IROA TECHNOLOGIES
...making metabolic profiling easy®

# ClusterFinder FAQ

## Set Up

**Q.** What are the requirements for running ClusterFinder™ software?

**A:** Hardware requirements:
1. Intel Pentium III /800 MHz or higher (or compatible); dual-core processor or higher;
2. 16 GB RAM minimum
3. Computer needs to be network-accessible.

**B.** Software requirements:
1. Java 8 must be installed and callable;
2. The computer should have at least 8 GB of RAM;
3. Windows 7 or higher.

**Q.** Can we use a VM (Virtual machine)?

**A.** The program should be run on a computer that has at least 8 GB memory, preferably more, and access to the internet.

## Loading data into the program

**Q.** What kind of data should be loaded into the program?

**A.** Best data input for Clusterfinder is mzXML or mzData. The data should be centroided, preferably as an mzXML file. Profile data will be accepted by the program but it will match each peak several thousand times and make many false matches to the side elements in each peak.

Note, all major instrument vendors provide software for converting their proprietary data files into one of the formats supported by ClusterFinder™ (e.g., SCIEX provides MS Data Converter software freely upon registration). Use ProteoWizard to convert files to mzXML with 'Peak picking' applied as the filter as centroid is not an option. They are easily loaded into ClusterFinder and added to the appropriate sample names. 'Create library from raw data' which takes you to the untargeted analysis review tab.

**Q.** Is there a way to determine if the files are correctly processed by ProteoWizard, independent of whether or not they run on ClusterFinder?

**A.** Size is the best indicator. They will be about 10% of the original size although this is highly vendor dependent since some MS vendors compress the raw data.

**Q.** Can AW data be imported directly into IROA or does it need to be converted to a different format first?

**A.** AW data format would need to be converted to centroided mzXML.

## Data

**Q.** How is normalization achieved in the IROA datasets?

**A:** *Spectral TIC normalization is achieved using the MS total useful signal (MSTUS) approach* (developed by Warrack, B.M. et al. **2009**. *Chromatogr B Analyt Technol Biomed Life Sci, 877,* 547–552). In this approach, only the components that are common to all signals are used after baseline correction and removal of the artifacts (from xenobiotics or chemical noise, for instance) and nonbiological compounds that carry no IROA signatures.

## Error messages

**Q.** Error message "Could not reserve enough space for object heap…"

**A.** Not enough memory, use computer with at least 8 GB memory

**Q.** Error message "Invalid Heap Size"

**A.** Computer is most likely running a 32-bit Java and not a 64-bit Java. The 32-bit has a memory maximum of 4G. Download java-64 bit.

**Q.** "Unable to contact to time server" and after hitting OK the second one comes up and says "Your license has expired. Please contact IROA technologies for license renewal and a more up-to-date version of the software." When you hit OK it closes the program. What does this mean?

**A.** The problem is most likely that that the corporate firewall (at user's location) whereby the server cannot verify the timeclock on the software and is not permitting any downloads to be run. The install may need to be done by their IT department who will download and most likely test it to ensure that it does not contain a virus.

**Q.** GC overhead memory limit exceeded – what does this mean?

**A.** Java has a Garbage Collection (GC) process that tries to free up memory. This message is saying that you are out of virtual memory. Manually change HEAP SIZE, depending on your capacity, for example change "HEAP_SIZE =8' to "HEAP_SIZE=32".

**Q.** What does the following error message mean? [15:56:24|SEVERE|WorkerThread]: Task error: Could not open file M:\Development\CF\Projects\T1\Libraries\t1.cflib.csv for writing.

**A.** You forgot to set the directory to your current project directory. You can set it by opening the advanced preferences, the first tab.

## Preferences

Changing the preference settings in ClusterFinder does drastically change the run time. Try different settings to get a good number of metabolites to create a library that can be used to identify metabolites from on-going samples. IF after you try to perform the IROA Unbiased Data Analysis and nothing happens look at the settings in Preferences. Your settings are likely too high.

**Q.** We also acquire with continuous reference mass correction, and so theoretically both the Agilent 6550 and 6545 Q-TOFs should be giving us <2ppm mass error, but we are also finding a 40ppm extraction window is needed to pull out all the correct 13C isotopomers (I think because of worse ion statistics on some of the low abundance mass isotopomers in any cluster?). What ppm window you recommend when working with Agilent data?

**A.** We see the same phenomenon. The very small and very large peaks (non-linear range peaks) show a fair bit of mass drift. If two peaks are "superimposed" then the algorithms for centering the apex are thrown off even in linear range peaks. While these may be rare events they are common enough that a ppm of between 35 and 50 can be used depending on the samples.

**Q.** In the settings, when I set the minimum peaks or noise to 25 (just an example) it will round to 2E1- is it actually set on 25 and the notation just shows 2E1 or is it actually rounding?

**A.** In the general settings uncheck the exponent box to allow it to show the actual number, rather than rounded number.

**Q.** What is the difference between minimum peak height and noise?

**A.** The minimum height is the smallest peak it will use to put together a cluster, and will be the largest peak in that cluster. The minimum will be the smallest peak it will use, i.e. below the minimum is "noise". Set the minimum at a value that looks right based on a visual examination of the spectra in a quiet region. The noise is set from the same place and represents the value below which there seems to be just too much noise.

## Curation

**Q.** How do you go about selecting certain metabolites to add (the ones that are green), versus confirming or throwing out the others (yellow or red ones).

**A.** Make sure the compound shows up in 2 or more samples, that it has a good chromatographic shape and the IROA peaks look good and as expected for the number of carbons. If they are unknowns, also see if there is a compound that has the same time-signature. If so, it is likely a fragment.

Note: ClusterFinder now supports moving bins by drag and drop to define parent-child relations. To remove the child you have to drag and drop it on the root of the tree. Parent-child relationships are preserved in the correlation analysis and also when you save and re-open the project.

## Saving the experimental data files

**Q.** I am having a problem with not being able to save the untargeted analysis runs, even after changing the directory I am having to rerun this step each time. The runs are much shorter **now** that I have altered the settings, but I was hoping to be able to not re-run each time. Right now, I am opening the project each time from the option on the startup screen.

**A.** What is saved are the files (which will reload), and the preferences you last used. When you have curated a particular setting in preferences write out a library for that setting combo. If you have enough signal you can run an analysis that will find several hundred IROA peaks in approximately 1 min. If this is not true next time concentrate the sample. During sample prep 1) the cells are extracted and the extract is dried down. The volume in which it is brought back can be adjusted down to increase the concentration. The sample may also need to be larger. The sample preparation is every bit as critical as the analysis.

**Q.** Is there another location that I should open from that may be saving the results in this tab? I've looked in the results folder for this project, but it is empty. Is there a particular location this would save to within a project folder?

**A.** Go into preferences access expert mode and look at the setting in the export tab. Reset the settings if necessary. This is where it will default to.

## Functional MS calculator

**Q.** What is the Functional MS calculator (called by calculator icon on toolbars)?

**A.** The Functional MS calculator It allows 2 things:
- calculate mass difference in ppm between two masses
- calculate isotope distribution from formula, adduct type and %C13

## Data Output

**Q.** We are running into some problems with the ClusterFinder output file. The current arrangement for an experiment with 8 samples is to have 8 rows per compound / IROA bin. This works well for most compounds that are detected in every sample, maybe with abundance differences between groups, but not for a compound like chorismate, which is only detected in the WT and is completely absent in the mutant. For chorismate (and a few others) there are only 4 rows and this difference makes it very hard to reshape the data table into a format: samples (rows) x compounds (columns) which is the required input for most statistical testing.

We've been trying to come up with ways to do this in Excel with the VLOOKUP function or variations on it, but it's proving challenging. There is a tab on your analysis spreadsheet there the table has been reshaped and missing values replaced with estimates, so we wondered if this means you already have a fast way to do this? Or could we request the initial ClusterFinder output table be reformatted so the number of rows per compound is always the same as the number of samples, even if there are empty cells / missing values when a compound is not detected.

**A.** All numbers are positive and usually have a minimum of, say, 50,000. We take all NAs and replace them with a random selection of numbers between 0 and 0.01. We understand this is a radical replacement but it creates a non-sparse dataset. Remember that we are always coming from a targeted analysis in which we are seriously looking for each compound.

**Q.** Was I mistaken in thinking that IROA/CF can identify the chemical formula for “13C orphan compounds”? I was thinking before that it would be possible (based on inferring the number of carbons and thus the chemical formula from the isotopic masses observed. Now it seems I was wrong to think that and in fact one should never run a 95% 13C sample, but only analyze it together with natural abundance or 5% 13C.

**A.** The number of carbons can be inferred from the height of either the C12 M+1 or the C13 M-1 but there can be enough error (insufficient labeling, instrument error, etc.) that we do not program it to do so. If you wish it is easy enough to manually look at the peak in question take a read and use it to reduce the number of other possible formulae at that mass. This will result should be unique. While there is always some error in the percent labeling if you look at the number of carbons in known molecules and the height of the M-1 then you should be able to figure out if the incorporation was good, i.e. that 95% is correct. We have seen experiments with more complex systems where despite feeding 95% media cells only attained 92% usually this means they were not grown long enough.

## IROA Publications

**Q.** I came across the following figure in the IROA publication, "
Addressing the current bottlenecks of metabolomics: Isotopic Ratio Outlier Analysis, an Isotopic labeling technique for accurate biochemical profiling"
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3696345/?tool=nihms
Have you ever done any comparison between a normal HR (high resolution) workflow to that of IROA to highlight its peak identification capabilities? From I understand the actual HR instruments are pretty powerful to provide the isotopic information and narrow a peak to a few molecular formulae candidates. What would it be the real gain of using IROA in this context compared to regular HR workflow? Of course, IROA technology has other great attributes of such as the elimination of artifacts and noise and also decreasing sample-to-sample variability, that are unquestionable. Still, the improvement in peak identification are somehow not very clear to me.

B

| Molecular Formula | Formula Mass | PPM Error |
|---|---|---|
| $C_{26}H_{37}N_2O_7$ | 479.3009 | 0.17 |
| $C_{11}H_{43}N_{15}O_2S_2$ | 479.3009 | 0.17 |
| $C_{18}H_{51}N_6OP_2S_2$ | 479.3010 | 0.38 |
| $C_{19}H_{50}N_6P_3$ | 479.3006 | 0.46 |
| $C_{18}H_{50}NP_3S_6$ | 479.3006 | 0.46 |
| $C_{17}H_{46}N_{13}OP_2$ | 479.3005 | 0.67 |
| $C_{23}H_{46}NO_7P$ | 479.3012 | 0.79 |
| $C_{28}H_{45}N_3OS_2$ | 479.3004 | 0.88 |
| $C_{33}H_{49}N_{11}OS_4$ | 479.3004 | 0.88 |
| $C_{15}H_{53}N_{14}O_3$ | 479.3014 | 1.21 |
| $C_{27}H_{47}N_6O_4P_2$ | 479.3002 | 1.29 |
| $C_{28}H_{51}N_6O_4S$ | 479.3002 | 1.29 |
| $C_{28}H_{53}N_6O_7P_2$ | 479.3015 | 1.42 |
| $C_{19}H_{57}N_{13}S$ | 479.3016 | 1.63 |
| $C_{10}H_{57}N_{15}O_7$ | 479.3000 | 1.71 |
| $C_{27}H_{48}NP_3$ | 479.3000 | 1.71 |
| $C_{17}H_{52}N_9P_3S_2$ | 479.3000 | 1.71 |

C

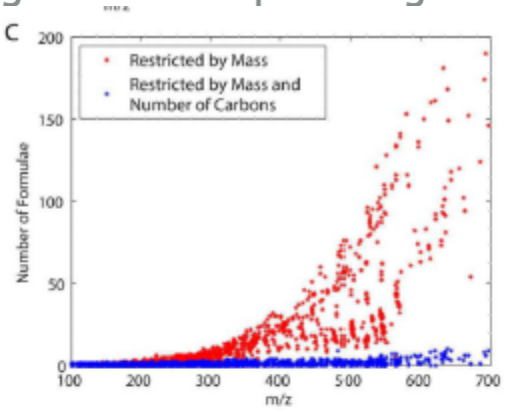


**Figure 2. IROA allows for the discrimination between biological molecules and artifacts and constrains the number of possible molecular formulae**
(A) Representative mass spectrum from a single scan in the IROA experiment. The blue peaks indicate isotope peaks originating from a single biological compound, tentatively identified as the $[M+H]^+$ of lysophosphatidylethanolamine 18:1. An $[M+Na]^+$ peak was also observed helping to confirm the protonated form. Red peaks originate from background (noise) or other biological compounds. The fold-change of this compound can be quantified by determining the ratio between the sum of the intensities of the unlabeled $^{12}C$ peak (480.3081) and its associated isotopic peaks (481.3108, 482.3140, etc…) to the sum of the intensities of the fully labeled $^{13}C$ base peak (503.3860) and its associated isotopic peaks (502.3807, 501.3781, etc…). (B) A table detailing the possible molecular formulae for the monoisotopic mass of this compound. Of the 17 possible molecular formulae within 2 ppm mass error for the compound in (A), only one has the correct number of carbons, $C_{23}H_{46}NO_7P$ (highlighted). (C) The number of possible molecular formulae for a compound is greatly restricted when exact number of carbons is used as a constraint. The possible molecular formulae within 2 ppm for 3131 IROA peaks were generated with (blue) or without (red) constraining for the number of carbons. For both (B) and (C), the formulae were generated using HR2, allowing the elements C, H, N, O, P, and S and a mass error of up to 2 ppm. Formulae were filtered using the seven golden rules with the exception of the isotopic pattern filter.[46]

**A.** The figure represents the following:

1) All possible formulae were generated using the elements listed. 2) These were then screened through the seven golden rules. 3) The graph was made under the assumption of no isotopic involvement, i.e. the monoisotopic mass of each formula was calculated. 4) The graph plots the number of options at each mass (assuming a 2 ppm window) using or not using the number of carbons constraint.

**Q.** Our work here is done on an Orbi at a resolution of ~70,000 (or ~35,000 if we do fragmentation). There are still way too many ambiguities as to what formulae are correct. Our point/belief is that is you have a mass and the number of carbons in the molecule then the dual constraint make the formula almost always unambiguous.

**A.** That was in fact the point of the Figure 2C above. The reality is that an exact mass is only an exact mass within a fairly tight window. If the peak is not in the middle of the linear range of the detector it will usually drift. It may not be much at first but as you get further out of the window it gets worse fast. The dual constraint relies less on the resolution. The fragmentation of IROA peaks is valuable. If you have fragmentation patterns for both monoisotopic peaks then you will always know the formula for the fragment which helps if the compound is an unknown and you are trying to do some structure elucidation.