ADAP
User Manual

Version 4.0.0

Du-Lab Team

Department of Bioinformatics and Genomics

University of North Carolina at Charlotte

xiuxia.du@uncc.edu

http://www.du-lab.org

April 15, 2019

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].

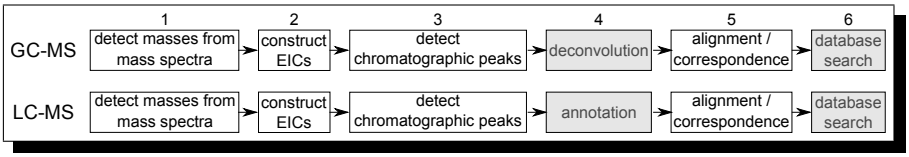


Figure 1: Workflows for pre-processing GC- and LC-MS data.

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.

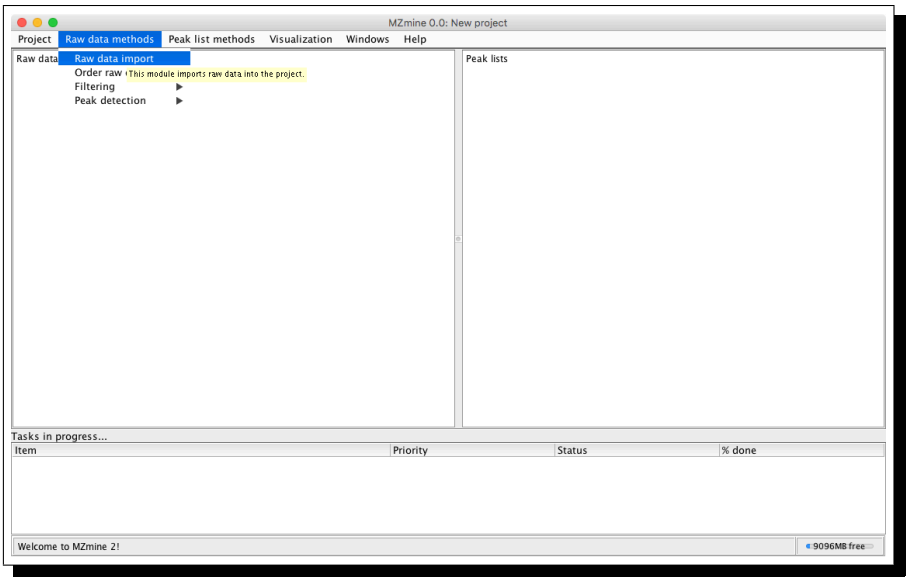


Figure 2: Import the raw data file.

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.

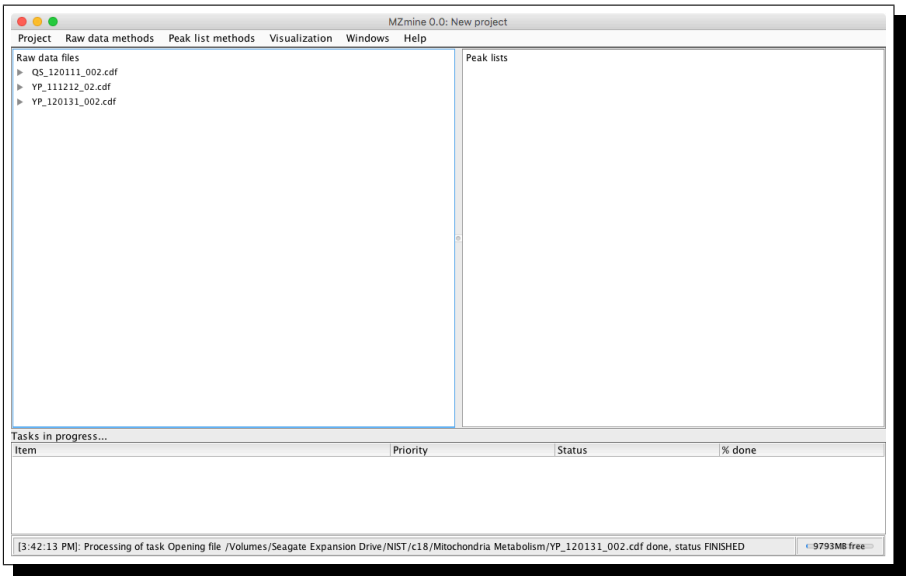


Figure 3: Imported data files.

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.

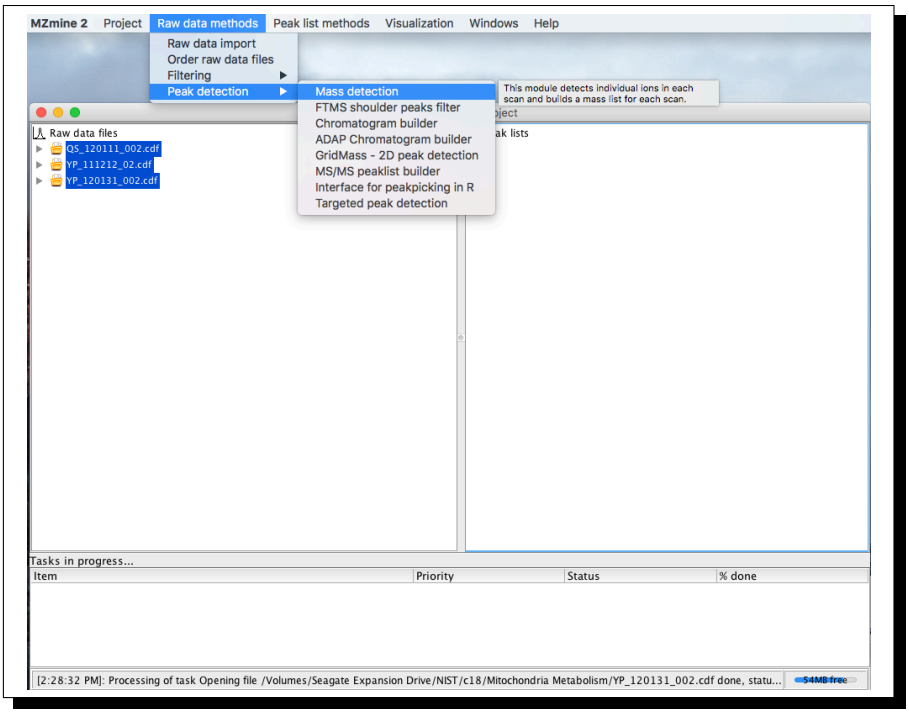


Figure 4: Mass detection from profile mass spectra

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.

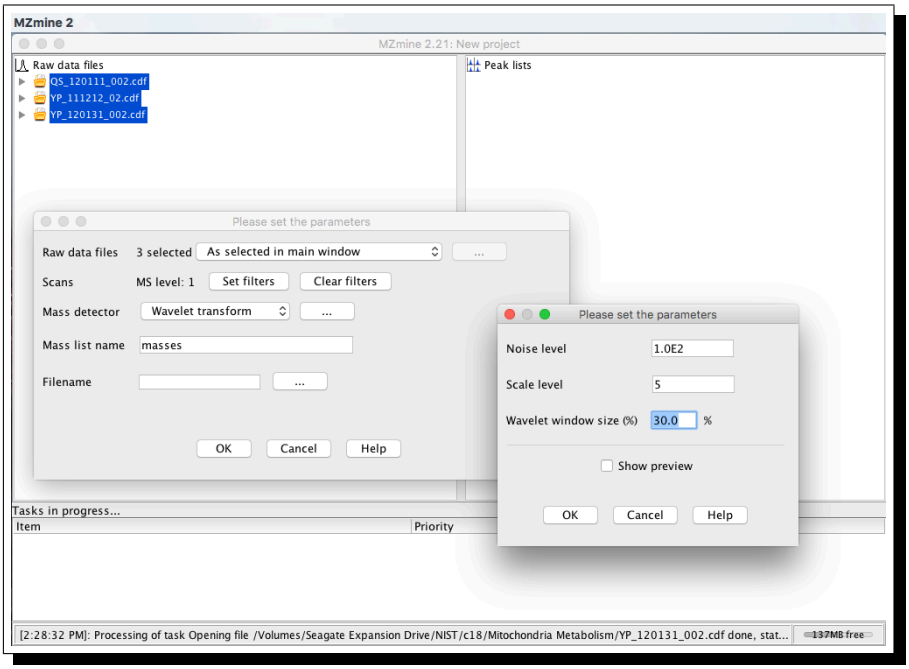


Figure 5: Mass detection by continuous wavelet transform.

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.

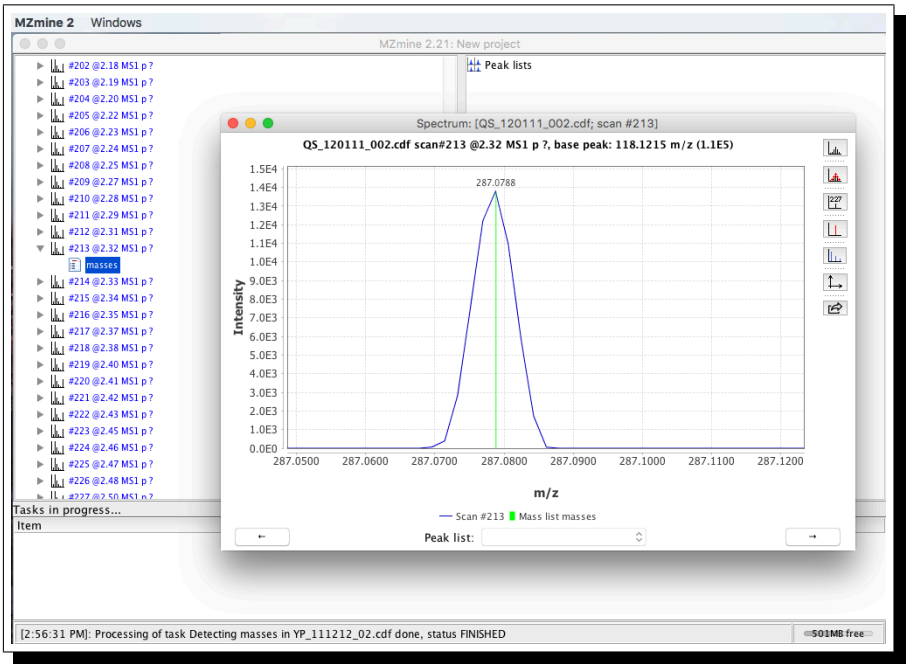


Figure 6: Mass detection result.

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.

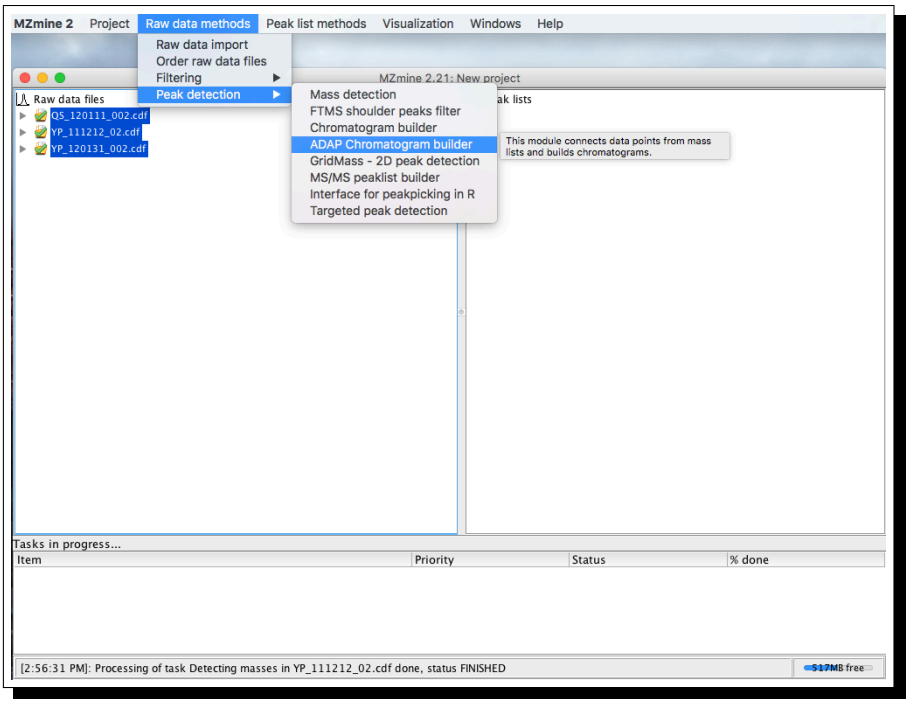


Figure 7: Selecting the ADAP chromatogram building.

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.

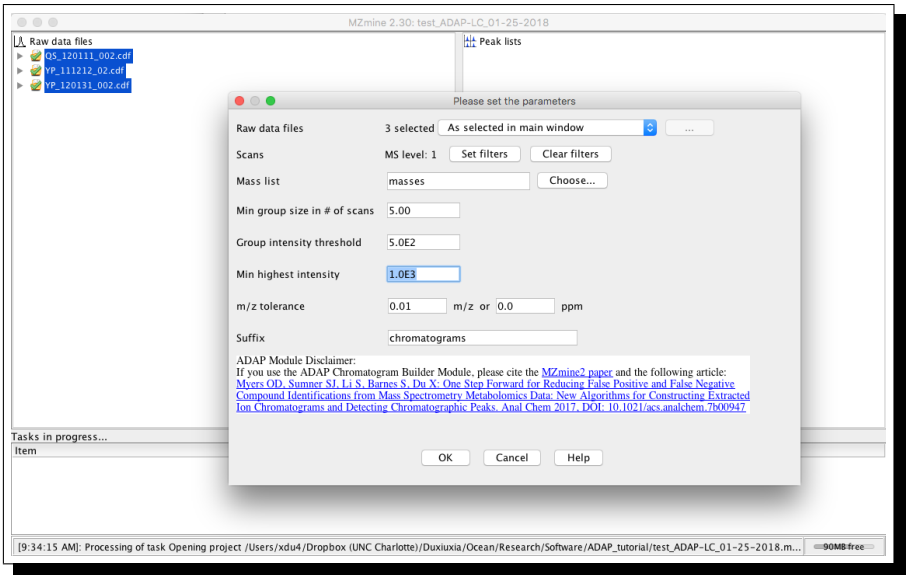


Figure 8: Example of ADAP chromatogram building parameters.

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.

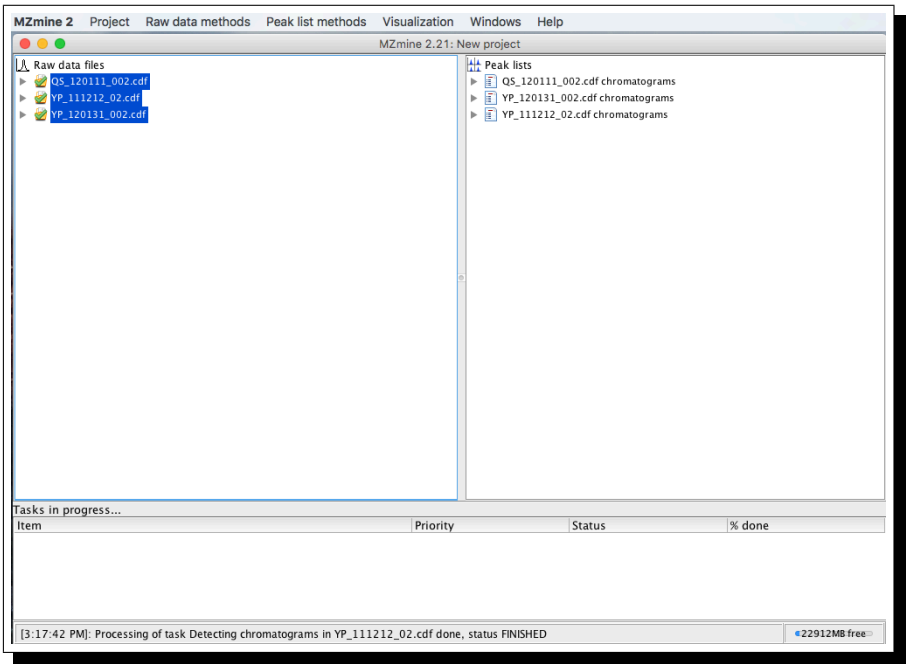


Figure 9: Results of chromatogram building.

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

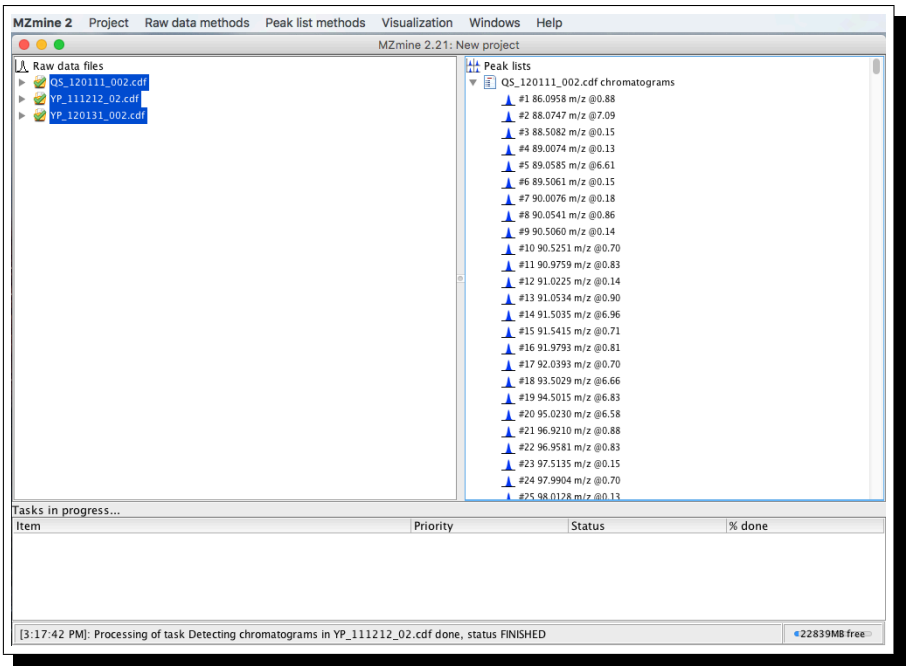


Figure 10: List of EICs that have been constructed.

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.

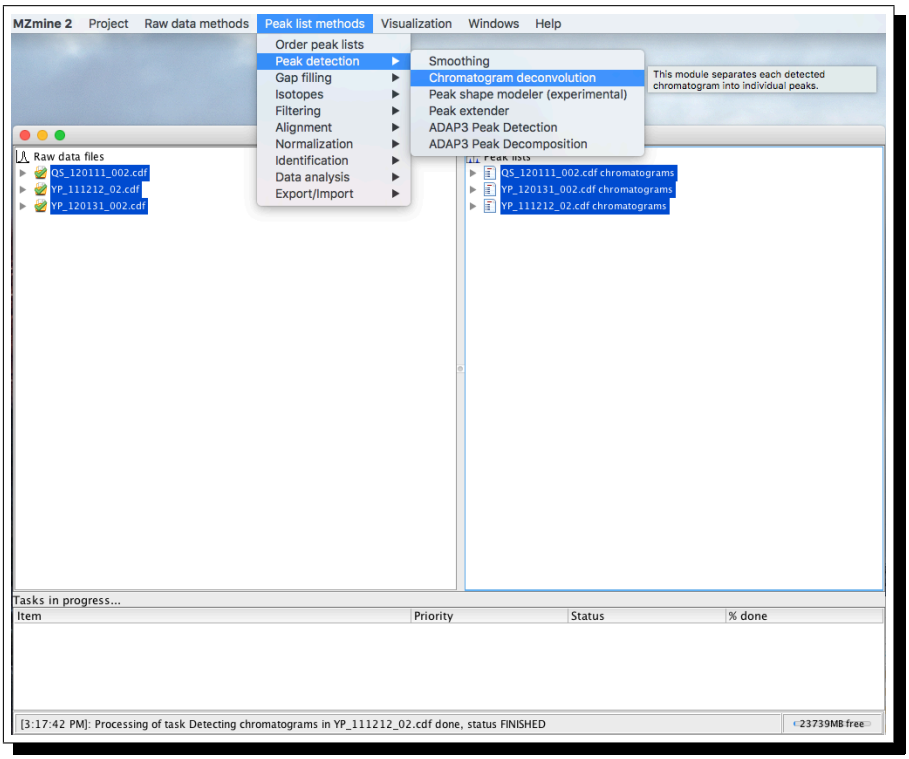


Figure 11: Detect peaks from EICs.

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.

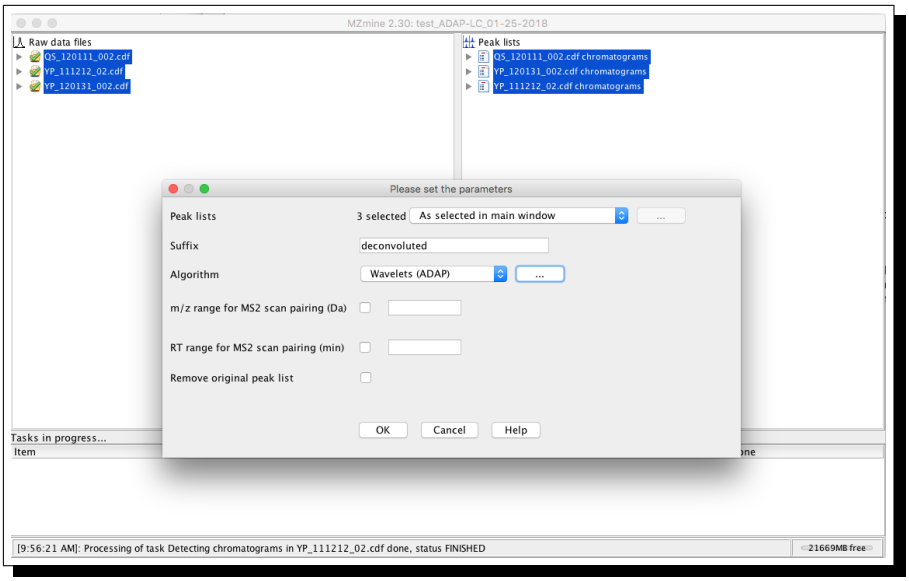


Figure 12: Select ADAP peak detection.

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.

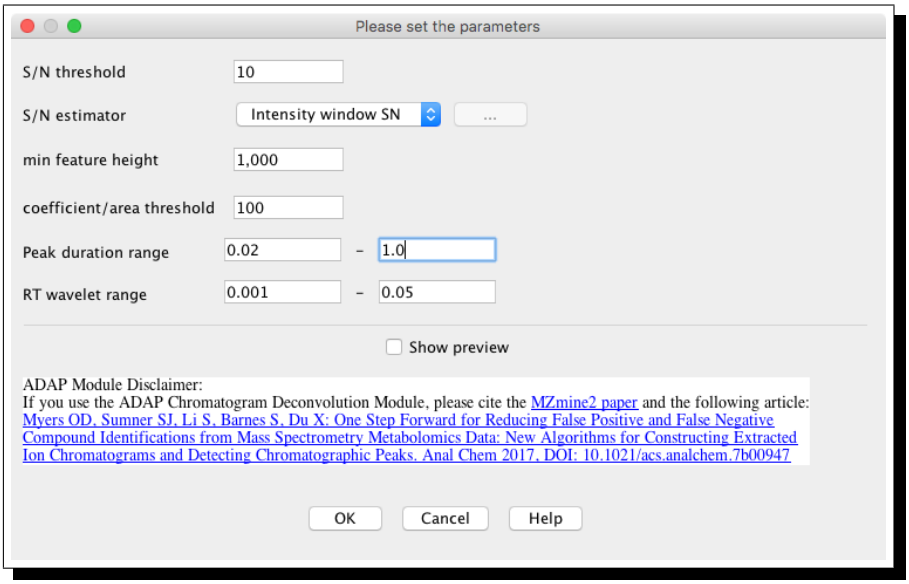


Figure 13: EIC peak detection parameters.

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.

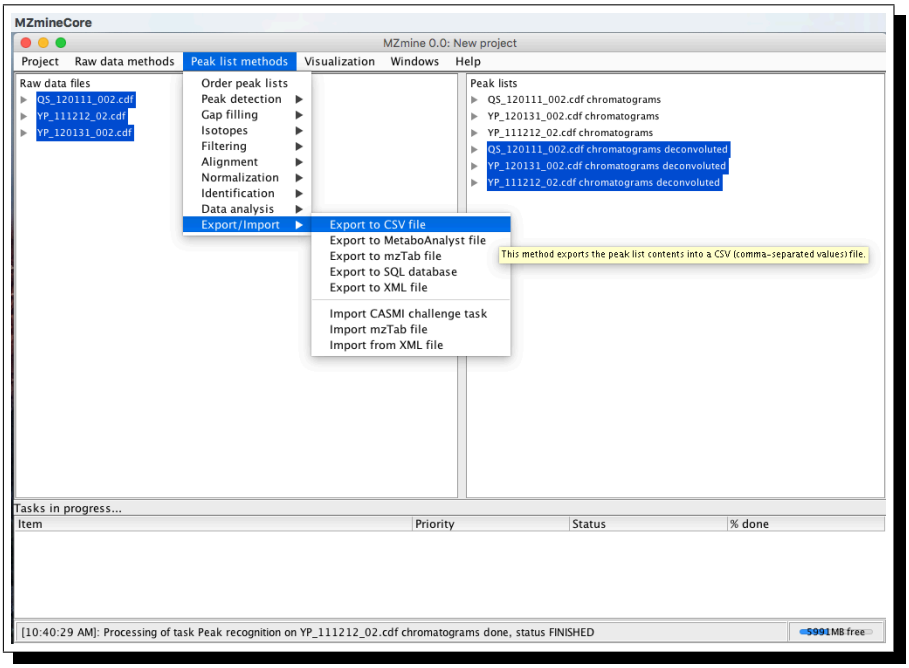
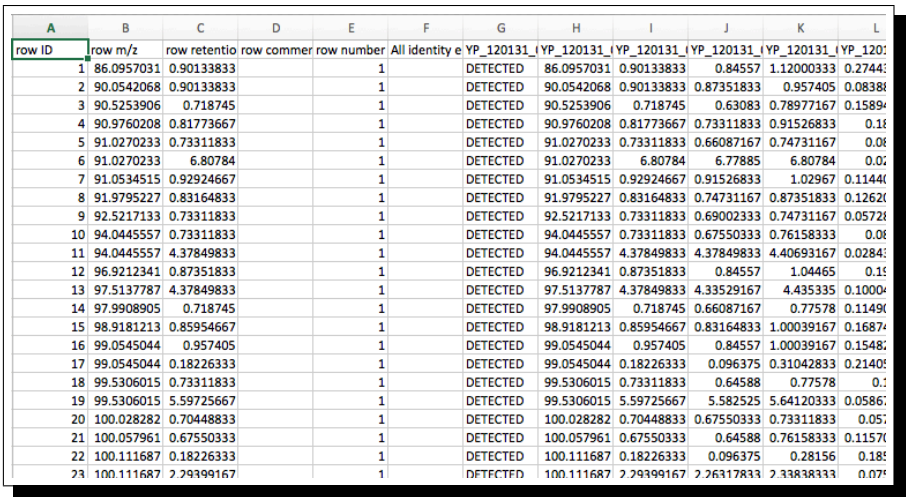


Figure 14: Export results from chromatographic peak detection.



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).

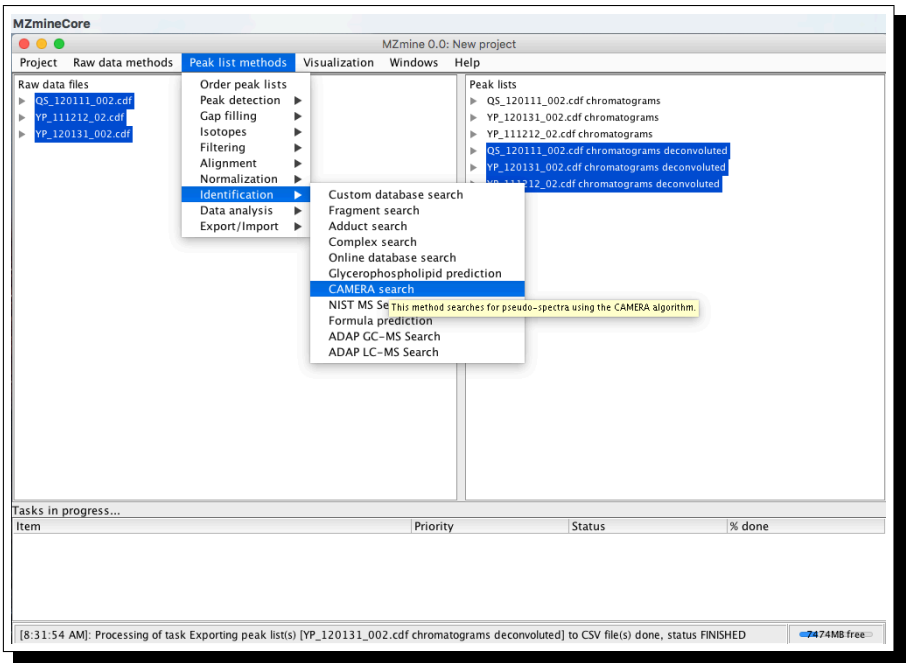


Figure 16: Use CAMERA for annotation of EIC peaks.

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

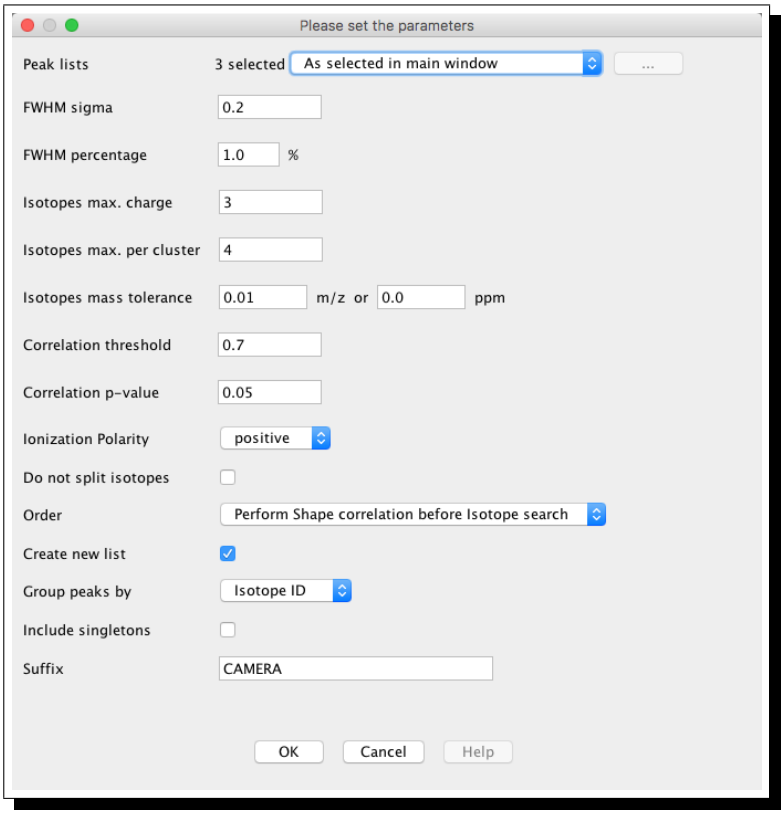


Figure 17: Specify parameters for CAMERA.

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.

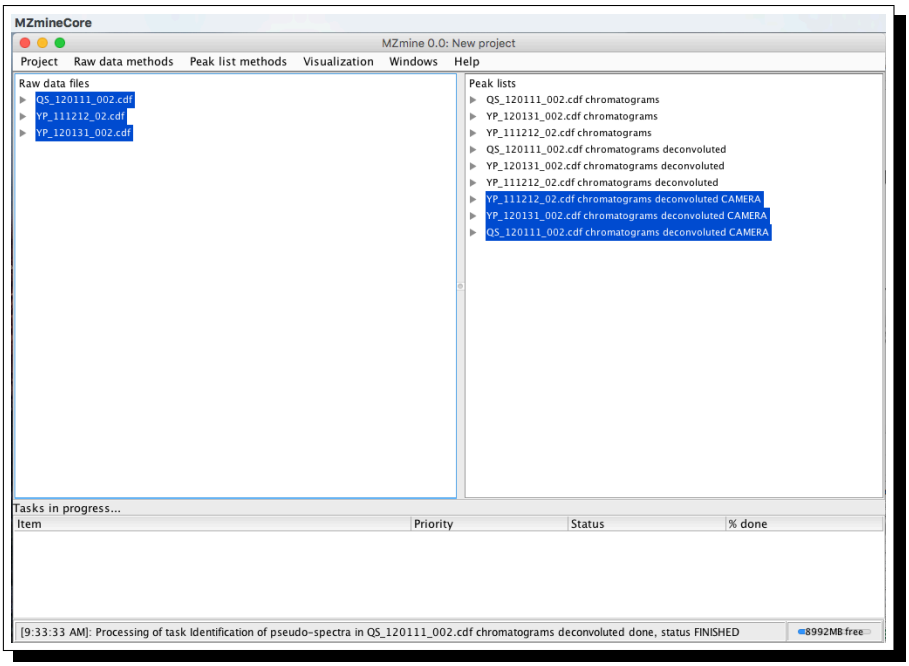


Figure 18: CAMERA finishes searches and results are displayed.

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*.

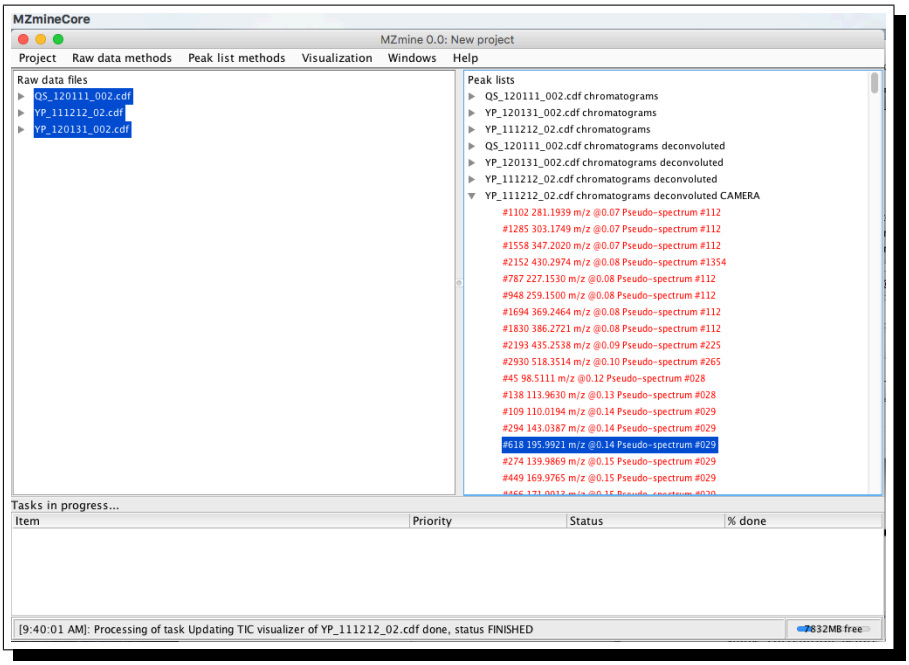


Figure 19: List of pseudo-spectra are displayed.

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

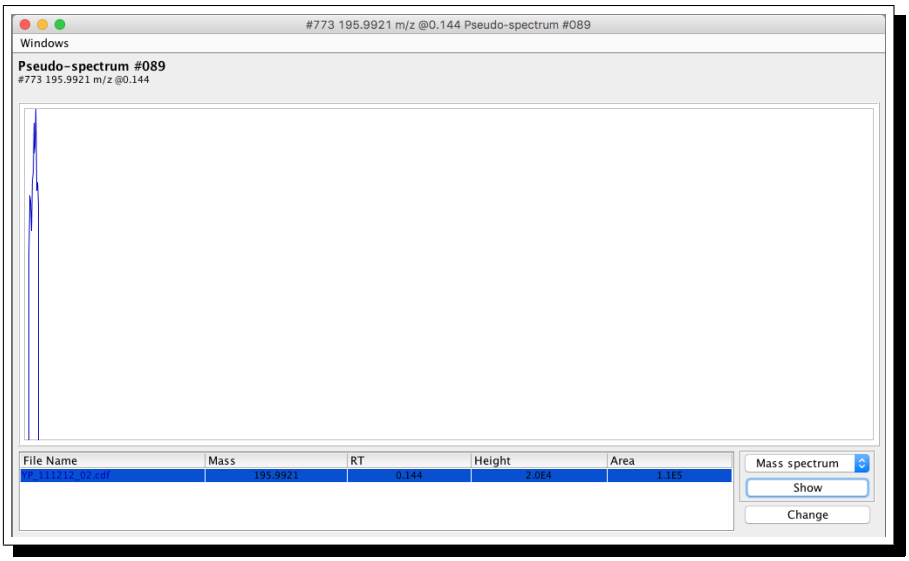


Figure 20: First step of visualizing a pseudo spectrum.

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).

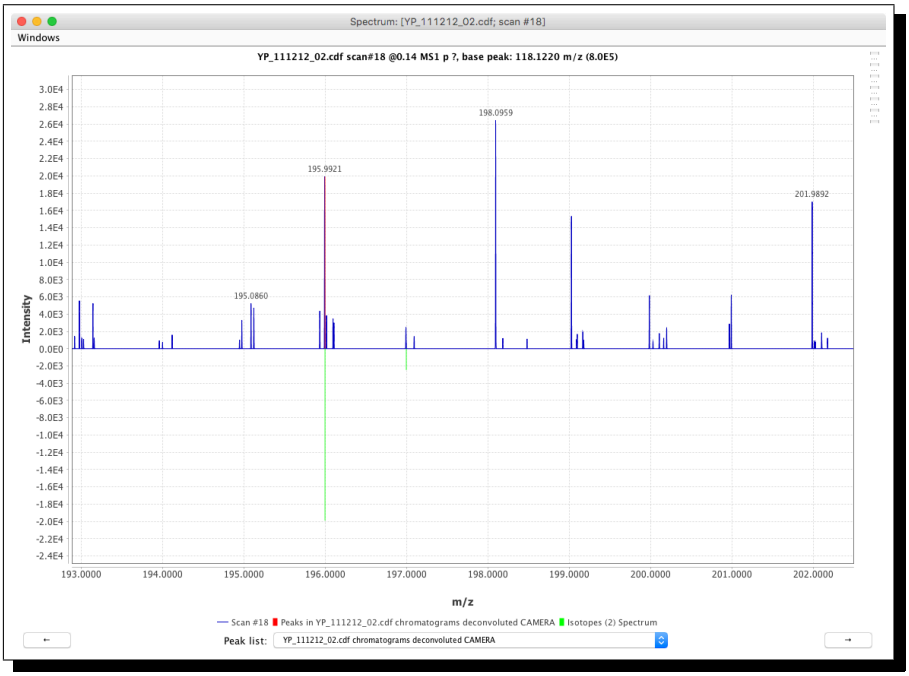


Figure 21: Second step of visualizing a pseudo spectrum.

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.

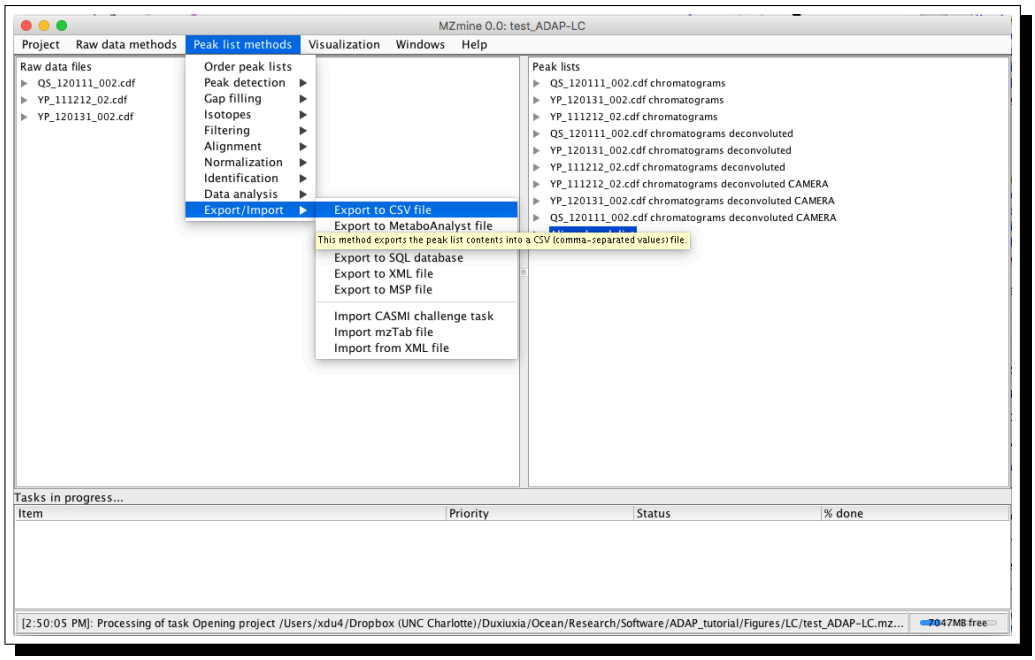


Figure 22: Export results.

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

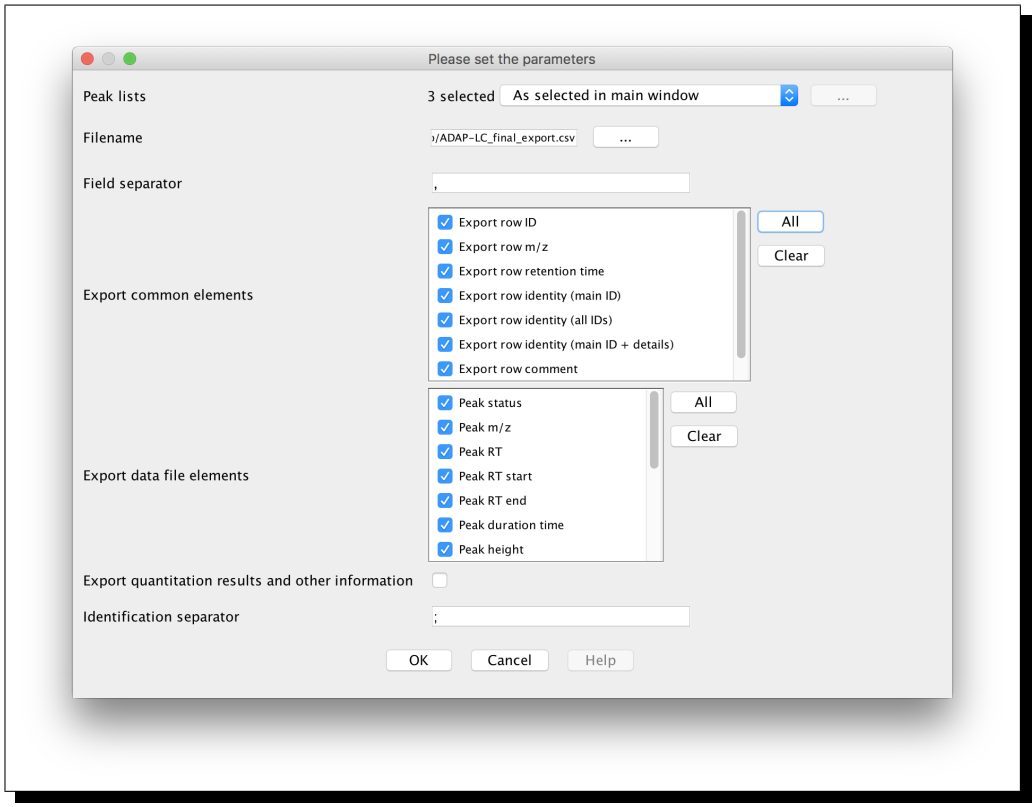


Figure 23: Select what to be exported to a CSV file.

Figure 24 shows part of the exported results. The CAMERA results can be found in the column "row identity (main ID + details)".

row ID	row m/z	row retention time	row identity (main ID)	row identity (all IDs)	row identity (main ID + details)	row i	YP_111212_0	YP_111212_02.i
1314	303.1749268	0.07006	Pseudo-spectrum #131	Pseudo-spectrum #131	Name: Pseudo-spectrum #131;Isotope: [44][M]+;Adduct: [M+Na]+ 280.186;Identification method: Bioconductor CAMERA	1	DETECTED	303.1749268
1632	347.2019653	0.07006	Pseudo-spectrum #131	Pseudo-spectrum #131	Name: Pseudo-spectrum #131;Isotope: [61][M]+;Adduct: [M+Na]+ 324.213;Identification method: Bioconductor CAMERA	1	DETECTED	347.2019653
1593	342.2467957	0.078593333	Pseudo-spectrum #150	Pseudo-spectrum #150	Name: Pseudo-spectrum #150;Isotope: [57][M]+;Adduct: [M+H+NH3]+ 324.212;Identification method: Bioconductor CAMERA	1	DETECTED	342.2467957
1780	369.2464294	0.078593333	Pseudo-spectrum #150	Pseudo-spectrum #150	Name: Pseudo-spectrum #150;Isotope: [68][M]+;Adduct: [M+H]+ 368.238;Identification method: Bioconductor CAMERA	1	DETECTED	369.2464294
1930	386.2721252	0.078593333	Pseudo-spectrum #150	Pseudo-spectrum #150	Name: Pseudo-spectrum #150;Isotope: [79][M]+;Adduct: [M+H+NH3]+ 368.238;Identification method: Bioconductor CAMERA	1	DETECTED	386.2721252
1970	391.2273865	0.078593333	Pseudo-spectrum #150	Pseudo-spectrum #150	Name: Pseudo-spectrum #150;Isotope: [81][M]+;Adduct: [M+Na]+ 368.238;Identification method: Bioconductor CAMERA	1	DETECTED	391.2273865
3162	518.3513794	0.103911667	Pseudo-spectrum #316	Pseudo-spectrum #316	Name: Pseudo-spectrum #316;Isotope: [163][M]+;Identification method: Bioconductor CAMERA	1	DETECTED	518.3513794
938	259.1500244	0.160346667	Pseudo-spectrum #135	Pseudo-spectrum #135	Name: Pseudo-spectrum #135;Isotope: [23][M]+;Adduct: [M+Na]+ 236.161;Identification method: Bioconductor CAMERA	1	DETECTED	259.1500244
2783	488.341217	0.168388333	Pseudo-spectrum #147	Pseudo-spectrum #147	Name: Pseudo-spectrum #147;Isotope: [138][M]+;Adduct: [M+H+NH3]+ 470.305;Identification method: Bioconductor CAMERA	1	DETECTED	488.341217
2826	493.2962036	0.168388333	Pseudo-spectrum #147	Pseudo-spectrum #147	Name: Pseudo-spectrum #147;Isotope: [140][M]+;Adduct: [M+Na]+ 470.305;Identification method: Bioconductor CAMERA	1	DETECTED	493.2962036
1417	317.1924744	0.168388333	Pseudo-spectrum #1757	Pseudo-spectrum #1757	Name: Pseudo-spectrum #1757;Isotope: [47][M]+;Identification method: Bioconductor CAMERA	1	DETECTED	317.1924744
1731	361.216217	0.168388333	Pseudo-spectrum #147	Pseudo-spectrum #147	Name: Pseudo-spectrum #147;Isotope: [63][M]+;Adduct: [M+Na]+ 338.228;Identification method: Bioconductor CAMERA	1	DETECTED	361.216217
3347	532.3635864	0.168388333	Pseudo-spectrum #147	Pseudo-spectrum #147	Name: Pseudo-spectrum #147;Isotope: [175][M]+;Adduct: [M+H+NH3]+ 514.333;Identification method: Bioconductor CAMERA	1	DETECTED	532.3635864
4324	664.4437866	0.176581667	Pseudo-spectrum #366	Pseudo-spectrum #366	Name: Pseudo-spectrum #366;Isotope: [223][M]+;Identification method: Bioconductor CAMERA	1	DETECTED	664.4437866
401	166.1107025	0.209265	Pseudo-spectrum #024	Pseudo-spectrum #024	Name: Pseudo-spectrum #024;Isotope: [9][M]2+;Adduct: [M+2H]2+ 330.207;Identification method: Bioconductor CAMERA	1	DETECTED	166.1107025
1514	331.2145081	0.209265	Pseudo-spectrum #024	Pseudo-spectrum #024	Name: Pseudo-spectrum #024;Isotope: [54][M]+;Adduct: [M+H]+ 330.207;Identification method: Bioconductor CAMERA	1	DETECTED	331.2145081
1301	301.1683655	0.21757	Pseudo-spectrum #024	Pseudo-spectrum #024	Name: Pseudo-spectrum #024;Isotope: [43][M]+;Adduct: [M+Na]+ 278.184;Identification method: Bioconductor CAMERA	1	DETECTED	301.1683655
2148	414.295929	0.292796667	Pseudo-spectrum #2027	Pseudo-spectrum #2027	Name: Pseudo-spectrum #2027;Isotope: [92][M]+;Identification method: Bioconductor CAMERA	1	DETECTED	414.295929
1418	317.1924744	0.301091667	Pseudo-spectrum #2026	Pseudo-spectrum #2026	Name: Pseudo-spectrum #2026;Isotope: [48][M]+;Identification method: Bioconductor CAMERA	1	DETECTED	317.1924744
1840	375.2325439	0.301091667	Pseudo-spectrum #2024	Pseudo-spectrum #2024	Name: Pseudo-spectrum #2024;Isotope: [72][M]+;Adduct: [M+Na]+ 352.244;Identification method: Bioconductor CAMERA	1	DETECTED	375.2325439
964	265.134613	0.309498333	Pseudo-spectrum #110	Pseudo-spectrum #110	Name: Pseudo-spectrum #110;Isotope: [25][M]+;Identification method: Bioconductor CAMERA	1	DETECTED	265.134613
1040	273.1654358	0.309498333	Pseudo-spectrum #110	Pseudo-spectrum #110	Name: Pseudo-spectrum #110;Isotope: [29][M]+;Identification method: Bioconductor CAMERA	1	DETECTED	273.1654358
2949	502.3528137	0.428845	Pseudo-spectrum #353	Pseudo-spectrum #353	Name: Pseudo-spectrum #353;Isotope: [149][M]+;Identification method: Bioconductor CAMERA	1	DETECTED	502.3528137
2545	458.3284302	0.446018333	Pseudo-spectrum #2208	Pseudo-spectrum #2208	Name: Pseudo-spectrum #2208;Isotope: [126][M]+;Identification method: Bioconductor CAMERA	1	DETECTED	458.3284302
1149	285.1659241	0.454478333	Pseudo-spectrum #254	Pseudo-spectrum #254	Name: Pseudo-spectrum #254;Isotope: [36][M]+;Adduct: [M+Na]+ 262.176;Identification method: Bioconductor CAMERA	1	DETECTED	285.1659241
2988	504.3412476	0.46305	Pseudo-spectrum #2209	Pseudo-spectrum #2209	Name: Pseudo-spectrum #2209;Isotope: [152][M]+;Identification method: Bioconductor CAMERA	1	DETECTED	504.3412476
1839	375.2325439	0.471685	Pseudo-spectrum #306	Pseudo-spectrum #306	Name: Pseudo-spectrum #306;Isotope: [71][M]+;Adduct: [M+Na]+ 352.244;Identification method: Bioconductor CAMERA	1	DETECTED	375.2325439
506	182.9840851	0.602906667	Pseudo-spectrum #045	Pseudo-spectrum #045	Name: Pseudo-spectrum #045;Isotope: [12][M]+;Adduct: [M+Na+NaCOOH]+ 92.0052;Identification method: Bioconductor CAMERA	1	DETECTED	182.9840851
3979	614.3139038	0.645143333	Pseudo-spectrum #049	Pseudo-spectrum #049	Name: Pseudo-spectrum #049;Isotope: [209][M]3+;Identification method: Bioconductor CAMERA	1	DETECTED	614.3139038
7173	1106.085693	0.645143333	Pseudo-spectrum #049	Pseudo-spectrum #049	Name: Pseudo-spectrum #049;Isotope: [334][M]2+;Identification method: Bioconductor CAMERA	1	DETECTED	1106.085693
4585	714.5251465	0.645143333	Pseudo-spectrum #049	Pseudo-spectrum #049	Name: Pseudo-spectrum #049;Isotope: [233][M]3+;Identification method: Bioconductor CAMERA	1	DETECTED	714.5251465
4618	717.0214844	0.645143333	Pseudo-spectrum #049	Pseudo-spectrum #049	Name: Pseudo-spectrum #049;Isotope: [235][M]3+;Identification method: Bioconductor CAMERA	1	DETECTED	717.0214844
4661	722.1855469	0.645143333	Pseudo-spectrum #049	Pseudo-spectrum #049	Name: Pseudo-spectrum #049;Isotope: [237][M]3+;Adduct: [2M+Na+2K]3+ 1032.82;Identification method: Bioconductor CAMERA	1	DETECTED	722.1855469
3294	527.1071777	0.645143333	Pseudo-spectrum #049	Pseudo-spectrum #049	Name: Pseudo-spectrum #049;Isotope: [168][M]3+;Identification method: Bioconductor CAMERA	1	DETECTED	527.1071777
4113	630.3443604	0.653176667	Pseudo-spectrum #049	Pseudo-spectrum #049	Name: Pseudo-spectrum #049;Isotope: [214][M]3+;Identification method: Bioconductor CAMERA	1	DETECTED	630.3443604
4431	682.0994263	0.653176667	Pseudo-spectrum #049	Pseudo-spectrum #049	Name: Pseudo-spectrum #049;Isotope: [225][M]+;Adduct: [M+H]+ 681.096;Identification method: Bioconductor CAMERA	1	DETECTED	682.0994263
4617	716.8546753	0.653176667	Pseudo-spectrum #049	Pseudo-spectrum #049	Name: Pseudo-spectrum #049;Isotope: [234][M]3+;Adduct: [2M+2Na+K]3+ 1032.82;Identification method: Bioconductor CAMERA	1	DETECTED	716.8546753

Figure 24: Exported results.

4 ADAP-GC

Data: We use the standard mixture GC-MS data provided by Dr. Wei Jia. The data have been produced by Agilent 6890N GC System (Santa Clara, CA, USA) coupled with Pegasus HT TOF-MS (LECO Corporation, St. Joseph, MI, USA). Seven samples with each containing Piruvic Acid (at 5.17 min) and Propanoic Acid (at 5.34 min) were prepared at concentrations 0.2, 0.4, 0.6, 0.8, 1.0, 2.0 and 5.0 µg/mL. For demonstration purposes, the data files were trimmed to the retention time range 5.12—5.49 min and split into two groups: high concentration (1.0, 2.0 and 5.0 µg/mL) and low concentration (0.2, 0.4, 0.6 and 0.8 µg/mL). These data files and an MZmine 2 project file can be found at https://drive.google.com/drive/folders/1hDS1u7LeA5aF4Rg89yY351m9cHMrq_sy?usp=sharing.

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 lies in performing the deconvolution. Therefore, we will only describe in detail the deconvolution step. For a detailed description of the chromatogram construction and peak detection steps, see the corresponding sections of pre-processing LC-MS data.

4.1 Detection of Masses and Construction of EICs

The seven data files are in the centroid mode already, so the *Centroid* method in MZmine 2 will be used for mass detection. The mass detection window is invoked by *Raw data methods* → *Peak detection* → *Mass detection* and shown in Figure 25.

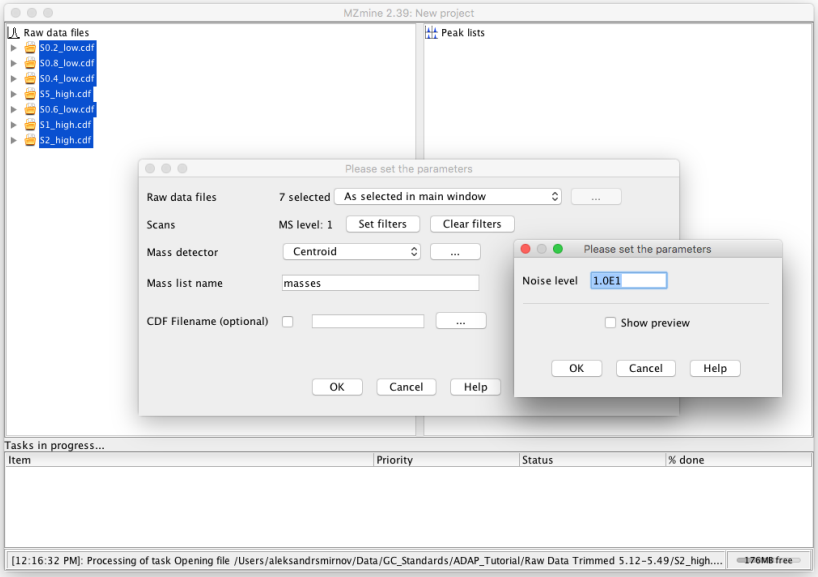


Figure 25: Mass detection of centroid data.

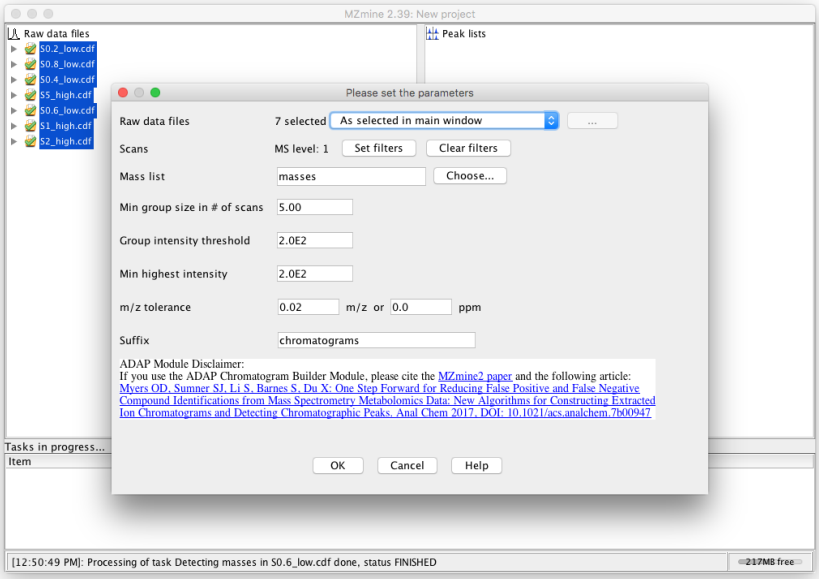


Figure 26: Example parameters for constructing EICs from GC-Orbitrap data.

The next step, construction of extracted ion chromatograms (EICs), is performed by *Raw data methods* → *Peak detection* → *ADAP Chromatogram builder*. Parameters for constructing EICs are shown in Figure 26. For parameter *Min group size in # of scans*, the value 5 is typically produces appropriate results. The next two parameters *Group intensity threshold* and *Min highest intensity* are chosen as follows: because the dynamic range of current mass spectrometry equipment is 4–5 orders of magnitude, we divide the maximum intensity of the signal 2×10^6 (see Figure 27 (left)) by the dynamic range 10^4. As a result, parameters *Group intensity threshold* and *Min highest intensity* are set to 200. Finally, parameter *m/z tolerance* is set to 0.02. The result of chromatogram construction can be observed in the (*m/z*, *ret time*) plane, using *Visualization* → *2D visualizer*, where the colored dots represent raw data points and the green lines represent constructed chromatograms (Figure 27 (right)).

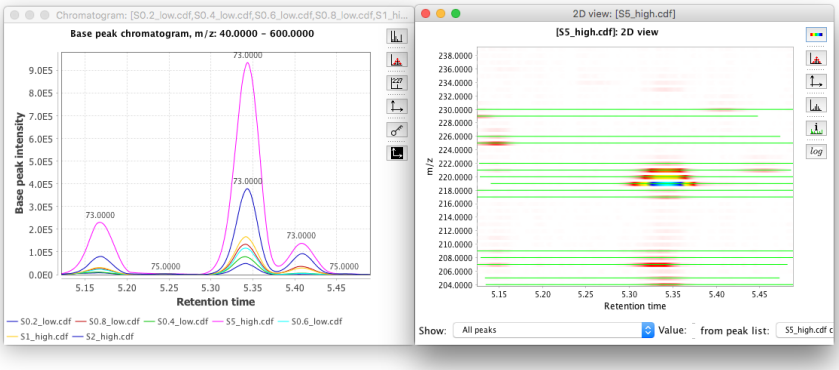


Figure 27: Total ion chromatograms (left) plotted using *Visualization* → *TIC/XIC visualizer*; and extracted ion chromatograms (right) plotted as green lines, using *Visualization* → *2D visualizer*.

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 28.

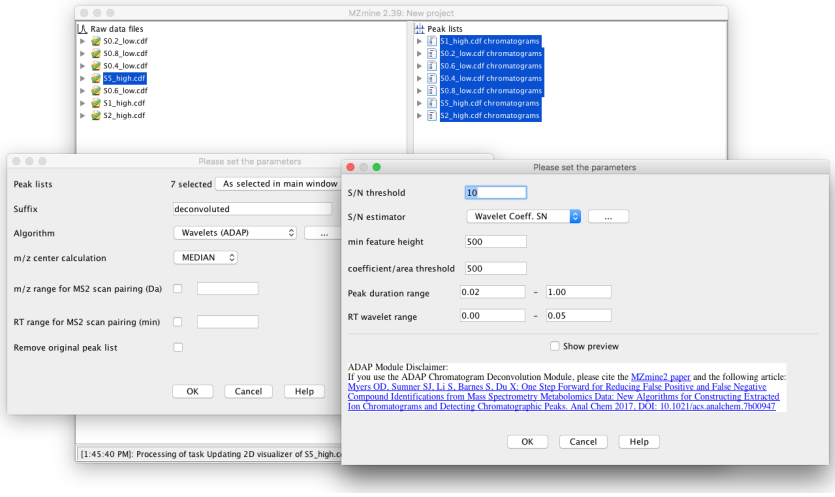


Figure 28: Select *Wavelet (ADAP)* for detecting peaks from EICs for GC-MS data.

Click the ellipse to open the parameter window. Figure 28 shows example parameters. Parameter *S/N threshold* is typically set to 10. Parameter *min feature height* is similar to parameters *Group intensity threshold* and *Min highest intensity* of ADAP Chromatogram builder, and should be chosen accordingly. Parameter *Peak duration range* defines the minimum and maximum peak width and can be estimated by looking at individual peaks (Figure 27 (left)). Finally, parameter *RT wavelet range* is somewhat challenging to estimate and is chosen by trying several different values and looking at peak-detection results in the preview pane.

The peak-detection is currently one of the most time-consuming steps in the ADAP-GC workflow. If the duration of chromatography is long, this step could take a while.

4.3 Spectral Deconvolution

In the ADAP-GC workflow, there are two spectral deconvolution algorithms available for users: **Hierarchical Clustering** and **Multivariate Curve Resolution**. Each method has certain advantages and disadvantages that are summarized in Table 1. Users may choose either one of these algorithms to perform spectral deconvolution.

Hierarchical Clustering	Multivariate Curve Resolution
Fast	Can be slow, depending on the size of deconvolution windows
Large number of user parameters	Small number of user parameters
The shape of model peaks is typically bell-like	Model peaks can have arbitrary shape

Table 1: Comparison of two spectral deconvolution methods.

4.3.1 Spectral Deconvolution / Hierarchical Clustering

Spectral Deconvolution detects 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. The first clustering combines peaks based on proximity of their retention times, while the second clustering refines the clusters by calculating the similarity of peaks' shapes.

After analytes are detected, a model peak in each cluster is chosen to represent the elution profile of the analyte. Choice of the model peak may affect the quality of the constructed fragmentation spectra. For details about the underlying algorithm, please refer to [4, 5, 6].

To perform deconvolution, select all of the chromatographic peaks detected from one or more data files, then click *Peak list methods* → *Spectral Deconvolution* → *Hierarchical clustering* as shown in Figure 29. To see preview of the deconvolution results, select *Show preview* option at the bottom of the parameter window.

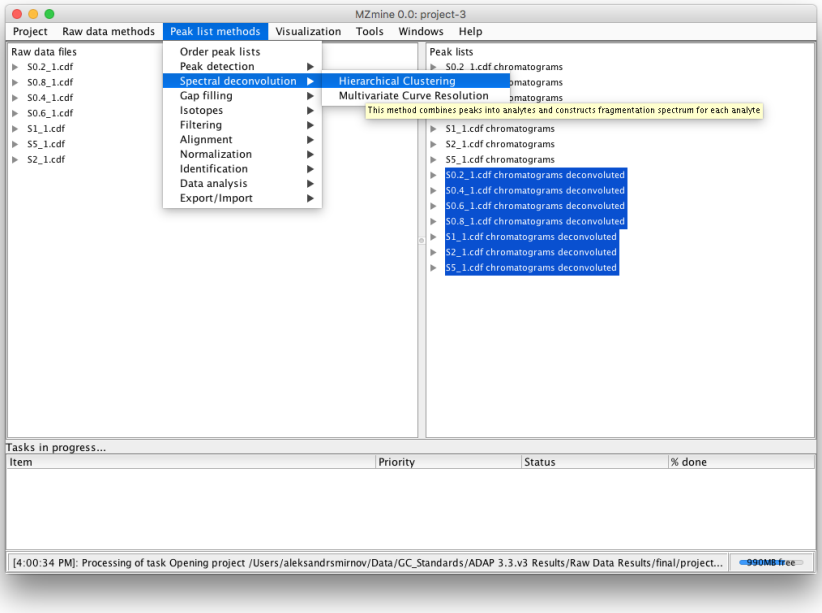


Figure 29: Deconvolution of chromatographic peaks.

A window as shown in Figure 30 pulls up allowing you to specify parameters for deconvolution.

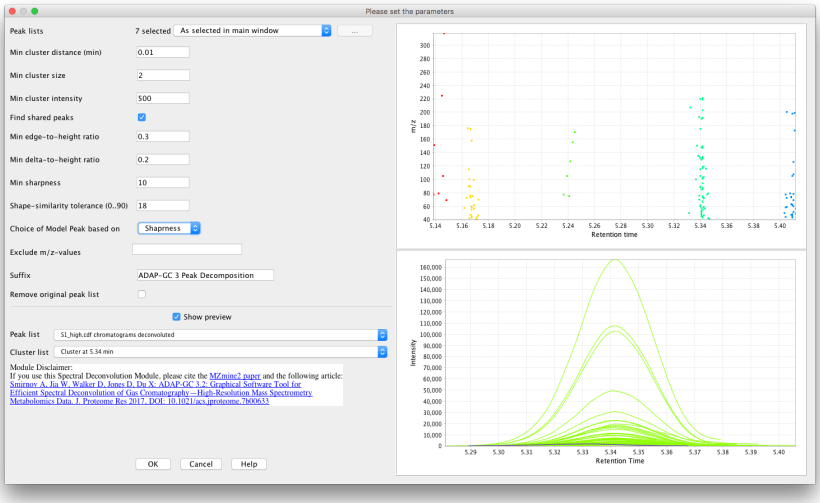


Figure 30: Specify parameters for decomposition of chromatographic peaks.

- First clustering parameters. The preview of the first clustering is displayed on the top right figure if the option *Show preview* is selected.
  - *Min cluster distance*: Minimum allowed time gap between any two clusters determined by the retention-time clustering.
  - *Min cluster size*: Minimum allowed size of a cluster determined by the retention-time clustering.
  - *Min cluster intensity*: Minimum allowed intensity of the highest peak in a cluster determined by the retention-time clustering.
- Filtering parameters. Peaks that passed the filter are displayed on the bottom right figure if the option *Show preview* is selected.
  - *Find shared peaks*: If selected, the algorithm makes an attempt to determine if a peak is shared, i.e. produced by more than one analyte. All shared peaks are filtered out. A peak is considered to be shared if (i) its chromatogram contains multiple local maxima, or (ii) the start and end intensities are sufficiently high relative to its apex intensity (see the next two parameters).
  - *Min edge-to-height ratio*. Peak is considered to be shared if the ratio of the start or end intensity to the apex intensity exceeds this value.
  - *Min delta-to-height ratio*. Peak is considered to be shared if the ratio of the difference between the start and end intensities to the apex intensity exceeds this value.
  - *Min sharpness*. All peaks with sharpness below this value are filtered out.
  - *Exclude m/z values*. Peaks with m/z from this list are filtered out. This list can be empty or contain both singular m/z values and ranges. Example: $1 - 73, 147, 221$.
- Second clustering parameters. The preview of the second clustering is displayed on the bottom right figure if the option *Show preview* is selected.

  - *Shape-similarity tolerance (0..90).* Threshold used in the second clustering. It represents the inverse cosine of the normalized dot-product of two elution profiles. Large values produce a few large clusters, while small values produce many small clusters.

- *Choice of Model Peak based on.* For each cluster, a representative model peak is chosen based on either *Sharpness* or *M/z value*. In both cases, a peak that has passed the filter and has the highest sharpness or m/z value respectively, will be chosen as the model peak for the cluster. In practice, *Shaprness* gives better quantitation results, while *M/z value* gives better identification results.

After spectral deconvolution is finished, the results are displayed as shown in Figure 31.

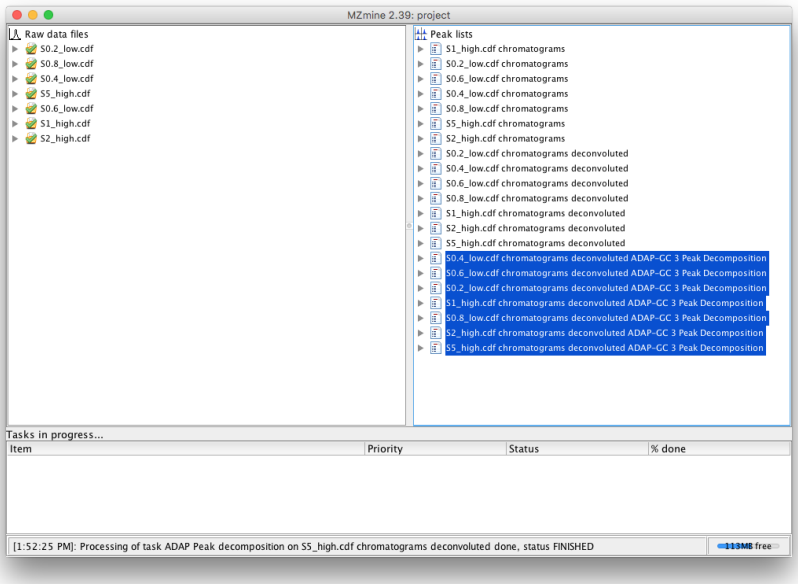


Figure 31: Decomposition results.

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 32). The m/z for each entry is the m/z of the model peak for this spectrum.

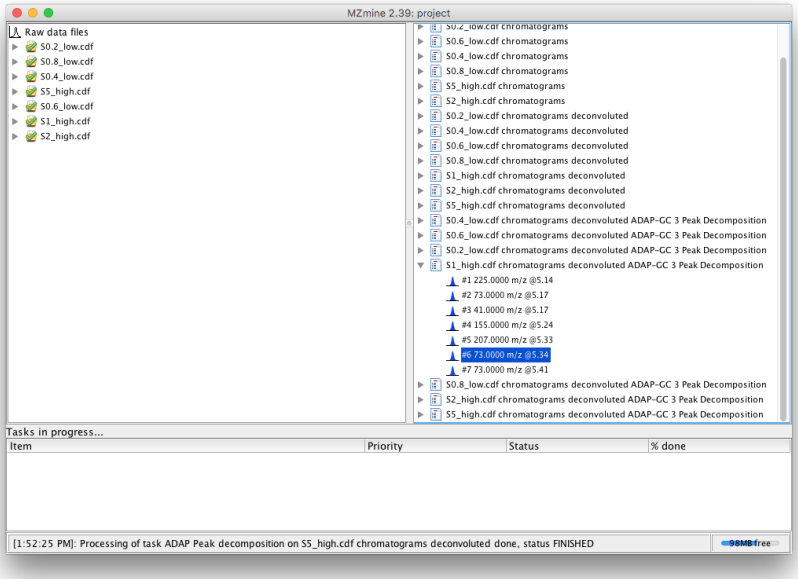


Figure 32: List of mass spectra constructed by the decomposition algorithm.

Double click on a particular mass spectrum will pull up a window as shown in Figure 33.

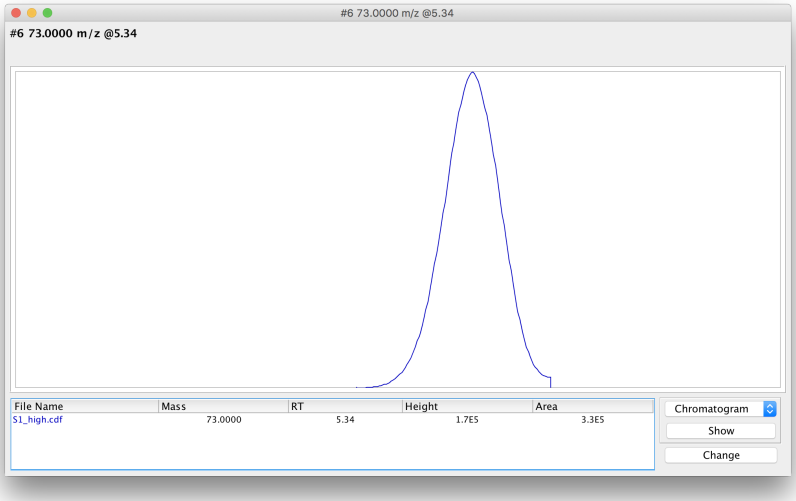


Figure 33: Peak information window.

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 34).

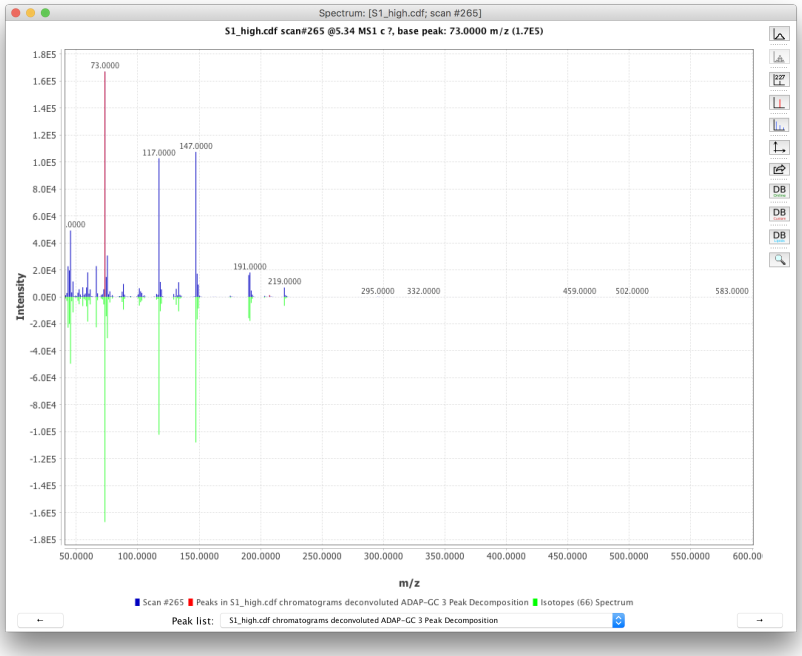


Figure 34: Mass spectra constructed by the decomposition algorithm.

4.3.2 Spectral Deconvolution / Multivariate Curve Resolution

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, click *Peak list methods* → *Spectral Deconvolution* → *Blind Source Separation* as shown in Figure 35. To see preview of the deconvolution results, select *Show preview* option at the bottom of the parameter window.

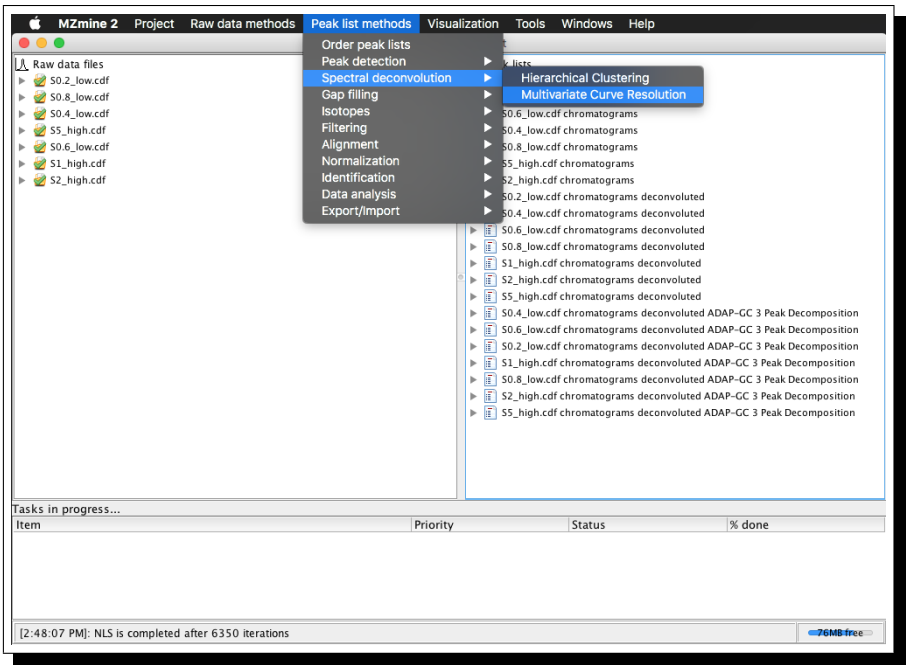


Figure 35: Deconvolution of chromatographic peaks.

A window as shown in Figure 36 pulls up allowing you to specify parameters for deconvolution.

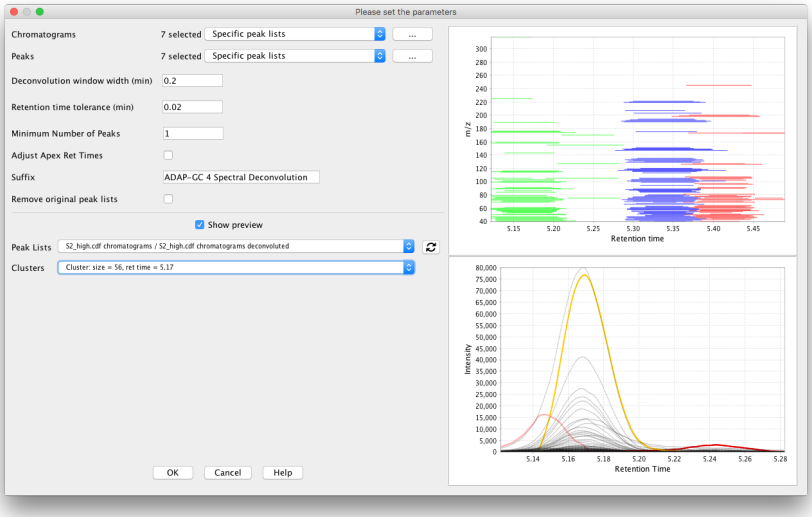


Figure 36: Specify parameters for decomposition of chromatographic peaks.

The spectral deconvolution uses both constructed chromatograms and detected peaks. The list of constructed chromatograms is specified by selecting *Specific peak lists* for parameter *Chromatograms*, clicking on the ellipsis button, and choosing one or more lists with chromatograms in the popup window (Figure 37). The list of detected peaks is specified by selecting *Specific peak lists* for parameter *Peaks*, clicking on the ellipsis button, and choosing one or more lists with detected peaks in the popup window (Figure 38).

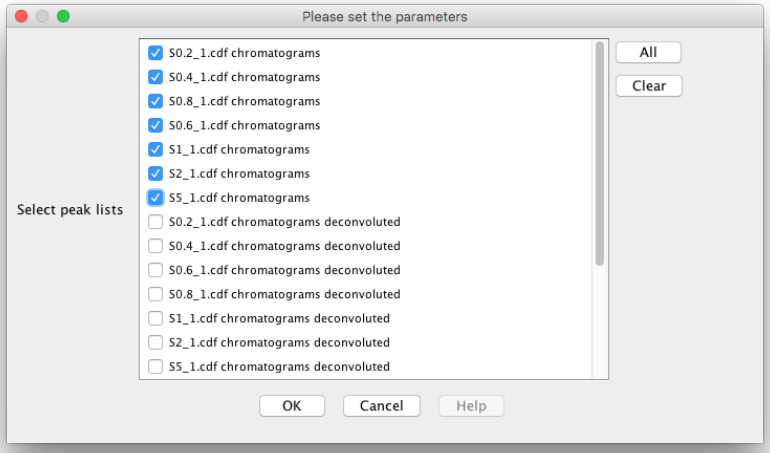


Figure 37: Choosing chromatograms for Spectral Deconvolution.

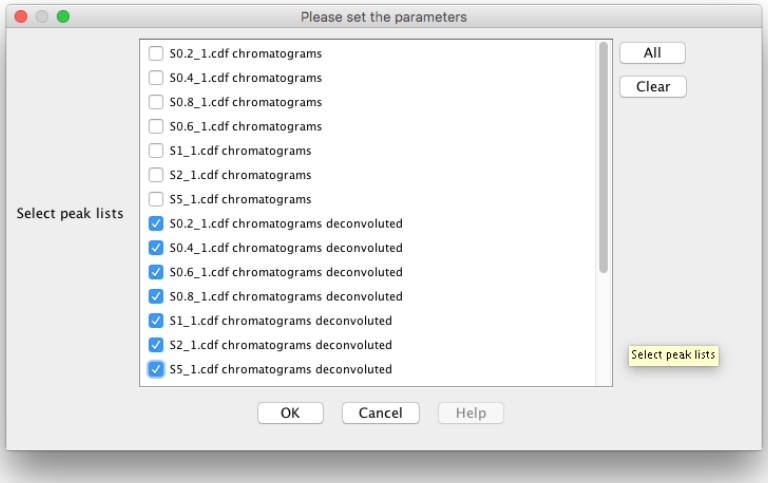


Figure 38: Choosing peaks for Spectral Deconvolution.

The Spectral Deconvolution consists of two steps:

1. Entire retention time interval is split into deconvolution windows so that
   - Peaks produced by the same component or by coeluting components belong to the same deconvolution window,
   - Number of peaks in deconvolution window is significantly smaller than the total number of peaks.

   The deconvolution windows are displayed in the top plot of the preview (see Figure 36), where lines denote peaks in the (retention time, m/z)-plane, and peaks located in the same

deconvolution window have the same color. The vertical sequences of peaks usually mark the presence of one or several compounds, so it is important that those peaks are assigned to the same deconvolution window, i.e. they have the same color on the plot. On the other hand, if deconvolution windows contain too many peaks, it will significantly slow down the spectral deconvolution computations, so the deconvolution windows should be as short (in the retention time domain) as possible.

Parameter *Deconvolution window width (min)* controls the window selection by specifying a window width in minutes. This window width can be chosen based on the width of peaks in a dataset. For GC/MS data, we use value 0.2 min in most cases.

2. The algorithm estimates the number of components in each deconvolution window and construct their model peaks and fragmentation spectra. The estimated number of components is controlled by parameters

   - *Retention time tolerance (min)*, which is the smallest time-gap between any two components. The value of this parameter should be a fraction of the average peak width. In our tests, we use 0.02 min.
   - *Minimum Number of Peaks*, which is the smallest number of peaks in a single component. This parameter depends on a dataset and on how many peaks were detected by the chromatogram deconvolution algorithm. Typically, its value would range from 1 (if only a few peaks are detected for some compounds) to 10 or more (if the number of detected peaks is large for all compounds).
   - *Adjust Apex Ret Times*. For a unit-mass-resolution data, where co-eluting compounds may be present, and a peak typically consists of tens and hundreds of points, this parameter should be off. For high-mass-resolution data, where co-eluting compounds are rare and a peak consists of a few points, this parameter should be on.

See section 4.3.1 for instructions on how to view the spectral deconvolution results.

4.4 Alignment

In ADAP-GC workflow, the alignment step uses similarity between fragmentation mass spectra to find similar components in several files. For that reason, the alignment is performed **after** the spectral deconvolution step.

For performing alignment, select the peak lists that need to be aligned, and then click *Peak list methods* → *Alignment* → *ADAP aligner* (Figure 39)

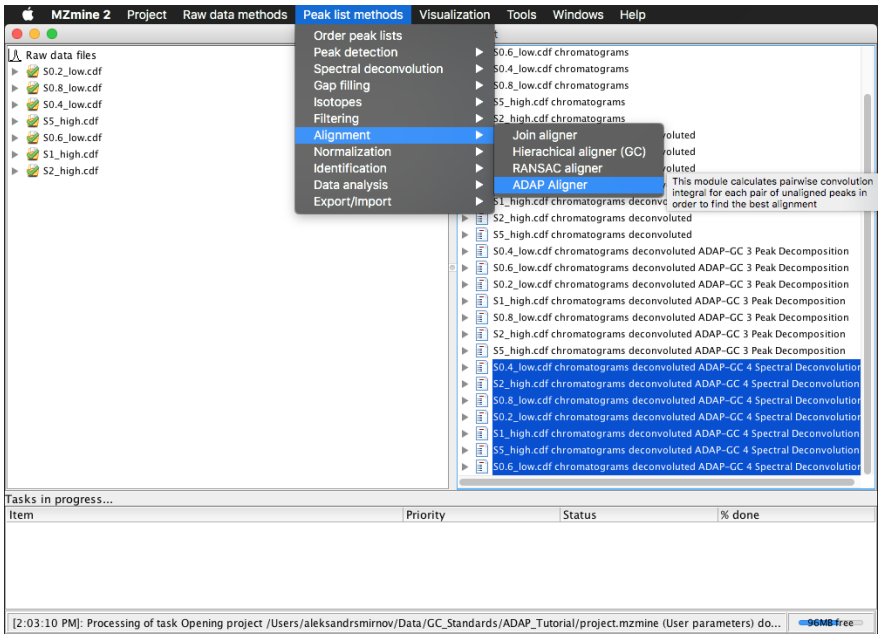


Figure 39: Alignment of components.

A window as shown in 40 pulls up allowing you to specify the alignment parameters.

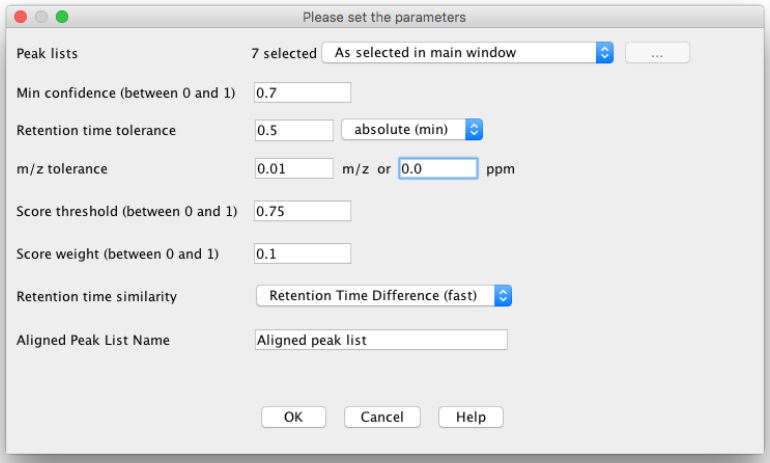


Figure 40: Specify parameters for alignment of components.

- *Min confidence* takes values between 0.0 and 1.0 and defines the minimum fraction of samples where aligned components must be present. For instance, if a dataset contains 10 samples, and the Min confidence is set to 0.7 (default value), then an aligned component must present in at least 7 samples.
- *Retention time tolerance* is the maximum time-gap between aligned compounds in different samples.

- *M/z tolerance* is used for comparing peaks in spectra of aligned compounds and choosing a quantitative mass.

- *Score threshold* takes values between 0.0 and 1.0. Similarity score between compounds in different samples is determined as follows:

$$Score(c_1, c_2) = wS_{time}(c_1, c_2) + (1 - w)S_{spec}(c_1, c_2),$$

  where S_{time} is the relative retention time difference between two compounds and S_{spec} is the spectrum similarity between two compounds. The score threshold defines the minimum similarity score between aligned compounds from different samples. The default value is 0.75.

- *Score weight* takes values between 0.0 and 1.0. This parameter is the value w that is used in the similarity score $Score(c_1, c_2)$. If $w = 0.0$, then only the spectrum similarity is used for calculating $Score(c_1, c_2)$. If $w = 1.0$, then only the retention time difference is used for calculating $Score(c_1, c_2)$. If $w \in (0.0, 1.0)$, then a weighted combination of the spectral similarity and the retention time difference is used. The default value of this parameter is 0.1.

- *Retention time similarity* Users can choose on of two options for calculating retention time similarity S_{time}: *retention time difference* and *cross-correlation*. As the second option is still in development, users are strongly advices to use the first option for now.

4.5 Student's T-test and Fold change

After alignment is complete, it is possible to calculate the significance of each component by performing Student's T-test and by calculating the logarithmic fold change. To calculate the significance, select peak list *Aligned peak list* and then click *Peak list methods* → *Data analysis* → *Student's t-test and fold change* as shown in Figure 41

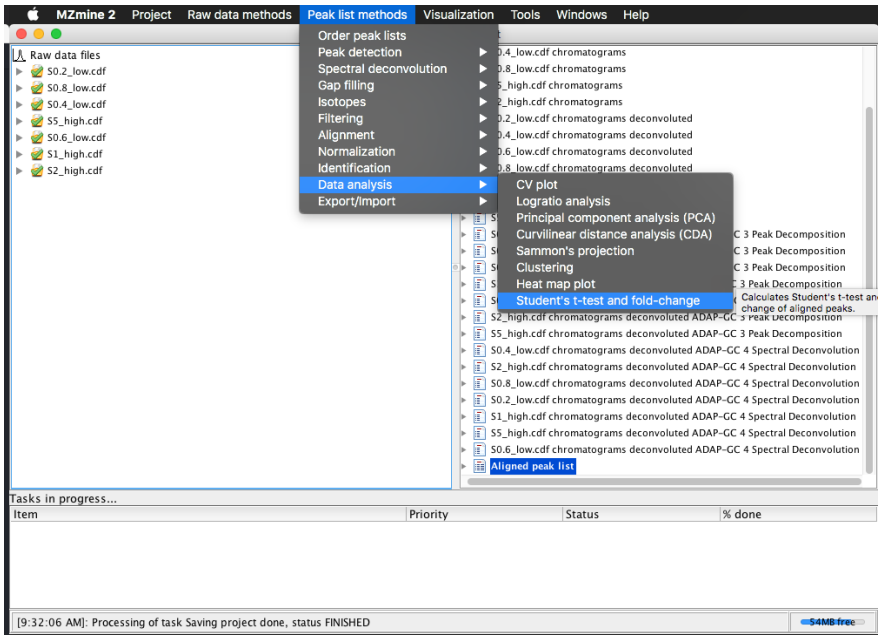


Figure 41: Student's T-test and Fold change.

A window as shown in Figure 42 will pull up. You will need to specify group IDs, so that the files from an experimental group would contain the experimental group ID in their names. Similarly, the files from a control group should contain the control group ID in their name.

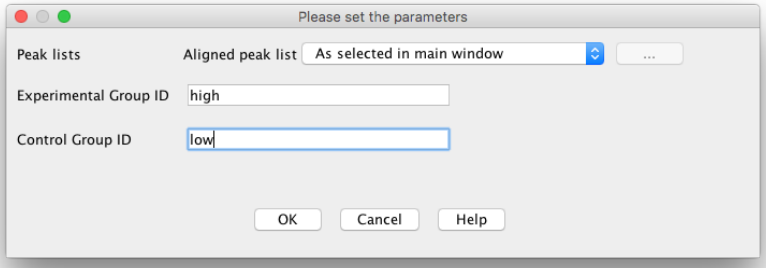


Figure 42: Student's T-test and Fold change.

Although, it is currently not possible to display the results of the significance calculation within MZmine 2, you can export a peak list into CSV format, and the csv file will contain columns STUDENT_P_VALUE, STUDENT_T_VALUE, and LOG2_FOLD_CHANGE with the corresponding significance values for each component.

To export results, select peak list *Aligned peak list* and then click *Peak list methods* → *Export/Import* → *Export to CSV file* as shown in Figure 43.

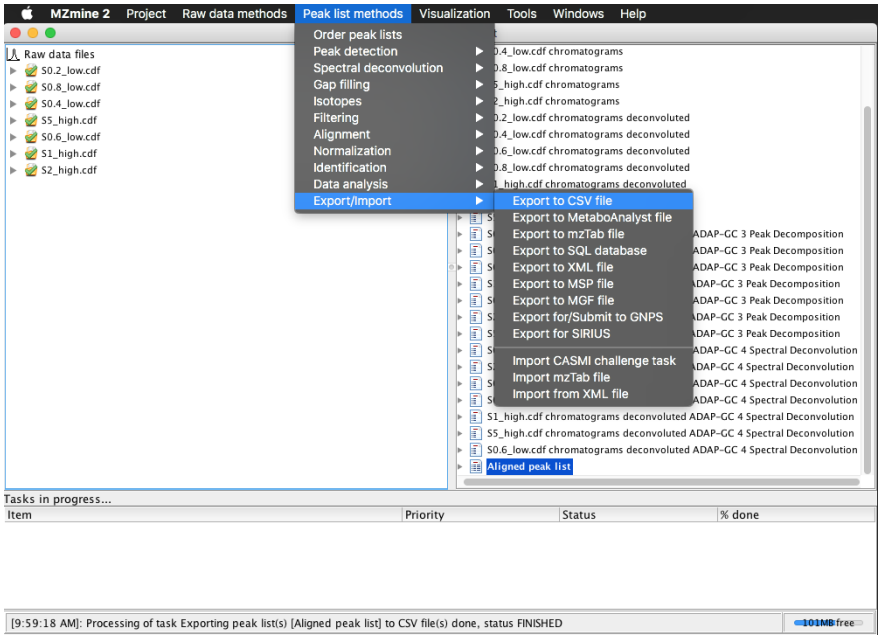


Figure 43: Export into CSV file.

A window as shown in Figure 44 will pull up. In addition to other options, make sure that you select *Export quantitation results and other information* to output the significance results.

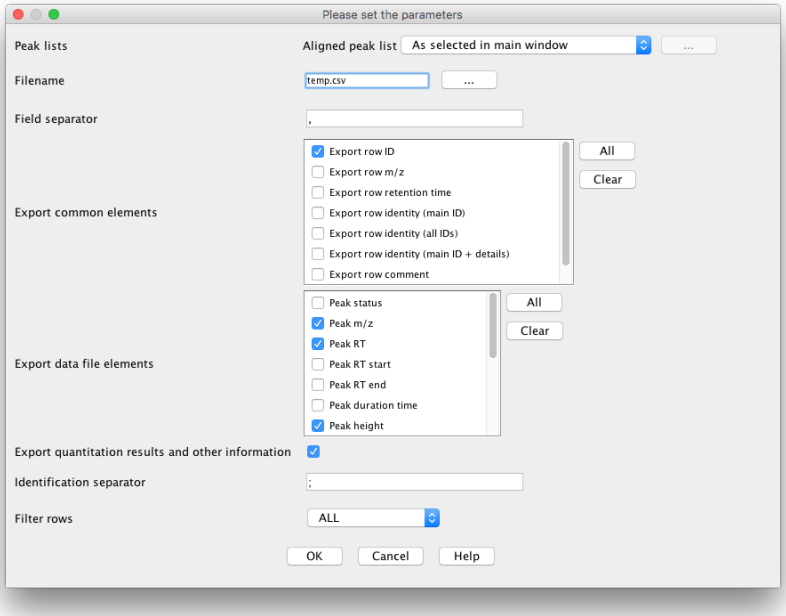


Figure 44: Export into CSV file.

The exported csv file is shown in Figure 45.

row ID	STUDENT_P_VALUE	STUDENT_T_VALUE	LOG2_FOLD_CHANGE	QUANTITATION INTEN	QUANTITATION INTE	QUANTITATION MASS
1	0.205645283746572	-1.84199063760559	1.15378259573839	6068.97675494103	16228.5631140642	73
2	0.24980691920291	-1.59756561321939	2.74028724071167	10127.3316653744	76895.6650501529	73
3	0.310443577297517	1.16721076016859	-0.135521200733597	2999.08751123155	2995.68786023116	75
4			1.00372036745727	18547.2773761625	39424.6504491954	73
5	0.227373446903107	-1.71566694736762	2.55217539662221	65895.5981651826	338331.129759904	73
6	0.141372138585872	-2.14443638837416	2.41530782116744	6016.59195323062	91343.4772104649	73
7			-0.483733580524133	1913.20869439238	901.792619346485	75

Figure 45: Export into CSV file.

4.6 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 46.

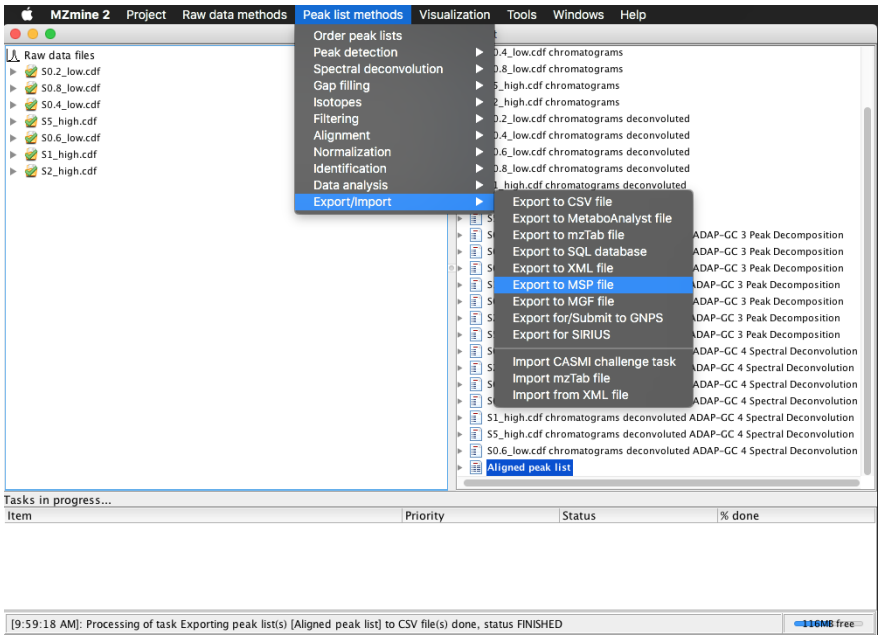


Figure 46: Export mass spectra to a MSP file.

A window as shown in Figure 47 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).

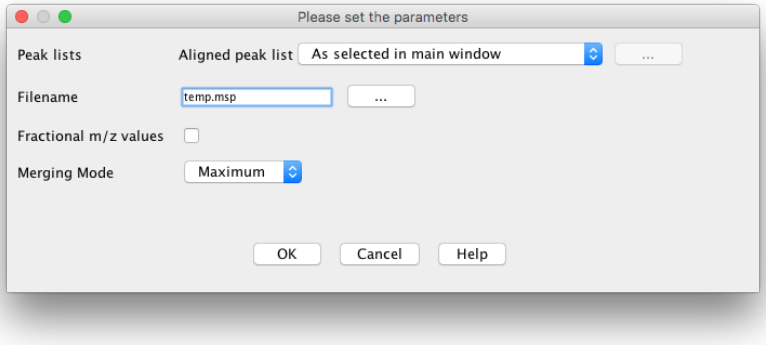


Figure 47: Export mass spectra to a MSP file.

Open the exported .msp file in a text editor. You will see that the mass spectra after alignment have been exported. Figure 48 shows a small portion of the .msp file.

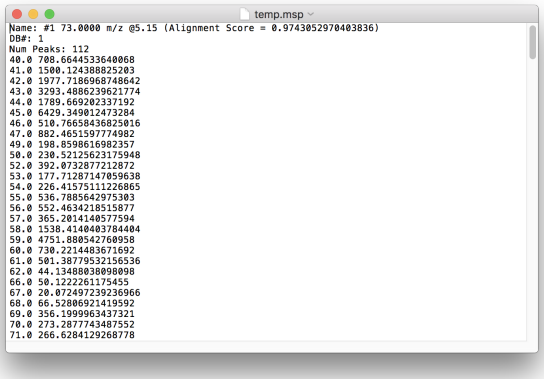


Figure 48: Example .msp file exported by ADAP-GC.

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 49.

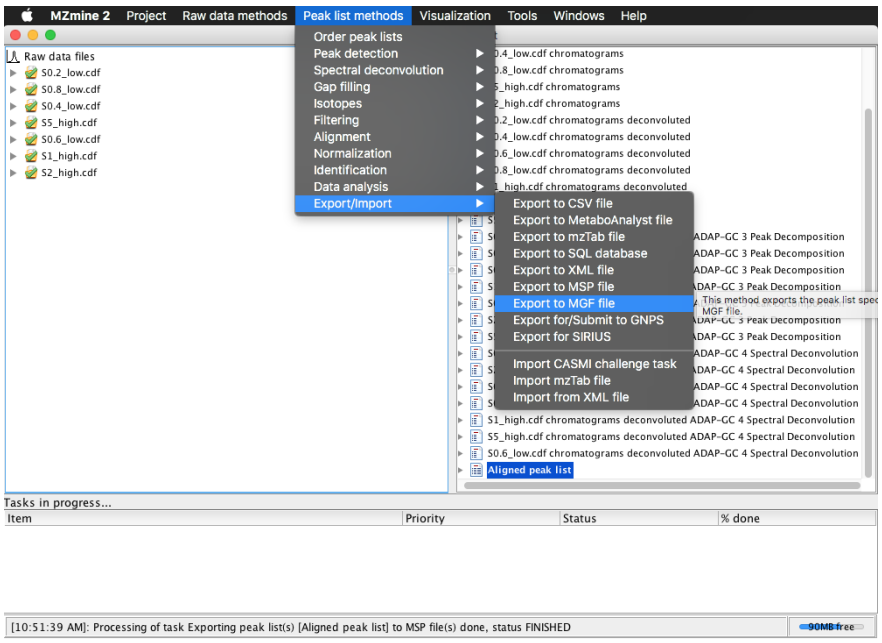


Figure 49: Export mass spectra to a MGF file.

A window as shown in Figure 50 is open allowing you to name the export file.

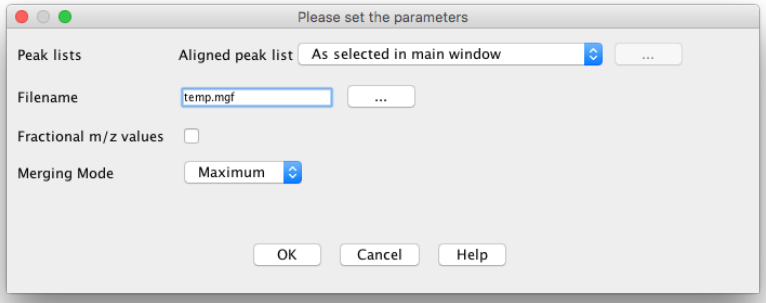


Figure 50: Export mass spectra to a MGF file.

Figure 51 shows part of a .mgf file exported from MZmine 2.

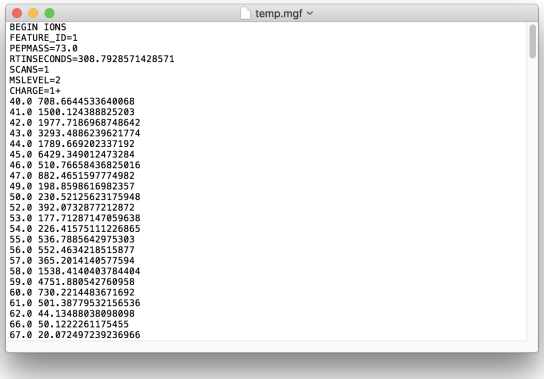


Figure 51: Example .mgf file exported by ADAP.

5 Batch Processing

Create the parameter file for batch processing.

On Mac, run `./startMZmine_MacOSX.command ‘‘path to and name of the batch processing file’’` in the terminal.

6 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.