Reproducible untargeted metabolomics data analysis workflow for exhaustive MS/MS annotation

Miao Yu[a*], Georgia Dolios[a], Lauren Petrick[a,b]

[a] Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, United States

[b] The Institute for Exposomic Research, Icahn School of Medicine at Mount Sinai, NY, 10029, United States

*Corresponding author: Email: miao.yu@mssm.edu Phone: +1-646-707-5791. Fax: +1-646-537-9654.

Abstract

Unknown features in untargeted metabolomics and non-targeted analysis (NTA) are identified using fragment ions from MS/MS spectra to predict the structures of the unknown compounds. The precursor ion selected for fragmentation is commonly performed using data dependent acquisition (DDA) strategies or following statistical analysis using targeted MS/MS approaches. However, the selected precursor ions from DDA only cover a biased subset of the peaks or features found in full scan data. In addition, different statistical analysis can select different precursor ions for MS/MS analysis, which make the *post-hoc* validation of ions selected by new statistical methods impossible for precursor ions selected by the original statistical method. Here we propose an automated, exhaustive, statistical model-free workflow: paired mass distance-dependent analysis (PMDDA), for untargeted mass spectrometry identification of unknown compounds. By removing redundant peaks and performing pseudo-targeted MS/MS analysis on independent peaks, we can comprehensively cover unknown compounds found in full scan analysis using a “one peak for one compound” workflow without a priori redundant peak information. We show that compared to DDA, PMDDA is more comprehensive and robust against samples' matrix effects. Further, more compounds were identified by database annotation using PMDDA compared with CAMERA and RAMClustR. Finally, compounds with signals in both positive and negative modes can be identified by the PMDDA workflow, to further reduce redundancies. The whole workflow is fully reproducible as a docker image xcmsrocker with both the original data and the data processing template.

Introduction

While metabolomics aims at revealing changes in levels of all possible metabolites in biological samples[1], non-targeted analysis (NTA) aims at comprehensive profiling of compounds in environmental samples[2]. To achieve these goals, both approaches use high-resolution mass spectrometry (HRMS) to perform unbiased measurement of small molecules followed by identification of unknowns[3]. In most HRMS-based workflows, small molecule profiles will first be extracted across samples as peaks or features[4]. Tens of thousands of features are typically extracted in each sample making it impractical to target every feature for MS/MS fragmentation[5]. For biological studies comparing subject groups, statistical analysis, machine learning algorithms and/or annotation can be performed to subset the features into peaks of interest[6,7]. Those selected peaks are then targeted for MS/MS fragmentation for identification. However, this approach is limited to a single research question and statistical analysis, as a new question or analysis would reveal different ions as targets for MS/MS analysis[8]. In contrast, group comparisons are not available in ecological study designs or environmental investigations for supervised statistical analysis[9]. In this case, an exhaustive identification strategy of all possible small molecules needs to be developed.

Automated untargeted MS/MS identification techniques such as data-independent acquisition (DIA) and data dependent acquisition (DDA) are powerful tools in qualitative untargeted analysis for identification of unknowns[10]. For DDA, precursor ions for MS/MS are selected during data collection by user-defined strategies. For DIA, all ions are sent into the collision cell for fragmentation, and deconvolution algorithms are used to connect the fragment ions to the parent compounds. However, DDA and DIA cover only a subset of the full scan features and the selected precursor ions may come from background instead of biologically relevant features[11]. In addition, DDA and DIA are designed for qualitative analysis instead of performing quantitative analysis with fragment ions[12], because a compromise must be made between more scan time for high quality fragment ions and well-shaped chromatography for precursor ions. Proposed solutions include time-staggered precursor ion lists as inclusion lists[13] or automated exclusion lists to cover more compounds during repeated DDA injections[14]. However, the sensitivity of DDA's precursor ions is comparable with full scan mass spectra[11] limiting the possibility to find extra precursor ions by DDA.

As an alternative to DDA or DIA, targeted MS/MS is a straightforward method for qualitative and quantitative analysis of known compounds. Since targeted MS/MS analysis requires a pre-defined peak list for both precursor and fragment ions[13], new strategies needed to be developed for implementation in untargeted analysis for unknown compounds. Mainly, since redundant peaks dominate full scan mass spectra, targeted MS/MS peak lists need to be refined by pseudo-spectra annotation, i.e., clustering all mass spectral signals stemming from each metabolite[15]. In practice, the number of unique compounds may be as little as twenty percent of the total feature numbers[16]. If only a single peak is selected as the precursor ion for each unknown compound, the numbers of precursors for targeted MS/MS are drastically reduced.

Such "one feature for one compound" strategy has been reported for several metabolomics studies[17,18], mainly using known adducts, neutral loss, and isotope pattern to detect the redundant peaks. Software packages such as CAMERA[19] and RamClustR[20] have been developed to annotate the pseudo-spectra for unknown full scan mass spectra algorithms that use correlation of peaks and pre-defined paired-mass distances for selecting redundant peaks to generate pseudo-spectra[7]. However, adducts or in-source reactions might be quite different among different sample matrices or instrument parameters[21], even for peaks from the same compound[22]. Therefore, a frequency-based paired-mass distances algorithm, such as the GlobalStd algorithm, could be an alternative solution to determine pseudo-spectra for exhaustive and local MS/MS analysis as it is designed to extract independent peaks without predefined redundant peaks information[3,16].

With such high complexity and no gold standard for metabolomics data pre-processing, reproducibility is important. Though raw metabolomics data can be uploaded and accessed through online databases such as MetaboLights[23] or metabolomics workbench[24], details of data analysis are not as transparent as data sharing, and reduce the ability to fully reproduce the reported findings[25]. Data analysis software with a graphic user interface (GUI) can be easy to use and document, but is also restricted to only defined operations[26]. An open source data process script can represent every step of the data analysis while still being flexible,[27] but researchers need to adopt specific software within an integrated development environment (IDE), which also reduces reproducibility due to the lack of experience with certain software[28]. To address these challenges, a system image with pre-installed open source software and data process templates for untargeted analysis should be developed to attain fully reproducible omics studies.

In this work, we developed an exhaustive and reproducible untargeted metabolomics data analysis workflow called paired-mass distance dependent analysis (PMDDA) to automatically list independent peaks as precursor ions for MS/MS annotation. We then compared PMDDA with DDA and the CAMERA and RamClustR precursor peaks selection algorithms using data acquired on standard reference material (NIST 1950) as demonstration. The utility of PMDDA was further demonstrated by finding the overlap in peaks between positive and negative mode analysis. All of the data and data processing scripts are reproducible by a publicly available docker image.

Methods

Sample preparation

NIST 1950 Frozen Human Plasma standard reference material (SRM), which documented 85 compounds in the sample, was used in this study for reproducibility. Aliquots of 50 µL of NIST SRM plasma were thawed on ice. Proteins were precipitated by the addition of 150 µL of ice-cold methanol containing isotope labelled internal standards, 10 sec of vortexing, and 30 min incubation at -80°C. The samples were then centrifuged at 13,000 g for 10 min at 4°C, and 70

µL of the supernatant was transferred to two 1.5 mL microcentrifuge tubes. The extracts were evaporated using a Savant SpeedVac concentrator at 35°C for 90 min and samples were stored at -80°C until analysis. Following the same protocol, 50 µL aliquots of a matrix blank (replacing the SRM plasma with water), were extracted.

Instrument analysis

Immediately prior to data acquisition, dried samples were reconstituted in 60 µL of methanol. Samples were analyzed using an ultra-high performance liquid chromatography (UHPLC) 1290 Infinity II system (including 0.3 µm inline filter, Agilent Technologies, Santa Clara, USA) with 1260 Infinity II isocratic pump (including 1:100 splitter) coupled to a 6545 quadrupole-time time of flight (Q-TOF) mass spectrometer with a dual AJS electrospray ionization source (Agilent Technologies, Santa Clara, USA). Samples were maintained at 4°C in the multisampler module. Reference masses included positive ionization mode: purine (m/z 121.0509), HP-0921 (m/z 922.0098); and negative ionization mode: purine (m/z 119.0363), HP-0921 (m/z 966.0007). Sheath and drying gas (Nitrogen purity >99.999%) flows were 12 L/min and 10 L/min, respectively. Drying and sheath gas was 250 °C, with the nebulizer pressure at 20 psig, and voltages for positive and negative ionization modes at +3000 V and -3000 V, respectively.

The extracts were injected onto a Zorbax Eclipse Plus C18, RRHD column (50 mm × 2.1 mm, 1.8 µm particle size, Agilent Technologies, Santa Clara, USA) coupled to a guard column (5 mm × 2 mm, 1.8 µm Agilent Technologies, Santa Clara, USA) maintained at 50°C. Separation occurred using Mobile phase A consisted of water with 0.1% formic acid and Mobile phase B consisted of 2-propanol:ACN (90:10, v/v) with 0.1% formic acid at a flow rate of 0.4 mL/min. A 15 min gradient was used (5% B for 2 min, increasing to 30 % B in 2 min, and increasing from 30 % to 98 % B in 9.5 min with a 1.5 min hold), followed by a column re-equilibration phase. Data was acquired with a mass range of 100-1000 m/z (MS1) and 20-1000 m/z (MS/MS).

Five SRM samples and five matrix blanks were analyzed. Data were collected in full scan positive and negative mode. Then, the precursor ions were selected for MS/MS fragmentation based on full scan data either via PMDDA, CAMERA, or RAMClustR. Peak lists for repeated injections of MS/MS analysis were automatically generated by an in-house script. Then, three DDA MS/MS data acquisitions were collected on both SRM samples and matrix samples. The collision energy was set at 20 eV for all MS/MS fragmentation.

Data analysis

Data analysis was performed in R (version 4.0.2)[29] according to the workflow described in Figure 1. Raw data were refined by retention time range between 30s and 930s for the positive and negative mode to remove both the void volume and the washing phase of the column. The peak picking parameters for xcms[30] were optimized by IPO[31] for the five SRM samples. After retention time correction and peak filling for the low abundance peaks, the features were further filtered by those with intensity fold change larger than three times that in the SRM than the matrix samples.

Peaks with relative standard deviation (RSD) larger than 30% in SRM samples were removed. The filtered peaks were processed by PMDDA, CAMERA, and RAMClustR to select the precursor ions for fragmentation. Repeated injections were designed to retain high sensitivity for exhaustive identification by MS/MS across the column gradient. The MS/MS data were then converted to open source format[32] and annotated using GNPS[33] for MS/MS annotation with default settings.

The whole PMDDA workflow (Fig. 1), including MS1 feature extraction and filtering, precursor ion selection, and injection peak table generation for MS/MS analysis has been included in the rmwf package's data processing template with links to download the original data via figshare[34]. In addition, the workflow and corresponding software were packaged into a docker image called xcmsrocker (https://hub.docker.com/repository/docker/yufree/xcmsrocker).

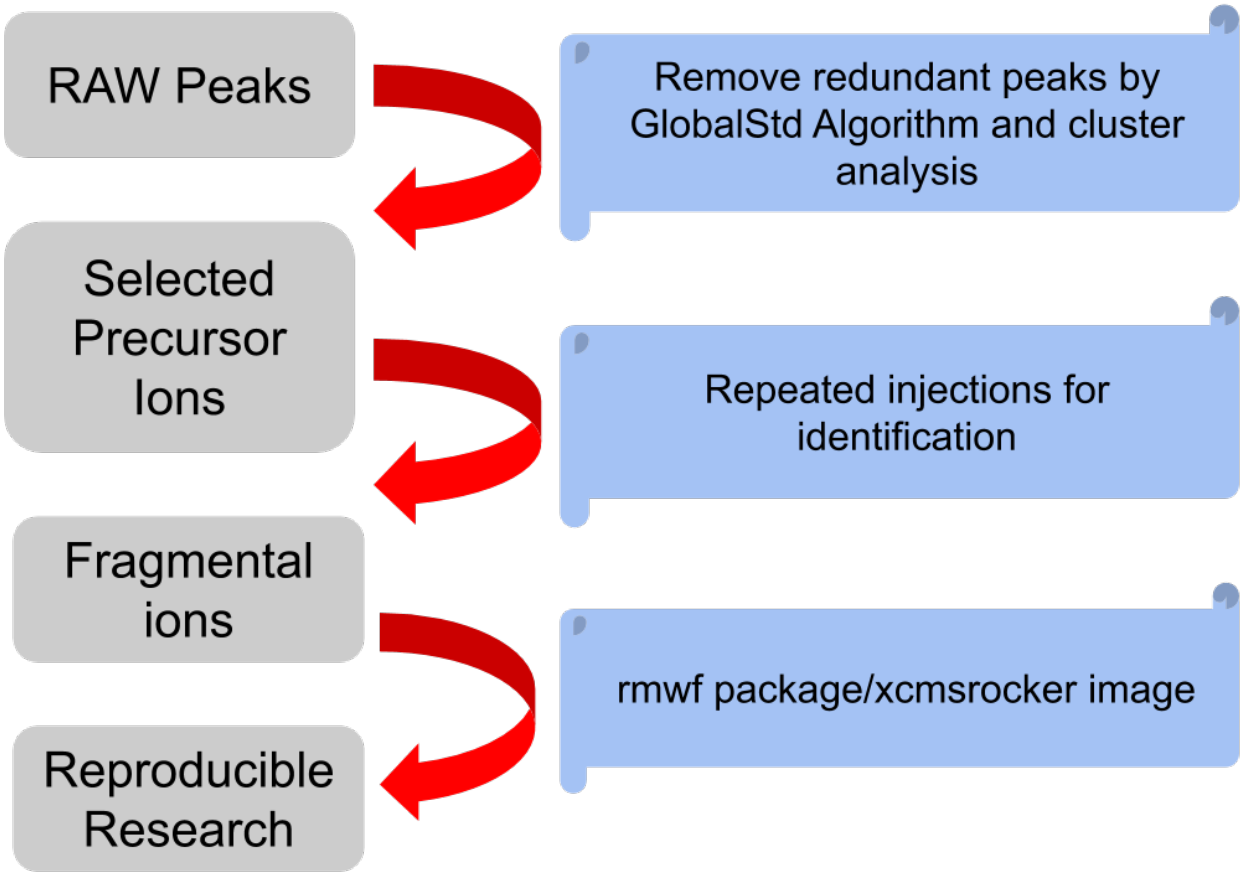

Figure 1. PMDDA workflow. Raw peaks are filtered by GlobalStd Algorithm to remove redundant peaks, then the remaining peaks are merged by cluster analysis to generate the precursor ion list. The selected peaks are assigned into multiple injections to collect the fragmental ions for structure identification. The whole analysis can be found as a data process template in the 'rmwf' package. The complete data analysis is reproducible as a xcmsrocker image.

Results and discussion

Precursor ion selection for MS/MS analysis

Using full scan mode, 6715 and 4666 features were measured in the NIST samples in positive and negative mode, respectively. After removal of peaks with fold change smaller than three times that of corresponding matrix samples and those peaks with a RSD less than 30%, 4711 and 3608 features remained in positive and negative mode, respectively, as potential precursor ions for MS/MS analysis.

For PMDDA, the GlobalStd algorithm was used to reduce the redundant peaks[16]. To select precursors for targeted analysis, each reduced independent peak was linked to their paired high frequency PMD ions as an ion cluster, or pseudo-spectra. Clusters were merged if independent peaks could be linked to the same paired ions. In addition, since ions within clusters should be highly correlated, Pearson correlation coefficients smaller than 0.9 between paired mass distances were used as a threshold to exclude unrelated peaks from the same compounds. For each merged ion cluster, the peak with the highest intensity was selected as the precursor ion for MS/MS analysis. For the SRM samples, in positive mode, 849 independent peaks were selected by the GlobalStd algorithm in which 780 precursor peaks were selected for targeted analysis after cluster analysis. In negative mode, 761 independent peaks generated 723 precursor peaks.

Precursor lists were generated for CAMERA and RAMClustR. For CAMERA[19], peak cluster groups following annotation of the feature table were treated as pseudo-spectra, and the proposed molecular weights for each pseudo-spectra were extracted. Then, the $[M+H]^+$ for positive mode and $[M-H]^-$ for negative mode were generated as precursor ions for targeted analysis. For the SRM samples, 862 and 710 precursor ions were generated for MS/MS annotation for positive and negative mode, respectively. Since RAMClustR[20] generated the molecular weight of each pseudo-spectra, the corresponding molecular ions ($[M+H]^+$ for positive mode and $[M-H]^-$ for negative mode) were generated for MS/MS analysis. For the SRM samples, 542 and 770 precursor ions were generated for positive and negative modes, respectively.

While several thousand features were measured in full-scan, the precursor ion selection process generated precursors for less than 1000 features, covering approximately 15% and 20% of the total feature numbers in positive and negative mode, respectively. Nevertheless, obtaining high quality MS/MS spectra for all of those features in a single injection with high sensitivity is challenging. In this case, the precursor ions were randomly assigned into multiple injections to make sure that no more than 6 ions were scanned within a retention time shift of 0.2 minutes of the original retention time from full scan. Such repeated injections for PMDDA, CAMERA, and RAMClustR were aimed to retain high sensitivity and compound coverage, and could be implemented into untargeted studies using pooled QC samples for untargeted MS/MS analysis.

Comparison with DDA, CAMERA and RamClustR

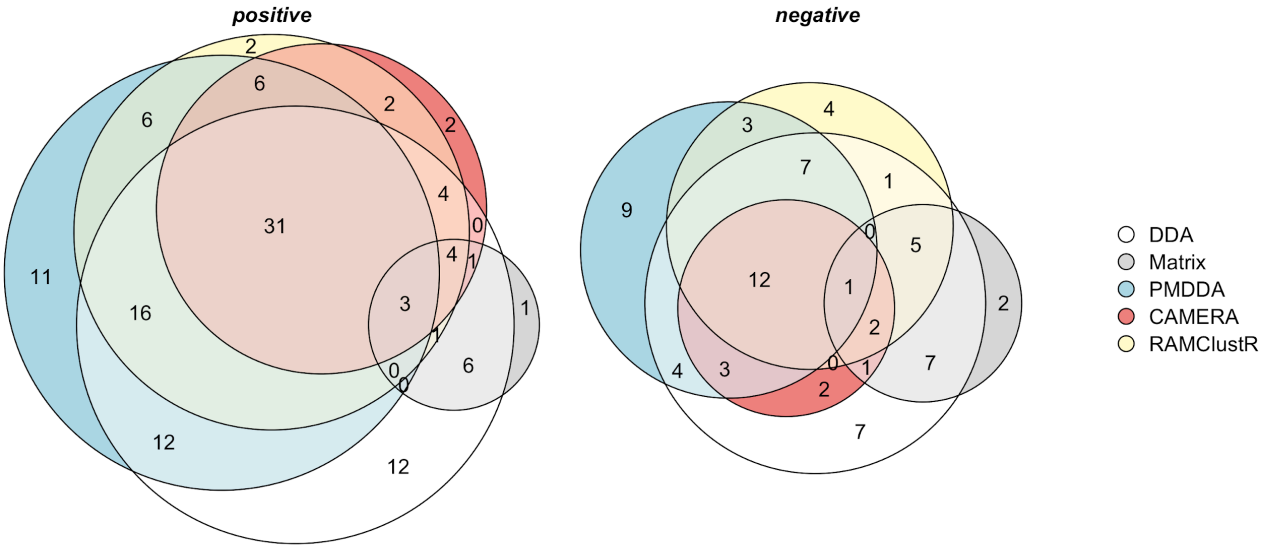

Figure 2. Euler diagram of identified compounds from DDA, CAMERA selected ions, RAMClustR selected ions, and PMDDA selected ions (left panel is positive mode data and right panel is negative mode data). The set of 'Matrix' means the identified compounds from matrix samples using DDA. The number of identified compounds that are overlapping in each analysis set is described.

Regular DDA was also performed for the SRM sample and matrix samples and the annotation results from GNPS compared to those obtained from PMDDA, CAMERA, and RAMClustR. As shown in figure 2, DDA identified 104 compounds and the DDA matrix identified 19 compounds in positive mode. Similarly, PMDDA identified 99 compounds. Both CAMERA and RAMclustR identified fewer compounds, 66 and 81, respectively. After removing compounds found in matrix samples, 118 unique compounds could be identified when DDA, PMDDA, CAMERA, and RAMClustR were used. However, only 31 of the compounds were identified in all four methods. Both PMDDA and DDA identified 11 unique compounds each, while CAMERA only identified 1 unique compound and RAMClustR only identified 2 unique compounds.

Results for negative mode were similar. DDA identified 52 compounds that included 3 compounds in the DDA matrix. PMDDA identified 41 compounds, CAMERA identified 25 compounds and RAMClustR identified 35 compounds. Among the 55 unique compounds found using all four methods after removal of compounds in matrix samples, only 13 compounds were overlapping between DDA, PMDDA, CAMERA, and RAMClustR. PMDDA identified 9 unique compounds similar to DDA (7). They both outperformed CAMERA (1) and RAMClustR (4).

SRM NIST 1950 contains 85 compounds with known exact masses including amino acids, fatty acids, clinical markers, etc. To compare the ability of each method to identify these known

compounds, protonated and deprotonated ions were generated as $[M+H]^+$ and $[M-H]^-$ for positive and negative modes, respectively. Then, the precursor ions selected from PMDDA, CAMERA, and RAMClustR were aligned among the m/z ions list for these known compounds within two decimal places. For positive mode, 0, 6, 3 and 5 ions matched in DDA, PMDDA, CAMERA and RAMClustR's precursor ions list while 1, 12, 9 and 4 ions matched in negative mode, respectively. This suggests that PMDDA performs as well or better than the other precursor selection algorithms for selecting biologically relevant compounds for MS/MS annotation.

Overall, PMDDA showed better coverage than both CAMERA or RAMClustR for untargeted annotation. This may be due to the fact that CAMERA and RAMClustR use pre-defined paired mass distances for adducts or redundant peaks, which may not accurately represent the specific sample type. PMDDA, on the other hand, employs a data-driven process to find high frequency paired mass distances within the samples, which may cover more unknown adducts or redundant peaks[16]. Another difference between PMDDA, CAMERA, and RAMClustR is the software design. The pmd package is designed to remove redundant peaks while CAMERA and RAMClustR are designed for annotation directly from the feature peak table. As such, the latter algorithms have not been optimized for generating a precursor list for MS/MS which may have decreased performance compared to PMDDA.

PMDDA showed complementary coverage to DDA. The precursor ion selection of PMDDA helped to identify 23 and 14 extra compounds not identified with DDA, in positive and negative mode, respectively. Although DDA methods can introduce a list of known contaminant peaks to exclude matrix compounds[14], the automatic data acquisition process of DDA suffers from collection of MS/MS of unknown background contaminants. DDA repeatedly collected MS/MS on background matrix ions (Figure 3) and contaminants (repeated compounds with same m/z, see Figure 3). However, when the ion list was pre-filtered by fold changes and RSD% filtering, precursor ion selection contained limited background ions and matrix compounds. In this case, precursor ion selection and DDA can be coupled together for an exhaustive annotation for unknown compounds.

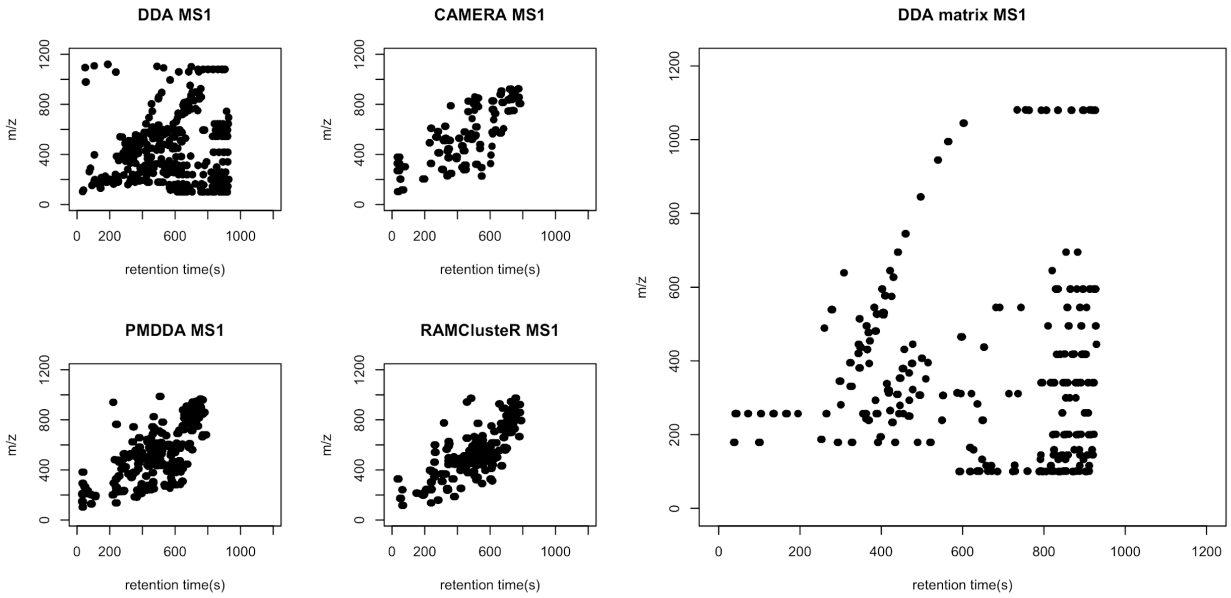

Figure 3. the metabolite profile of selected ions for MS/MS analysis (negative mode). DDA MS1 collected precursor ions from DDA. CAMERA, PMDDA, and RAMClustR displayed the selected precursor ions from corresponding software. DDA matrix MS1 shows the precursor ions from matrix samples, which includes probable contaminants (see horizontal repeated ions with the same m/z).

Compounds identified in both negative and positive ionization modes

To expand metabolite coverage, the same sample is typically analyzed under both negative and positive electrospray ionization modes for a given chromatography, and statistical analysis performed separately for both assays. However, compounds do not show the same ionization behavior in different modes, and respective peaks may be present in only one ionization mode or in both. This causes challenges for statistical analysis methods, such as false discovery rate control, which are highly dependent on the independent numbers of total compounds[35]. To overcome this, connections between negative and positive mode can be built after MS/MS annotation or identification, which might introduce bias on downstream statistical analysis. A previous study used correlation analysis to screen the same compounds in both modes[36], which can be influenced by redundant peaks from the same compounds. As an alternative, untargeted features present in both positive mode and negative mode can be determined using PMD.

Untargeted features present in both positive and negative mode can be linked by paired mass distance of 2.02 Da representing the difference between $[M+H]^+$ and $[M-H]^-$ in the two modes. For SRM samples, we found 100 peaks that could be linked with 2.02 Da within a retention time shift of 10s (see Figure 4). MS/MS annotation of those 100 peaks using PMDDA identified 31 unique compounds with GNPS, only 4 of which had the same annotation in both negative and positive mode due to the absence of a library spectra in the opposite mode. Since spectral

annotation databases might contain a more expansive coverage of only one ionization mode for certain compounds, linking through PMD could reduce the potential redundant annotations or facilitate annotation of unknowns. By linking features in positive and negative mode, the total number of independent metabolites was reduced for choosing the appropriate downstream statistical analysis. A limitation of the current algorithm is that this linkage only works on data analyzed on the same chromatography column and gradient.

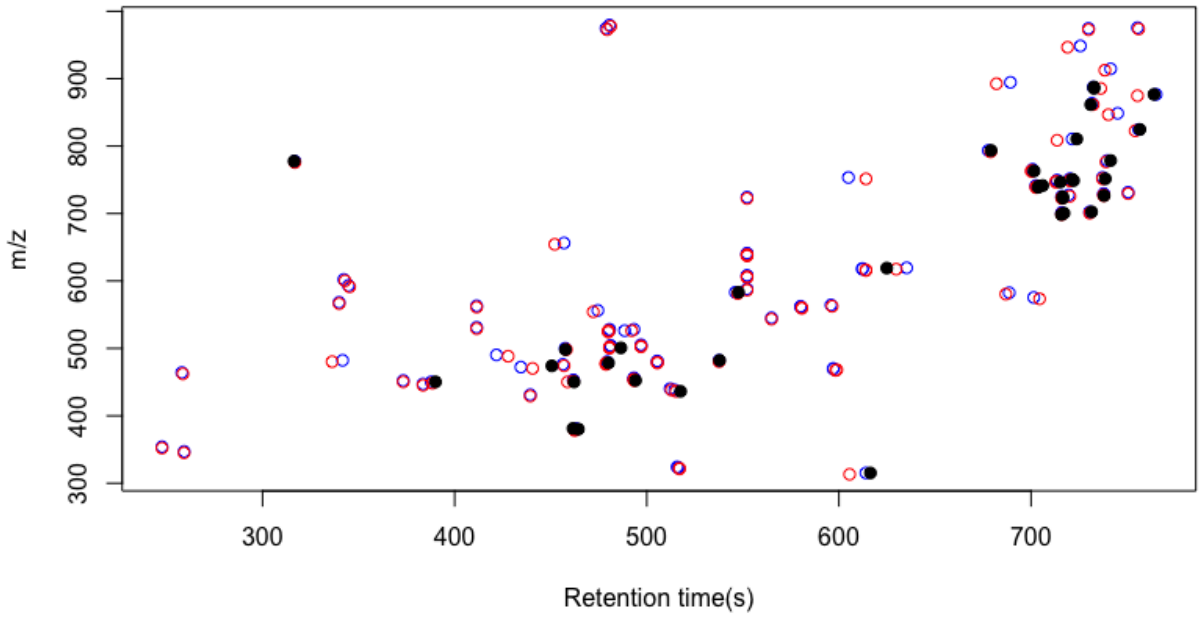

Figure 4. Features linked between positive and negative by PMD 2.02 Da within a retention time shift of 10s for positive and negative mode ionization. The red and blue circles represent positive and negative ions, respectively. Compounds with confirmed identities based on MS/MS annotation to GNPS are colored in black.

Reproducible research

We aimed to maximize reproducibility of this research. Therefore, we used SRM samples that are commercially available and commonly used in metabolomics workflows, and made the raw data accessible online for future potential research purposes. In order to provide full transparency on the data analysis, we choose a command line based script within a graphic user interface to make sure every step is recorded and reproducible by other researchers[26]. A docker image, xcmsrocker was created based on Rocker image[37], which pre-installs most of the R-based metabolomics and NTA data analysis software. This docker image is available online and can be installed on any personal computer, workstation, or cloud computation platform with RStudio as IDE[38]. Software used for this workflow such as IPO, xcms, pmd, CAMERA, and RAMClustR had been pre-installed. The R package rmwf is also included with the data processing script of this PMDDA workflow as a template, as well as other workflow templates

such as peak picking, annotation, or statistical analysis for different software. ‘xcmsrocker’ is freely available for download at https://hub.docker.com/r/yufree/xcmsrocker.

Conclusion

In this work, we propose an automated, reproducible, and exhaustive workflow to perform exhaustive MS/MS annotation based on precursor ions selection from full scan mode untargeted metabolomics data. We demonstrated that PMDDA outperforms both CAMERA and RAMClustR for breadth of pseudo-spectra precursor ions selection. In addition, this workflow can be coupled with typical DDA MS/MS analysis for even further annotation of unknown compounds. The PMDDA workflow was also able to identify features present in both negative and positive ionization modes, demonstrating the utility of the workflow to reduce duplicates for downstream statistical analysis. The PMDDA workflow is fully open source, reproducible, and includes all raw data and data processing scripts available online.

Acknowledgement

This work was supported by the National Institutes of Health/National Institute of Environmental Health Sciences grants U2CES030859, P30ES23515, R21ES030882, and R01ES031117.

References

1. Fessenden, M. Metabolomics: Small molecules, single cells. *Nature* **540**, 153–155 (2016).
2. Sobus, J. R. *et al.* Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA. *J. Expo. Sci. Environ. Epidemiol.* **28**, 411–426 (2018).
3. Yu, M. & Petrick, L. Untargeted high-resolution paired mass distance data mining for retrieving general chemical relationships. *Commun. Chem.* **3**, 1–6 (2020).
4. Tang, Y. *et al.* Advances in mass spectrometry-based omics analysis of trace organics in water. *TrAC Trends Anal. Chem.* **128**, 115918 (2020).
5. Barnes, S. *et al.* Training in metabolomics research. I. Designing the experiment, collecting and extracting samples and generating metabolomics data. *J. Mass Spectrom.* **51**, 461–475 (2016).
6. Mendez, K. M., Reinke, S. N. & Broadhurst, D. I. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics* **15**, 150 (2019).
7. Domingo-Almenara, X., Montenegro-Burke, J. R., Benton, H. P. & Siuzdak, G. Annotation: A Computational Solution for Streamlining Metabolomics Analysis. *Anal. Chem.* **90**, 480–489 (2018).
8. Chong, J., Wishart, D. S. & Xia, J. Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. *Curr. Protoc. Bioinforma.* **68**, e86 (2019).
9. Ljoncheva, M., Stepišnik, T., Džeroski, S. & Kosjek, T. Cheminformatics in MS-based environmental exposomics: Current achievements and future directions. *Trends Environ. Anal. Chem.* **28**, e00099 (2020).
10. Zhu, X., Chen, Y. & Subramanian, R. Comparison of Information-Dependent Acquisition, SWATH, and MSAll Techniques in Metabolite Identification Study Employing Ultrahigh-Performance Liquid Chromatography–Quadrupole Time-of-Flight Mass Spectrometry. *Anal. Chem.* **86**, 1202–1209 (2014).
11. Guo, J. & Huan, T. Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography–Mass Spectrometry Based Untargeted Metabolomics. *Anal. Chem.* **92**, 8072–8080 (2020).
12. Nash, W. J. & Dunn, W. B. From mass to metabolite in human untargeted metabolomics: Recent advances in annotation of metabolites applying liquid chromatography-mass spectrometry data. *TrAC Trends Anal. Chem.* **120**, 115324 (2019).
13. Wang, Y. *et al.* Enhanced MS/MS coverage for metabolite identification in LC-MS-based untargeted metabolomics by target-directed data dependent acquisition with time-staggered precursor ion list. *Anal. Chim. Acta* **992**, 67–75 (2017).
14. Koelmel, J. P. *et al.* Expanding lipidome coverage using LC-MS/MS data-dependent acquisition with automated exclusion list generation. *J. Am. Soc. Mass Spectrom.* **28**, 908–917 (2017).
15. Mahieu, N. G. & Patti, G. J. Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites. *Anal. Chem.* **89**, 10397–10406 (2017).

16. Yu, M., Olkowicz, M. & Pawliszyn, J. Structure/reaction directed analysis for LC-MS based untargeted analysis. *Anal. Chim. Acta* **1050**, 16–24 (2019).
17. Luo, P. *et al.* Multiple Reaction Monitoring-Ion Pair Finder: A Systematic Approach To Transform Nontargeted Mode to Pseudotargeted Mode for Metabolomics Study Based on Liquid Chromatography–Mass Spectrometry. *Anal. Chem.* **87**, 5050–5055 (2015).
18. Zeng, Z. *et al.* Ion fusion of high-resolution LC-MS-based metabolomics data to discover more reliable biomarkers. *Anal. Chem.* **86**, 3793–3800 (2014).
19. Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R. & Neumann, S. CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Anal. Chem.* **84**, 283–289 (2012).
20. Broeckling, C. D., Afsar, F. A., Neumann, S., Ben-Hur, A. & Prenni, J. E. RAMClust: A Novel Feature Clustering Method Enables Spectral-Matching-Based Annotation for Metabolomics Data. *Anal. Chem.* **86**, 6812–6817 (2014).
21. Sindelar, M. & Patti, G. J. Chemical Discovery in the Era of Metabolomics. *J. Am. Chem. Soc.* (2020) doi:10.1021/jacs.9b13198.
22. Liigand, P., Liigand, J., Kaupmees, K. & Kruve, A. 30 Years of research on ESI/MS response: Trends, contradictions and applications. *Anal. Chim. Acta* (2020) doi:10.1016/j.aca.2020.11.049.
23. Haug, K. *et al.* MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res.* **48**, D440–D444 (2020).
24. The Metabolomics Workbench. https://www.metabolomicsworkbench.org/.
25. Goodman, S. N., Fanelli, D. & Ioannidis, J. P. A. What does research reproducibility mean? *Sci. Transl. Med.* **8**, 341ps12-341ps12 (2016).
26. Hung, L.-H., Kristiyanto, D., Lee, S. B. & Yeung, K. Y. GUIdock: Using Docker Containers with a Common Graphics User Interface to Address the Reproducibility of Research. *PLOS ONE* **11**, e0152686 (2016).
27. Gandrud, C. *Reproducible Research with R and R Studio*. (CRC Press, 2013).
28. Boettiger, C. An introduction to Docker for reproducible research, with examples from the R environment. *ACM SIGOPS Oper. Syst. Rev.* **49**, 71–79 (2015).
29. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2020).
30. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **78**, 779–787 (2006).
31. Libiseller, G. *et al.* IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics* **16**, 118 (2015).
32. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
33. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
34. Reproducilble Metabolomics WorkFlow. https://figshare.com/projects/Reproducilble_Metabolomics_WorkFlow/59777.
35. Storey, J. D. False Discovery Rate. in *International Encyclopedia of Statistical Science* (ed. Lovric, M.) 504–508 (Springer Berlin Heidelberg, 2011). doi:10.1007/978-3-642-04898-

2_248.
36. Lee, H.-J., Kremer, D. M., Sajjakulnukit, P., Zhang, L. & Lyssiotis, C. A. A large-scale analysis of targeted metabolomics data from heterogeneous biological samples provides insights into metabolite dynamics. *Metabolomics* **15**, 103 (2019).
37. Boettiger, C. & Eddelbuettel, D. An Introduction to Rocker: Docker Containers for R. *R J.* **9**, 527–536 (2017).
38. RStudio Team. *RStudio: Integrated Development Environment for R.* (RStudio, PBC, 2020).