this challenge. In metabolomics, recently, MS-DIAL partially addresses this problem by mathematical deconvolution, and achieves the metabolite identification through a spectrum-centric library matching.¹⁸

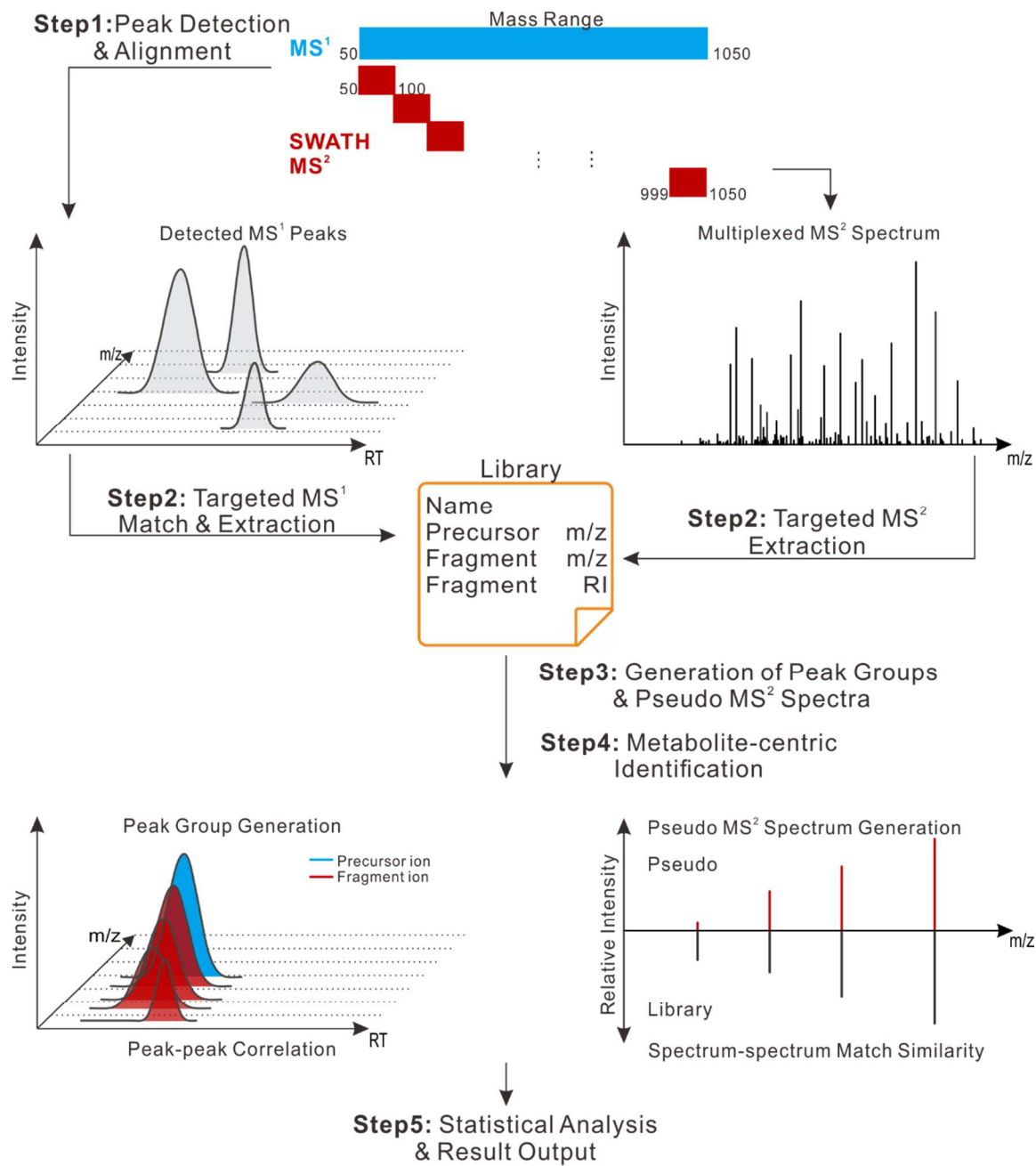



Figure 1. The schematic illustration of MetDIA workflow for the targeted extraction of metabolites from DIA dataset.

Here, we developed an alternative, fundamentally different DIA data analysis approach, namely MetDIA, for targeted extraction of metabolites from multiplexed MS² spectra generated by DIA. Compared with spectrum-centric DDA approach, MetDIA approach considers each metabolite in the metabolite spectral library as an analysis target. For each metabolite in the library, chemical information such as extracted ion chromatograms (EIC) of precursor ion and fragment ions, mass accuracy of precursor ion and fragment ions, MS² spectra, is readily detected, extracted, scored and statistically assessed for the purpose of metabolite identification, referred as metabolite-centric approach. As shown in Figure 1, the MetDIA workflow can be summarized into five steps: (1) peak detection and alignment; (2) targeted chromatogram extractions; (3) generations of peak-groups and pseudo MS² spectra; (4) metabolite-centric identification; (5) statistical analysis and result output. In this study, we first showcased our MetDIA method to process metabolite standard mixture samples consisting of 30 metabolites. Then, a minimum metabolite-centric identification score responsible for 1% false positive rate (FPR) of identification is determined as 0.8 using fully ¹³C labeled metabolite extracts from *Escherichia coli* (*E. coli*). We further compared our MetDIA method with DDA method and the previous reported MS-DIAL method,¹⁸ and results show that MetDIA achieves higher sensitivity and specificity on metabolite identification and wider metabolite coverage in a variety of different biological samples. The MetDIA program and the metabolite spectral library is freely available on the internet.

EXPERIMENTAL SECTION

The extraction of *E. coli* samples, human serum samples and Jurkat cell samples follows the protocol in our previous publication.²² Other experimental details about chemicals, metabolite extraction, LC-MS parameters and data acquisition are provided in the Supporting Information.

MetDIA data processing. MetDIA is developed in R programming environment and provided in our website (<http://www.metabolomics-shanghai.org/software.php>). The raw data were first converted to mzXML files using the “msconvert” program from ProteoWizard (version 3.0.6526), then loaded into R and processed by MetDIA. For the processing of multiple files, data files for different sample groups should be placed in different subfolders separately. The MetDIA program has five steps to process SWATH data and the detailed instructions were provided in Supporting Information.

First, data files in mzXML format were imported into R, and peak detection was operated on MS¹ data. htWave algorithm developed by Tautenhahn *et al.*²³ and included in XCMS, is used for peak detection. In peak detection, parameter “peakwidth” (see Supporting Information) is set as (5, 30) in unit of seconds, referring to the minimum and maximum peak widths for feature detection. Parameter “sn”, referring to signal to noise ratio, is set as 6. For the processing of multiple files, ordered bijective

interpolated warping (OBI-Warp) algorithm, developed by Prince *et al.*²⁴ and included in XCMS, is used for retention time correction and alignment.

Secondly, an in-house metabolite spectral library (see Supporting Information) is used for targeted chromatogram extractions. Targeted extractions is separated as two steps: (1) targeted MS¹ match and extraction; (2) targeted MS² extraction using MS² ions in the spectral library. First, accurate mass match is performed to match all metabolites in the library to the detected peaks with a mass tolerance of ± 12.5 ppm, with a low limit of 0.005 Da (for ions with $m/z < 400$ Da).²⁵ For matched metabolites, ion chromatograms for both MS¹ and MS² ions are extracted using the original retention time, and ion chromatograms between the starting and ending times are recorded and used for peak shape similarity calculation. For MS² ions, ion chromatograms are only extracted in the corresponding SWATH window. The extraction mass tolerance for both MS¹ and MS² ions are defined as ± 12.5 ppm and ± 17.5 ppm, respectively, with low limits of 0.005 Da or 0.007 Da for ions with $m/z < 400$ Da in MS¹ or MS² scans, respectively.²⁵

Thirdly, the extracted ion chromatograms of precursor ions and their product ions are grouped to generate peak-groups. Each peak-group must contain at least one fragment ion with the intensities above 0. The corresponding pseudo MS² spectrum is extracted from multiplexed MS² spectrum at the apex of the peak-group.

Fourthly, metabolite-centric identification is performed by calculating two orthogonal scores: peak-peak correlation (PPC) and spectrum-spectrum match (SSM). PPC score is used to characterize the similarity of ion chromatograms between the precursor ion and its fragment ions, and calculated using Pearson correlation coefficient (equation 1). Therefore, each MS² ion has a calculated correlation value, and highest correlation values from half of MS² ions are averaged as PPC score for metabolite identification.

$$\text{PPC score} = \frac{\sum (I_{Pi} - \bar{I}_P)(I_{Fi} - \bar{I}_F)}{\sqrt{\sum (I_{Pi} - \bar{I}_P)^2} \sqrt{\sum (I_{Fi} - \bar{I}_F)^2}} \quad (1)$$

Where, P = Precursor ion, F = Fragment ion; I_{Pi} refers to the peak intensity of precursor ion in MS¹ scan i , ($i = 1, 2, \dots, n$); \bar{I}_P refers to the averaged peak intensity of precursor ion. I_{Fi} refers to the peak intensity of fragment ion in SWATH-MS² scan i , ($i = 1, 2, \dots, n$); \bar{I}_F refers to the averaged peak intensity of fragment.

SSM score is used to characterize the similarity between experimental pseudo spectra and standard metabolite spectra in the library, and is calculated with dot product function.²⁶

$$\text{SSM} = \frac{(\sum W_L W_Q)^2}{\sum W_L^2 \sum W_Q^2} \quad (2)$$

Where, Weighted intensity, $W = [\text{Peak Intensity}]^n [\text{Mass}]^m$, $n=1$, $m=0$; L = Library; Q = Query.

Then, metabolite-centric identification (MCI) score, is calculated by averaging PPC and SSM as an indicative score:

$$\text{MCI} = (\text{SSM} + \text{PPC}) / 2 \quad (3)$$

Finally, different statistical methods are used for statistical analysis. Specifically, Welch's *t*-test and ANOVA test are used for two and multiple groups, respectively. Fold changes and *p* values are calculated and output in the final report. For single file processing mode, statistical analysis is skipped (Scheme S2 in the Supporting Information).

Results and Discussion.

MetDIA workflow. Metabolomics data were acquired using SWATH technique and processed using MetDIA for targeted metabolite extraction (see Experimental Section). We first showcased our MetDIA approach using a set of metabolite standard mixture that contains 30 metabolites (namely 30STD_mix, Table S1 in Supporting Information) and three biological samples.

(i) Peak detection and alignment. Peak detection was first operated on MS¹ data. For multiple data files, other additional functions required for untargeted metabolomics data analysis such as peak alignment and grouping across multiple samples are also performed (Scheme S1 in the Supporting Information). Most of these common data processing functions are modified and implemented using existing functions in popular metabolomics software XCMS.^{27,28} For the 30STD_mix samples, a total of 1,610 peaks were detected, which were then used for precursor ion match and quantitative analyses.

(ii) Targeted chromatogram extractions. An in-house metabolite spectral library is used for targeted chromatogram extractions. Now the library contains 786 metabolites in total (765 in positive mode and 757 in negative mode), and each metabolite is associated with experimental MS² spectra acquired using DDA approach with collision energy at 35 ± 15 eV (see Supporting Information). This step is divided into two parts: (1) targeted MS¹ match and extraction; (2) targeted MS² extraction using MS² ions in the spectral library. First, accurate mass match was performed to match all of the 765 metabolites in positive mode to the detected 1,610 peaks. The result shows that 80 out of 765 metabolites has at least one match to the detected peaks (109 peaks in total). Then, for each matched metabolite peak, MetDIA automatically extracts ion chromatograms of MS¹ ions and corresponding MS² ions from MS¹ and SWATH-MS² scans, producing integrated, continuous MS¹ and MS² ion count plots versus retention time (Figure 1). The extraction *m/z* window-width can be specifically defined according to the instrument-specific MS² resolution. With the restriction of precursor ions, MetDIA could effectively filter a large number of false positive features. Here, in MetDIA program, only precursor ion matched

metabolites in the library are further proceed for targeted chromatogram extractions.

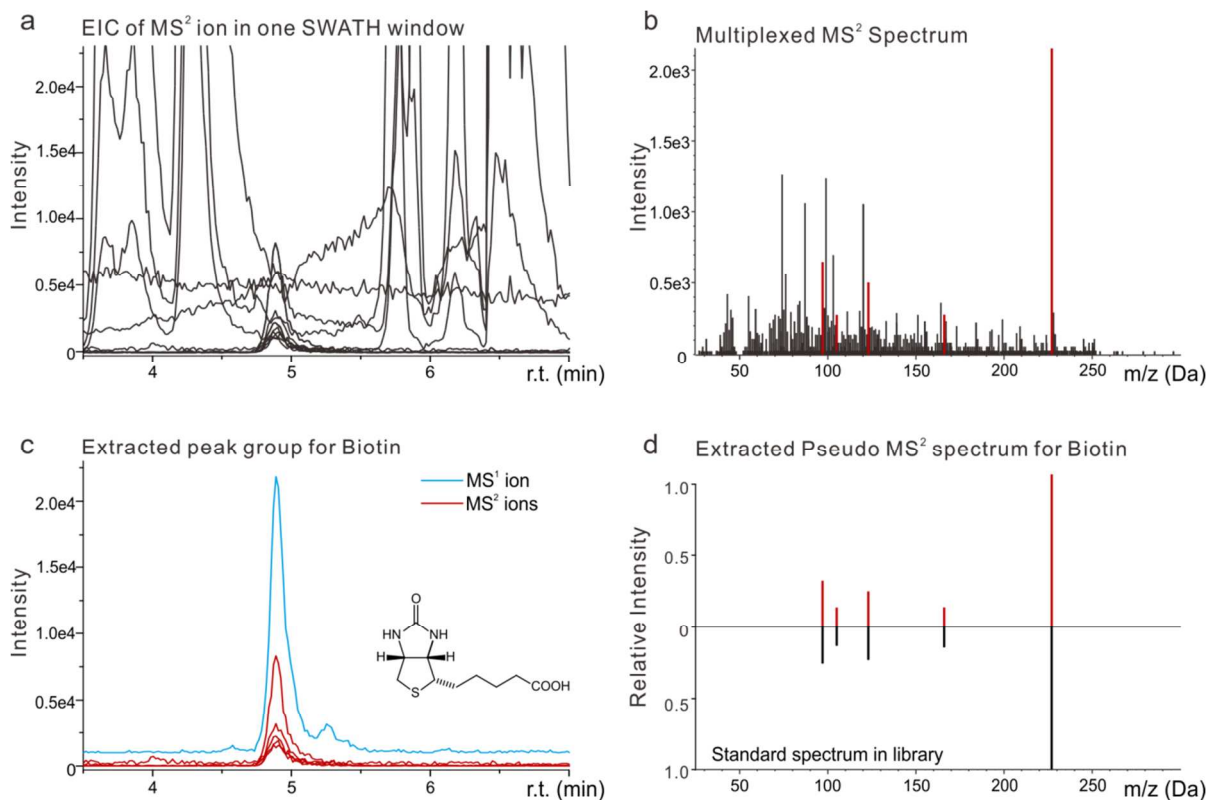


Figure 2. Generation of peak-groups and pseudo MS² spectra for metabolite-centric identification. Taken metabolite Biotin (accurate mass 244.0882) as an example: (a) EIC of all MS² ions in the corresponding SWATH window (199-250 Da); (b) multiplexed MS² spectrum acquired in the SWATH window 199-250 Da; (c) extracted peak-group of metabolite Biotin using its precursor ion and five fragment ions; the peak-peak correlation score for the extracted peak-group is 0.963; (d) extracted pseudo MS² spectrum using five fragment ions of Biotin from the multiplexed MS² spectrum, and the spectrum-spectrum match similarity score is 0.998.

(iii) Generation of peak-groups and pseudo MS² spectra. After extraction of both MS¹ and MS² ions using the spectral library, precursor ion and fragment ion peaks are aligned to generate peak-group (Figure 1). Then the corresponding pseudo MS² spectrum is extracted from multiplexed MS² spectrum at the apex of the peak-group. Taken metabolite Biotin as an example (Figure 2), the MS¹ ion of Biotin (m/z 245.0955 Da, $[M+H]^+$) was first detected and extracted in the MS¹ scan. In the corresponding SWATH acquisition window (199-250 Da), five fragment ions (MS²) including m/z 97.0394 Da, 105.0689 Da, 123.0247 Da, 166.0683 Da, and 227.0843 Da were extracted and formed a metabolite peak-group with their precursor ion (Figure 2c). Although only a small portion of precursor ions in the full scan were

isolated and fragmented in one SWATH acquisition window, the generated MS^2 spectra contain hundreds of fragment ions (Figure 2b), and could not be directly used for spectral match and metabolite identification. In this work, MetDIA performs target extractions of fragment ions (maximum of 5 ions) from the multiplexed MS^2 spectrum, and effectively reduces the complexity of the MS^2 spectra, referred as pseudo MS^2 spectrum. The extracted pseudo MS^2 spectrum can directly be used for spectral match (Figure 2d). Using the similar way, only 97 peak-groups out of 109 peaks were generated in 30STD_mix sample (Figure 3a), the rest 12 peaks have no available MS^2 ions in the multiplexed spectra for extraction, therefore, no peak-groups or pseudo MS^2 spectrum were generated.

(iv) Metabolite-centric identification. Metabolite-centric identification is achieved through scoring peak-groups and pseudo MS^2 spectra with two orthogonal scores. First, the spectral similarity between pseudo MS^2 spectrum and standard spectrum in library is calculated using dot product function, and defined as spectrum-spectrum match (SSM) score (see Experimental Section). SSM is standardized ranging from 0 to 1, which refers to no similarity and an exact match, respectively. A SSM score larger than 0.8 is required for a true metabolite identification. Then, peak-peak correlation (PPC) score is calculated to characterize the similarity of elution profiles between the precursor ion and fragment ions, and is a quantitative measurement of true existence for a given metabolite in the library. In traditional untargeted metabolomics using DDA technique, as a result of random selection of precursor ions, the loss of peak shape information of product ions leads to the use of SSM, rather than PPC for metabolite identification. Here, in SWATH based untargeted metabolomics, PPC could be used as an additional diagnostic criterion for metabolite identification. Pearson correlation coefficient is used for the PPC score calculation. PPC is standardized ranging from -1 to 1, referring to negative and positive correlations. Similarly, a positive PPC score larger than 0.8 is required for a true existence for a given metabolite.

In one 30STD_mix sample, a total of 34 metabolites were identified with both PPC and SSM scores larger than 0.8 (Figure 3a and 3b). All of the spiked 30 metabolites are identified, making a 100% true positive rate. In metabolomics, according to Metabolomics Standard Initiative, orthogonal criteria could increase identification accuracy. MS^1 information is complementary to MS^2 information for metabolite identification. As shown in Figure 3a, the restriction of precursor ion could effectively reduce false positive identification in metabolomics. In addition, an example of true positive identification is given in Figure 2c and 2d, the calculated PPC and SSM scores for metabolite Biotin are as high as 0.963 and 0.998, respectively. Other metabolite examples for true negative identifications are given in Figure S1 in the Supporting Information. For example, metabolite L-lysine has a high SSM score (0.995) by chance, but a low PPC score (0.575). Differently, metabolite indole-3-carboxylic acid has a high PPC score (0.962), but a low SSM score (0.337); metabolite dimethylglycine has both low scores in PPC (0.653)

and SSM (0.407). The results demonstrate the importance of PPC and SSM scores for metabolite identifications.

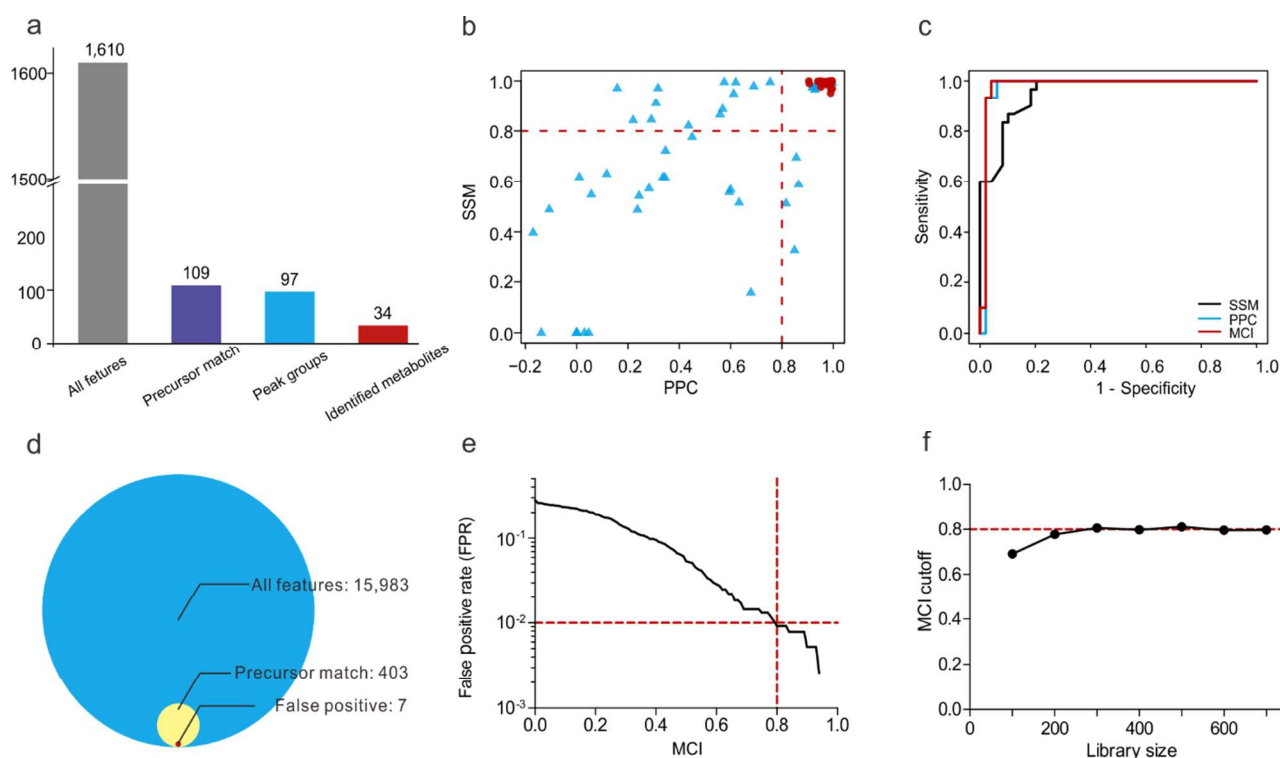


Figure 3. (a) The stepwise processing of the standard mixture sample that contains 30 metabolites (namely 30STD_mix). (b) The distributions of peak-peak correlation scores and spectrum-spectrum match scores for the 30STD_mix sample. The red circles represent for the true identifications and the blue triangles represent the false identifications. (c) Receiver operating characteristic analysis using SSM, PPC, and MCI scores. Areas under curve for SSM, PPC, and MCI scores are 0.955, 0.976, and 0.981, respectively. (d) Number of false positive identifications of metabolites in fully ¹³C labeled metabolite extracts from bacteria. (e) The determination of minimum MCI score (0.8) for 1% false positive rate (FPR) using fully ¹³C labeled metabolite extracts from bacteria. (f) The relationship of minimum MCI scores and the library sizes with 1% FPR.

Finally, we evaluated the mathematical combinations of PPC and SSM scores toward metabolite-centric identification (MCI) score. The PPC and SSM score distributions of metabolite extracted in 30STD_mix samples (Figure 3b) and fully ¹³C labeled metabolite extracts from *E.coli* bacteria (Figure S2a in Supporting Information) are both used. In fully ¹³C labeled metabolite extracts, theoretically, no metabolite from the metabolite library should be identified. Both results demonstrated that PPC performs better classification than SSM on metabolite identification. In both cases, with the similar cutoff scores, SSM gives more false identifications than PPC. In addition, receiver operating

characteristic analysis (ROC, Figure 3c) provides the similar result that PPC has higher predictive power than SSM. We further tested different combinations of PPC and SSM (Figure S2b). The result shows that the rate of false positive identification is relatively low when the percentage of PPC is more than 0.3. Therefore, in our work, we choose the percentage of PPC score is 0.5 for combination, namely metabolite-centric identification (MCI) score. MCI score provides even better identification (AUC, 0.981) than PPC and SSM alone (Figure 3c), with a high sensitivity (100%) and specificity (99.5%) at the score of 0.8.

Unlike proteomics, the determination of false discovery rate (FDR) is not feasible for metabolite identification in metabolomics as no appropriate decoy metabolite databases can be constructed.²⁹ Instead, here, we evaluated false positive rate (FPR, $FP/(FP + TN)$) with fully ^{13}C labeled metabolite extracts from *E.coli* bacteria (Figure 3d and Figure S2a in Supporting information). As seen in Figure 3d, MetDIA program identified 7 metabolites using our metabolite library (765 metabolites for positive mode in total) with MCI scores larger than 0.8, corresponding to FPR as low as 1% ($7/(7 + 758)$). Finally, we plotted the relationship of FPR values versus MCI scores in Figure 3e, and determined the minimum MCI score as 0.8 for 1% FPR in metabolite identification. To demonstrate the relationship of minimum MCI scores required for 1% FPR and the library sizes (i.e., number of metabolites in the library), we randomly selected different numbers of metabolites ranging from 100 to 700 metabolites for MetDIA analysis, and determined minimum MCI scores corresponding to 1% FPR. The data analyses were randomly repeated for ten times, and results prove that the number of metabolites in library has little effect on the minimum MCI score for 1% FPR when library size exceeding 300 metabolites (Figure 3f). Therefore, the minimum MCI score required for 1% FPR is set as 0.8.

Current SWATH data processing programs are mostly developed and optimized for proteomics. Tools such as OpenSWATH and AB Sciex SWATHTM (a software plugin in Peakview to process SWATH data) are excellent for proteomics data processing. However, due to the lack of valid decoy databases in metabolomics, we found these tools are not effective for metabolite identification (Figure S3 and S4 in the Supporting Information).

(v) Statistical analysis and result output. After metabolite identification, statistical analysis is used to compare the significance changes of quantities for metabolites among two or more sample groups. The example output result contains fold change, *p* value and identification information (see Supporting Information).

Comparison of DIA and DDA methods. In order to validate our MetDIA approach, we further acquired two concentrations of 30STD_mix samples (330 ppb and 33 ppb) and three different biological samples

(human serum, *E. coli* bacteria, and rat liver tissue) using both DIA (SWATH) and DDA methods, and compared the sensitivity and selectivity of both methods on metabolite identification. The summary of the results is shown in Figure 4a and 4b. Although both DDA and DIA (using MetDIA) could successfully identify 30 true positive identifications of metabolites in high concentration 30STD_mix sample (330 ppb), there are less false positives using MetDIA method (5 false positives) than DDA method (9 false positives). When the concentration of 30STD_mix sample decreased to 33 ppb, the DDA method could only identify 19 metabolites, fewer than 24 metabolites using MetDIA. These results suggest that MetDIA, providing more true positive and less false positive identifications of metabolites, is more accurate and sensitive than DDA method. Similar results are obtained with biological samples, MetDIA identified 152, 138 and 192 metabolites in human serum, *E. coli* bacteria, and rat liver tissue, respectively, while DDA method only detects 85, 79, 126 metabolites, respectively. Further analysis proves that MetDIA could successfully identify more low abundant metabolites (Figure S5 in Supporting Information).

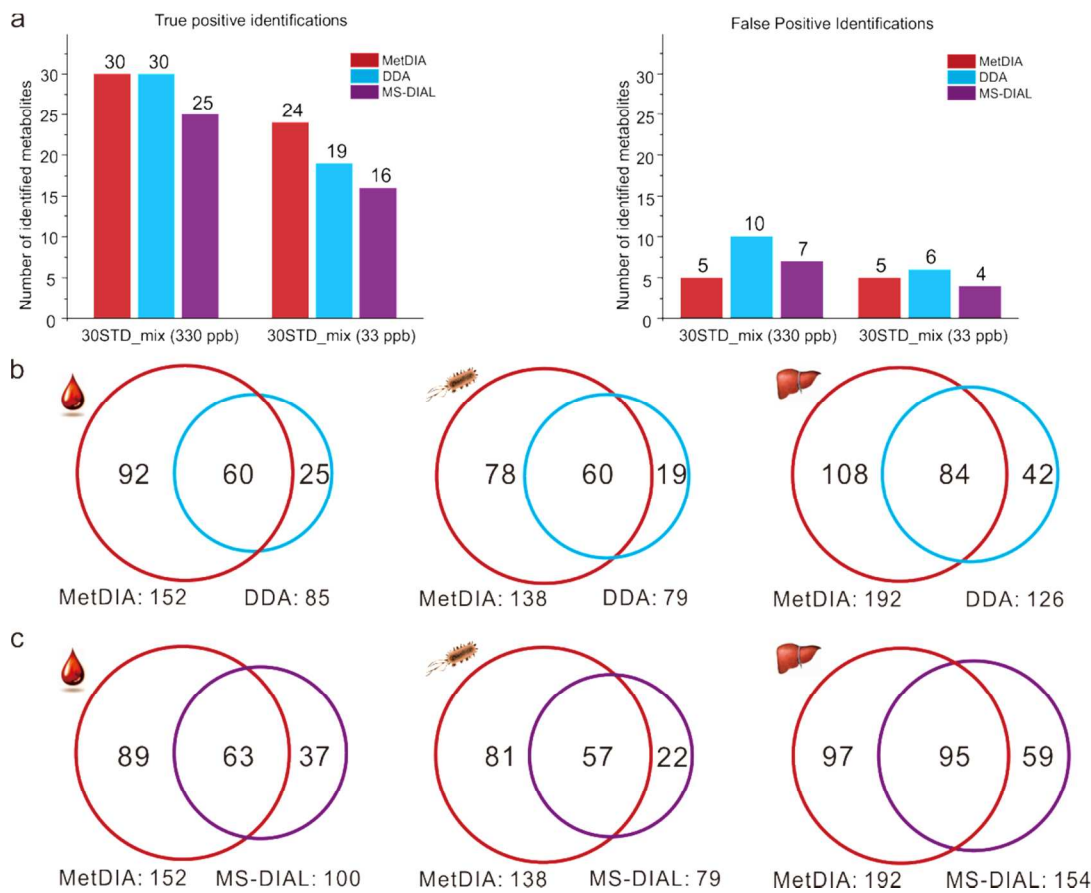


Figure 4. (a) Comparison of identified and misidentified metabolites using MetDIA, DDA and MS-DIAL methods in 30STD_mix samples (3 replicates in each group). (b) Comparison of identified metabolite numbers in biological samples using MetDIA and DDA methods. (c) Comparison of identified metabolite

numbers in biological samples using MetDIA and MS-DIAL methods (ten replicates in each group).

Comparison of MetDIA and MS-DIAL. We further compared our MetDIA method with the recent published MS-DIAL method¹⁸ regarding to the performance of metabolite identification. MS-DIAL program is an excellent and the first software tool for SWATH based untargeted metabolomics, which implemented spectral deconvolution method from GC-MS to purify MS² spectra for spectrum-centric match and metabolite identification. It aims to identify as many metabolites as possible. The same datasets of 30STD_mix samples (330 ppb and 33 ppb) and biological samples were processed using both MetDIA and MS-DIAL, and comparative results are shown in Figure 4a and 4c. In 30STD_mix samples, MS-DIAL fails to identify all of the 30 true positive metabolites at either high or low concentrations. Similar results were obtained for biological samples that MS-DIAL detected 100, 79 and 154 metabolites in human serum, E. coli bacteria, and rat liver tissue, respectively, 20-75% fewer than our MetDIA method. The parameters and their optimizations for MS-DIAL data processing were listed in Supporting Information. Again, we have to emphasize that although MetDIA is more sensitive on metabolite identification, MS-DIAL adopts more complicated and fully untargeted approach to identify as many metabolites as possible. Here, we developed MetDIA as an alternative, targeted extraction based approach to process DIA based metabolomics data, which is complementary to MS-DIAL program. We share the common desire with the authors of MS-DIAL to promote the DIA application in metabolomics.

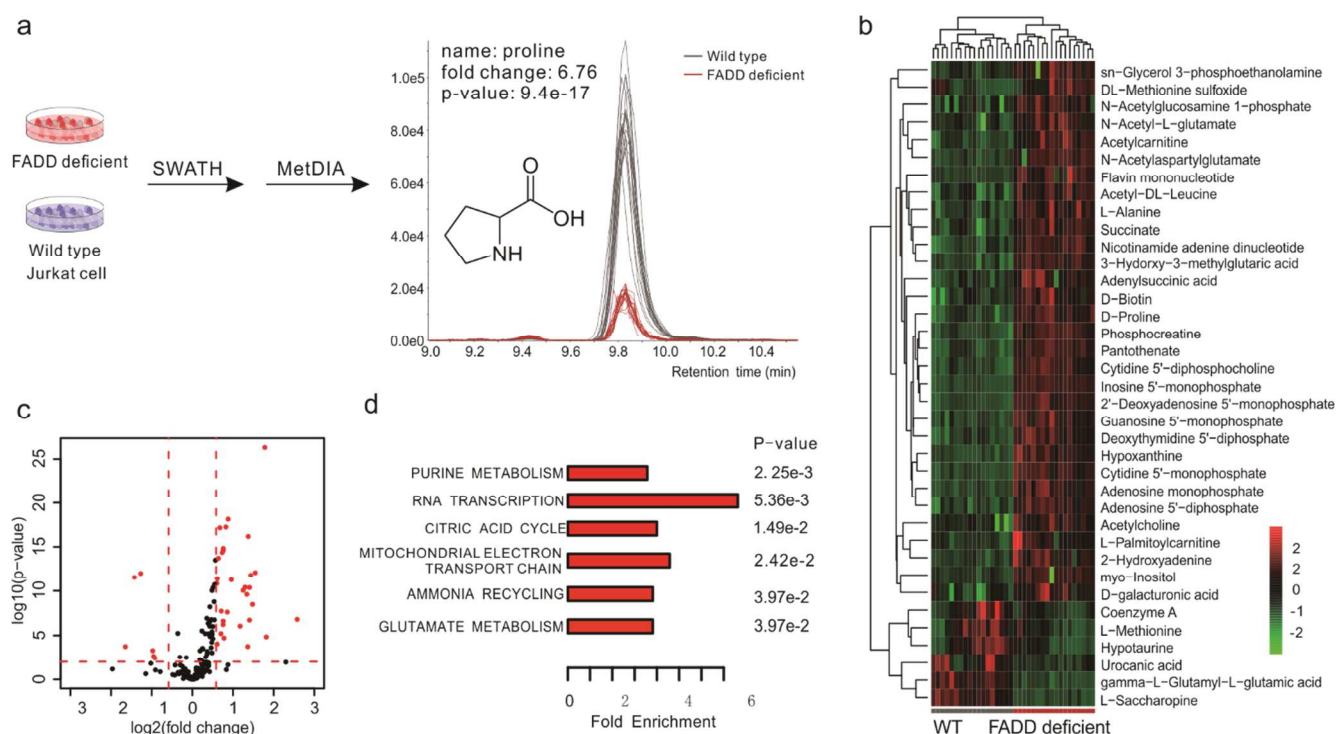


Figure 5. (a) Demonstration of data acquisition, processing, and statistical analysis workflow for

metabolomic profiling using DIA method. (b) Heat map of 60 significantly changed metabolites detected by MetDIA. (c) Volcano plot of all 156 identified metabolites from MetDIA. Red points represent significantly changed metabolites and black points represent insignificantly changed ones. (d) Metabolic pathway enrichment analysis of 37 significantly changed metabolites using MetaboAnalyst.³⁰

Metabolomic profiling of biological samples. We demonstrated the utility of our MetDIA method for the functional study of protein FADD (Fas-associated protein with death domain) by metabolomic profiling of wild type and FADD deficient Jurkat cells (Figure 5). FADD is an important protein regulating a lot of biological events, such as apoptosis, necroptosis and lymphocyte proliferation.³¹ Specifically, wild type and FADD deficient cell samples were analyzed using SWATH technique. All acquired SWATH datasets were converted and analyzed by MetDIA program. In total, 156 metabolites were identified in the cell metabolome (Figure 5b and 5c). For each of detected metabolite, the qualitative information such as metabolite name, KEGG ID, and quantitative information such as intensity, fold change, and *p* value calculated with MS¹ peak area was output as a .csv file (see Supporting Information). Among these, 37 metabolites significantly changed ($p < 0.05$, fold change > 1.5). These metabolites were further submitted to pathway enrichment analysis using MetaboAnalyst.³⁰ The analysis demonstrated that six pathways changed significantly, such as purine metabolism and RNA transcription with the deficiency of protein FADD (Figure 5d).

CONCLUSIONS

In conclusion, a new DIA data analysis approach, namely MetDIA, is developed for targeted extraction of metabolites from DIA datasets. Compared with spectrum-centric DDA approach, MetDIA approach considers each metabolite in the metabolite spectral library as an analysis target, achieving the data processing automatically within five steps, including (1) peak detection and alignment; (2) targeted chromatogram extractions; (3) generations of peak-groups and pseudo MS² spectra; (4) metabolite-centric identification; and (5) statistical analysis and result output. The metabolite-centric identification is achieved through scoring peak-group and pseudo MS² spectra with two orthogonal scores, PPC and SSM. A minimum metabolite-centric identification score responsible for 1% false positive rate of identification is determined as 0.8 using fully ¹³C labeled biological extracts. The comparisons of our MetDIA method with DDA and MS-DIAL¹⁸ methods demonstrated that MetDIA could significantly detect more metabolites in biological samples, and is more accurate and sensitive for metabolite identifications. Therefore, we believe the MetDIA method should have a wide application in metabolomics. Finally, the program MetDIA and associated metabolite spectral library is freely available

on the internet.

ASSOCIATED CONTENT

Supporting Information

Additional information as noted in main text is available free of charge via the Internet at <http://pubs.acs.org/>.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jiangzhu@sioc.ac.cn, phone: 86-21-68582296

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The work is financially supported by National Natural Science Foundation of China (Grant No. 21575151). Z.-J. Z. is also supported by Thousand Youth Talents Program (The Recruitment Program of Global Youth Experts from Chinese government).

Reference

- (1) Cajka, T.; Fiehn, O. *Anal. Chem.* **2016**, *88*, 524-545.
- (2) Li, L.; Li, R.; Zhou, J.; Zuniga, A.; Stanislaus, A. E.; Wu, Y.; Huan, T.; Zheng, J.; Shi, Y.; Wishart, D. S.; Lin, G. *Anal. Chem.* **2013**, *85*, 3401-3408.
- (3) Peng, J.; Chen, Y. T.; Chen, C. L.; Li, L. *Anal. Chem.* **2014**, *86*, 6540-6547.
- (4) Tsugawa, H.; Arita, M.; Kanazawa, M.; Ogiwara, A.; Bamba, T.; Fukusaki, E. *Anal. Chem.* **2013**, *85*, 5191-5199.
- (5) Patti, G. J.; Yanes, O.; Siuzdak, G. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 263-269.
- (6) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. *Ther. Drug Monit.* **2005**, *27*, 747-751.
- (7) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom.* **2010**, *45*, 703-714.

- (8) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M. A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; Macinnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L. *Nucleic Acids Res.* **2007**, *35*, D521-526.
- (9) Zhu, X.; Chen, Y.; Subramanian, R. *Anal. Chem.* **2014**, *86*, 1202-1209.
- (10) Roemmelt, A. T.; Steuer, A. E.; Poetzsch, M.; Kraemer, T. *Anal. Chem.* **2014**, *86*, 11742-11749.
- (11) Neumann, S.; Bocker, S. *Anal. Bioanal. Chem.* **2010**, *398*, 2779-2788.
- (12) Silva, J. C.; Gorenstein, M. V.; Li, G. Z.; Vissers, J. P.; Geromanos, S. J. *Mol. Cell. Proteomics* **2005**, *5*, 144-156.
- (13) Panchaud, A.; Scherl, A.; Shaffer, S. A.; Haller, P. D.; Kulasekara, H. D.; Miller, S. I.; Goodlett, D. R. *Anal. Chem.* **2009**, *81*, 6481-6488.
- (14) Gillet, L. C.; Navarro, P.; Tate, S.; Rost, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. *Mol. Cell. Proteomics* **2012**, *11*, O111.016717- O111.016717.
- (15) Wang, J.; Tucholska, M.; Knight, J. D. R.; Lambert, J.-P.; Tate, S.; Larsen, B.; Gingras, A.-C.; Bandeira, N. *Nat. Methods* **2015**.
- (16) Li, Y.; Zhong, C.-Q.; Xu, X.; Cai, S.; Wu, X.; Zhang, Y.; Chen, J.; Shi, J.; Lin, S.; Han, J. *Nat. Methods* **2015**.
- (17) Tsou, C. C.; Avtonomov, D.; Larsen, B.; Tucholska, M.; Choi, H.; Gingras, A. C.; Nesvizhskii, A. I. *Nat. Methods* **2015**, *12*, 258-264.
- (18) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. *Nat. Methods* **2015**, *12*, 523-526.
- (19) Venable, J. D.; Dong, M. Q.; Wohlschlegel, J.; Dillin, A.; Yates, J. R. *Nat. Methods* **2004**, *1*, 39-45.
- (20) Rost, H. L.; Rosenberger, G.; Navarro, P.; Gillet, L.; Miladinovic, S. M.; Schubert, O. T.; Wolski, W.; Collins, B. C.; Malmstrom, J.; Malmstrom, L.; Aebersold, R. *Nat. Biotechnol* **2014**, *32*, 219-223.
- (21) Egertson, J. D.; MacLean, B.; Johnson, R.; Xuan, Y.; MacCoss, M. J. *Nat Protoc* **2015**, *10*, 887-903.
- (22) Cai, Y.; Weng, K.; Guo, Y.; Peng, J.; Zhu, Z.-J. *Metabolomics* **2015**, *11*, 1575-1586.
- (23) Tautenhahn, R.; Bottcher, C.; Neumann, S. *BMC Bioinformatics* **2008**, *9*, 504.
- (24) John, T. P.; Edward, M. M. *Anal. Chem.* **2006**, *78*, 6140-6152.
- (25) Yang, X.; Neta, P.; Stein, S. E. *Anal. Chem.* **2014**, *86*, 6393-6400.
- (26) Stein, S. E.; Scott, D. R. *J. Am. Soc. Mass Spectr.* **1994**, *5*, 859-866.
- (27) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779-787.
- (28) Benton, H. P.; Wong, D. M.; Trauger, S. A.; Siuzdak, G. *Anal. Chem.* **2008**, *80*, 6382-6389.
- (29) Stein, S. *Anal. Chem.* **2012**, *84*, 7274-7282.
- (30) Xia, J.; Psychogios, N.; Young, N.; Wishart, D. S. *Nucleic Acids Res* **2009**, *37*, W652-660.
- (31) Tourneur, L.; Chiocchia, G. *Trends Immunol.* **2010**, *31*, 260-269.

for TOC only

