analytical chemistry

Article

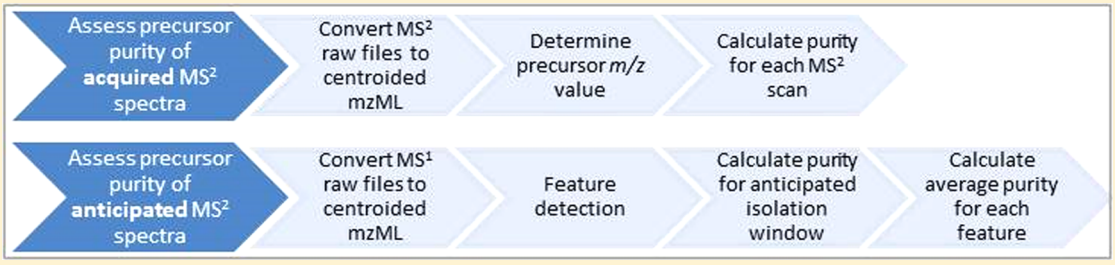
pubs.acs.org/ac

msPurity: Automated Evaluation of Precursor Ion Purity for Mass Spectrometry-Based Fragmentation in Metabolomics

Thomas N. Lawson,[†] Ralf J. M. Weber,[†,‡] Martin R. Jones,[†] Andrew J. Chetwynd,[†,‡] Giovanny Rodríguez-Blanco,[†,‡] Riccardo Di Guida,[†] Mark R. Viant,[†,‡] and Warwick B. Dunn*[,†,‡]

[†]School of Biosciences and [‡]Phenome Centre Birmingham, College of Life and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

Supporting Information



ABSTRACT: Tandem mass spectrometry (MS/MS or MS^2) is a widely used approach for structural annotation and identification of metabolites in complex biological samples. The importance of assessing the contribution of the precursor ion within an isolation window for MS^2 experiments has been previously detailed in proteomics, where precursor ion purity influences the quality and accuracy of matching to mass spectral libraries, but to date, there has been little attention to this data-processing technique in metabolomics. Here, we present msPurity, a vendor-independent R package for liquid chromatography (LC) and direct infusion (DI) MS^2 that calculates a simple metric to describe the contribution of the selected precursor. The precursor purity metric is calculated as "intensity of a selected precursor divided by the summed intensity of the isolation window". The metric is interpolated at the recorded point of MS^2 acquisition using bordering full-scan spectra. Isotopic peaks of the selected precursor can be removed, and low abundance peaks that are believed to have limited contribution to the resulting MS^2 spectra are removed. Additionally, the isolation efficiency of the mass spectrometer can be taken into account. The package was applied to Data Dependent Acquisition (DDA)-based MS^2 metabolomics data sets derived from three metabolomics data repositories. For the 10 LC-MS^2 DDA data sets with > ±1 Da isolation windows, the median precursor purity score ranged from 0.67 to 0.96 (scale = 0 to +1). The R package was also used to assess precursor purity of theoretical isolation windows from LC-MS data sets of differing sample types. The theoretical isolation windows being the same width used for an anticipated DDA experiment (±0.5 Da). The most complex sample had a median precursor purity score of 0.46 for the 64,498 XCMS determined features, in comparison to the less spectrally complex sample that had a purity score of 0.66 for 5071 XCMS features. It has been previously reported in proteomics that a purity score of <0.5 can produce unreliable spectra matching results. With this assumption, we show that for complex samples there will be a large number of metabolites where traditional DDA approaches will struggle to provide reliable annotations or accurate matches to mass spectral libraries.

Mass spectrometry (MS) is routinely used to (semi)-quantify, annotate, and identify small molecules (metabolites) in complex biological matrices. An MS experimental workflow can consist of infusing a sample directly into a mass spectrometer without any prior chromatographic separation (direct infusion mass spectrometry; DIMS),[1] but more often, the sample components are spatially separated via either gas or liquid chromatography (LC).[2] The predictable mass fragmentation patterns observed from the electron ionization (EI) technique commonly used with gas chromatography allows for reliable matching to mass spectral libraries such as NIST[3] and the Golm Metabolome Database.[4] So-called soft ionization techniques, including electrospray ionization (ESI), used in conjunction with LC methods provide minimal fragmentation because of the ionization process. In these cases, gas-phase fragmentation can be applied through collision-induced dissociation in hybrid quadrupole, ion trap, quadrupole time-of-flight (Q-TOF), or Orbitrap systems. The term "tandem mass spectrometry" (MS/MS or MS^2) is used when a single collision step is used, but product ions can be isolated for further collision to provide MS^n spectra where $n \geq 3$. The focus of this paper is on fragmentation resulting from tandem mass spectrometry.

A key component of any MS^2 (or higher) technology is the isolation of selected m/z windows for gas-phase fragmentation and the mapping back of the fragmentation (product) spectrum

Received: November 6, 2016
Accepted: January 24, 2017
Published: January 24, 2017

DOI: 10.1021/acs.analchem.6b04358
Anal. Chem. 2017, 89, 2432−2439

to the selected m/z window. In targeted and data-dependent acquisition (DDA)-based experiments, where MS^2 is performed on a dynamic list of precursor ions, as often determined by a preceding MS full scan, an isolation window is centered on the targeted precursor peak (m/z value). However, the isolation window can often contain more than one distinct peak, fragmentation spectra resulting from these situations being termed "chimeric"[5] and can be problematic for interpretation of the spectra and mass spectral library searching.

The more conservative estimates of chimeric spectra in proteomic DDA-based research are calculated as over 11.2% of all MS^2 spectra acquired,[6] but others have calculated this to be as high as 50% or more.[5,7] A direct comparison between different studies is difficult though as the number of chimeric spectra will vary considerably based upon sample complexity, instrumental configuration, and the procedure by which spectral chimerism is calculated.

The impact of chimeric spectra on spectral matching and annotation depends on the purity of the isolation window fragmented (i.e., the ratio between the relative intensity of the precursor divided by the summed intensity of all ions within the isolation window). If the purity of the precursor ion is sufficiently low, it can often be difficult to determine the origin of the resulting product ion(s). This in turn can lead to erroneous spectral matching results or no spectral matches.

Deconvolution of chimeric spectra however is possible and forms the basis of the data-analysis procedures applied to data independent acquisition (DIA) experiments. Where previous approaches in proteomics relied on comparison of fragmentation spectra to *in silico*-generated peptide databases,[8] alternative approaches have been derived from selected reaction monitoring data analysis methods (MS-SWATH,[9] OpenSWATH[10]). Although not as prevalent as DDA-based analyses, DIA-based approaches are also available in the field of metabolomics (MS-DIAL[11]).

A metabolomics analytical and data analysis workflow that directly takes into consideration the purity of an isolation window has also been developed, enabling deconvolution of MS^2 spectra.[12] The approach, demonstrated using an Agilent 6520 Q-TOF instrument, requires sliding isolation windows to be acquired surrounding the precursor of interest.

However, standard DDA-based approaches are still widely used and are popular in metabolomics,[13−15] and the deconvolution techniques detailed above are not always applicable to standard DDA-based acquisition techniques. In these cases, simply assessing the targeted precursor purity can be useful in interpreting the MS^2 spectra and aid in assessing the reliability of any subsequent annotation.

Some of the first software that had the functionality to count the number of chimeric spectra included the C/C++ software "hardklör"[6] and the python software "ChimeraCounter".[5] ChimeraCounter uses the "percent chimera intensity (PCI)" metric to count chimeric spectra when peaks that are not isotopes of the targeted precursor are within the isolation window and are above a user-defined percentage of the peak height of the precursor. Another metric has been previously described to calculate the "Precursor Ion Fragment" (PIF).[16] This is calculated as the targeted precursor ion intensity divided by the overall intensity in the isolation window and essentially gives a score of the purity of the precursor at the closest full scan (MS^1) to the MS^2 scan. The PIF implementation is provided within the quantitative proteomics software package MaxQuant.[17]

An automated computational method is detailed here to calculate the purity of fragmentation spectra post-acquisition from either a LC-MS^2 or DIMS^2 metabolomics experiment. What we call here "precursor purity" is calculated with a revised Michalski approach.[16] Advances include that the metric is interpolated at the recorded time of the MS^2 acquisition using bordering full scan spectra, the isolation efficiency of the mass spectrometer can be included within the calculation, and as per the PCI approach, isotopic peaks of the targeted precursor can be removed and peaks with intensities below a minimum percentage of the precursor peak intensity are removed from the calculation. The software has been applied to investigate 12 DDA and one DIA metabolomics data sets for different biological samples retrieved from the data repositories MetaboLights,[18] Metabolomics Workbench,[19] and PRIMe Data Resource of Plant Metabolomics (DROP Met) (http://prime.psc.riken.jp/). We also detail how theoretical isolation windows can be assessed using MS^1 data sets collected independent of MS^2 acquisitions. The computational methods detailed in this paper are available in the R package "msPurity". The package has been developed to work as a standalone or to be used in conjunction with the metabolomics peak detection and processing R package XCMS.[20]

MATERIALS AND METHODS

Overview of Software. An R package has been developed to assess the contribution of the targeted precursor in a fragmentation isolation window using a metric called "precursor purity". The use cases of the package can be divided into two types:

1. Assessing precursor purity of previously acquired MS^2 spectra: A user has acquired either LC-MS^2 or DIMS^2 spectra and an assessment is made of the precursor purity for each MS^2 scan.
2. Assessing precursor purity of anticipated isolation windows for MS^2 spectra: A user has acquired either LC-MS or DIMS full scan (MS^1) data and an assessment is to be made of the precursor purity of detected features using anticipated or theoretical isolation windows. This information can then be used to guide further targeted MS^2 experiments.

Availability. The msPurity package is available for Windows (both 32 and 64 bit), Mac OS, and Linux from the Bioinformatics software repository for R: Bioconductor $\geq$ 3.4. The code can also be downloaded directly from Github: https://github.com/Viant-Metabolomics/msPurity.

File Preparation. The input for the msPurity R package requires the raw vendor format of the mass spectrometry data to be converted into the mzML file format.[21] The controlled vocabulary used for the mzML format allows the isolation window widths used by the instrument to be parsed out of the mzML file from either Thermo, Agilent, AB Sciex, or Waters instruments. The conversion of instrument raw files to mzML can be achieved with the Proteowizard conversion tool MSconvert[22] or RawConvertor.[23]

If spectra were not collected in centroid mode, it is essential that the peaks are converted from profile to centroid format before processing within msPurity.

Assessing Purity of Acquired MS^2 Spectra. To assess the precursor purity of a MS^2 data set, the m/z value of the precursor ion associated with each MS^2 spectrum must first be determined. With the assumption that the most abundant peak

DOI: 10.1021/acs.analchem.6b04358
Anal. Chem. 2017, 89, 2432−2439

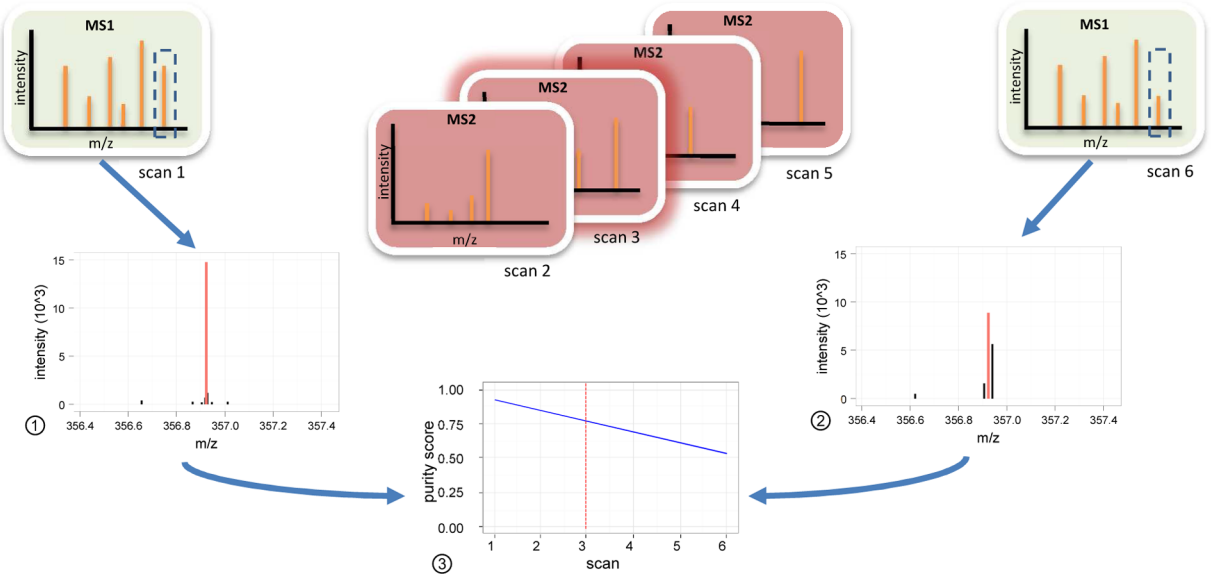


Figure 1. Assessment of acquired MS^2 precursor purity for a standard DDA-based experiment. Scans 1 and 6 are MS^1 scans, and scans 2 to 5 are MS^2 scans for different selected precursor ions where a DDA top 4 method (fragment the top 4 most intense ions) has been applied. The example in the figure calculates the precursor purity for scan 3. The precursor purity is calculated prior to the MS^2 acquisition (1) and after MS^2 acquisition (2), and the final score is then interpolated at the time of the MS^2 acquisition using bordering full-scan spectra (3).

in the window will have the highest contribution to the fragmentation spectrum, the most abundant peak within the isolation window is used to define the precursor m/z value. This is determined for the MS^1 scan closest to the MS^2 scan.

The precursor purity score of the targeted precursor ion is then calculated directly from the mzML spectra of the MS^1 scan as "target precursor intensity/summed intensity in isolation window". This simple ratio provides an easy to interpret score bounded between 0 and 1. A value of 1 should signify only the selected precursor has contributed to the resulting spectra, and values closer to 0 should signify that there is little or no contribution from the selected precursor ion to the summed intensity in the isolation window. It should be noted that this approach is unable to distinguish isobaric species.

It is not possible to calculate the score at the exact time of the MS^2 acquisition as the MS^1 spectra is typically not acquired simultaneously for DDA approaches. To account for this, the precursor purity is calculated using two MS^1 scans, one prior to and one post MS^2 acquisition, then interpolated at the point of MS^2 acquisition (Figure 1).

Assessing Purity of Anticipated Isolation Windows for MS^2 Spectra. The precursor purity calculation is performed using information derived solely from the MS^1 spectra; therefore, the calculation can be performed independently from any MS^2 acquisition.

The steps for assessing the precursor ion purity of anticipated MS^2 spectra using an LC-MS data set are as follows:

1. Feature detection is carried out within XCMS.
2. The full width half-maximum (fwhm) of the chromatographic peak is calculated for each XCMS feature.
3. The median precursor purity score (intensity of target/summed intensity in anticipated isolation window) is then calculated for the fwhm of the chromatographic peak of each feature (Figure 2). The median score is then calculated across any chosen samples or technical replicates.

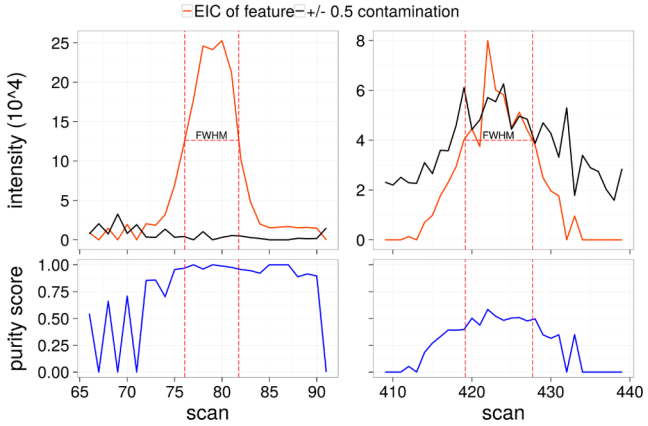


Figure 2. Assessing precursor purity of anticipated isolation windows for two XCMS features derived from a MS^2 data set. The extracted ion chromatogram (EIC) of each feature (red) and the intensity of contaminating peaks (blue) within a ±0.5 Da window are shown, scan range determined by XCMS, approximate full width half-maximum (fwhm) highlighted with red vertical lines.

The steps for assessing the precursor purity of anticipated MS^2 spectra using DIMS data sets consists of the following:

1. Basic processing centroid peaks performed within msPurity. This involves filtering out peaks below a user determined signal-to-noise level, averaging between multiple scans, filtering peaks not present in a minimum number of scans, filtering of peaks where the relative standard deviation of the intensity is below a given threshold, and removal of blank peaks from the sample peak list.
2. The median precursor purity score is determined for each feature across all scans in the data set.

The predicted purity scores can be used to select features for MS^2 in a targeted or semi-targeted experiment, e.g., using a "nearline" approach.[24] Additionally, the information can be

DOI: 10.1021/acs.analchem.6b04358
Anal. Chem. 2017, 89, 2432−2439

used to assess what isolation window width would be suitable for any subsequent MS^2 experiments.

Removal of Isotopic Peaks. The msPurity package can determine if the ^{13}C isotopic peak of the targeted precursor is within the isolation window or if the targeted precursor is the ^{13}C isotopic peak. This is performed for single, doubly, and triply charged species and applies the m/z difference (1.0033, 0.5017, and 0.3344, respectively) to identify the ^{13}C isotopic peaks. The isotopic related peak that is not the precursor can then be omitted from any of the purity calculations. See Supporting Information, Section 1, for further details. Whether or not isotopic peaks are to be removed should depend on the downstream analysis to be performed; for consistency, all results shown in the paper do not omit the isotopic peaks from the purity calculations.

Removal of Peaks of Relative Low Abundance. When using larger isolation windows, there can often be numerous peaks that have a low intensity relative to the targeted precursor peak and that have little impact on the product ion fragmentation spectrum. When these peaks are included in the purity calculation, a lower precursor purity score can be calculated, which is misleading. To accommodate for this, only peaks that have an intensity greater than 5% of the targeted precursor ion peak are included in the precursor purity calculation (the threshold of 5% can be adjusted by the user). A similar approach is used in the PCI calculation.[5]

Isolation Efficiency. Isolation efficiency is defined here as the effect of the position of an ion within an isolation window on its relative intensity in corresponding fragmentation spectra. When the isolation efficiency of an instrument is known, the peak intensities within an isolation window can be normalized for the precursor purity calculation.

The isolation efficiency can be estimated experimentally by acquiring MS^2 spectra where the precursor ion of interest is experimentally measured at different locations within the isolation window. The precursor ion peak, or a known product ion peak, can then be monitored across the isolation window. The resulting intensity profile is then converted to be within a scale of 0 to 1, where 1 is the position in the isolation window with highest response and 0 is where the precursor ion is not detected. This efficiency profile can then be used to normalize the peak intensities based on their position within the isolation window.

For this study, the isolation efficiency experiment was performed using a Thermo Scientific Q Exactive Focus mass spectrometer operated with a ±0.5 Da isolation window. Isolated ions had m/z values of 195.0875, 524.2649, and 922.0086 and were generated from constituents of a Pierce LTQ Velos ESI positive ion calibration solution (product number #88323). The mean isolation efficiency profile was then used to create a simple linear model using B-spline polynomials (five anchors) to predict the isolation efficiency based on isolation window position (adjusted R^2: 0.9812; F: 210; p-value < 0.001) (Figure 3).

In addition, an assessment of isolation windows for Waters instruments was also carried out using a Xevo G2 XS QTOF mass spectrometer tuned to a mass resolution of 40,000 and run in positive ESI mode. When operated in MS^2 mode, the quadrupole low mass resolution was manually tuned to 23 providing a ±0.5 Da isolation window. Isolation windows surrounding ions at m/z 269.20, 362.95, and 906.81 from a solution of 0.5 mM sodium formate in 90:10 isopropanol:water

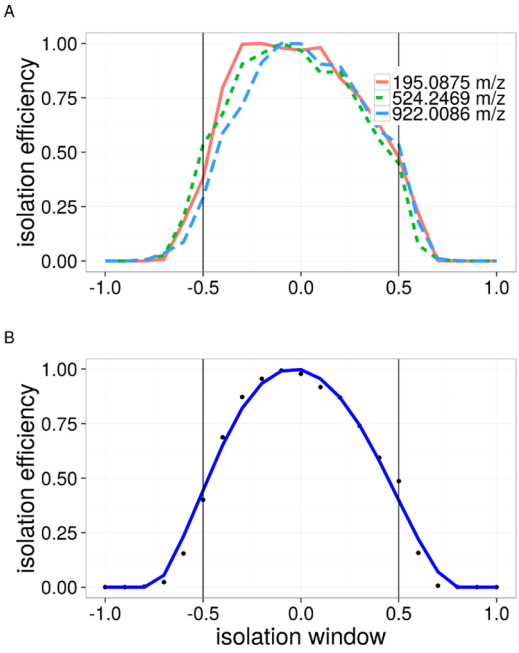


Figure 3. Overview of isolation efficiency experiment determined using a Thermo Scientific Q Exactive Focus. (A) Isolation efficiency profile of Thermo Scientific Q Exactive Focus mass spectrometer on the following ions: 195.0875, 524.2649, and 922.0086 m/z. (B) A simple linear model using B-spline polynomials to predict isolation efficiency based on isolation window position for the Thermo Scientific Q Exactive Focus mass spectrometer (negative values have been zeroed).

were assessed. The mean isolation efficiency profile observed is similar in shape to a Gaussian curve (Figure S1)

Public Metabolomic Data Sets. A total of 12 DDA and one DIA data sets were obtained from the public data repositories MetaboLights,[18] Metabolomics Workbench,[19] and PRIMe Data Resource of Plant Metabolomics (DROP Met) (http://prime.psc.riken.jp/) to be assessed for their precursor purity. See Table S2 for a summary of the studies. As of December 31, 2015, this is an exhaustive list of DDA and DIA data sets that include either the RAW, mzML, or other XML open source data formats. Further details of the experimental setup for each study can be located in the corresponding submission entry on the relevant data repository web site and within the associated papers for each study.

If not already in a mzML centroided data format, the files were converted using Proteowizard MSconvert (v3.0).[22] The purity assessments were made for all MS^2 scans for every sample using the msPurity package using the function "purityA" with standard parameter settings. Importantly though, the precursor purity was calculated for all studies using the most intense peak within the isolation window as the target precursor ion, isotopic peaks were not removed, and no isolation efficiency normalization was performed. The efficiency normalization was omitted from the public data as it was not possible to retrospectively produce comparable isolation efficiency profiles for the multiple instruments used.

If possible, the isolation window widths were extracted from the mzML file. However, for data sets derived from the Agilent Q-TOF, the isolation window widths are not included in the mzML file. For Agilent Q-TOF instrumentation, the isolation windows are determined as either narrow (≈1.3 amu), medium (≈4 amu), or large (≈9 amu), with the exact widths changing

DOI: 10.1021/acs.analchem.6b04358
Anal. Chem. 2017, 89, 2432−2439

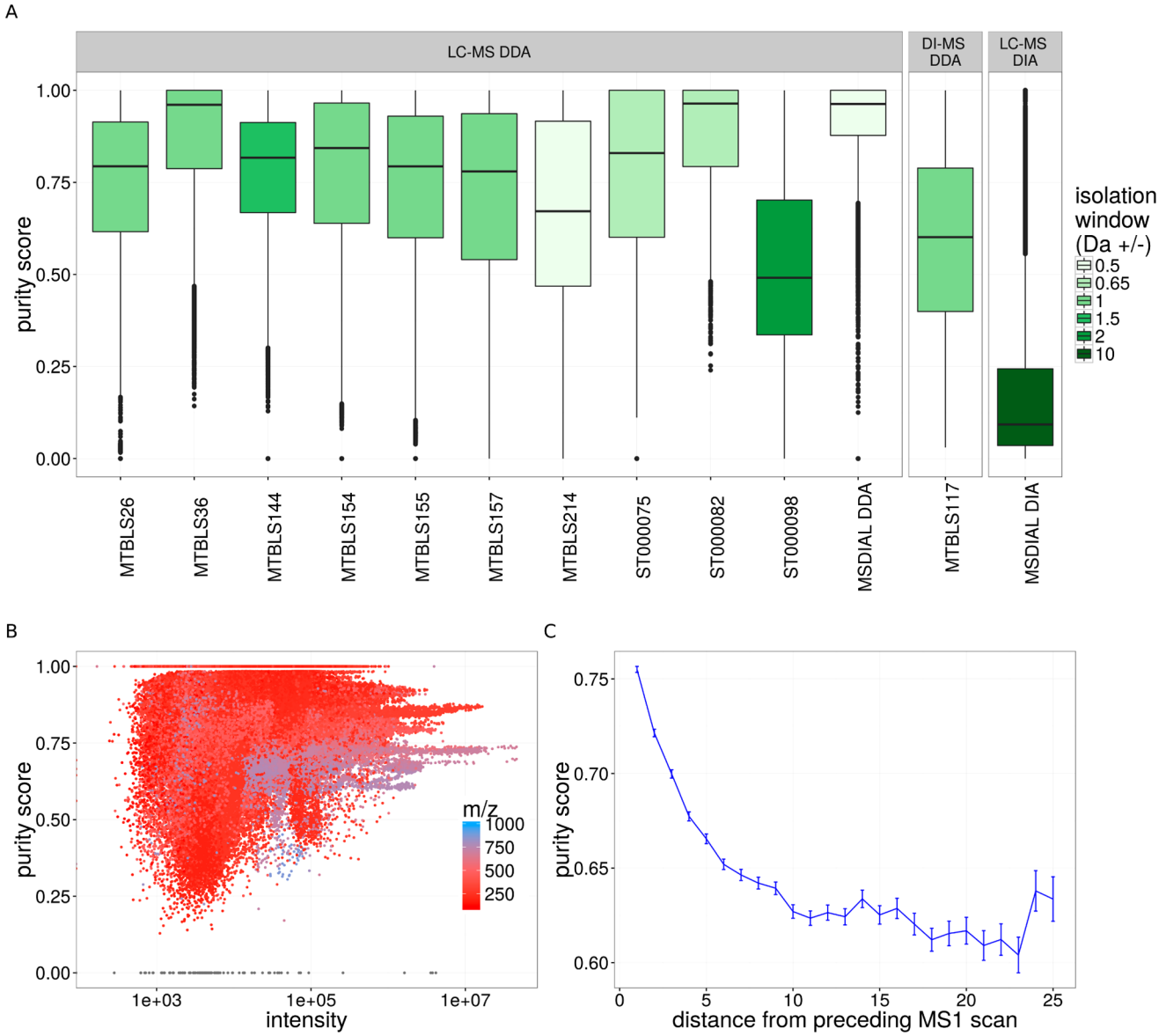


Figure 4. Overview of assessment of precursor purity for public DDA and DIA metabolomic data sets. (A) Boxplots of calculated purity scores of all fragmentation spectra. See Figure S2 for violin plots derived from the same data. (B) Summary of purity score relationship with selected precursor intensity and m/z value for study MTBLS144; x axis is in $\log_{10}$ scale. (C) The mean purity relationship with the distance from the preceding MS1 scan for study MTBLS214. Error bars represent one standard error of the mean.

depending on the instrument. In the case of the Agilent data sets used here (ST000075 and ST000085), a narrow resolution of ≈1.3 amu was used. For these files, an approximate isolation window of ±0.65 Da centered on the precursor ion m/z was used for the purity assessments. For Agilent Q-TOF instruments, the medium and large window widths are not currently supported within msPurity, as these larger isolation window widths are not evenly centered on the recorded precursor m/z in the mzML file.

In-House Metabolomic Data Sets. Two LC-MS metabolomic data sets were acquired to assess XCMS-derived metabolite features for precursor purity.

Daphnia magna extract (using 2.5:1 v/v methanol:water) was separated using a reversed-phase Syncronis Phenyl LC column, and mass spectral detection of the LC column eluates was performed using a Thermo Scientific Q Exactive mass spectrometer equipped with a H-ESI II source acquiring in positive ionization mode.

Additionally, *Ovis aries* (sheep) kidney tissue extracts (using 70/30 methanol/water) were separated using a Hypersil Gold C18 reversed phase column and analyzed using a Thermo Scientific Q Exactive Focus mass spectrometer. Feature detection was carried out using the XCMS package[20] with the following parameters: peak picking algorithm = centWave, snthrs = 5, ppm = 5, prefilter = c(3, 5000), integrate = 1, and mzdiff = 0.001. Peaks were aligned by retention time using the retcor.loess algorithm and grouped using the group.density algorithm with the following settings: bw = 5, mzwid = 0.01, and minfrac = 0.5. The msPurity function "purityX" was run on the xcmsSet objects to assess the precursor purity of theoretical isolation windows of ±0.5 Da for XCMS features using the isolation efficiency normalization predicted for the Thermo Fisher Q Exactive mass spectrometer. See Supporting

DOI: 10.1021/acs.analchem.6b04358
Anal. Chem. 2017, 89, 2432−2439

Information, Section 3, for a more extensive description of the materials and methods for the in-house metabolomics data sets.

RESULTS AND DISCUSSION

Assessment of Precursor Ion Purity for Publically Available DDA and DIA Data Sets. Figure 4 and Table 1 summarize the purity of available LC-MS and DIMS-based DDA data sets and a single LC-MS DIA data set.

A direct comparison of precursor ion purity between the different data sets is difficult as the score will vary considerably based on the complexity of the sample, instrument used, isolation window width, and the exact details of how the DDA was performed (e.g., if blank exclusion was applied, length of time on dynamic exclusion and number of MS^2 scans per MS^1 scan (top *n* method)). However, by using this wide range of studies, a broad overview can be shown of the extent of precursor ion purity in metabolomics.

For nearly all of the LC-MS DDA-based studies where the isolation window width was $\leq \pm 1$ Da, the median precursor purity score was >0.7; the exception is study MTBLS214 where the median precursor purity was 0.67. One factor contributing to the lower score will be the larger number of consecutive MS^2 scans (up to 26) being performed between MS^1 scans, whereas for all other DDA studies the maximum number of consecutive MS^2 scans was 12. The large number of MS^2 scans acquired consecutively can lead to lower intensity precursor ions selected for fragmentation that would typically be ignored with a smaller number of consecutive MS^2 scans. This is demonstrated in Figure 4C, where the purity of the MS^2 spectra decreases with the distance from the preceding MS^1 scan. The study MTBLS214 also highlighted the need for caution when the number of consecutive MS^2 scans is high. For this study, the nearest MS^1 scan used to calculate the purity score was up to ≈5.2 s (12 scans) away from the MS^2 acquisition.

The precursor purity scores reported here are in general agreement with the field of proteomics, where the median precursor purity score (or PIF, within this field) of targeted peptides was calculated as 0.85 and 0.73 for isolation windows of ±0.5 Da and ±1 Da, respectively.[16]

Data were acquired in both positive and negative ionization methods for 11 of the public data sets and allowed a comparison between the precursor ion purity scores derived from positive and negative methods. In general, the positive ionization data had higher precursor purity scores than negative ionization data (mean difference of +0.04), but this was not always the case. For 7 out of the 11 data sets, the purity score was significantly higher for the positive ionization mode (unpaired Wilcoxon rank sum adjusted p-value using Benjamini & Hochberg approach < 0.0001), but for 3 out of the 11 data sets, the purity score was significantly smaller for the positive ionization mode (unpaired Wilcoxon rank sum adjusted p-value using Benjamini & Hochberg approach < 0.0001). See Table S2 for further details.

It should be noted that all of the scores presented in Figure 4 used the most intense peak within the isolation window for the target precursor peak. When the peak that is at the center of the isolation window is chosen for the calculation, then the precursor purity scores were significantly lower for all studies (mean difference for all studies: −0.04, paired Wilcoxon rank sum adjusted p-value using Benjamini & Hochberg approach < 0.0001). For the smaller isolation windows, the impact is marginal, but for the larger isolation windows observed for DIA experiments this has a much larger impact on the purity score

Table 1. Summary of Public Metabolomics Data Sets Used To Assess Precursor Purity[a]

	study							precursor purity score		
MS^2 method	ID	ref	species	column	MS vendor	MS Instrument	isolation window width (Da)	min	median	max
DDA	MTBLS26	25	*M. musculus*	Hypersil GOLD C18 HPLC	Thermo Scientific	LTQ-FT	±1	0.00	0.79	1.00
DDA	MTBLS36	13	*S. lycopersicum*	Waters Acquity UPLC	Thermo Scientific	LTQ Orbitrap Velos	±1	0.14	0.96	1.00
DDA	MTBLS144	NA	*T. pseudonana*	Synergi Fusion reversed phase	Thermo Scientific	LTQ-FT	±1.5	0.00	0.82	1.00
DDA	MTBLS154	14	*T. pseudonana*	Synergi Fusion reversed phase	Thermo Scientific	LTQ FT Ultra	±1	0.00	0.84	1.00
DDA	MTBLS155	NA	*S. elongatus*	Synergi Fusion reversed phase	Thermo Scientific	LTQ FT Ultra	±1	0.00	0.79	1.00
DDA	MTBLS157	NA	*R. pomeroyi* DSS-3	Synergi Fusion reversed phase	Thermo Scientific	LTQ FT Ultra	±1	0.00	0.78	1.00
DDA	MTBLS214	26	*H. sapiens*	Prominence UFLC (Shimadzu)	AB Sciex	TripleTOF 5600	±0.5	0.00	0.67	1.00
DDA	ST000075	15	*M. musculus*	Waters Acquity CSH C18/ Waters Acquity VanGuard CSH C18	Agilent	6530 QTOF	±0.65	0.00	0.83	1.00
DDA	ST000082	NA	*H. sapiens*	Waters Acquity CSH C18/ Waters Acquity VanGuard CSH C18	Agilent	6530 QTOF	±0.65	0.24	0.96	1.00
DDA	ST000098	NA	*C. elegans*	Waters Acquity UPLC	Thermo Scientific	Q Exactive	±2	0.00	0.49	1.00
DI-MS DDA	MTBLS117	NA	*G. max*	NA	Thermo Scientific	LTQ Orbitrap Classic	±1	0.03	0.60	1.00
DIA	MS-DIAL: DIA	11	*Chlamydomonas/Chrollera*	HILIC	AB Sciex	TripleTOF 5600+	≈ ±10	0.00	0.09	1.00
DDA	MS-DIAL: DDA	11	*Chlamydomonas/Chrollera*	HILIC	AB Sciex	TripleTOF 5600+	±0.5	0.00	0.96	1.00

[a]See Table S1 for more information regarding each study.

DOI: 10.1021/acs.analchem.6b04358
Anal. Chem. 2017, 89, 2432−2439

reported (−0.16 difference in purity). See Table S2 for full details. The Raw Convertor[23] tool also uses the most intense peak in the isolation window as the precursor for the same assumption used here; the most abundant peak within the isolation window will contribute most to the fragmentation spectra.

For the only public DIA data set available, the median purity score is <0.1. The lower purity score for the DIA data set is expected due to the much larger isolation window width. As the principle behind DIA is to deconvolute spectra derived from these large isolation windows, the purity score for DIA-based data sets is in some regards inconsequential. Nevertheless, this data set provides a good example to demonstrate that small precursor ion purities are observed when very large isolation window widths are employed.

For the one public DIMS DDA data set analyzed, the median precursor purity score was 0.54. A lower precursor purity score was expected due to the lack of chromatography and, therefore, more complex MS^1 scans as well as a longer period to acquire MS^2 spectra. This leads to a higher number of low intensity precursor ions being fragmented that typically have lower precursor purity scores. It should be acknowledged here that the isobaric overlap from DIMS would also typically be higher than LC-MS if no prior fractionation has been performed.

These assessments confirm that the precursor ion purity calculation can be performed on various types of MS^2 experiment (DDA, DIA, LC-MS, and DIMS) from several instrument manufactures (Thermo Scientific, AB Sciex, and Agilent) and can provide a simple assessment of "chimeric" spectra in a data set. If no deconvolution of the chimeric spectra is to be performed, then spectra below a certain precursor purity score can be removed to improve the accuracy of any later spectral matching and annotation.

A plot of the intensity of the selected precursor ion against the precursor purity shown in Figure 4B highlights a trend where the most intense precursors will have higher purity (closer to 1) but also shows that many precursors of lower intensity can still have high purity scores. This implies that simply having a higher threshold for the intensity of the precursor ion to be chosen for fragmentation is insufficient in many cases to improve the precursor ion purity of MS^2 spectra.

An acceptable level of precursor purity for DDA or targeted analysis could be considered to be at a level at where spectral matching to a library compound provides successful annotation or identification. In the field of proteomics, when the precursor purity score is at a relatively high level (≥0.5), the identification rate of proteins stays at nearly 60%.[16] Similar rates are difficult to calculate for metabolomics data sets due to the lack of a comprehensive database of metabolites and associated spectra. However, it is reasonable to be cautious of any MS^2 mass spectra where the target precursor ion is thought to have contributed less than half of the ions that have been fragmented. We acknowledge though that the number of other ions present in the isolation window, their relative intensity compared to the chosen precursor ion, and the complexity of the fragmentation mass spectra for each non-precursor ion will also have a significant contribution to the ability to accurately apply the MS^2 spectra for mass spectral matching.

Assessing Precursor Purity of Anticipated MS^2 Spectra. The anticipated precursor purity calculated for the fwhm of the chromatographic peak for all XCMS features (following blank subtraction) determined from two in-house LC-MS data sets is summarized in Figure 5.

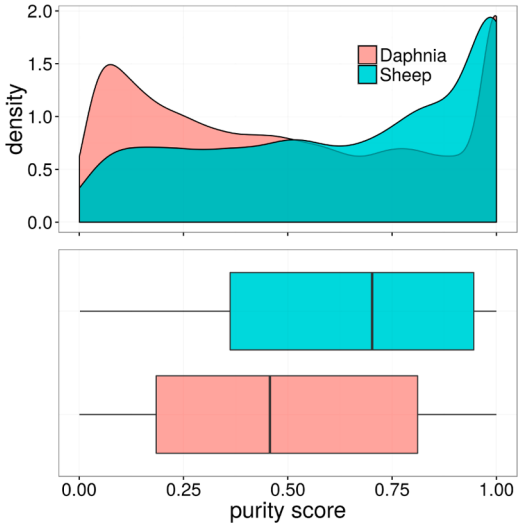


Figure 5. Overview of the precursor purity of XCMS features calculated for theoretical isolation windows of an anticipated MS^2 experiment. Calculated for *D. magna* extracts (highly complex sample) and sheep kidney extracts (less complex sample).

The complex *D. magna* sample consisting of 64,498 XCMS features had a median purity score of 0.46. The less spectrally complex sheep kidney (5071 XCMS features) had a median precursor purity score of 0.66. The number of XCMS features with anticipated precursor purity scores <0.5 was 53% for the *D. magna* sample and 39% for the sheep kidney sample. If we assume that reliable spectral matching requires a precursor purity score >0.5, the results demonstrate there would be a considerable number of XCMS features where the precursor purity score is expected to be so low that it will be problematic for standard MS^2 spectral matching. If further MS^2 experiments were to be performed on a targeted group of these features, the precursor purity score could then be used to direct the acquisition of XCMS features where reliable MS^2 matching could be performed.

Validation of Anticipated Score of XCMS Feature. To show how stable the anticipated precursor purity scores were for the *D. magna* sample, an LC-MS^2 DDA experiment was conducted on the same data set as the original LC-MS experiment. When there was an XCMS determined feature observed between both the LC-MS and LC-MS^2 data sets, it allowed a comparison to be made of both the anticipated (i.e., predicted) precursor purity score and the precursor purity score from the acquired MS^2 data. A total of 2873 features were assessed with a median of four MS^2 spectra used for each feature. The median difference between the anticipated and observed precursor purity score was 0.05.

CONCLUSION

The msPurity package has been designed to evaluate precursor purity or chimera in metabolomics MS^2 data sets by calculating an easily interpretable metric (intensity of selected precursor/summed intensity of the isolation window) similar to the previously described method used in proteomics.[16] However, the method described here differs from the Michalski[16] approach in that we interpolate peak intensity of MS^2

DOI: 10.1021/acs.analchem.6b04358
Anal. Chem. 2017, 89, 2432−2439

acquisitions using bordering full-scan spectra, and we remove low abundant peaks that are believed to have limited contribution to resulting MS^2 spectra and can optionally remove isotopic peaks and take into account the isolation efficiency of the mass spectrometer.

Using the msPurity package, we present here the first comprehensive characterization of precursor purity or chimera in metabolomics and show that for complex samples a high proportion of metabolite features will have low (<0.5) precursor purity, which could be potentially problematic for standard DDA-based spectral matching analysis.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b04358.

Isolation efficiency profile of Waters Xevo G2 XS QTOF. Further details for public data sets. Further details for isotopic peak calculation. "Violin plots" for public precursor purity results. Positive and negative ionization comparison results. Peak center vs most intense peak in isolation window comparison. Further materials and methods for in-house experiments. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: w.dunn@bham.ac.uk.

ORCID

Thomas N. Lawson: 0000-0002-5915-7980
Andrew J. Chetwynd: 0000-0001-6648-6881
Giovanny Rodríguez-Blanco: 0000-0002-7154-7244
Warwick B. Dunn: 0000-0001-6924-0027

Author Contributions

All authors have given approval to the final version of the manuscript.

Funding

This work was supported financially through two NERC CASE Ph.D. studentships in collaboration with GigaScience (NE/L002493/1) and Thermo Fisher Scientific (NE/J017442/1).

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Kirwan, J. A.; Weber, R. J. M.; Broadhurst, D. I.; Viant, M. R. *Sci. Data* **2014**, *1*, 1−13.

(2) Dunn, W. B.; Broadhurst, D.; Begley, P.; Zelena, E.; Francis-McIntyre, S.; Anderson, N.; Brown, M.; Knowles, J. D.; Halsall, A.; Haselden, J. N.; Nicholls, A. W.; Wilson, I. D.; Kell, D. B.; Goodacre, R. *Nat. Protoc.* **2011**, *6* (7), 1060−1083.

(3) Lemmon, E. W.; McLinden, M. O.; Friend, D. G.; Linstrom, P. J.; Mallard, W. G. *NIST Chemistry WebBook*National Institute of Standards and Technology: Gaithersburg, MD, 2011.

(4) Hummel, J.; Selbig, J.; Walther, D.; Kopka, J. In *Metabolomics*; Springer, 2007; pp 75−95.

(5) Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-Arendt, K.; Ahn, N. G.; Old, W. M. *J. Proteome Res.* **2010**, *9* (8), 4152−4160.

(6) Hoopmann, M. R.; Finney, G. L.; MacCoss, M. J. *Anal. Chem.* **2007**, *79* (15), 5620−5632.

(7) Luethy, R.; Kessner, D. E.; Katz, J. E.; MacLean, B.; Grothe, R.; Kani, K.; Faça, V.; Pitteri, S.; Hanash, S.; Agus, D. B.; Mallick, P. *J. Proteome Res.* **2008**, *7* (9), 4031−4039.

(8) Li, G. Z.; Vissers, J. P. C.; Silva, J. C.; Golick, D.; Gorenstein, M. V.; Geromanos, S. J. *Proteomics* **2009**, *9* (6), 1696−1719.

(9) Gillet, L. C.; Navarro, P.; Tate, S.; Rost, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R. *Mol. Cell. Proteomics* **2012**, *11* (6), O111.016717.

(10) Röst, H. L.; Rosenberger, G.; Navarro, P.; Gillet, L.; Miladinović, S. M.; Schubert, O. T.; Wolski, W.; Collins, B. C.; Malmström, J.; Malmström, L.; Aebersold, R. *Nat. Biotechnol.* **2014**, *32* (3), 219−223.

(11) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. *Nat. Methods* **2015**, *12* (6), 523−526.

(12) Nikolskiy, I.; Mahieu, N. G.; Chen, Y. J.; Tautenhahn, R.; Patti, G. J. *Anal. Chem.* **2013**, *85* (16), 7713−7719.

(13) Beisken, S.; Earll, M.; Baxter, C.; Portwood, D.; Ament, Z.; Kende, A.; Hodgman, C.; Seymour, G.; Smith, R.; Fraser, P.; Seymour, M.; Salek, R. M.; Steinbeck, C. *Sci. Data* **2014**, *1*, 140029.

(14) Longnecker, K.; Kido Soule, M. C.; Kujawinski, E. B. *Mar. Chem.* **2015**, *168*, 114−123.

(15) Chen, L.; Li, J.; Guo, T.; Ghosh, S.; Koh, S. K.; Tian, D.; Zhang, L.; Jia, D.; Beuerman, R. W.; Aebersold, R.; Chan, E. C. Y.; Zhou, L. *J. Proteome Res.* **2015**, *14* (9), 3982−3995.

(16) Michalski, A.; Cox, J.; Mann, M. *J. Proteome Res.* **2011**, *10* (4), 1785−1793.

(17) Tyanova, S.; Temu, T.; Carlson, A.; Sinitcyn, P.; Mann, M.; Cox, J. *Proteomics* **2015**, *15* (8), 1453−1456.

(18) Haug, K.; Salek, R. M.; Conesa, P.; Hastings, J.; de Matos, P.; Rijnbeek, M.; Mahendraker, T.; Williams, M.; Neumann, S.; Rocca-Serra, P.; Maguire, E.; González-Beltrán, A.; Sansone, S.-A.; Griffin, J. L.; Steinbeck, C. *Nucleic Acids Res.* **2013**, *41* (Database issue), D781−D786.

(19) Sud, M.; Fahy, E.; Cotter, D.; Azam, K.; Vadivelu, I.; Burant, C.; Edison, A.; Fiehn, O.; Higashi, R.; Nair, K. S.; Sumner, S.; Subramaniam, S. *Nucleic Acids Res.* **2016**, *44*, D463−D470.

(20) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78* (3), 779−787.

(21) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Römpp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.; Souda, P.; Hermjakob, H.; Binz, P.-A.; Deutsch, E. W. *Mol. Cell. Proteomics* **2011**, *10* (1), R110.000133.

(22) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. a; Brusniak, M.-Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. *Nat. Biotechnol.* **2012**, *30* (10), 918−920.

(23) He, L.; Diedrich, J. K.; Chu, Y.-Y.; Yates, J. R. *Anal. Chem.* **2015**, *87*, 11361−11367.

(24) Neumann, S.; Thum, A.; Böttcher, C. *Metabolomics* **2013**, *9*, 84−91.

(25) Chitraju, C.; Trotzmuller, M.; Hartler, J.; Wolinski, H.; Thallinger, G. G.; Lass, a.; Zechner, R.; Zimmermann, R.; Kofeler, H. C.; Spener, F. *J. Lipid Res.* **2012**, *53* (10), 2141−2152.

(26) Fiore, C. L.; Longnecker, K.; Kido Soule, M. C.; Kujawinski, E. B. *Environ. Microbiol.* **2015**, *17*, 3949.

DOI: 10.1021/acs.analchem.6b04358
Anal. Chem. 2017, 89, 2432−2439