ADAP
User Manual

Version 3.3.0

Du-Lab Team

Department of Bioinformatics and Genomics

University of North Carolina at Charlotte

xiuxia.du@uncc.edu

http://www.du-lab.org

January 25, 2018

Contents

1 Introduction

ADAP (Automated Data Analysis Pipeline) was developed for pre-processing untargeted mass spectrometry-based metabolomics data. It consists of two components: ADAP-GC and ADAP-LC for pre-processing GC-MS and LC-MS data, respectively. Figure 1 depicts the workflows of the two pipelines. The two pipelines share modules 1, 2, 3, and 5. The differences between the two pipelines lie in modules 4 and 6. Deconvolution is unique to ADAP-GC while peak annotation is unique to ADAP-LC. Compound identification in ADAP-GC is achieved by comparing spectral similarity while compound identification in ADAP-LC is achieved by comparing experimental masses and isotopic distributions against exact masses and theoretical isotopic distributions.

The computing modules for construction of EICs, detecting peaks, and deconvolution have been developed by Du-lab team, implemented in Java, and incorporated into the framework of MZmine 2. Next we describe how to use ADAP-GC and ADAP-LC. For other capabilities of MZmine 2, please refer to the MZmine 2 website [1].

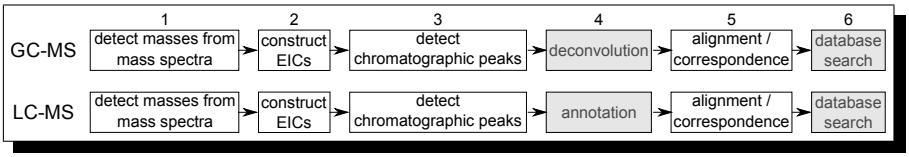


Figure 1: Workflows for pre-processing GC- and LC-MS data.

2 Download and Installation

ADAP computational modules have been part of MZmine 2 since version MZmine 2.24. No installation of extra packages is required. For description on how to download and install MZmine 2, please refer to the MZmine 2 manual [1].

3 ADAP-LC

We will illustrate how to use the ADAP-LC workflow using three data files. The data is in profile mode and so we will start with detecting masses from the mass spectra, i.e. centroiding.

3.1 Detection of Masses from Mass Spectra

Click on *Raw data methods* → *Raw data import*, shown in Figure 2.

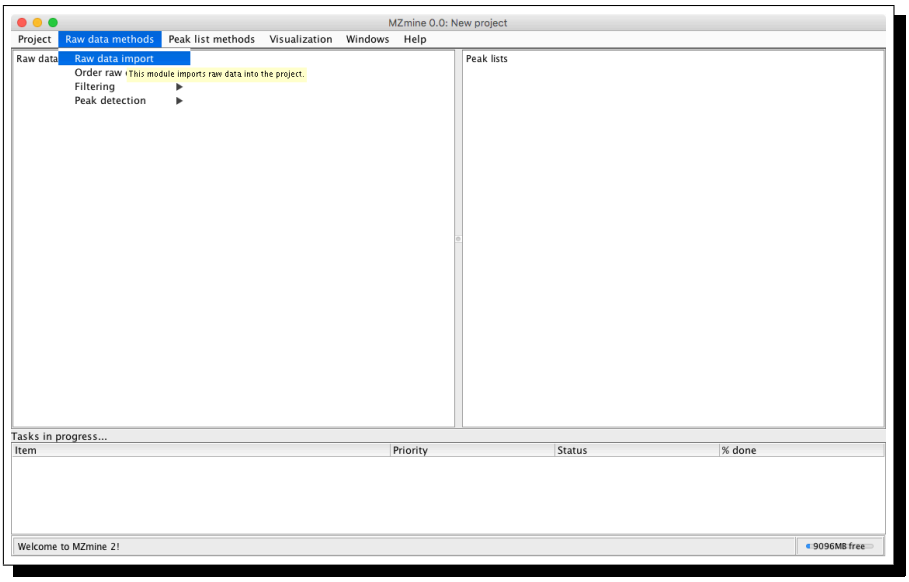


Figure 2: Import the raw data file.

This will open a window from which the desired data files may be chosen. The imported data files will appear in the left hand window of the GUI, labeled *Raw data files*, as shown in Figure 3.

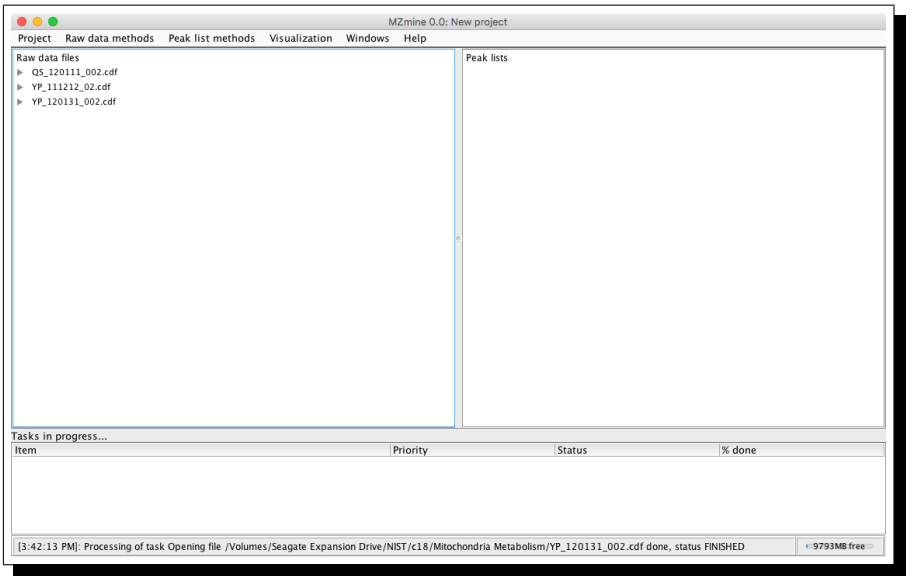


Figure 3: Imported data files.

To detect masses from the profile mass spectra, select the files that have been imported and then click *Raw data methods* → *Peak detection* → *Mass detection* as shown in Fig. 4.

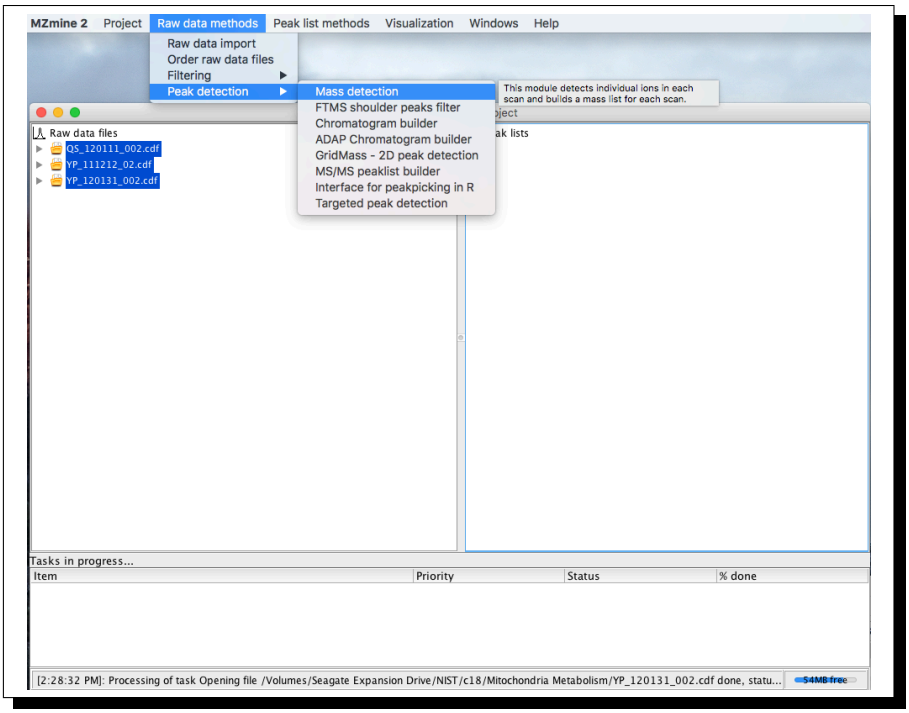


Figure 4: Mass detection from profile mass spectra

This will open a window with several options. From this window click on the *Mass detector* drop down box and choose *Wavelet transform*, then click on the ellipsis box directly to the right of the drop down box. The ellipsis button opens up a parameter selection window for the wavelet transform parameters. Both of these windows and the good parameters for these data files are shown in Figure 5.

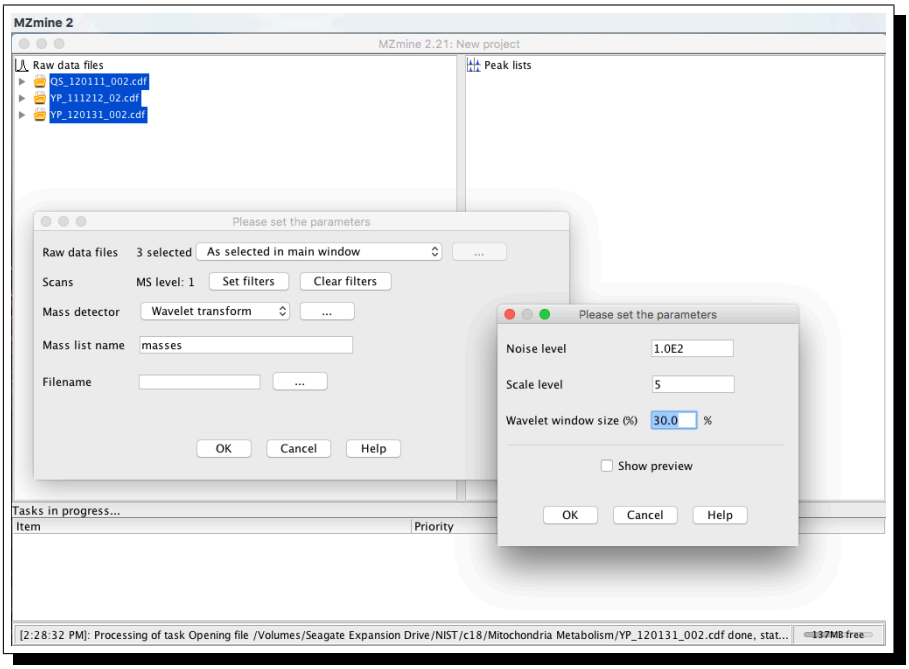


Figure 5: Mass detection by continuous wavelet transform.

Click *OK* in both windows in Figure 5 and start the mass detection process. The process status will be shown in the bottom panel. After the process is finished, click on the triangle immediately to the left of each data file and you will see the list of the profile spectra. Then click on the triangle to the left of each profile spectrum and you will find that the centroid spectrum labelled as *masses* is shown immediately below the corresponding profile spectrum. Double click on the *masses* brings up a window displaying the profile spectra in blue and centroid masses that have been detected in green as shown in Figure 6. By stacking together the centroid spectrum and the profile spectrum, you can check how well the mass detection works.

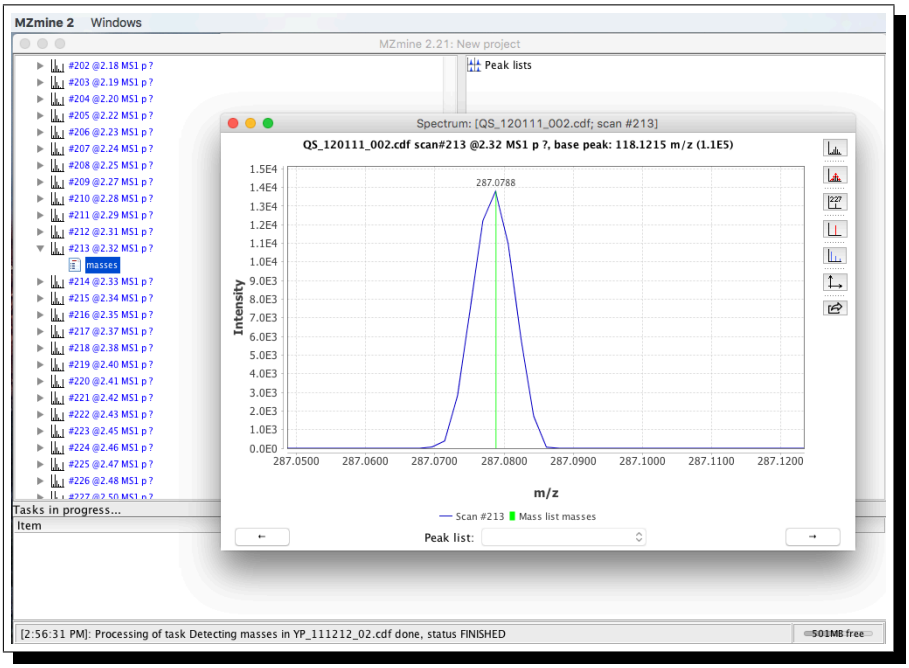


Figure 6: Mass detection result.

You can also use a third party software package, for example `msConvert`, for detecting masses and then import the centroid data into MZmine 2.

3.2 Construction of Extracted Ion Chromatograms

Chromatogram building builds extracted ion chromatograms (EIC) for masses that have been detected by the mass spectrometry continuously over a certain duration of time. To perform chromatogram building using the ADAP method, click *Raw data methods* → *Peak detection* → *ADAP chromatogram builder* as shown in Figure 7.

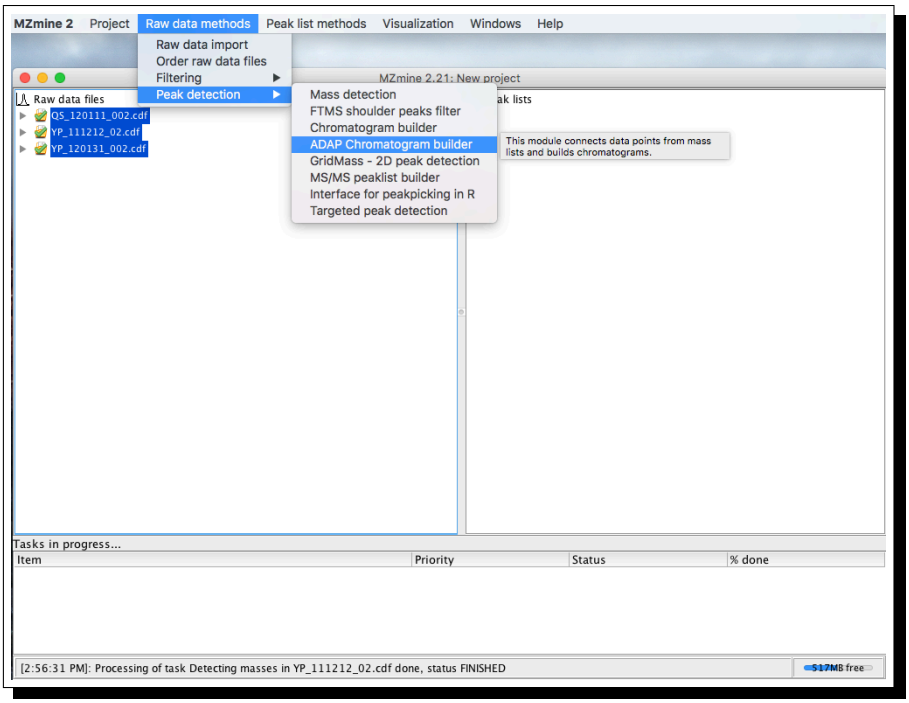


Figure 7: Selecting the ADAP chromatogram building.

This will pull up a window to set the parameters for the ADAP chromatogram building. The window and an example of the good parameters for the example file are shown in Figure 8.

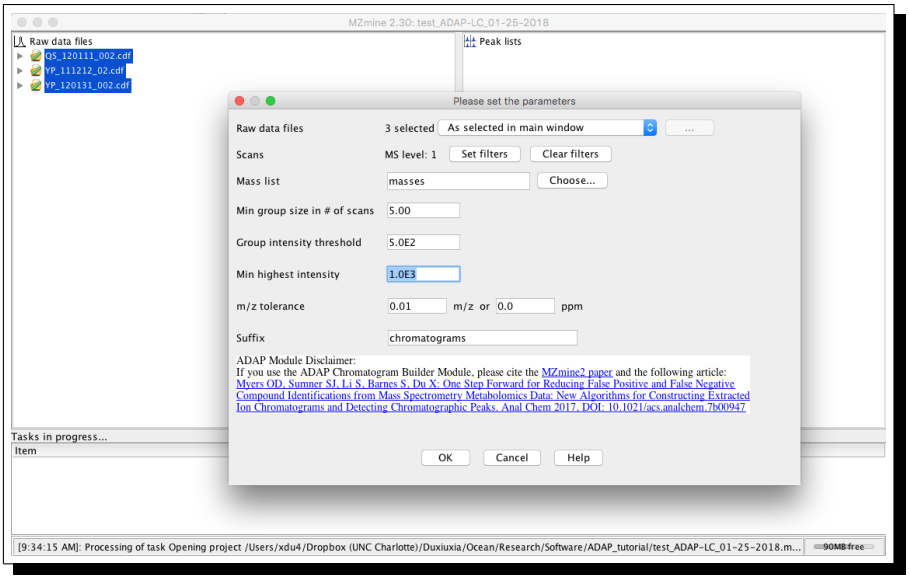


Figure 8: Example of ADAP chromatogram building parameters.

Description of parameters:

- *Min group size in # of scans*: In the entire chromatogram there must be at least this number of sequential scans having points above the *Group intensity threshold* set by the user.

The optimal value depends on the chromatography system setup. The best way to set this parameter is by studying the raw data and determining what is the typical time span of chromatographic peaks.

- *Group intensity threshold*: See above.
- *Min highest intensity*: There must be at least one point in the chromatogram that has an intensity greater than or equal to this value.
- *m/z tolerance*: Maximum m/z difference of data points in consecutive scans in order to be connected to the same chromatogram. Twice the *m/z tolerance* set by the user is the maximum width of a mass trace. We strongly recommend setting the *m/z* value and **not** the ppm value. Whichever value is set to 0.0 will not be used.
- *Suffix*: The resulting chromatogram will be named *file name + suffix*.

Click *OK* starts the chromatogram building process. After the process is complete, a list of chromatograms will appear in the right hand window of the GUI labeled *Peak Lists* as shown in Figure 9.

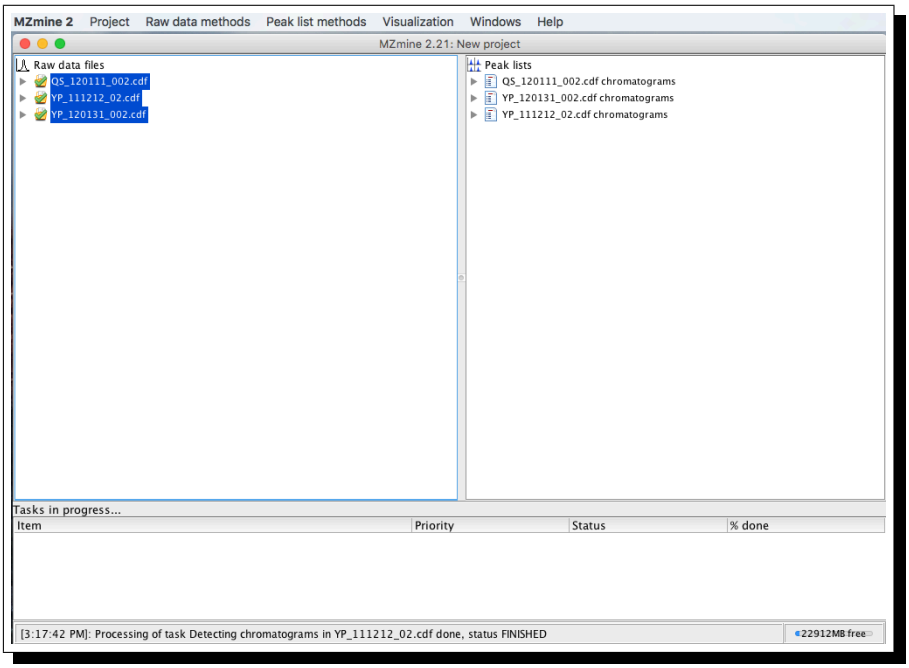


Figure 9: Results of chromatogram building.

Click the triangle to the left of each data file expands the list of EICs as shown in Figure 10.

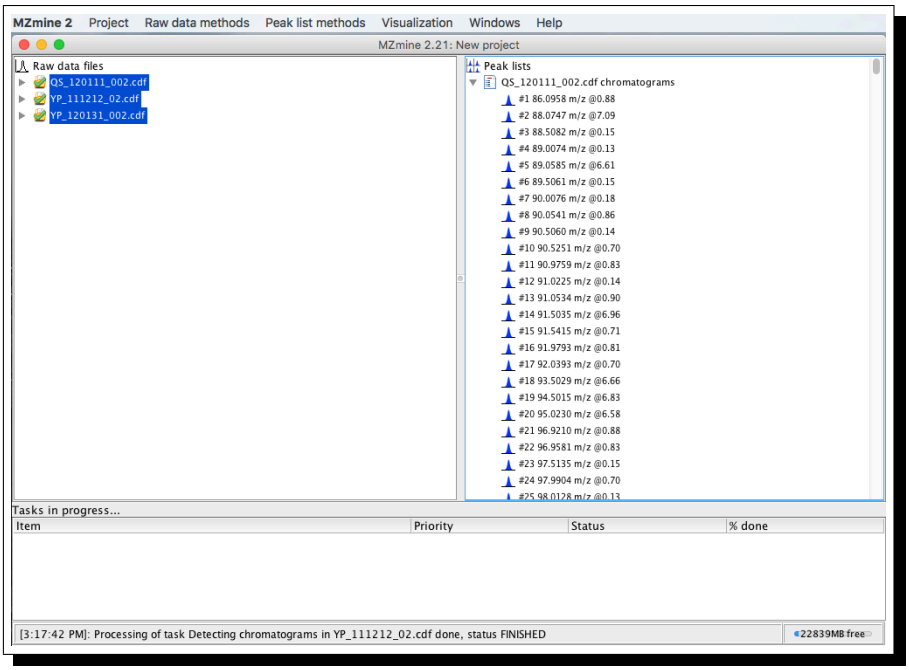


Figure 10: List of EICs that have been constructed.

3.3 Detection of Peaks from EICs

Each EIC that has been constructed spans the entire duration of the chromatography. To detect the peaks from all of the EICs, select the EICs and click *Peak list methods* → *Peak detection* → *Chromatogram deconvolution* as shown in Figure 11.

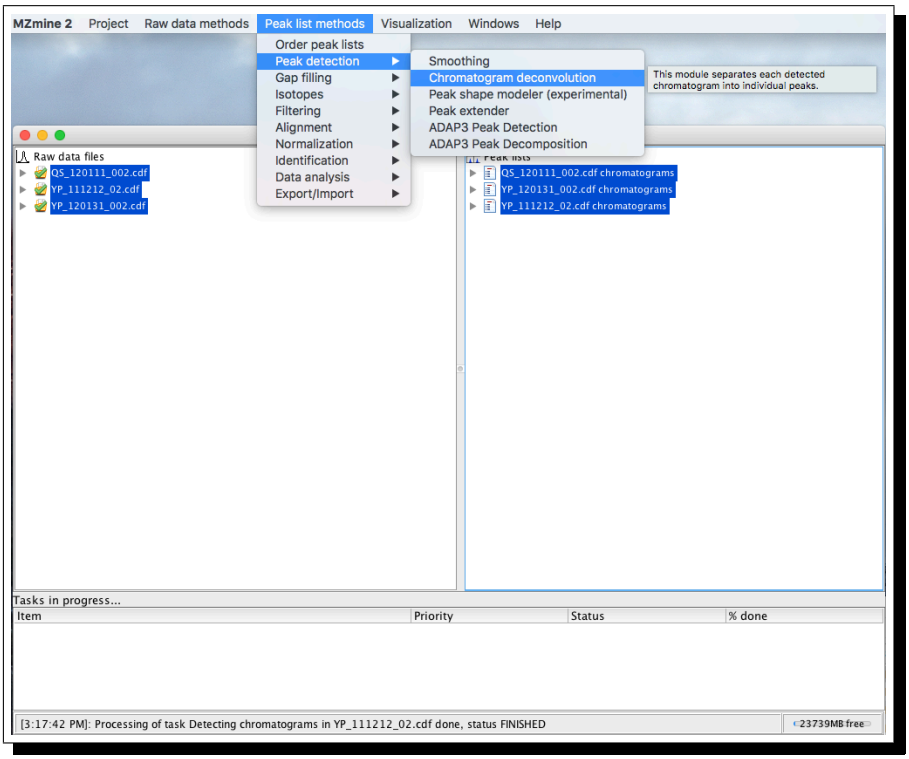


Figure 11: Detect peaks from EICs.

This will open a window with a drop down box for selecting the peak detection method. From the drop down box choose the *Wavelets (ADAP)* option as shown in Fig. 12.

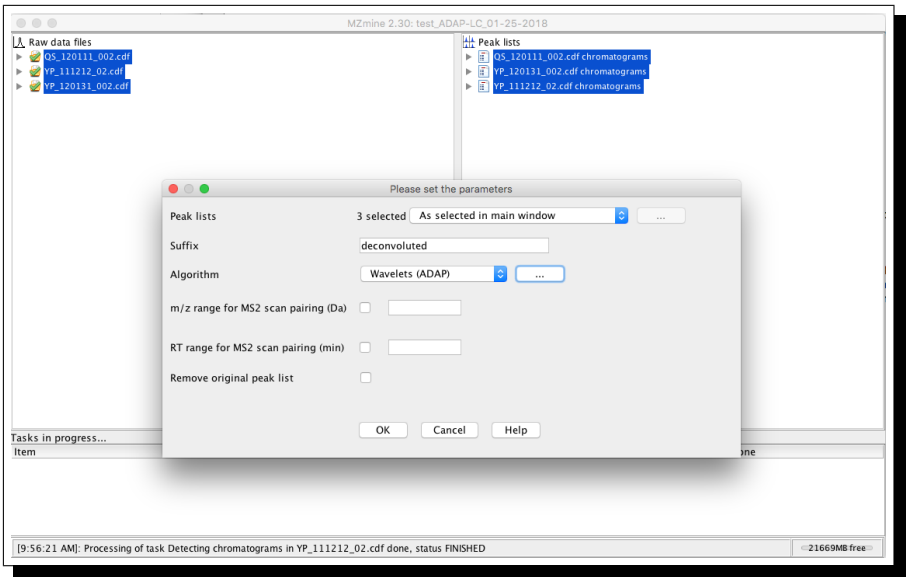


Figure 12: Select ADAP peak detection.

Click on the ellipsis box/button next to the drop down box. The ellipsis button will open a window for setting the peak detection parameters. Both windows are shown in Fig. 13.

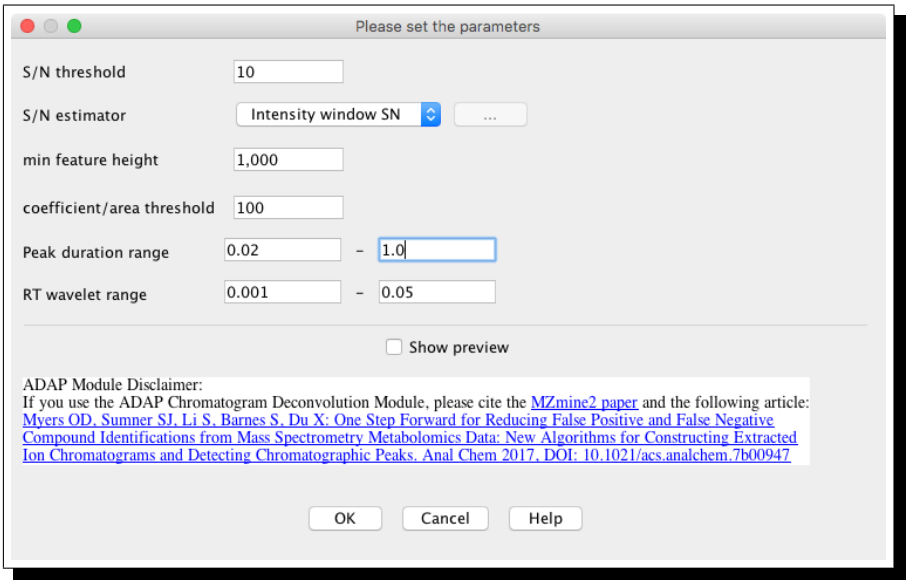


Figure 13: EIC peak detection parameters.

Description of parameters:

- *S/N threshold*: The minimum signal to noise ratio a peak must have to be considered a real feature. Values greater than or equal to 7 will work well and will only detect a very small number of false positive peaks.

- *S/N estimator*: User can choose one of two estimators of the signal-to-noise ratio
  - *Intensity window SN* (tested on LC-MS datasets) uses the peak height as the signal level and the standard deviation of intensities around the peak as the noise level;
  - *Wavelet Coeff. SN* (tested on GC-MS datasets) uses the continuous wavelet transform coefficients to estimate the signal and noise levels. Analogous approach is implemented in R-package *wmtsa*.
- *min feature height*: The smallest intensity a peak can have and be considered a real feature.
- *coefficient/area threshold*: This number must be chosen by looking at examples using the *show preview button* at the bottom of the window. This is the best coefficient found by taking the inner product of the wavelet at the best scale and the peak, and then dividing by the area under the peak. Values around 100 work well for most data.
- *Peak duration range*: Minimum and maximum widths allowed for a peak.
- *RT wavelet range*: Minimum and maximum widths of the wavelets used for detecting peaks.

After the detection of chromatographic peaks is complete, a list of chromatographic peaks will appear below the list of chromatograms in the *Peak lists* window for each data file. Each list of peaks can be exported, separately, by selecting the peaks detected from one data file and clicking on *Peak list methods*, mousing over the *Export/Import* option and then selecting the desired export method (Figure 14). Figure 15 shows a sample export of the chromatographic peak detection results.

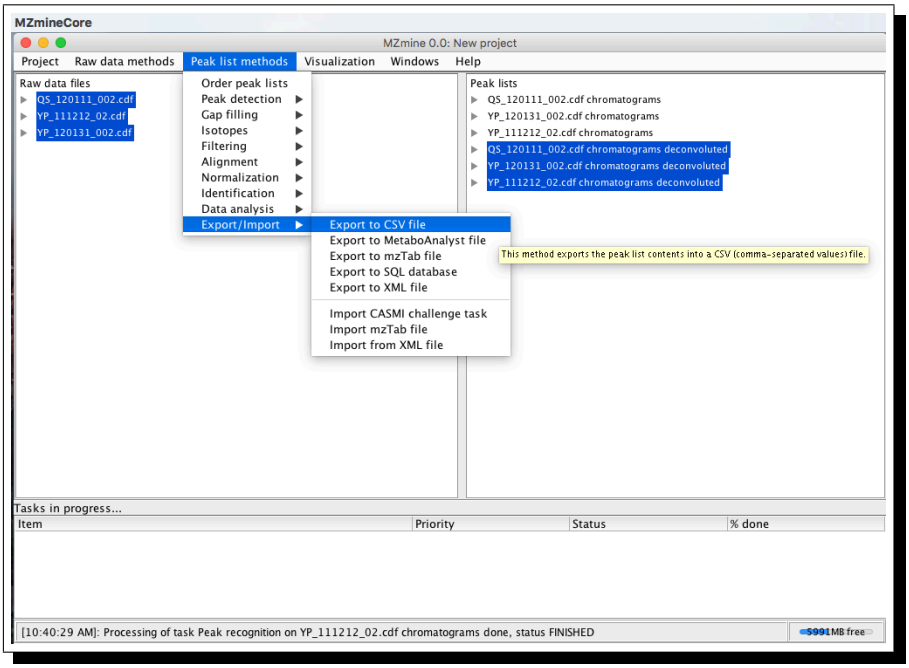
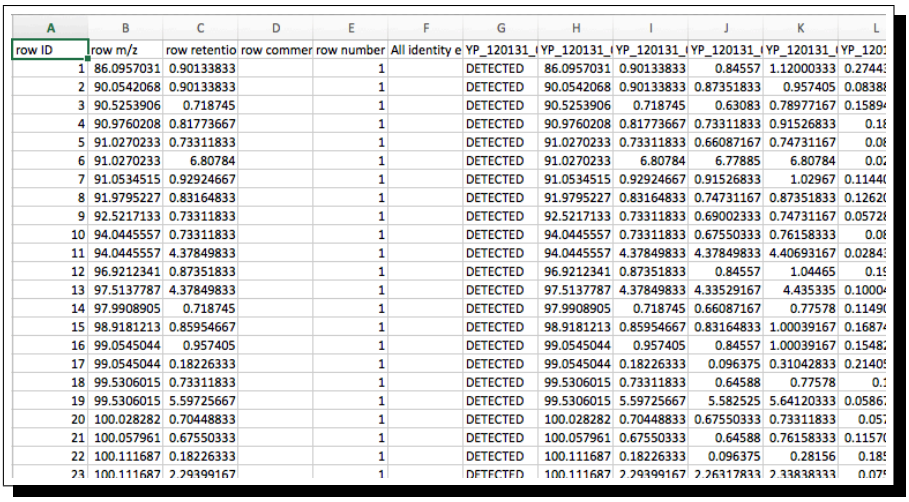


Figure 14: Export results from chromatographic peak detection.



row ID	row m/z	row retentio	row commer	row number	All identity e	YP_120131_(	YP_120131_(	YP_120131_(	YP_120131_(	YP_120131_(	YP_120:
1	86.0957031	0.90133833		1		DETECTED	86.0957031	0.90133833	0.84557	1.12000333	0.2744:
2	90.0542068	0.90133833		1		DETECTED	90.0542068	0.90133833	0.87351833	0.957405	0.0838(
3	90.5253906	0.718745		1		DETECTED	90.5253906	0.718745	0.63083	0.78977167	0.1589(
4	90.9760208	0.81773667		1		DETECTED	90.9760208	0.81773667	0.73311833	0.91526833	0.1(
5	91.0270233	0.73311833		1		DETECTED	91.0270233	0.73311833	0.66087167	0.74731167	0.0(
6	91.0270233	6.80784		1		DETECTED	91.0270233	6.80784	6.77885	6.80784	0.0:
7	91.0534515	0.92924667		1		DETECTED	91.0534515	0.92924667	0.91526833	1.02967	0.1144(
8	91.9795227	0.83164833		1		DETECTED	91.9795227	0.83164833	0.74731167	0.87351833	0.1262(
9	92.5217133	0.73311833		1		DETECTED	92.5217133	0.73311833	0.69002333	0.74731167	0.0572(
10	94.0445557	0.73311833		1		DETECTED	94.0445557	0.73311833	0.67550333	0.76158333	0.0(
11	94.0445557	4.37849833		1		DETECTED	94.0445557	4.37849833	4.37849833	4.40693167	0.0284:
12	96.9212341	0.87351833		1		DETECTED	96.9212341	0.87351833	0.84557	1.04465	0.1(
13	97.5137787	4.37849833		1		DETECTED	97.5137787	4.37849833	4.33529167	4.435335	0.1000(
14	97.9908905	0.718745		1		DETECTED	97.9908905	0.718745	0.66087167	0.77578	0.1149(
15	98.9181213	0.85954667		1		DETECTED	98.9181213	0.85954667	0.83164833	1.00039167	0.1687(
16	99.0545044	0.957405		1		DETECTED	99.0545044	0.957405	0.84557	1.00039167	0.1548:
17	99.0545044	0.18226333		1		DETECTED	99.0545044	0.18226333	0.096375	0.31042833	0.2140(
18	99.5306015	0.73311833		1		DETECTED	99.5306015	0.73311833	0.64588	0.77578	0.:
19	99.5306015	5.59725667		1		DETECTED	99.5306015	5.59725667	5.582525	5.64120333	0.0586:
20	100.028282	0.70448833		1		DETECTED	100.028282	0.70448833	0.67550333	0.73311833	0.05:
21	100.057961	0.67550333		1		DETECTED	100.057961	0.67550333	0.64588	0.76158333	0.1157(
22	100.111687	0.18226333		1		DETECTED	100.111687	0.18226333	0.096375	0.28156	0.18(
23	100.111687	2.29399167		1		DETECTED	100.111687	2.29399167	2.26317833	2.33838333	0.07(

Figure 15: Sample export of results from chromatographic peak detection.

3.4 Annotation of EIC Peaks Using CAMERA

CAMERA is an R package that provides a strategy for compound spectra extraction and annotation of LC-MS datasets. It has been implemented by the MZmine 2 team into MZmine 2. The Du-lab team modified the CAMERA process slightly for extracting experimental isotopic patterns. The isotopic patterns will be used for identifying the analytes. For details about CAMERA, refer to [2, 3].

To do the annotation using CAMERA, click *Peak list methods* → *Identification* → *CAMERA search* (Figure 16).

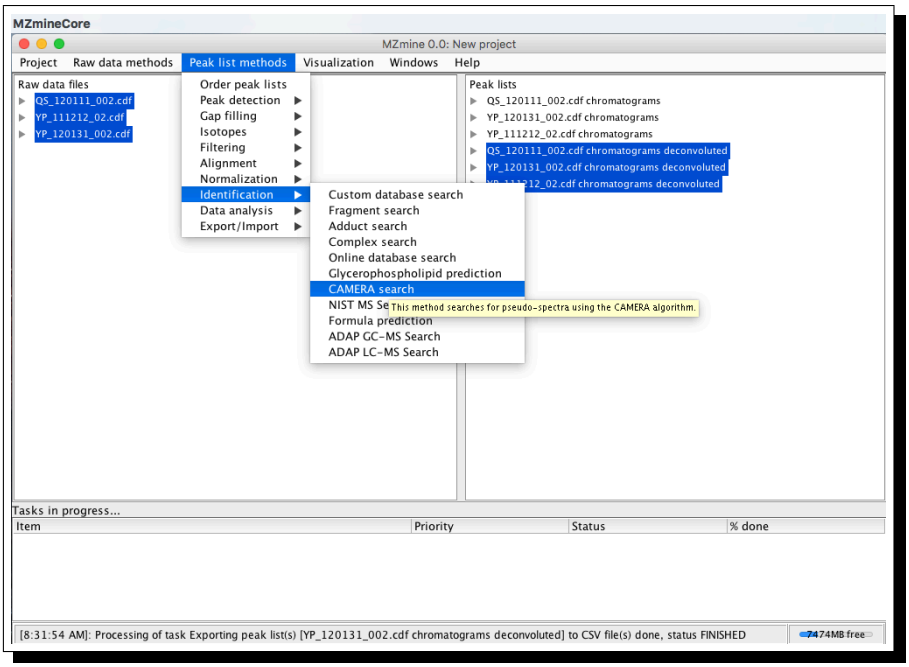


Figure 16: Use CAMERA for annotation of EIC peaks.

A window will pull up as shown in Figure 17 allowing users to specify parameters.

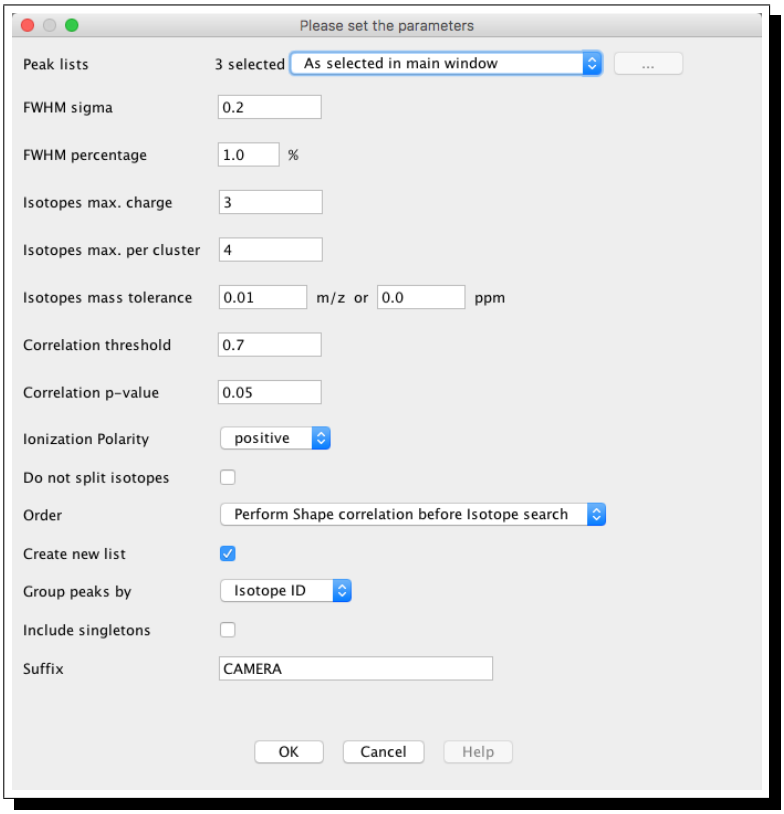


Figure 17: Specify parameters for CAMERA.

With the slight modification by the Du-lab team, an option (item *Order* in Figure 17) is provided to *perform shape correlation before isotope search* for stricter requirement of determining an isotopic pattern. With this stricter requirement, the mass peaks that form an isotopic pattern will have to meet not only the m/z requirement, but peak shape similarity as well. You can use the original CAMERA too by selecting *Perform Isotope search before Shape correlation*. Be aware that it could take a while for a CAMERA search to finish.

After CAMERA does finish the search, the results are displayed as shown in Figure 18.

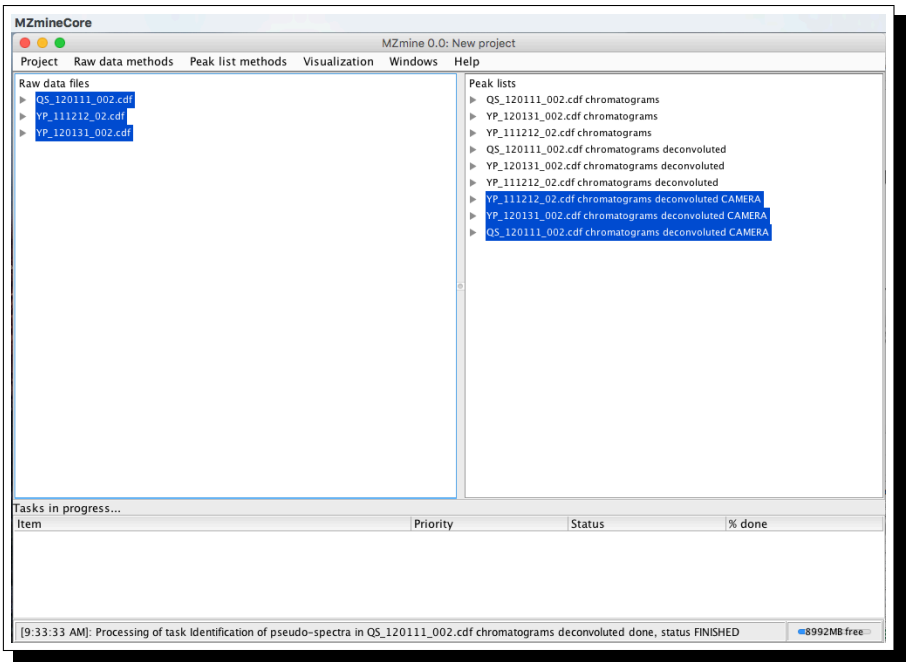


Figure 18: CAMERA finishes searches and results are displayed.

Click on the triangle immediately to the left of *YP_111212_02.cdf chromatograms deconvoluted CAMERA* will display the CAMERA search results (Figure 19) for data file *YP_111212_02.cdf*.

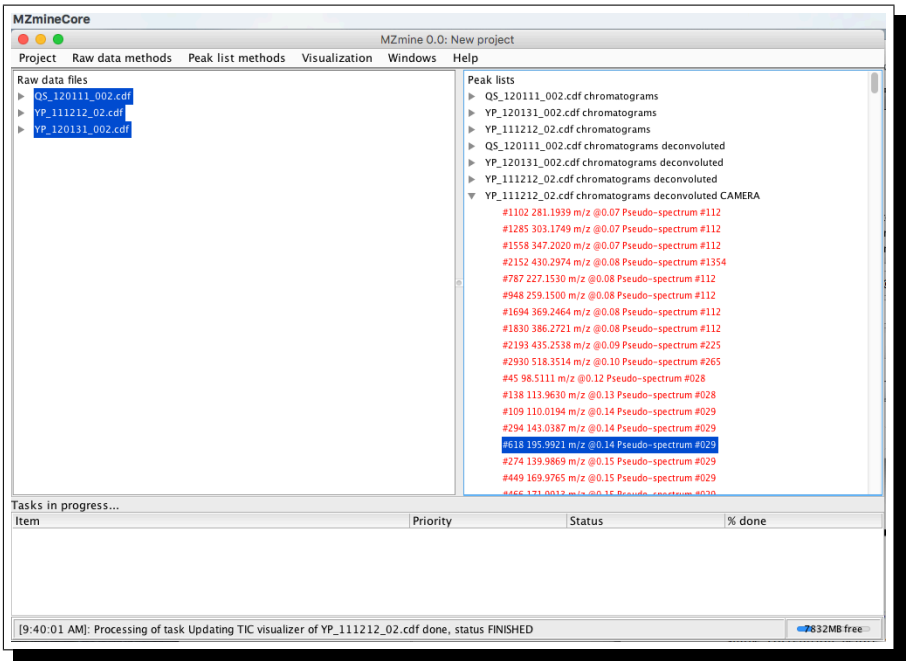


Figure 19: List of pseudo-spectra are displayed.

Each pseudo-spectrum can be displayed in the context of the raw spectrum. For example, to display pseudo-spectrum #029 in data file *YP_111212_02.cdf*, double click the pseudo-spectrum. A window will pull up as shown in Figure 20

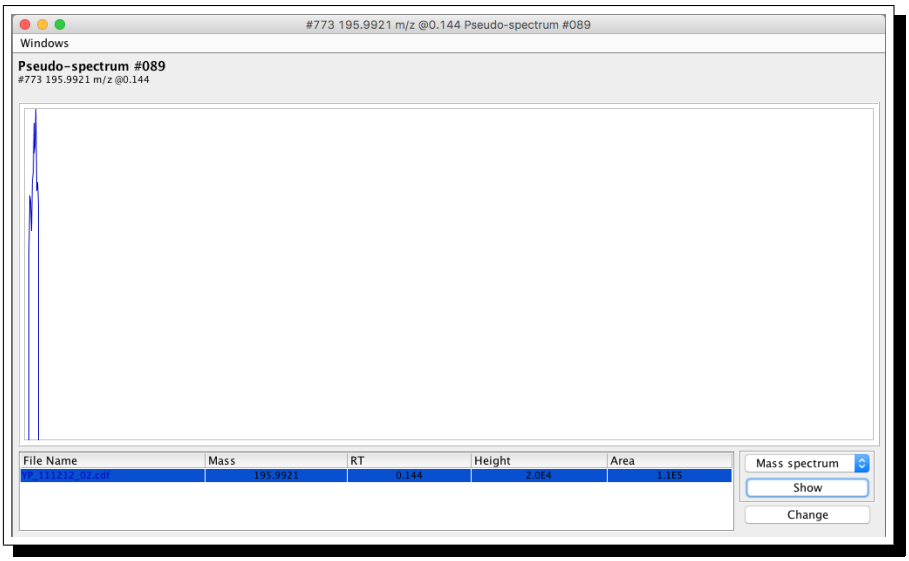


Figure 20: First step of visualizing a pseudo spectrum.

Select *Mass spectrum* in the bottom-right corner and then click on *Show* will pull up a window displaying the pseudo spectrum (green sticks) in the context of the raw spectrum (Figure 21).

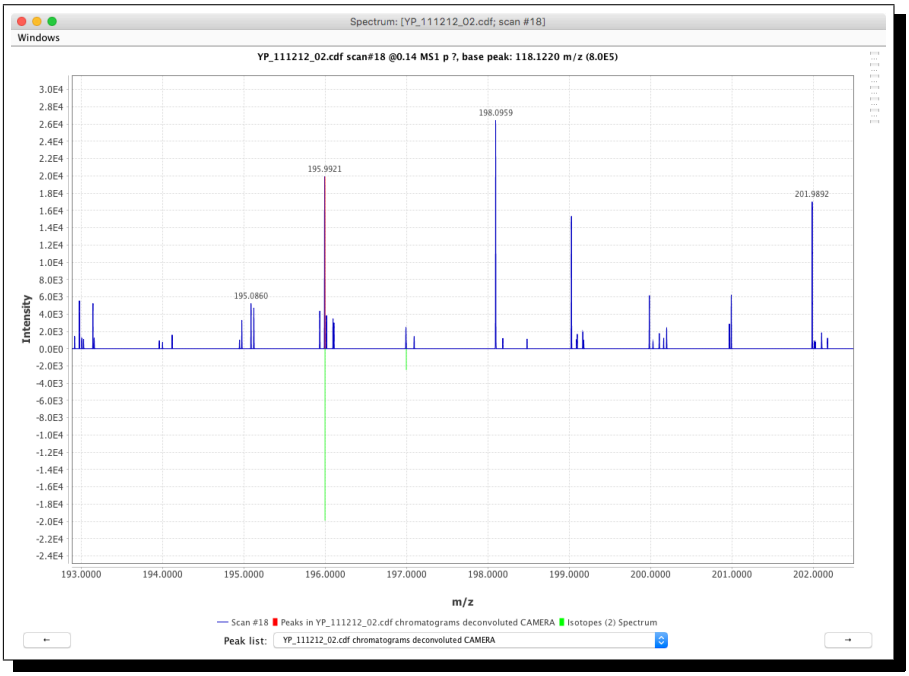


Figure 21: Second step of visualizing a pseudo spectrum.

3.5 Results Export

The final results after detection of EIC peak detection can be exported. Click *Peak list methods* → *Export/Import* → *Export to CSV file* as shown in Figure 22.

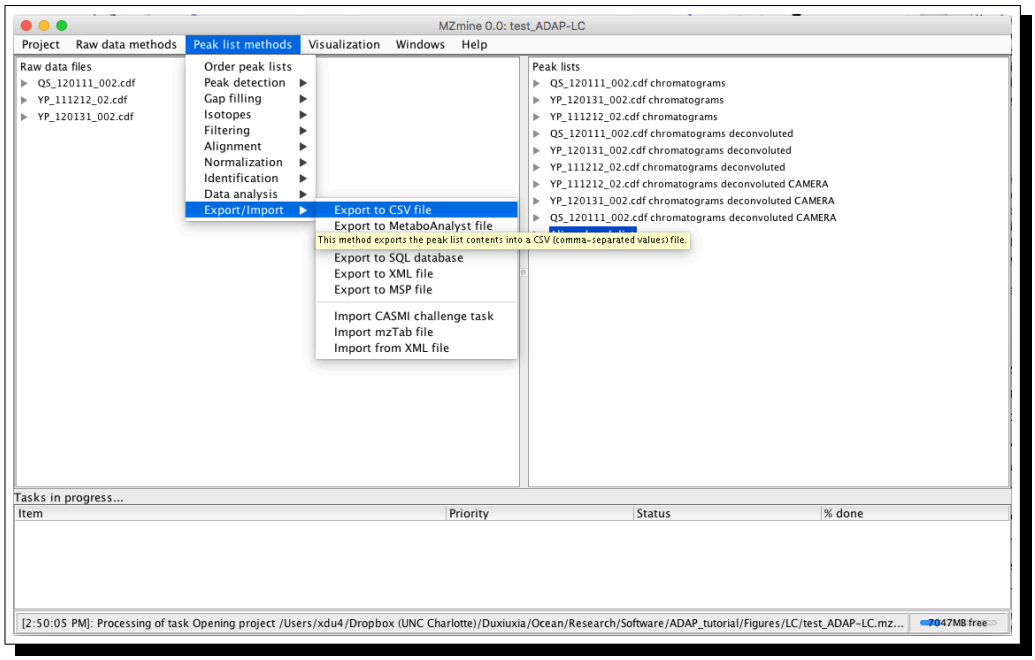


Figure 22: Export results.

A window pulls up as shown in Figure 23 allowing to select what to export.

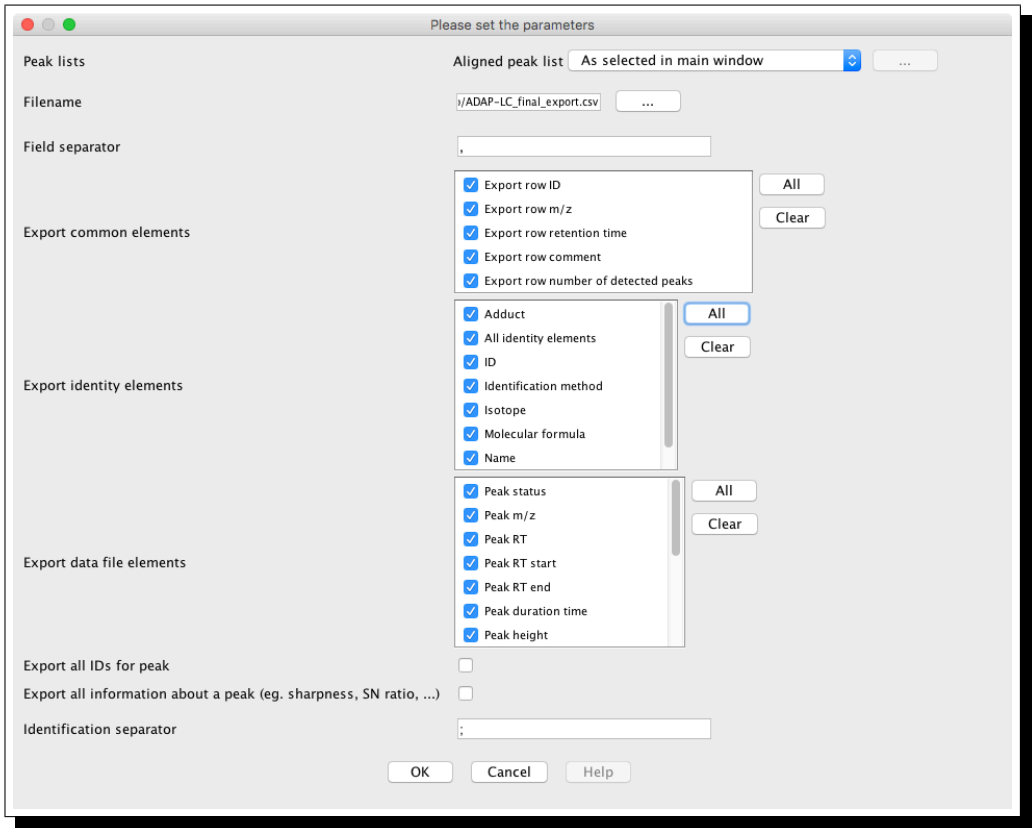
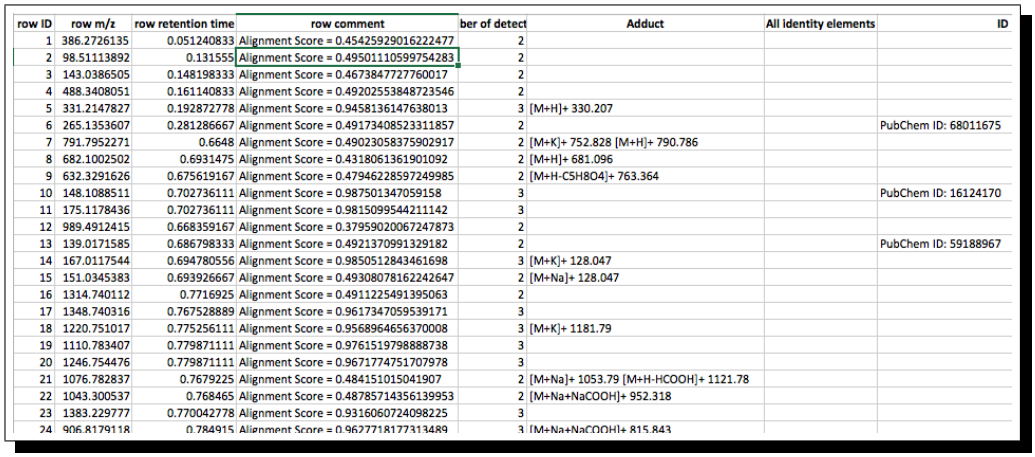


Figure 23: Select what to be exported to a CSV file.

Figure 24 shows part of the exported results.



row ID	row m/z	row retention time	row comment	ber of detect	Adduct	All identity elements	ID
1	386.2726135	0.051240833	Alignment Score = 0.45425929016222477	2			
2	98.51113892	0.131555	Alignment Score = 0.49501110599754283	2			
3	143.0386505	0.148198333	Alignment Score = 0.4673847727760017	2			
4	488.3408051	0.161140833	Alignment Score = 0.49202553848723546	2			
5	331.2147827	0.192872778	Alignment Score = 0.9458136147638013	3	[M+H]+ 330.207		
6	265.1353607	0.281286667	Alignment Score = 0.49173408523311857	2			PubChem ID: 68011675
7	791.7952271	0.6648	Alignment Score = 0.49023058375902917	2	[M+K]+ 752.828 [M+H]+ 790.786		
8	682.1002502	0.6931475	Alignment Score = 0.4318061361901092	2	[M+H]+ 681.096		
9	632.3291626	0.675619167	Alignment Score = 0.47946228597249985	2	[M+H-C5H8O4]+ 763.364		
10	148.1088511	0.702736111	Alignment Score = 0.98750134705915 8	3			PubChem ID: 16124170
11	175.1178436	0.702736111	Alignment Score = 0.9815099544211142	3			
12	989.4912415	0.668359167	Alignment Score = 0.37959020067247873	2			
13	139.0171585	0.686798333	Alignment Score = 0.4921370991329182	2			PubChem ID: 59188967
14	167.0117544	0.694780556	Alignment Score = 0.9850512843461698	3	[M+K]+ 128.047		
15	151.0345383	0.693926667	Alignment Score = 0.49308078162242647	2	[M+Na]+ 128.047		
16	1314.740112	0.7716925	Alignment Score = 0.4911225491395063	2			
17	1348.740316	0.767528889	Alignment Score = 0.9617347059539171	3			
18	1220.751017	0.775256111	Alignment Score = 0.9568964656370008	3	[M+K]+ 1181.79		
19	1110.783407	0.779871111	Alignment Score = 0.9761519798888738	3			
20	1246.754476	0.779871111	Alignment Score = 0.9671774751707978	3			
21	1076.782837	0.7679225	Alignment Score = 0.484151015041907	2	[M+Na]+ 1053.79 [M+H-HCOOH]+ 1121.78		
22	1043.300537	0.768465	Alignment Score = 0.48785714356139953	2	[M+Na+NaCOOH]+ 952.318		
23	1383.229777	0.770042778	Alignment Score = 0.9316060724098225	3			
24	906.8179118	0.784915	Alignment Score = 0.9627718177313489	3	[M+Na+NaCOOH]+ 815.843		

Figure 24: Exported results.

4 ADAP-GC

The first three steps of pre-processing GC-MS data are the same as those for LC-MS data. The major difference between the two pipelines lie in performing the deconvolution. Therefore, we will only describe in detail the deconvolution step. We will use three example high mass measurement accuracy data files that were acquired on ThermoFisher GC-Orbitrap to illustrate.

4.1 Detection of Masses and Construction of EICs

The three data files are in centroid mode already, so the *Centroid* method in MZmine 2 will be used for mass detection as shown in Figure 25.

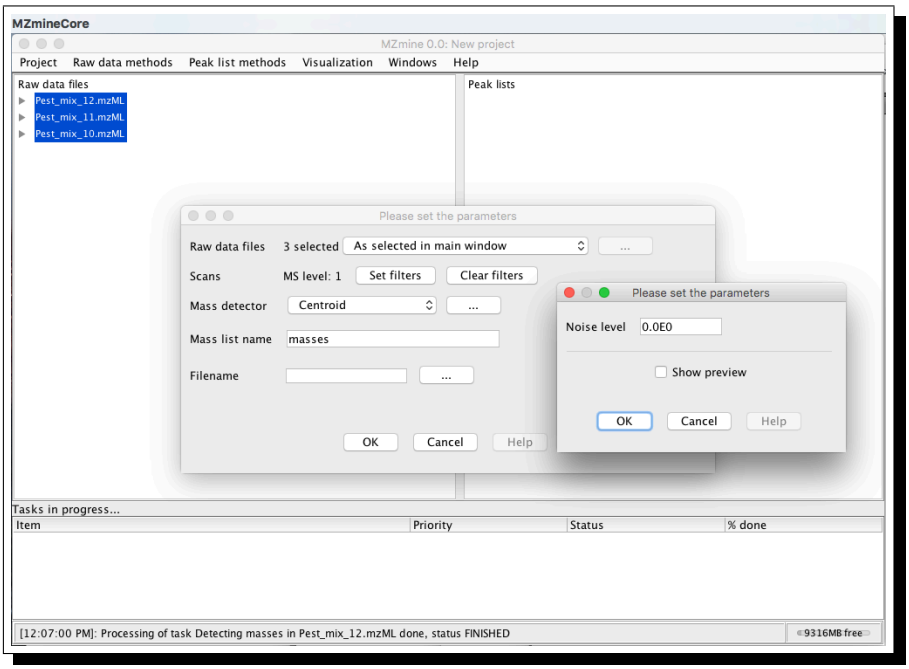


Figure 25: Mass detection of centroid data.

Parameters for constructing EICs are shown in Figure 26.

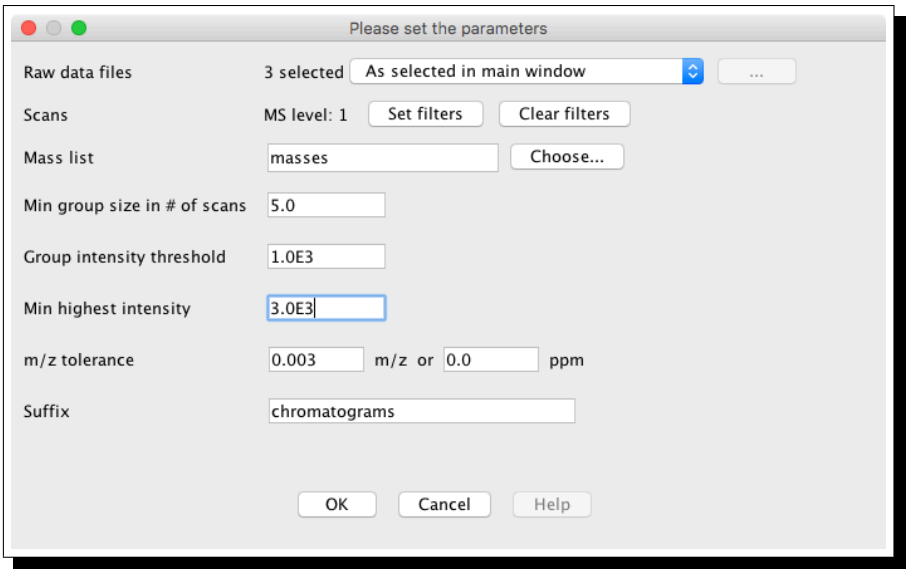


Figure 26: Example parameters for constructing EICs from GC-Orbitrap data.

4.2 Detection of Peaks from EICs

Detection of chromatographic peaks is invoked by clicking *Peak list methods* → *Peak detection* → *Chromatogram deconvolution*. A window will open. Select the *Wavelets* algorithm as shown in Figure 27.

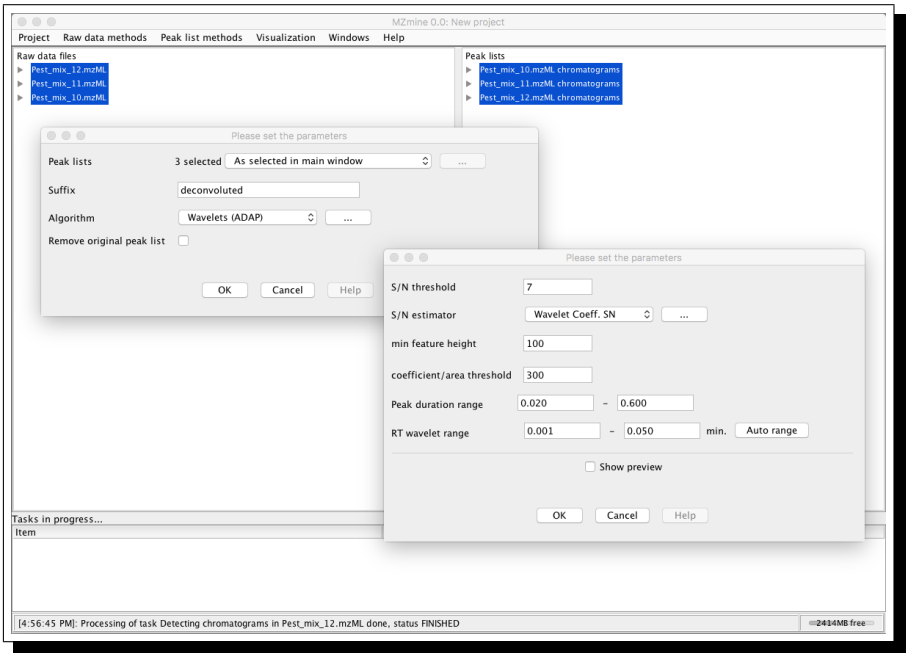


Figure 27: Select *Wavelet (ADAP)* for detecting peaks from EICs for GC-MS data.

Click the ellipse to open the parameter window. Figure 27 shows example parameters. Click *OK* to start peak detection. If the duration of chromatography is long, this step could take a while.

4.3 Sectral Deconvolution

The term *Spectral Deconvolution* refers to detecting analytes by combining similar peaks into clusters and using their intensities to construct fragmentation mass spectra. Detection of analytes is performed by two clustering steps and one filtering step in between. Correspondingly, first all peak are combined into clusters based on their retention times. Then, model peaks are determined, that best describe the peaks in a cluster. Finally, all peaks in a cluster are decomposed into linear combination of the model peaks and their fragmentation mass spectra are constructed. Choice of the model peaks may affect the quality of the constructed fragmentation spectra.

To perform deconvolution, select one or more peak lists detected from data files, then click *Peak list methods* → *Peak detection* → *New Spectral Deconvolution* as shown in Figure 28. To see preview of the deconvolution results, select *Show preview* option at the bottom of the parameter window.

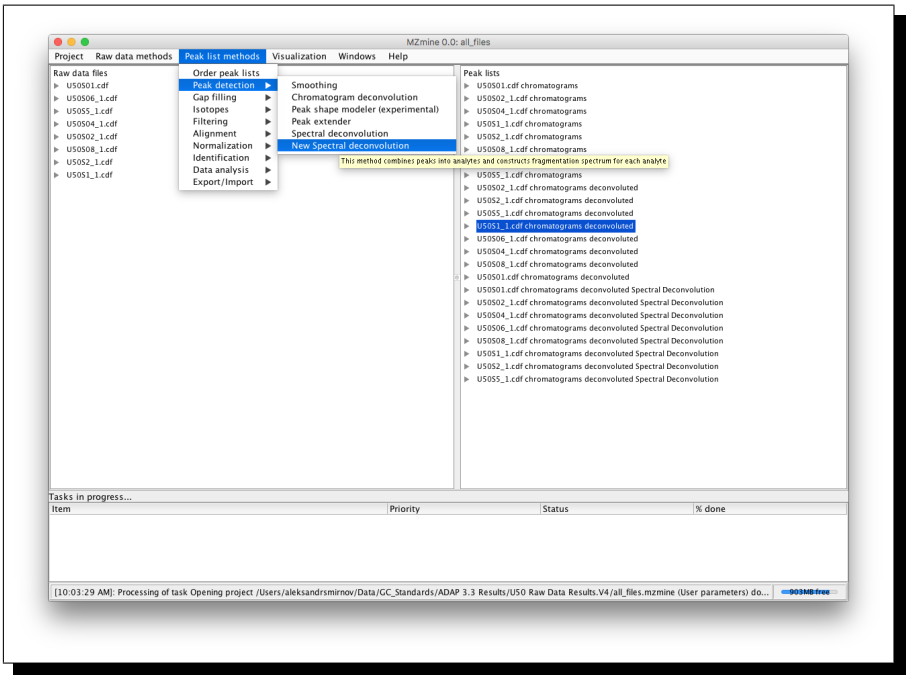


Figure 28: Deconvolution of chromatographic peaks.

A window as shown in Figure 29 pulls up allowing you to specify parameters for deconvolution.

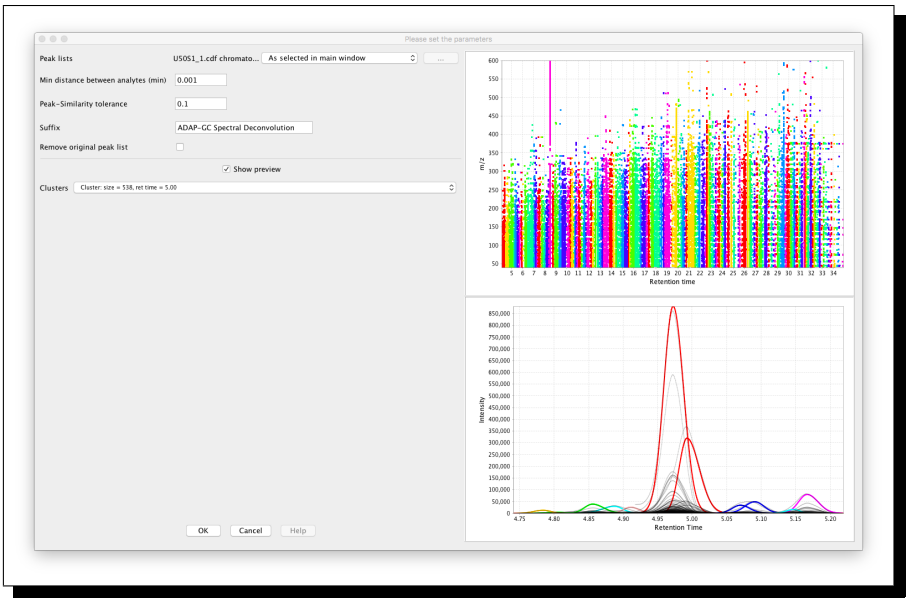


Figure 29: Specify parameters for decomposition of chromatographic peaks.

- *Min distance between analytes*: Minimum allowed retention-time gap between two clusters (analytes). Small values correspond to a big number of narrow clusters, while high values correspond to a small number of wide clusters. If the option *Show preview* is selected, the clusters of peaks are shown on the top right figure. Be aware that wide clusters significantly slow down the algorithm, so try to keep clusters relatively narrow. In our tests, we used the values about 0.005 min.

- *Peak-similarity tolerance*: Every model peak has a bi-gaussian shape and is required to have at least one similar real EIC peak. If the option *Show preview* is selected, model peaks (colored) and real EIC peaks (gray) of a cluster are shown on the bottom right figure. The number of model peaks is controlled by the peak-similarity tolerance: small tolerance corresponds to fewer model peaks and large tolerance corresponds to higher number of model peaks. In our tests, we used the values about 0.1.

After spectral deconvolution is finished, the results are displayed as shown in Figure 30.

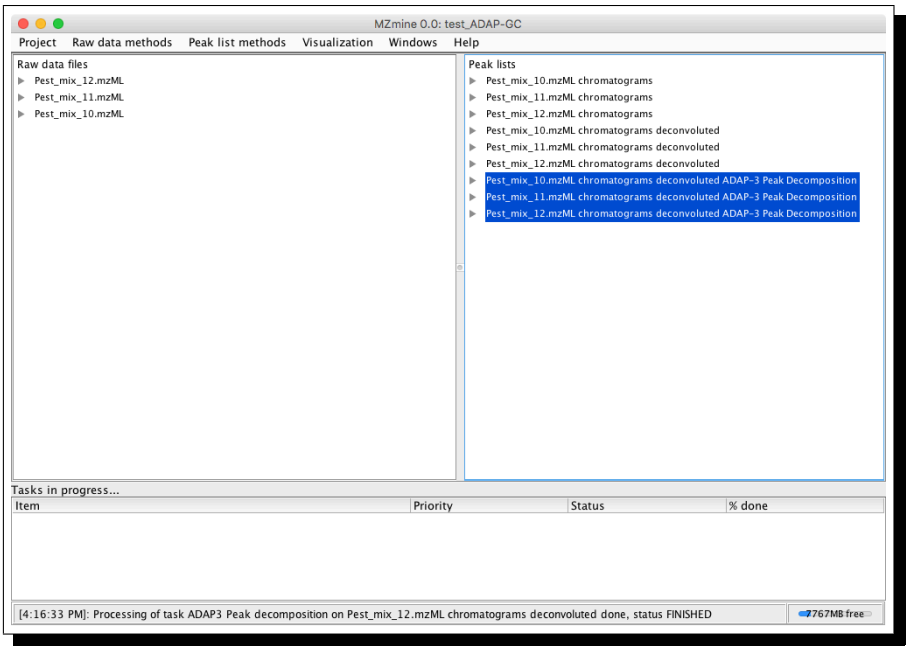


Figure 30: Decomposition results.

Expand the results for each data file by clicking on the left triangle, you will see a list of mass spectra that have been constructed by the deconvolution algorithm (Figure 31). The m/z for each entry is the m/z of the model peak for this spectrum.

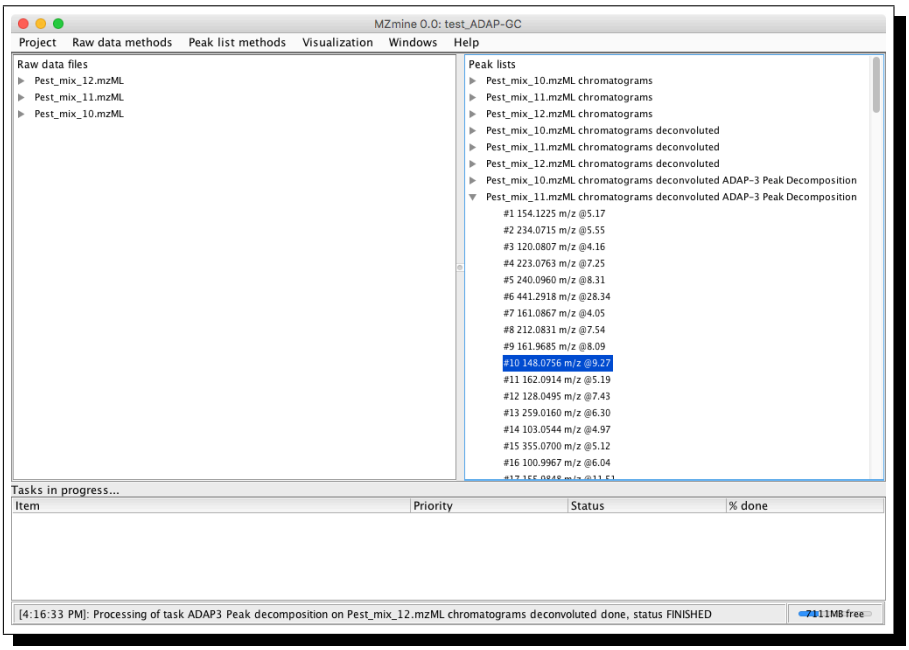


Figure 31: List of mass spectra constructed by the decomposition algorithm.

Double click on a particular mass spectrum will pull up a window as shown in Figure 32.

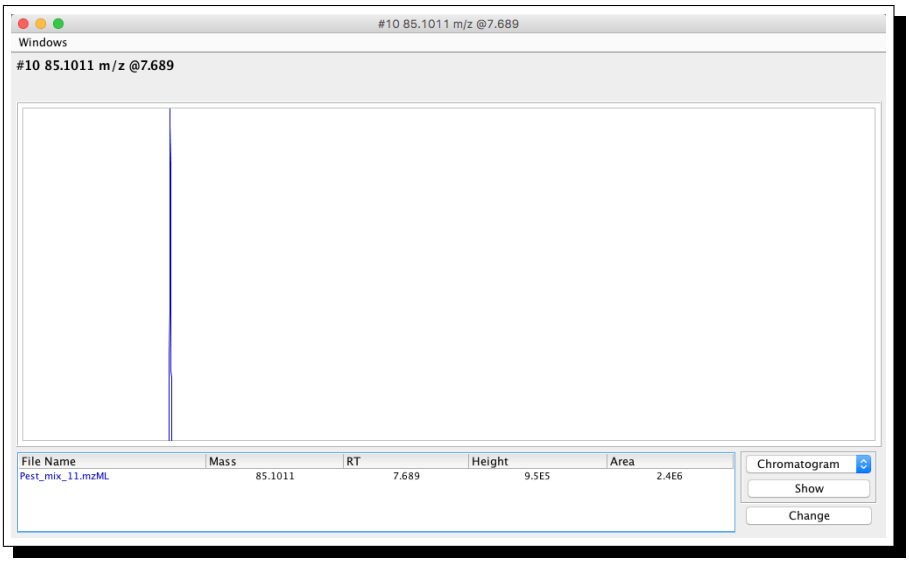


Figure 32: Peak information window.

Click on the data file name and then select *Mass spectrum* in the drop-down menu on the right. The spectrum that has been constructed (green) in the context of the raw spectrum (blue) is displayed (Figure 33).

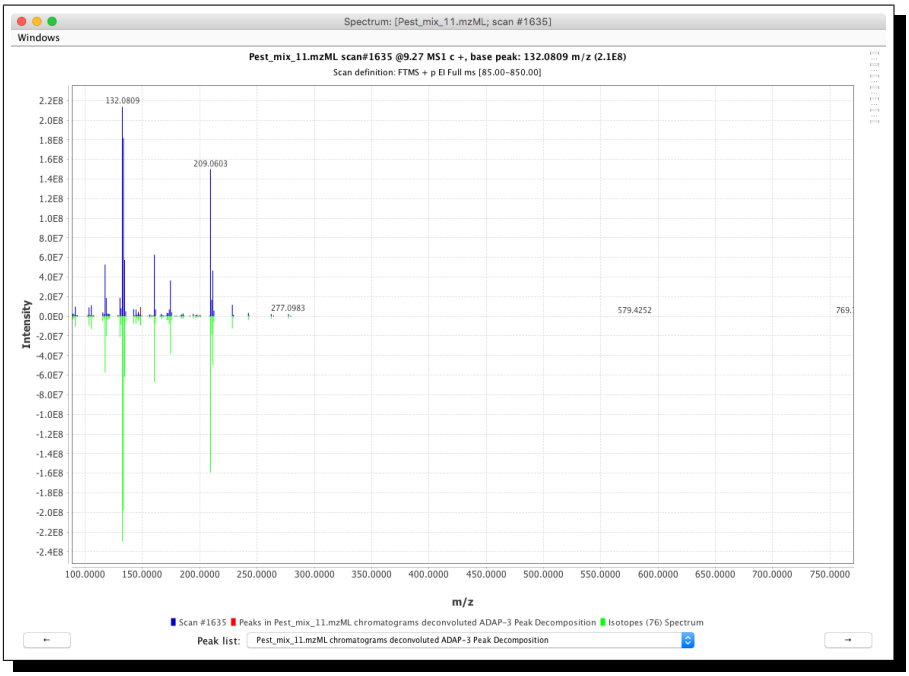


Figure 33: Mass spectra constructed by the decomposition algorithm.

4.4 Spectra Export

The mass spectra that have been constructed can be exported in .msp format and then imported to *NIST MS Search* for identification. To export the spectra, select the *Aligned peak list* and then click *Peak list methods* → *Export/Import* → *Export to MSP file* as shown in Figure 34.

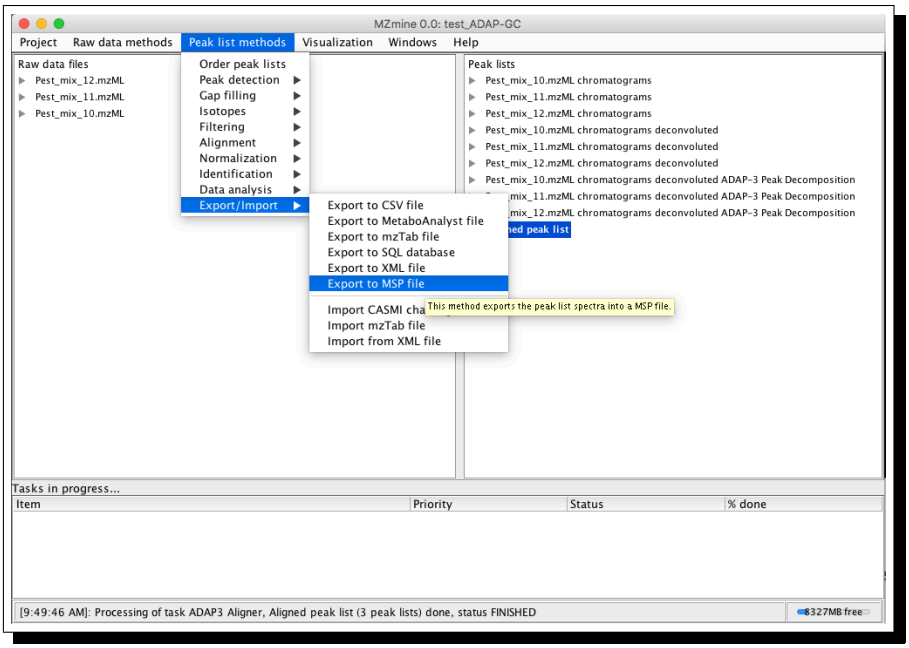


Figure 34: Export mass spectra to a MSP file.

A window as shown in Figure 35 will pull up. You will need to choose a location and file name for the .msp file, check whether or not to round the m/z values for searching against unit-mass spectral libraries, and the merging mode when rounding is selected (i.e. two or more peaks exist within a 1 dalton window).

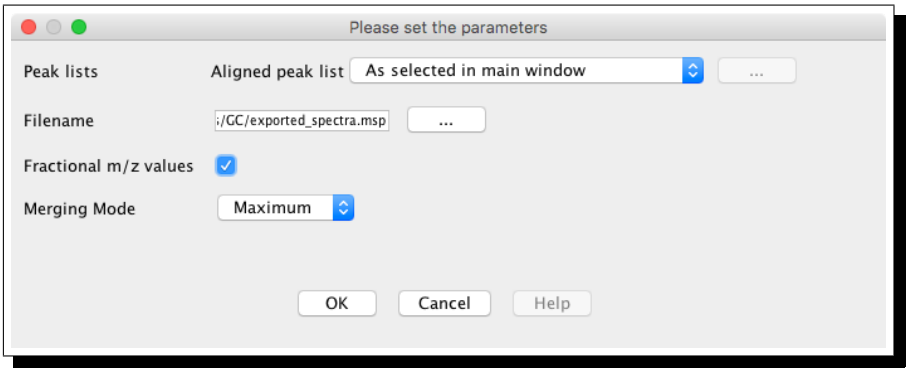


Figure 35: Export mass spectra to a MSP file.

Open the exported .msp file in a text editor. You will see that the mass spectra after alignment have been exported. Figure 36 shows a small portion of the .msp file.

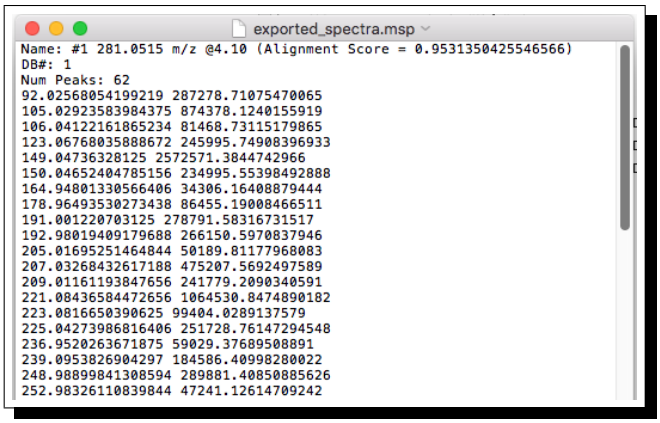


Figure 36: Example .msp file exported by ADAP-GC.

The constructed mass spectra can also be exported in .mgf format. To do so, select the *Aligned peak list* and then click *Peak list methods* → *Export/Import* → *Export to MGF file* as shown in Figure 37.

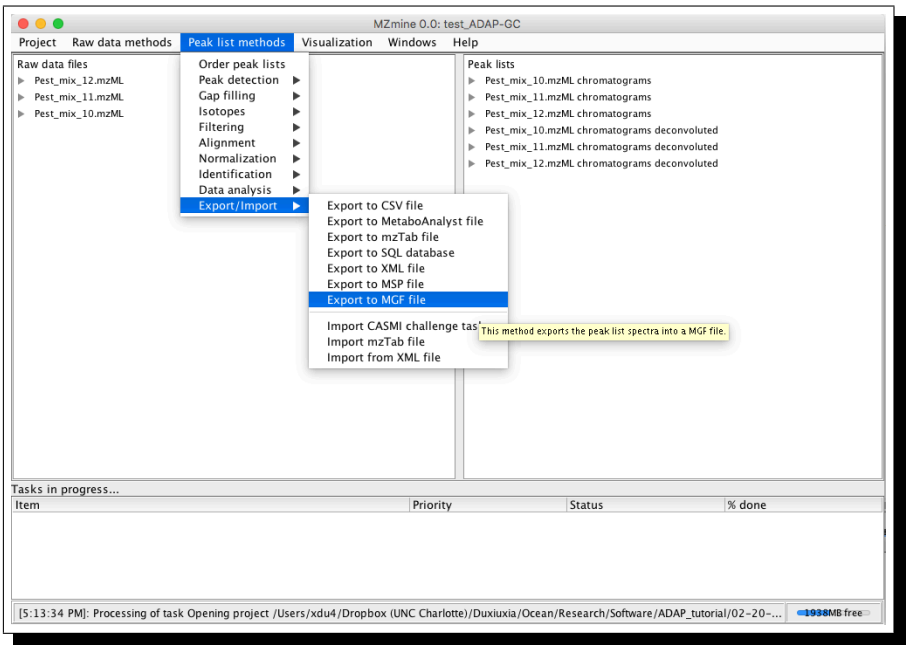


Figure 37: Export mass spectra to a MGF file.

A window as shown in Figure 38 is open allowing you to name the export file.

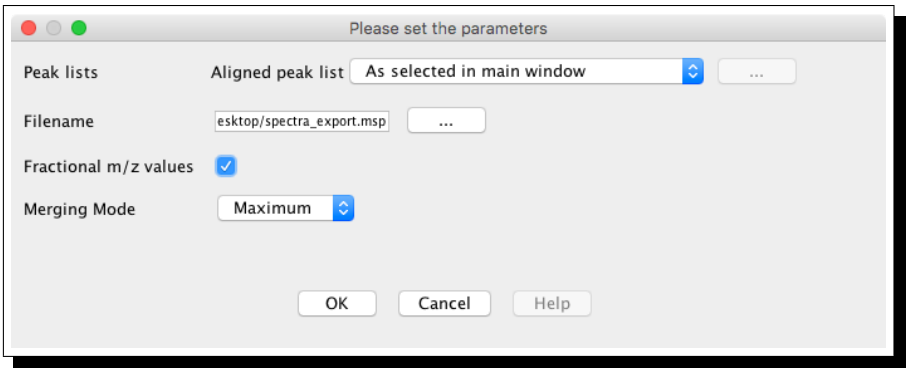


Figure 38: Export mass spectra to a MGF file.

Figure 39 shows part of a .mgf file exported from MZmine 2.

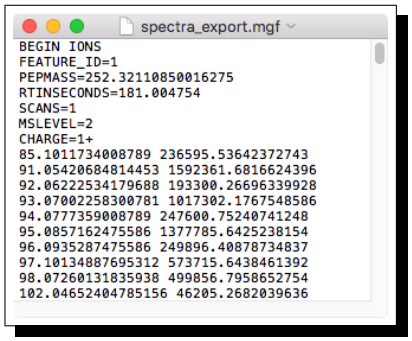


Figure 39: Example .mgf file exported by ADAP.

5 List of Additions and Changes Du-lab Team Made to MZmine 2

For details about the following changes and addition, please refer to the main text of the tutorial.

- Category: *Raw data methods* → *Peak detection*
  - **Mass detection**: added *Filename* for choosing the directory and filename to output detected masses to. The checkbox allows the user to choose if they would like to output this file or not.
  - **ADAP Chromatogram builder**: a new method of chromatogram building.
- Category: *Peak list methods* → *Peak Detection*
  - **Chromatogram Deconvolution**: **Wavelets (ADAP)**.
  - **Spectral deconvolution**: a new method for pre-processing GC-MS data by detecting analytes and constructing their fragmentation spectra.
- Category: *Peak list methods* → *Identification*
  - **CAMERA search**: Modified CAMERA search.
- Category: *Peak list methods* → *Export / Import*
  - **Export to MSP file**: exporting constructed spectra to a file in MSP format
  - **Export to MGF file**: exporting constructed spectra to a file in MGF format
- Category: *Visualization*
  - **Point 2D visualizer**: Heat map visualization of intensities in RT and m/z domain.

References

[1] **MZmine 2** [http://mzmine.github.io/]

[2] Kuhl C, Tautenhahn R, Bottcher C, Larson TR, Neumann S: `CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets.` *Anal Chem* 2012, 84(1):283-289.

[3] **CAMERA** [https://bioconductor.org/packages/release/bioc/html/CAMERA.html]

[4] Jiang W, Qiu Y, Ni Y, Su M, Jia W, Du X: **An automated data analysis pipeline for GC-TOF-MS metabonomics studies.** *J Proteome Res* 2010, 9(11):5974-5981.

[5] Ni Y, Qiu Y, Jiang W, Suttlemyre K, Su M, Zhang W, Jia W, Du X: **ADAP-GC 2.0: deconvolution of coeluting metabolites from GC/TOF-MS data for metabolomics studies.** *Anal Chem* 2012, 84(15):6619-6629.

[6] Ni Y, Su M, Qiu Y, Jia W, Du X: **ADAP-GC 3.0: Improved Peak Detection and Deconvolution of Co-eluting Metabolites from GC/TOF-MS Data for Metabolomics Studies.** *Anal Chem* 2016, 88(17):8802-8811.