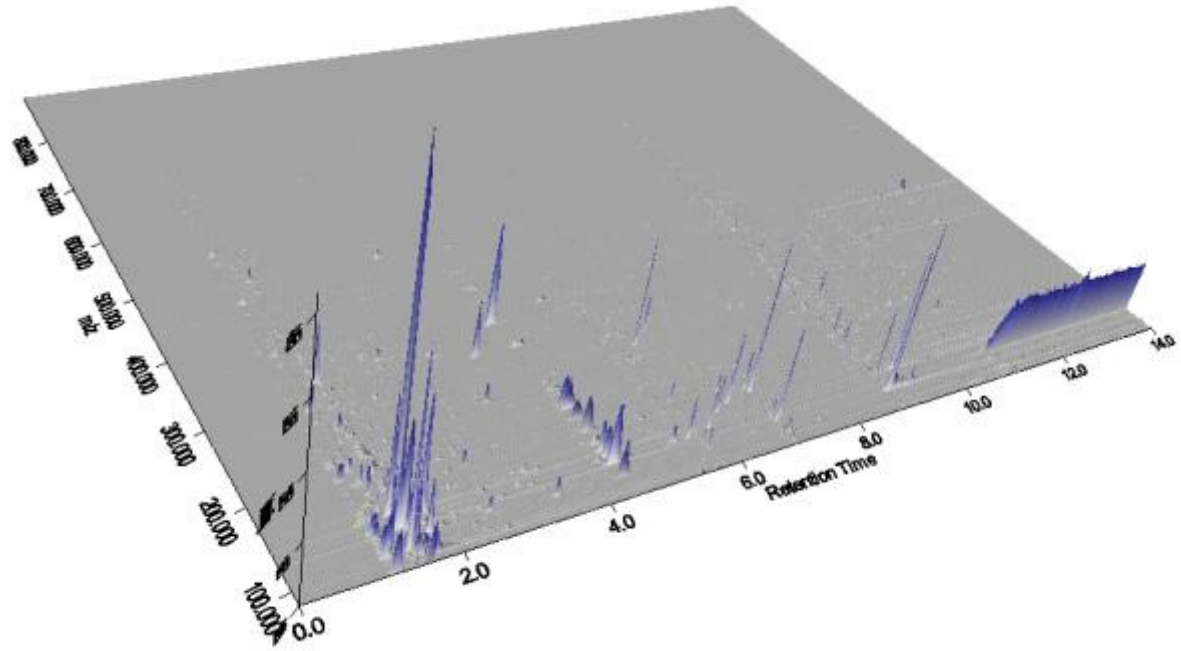
# MZmine Tutorial

First presented at EMBL-EBI Industry Programme & Metabolights Project Workshop

22nd May 2012

Mark Earll - Syngenta UK



## About MZmine

MZmine was developed at Okinawa Institute of Science and Technology, Japan and VTT Finland.More recently some development has been sponsored by Syngenta. It is a Java based program and is therefore platform independent. You may download it from the following website however it should be pre-installed on the EBI machines for this tutorial.

*http://MZmine.sourceforge.net/*

MZmine will import the following filetypes: Net CDF, mzData, mzML, mzXML, Xcalibur Raw files, Agilent CSV files. (For the Thermo Xcalibur files it is necessary either to have the Thermo Xcalibur software installed on the same machine or to have downloaded and installed the free ThermoMSFilereader software to be found at

*http://sjsupport.thermofinnigan.com/public/detail.asp?id=586.*

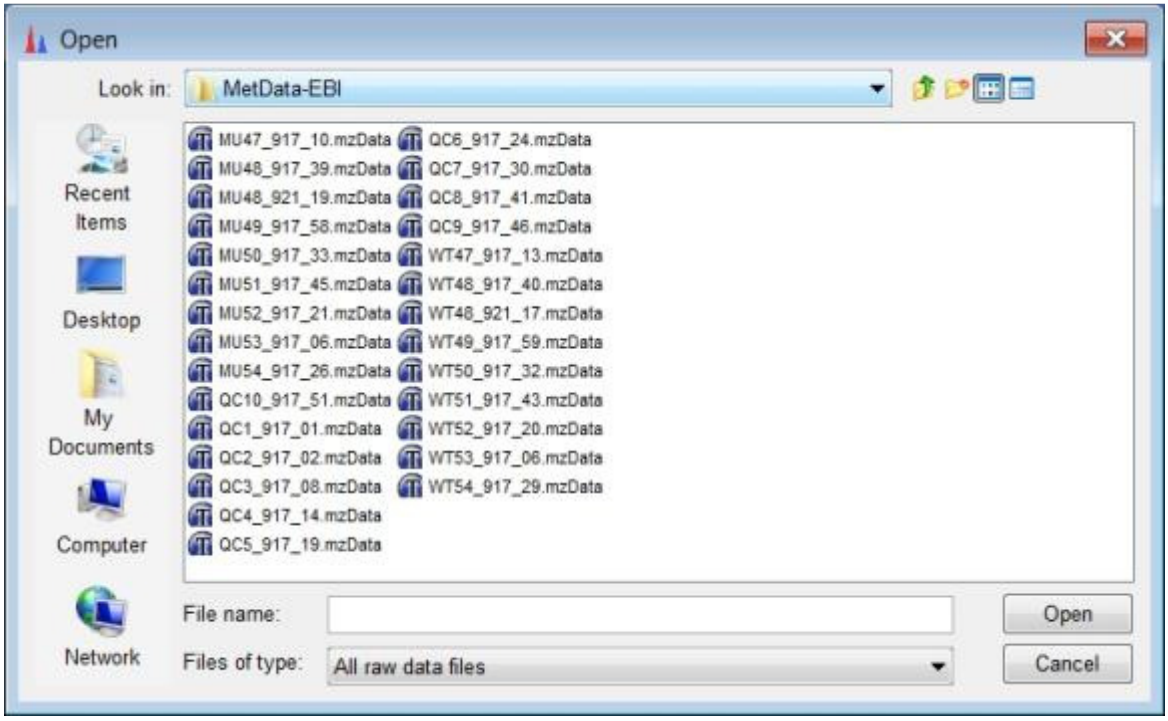
The version of MZmine used in the following examples was 2.8

## Processing a simple Metabolomics dataset in MZmine

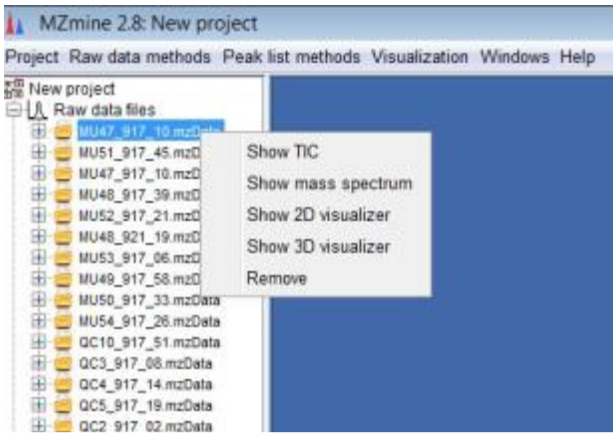
In the example dataset we have an excerpt of a metabolomic study on the ripening of fruits. We have nine samples of two different varieties, Wild-type and non-ripening Mutant plus ten control samples which consist of a large batch of identical fruit extract that are run at every fifth sample. In addition the fruit are sampled everyday from the onset of ripening between 47 and 54 days. (This example datasets is only a small excerpt of a larger replicated study). The data were collected on a Thermo Velos Orbitrap running in ESI+ mode with a UPLC column.

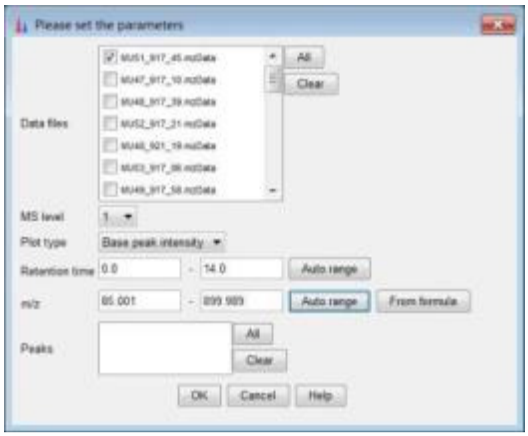
## Loading the data

One of the great advantages of MZmine is its interactivity. Firstly we begin by importing the data.
*Raw Data Methods / Import*

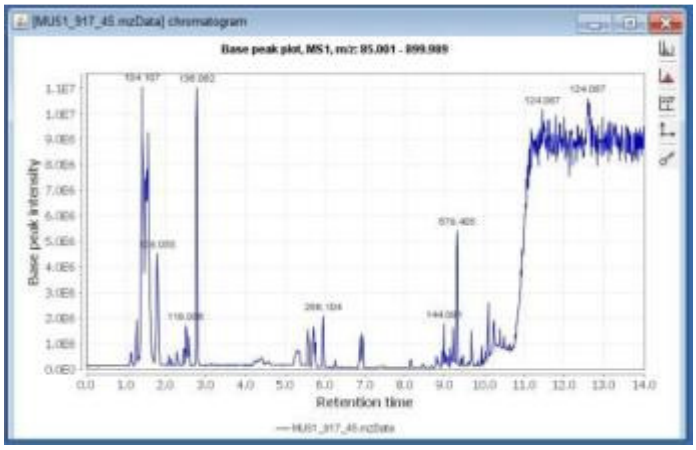


Once the data is imported we can right click on the data file to reveal several display options



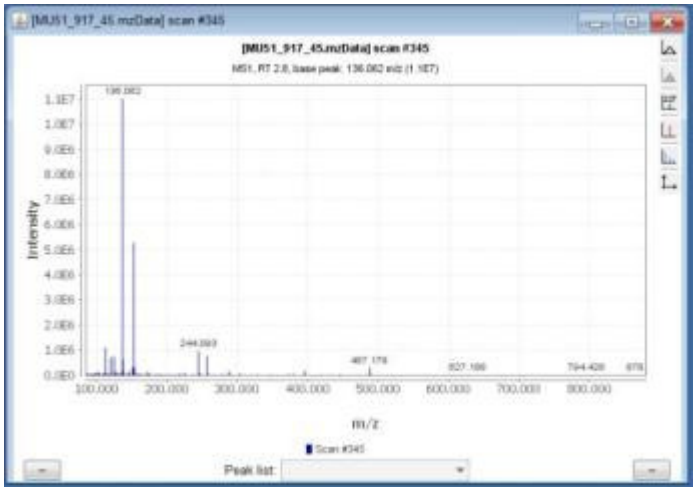




The TIC option offers the option of Base peak or TIC and allows you to set various ranges. Clicking OK leads to a high quality spectrum plot. The plot is fully zoomable and interactive, and double clicking a peak leads to its mass spectrum. Clicking and dragging upward or to the left is a gesture which results in zooming back out to maximum zoom. Clicking and dragging downwards or to the right zooms in.




***NB: Make sure you take a note of the height of the baseline and the height of the smallest peaks. This will be useful later.***

The mass spectrum plot also enables you to see associated ms-ms data. (In this dataset the MS-MS information has been removed).




Peak detection is a three step process:

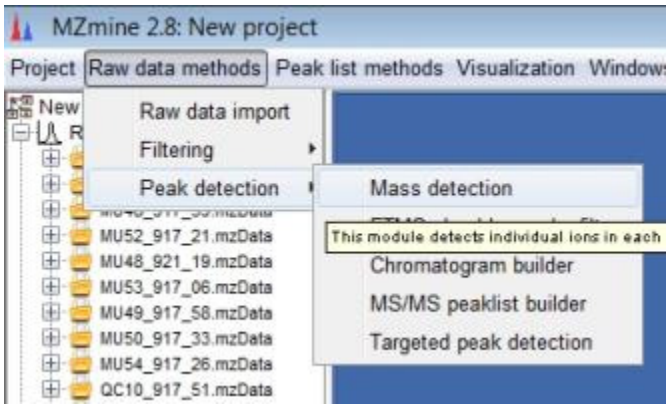
(1) Mass detection

(2) Chromatogram building

(3) Peak Deconvolution

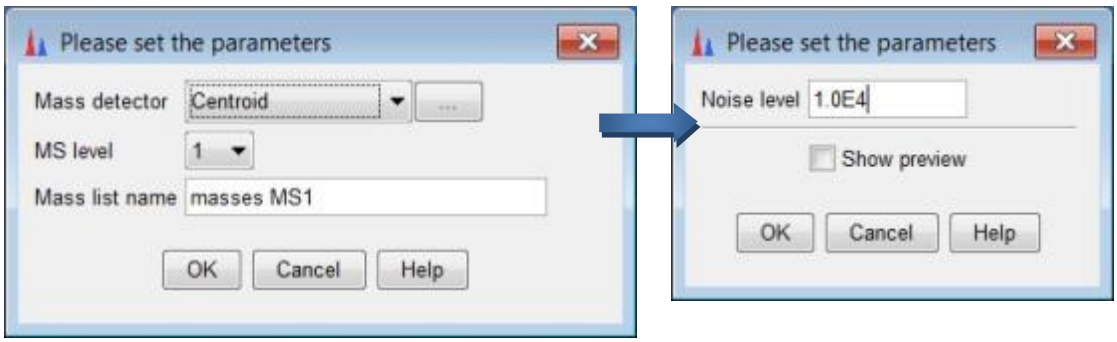
## Mass Detection

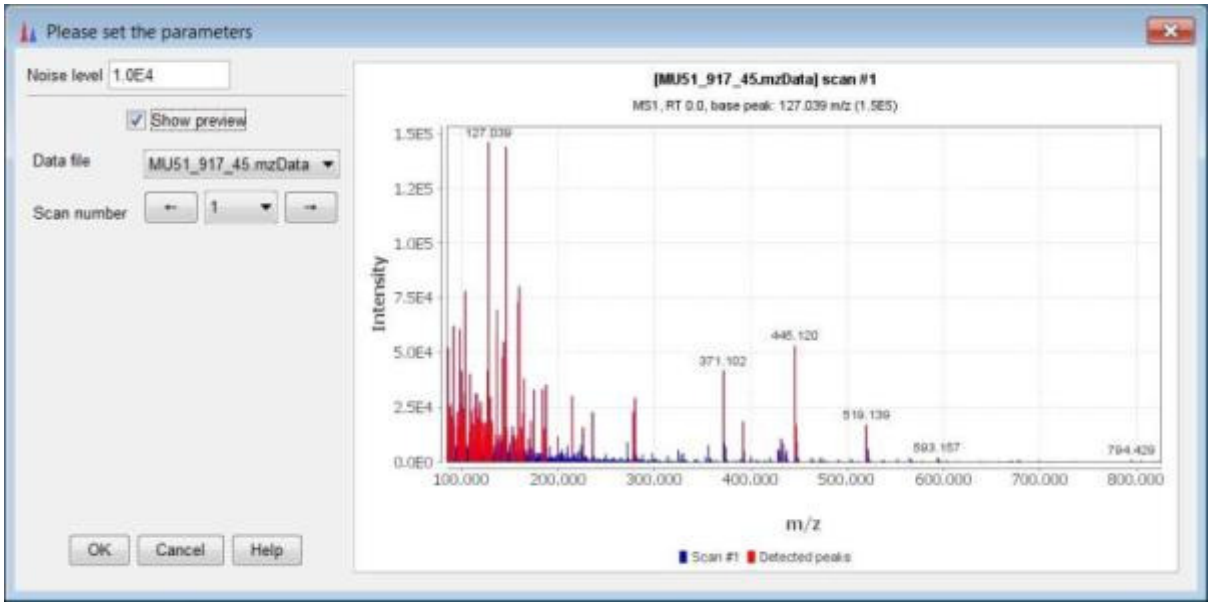
Click on *Raw data methods/Peak Detection/Mass detection*



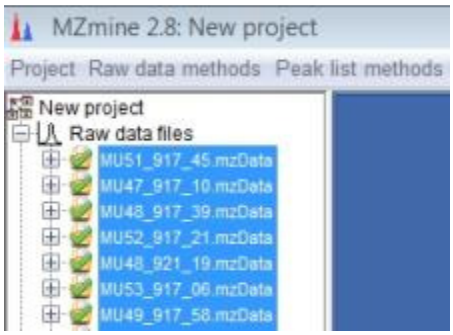

In this case we have imported mzData files which are centroided during the conversion from .RAW so the only option is Centroid mode. (If you have imported Thermo .RAW files then the data is continuous and you can use the exact mass, local maxima, recursive threshold or wavelet methods).

The *'Show preview'* option allows you to interactively set the threshold for peak detection in the mass dimension. The aim is to detect peaks but not too many noisy features.



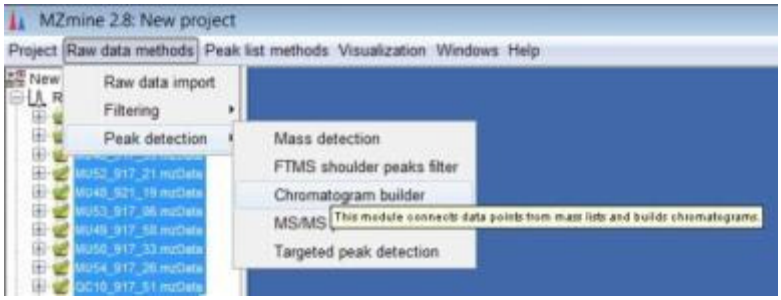



After clicking OK MZmine will build the mass list. Depending on the speed of your computer this may take some time. When the mass list is built the icon will show a green tick mark.

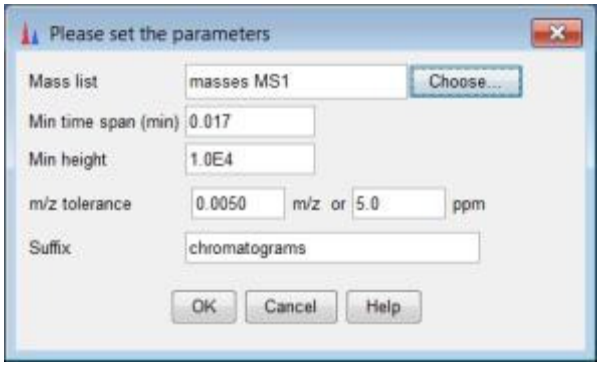



## Chromatogram Builder

The Chromatogram builder is found under *Raw data methods/Peak detection/chromatogram builder*

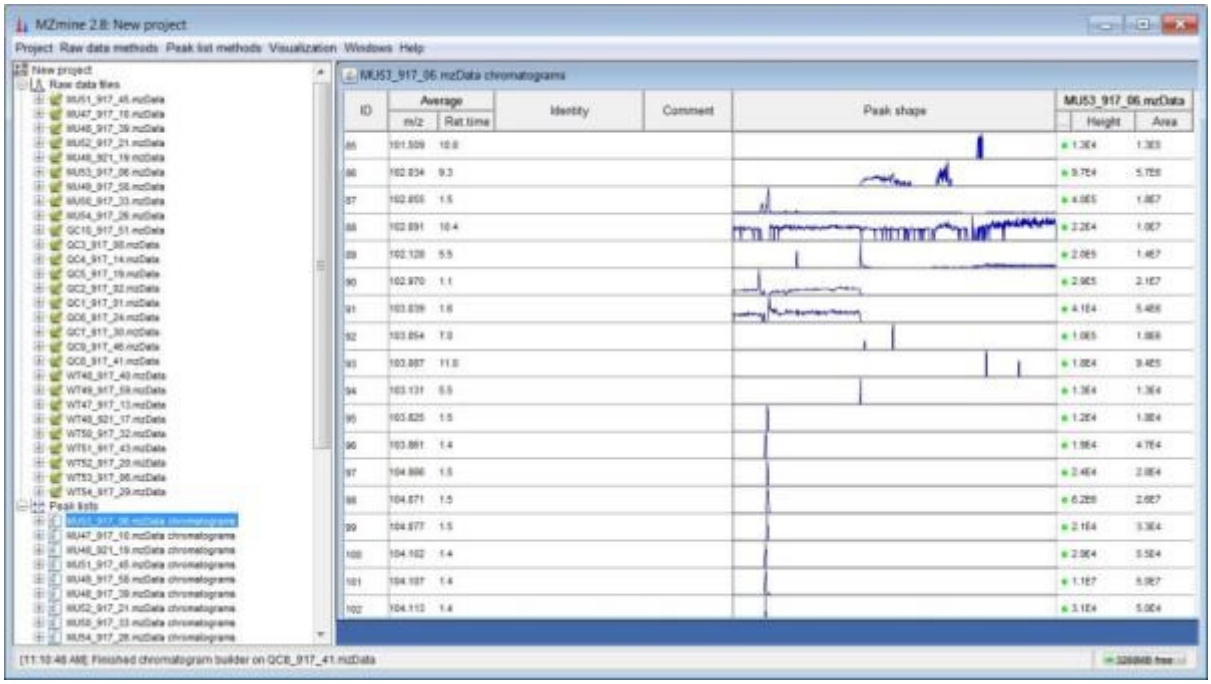



Click on the *Choose...* button to select the mass list just generated and fill in the parameters as shown (for your own data these will vary)




You will then see a number of chromatograms listed in the left hand pane

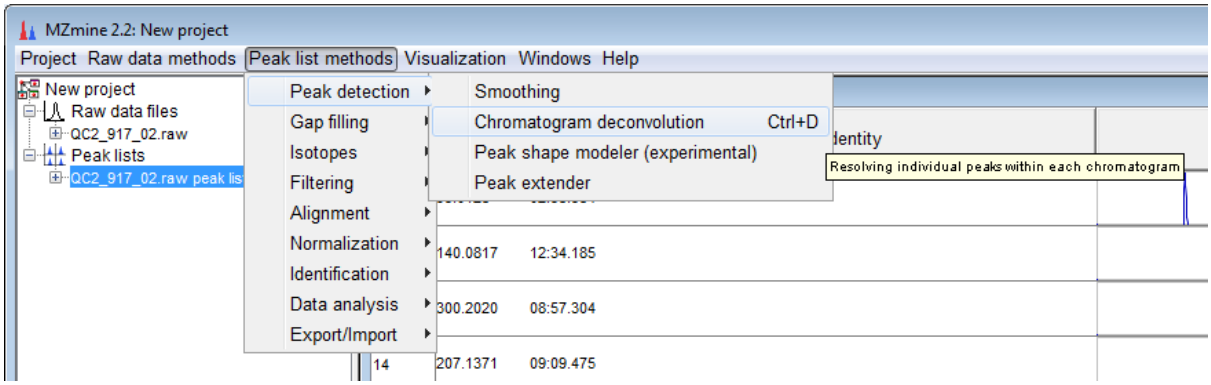

Double clicking on a chromatogram will bring up the results:



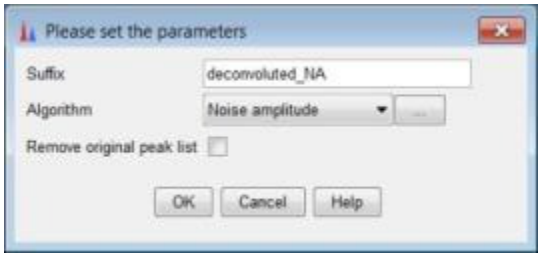

The peak list is comprised of a series of ion chromatograms taken at each point that was detected in the chromatogram builder. Some ion chromatograms may contain more than one peak so a second peak deconvolution stage is required.

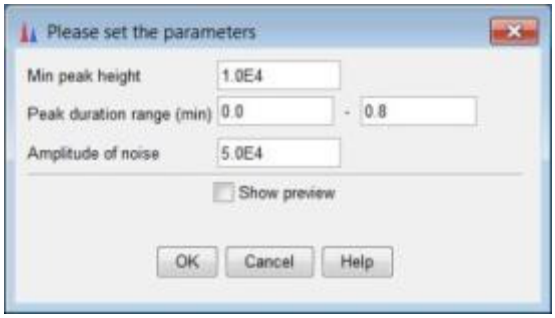
## Peak Deconvolution

Make sure you highlight the chromatograms in the left hand pane. Click *Peak list methods/Chromatogram deconvolution*. You have a choice between Baseline Cut-off, Noise Amplitude, Savitsky-Golay, Local minimum search [and Wavelets (XCMS) in a future version]

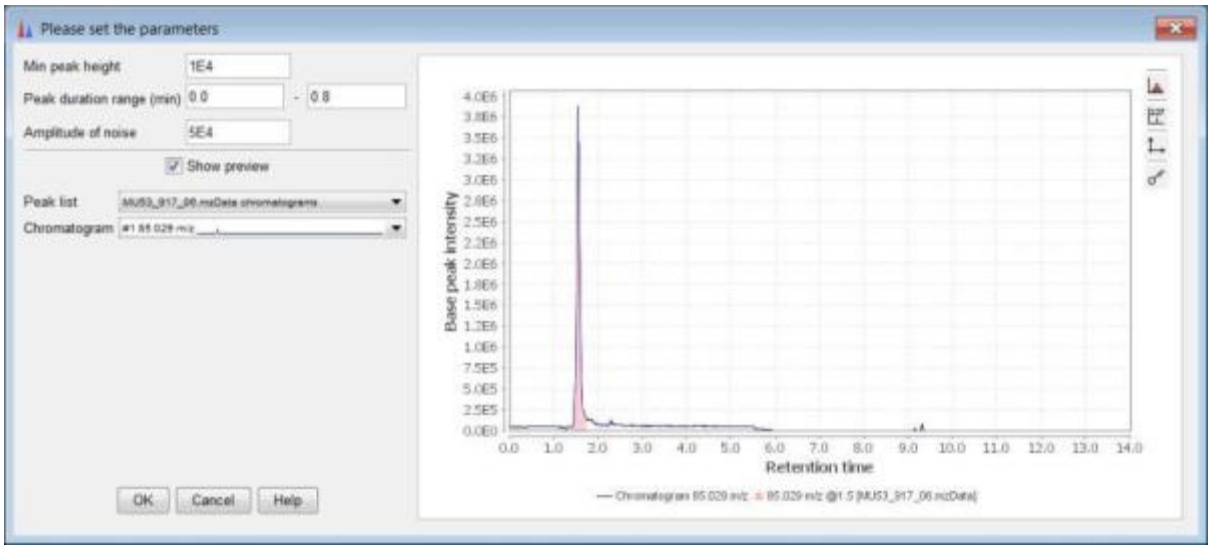



Here we use *Noise Amplitude* [My personal preference is Wavelets XCMS but it may not be implemented in this demonstration version]

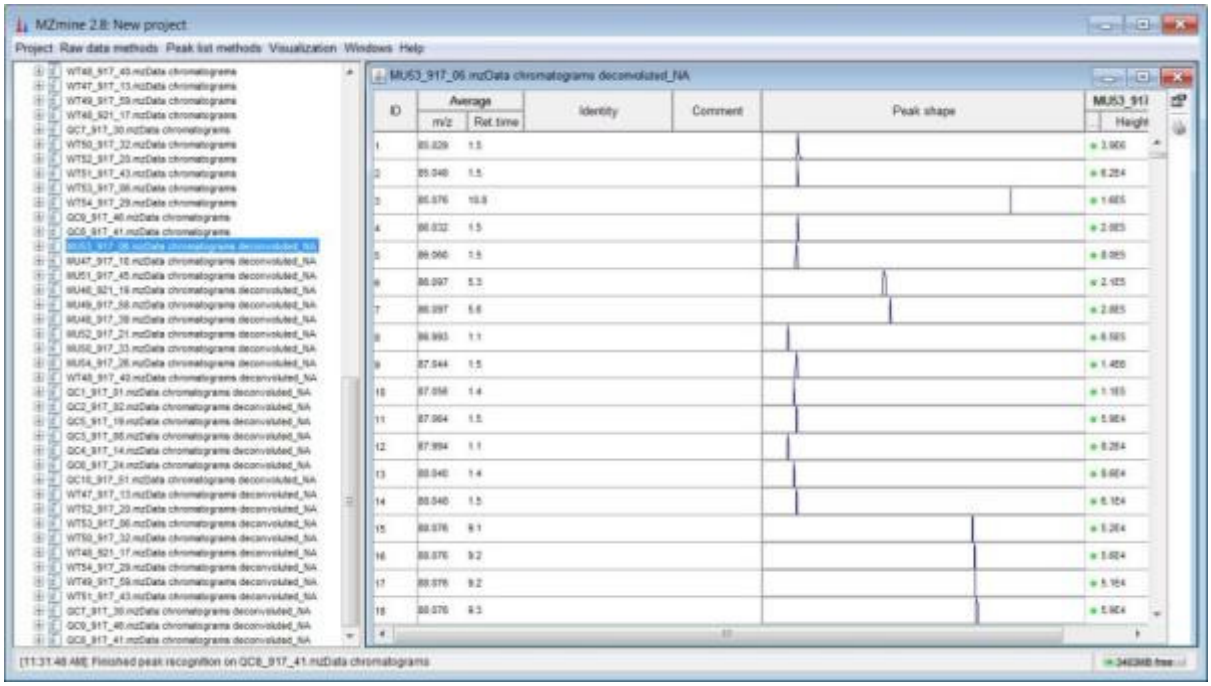


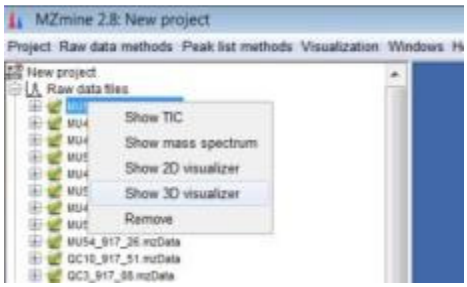

Fill in the boxes with the appropriate values. One of the recent key improvements to MZmine is the ability to specify a maximum for peak duration. This is very useful for removing some of the artefact peaks caused by column bleed. Here we use 0.8 mins. Another trick used here is to set the amplitude of noise slightly higher than the min peak height. This means that peaks with raised baselines as in the example get detected. The downside is a slight loss in integration accuracy. (The alternative is to baseline correct first - see later, or use a more robust peak picker such as the wavelet option). Peak picking is always a compromise and requires a lot of experimentation for optimal results.



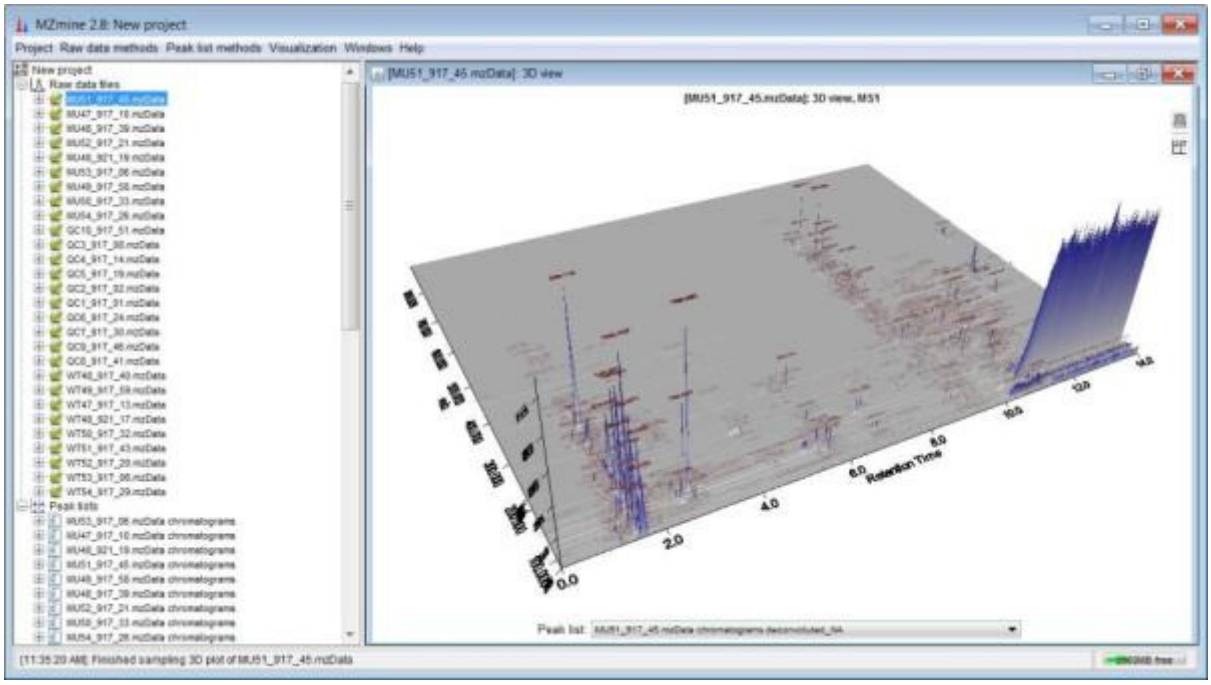

After the peak deconvolution step MZmine produces a resolved peak list with one peak per row:



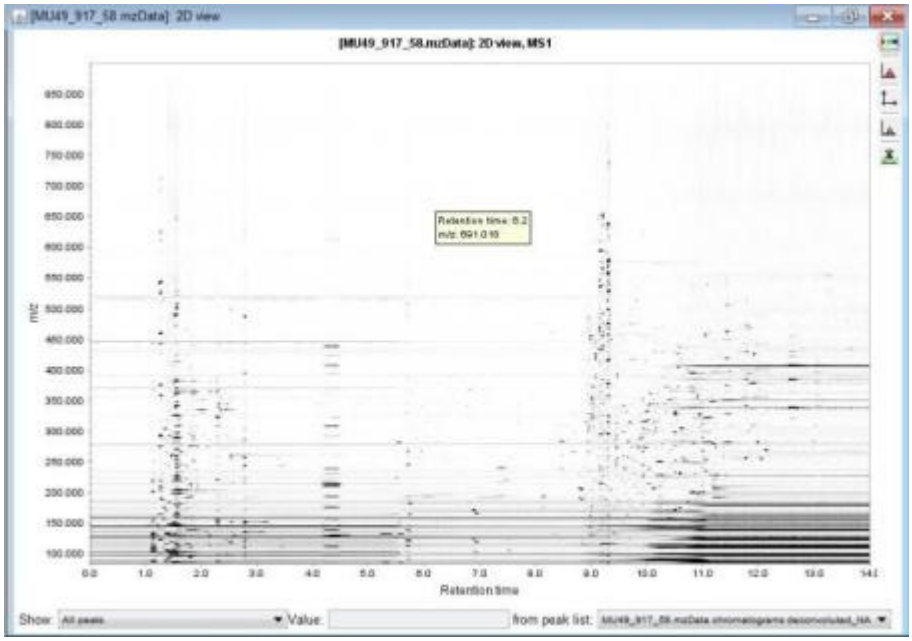

We can visualise the peaks using the 3D visualiser plot on the raw data. This is a useful check of the accuracy of peak picking.



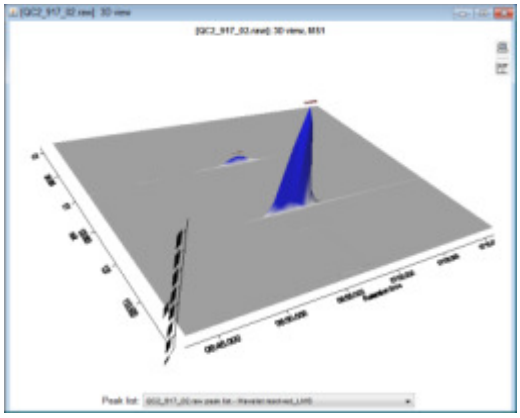




There is also a 2D "gel view" of the data - click *Show 2D visualiser*.

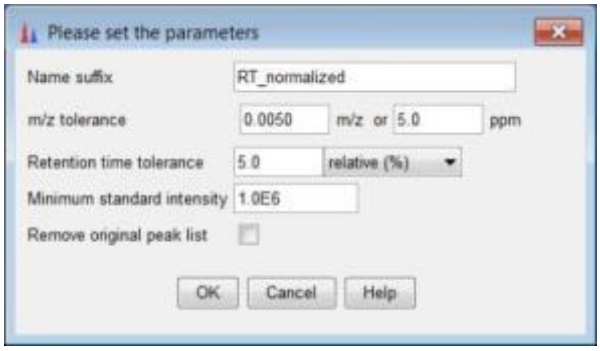



Right clicking a peak in the peak list From the peak list and selecting (*show... chromatogram quick*) shows the peak and the peak integration in pink. (There is also an option to see the peak in 3D but this appears to be broken in 2.8)

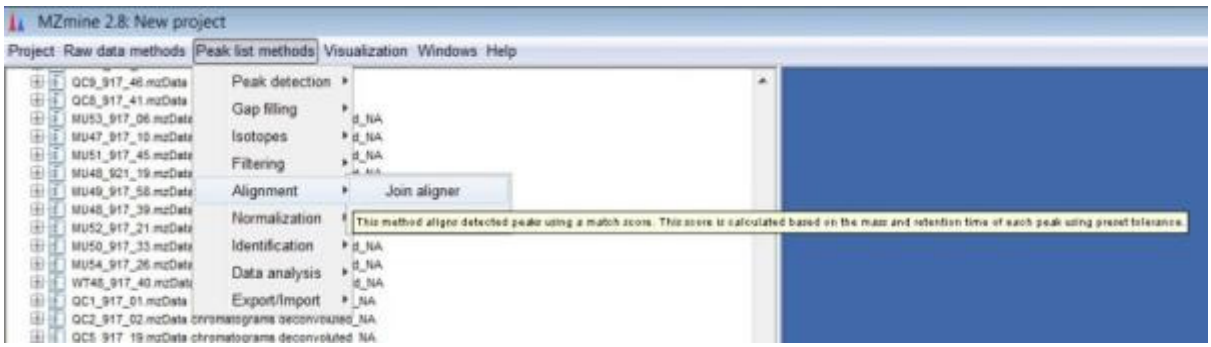


## Peak Alignment

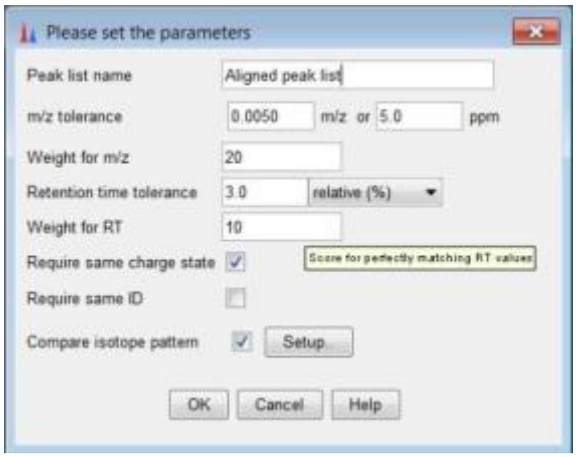
In MZmine peak alignment is done after the peaks are picked. To adjust for any slight variation in retention time a retention time normaliser is provided. Click on *Peak list methods /Normalisation /Retention time normaliser*.

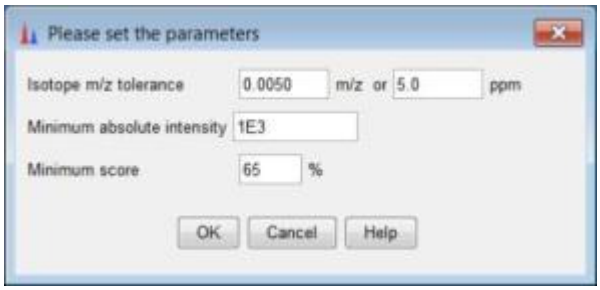



Next we will combine the peaks using the *Peak list methods/Alignment/Join aligner.*

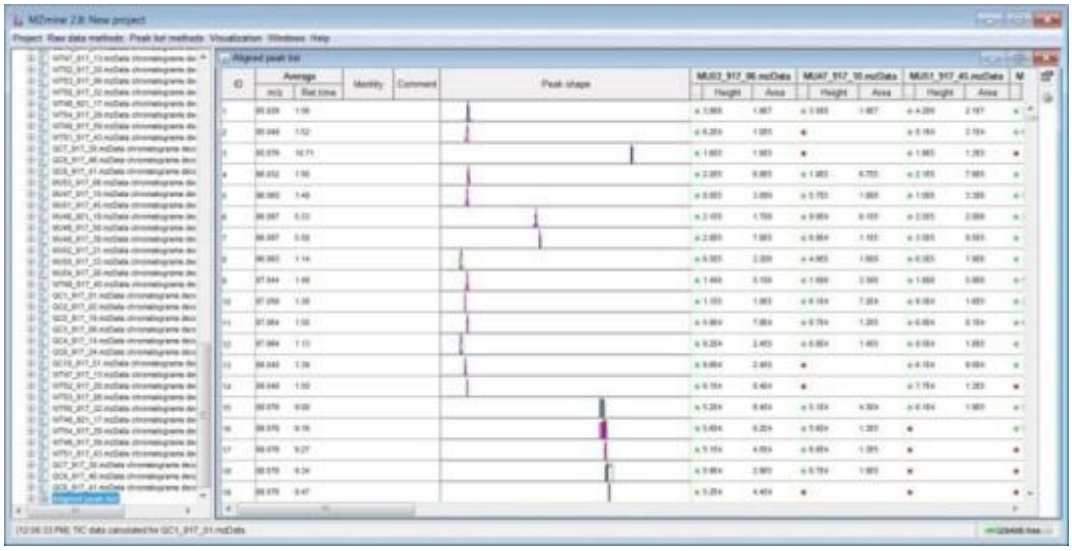



The alignment is based upon RT and m/z tolerance. There are options to only merge ions with the same charge state, the same ID or by isotope pattern. We will not use "Require same ID" because we have not identified any compounds yet.




You should now have an aligned peak list. Green dots indicate the presence of that peak in the scan. A red dot indicates the peak was not detected. After the identification process we will return to fill in these gaps with baseline levels from the other scans. (Gap-filling may alter the accuracy of the m/z value due to averaging)

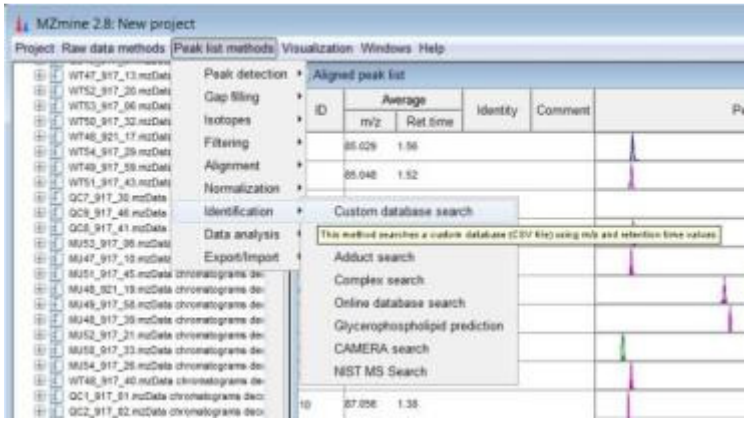


## Identification

In order to identify peaks a database of m/z or m/z and Retention times are required. There are two options. One is a custom database compiled on your own instrument based upon the measured masses of the molecular ion and any adducts or fragments. The other option is an online database search based purely on the accurate mass and isotopic pattern matching.

## Custom database search

Peak list methods/Custom database search

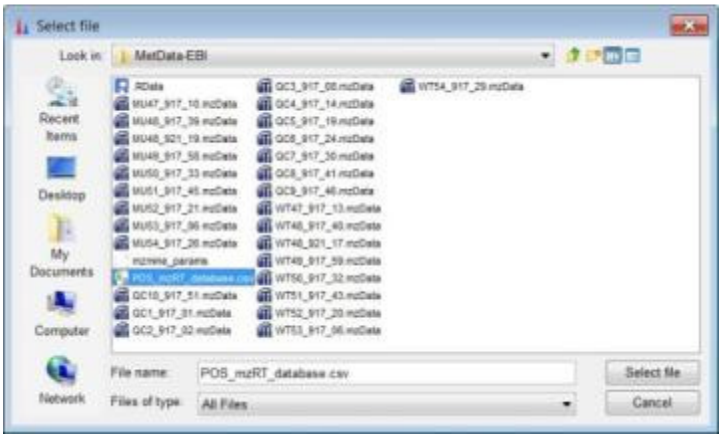



In the dialog box select POS_mzRT_database.csv (which will be with the demo datasets).

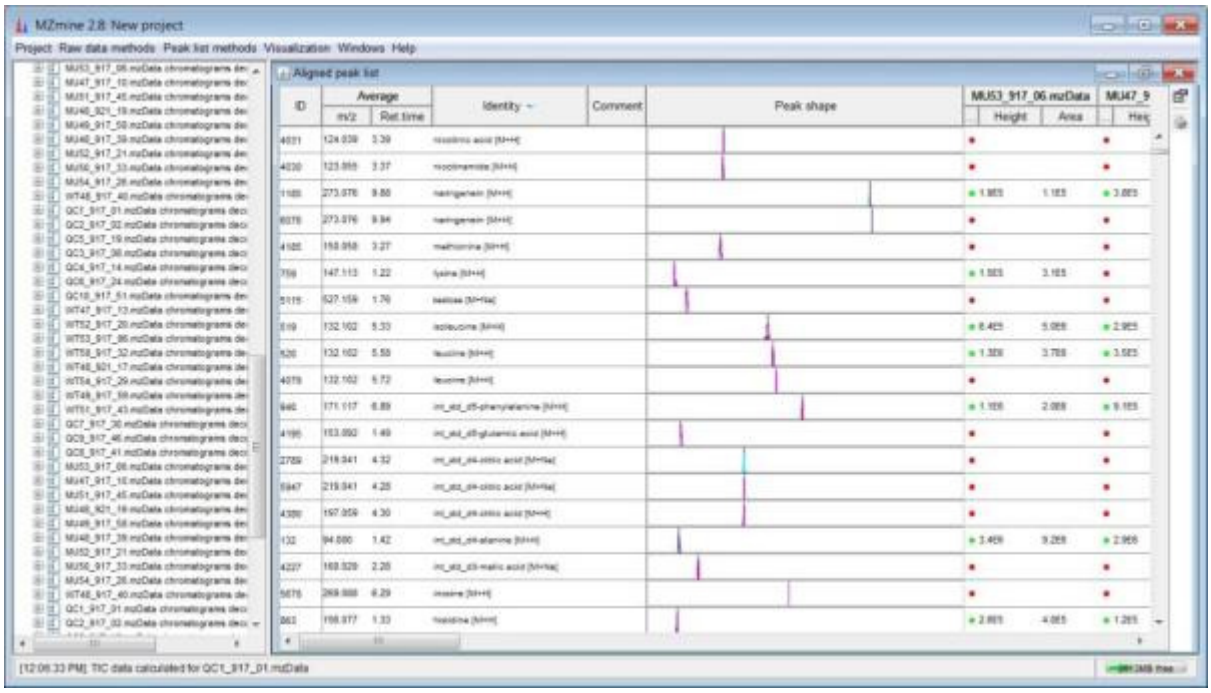
This is a database we have compiled for our library using our UPLC-MS system in positive ESI+ mode. (NB: Your own data will differ in RT and possibly the ionisation profiles - you will need to compile your own library database relevant to your own system!)

Here is an excerpt of our custom database. The first column is KEGG ID, then accurate m/z, Retention time, Identity and Formula. We have also included common adducts and dimers in the list such as [M+K], [M+Na], [2M+H]

| 1 | ID | m/z | Retention | Identity | Formula |
|---|---|---|---|---|---|
| 28 | C00041 | 90.05496 | 1.45 | alanine [M+H] | C3H7NO2 |
| 29 | na | 161.0921 | 1.71 | alanine-alanine [M+H] | C6H12N2O3 |
| 30 | C01551 | 159.0513 | 1.62 | allantoin [M+H] | C4H6N4O3 |
| 31 | C06464 | 181.0707 | 1.52 | altrose [M+H] | C6H12O6 |
| 32 | C00216 C00259 | 151.0601 | 1.62 | arabinose [M+H] | C5H10O5 |
| 33 | C01112 | 231.0264 | 1.56 | arabinose 5 phosphate [M+H] | C5H11O8P |
| 34 | C00532 | 153.0758 | 1.51 | arabitol [M+H] | C5H12O5 |
| 35 | C00792 | 175.119 | 1.37 | arginine [M+H] | C6H14N4O2 |
| 36 | C00049 | 134.0448 | 1.45 | aspartic acid [M+H] | C4H7NO4 |
| 37 | C00099 | 90.05496 | 1.38 | beta-alanine [M+H] | C3H7NO2 |
| 38 | C02512 | 115.0502 | 1.48 | beta-cyano-l-alanine [M+H] | C4H6N2O2 |
| 39 | C00719 | 118.0863 | 1.57 | trimethylglycine [M+H] | C5H11NO2 |
| 40 | C00308 | 177.0982 | 1.34 | canavanine [M+H] | C5H12N4O3 |
| 41 | C09773 | 363.1286 | 5.98 | catalpol [M+H] | C15H22O10 |
| 42 | C00185 | 343.1235 | 1.77 | cellobiose [M+H] | C12H22O11 |
| 43 | C01484 | 209.0961 | 10.94 | chalcone [M+H] | C15H12O |
| 44 | C00852 | 355.1024 | 9.1 | chlorogenic acid [M+H] | C16H18O9 |

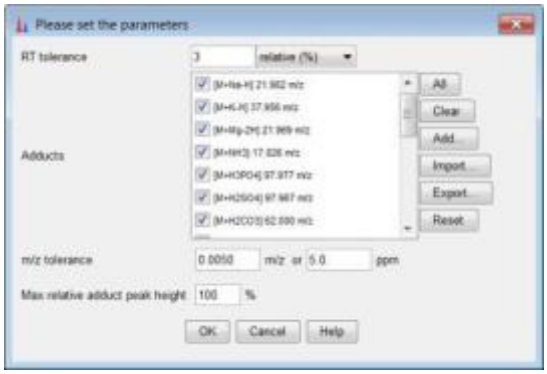




Clicking the Identity tab twice should bring all the identified peaks to the top of the list. Notice that some of the identified component are isotopic internal standards which we use to normalise the data (outside MZmine).




## Adduct search

Under peak list methods/ identification there is an adduct search option. Note you can now load or save a customised list of adducts. Set the RT and m/z tolerance and the relative adduct peak height.





There is a similar option for fragment search (based upon MS/MS data). This cannot be used on the current dataset as the MS2 information has been removed in the conversion from RAW to mzData. There is also a method for removing isotopic Peaks *list methods/Isotopes/Isotopic peaks grouper* (not shown).

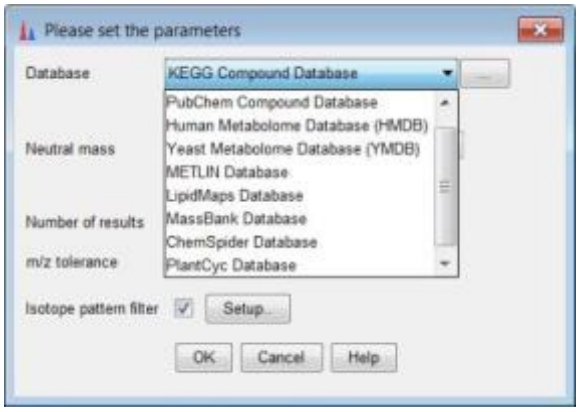
## Online Search

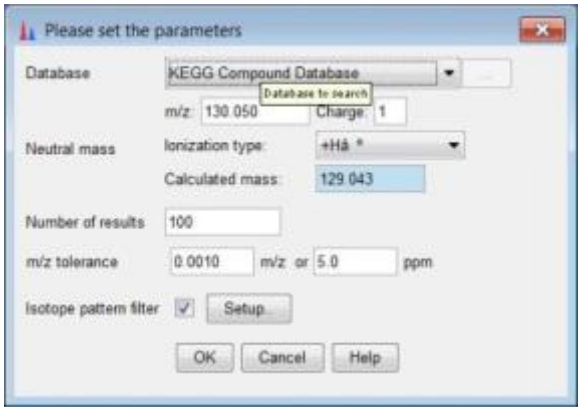
Identification by online searching should be done ***with caution***. We have found that this method often returns many research compounds, drugs and pharmaceuticals which are irrelevant to our plant based studies. For this reason we recommend searching ***individual peaks*** using the peak list.

Let's keep things simple and search a single peak:

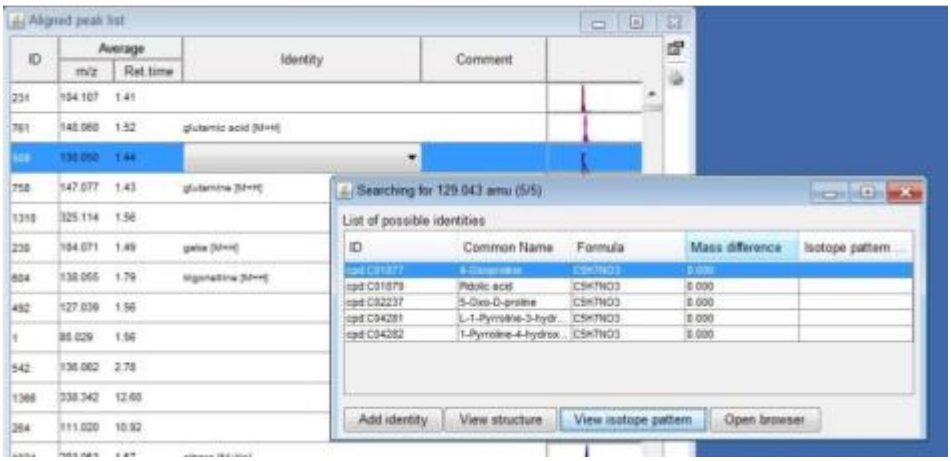
*Right click/Search/Search online database:*

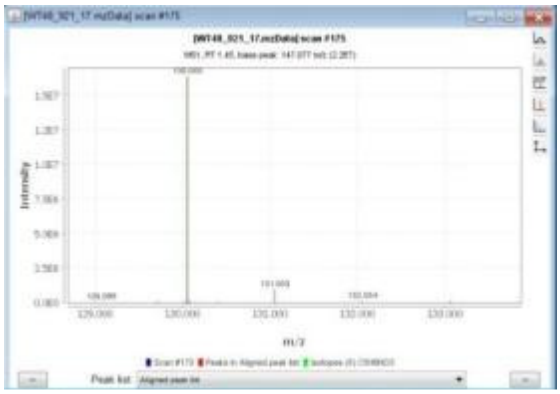

You are presented with a number of online database options: Let's try a search of KEGG. Set the charge appropriate to the technique (here it is ESI+ so we set +H). Set the m/z tolerance and the isotope pattern filter (as before)

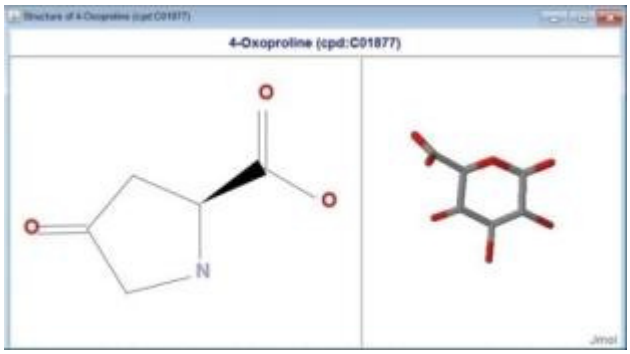



MZmine will start the search and any hits are displayed in a new window. The isotope pattern and structure may be viewed. If you think the structure is an appropriate match then the identity may be added using the *Add identity* button.



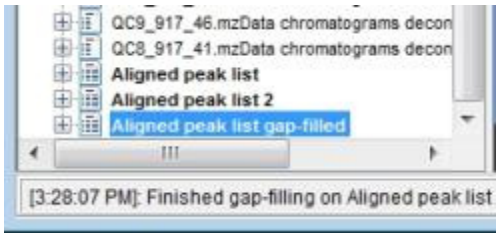



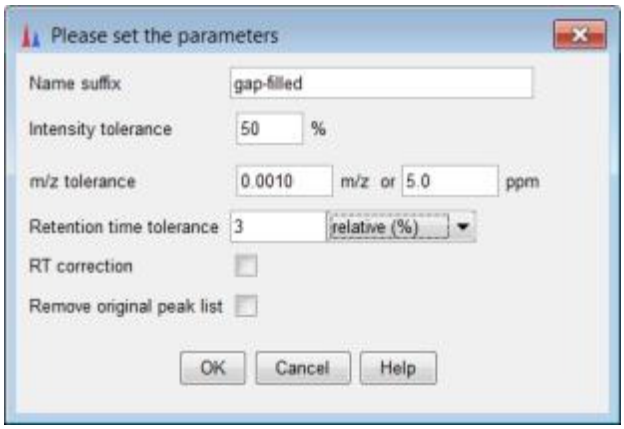

The other alternative is to search the whole list from the *Peak list methods/Identification/Online database search* menu**. <u>WARNING If you do this be prepared for many hours of deleting irrelevant peaks!</u>**

**<u>[PLEASE DO NOT DO THIS DURING THE DEMO FOR BANDWIDTH REASONS !!]</u>**

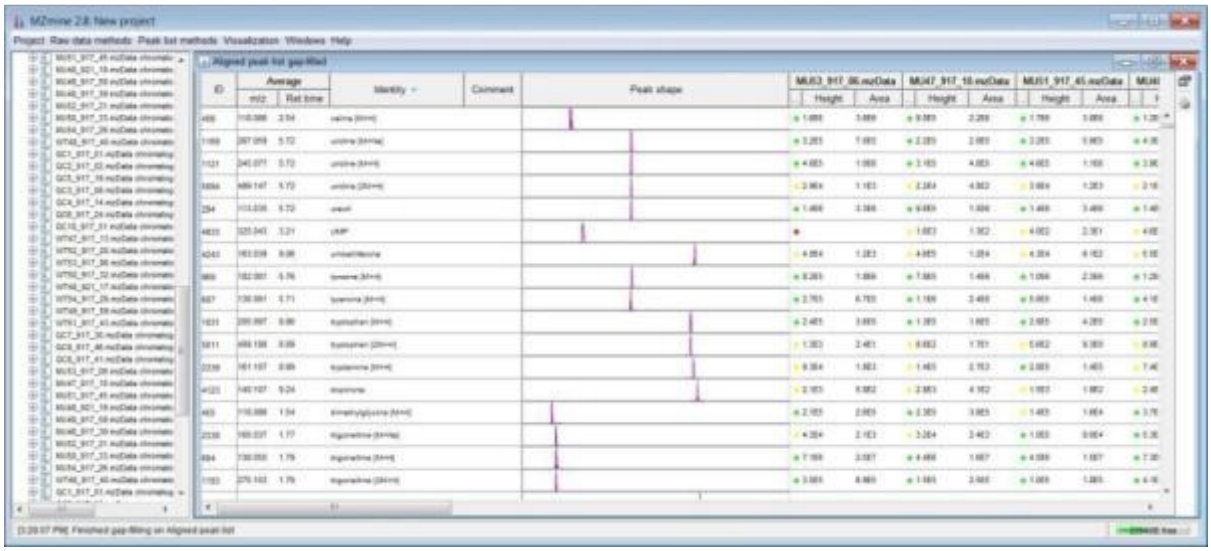
## Gap Filling

We will now try to fill in the gaps where peaks were detected in some scans but not others. There are a number of occasions where a peak may be present but not detected well due to being close to the detection limit in some samples. Gap filling is done by searching the target window where a peak was detected and looking for appropriate peak features in that window. There are two options "Peak Finder" or "Same mz and RT range gap filler". Let's first use the "Peak Finder" option. The gap filled peak list will appear as a new item in the left hand pane.



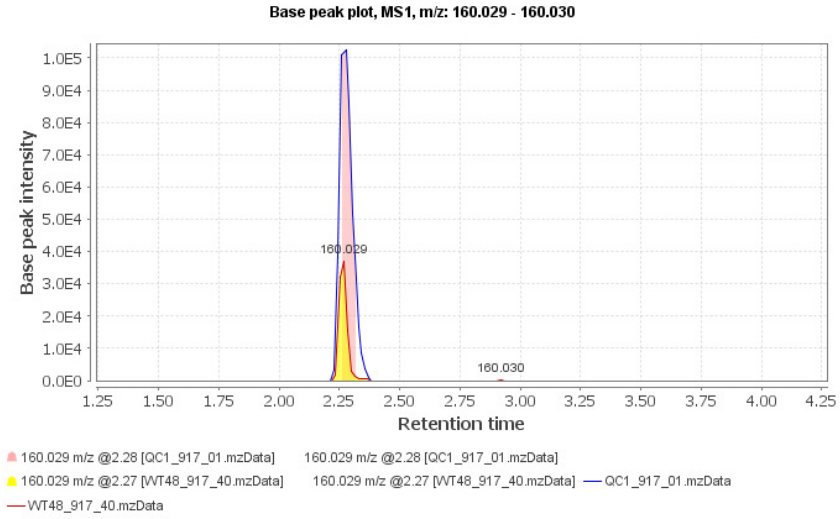




Here is the final gap filled spreadsheet peak list. Filled gaps are shown with a yellow icon. There may still be some gaps in which no evidence for the peak was found.



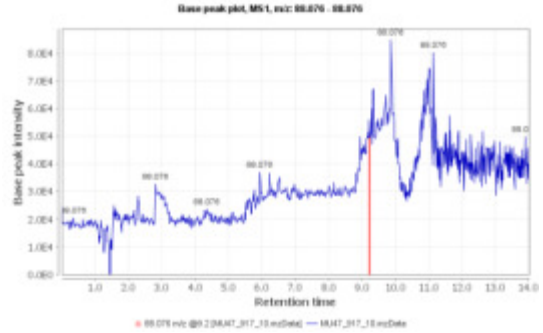
By right clicking on the peak list and *Show/Chromatogram (dialog)* and carefully selecting a scan with a detected peak and a scan with a filled peak you can see if the filling has been done in a sensible way. (Right clicking on the chromatogram and selecting *Show /Chromatogram (dialog)* brings up all

peaks overlaid but right clicking on an individual peak column and selecting *Show/Chromatogram (quick)* brings up just that peak). In the chromatogram the pink peak is the original and the yellow is the Gap filled by Peak finder. Note: in this case the earlier decision to set the baseline high may be causing a non-optimal integration of larger peaks. Peak picking is always a compromise between detection and accurate peak representation. (From the authors personal experience the XCMS wavelet method seems to be a more robust peak picker in practice).

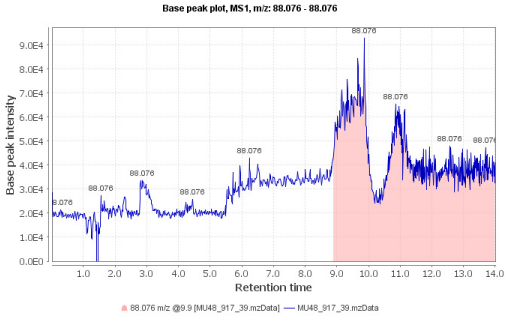



However the peak finder option can often backfire because it may detect previously removed broad artefact peaks.

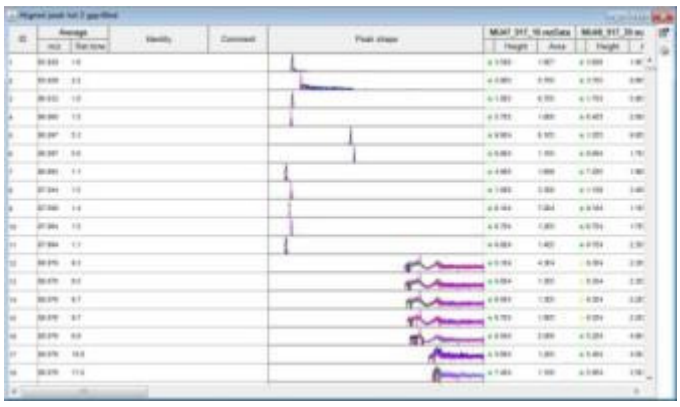
Originally detected peak



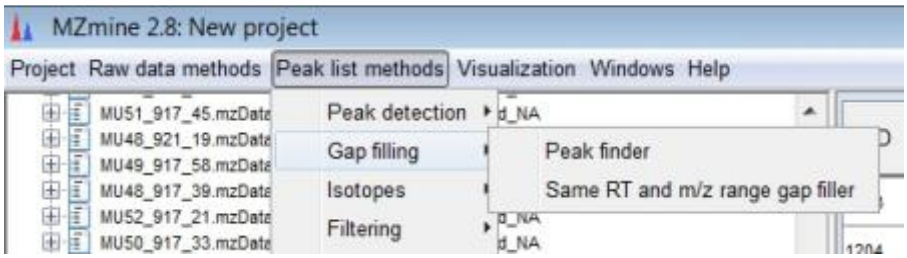
Gap filled peak

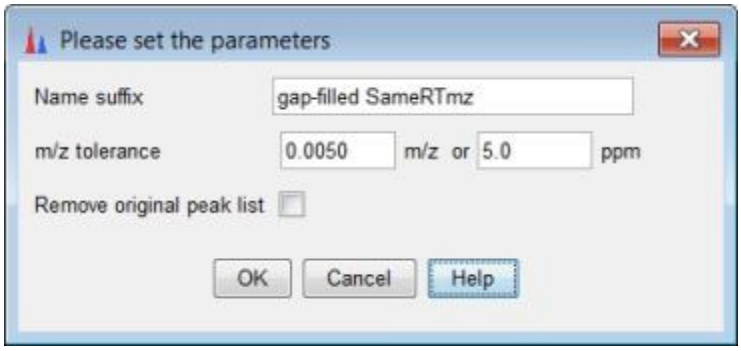


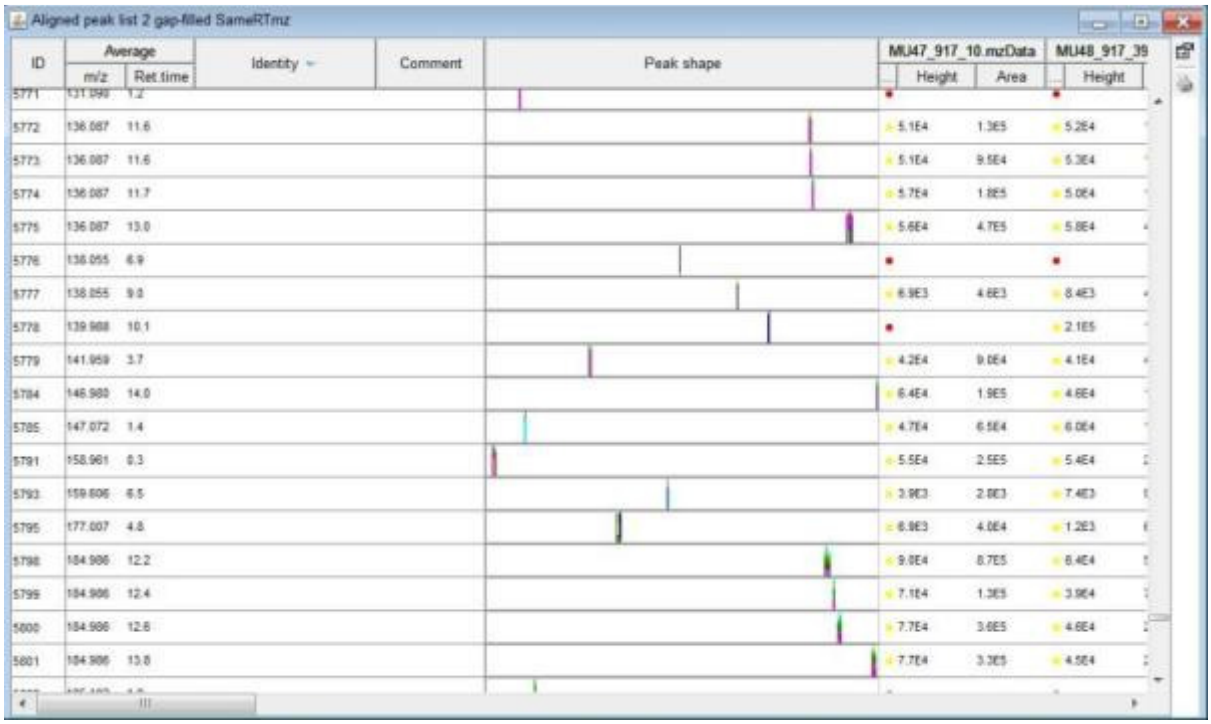

This means our carefully removed artefacts are now back with us! An alternative way of looking at this is that we can used this as a way to detect peaks that may be artefacts anyway!



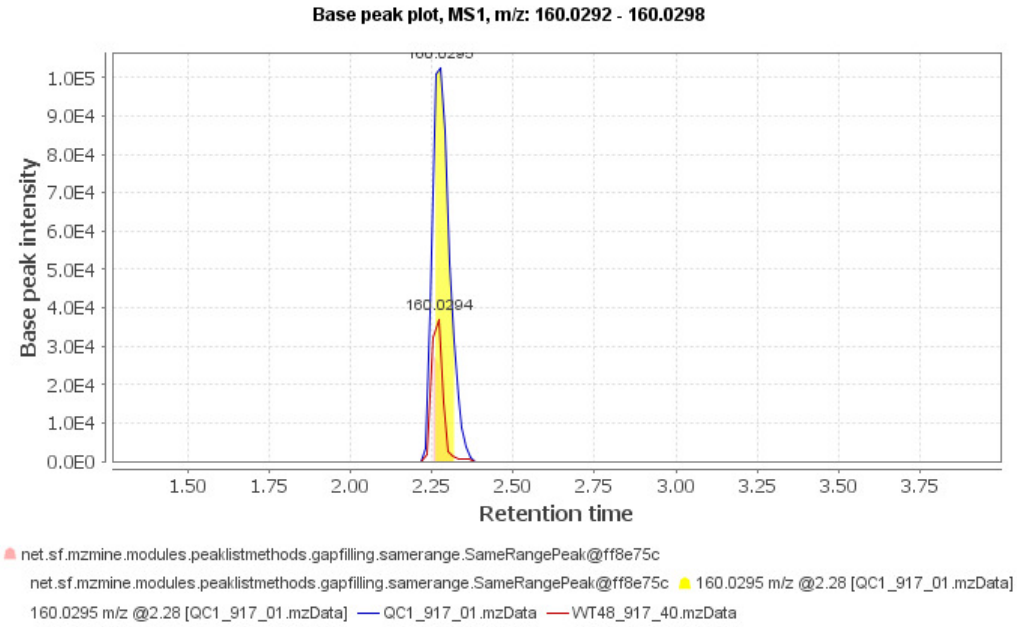
The other alternative for gap filling is the "Same RT and m/z range gap filler" this limits the gap fill to features within the original detected peak window. This results in much cleaner results.



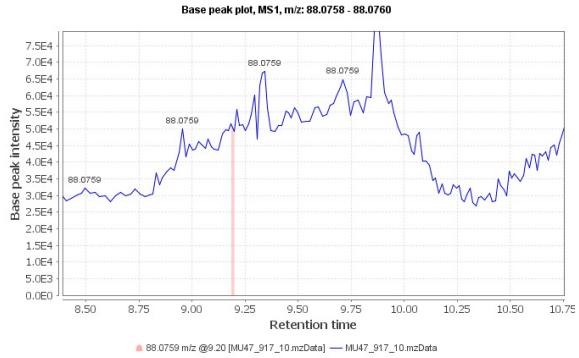



If we look at the same example as above we can now see the detected peak is now cut rather than detected in full. In such cases the hope is that the peak cut-off is applied consistently across all peaks to preserve relative quantitation. Again a compromise is made.

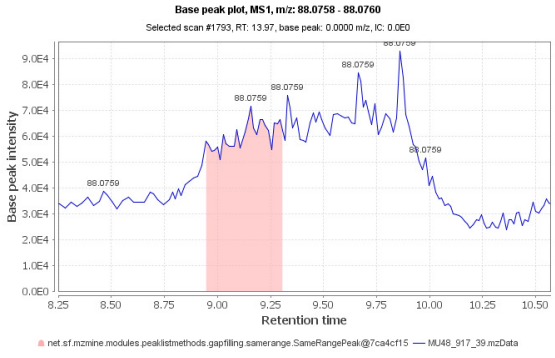



Looking again at one of the artefact peaks notice the gap filled peak is now defined by the mz tolerance and RT tolerance. The gap filled is broader than the original peak due to the RT window which is defined by all the peaks in the row. The variation in the retention time across the row is a function of the earlier tolerances used in both RT Normalisation and Join Align. This demonstrates the need to be careful when setting up the parameters from the very beginning.

Originally detected peak




Gap filled peak

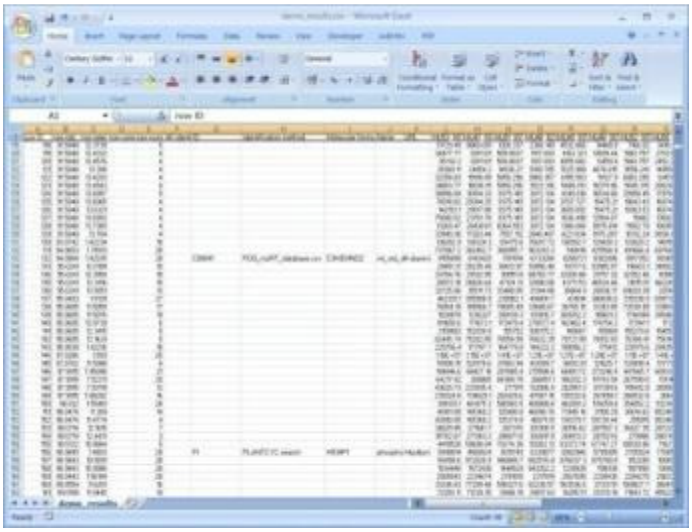



Despite the limitations of gap filling it is far preferable to have some estimate of baseline levels than to report the value as missing for later statistical analysis.

## Export of results

The Export to CSV option allows the export of the final spreadsheet.


The final exported data is displayed in excel. At this stage it is probable that the data will be processed further in a commercial data analysis package but as we will see in later examples there are some possibilities to use open source tools to analyse metabolomics data.

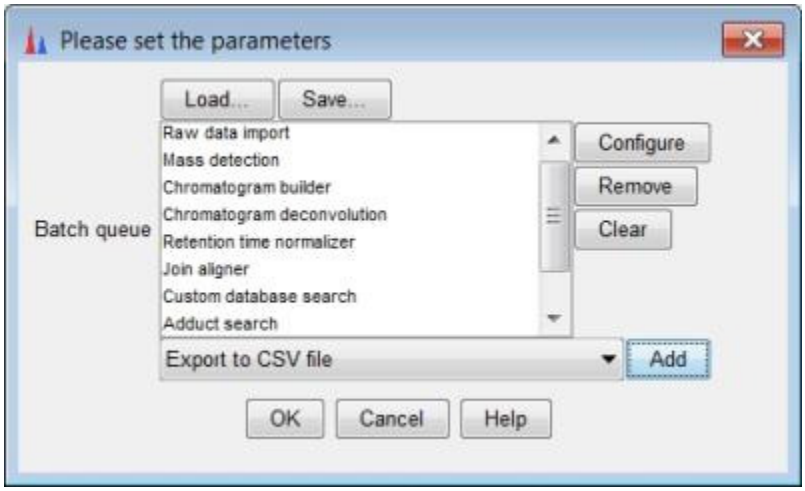


## Batch analysis

MZmine contains a Batch mode tool which allows a chain of processes to be set up which is a very useful feature with large datasets as the processing can be set to run overnight in unattended operation. The output of the previous operation is fed to the next operation.

Another useful feature is that once the parameters for a particular operation have been set up MZmine remembers the last used settings so we can apply the peak picking we developed above to every sample in a study.

The MZmine Batch command is to be found under the Project menu. The screenshot below shows a typical sequence.

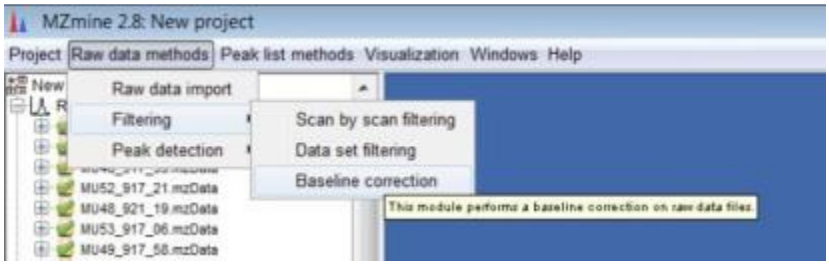



A recently added feature is the ability to save or load batch sequences.

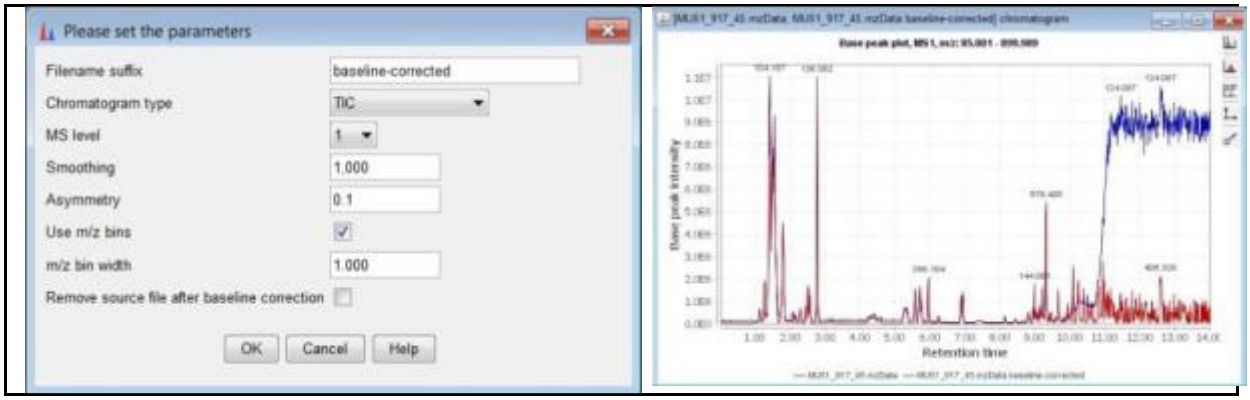
As an exercise try adding the steps you went through above to a new batch sequence.

## Baseline Correction

Another new feature is baseline correction. This is applied to the raw data before peak detection. It is found under *Raw data methods/Baseline correction*.




The baseline correction dialog box has two main options, the Smoothing and Asymmetry factors. Try playing with different settings of these factors and comparing the TIC plots before and after (NB: Ensure the "Remove source file after baseline correction" is switched OFF

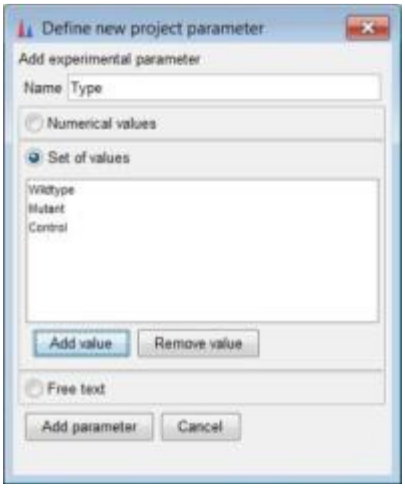


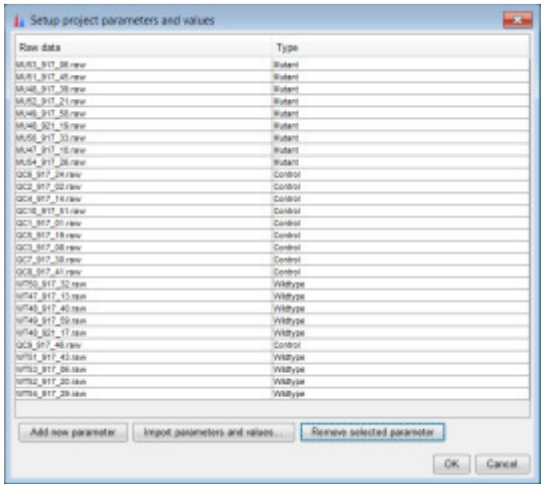
## Initial data analysis in MZmine

Limited data analysis tools are included in MZmine. In order to use them it is necessary to 'Set Sample Parameters' under the project menu. You can define as many parameters as you wish.

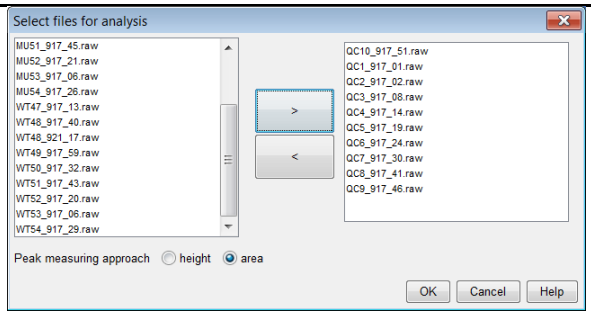
In the example data we have an excerpt of a metabolomic study on the ripening of fruits. We have nine samples of two different varieties, Wild-type and non-ripening Mutant plus ten control samples which consist of a large batch of identical fruit extract that are run at every fifth sample. In addition the fruit are sampled everyday from the onset of ripening between 47 and 54 days. (This example datasets is only a small excerpt of a larger replicated study).

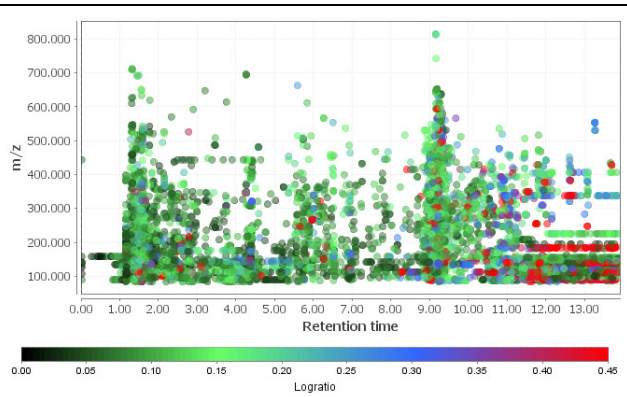
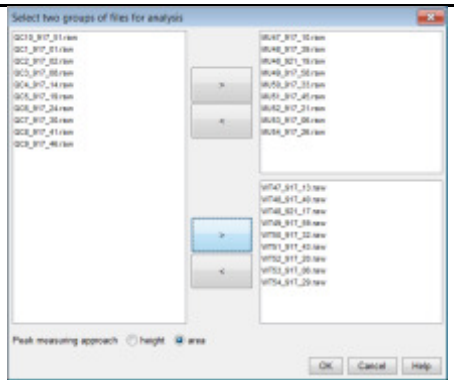
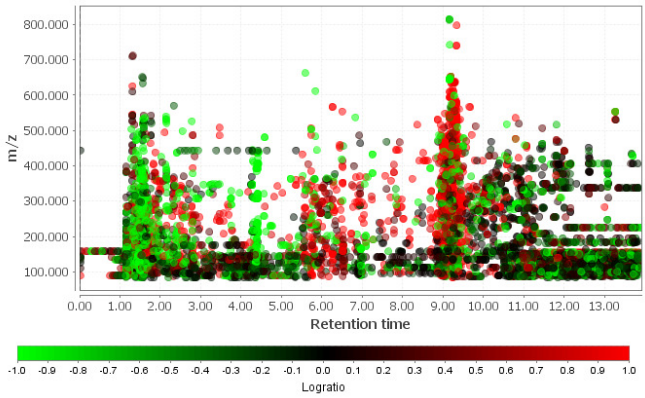
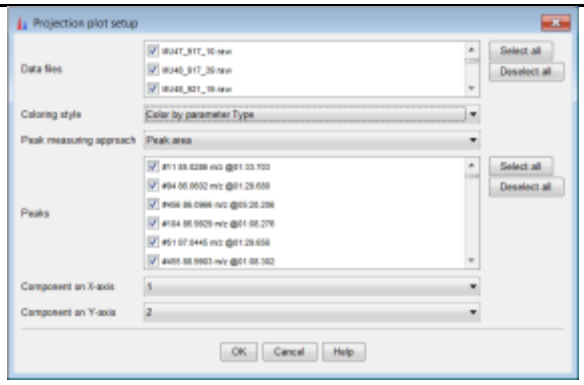
We now set a new experimental parameter called Type with the values "Wildtype", "Mutant" and "Control"

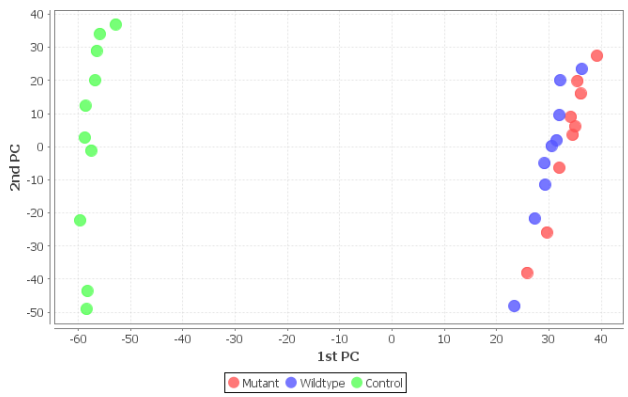
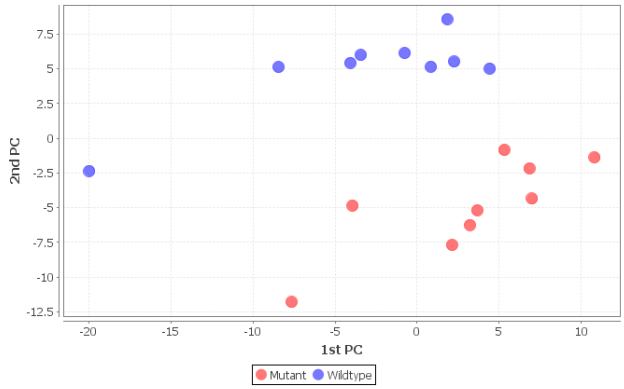
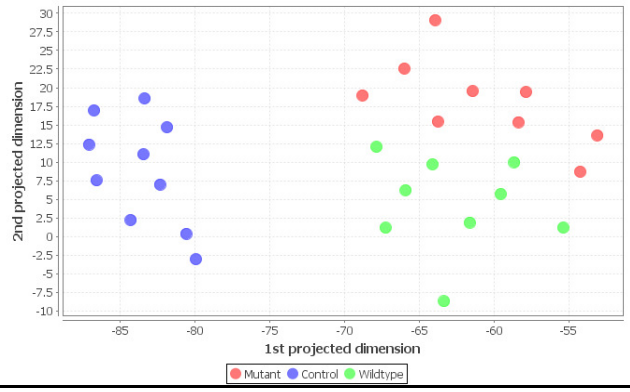



MZmine has a number of data analysis options, *Coefficient of variation (CV) analysis*, *Log Ratio Analysis, Principal Component Analysis, Curvilinear Distance Analysis, Sammon's projection*, and *clustering*. The most useful of these options are described below:

| | |
|---|---|
| **Coefficient of variation analysis** calculates the coefficient of variation of each peak and displays the result as a colour coded plot. *(Ensure you have your final peak list highlighted)*<br><br>In this case we have selected the 10 control samples. The graph shows that some of the later peaks are showing some unexpectedly high variation. |  |

| | |
|---|---|
| |  |
| **Logratio analysis** looks at the difference between two groups. It is the ratio of the natural logarithm of the ratio of each group average to the natural logarithm of 2<br><br>Here we have selected to compare the Mutant to the Wild-type. The display shows some metabolite peaks are quite differently expressed between the groups |  <br>  |
| **Principal component analysis** is a dimension reduction method of multivariate analysis.<br><br>The first principal component show a separation of the control samples from the Wildtype and Mutant<br><br>The second principal component shows a vertical trend. This at first glance appears to be related to ripening but on closer inspection the control samples also exhibit the same effect, showing that the variation is almost certainly due to spectrometer drift. The data has not yet been normalised to the isotopic internal standards which were spiked into the samples. (This |  |

| | |
|---|---|
| operation at the moment has to be done outside MZmine).<br><br>Removing the control samples we begin to see some separation of Wild-type from Mutant.<br><br>Unfortunately there is no corresponding loadings plot with which to interpret the results. The PCA function is purely a quick visualisation tool, more sophisticated analysis is left for specialist chemometrics software outside MZmine. |  <br>  |
| **Sammon's projection** is a non-linear multidimensional scaling method which projects multidimensional data down to just two dimensions. It is a useful as a method to examine approximate clustering in data but offers no interpretability. |  |

## Other features in MZmine

There are many more features in MZmine, including some support for ms/ms data and formula prediction. More features are being added all the time, recent developments in this area include links to the NIST MS Search program to allow the use of MZmine for GC-MS data.

## Get Involved !

Please join the community! Not just for programmers - testers and document authors always appreciated!

## Developers Mailing List:

http://sourceforge.net/mailarchive/forum.php?forum_name=MZmine-devel

## Acknowledgements:

I would like to acknowledge the following people for help with the current round of MZmine development.

Tomas Pluskal, Matej Orešič ,Mikko Katajamaa, Carsten Kuhl, Ralf Tautenhahn, Steffen Neumann Christoph Steinbeck, Stephan Beisken, Chris Pudney, Mark Seymour, Mark Forster, Martin Cip, Dave Portwood, Aniko Kende , Madalina Oppermann, Erik Johansson, Johan Trygg.

I would particularly like to thank those in the XCMS community who were very gracious in allowing us to combine some of the features into MZmine and to Chris Pudney our hardworking programmer.

Mark Earll mark.earll@syngenta.com

Wednesday, 16 May 2012