Article

One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks

Owen D. Myers, Susan J. Sumner, Shuzhao Li, Stephen Barnes, and Xiuxia Du

Anal. Chem., **Just Accepted Manuscript** • DOI: 10.1021/acs.analchem.7b00947 • Publication Date (Web): 28 Jul 2017

One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks

Owen D. Myers,[†] Susan J. Sumner,[‡] Shuzhao Li,[¶] Stephen Barnes,[§] and Xiuxia Du[*,†]

[†]*University of North Carolina at Charlotte, Charlotte, NC 28223, USA*

[‡]*University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA*

[¶]*Emory University, Atlanta, GA 30322, USA*

[§]*University of Alabama at Birmingham, Birmingham, AL 35294, USA*

E-mail: xiuxia.du@uncc.edu

Phone: (704) 687-7307

9201 University City Blvd., Charlotte, NC 28223, USA

Abstract

False positive and false negative peaks detected from extracted ion chromatograms (EIC) are an urgent problem with existing software packages that preprocess untargeted liquid or gas chromatography-mass spectrometry metabolomics data because they can translate downstream into spurious or missing compound identifications. We have developed new algorithms that carry out the sequential construction of EICs and detection of EIC peaks. We compare the new algorithms to two popular software packages XCMS and MZmine 2, and present evidence that these new algorithms detect significantly fewer false positives. Regarding the detection of compounds known to be present in the data, the new algorithms perform at least as well as XCMS and MZmine 2. Furthermore, we present evidence that mass tolerance in m/z should be favored rather than mass tolerance in ppm in the process of constructing EICs. The mass tolerance parameter plays a critical role in the EIC construction process and can have immense impact on the detection of EIC peaks.

Introduction

Mass spectrometry (MS) coupled to liquid chromatography (LC) or gas chromatography (GC) have become indispensable analytical platforms for untargeted metabolomics.[1–6] With sensitivity, chromatographic resolution, and mass measurement accuracy continuously improving, more and more analytes are being detected.[7] As a result, raw untargeted LC/MS and GC/MS data have become more complex and data preprocessing has become more challenging. For both LC/MS-based and GC/MS-based untargeted metabolomics, data preprocessing extracts qualitative and quantitative information of metabolites from raw mass spectrometry data by carrying out a sequence of computational tasks.[8,9]

Construction of EICs and detection of chromatographic peaks from the EICs are the first two tasks in the preprocessing workflow and are critical for the success of the entire metabolomics study. This is because both steps are essential for the identification and relative quantitation of analytes. Also, errors in these two steps propagate not only throughout the entire data preprocessing process, but affect subsequent statistical analysis and metabolic pathway analysis as well.

Though many software packages, both commercial[10–16] and open source,[17–31] have been developed to perform these critical steps of the metatabolomics analysis workflow, XCMS[32–34] and MZmine 2[9,35] have become arguably the most widely used free software tools for preprocessing untargeted metabolomics data. Specifically, the *centWave* algorithm originally developed in XCMS has now become the primary chromatographic peak detection method for XCMS and MZmine 2. However, Coble et. al.[36] compared the performance of MetAlign, XCMS, and MZmine 2 and concluded that far too many false positives are being reported. Through our own use of XCMS and MZmine 2 we have also noticed considerable numbers of false positive EIC peaks and see this as an area of analysis that requires serious attention.

In this paper we present new EIC construction and chromatographic peak detection algorithms that we have developed as part of the Automated Data Analysis Pipeline (ADAP) workflow[8,37,38] for both GC/MS and LC/MS data preprocessing. We show evidence that these ADAP algorithms are able to detect peaks of known compounds at least as well as XCMS and MZmine 2 while

comparatively reducing the detection of false EIC peaks by a substantial margin. ADAP's EIC construction algorithm works from the most intense data points down to the least, taking advantage of the higher accuracy in mass-to-charge ratio, m/z, of the more intense data points. ADAP's peak detection algorithm addresses some of the inadequate filtering methods in *centWave* that result in the detection of false peaks in EICs by introducing a new method of signal-to-noise ratio estimation as well as several other filtering steps.

The new ADAP algorithms for constructing EICs and detecting EIC peaks have been implemented in Java and incorporated into the MZmine 2 framework.[39] Henceforth when we refer to "ADAP" we are referring to the sequential use of the EIC construction and EIC peak detection modules in MZmine 2. ADAP is open source, freely available, and combined with the preexisting strengths of MZmine 2. All of the beneficial aspects of MZmine 2 such as modularity, visualization, portability, and flexibility are available when using ADAP.

While developing the new algorithm for constructing EICs, we asked what unit of mass tolerance should be preferred: m/z or ppm? We have found that mass tolerance in m/z should be used, despite the fact that mass tolerance in ppm is often used in many software packages. We will present evidence for favoring mass tolerance in m/z.

We have made the following materials available online so that readers may test ADAP and scrutinize the results reported herein: 1) raw data that was used for developing and testing ADAP; 2) MZmine 2 package that includes ADAP (for details see Section "ADAP in MZmine 2" of the Supporting Information); 3) images of thousands of EIC peaks detected by XCMS, MZmine 2, and ADAP; and 4) modified XCMS source code to allow it to use mass tolerance in m/z for EIC construction. All these materials can be found at `http://www.du-lab.org/publications.html`.

Experimental Section

Here we give a brief overview of the data files used in this paper. For specific details on each data file and how it was produced see Section "Experimental Procedures" of the Supporting Information. We work with five separate data files and give them the following abbreviated names: DCSM, YP01, YP02, VT001, and MAR17. These data files are produced on LC/MS analytical platforms.

DCSM, YP01, YP02, VT001 are used for our primary results and all contain some known compounds. DCSM is a standard mixture file that was generated from a mixture of 21 standard compounds. Section "Compounds Manually Confirmed in the DCSM File" of the Supporting information lists these compounds. YP01, YP02, and VT001 were all generated from NIST Standard Reference Material (SRM) 1950, a representation of human plasma.[40] We show the compounds manually confirmed in each data file in Section "Compounds Manually Confirmed in the YP01, YP02, and VT001 Files" of the Supporting Information. These manually confirmed compounds are used to benchmark the false negative rate of EIC peak detection by XCMS, MZmine 2, and ADAP. The MAR17 data is also generated from a blood plasma sample but comes from study ST000045 in the Metabolomics Workbench.[41] This data file is used to investigate what unit of mass tolerance should be used in the construction of EICs: m/z or ppm.

Results and Discussion

Principles of EIC Construction and EIC Peak Detection by ADAP. For EIC construction, ADAP works from the largest intensity data point in the entire data file down to the smallest. The reasoning is that it is beneficial to begin EICs with larger intensity data points because their m/z values tend to be more accurate. In this way the ADAP EIC construction algorithm is similar to that proposed by Kenar et al.[42] Other than beginning with the most intense peaks, the rest of the two algorithms are very different. For detecting peaks in these EICs, ADAP uses the continuous wavelet transform (CWT) and ridgeline detection approach (to be described in detail), just as what *centWave* does but with some very important differences in aspects such as filtering by ridgeline

length. Also, to selectively eliminate false peaks which can result from imperfections of the CWT method we implement several peak filtering steps including a simpler signal-to-noise estimation compared to that used by *centWave*.

Construction of EICs by ADAP. We define $\epsilon_{m/z}$ to be the mass tolerance parameter m/z. Figure 1 shows a simplified flow diagram of the ADAP EIC construction process. This process consists of the following sequential steps:

(1) Take *all* the data points in a data file, sort them by their intensities, and remove those points (mostly noise) below a certain intensity threshold.

(2) Starting with the most intense data point, the first EIC is created.

(3) For this EIC, establish an immutable m/z range that is the data point's $m/z \pm \epsilon_{m/z}$ where $\epsilon_{m/z}$ is specified by the user.

(4) The next data point, which will be the next most intense, is added to an existing EIC if its m/z value falls within its m/z range.

(5) If the next data point does not fall within an EICs m/z range, a new EIC is created. New EICs are only created if the point meets the minimum start intensity requirement set by the user.

(6) An m/z range for a new EIC is created the same way as in step (3) except the boundaries will be adjusted to avoid overlapping with pre-existing EICs. As an example consider an existing EIC with m/z range $(100.000, 100.020)$ for $\epsilon_{m/z} = 0.01$. If the new EIC is initialized with a data point having an m/z value of 100.025, then this new EIC will have a m/z range set to $(100.020, 100.035)$ rather than $(100.015, 100.035)$.

(7) Repeat steps (4)-(6) until all the data has been processed.

(8) Finally, a post processing step is implemented. Only EICs with a user defined number of continuous points above a user defined intensity threshold are kept.

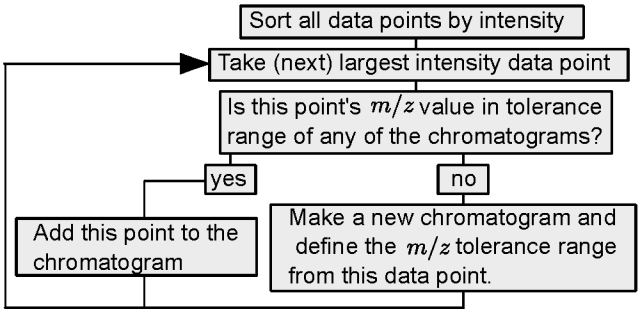


Figure 1: Simplified flow diagram of the ADAP EIC construction process.

Detection of EIC Peaks by ADAP. ADAP detects EIC peaks by using a CWT method. Wavelet coefficients are first calculated as the inner product between the EIC and wavelets at different scales and locations.[33,43] Subsequently, peak location and boundaries are determined through a ridgeline detection (to be described next).

Before we describe the details of the peak detection, it is worth noting that ADAP does not perform any type of baseline correction, because we have found that imperfections in baseline correction methods can produce convincing false positive peaks. In Section "Example of Baseline Removal Creating False Positives" of the Supporting information, we show an example of how large asymmetric chemical noise peaks can be altered by a baseline correction to appear more like real peaks. Therefore, we believe it is best to not use any such correction in order to ensure the smallest possible number of false positives.

In terms of the implementation of peak detection, the process of calculating wavelet coefficients in ADAP was created through the alteration of *msConvert* (part of ProteoWizard)[44,45] source code in C++. *msConvert* does not have a ridgeline detection module. *centWave*, created through adapting the R package MassSpecWavelet,[43,46] has the ridgeline detection module, however *centWave* is primarily written in R and ADAP is primarily in Java. In order to avoid the need to interface with R and the possible problems that can arise from a Java/R interface such as connection failures with the virtual R server (Rserve) that MZmine 2 uses, we have developed a separate ridgeline detection method for ADAP in Java. Due to specific details that are unique to ADAP, we describe the ADAP ridgeline detection next.

Ridgeline Detection. A real peak in an EIC should create a local maxima in the wavelet coefficients at multiple scales. The wavelet scale for which the wavelet most closely matches the shape of the peak, the best scale, will create the largest coefficient. Scales close to the best scale should also have reasonably similar shapes to the peak and therefore create adjacent maxima between those scales. Ridgelines are the series of connected local maxima of the wavelet coefficients across scales which are indicative of a real peak.

Our procedure for detecting the ridgelines is similar to that described by Du et al.[43] and Wee et al.[47] and is as follows.

(1) Begin with the coefficients corresponding to the largest wavelet scale. The largest wavelet scale is currently hard coded to be 10.

(2) Find the (next) largest coefficient at this scale and initialize a ridgeline. Then remove all coefficients that are within half the estimated compact support of the Ricker wavelet ($2.5\times$ the current scale).

(3) Repeat step (2) until there are no more coefficients remaining for this wavelet scale.

(4) Move to the next scale (decrease by one) and repeat (1)-(4). Add new coefficients to an existing ridgeline if they are close in RT. We define close to be a difference in their indices that is less than or equal to the current scale being investigated. The smallest wavelet scale is hard coded to be 1.

(5) After all scales have been processed, ridgelines must have a length, i.e., the total number of scales represented in the ridgeline, greater than or equal to 7, and not more than 2 gaps (missing scales) total.

The removal of local maxima (For examples, please refer to Section "ADAP Determining Local Maxima in Wavelet Coefficients" of the Supporting information) nearby a given coefficient (step (2) above) can potentially create maxima at the boundary of the removed data and the remaining data. An artificial maxima could be created in a similar way during the processing of the next scale, likely being close to the first maxima. Such a scenario potentially creates artificial ridgelines.

However, we assume that even if such a ridgeline is created it will not meet any of the other required criteria, described below, in the peak picking process and there is therefore little danger of producing false positives. The benefit of such a method lies in valuing the largest coefficients, which most probably are present because of real EIC peaks.

Determination of Peak Location and Boundaries. The location of the peak is taken to be the RT of the largest coefficient in the ridgeline. The left (right) boundary of the peak is taken to be the RT of the peak minus (plus) the best scale multiplied by the time between scans.

Peak boundaries should be close to local minima. However, the boundaries determined above often do not coincide with the local minima. We correct the boundaries to the first local minima on both sides of the EIC.

Noise Estimation and Signal-to-Noise Ratio Calculation. We estimate the signal-to-noise ratio, S/N, for each peak detected by the CWT. The user may then specify a S/N threshold that can filter many of the false positives found from the CWT method. This can only be done if a reasonable estimate of the noise is made. This is an important step because a poor estimate of the noise could prevent false peaks from being filtered and filter out real peaks.

To calculate S/N, we choose S to be the maximum intensity between the boundaries of the peak under investigation. Noise, N, is estimated using two different methods. The final estimate of N is the smaller value found from the two methods and is used to calculate S/N. Each estimation of the noise attempts to avoid overestimate from the accidental inclusion of other real peaks that may be close in RT.

Method 1:

(1) Set two windows, one on each side of the peak in the EIC. The windows begin at the left and right peak boundaries and end at the peak boundaries $\pm(2 \times PW)$, respectively, where PW stands for peak width and is defined to be the number of scans between the two boundaries of a peak.

(2) Combine the two windows, calculate the standard deviation of the intensities, and store it as one possible value of the noise.

(3) Expand both windows out from the peak by a single scan. The boundaries closest to the peak remain the same. After the first expansion, each window has a length of $2 \times PW + 1$.

(4) Combine the two windows and calculate and store the standard deviation of the intensities.

(5) Repeat steps (3)-(4) until each window has a length of $8 \times PW$.

(6) Incrementally shrink each window by one scan, combine the two windows, and calculate and store the standard deviation. The windows are shrunk by moving the boundary closest to the peak toward the boundary furthest from it.

(7) Repeat step (6) until the window size is $2 \times PW$.

(8) The final noise estimate is taken to be the smallest stored standard deviation.

Method 2:

(1) Same as (1) in method 1.

(2) Same as (2) in method 1.

(3) Shift each entire window away from the EIC peak by one scan; the window lengths do not change.

(4) Repeat steps (2)-(3) until each window's boundary furthest from the feature is $8 \times PW$ from the closest boundary of that feature.

(5) The final noise is taken to be the smallest stored standard deviation.

Figure 2 shows a cartoon of the two different window size variations.

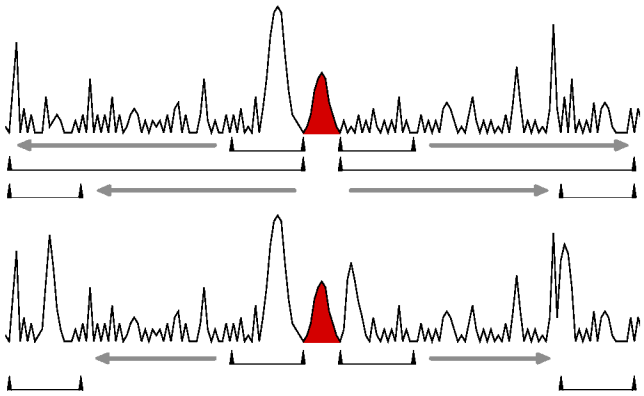


Figure 2: Cartoon of the two window size variations used to estimate EIC noise by ADAP. Top panel illustrates method 1. Bottom panel illustrates method 2. The black carrots represent the boundaries of the windows at the beginning and end stages of the window size changes. The grey arrows show the direction in which either one or both of the boundaries is (are) changed one scan at a time.

Wavelet Coefficient Filter. The magnitude of the wavelet coefficient alone is not sufficient for determining whether or not a peak is real because it has a strong dependence on the intensities of the data points the inner product is being taken with. To ensure that low intensity peaks can be reliably detected and that poorly shaped peaks can be reliably filtered out, we take the largest coefficient, $C_{\max}$, found for a given peak and divide it by the area, A, under the curve between the two boundaries of the peak. We calculate the area using a trapezoidal method so that A is exactly the area under the curve created by connecting adjacent data points with straight lines. The result is a measure for which large values correspond to peaks similar in shape to the wavelet. Figure 3 shows several example peaks in the YP01 data file and lists their corresponding $C_{\max}/A$. One important property of $C_{\max}/A$, which can be seen by comparing the values in Figure 3 (A) and (B), is that intermittent dips in the intensity can increase the value due to the reduced area. This is beneficial for finding messy low intensity peaks but can also be problematic if the area is so small it results in the detection of a peak with a very bad shape.

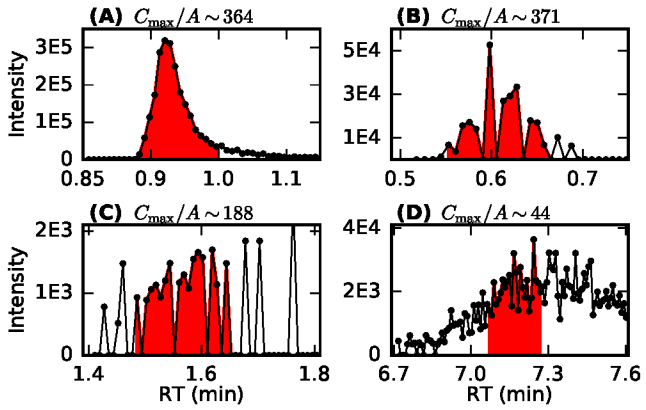


Figure 3: Examples of EIC peaks in data file YP01 and the corresponding $C_{\max}/A$ (maximum wavelet coefficient divided by the area of the peak colored red). Each peak's representative m/z value is (A) 106.049, (B) 659.450, (C) 297.2417, D) 288.133.

Peak Property Filters. For each EIC, the initializing point is the highest intensity among all points in the EIC and has satisfied the minimum intensity condition (step 1 of the EIC construction process). This minimum intensity threshold is for filtering out noise and should be set at a relatively low value to avoid missing real peaks with low intensity. It is worth mentioning that even if this intensity threshold is quite low compared to the highest intensity in the entire data file, generally more than 50% of the data points could be filtered, which improves the speed of EIC construction considerably.

After peaks have been detected through CWT and ridgeline detection, it could be useful to discount low intensity peaks or keep only the highest intensity peaks. ADAP includes a second intensity threshold to allow users to make a choice pertaining to the heights of peaks after EIC peak detection. Peaks with heights less than this threshold will be discarded.

It is not uncommon to see EIC peaks that contain zero intensity points because of missing mass values in some scans. We assume these zero intensity points are missed because of the instrument and, in general, would like to detect peaks with missing points as long as their overall profile suggests that they correspond to real compounds. Figure 3B shows one example. ADAP considers these peaks to be real as long as the number of non-zero intensity data points is greater than the number of zero-intensity data points. We consider the ability of ADAP to detect these peaks a

strength, and it is likely that such peaks will be missed by XCMS and MZmine 2. An important note, the area in the final results is calculated beneath the curve created by connecting adjacent data points with straight lines (trapezoidal integration).

m/z vs. ppm as mass tolerance for EIC construction. Does the choice between m/z and ppm as the unit for mass tolerance substantially influence the final peak detection results? This is an important question because the mass tolerance parameter plays such an important role in the EIC construction process.

In order to answer this question, EIC peaks that have been detected by ADAP from the centroided MAR17 file were studied. To ensure the smallest number of false positive peaks, very strict parameters (as shown in Section "Parameters used by ADAP for Constructing EICs and Detecting EIC Peaks from File MAR17" of the Supporting information) were used for constructing EICs and detecting EIC peaks. Eventually, a total of 441 EIC peaks were selected.

For each of the 441 EIC peaks, the mass range in m/z of the data points that form the peak was found. Furthermore, the representative m/z of the EIC peak is determined. This representative m/z is chosen to be the initializing m/z of the entire EIC though other representative values, such as the average m/z would be valid as well. Three examples of EIC peaks and the m/z values of data points along each EIC peak are shown in Figure 4. The range of possible representative m/z values in the 441 EIC peaks spans $[58.0, 1152.0]$, and the mass ranges of the 441 EIC peaks spans $[0, 0.039]$.

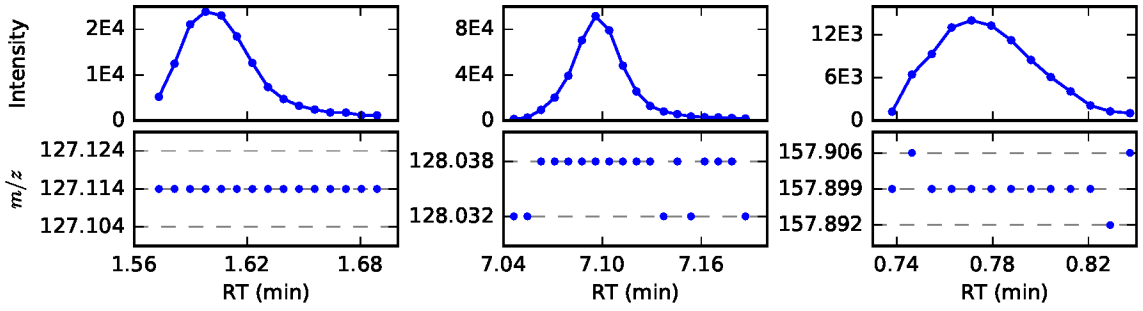


Figure 4: Examples of EIC peaks with centroid m/z values spanning different ranges in the m/z domain. (Left) Centroid m/z values are exactly the same across consecutive scans. (Middle) Centroid m/z values across consecutive scans span one m/z sampling interval. (Right) Centroid m/z values across consecutive scans span two m/z sampling intervals.

Based on the representative m/z values and the mass ranges in m/z of the EIC peaks, the corresponding mass ranges in ppm of the EIC peaks were calculated as

$$\text{mass range in } ppm = \frac{\text{mass range in } m/z}{\text{representative } m/z} \times 10^6 \quad (1)$$

In Section "Comparing ppm and m/z" of the Supporting information we show examples of the calculation using the example EICs in Figure 4. Figure 5 shows range$_{m/z}$ vs. the representative m/z in blue, and range$_{ppm}$ vs. the representative m/z in red as a scatter plot for all of the 441 EIC peaks. Each point corresponds to one EIC peak. An important aspect of Figure 5 is that a mass range in m/z of 0.02 could ensure that most of the EIC peaks would include the majority of the data points forming each peak, whereas a mass range in ppm needs to be about $100ppm$ to achieve the same goal. However, such a large ppm value will almost certainly cause the issue of merging two or more EICs for large masses. On the other hand, a much smaller ppm tolerance will almost certainly cause the issue of splitting two or more EICs for small masses. Therefore, we believe that a mass range in m/z is more appropriate for the construction of EICs if one single mass tolerance is to be used for all of the m/z values in a data file.

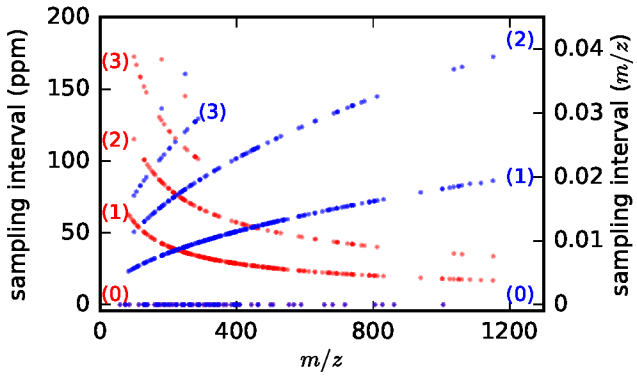


Figure 5: Relationship between mass range and representative m/z for 441 EIC peaks in data file MAR17. The relationship between mass range in m/z and representative m/z is shown in blue (right vertical axis). The relationship between mass range in ppm and representative m/z is shown in red (left vertical axis). Each point in the figure corresponds to one EIC peak. Points along curve (0) correspond to EIC peaks where centroid m/z values remain the same across consecutive scans for a given peak and therefore both the mass range in m/z and mass range in ppm are 0. As a result of the overlapping red and blue, curve 0 appears as purple. Figure 4 (left) shows one example of such an EIC peak. EIC peaks represented by points along curve (1) have a mass range of one m/z sampling interval in the profile mass spectra and Figure 4 (middle) shows one such example. Points along curve (2) have a mass range of two m/z sampling intervals and Figure 4 (right) shows one such example.

In order to understand why there are several distinct red and blue curves in Figure 5 and why the mass range in m/z increases with the representative m/z values whereas the mass range in ppm decreases with the representative m/z, we need to understand the varying sampling intervals in a single profile mass spectrum and the process of spectrum centroiding. In a profile mass spectrum, the magnitude of the sampling interval generally increases with m/z. As a result, the larger the m/z, the larger the sampling interval. Figure 6(A) shows the sampling interval vs. the centroid m/z for mass spectral peaks in scan number 200 of the MAR17 file and Figure 6(B) shows the sampling points for two specific mass spectral peaks in this scan. The sampling interval for the blue mass peak with the centroid m/z around 146.95 is 0.007 m/z, while the sampling interval for the red mass peak with the centroid m/z around 957.95 is about 0.018 m/z. The file MAR17 was generated on an Agilent Q-TOF MS (Agilent 6220).[48] Similar relationships have been observed for data generated on Waters[49] and ThermoFisher[50] mass spectrometers and we show this relationship for two other data files in Section "Comparing ppm and m/z" of the Supporting information.

To centroid a profile mass spectrum, local maximum or wavelet transform methods are usually used (Both *msConvert* and MZmine 2 include these two methods) though other approaches exist. When the local maximum or wavelet transform is used, the resulting centroid m/z is one of the sampling m/z values. That is to say, the centroid m/z is latched to an existing m/z value in the profile mass spectrum, rather than being somewhere between the sampling values. Figure 6(B) shows the centroid m/z values indicated by stars when local maximum method is used for centroiding.

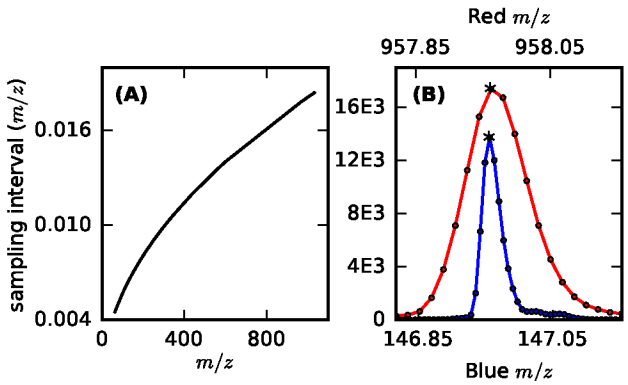


Figure 6: (A) Relationship between m/z sampling intervals and centroid m/z values for mass spectral peaks in the profile mass spectrum scan 200 in the MAR17 data file. The sampling interval increases with m/z values. The same relationship holds for all other profile mass spectra in this data file. (B) Sampling points and sampling intervals for two specific mass spectral peaks in the profile mass spectrum scan 200. One spectral peak is in blue and the other is in red. Black dots along the two curves are the sampling points. The m/z axis for the blue curve is at the bottom and the m/z axis for the red curve is at the top. Both m/z axes are scaled the same. Stars indicate the centroid m/z when local maximum method is used for centroiding the blue and the red mass spectral peaks. The sampling interval for the blue mass peak is ~ 0.007 wide in the m/z domain whereas the sampling interval for the red mass peak is ~ 0.018.

When an EIC is constructed from centroid m/z values, the centroid m/z values of the data points that form the EIC peak could be exactly the same across consecutive scans regardless of the representative m/z value of the EIC. Alternatively they could span one, two, or more sampling intervals in the m/z domain. Figure 4 shows examples of EICs that span zero (left), one (middle), or two (right) m/z sampling intervals. This explains why there are several distinct red and blue curves (formed by dots) in Figure 5. Specifically, points along blue curve 0, 1, 2, or 3 represent EIC peaks whose data points span 0, 1, 2, or 3 sampling intervals in the m/z dimension. The reason that

the mass range in m/z for EIC peaks along each blue curve increases with the representative m/z is due to the increasing sampling intervals in each profile mass spectrum as shown in Figure 6(A). In fact, the curve in Figure 6(A) is exactly the blue curve 1 in Figure 5.

For each EIC peak, the mass range in ppm is calculated as the ratio of the mass range in m/z to the representative m/z as shown in Equation (1). Even though the numerator slightly increases with the representative m/z as depicted in Figure 6(A), the denominator increases much faster. This causes the mass range in ppm to decrease relatively rapidly when the representative m/z values increase. The maximum mass range in ppm for the 441 EIC peaks reaches almost 200 ppm for peaks with small representative m/z values. From a compound identification point of view, a mass error as small as 10 ppm could mean completely different compounds, let alone 200 ppm. Therefore, it is worth reiterating that mass tolerance in m/z should be favored over ppm.

Modified XCMS to use mass tolerance in m/z. Due to the aforementioned problems associated with using mass tolerance in ppm for EIC construction, we exclusively set the mass tolerance in m/z for the work presented herein. In order to appropriately compare the results found using ADAP and those found using XCMS, which only uses ppm, we modified XCMS to use mass tolerance in m/z. This modified version of XCMS can be found at `http://www.du-lab.org/publications.html` so that readers may test it on their own.

Running and comparing EIC construction and EIC peak picking on data. We have run EIC construction and EIC peak detection on four data files DCSM, YP01, YP02, and VT001 using XCMS, MZmine 2, and ADAP (all processing parameters can be found in Section "Preprocessing Parameters used by ADAP, XCMS, and MZmine 2" of the Supporting information). Each software package detects the majority of the monoisotopic peaks of the compounds manually identified to be present in the data. Due to the similarity in the detection of these compounds we push the details about the compounds and their detection rate by each package to the Supporting Information (sections "Summary of Compound Detection Results" through "Compounds Manually Confirmed in the YP01, YP02, and VT001 Files"). Figure 7 shows Venn diagrams with sections representing the number of peaks found by all three, overlap between two, and found by each individual

software package only. Peaks are compared using a 2-dimensional window of 0.01 m/z unit and 1.5 seconds. Peaks detected by all three packages are more likely real peaks than peaks that have been detected by one software package only. Therefore, we first examine the lobes of the diagrams where EIC peaks have been detected by one software package only for possible false positive peaks.

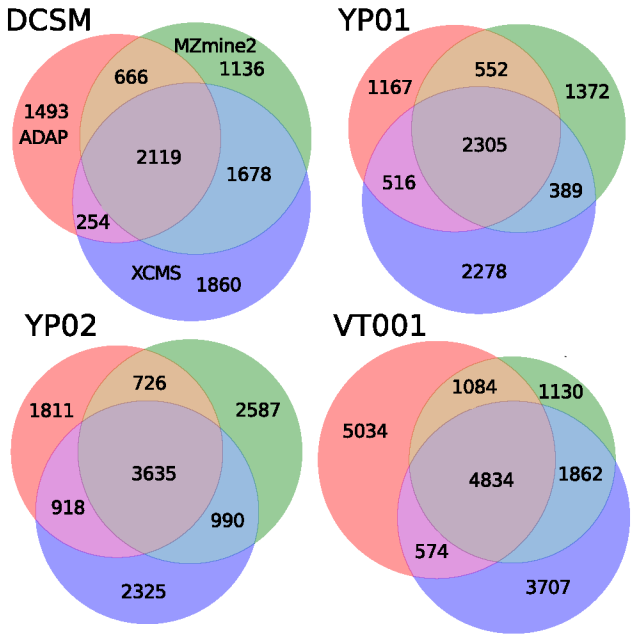


Figure 7: Venn diagrams showing the number of peaks found using ADAP, XCMS, and MZmine 2. Determination of overlap in the Venn diagram was done by checking the proximity of peaks in retention time (within 1.5 seconds) and m/z (within 0.01 unit).

We randomly sample peaks from each lobe of the Venn diagram and visually inspect each of the samples. Using several criteria we count the number of "good" peaks in each sample. For details on the criteria used to determine the good peaks, and of the random sampling, please see section "Details Regarding Random Sampling and Sorting of Detected Peaks" of the Supporting information. The count of good peaks in the random sample gives us an estimate of the proportion of good peaks to false positives in each lobe. We use the Clopper-Pearson method[51] to determine the 95% confidence interval (CI) for each estimate. The results are summarized in Table 1 from which we can see that the proportion of good peaks in the ADAP only lobe are considerably higher than the proportions in the XCMS and MZmine 2 only lobes. This demonstrates that the rate of

false positive EIC peaks detected by ADAP is much lower than XCMS and MZmine 2.

Table 1: (*top number*) Proportion, shown in red, of peaks in the ADAP, XCMS, or MZmine 2-only lobe of Figure 7 that are in the good category. (*bottom range*) 95% confidence interval, shown in black, of the proportion.

Data File	ADAP (%)	XCMS (%)	MZmine 2 (%)
DCSM	94.5 91.8-96.5	16.5 13.0-20.5	43.3 38.3-48.3
YP01	67.3 62.4-71.8	18.0 14.4-22.1	6.5 4.3-9.4
YP02	52.3 47.2-57.2	36.8 32.0-41.7	3.0 1.6-5.2
VT001	46.8 41.8-51.8	3.5 1.9-5.8	2.0 0.9-4.0

Furthermore, we examined EIC peaks in the XCMS and MZmine 2 overlapping region of the Venn diagram in Figure 7 and the overlapping region of all three software packages. Table 2 shows an estimate of the proportion of the peaks that are good in the corresponding portions of the Venn diagram for the YP01 dataset as well as the confidence interval for the proportions. The proportion of EIC peaks that have been detected by both XCMS and MZmine 2 is about 68%. The proportion for the overlap region of all three software packages is higher at about 83%, indicating again that ADAP is able to reduce the false positive rate of EIC peak detection.

Table 2: (*top number*) Proportion of peaks, shown in red, in the overlapping regions of Figure 7 for data file YP01 that are good. (*bottom range*) 95% confidence interval, shown in black, of the proportions.

Data File	XCMS and MZmine 2 Overlap	ADAP, XCMS and MZmine2 Overlap
YP01	67.8 62.9-72.3	83.3 79.2-86.8

Conclusion

Accurate construction of EICs and detection of peaks from EICs are critical for the success of any untargeted, mass spectrometry-based metabolomcis studies. This is because false positive and

false negative EIC peaks can turn into false and missing compound identifications. The high rate of false positive and false negative peak detection by existing software packages motivated us to improve the underlying algorithms of these packages. These efforts produced new algorithms for constructing EICs and detecting EIC peaks. We have compared peak picking of the new algorithms with that of XCMS and MZmine 2 and demonstrated that the percentage of false positive peaks detected by ADAP is significantly lower. In addition, we have shown that the reduction of false positives does not come at the cost of poor sensitivity. The new algorithms we developed have been implemented in Java as part of the ADAP package that consists of a complete workflow for preprocessing raw LC/MS and GC/MS metabolomics data. ADAP has been incorporated into MZmine 2 to take advantage of the latter's strengths including modularity, visualization, and flexibility.

In addition, we investigated which unit of mass tolerance should be favored in the process of constructing EICs: m/z or ppm. Mass tolerance is the most important parameter in this process and has an immense impact on the subsequent peak detection. Our investigation shows that no reasonable mass tolerance in ppm could be chosen that will not cause serious problems of EIC merging or splitting. Therefore, we conclude that mass tolerance in m/z is a better way of setting mass tolerance for EIC construction.

With LC and GC/MS platforms becoming increasingly sensitive, more and more compounds are now detectable in biological samples. The development of automated data preprocessing pipelines that can distinguish real peaks in the data from noise or other artifacts is important for obtaining a clear picture of the role of metabolites in biological processes. Currently no software package performs well enough that the results can be trusted without a significant amount of human oversight and involvement. We hope that the ADAP (Automated Data Analysis Pipeline) peak peaking is one step forward in the direction of reducing false positive and false negative chromatographic peaks and in the direction of fully automated data preprocessing.

Supporting Information Available

The Supporting Information (PDF) contains the following information: Experimental Procedures; ADAP Implementation of EIC Construction; Example Baseline Removal Creating False Positive; ADAP Determining Local Maxima in Wavelet Coefficients; Details Regarding Random Sampling and Sorting of Detected Peaks; Important Details of Venn Diagram Counting; Comparing ppm and m/z; Summary of Compound Detection Results; Compounds Manually Confirmed in the DCSM File; Compounds Manually Confirmed in the YP01, YP02, and VT001 Files; Preprocessing Parameters used by ADAP, XCMS, and MZmine 2; Parameters used by ADAP for Constructing EICs and Detecting EIC Peaks from File MAR17. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

Acknowledgement

We would like to acknowledge: 1) Dr. David A. Horita at the University of North Carolina at Chapel Hill for many insightful discussions; 2) Dr. Aleksandr Smirnov at the University of North Carolina at Charlotte for many discussions and for helping with compound identifications, and 3) Dr. Tomáš Pluskal at the Whitehead Institute for Biomedical Research for his big help with incorporating ADAP into MZmine 2.

References

(1) Scalbert, A.; Brennan, L.; Fiehn, O.; Hankemeier, T.; Kristal, B. S.; van Ommen, B.; Pujos-Guillot, E.; Verheij, E.; Wishart, D.; Wopereis, S. *Metabolomics* **2009**, *5*, 435–458.

(2) Dunn, W. B.; Broadhurst, D. I.; Atherton, H. J.; Goodacre, R.; Griffin, J. L. *Chem Soc Rev* **2011**, *40*, 387–426.

(3) 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.; Human Serum Metabolome, C. *Nat Protoc* **2011**, *6*, 1060–83.

(4) Yin, P.; Xu, G. *J Chromatogr A* **2014**, *1374*, 1–13.

(5) Jorge, T. F.; Rodrigues, J. A.; Caldana, C.; Schmidt, R.; van Dongen, J. T.; Thomas-Oates, J.; Antonio, C. *Mass Spectrom Rev* **2016**, *35*, 620–49.

(6) Fiehn, O. *Curr Protoc Mol Biol* **2016**, *114*, 30 4 1–30 4 32.

(7) Jones, D. P. *Toxicol Rep* **2016**, *3*, 29–45.

(8) Jiang, W.; Qiu, Y.; Ni, Y.; Su, M.; Jia, W.; Du, X. *J Proteome Res* **2010**, *9*, 5974–81.

(9) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinformatics* **2010**, *11*, 395.

(10) http://www.nonlinear.com/progenesis/qi/.

(11) https://www.thermofisher.com/order/catalog/product/IQLAAEGABSFAHSMAPV.

(12) http://www.agilent.com/en-us/products/software-informatics/masshunter-suite/masshunter/masshunter-software.

(13) http://sciex.com/products/software/markerview-software.

(14) http://www.leco.com/products/separation-science/software-accessories/chromatof-software.

(15) http://www.spectralworks.com/products/analyzerpro/.

(16) http://www.prs.no/MS%20Resolver/MS%20Resolver.html.

(17) Lommen, A. *Anal Chem* **2009**, *81*, 3079–86.

(18) Lommen, A. *Methods Mol Biol* **2012**, *860*, 229–53.

(19) Lommen, A.; Kools, H. J. *Metabolomics* **2012**, *8*, 719–726.

(20) Yu, T.; Park, Y.; Johnson, J. M.; Jones, D. P. *Bioinformatics* **2009**, *25*, 1930–6.

(21) Yu, T.; Peng, H. *BMC Bioinformatics* **2010**, *11*, 559.

(22) Yu, T.; Park, Y.; Li, S.; Jones, D. P. *J Proteome Res* **2013**, *12*, 1419–27.

(23) Yu, T.; Jones, D. P. *Bioinformatics* **2014**, *30*, 2941–8.

(24) Clasquin, M. F.; Melamud, E.; Rabinowitz, J. D. *Curr Protoc Bioinformatics* **2012**, *Chapter 14*, Unit14 11.

(25) `http://genomics-pubs.princeton.edu/mzroll/index.php`.

(26) Wei, X.; Sun, W.; Shi, X.; Koo, I.; Wang, B.; Zhang, J.; Yin, X.; Tang, Y.; Bogdanov, B.; Kim, S.; Zhou, Z.; McClain, C.; Zhang, X. *Anal Chem* **2011**, *83*, 7668–75.

(27) Rost, H. L. et al. *Nat Methods* **2016**, *13*, 741–8.

(28) Sturm, M.; Bertsch, A.; Gropl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; Kohlbacher, O. *BMC Bioinformatics* **2008**, *9*, 163.

(29) `http://www.openms.de/`.

(30) Baran, R.; Kochi, H.; Saito, N.; Suematsu, M.; Soga, T.; Nishioka, T.; Robert, M.; Tomita, M. *BMC Bioinformatics* **2006**, *7*, 530.

(31) `http://mathdamp.iab.keio.ac.jp/`.

(32) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal Chem* **2006**, *78*, 779–87.

(33) Tautenhahn, R.; Bottcher, C.; Neumann, S. *BMC Bioinformatics* **2008**, *9*, 504.

(34) Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G. *Anal Chem* **2012**, *84*, 5035–9.

(35) Katajamaa, M.; Miettinen, J.; Oresic, M. *Bioinformatics* **2006**, *22*, 634–6.

(36) Coble, J. B.; Fraga, C. G. *J Chromatogr A* **2014**, *1358*, 155–64.

(37) Ni, Y.; Qiu, Y.; Jiang, W.; Suttlemyre, K.; Su, M.; Zhang, W.; Jia, W.; Du, X. *Anal Chem* **2012**, *84*, 6619–29.

(38) Ni, Y.; Su, M.; Qiu, Y.; Jia, W.; Du, X. *Anal Chem* **2016**, *88*, 8802–11.

(39) http://mzmine.github.io/.

(40) NIST SRM 1950. https://www-s.nist.gov/srmors/certificates/1950.pdf, [Accessed January 30, 2017].

(41) http://www.metabolomicsworkbench.org/.

(42) Kenar, E.; Franken, H.; Forcisi, S.; Wormann, K.; Haring, H. U.; Lehmann, R.; Schmitt-Kopplin, P.; Zell, A.; Kohlbacher, O. *Mol Cell Proteomics* **2014**, *13*, 348–59.

(43) Du, P.; Kibbe, W. A.; Lin, S. M. *Bioinformatics* **2006**, *22*, 2059–65.

(44) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. *Bioinformatics* **2008**, *24*, 2534–6.

(45) http://proteowizard.sourceforge.net/.

(46) https://www.bioconductor.org/packages/release/bioc/html/MassSpecWavelet.html.

(47) Wee, A.; Grayden, D. B.; Zhu, Y.; Petkovic-Duran, K.; Smith, D. *Electrophoresis* **2008**, *29*, 4215–25.

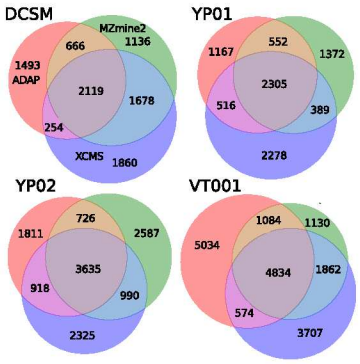
(48) http://www.agilent.com/en-us/products/mass-spectrometry/lc-ms-instruments/6200-series-accurate-mass-time-of-flight-(tof)-lc-ms.

(49) www.waters.com.

(50) http://www.thermofisher.com/.

(51) https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval.

TOC graphic


DCSM
MZmine2
1136
666
1493
ADAP
2119
1678
254
XCMS
1860
YP01
552
1167
1372
2305
516
389
2278
YP02
726
1811
2587
3635
918
990
2325
VT001
1084
1130
5034
4834
1862
574
3707