analytical chemistry

Article

pubs.acs.org/ac

Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data

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

[†]University of North Carolina at Charlotte, Charlotte, North Carolina 28223, United States

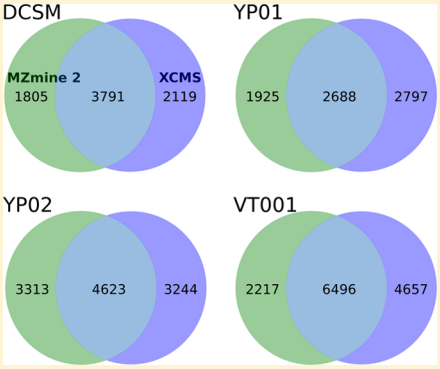
[‡]University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27514, United States

[§]Emory University, Atlanta, Georgia 30322, United States

[∥]University of Alabama at Birmingham, Birmingham, Alabama 35294, United States

Supporting Information

ABSTRACT: XCMS and MZmine 2 are two widely used software packages for preprocessing untargeted LC/MS metabolomics data. Both construct extracted ion chromatograms (EICs) and detect peaks from the EICs, the first two steps in the data preprocessing workflow. While both packages have performed admirably in peak picking, they also detect a problematic number of false positive EIC peaks and can also fail to detect real EIC peaks. The former and latter translate downstream into spurious and missing compounds and present significant limitations with most existing software packages that preprocess untargeted mass spectrometry metabolomics data. We seek to understand the specific reasons why XCMS and MZmine 2 find the false positive EIC peaks that they do and in what ways they fail to detect real compounds. We investigate differences of EIC construction methods in XCMS and MZmine 2 and find several problems in the XCMS *centWave* peak detection algorithm which we show are partly responsible for the false positive and false negative compound identifications. In addition, we find a problem with MZmine 2's use of *centWave*. We hope that a detailed understanding of the XCMS and MZmine 2 algorithms will allow users to work with them more effectively and will also help with future algorithmic development.



Mass spectrometry (MS) coupled to liquid or gas chromatography (LC or GC) are analytical platforms that have been improving quickly in sensitivity, chromatographic resolution, and mass measurement accuracy.[1−6] Such improvements have led to extremely complex data sets, and preprocessing of untargeted matabolomics data have therefore become more challenging.

Current preprocessing pipelines comprise five sequential steps: (1) construction of extracted ion chromatograms (EIC), (2) detection of chromatographic peaks (boundaries and apex) from the EICs, (3) annotation of EIC peaks for LC/MS data by assigning them to molecular ions or adducts or deconvolution of EIC peaks for GC/MS data for coeluting analytes, (4) alignment of analytes across samples, and (5) identification and relative quantitation of analytes. The first two steps are particularly important because errors in these steps propagate not only through the entire data preprocessing steps but affect subsequent statistical analysis and metabolic pathway analysis as well.

EIC construction determines mass-to-charge ratios (m/z) that have been measured repeatedly by MS over a duration of time, while peak detection determines the retention time (RT) and elution profile of analytes. Many commercial and free software packages have been developed to perform these two tasks (see the section Non-Comprehensive List of Software Packages Developed for Preprocessing Metabolomics Data of the Supporting Information for examples). However, no algorithm has been accepted as the benchmark in the metabolomics field, and all existing methods produce different results.

A number of algorithms have been proposed for detecting chromatographic peaks from EICs,[7−9] but the continuous wavelet transform (CWT)-based *centWave* algorithm has become particularly prevalent in the metabolomics community.[10] Furthermore, XCMS (including the command line version[11] and XCMS Online[12]) and MZmine 2 have integrated the *centWave* algorithm into their modular framework, which further increased the usage of *centWave* due to the ease of use of these two open source software packages.

XCMS and MZmine 2 have become arguably the most widely used free software tools for preprocessing untargeted metabolomics data. Given the prevalence of XCMS and MZmine 2, an important question is how do the EIC construction and peak detection results of each package

Received: March 22, 2017
Accepted: July 28, 2017
Published: July 28, 2017

DOI: 10.1021/acs.analchem.7b01069
Anal. Chem. 2017, 89, 8689−8695

compare and how well do they each perform. Coble et al. compared the performance of MetAlign, XCMS, and MZmine and concluded that "there is a pressing need to improve the preprocessing tools to reduce the percentage of false peaks..."[13] Another comparative evaluation of preprocessing packages was carried out by Rafiei et al.[14] They compared the peak picking workflows in three commercial software packages PeakView,[15] Markerview,[16] and MetabolitePilot[17] from the MS vendor Sciex[18] as well as the freeware XCMS Online. They observed vast differences in resulting peak lists after the four packages were employed on identical LC/MS urine and bile samples as well as on a standard metabolite mix. Furthermore, they observed that "there were a number of standard metabolites undetected by all the four workflows."

In the course of using XCMS and MZmine 2 (version MZmine 2.21), we have also noticed a considerable number of false positive peaks and significant differences in peak lists from the two packages. This motivated us to investigate the causes of false positive and false negative peaks. We have found that two key aspects of the *centWave* algorithm are responsible for many of the false positive peaks: signal-to-noise ratio estimation and CWT ridgeline detection. To understand why these portions of the *centWave* algorithm could produce a false positive or false negative peak, we elucidate the details and inner workings of the *centWave* algorithm herein. In addition, we compare peak picking results produced by XCMS and MZmine 2 from four individual data files to see peak picking differences for each of the individual files.

EXPERIMENTAL SECTION

Four data files, named DCSM, YP01, YP02, and VT001, were used to test the performance of XCMS and MZmine 2. All of these files can be found at http://www.du-lab.org/publications.html and detailed information on the experimental procedures can be found in the section Experimental Procedures of the Supporting Information.

RESULTS AND DISCUSSION

Comparison of Peaks Detected by XCMS and MZmine 2.

We used XCMS and MZmine 2 to construct EICs, detect EIC peaks from the four data files, and then compare the lists of EIC peaks. All of the parameters used for the preprocessing steps with each software package can be found in the section Preprocessing Parameters used by XCMS and MZmine 2 of the Supporting Information. Figure 1 shows Venn diagrams of the peaks found by XCMS and MZmine 2. We used a 2-dimensional window of 1.5 s in RT and 0.01 m/z to determine if two peaks were the same and therefore belonged to the overlapping region of the diagram. The substantial size of the individual XCMS and MZmine 2 lobes indicates the large differences found in the peak lists.

We analyzed the number of false positive peaks in different portions of the Venn diagram in the following way. A random sample of 400 peaks, with replacement, is chosen from each lobe of each Venn diagram and visually inspected. We scrutinized each peak shape and the boundaries defining each peak to sort the random sample of peaks into three categories. The first category is peaks that correspond possibly to real compounds, the second category is peaks that clearly do not correspond to real compounds, and the third is peaks that cannot easily be placed into the other two categories. From here on we refer to these three categories as good, bad, and uncertain in that order. More specifically, a peak was considered good if it met the following criteria:

(1) The boundaries of the peak appear to encapsulate the majority of the peak.

(2) The boundaries of the peak encapsulate the maximum of the peak.

(3) The peak is not immediately surrounded by peaks of similar intensity and shape that make the peak itself look like noise.

(4) The peak has a good shape or, if not, strongly meets criteria 3.

We recognize that these criteria and the choices based off of them are subjective. However, we believe the results presented here do not strongly depend on the person performing the categorization and have made all of the images from which the visual inspection was performed available for scrutiny at http://www.du-lab.org/publications.html. We must state that the researcher sorting the peaks was not blind to the analysis methods during the sorting.

From the categorized random samples of peaks found in each lobe of the Venn diagrams, we estimated the proportion of these peaks that are good as well as a 95% confidence interval (CI) for this proportion using the Clopper-Pearson method.[19] Briefly, if n is the number of sampled peaks, r is the number of those peaks which are considered good, and α is the acceptable error (0.05 for a 95% CI), then the upper and lower bounds of the Clopper-Pearson method are most concisely stated as

$$\text{lower bound} = \text{QB}(r, n - r + 1, \alpha/2),$$
$$\text{upper bound} = \text{QB}(r + 1, n - r, 1 - \alpha/2) \quad (1)$$

where $\text{QB}(a, b, q)$ is the q quantile of the beta distribution,

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)} \quad (2)$$

with shape parameters a and b.[20]

The proportion of good peaks and the CI estimated from the random samples are shown in Table 1. From the table, we can see that the majority of the EIC peaks detected by XCMS or MZmine 2 only are false positive peaks. We also present the results of the same analysis performed on the peaks in the overlapping region of the Venn diagram (last column of Table

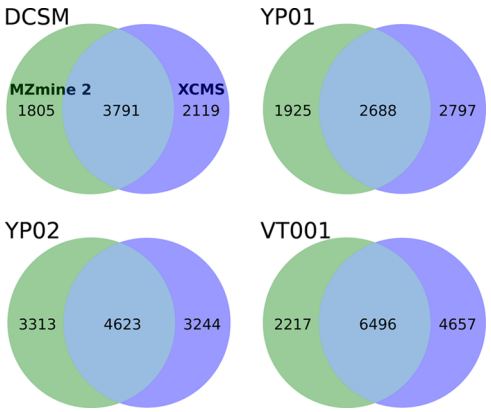


Figure 1. Venn diagrams showing the number of peaks found using XCMS and MZmine 2. Determination of overlap in the Venn diagram was done by checking the proximity of peaks in retention time (within 1.5 s) and m/z (within 0.01).

DOI: 10.1021/acs.analchem.7b01069
Anal. Chem. 2017, 89, 8689−8695

Table 1. (Top Number) Proportion of Peaks in the XCMS-Only Lobe, MZmine 2-Only Lobe, and Overlapping Sections of Figure 1 That Are Considered Good Peaks and (Bottom Range) 95% Confidence Interval of the Proportion

data file	XCMS only	MZmine 2 only	overlap
DCSM	0.20	0.61	0.73
	0.16−0.24	0.56−0.66	0.69−0.78
YP01	0.36	0.25	0.68
	0.31−0.41	0.20−0.29	0.63−0.72
YP02	0.61	0.15	0.71
	0.56−0.65	0.11−0.18	0.66−0.75
VT001	0.13	0.26	0.41
	0.96−0.16	0.22−0.30	0.36−0.46

1). The number of false positives in the overlapping region is significantly better but still quite high.

Figure 2 shows four specific examples of false positive peaks. Panels A and B show peaks detected by XCMS-only and panels

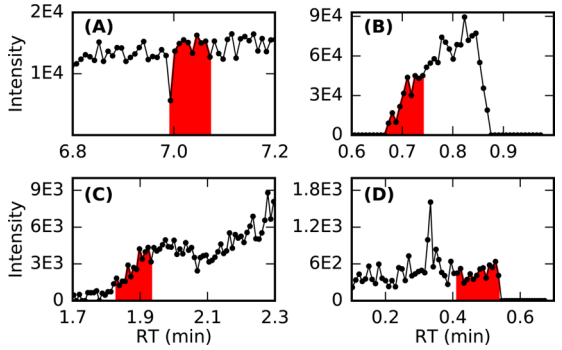


Figure 2. Examples of false positive peaks, shaded in red, detected by XCMS or MZmine 2 from file YP01. (A, B) Examples of false positives found by XCMS-only for m/z being 149.053 and 1043.796, respectively. (C, D) Examples of false positives found by MZmine 2 only for m/z being 498.323 and 223.942, respectively.

C and D show peaks detected by MZmine 2-only. These peaks were found in data file YP01. These examples illustrate common recurring themes in the detection methods. Panels B−D show that the detected boundaries do not encompass the entire peak. Panel A shows another prevalent type of false positive peak found using XCMS and MZmine 2 where portions of flat plateaus are detected as peaks.

In order to understand these false positives and the causes of the large differences between the peaks detected by XCMS and MZmine 2, we need to have a basic understanding of the principles of EIC construction and EIC peak detection in these two software packages.

Construction of EICs by XCMS and MZmine 2. Both LC/MS and GC/MS data are organized as chronological mass spectrometry scans, henceforth simply referred to as scans. Each scan occurs at a RT in the chromatography process. We define s_i to be the ith scan. For this discussion we will assume that all the data is centroided, i.e., the m/z value of each mass spectral peak has been determined.

XCMS EIC Construction. In XCMS, EICs are built chronologically in RT, starting with the first scan and progressing incrementally. Figure 3 shows a flow diagram of the logic that is used to build the EICs. With the figure as a reference, we will describe each step of the building process

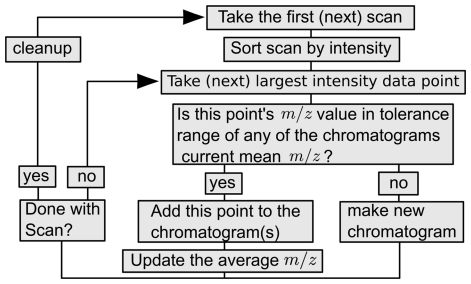


Figure 3. Simplified flow digram of XCMS EIC construction.

through an example, starting with none of the data having been processed.

The first step is to collect the data from the first scan s_1. Next, the data points in the scan are sorted by their intensities from largest to smallest. Then the first, and therefore largest intensity, data point is checked to determine if it is within the user defined m/z tolerance range of an existing EIC. We define the user defined tolerance in ppm to be ϵ_{ppm} and in m/z as $\epsilon_{m/z}$. Since no EICs have yet been initialized, a new EIC is initialized with this data point. Whenever an EIC is initialized, a running average of the EIC's m/z values is begun. The running average is an important aspect of XCMS's EIC building because it serves as the reference point from which it is determined if other data points belong to an EIC.

In our example, there is currently one EIC with a single data point. The average m/z of this EIC is just the m/z value of the one data point. The next largest intensity data point in the scan is then considered. There are only two possibilities. The first possibility (the "yes" case in Figure 3), is that the new data point falls within $\epsilon_{m/z}$ of the EIC's running average m/z and thus belongs in the EIC. After this new data point is added, the average m/z of the EIC is updated and the intensities of the points are summed together. If more points are added to the same EIC from this scan, then all their intensities are summed and they contribute to the average m/z as well.

The second possibility (the "no" case in Figure 3) is that the new data point is a part of a different and unmade EIC. This is the case if its m/z value is not within $\epsilon_{m/z}$ of the preexisting EIC's running average m/z, assuming a reasonable tolerance is chosen. If the data point is outside the tolerance, a new EIC is started.

These above steps are continued until all data points in s_1 have been processed. The following steps occur in the "cleanup" block of Figure 3, before moving on to s_2. The EICs that are long enough but have not had a point added to them during the processing of the current scan are stored. By default, an EIC that is too short is thrown away if it has not had a point added to it during the last processed scan. If an EIC is in the list of completed EICs, it can no longer have points added to it. Next, the intensities of the points in the completed EICs are checked. If the number of points above a minimum intensity (second value in the user-defined prefilter parameter) falls below a certain value (first value in the user-defined prefilter parameter), the EIC will be removed.

Next, XCMS will move to s_2 and repeat the above procedure for the new scan. New points added to EICs will be adjacent to the previous points in the RT domain. For example, if s_2 is being processed and a data point is found to fall within $\epsilon_{m/z}$ of an existing EIC, the data point is then added to the EIC and the EIC now spans two points in RT.

DOI: 10.1021/acs.analchem.7b01069
Anal. Chem. 2017, 89, 8689−8695

MZmine 2 EIC Construction. Similar to XCMS, MZmine 2 builds EICs chronologically in RT. Figure 4 shows the MZmine

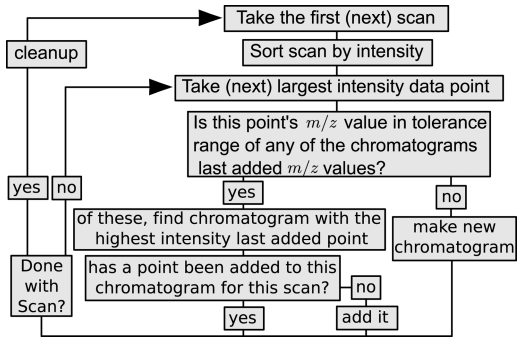


Figure 4. Simplified flow digram of MZmine 2's EIC building.

2 EIC building logic. The first three steps are exactly the same as those for XCMS. We note here that the mass tolerance may be set either in ϵ_{ppm} or $\epsilon_{m/z}$ whereas XCMS only accepts a user-defined ϵ_{ppm} and then uses ϵ_{ppm} to calculate the corresponding $\epsilon_{m/z}$ for each EIC.

The logic begins to differ at the fourth step where instead of checking the current data point's m/z value against existing EIC's average m/z, MZmine 2 checks if the data point falls within ϵ of the point that was last added to a given EIC. If the new point is outside ϵ of each of the current EIC's last-added-points, a new EIC is initialized. If the current point is found to be within ϵ of more than one last-added-point, then it is added to the EIC that has the largest intensity last-added-point. Only one point per scan can be added to any given EIC. If an EIC is selected but has already had a point added to it during the processing of the current scan, the current point is discarded.

After a scan is processed, MZmine 2 also performs several checks which occur in the cleanup block of Figure 4. MZmine 2 keeps all EICs that have had a point added to them during the processing of the current scan. It will also keep EICs that have not had a point added but only if they meet a length requirement set by the user (Min time span parameter). After all scans have been processed, a final post processing step is performed which throws away EICs with intensities that are too small (using the Min height parameter).

Similarities and Differences between XCMS and MZmine 2 EIC Construction. An important similarity between XCMS and MZmine 2 is that both sort each scan by the intensity of the data points and discard short EICs if no points were added to them during the processing of a scan. The largest difference between the XCMS and MZmine 2 EIC construction methods lies in determining if a point will be added to any given EIC. XCMS keeps track of the average m/z over all points and the tolerance range is found using this average as the reference point. The XCMS method allows multiple points from the same scan added to a single EIC and their intensities will be added together. In contrast, the MZmine 2 tolerance range is found using the last point added to the EIC as the reference point.

The second difference between XCMS and MZmine 2 is that one data point can be found in multiple EICs in XCMS. In MZmine 2, a point is added only to the EIC with the last-added-point having the largest intensity. There are differences in the cleanup functions as well. XCMS looks at the intensities after processing every scan whereas MZmine 2 looks at the intensities after all scans have been processed. For additional details of the implementation of the EIC construction procedures for both XCMS and MZmine 2, please see the pseudocode of the cleanup functions in sections EIC Construction by XCMS and EIC Construction by MZmine 2 in the Supporting Information.

Detection of EIC Peaks by XCMS and MZmine 2. Following EIC construction is the detection of peaks in them. Both XCMS and MZmine 2 include at least two different algorithms for detecting EIC peaks. Among the algorithms, *centWave*, originally developed for XCMS but also used by MZmine 2, is probably the best performing one in terms of sensitivity and robustness. However, key differences in the use of *centWave* cause differences in the detected peaks when comparing results from XCMS and MZmine 2.

Peak detection is a crucial part of any preprocessing workflow of untargeted metabolomics data but is also primarily responsible for false negative peaks (i.e., missing peaks) and false positive peaks. It is only through understanding the details of *centWave* that we can understand how false peaks are produced. Next we describe the *centWave* method and the MZmine 2's use of *centWave* at a level of detail that is descriptive without being too cumbersome.

XCMS Detection of EIC Peaks. The XCMS *centWave* algorithm detects peaks primarily by using a continuous wavelet transform method. However, not all EICs are passed to the wavelet transform method for peak detection. Before the wavelet transform, XCMS discards EICs in two different steps based on estimates of the EIC baseline and noise. After the wavelet transform, XCMS examines the wavelet coefficients and the intensities of the detected peaks and determines whether or not each peak is valid.

A total of five sequential filters are used to eliminate bad EICs and peaks. Several variants of the EIC constructed by the *centWave* EIC construction algorithm are used in the filtering steps. For easy understanding, we describe these variants plainly here but include the exact details of these EIC variants and how they are used in the section Detailed Description of XCMS EIC Peak Detection of the Supporting Information. This section in the Supporting Information is parallel to the section here but contains many of the details that have been omitted herein in favor of readability. Next, we describe the five sequential filters that *centWave* uses.

(1) Number of consecutive data points above an estimate of the EIC baseline. XCMS estimates the baseline of a given EIC by finding the arithmetic mean of the EIC intensities after the EIC is extended in retention time on both sides. Such an extension is done using the m/z range established by the original EIC. With the EIC baseline estimated, XCMS counts the maximum number of sequential points above the baseline. If the largest number of sequential points above the baseline is found to be greater than or equal to a threshold value, the EIC continues to the next step, otherwise it is discarded.

(2) Signal-to-noise ratios of EICs. An EIC must also pass a S/N (signal-to-noise ratio) check in order to be passed to the wavelet transform. Unless the EIC is very long, XCMS finds two values important to its estimation of S/N. We call these $lnoise_{\text{std}}$ and $lnoise_{\text{mean}}$ which are the standard deviation and mean of two sets of points located immediately outside the original EIC's RT boundaries, one set on each side. The name of these values comes from the *centWave* source code, and we believe *lnoise* stands for "local noise". The mean and standard deviations are calculated for each set and the smaller mean and standard deviation are take to be the final $lnoise_{\text{mean}}$ and $lnoise_{\text{std}}$,

DOI: 10.1021/acs.analchem.7b01069
Anal. Chem. 2017, 89, 8689−8695

respectively. An important detail is that the mean and standard deviation are taken after certain groups of consecutive data points above the baseline have been removed. Therefore, it is not always clear by looking at the data what $lnoise_{std}$ and $lnoise_{mean}$ will be.

A serious problem exists in XCMS's determination of $lnoise_{std}$ and $lnoise_{mean}$. If no points or only a single point is found to be below the baseline in either of the two sets of points, then the final $lnoise_{std}$ and $lnoise_{mean}$ will both be set to one. As we will show in the proceeding section, this can cause false positive EIC peaks because it results in a great underestimate of the local noise.

Eventually, XCMS determines whether or not an EIC will be discarded based on two values: $lnoise_{mean}$ and a value called *sdthr* (variable name in *centWave* code). The *sdthr* value is found by taking the *S*/*N* threshold set by the user and multiplying it by $lnoise_{std}$. If the maximum intensity of the EIC minus $lnoise_{mean}$ is not greater than or equal to the *sdthr* value, the EIC is discarded.

(3) Wavelet transform and the first wavelet coefficient check. For EICs that have passed both of the two filters described so far, XCMS passes them to the wavelet transform method. After the CWT has been performed, another *S*/*N* check is done using the wavelet transform coefficients. Specifically, $lnoise_{mean}$ is subtracted from the largest wavelet coefficient for each EIC. Only if this value is greater than or equal to *sdthr* will this EIC be further considered.

We must note that this comparison of wavelet coefficients with $lnoise_{mean}$ and *sdthr* is not valid as it mixes two entirely separate quantities. The coefficients from a CWT are found through an integral transform which produces a quantity with different units from that of the intensities in an EIC. Even if this check correctly filters out a select few false positives, it is not a valid way of filtering and its results are unpredictable. In the section Detailed Explanation of Examples of False Positive Detection of the Supporting Information, we show an example of it producing a false negative peak.

(4) Ridgeline detection and the second wavelet coefficient check. True EIC peaks should produce local maxima in the wavelet coefficients at multiple wavelet scales, and these maxima should be relatively close together in the RT domain. As a result, these maxima form the so-called ridgeline. Ridgeline detection is the process that connects local maximum wavelet coefficients across different wavelet scales. Requiring the local maxima of wavelet coefficients to form a ridgeline across a number of consecutive scales is a robust way of detecting EIC peaks. However, *centWave* does not impose a length requirement on the ridgelines. Instead, ridgelines of any length are accepted. Consequently, many false positive peaks can be produced.

After ridgelines are detected, *centWave* carries out the fourth filtering step which is a more exclusive version of the wavelet coefficient check from above as it is performed using a restricted set of coefficients. For all points in a ridgeline, the coefficients at the smallest scale corresponding to those points' RT are considered. For example, if the smallest scale of a CWT is 3 and a ridgeline contains 3 connected local maxima at 1.00, 1.01, and 1.00 min, for scales 4, 5, and 6, respectively, the coefficients at scale 3 (smallest scale) for times 1.00, 1.01, and 1.00 are considered. Then $lnoise_{mean}$ is subtracted from these values and the result is checked to see if it is greater than *sdthr*. Each ridgeline must have one point satisfying this condition to be further considered. We do not believe that this is a meaningful comparison and have no insight into the reasoning behind it.

(5) EIC intensity check. After ridgeline detection, a final *S*/*N* check is performed. For each ridgeline, the maximum wavelet coefficient at different scales may be located at slightly different RT. *centWave* checks the intensity of the EIC at all of the corresponding RTs. At least one of the data points in the EIC must have an intensity which after subtracting $lnoise_{mean}$ is greater than *sdthr*.

MZmine 2 Implementing XCMS centWave. MZmine 2 does not have its own implementation of *centWave*, so it calls the XCMS *centWave* function directly. However, MZmine 2 does not directly pass the EICs it constructs to *centWave*. Each EIC is altered in two ways before being passed into *centWave*. The first alteration is intentional, where the EICs are split up into segments and then each segment is passed to *centWave* separately. MZmine 2 splits an EIC into nonzero sections and passes individual nonzero pieces to *centWave*. This segmentation of EICs could produce false peaks or disrupt the detection of decent peaks if just one scan in an EIC peak has zero intensity. After *centWave*, MZmine 2 collects the boundaries of the peaks detected by the *centWave* algorithm (For details on how peak boundaries are determined in XCMS, please refer to the section Determination of Peak Boundaries by XCMS in the Supporting Information). The maximum intensity data point between the boundaries is taken to be the location of the peak.

The second alteration results in a distorted EIC. To explain this distortion we define I_i to be the ith intensity of the EIC. When MZmine 2 attempts to pass an EIC with an array of intensities (I_0, I_1, I_2,...) to *centWave*, the final EIC that is passed to the wavelet transform method is actually ($I_0 + I_1$, $I_1 + I_2$, $I_2 + I_3$,...) . A specific example of such a distortion, as well as the details on how it occurs, can be found in the section MZmine 2: Problem Using *centWave* of the Supporting Information.

Detailed Explanation of Examples of False Positive Peaks. As we have described in the above sections, there are both similarities and differences between the XCMS and MZmine 2 EIC construction and peak detection processes. The differences between them contribute to the differences found in the results, namely, the lists of EIC peaks detected. Many of the differences in the peak lists come from the false positive peaks that are detected. We have seen several representative examples of false positives in Figure 2. Now that we have introduced the algorithms that XCMS and MZmine 2 use, we will give an overview of how these specific false positive peaks were produced. For more details, please see the section Detailed Explanation of Examples of False Positive Detection of the Supporting Information.

Figure 2A was incorrectly detected as an EIC peak because of problems with *centWave*'s *S*/*N* estimation. The calculated baseline is reasonable (close to what one might expect by inspecting the data visually) and a large number of points exceeds this value; therefore, it passes the continuous-points-above-baseline test. A more serious problem arises in the calculation of $lnoise_{std}$ and $lnoise_{mean}$, which are both found to be 1 because of problems in the signal-to-noise estimation as discussed in the section Detailed Description of XCMS EIC Peak Detection in the Supporting Information. With such a small $lnoise_{mean}$, this peak passes the *S*/*N* check.

Since this EIC passes both the *S*/*N* check and the check for the number of consecutive data points above the EIC baseline, it is passed to the subsequent wavelet transform. It was surprising that this portion of the EIC should have been found

DOI: 10.1021/acs.analchem.7b01069
Anal. Chem. 2017, 89, 8689−8695

as a peak by the CWT since it is unlikely that it could produce a large inner product with the wavelet at a variety of scales. However, from our investigation of *centWave*, we found that XCMS considers ridgelines with the minimum length of 1 as possible peaks (in the XCMS package that we have posted on http://www.du-lab.org/publications.html, this can be found in the *findPeaks.centWave* function on line 1472 of the *xcmsRaw.R* file). We consider this requirement of the minimum length of ridgeline lines to be 1 is a mistake because the purpose of the CWT and ridgeline method is to detect peaks that produce maxima in the wavelet coefficients at a large enough number of consecutive wavelet scales so that some noise peaks can be filtered out. For a much more detailed description about how the false positive peaks in Figure 2 are detected, please see the section Detailed Explanation of Examples of False Positive Detection of the Supporting Information.

Raw mass spectrometry data is notoriously noisy. The aforementioned examples demonstrate the importance of proper noise analysis. As the sensitivity of LC/MS and GC/MS platforms is continuously increasing, noise problems are becoming more and more serious. Therefore, proper handling of noise is becoming critical for successful data preprocessing. Furthermore, ridgelines must be of sufficient length. Improper noise analysis together with the lack of ridgeline length requirement causes a large portion of the false positive EIC peaks detected by *centWave*.

Effect of False Positive Peaks on Compound Identification. Many of the false positive EIC peaks will propagate through analysis steps after peak detection and might be identified as spurious compounds. To show that this occurs in practice we run the full MZmine 2 preprocessing pipeline on all four data files. With each list of EIC peaks detected by MZmine 2, we run CAMERA[21] to find the isotopes and adducts. Next we search each monoisotopic mass and the corresponding experimental isotopic distribution (at least one isotope required) against the PubChem database to tentatively determine the identity of the monoisotopic mass. The CAMERA algorithm and the database search are performed with strict parameters to try to minimize false positives. By visual inspection of the peaks associated with each of the detected compounds, we determine (using the criteria described in the Results and Discussion subsection Comparison of Peaks Detected by XCMS and MZmine 2) if the identification is the product of a false positive or real peak. The specific parameters used in CAMERA and the database search can be found in the section Methods and Parameters for Compound Identification of the Supporting Information, where we also show an example of a false positive peak and its isotopic distribution. From our visual inspection of peaks, we find the percentage of compounds identified from real peaks for each data file: DCSM 91.7%, YP01 75.0%, YP02 71.3%, VT001 62.9%. Though the majority of compounds detected are determined from real peaks, there still remain a troubling number that are determined from false positives.

False Negatives. In Figure 5A,B we show two peaks that have been missed by both XCMS and MZmine 2. Though the peak in panel A does not have a smooth profile, it has been manually confirmed to be the first isotope (^{13}C) of a compound whose monoisotopic mass produces a clear high intensity EIC.

Figure 5 panel C illustrates the reason why XCMS missed the peak shown in panel A. For this peak, XCMS does not find a coefficient at the smallest wavelet scale in the ridgeline that passes the second wavelet coefficient check. Figure 5C shows all

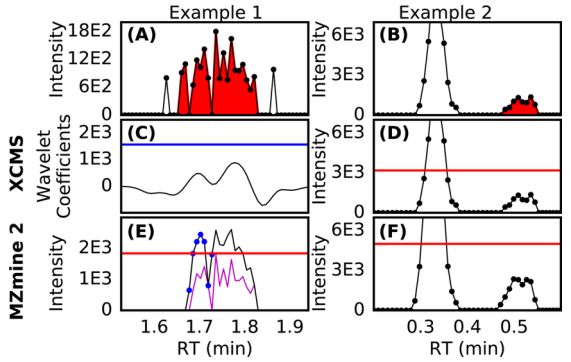


Figure 5. Two examples of EIC peaks missed by XCMS and MZmine2, shaded in red, are shown in parts A and B. Both are from data file YP01. The representative m/z values for the left and right red peak are 300.148 and 123.954, respectively. (C) The magnitude of the wavelet coefficients at the smallest wavelet scale (black) and *sdthr* (blue). (D) XCMS EIC (black) and $lnoise_{mean}$ (red). (E) MZmine 2 EIC (magenta), distorted EIC (black), EIC portion checked for number of points above baseline (blue dots), and the estimated baseline (red). (F) MZmine 2 distorted EIC (black) and $lnoise_{mean}$ (red).

of the coefficients at the smallest wavelet scale across the same RT as the displayed peak as well as the *sdthr* value as a blue line. The $lnoise_{mean}$ is found to be ~26 so clearly any of these coefficients minus $lnoise_{mean}$ are not greater than *sdthr*. However, this peak actually creates a very clear set of maxima in the coefficients. The problem is that *centWave* compares the wavelet coefficients at the smallest scale against *sdthr* for this wavelet coefficient check. Consequently, the check failed and the EIC peak was discarded.

It is worth reiterating that comparing wavelet coefficients with *sdthr* is invalid because they are quantities of different units. Comparing wavelet coefficients at the smallest wavelet scale with *sdthr* is even worse since false negative peaks could result, as demonstrated in Figure 5.

Figure 5 panel E depicts why MZmine 2 missed the peak in panel A. The magenta line is the relevant portion of the EIC from the actual data. The black line is the distorted EIC that MZmine 2 actually passes into *centWave*. The red line is the baseline estimated from the distorted EIC. The blue circular markers denote one of the temporary EICs used for peak detection. For an EIC to be passed to the wavelet transform, there must be a certain number of consecutive data points (4 in this case) in the temporary EIC that are above the baseline. However, this temporary EIC has only three consecutive data points above the baseline. The leftmost marker that appears to intersect the line has an intensity that is in fact below it. With only three consecutive points above the baseline, this EIC is discounted.

Figure 5 panel D shows the reason why the peak highlighted in red in panel B is missed by XCMS. The red line again shows the baseline value used in the consecutive-points-above-baseline check. This peak is not found because there are no points above the baseline. In this case the baseline for the red peak in (B) is overestimated due to the presence of the left peak that is of much higher intensity. For the same reason, the peak was missed by the *centWave* algorithm used by MZmine 2, as shown in panel F. Comparison of panel F to panels B and D reveals the distortion of EICs in MZmine 2's use of *centWave*.

DOI: 10.1021/acs.analchem.7b01069
Anal. Chem. 2017, 89, 8689−8695

CONCLUSION

To prevent the propagation of peak detection errors through an untargeted matabolomics data preprocessing pipeline, the EIC construction and EIC peak picking steps must be done correctly. Because of the importance of these two steps, we have performed a careful study of their implementation in XCMS and MZmine 2. The detailed description of the EIC construction and chromatographic peak picking methods used by XCMS and MZmine 2 focuses on aiding the understanding of their differences and the causes of false positive and false negative EIC peaks detected by them.

Our investigation of the XCMS and MZmine 2 has led to the discovery of four important issues. First, due to the incorrect passing of data between MZmine 2 and *centWave*, MZmine 2's EICs are distorted in the *centWave* algorithm. Second, in the *centWave* algorithm there is no ridgeline length requirement to ensure that peaks must produce maxima in the wavelet coefficients at a variety of scales. Third, the *centWave* algorithm incorrectly directly compares wavelet coefficients to the intensities of the EIC. Fourth, the estimation of baseline and noise of an EIC is convoluted and circumstantial, which can cause ineffective filtering of bad EICs.

Even though we find several problems in *centWave* and in the MZmine 2 use of *centWave*, there are many circumstances in which the peak detection algorithms perform well. We hope that our investigation of the peak picking methods and the discovered problems will help guide their future improvements and improve the ability of users to effectively use them.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.7b01069.

Noncomprehensive list of software packages developed for preprocessing metabolomics data; experimental procedures for the data files; details on EIC construction by XCMS; details on EIC construction by MZmine 2; detailed description of XCMS EIC peak detection; determination of peak boundaries by XCMS; MZmine 2: problem using *centWave*; detailed explanation of examples of false positive detection; effect of false positive peaks on compound identification; compounds manually confirmed in the DCSM file; compounds manually confirmed in the YP01, YP02, and VT001 files; preprocessing parameters used by XCMS and MZmine 2; results produced by XCMS and MZmine 2 with different parameters; methods and parameters for compound identification; and XCMS *centWave* variables (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: xiuxia.du@uncc.edu. Phone: (704) 687-7307.

ORCID

Xiuxia Du: 0000-0002-3468-9585

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank the National Science Foundation Award 1262416 for funding this research. In addition, the authors thank Dr. David A. Horita at the University of North Carolina at Chapel Hill for his insightful discussions.

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*, 1060−1083.

(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) Stolt, R.; Torgrip, R. J.; Lindberg, J.; Csenki, L.; Kolmert, J.; Schuppe-Koistinen, I.; Jacobsson, S. P. *Anal. Chem.* **2006**, *78*, 975−83.

(8) Takahashi, H.; Morimoto, T.; Ogasawara, N.; Kanaya, S. *BMC Bioinf.* **2011**, *12*, 259.

(9) Wei, X.; Shi, X.; Kim, S.; Zhang, L.; Patrick, J. S.; Binkley, J.; McClain, C.; Zhang, X. *Anal. Chem.* **2012**, *84*, 7963−71.

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

(11) *XCMS R package*, http://bioconductor.org/packages/release/bioc/html/xcms.html (Accessed January 30, 2017).

(12) *XCMS Online*, https://xcmsonline.scripps.edu/landing_page.php?pgcontent/mainPage (Accessed January 30, 2017).

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

(14) Rafiei, A.; Sleno, L. *Rapid Commun. Mass Spectrom.* **2015**, *29*, 119−27.

(15) *PeakView*, http://sciex.com/products/software/peakview-software (Accessed January 30, 2017).

(16) *MarkerView*, http://sciex.com/products/software/markerview-software (Accessed January 30, 2017).

(17) *Metabolite Pilot*, http://sciex.com/products/software/metabolitepilot-software (Accessed January 30, 2017).

(18) *Sciex Mass Spectrometry*, https://sciex.com/products/mass-spectrometers (Accessed January 30, 2017).

(19) *Binomial Proportion Confidence Interval*, https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval (Accessed January 30, 2017).

(20) Newcombe, R. *Confidence Intervals for Proportions and Related Measures of Effect Size*; Chapman & Hall/CRC Biostatistics Series; CRC Press, 2012.

(21) Kuhl, C.; Tautenhahn, R.; Bottcher, C.; Larson, T. R.; Neumann, S. *Anal. Chem.* **2012**, *84*, 283−9.

DOI: 10.1021/acs.analchem.7b01069
Anal. Chem. 2017, 89, 8689−8695