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

Motivation

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. 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. Here we propose an reproducible, automated, exhaustive, statistical model-free workflow: paired mass distance-dependent analysis (PMDDA), for untargeted mass spectrometry identification of unknown compounds found in MS1 full scan.

Results

More annotated compounds/molecular networks/spectrum were found using PMDDA compared with CAMERA and RAMClustR. Meanwhile, PMDDA can generate the preferred ions list for iterative DDA to cover more compounds when instruments support such functions.

Availability and implementation

The whole workflow is fully reproducible as a docker image xcmsrocker with both the original data and the data processing template. https://hub.docker.com/r/yufree/xcmsrocker A related R package is developed and released online: https://github.com/yufree/rmwf. R script, data files and links of GNPS annotation results including MS1 peaks list and MS2 MGF files were provided in supplementary information.

Contact

miao.yu@mssm.edu

Supplementary information

Supplementary data are available online.

Introduction

While metabolomics often aims at revealing changes in levels of all possible metabolites in biological samples(Fessenden, 2016), non-targeted analysis (NTA) usually aims at comprehensive profiling of compounds in environmental samples(Sobus *et al.*, 2018). To achieve these goals, both approaches use high-resolution mass spectrometry (HRMS) to perform unbiased measurement of small molecules followed by identification of unknowns(Yu and Petrick, 2020). In most HRMS-based workflows, small molecule profiles will first be extracted across samples as peaks or features(Tang *et al.*, 2020). Tens of thousands of features are typically extracted in each sample making it impractical to target every feature for MS/MS fragmentation(Barnes *et al.*, 2016). 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(Mendez *et al.*, 2019; Domingo-Almenara *et al.*, 2018). 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(Chong *et al.*, 2019). In contrast, group comparisons may not be available in ecological study designs or environmental investigations for supervised statistical analysis(Ljoncheva *et al.*, 2020). In this case, an exhaustive identification strategy of all possible small molecules needs to be developed to maximize the matching for quantification results from MS1 with annotation results from MS/MS analysis.

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(Zhu *et al.*, 2014). 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(Guo and Huan, 2020). In addition, DDA and DIA are designed for qualitative analysis instead of performing quantitative analysis with fragment ions(Nash and Dunn, 2019), 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(Wang *et al.*, 2017) or automated exclusion lists to cover more compounds during repeated DDA injections(Koelmel *et al.*, 2017). DDA will prefer precursors with higher intensity to reach similar sensitivity of MS1 full scan. A better preferred list of precursor ions can extend the coverage of DDA(Ten-Doménech *et al.*, 2020).

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(Wang *et al.*, 2017), new strategies needed to be developed for implementation in untargeted analysis for discovery and hypothesis generation. 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(Domingo-Almenara *et al.*, 2018; Mahieu and Patti, 2017). In practice, the number of unique compounds may be as little as twenty percent of the total feature numbers(Yu *et al.*, 2019). 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(Luo *et al.*, 2015; Zeng *et al.*, 2014), mainly using known adducts, neutral loss, and isotope pattern to detect the redundant peaks. Software packages such as CAMERA(Kuhl *et al.*, 2012) and RamClustR(Broeckling *et al.*, 2014) 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(Domingo-Almenara *et al.*, 2018). However, adducts or in-source reactions might be quite different among different sample matrices or instrument parameters(Sindelar and Patti, 2020), even for peaks from the same compound(Liigand *et al.*, 2020). 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(Yu and Petrick, 2020; Yu *et al.*, 2019). For example, sodium adducts should be considered only if paired mass distance (PMD) 21.98 Da appeared in high frequency. Some of the high frequency PMDs belong to known adducts while others might belong to unknown adducts, oligomers or combinations of known adducts. GlobalStd will try to remove the study specific redundant peaks as much as possible instead of using predefined adducts or reactions lists.

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(Haug *et al.*, 2020) or Metabolomics Workbenchs (https://www.metabolomicsworkbench.org/), details of data analysis are not as transparent as data sharing, and reduce the ability to fully reproduce the reported findings(Goodman *et al.*, 2016). Data analysis software with a graphic user interface (GUI) can be easy to use and document, but is also restricted to only defined operations(Hung *et al.*, 2016). An open source data process script can represent every step of the data analysis while still being flexible,(Gandrud, 2013) 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(Boettiger, 2015). 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 and link them with MS1 full scan data as much as possible. We then compared PMDDA with CAMERA and RamClustR precursor peaks selection algorithms using data acquired on standard reference material (NIST 1950) as demonstration. We also integrated PMDDA selected precursor ions with iterative DDA as a

preferred ions list to expand the compound's coverage of MS1 features. All of the data and data processing scripts are reproducible by a publicly available docker image.

Data and 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). The scan rate for MS1 full scan is 1.67spectra/s. The targeted analysis/ iterative DDA scan rate for MS1 is 4 spectra/s and 2 spectra/s for MS2 and 4 max precursors per cycle was set for iterative DDA.

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. The collision energy was set at 20 eV for all MS/MS fragmentation. In addition, eight injections of iterative DDA with PMDDA selected precursors as the preferred ions list were performed(Ten-Doménech *et al.*, 2020). For iterative DDA, ions selected as precursors in previous injections are removed from the list in the following injections. Use of a preferred ions list ensures the selected ions were fragmented if they were in the samples.

Data analysis

Data analysis was performed in R (version 4.0.2)(R Core Team, 2020) 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(Smith *et al.*, 2006) were optimized by IPO(Libiseller *et al.*, 2015) 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 linked with the MS2 annotation results from PMDDA, CAMERA, and RAMClustR selected precursor ions for comparison. 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(Chambers *et al.*, 2012) and annotated using GNPS(Wang *et al.*, 2016) molecular networking for MS/MS annotation with default settings (2Da shift for precursor ions to include isotope and 0.5Da shift for fragmental ions). Then annotation results were linked back to MS1 full scan filtered data for further investigation with <5ppm shift of mass-to-charge ratio and <5 second shift of retention time. Then the molecular networks and annotation results were compared among different methods.

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, theoretical m/z for 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.

MS/MS spectra of the peaks matched to the filtered MS1 features list as MGF files were extracted for further investigation or improved matching to the algorithm/database. The MS2 spectra were extracted by combining spectrum from similar precursor ions within 0.02Da, with fragmental ions shifted < 5 ppm, and only including peaks that were larger than 60% of the remaining spectra.

The whole PMDDA workflow (Figure 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 (https://figshare.com/projects/Reproducible_Metabolomics_WorkFlow/59777). Here, the MS/MS analysis can be targeted analysis with the selected precursor ions and/or various data-dependent acquisition modes with selected precursor ions as preferred ions when the instrument supports this feature. In addition, the workflow and corresponding software were packaged into a docker image called xcmsrocker (https://hub.docker.com/repository/docker/yufree/xcmsrocker). We also supplied the script, data files and links of GNPS annotation results for this study including MS1 peaks list and MS2 MGF files as supplementary information for reproducible purpose.

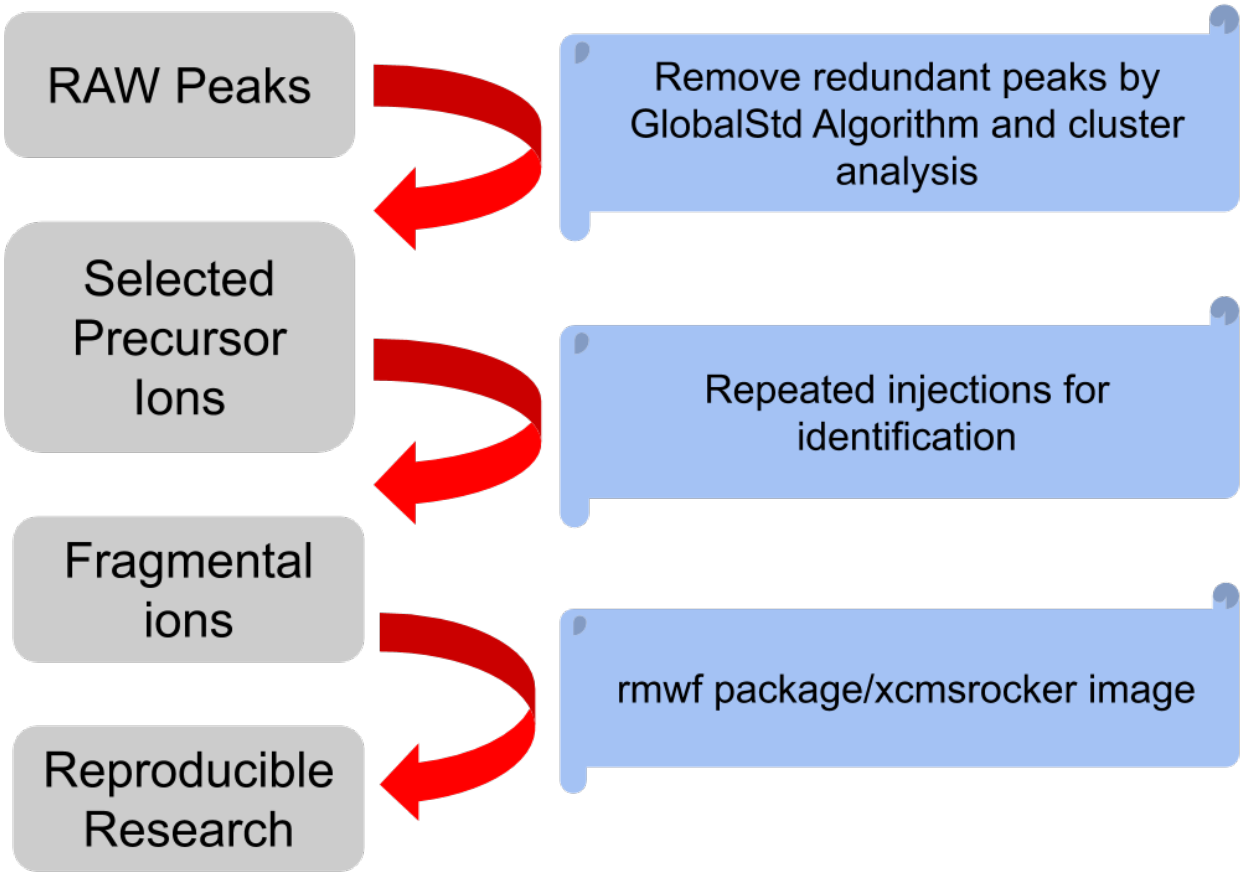

Figure 1. PMDDA workflow. Raw peaks are filtered by GlobalStd Algorithm to remove redundant peaks, then the remaining peaks are merged by cluster analysis to generate the precursor ion list. The selected peaks are assigned into multiple injections to collect the fragmental ions for structure identification. 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

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 larger 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(Yu *et al.*, 2019). 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 also generated for CAMERA and RAMClustR. For CAMERA(Kuhl *et al.*, 2012), 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(Broeckling *et al.*, 2014) 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 CAMERA and RamClustR

We compared the molecular networking results from GNPS for MS/MS collected using the PMDDA, CAMERA, and RAMClustR workflows (see Supplementary Materials for GNPS links). Here, only the results with precursor ions found in MS1 full scan were kept for comparison.

Figure S1 and S2 shows the MS1 full scan peaks covered by MS2 precursor ions using different methods. We find that only a subset of the MS2 precursor ions can be linked back to MS1 full scan data for iDDA. In this case, some peaks that can be annotated from MS2 data do not have available MS1 data for quantitative analysis. However, targeted analysis such as PMDDA can link MS1 and MS2 data as comprehensive studies.

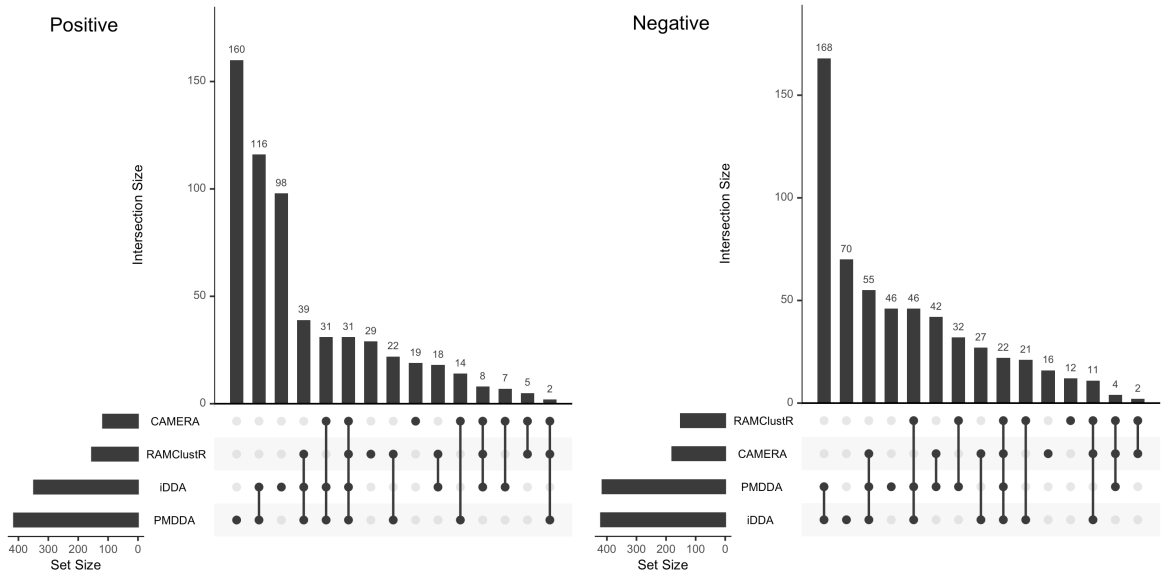

Figure 2. UpSet plot of metabolites networks found from CAMERA selected ions, RAMClustR selected ions, PMDDA selected ions, and iterative DDA (left panel is positive mode data and right panel is negative mode data). The set of ‘iDDA’ means iterative DDA with PMDDA selected precursor ions as the preferred list.

The chemical coverage of different methods were compared based on molecular networks found by GNPS, as well as annotation results. As shown in figure 2, PMDDA found 160 unique molecular networks and iDDA found 98 unique molecular networks and shared 116 unique molecular networks with PMDDA. Both CAMERA and RAMclustR identified fewer unique molecular networks compared with PMDDA, 19 and 29, respectively. However, only 31 molecular networks were identified in all four methods. For annotation results, as shown in figure S3, PMDDA found 73 compounds and iDDA found 77 compounds. Both CAMERA and RAMclustR identified fewer compounds, 29 and 41, respectively. However, only 16 compounds were identified in all four methods. PMDDA identified 6 unique compounds and another 23

compounds shared with iDDA while RAMClustR only identified 3 unique compounds and CAMERA didn't have any unique annotated compounds.

Results for negative mode were similar. As shown in figure 2, PMDDA found 46 unique molecular networks and iDDA found 70 unique molecular networks. PMDDA and iDDA shared 168 molecular networks. Both CAMERA and RAMclustR identified fewer unique molecular networks compared with PMDDA, 16 and 12, respectively. However, only 22 unique molecular networks were identified in all four methods. As shown in figure S3, PMDDA identified 113 compounds and iDDA identified 122 compounds. PMDDA and iDDA shared 18 compounds and iDDA found 6 unique compounds. CAMERA identified 31 compounds and RAMClustR identified 76 compounds. Only 4 compounds were overlapping between PMDDA, iDDA, CAMERA, and RAMClustR. Both CAMERA and RAMclustR had no unique compounds found. In this case, PMDDA outperformed CAMERA and RAMclustR and it would be helpful to perform iDDA to extend the coverage of molecular networks.

Known compounds in NIST 1950 were also compared among different methods. For positive mode, 6, 3 and 5 ions matched in PMDDA, CAMERA and RAMClustR's precursor ions list while 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.

Since the database-based annotation is biased towards compounds with available spectral data, and GNPS molecular networks may have multiple spectra from the same compounds, we also compared, by open source software, the number of unique MS1 compounds for which there was MS2 spectral data collected for CAMERA, RAMClustR, PMDDA. For positive mode, PMDDA could extract 293 spectra for unique MS1 compounds, more than CAMERA (34), RAMClustR (163). For negative mode, PMDDA found 254 spectra matching to unique MS1 data while CAMERA (46) and RAMClustR (150) have less spectra extracted.

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 pseudo spectra, which may cover more unknown adducts or redundant peaks(Yu *et al.*, 2019). As shown in Figure S4 and S5, some of the high frequency PMDs belong to known adducts while others might belong to unknown adducts, oligomers or combinations of known adducts. 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.

When we include the results from iterative DDA with the PMDDA selected precursor as the preferred list, the annotation performance can be further improved. However, PMDDA contains

some unique annotations missing by iterative DDA (see figure S3). On the other hand, as shown in figure S6 iterative DDA can cover compounds with lower intensity missing by other methods on MS1 full scan data. A combination of PMDDA as preferred ions list and iterative DDA data collection should be considered to reach a larger coverage of peaks found in MS1 full scan when the hardware supports such data acquisition mode.

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(Hung *et al.*, 2016). A docker image, xcmsrocker was created based on Rocker image(Boettiger and Eddelbuettel, 2017), 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(RStudio Team, 2020). 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 iterative DDA to cover more compounds found in MS1 full-scan. The PMDDA workflow demonstrates 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

Barnes,S. *et al.* (2016) Training in metabolomics research. I. Designing the experiment, collecting and extracting samples and generating metabolomics data. *J. Mass Spectrom.*, **51**, 461–475.

Boettiger,C. (2015) An introduction to Docker for reproducible research, with examples from the R environment. *ACM SIGOPS Oper. Syst. Rev.*, **49**, 71–79.

Boettiger,C. and Eddelbuettel,D. (2017) An Introduction to Rocker: Docker Containers for R. *R J.*, **9**, 527–536.

Broeckling,C.D. *et al.* (2014) RAMClust: A Novel Feature Clustering Method Enables Spectral-Matching-Based Annotation for Metabolomics Data. *Anal. Chem.*, **86**, 6812–6817.

Chambers,M.C. *et al.* (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.*, **30**, 918–920.

Chong,J. *et al.* (2019) Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. *Curr. Protoc. Bioinforma.*, **68**, e86.

Domingo-Almenara,X. *et al.* (2018) Annotation: A Computational Solution for Streamlining Metabolomics Analysis. *Anal. Chem.*, **90**, 480–489.

Fessenden,M. (2016) Metabolomics: Small molecules, single cells. *Nature*, **540**, 153–155.

Gandrud,C. (2013) Reproducible Research with R and R Studio CRC Press.

Goodman,S.N. *et al.* (2016) What does research reproducibility mean? *Sci. Transl. Med.*, **8**, 341ps12-341ps12.

Guo,J. and Huan,T. (2020) Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography–Mass Spectrometry Based Untargeted Metabolomics. *Anal. Chem.*, **92**, 8072–8080.

Haug,K. *et al.* (2020) MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res.*, **48**, D440–D444.

Hung,L.-H. *et al.* (2016) GUIdock: Using Docker Containers with a Common Graphics User Interface to Address the Reproducibility of Research. *PLOS ONE*, **11**, e0152686.

Koelmel,J.P. *et al.* (2017) Expanding lipidome coverage using LC-MS/MS data-dependent acquisition with automated exclusion list generation. *J. Am. Soc. Mass Spectrom.*, **28**, 908–917.

Kuhl,C. *et al.* (2012) CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Anal. Chem.*, **84**, 283–289.

Libiseller,G. *et al.* (2015) IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics*, **16**, 118.

Liigand,P. *et al.* (2020) 30 Years of research on ESI/MS response: Trends, contradictions and applications. *Anal. Chim. Acta*.

Ljoncheva,M. *et al.* (2020) Cheminformatics in MS-based environmental exposomics: Current achievements and future directions. *Trends Environ. Anal. Chem.*, **28**, e00099.

Luo,P. *et al.* (2015) 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.

Mahieu,N.G. and Patti,G.J. (2017) Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites. *Anal. Chem.*, **89**, 10397–10406.

Mendez,K.M. *et al.* (2019) 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.
Nash,W.J. and Dunn,W.B. (2019) 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.
R Core Team (2020) R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.
RStudio Team (2020) RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA.
Sindelar,M. and Patti,G.J. (2020) Chemical Discovery in the Era of Metabolomics. *J. Am. Chem. Soc.*
Smith,C.A. *et al.* (2006) XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.*, **78**, 779–787.
Sobus,J.R. *et al.* (2018) Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA. *J. Expo. Sci. Environ. Epidemiol.*, **28**, 411–426.
Tang,Y. *et al.* (2020) Advances in mass spectrometry-based omics analysis of trace organics in water. *TrAC Trends Anal. Chem.*, **128**, 115918.
Ten-Doménech,I. *et al.* (2020) Comparing Targeted vs. Untargeted MS2 Data-Dependent Acquisition for Peak Annotation in LC–MS Metabolomics. *Metabolites*, **10**, 126.
Wang,M. *et al.* (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.*, **34**, 828–837.
Wang,Y. *et al.* (2017) 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.
Yu,M. *et al.* (2019) Structure/reaction directed analysis for LC-MS based untargeted analysis. *Anal. Chim. Acta*, **1050**, 16–24.
Yu,M. and Petrick,L. (2020) Untargeted high-resolution paired mass distance data mining for retrieving general chemical relationships. *Commun. Chem.*, **3**, 1–6.
Zeng,Z. *et al.* (2014) Ion fusion of high-resolution LC-MS-based metabolomics data to discover more reliable biomarkers. *Anal. Chem.*, **86**, 3793–3800.
Zhu,X. *et al.* (2014) 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.