Input files preparation

LipidFinder v2.0 has improved its flexibility towards the layout of input CSV files. The first stage, PeakFilter, expects input data to be pre-aligned and framed using XCMS and stored in one or more comma-separated values (CSV) files. Moreover, we have designed the parameters to allow input files from other pre-processing software tools (e.g. $SIEVE^{TM}$). Amalgamator and MSSearch expect the input CSV file to be as the output file of their corresponding previous stage. It is important that the input files are in the expected format to avoid errors or misguided processing. This guide details how to prepare and format your data to leave it ready to be processed with $PeakFilter^{1}$, Amalgamator and MSSearch.

XCMS file pre-processing

The following steps will guide you through the pre-processing phase with XCMS:

- 1. Generate **mzXML** files from your **raw** files with a software like <u>ProteoWizard</u>.

 There is a useful guide provided by XMCS about this topic <u>here</u>.
- 2. Next, you can use either <u>XCMS online</u> or the R script we provide in the same directory as this document to pre-process your dataset. The following steps are for the R script option.
- 3. The mzXML files must then be located in **two subfolders**: one must contain the biological samples (and quality control samples if any), and the other must contain the blank (solvent) samples.
- 4. Run the **R script** we have developed to pre-process your files with XCMS (included in the same folder as this document). The script will ask for the folder where the two previous subfolders are located, as well as the name to give to the output CSV file, and if you want to just annotate or also remove the isotopes.² Depending on the number of files, their size and the computer being used, the pre-processing might take several hours to finish.
- 5. Remember to check and select the corresponding **time unit** of the **retention time** before running *PeakFilter*. If it is in seconds, it will automatically convert this column to minutes before processing the data.

PeakFilter's input CSV file layout

All restrictions are already met by any output CSV file generated from the preprocessing options:

- The **first column** must always be the **frame ID**: a number that uniquely identifies each row.
- There must be an **m/z and retention time columns** anywhere in the file. Their name is now part of the parameters (*mzCol* and *rtCol*).
- All **intensity columns** must be *together* but *anywhere* in the file, preserving the following order:
 - i. **Sample replicates**: n biological samples with m replicates **each**. When m>1, all replicates for the same sample must be together and they should have a column name in the format name1, name2, etc. such that name is

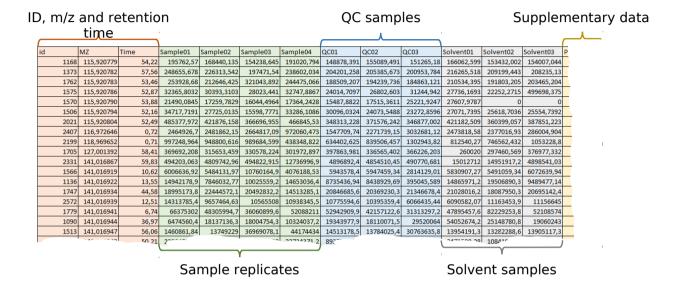
¹ Note that every parameter in both pre-processing tools has been devised for high resolution liquid chromatography/mass spectrometry (LC/MS) data.

² Note that the functions included in the R script have **several parameters that should be tailored** to your experimental conditions.

unique for each sample and replicates are suffixed with numbers in ascending order starting from 1.

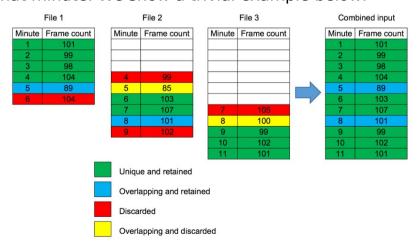
- ii. Quality control (QC) samples [optional].
- iii. Solvent (blank) samples [optional].

There can be any number of additional columns (uniquely labelled) with supplementary data. Below you can find an example of how the input CSV file should look like.



Other considerations:

- 1. All retention times should be greater than zero.
- 2. Column names should **only** consist of alphanumeric characters, hyphens ("-") and underscores (" ").
- 3. PeakFilter supports the use of multiple files split by time ranges to represent a single run. However, note that except for the first retention time minute of the first file and the last retention time minute of the last file, all first and last minutes are trimmed from the data since they are unreliable. The file import procedure supports overlap (after trimming). Where retention times overlap (in minute chunks), the frames retained are those from the file with the most frames for that minute. We show a trivial example below:



Amalgamator's input CSV file layout

The following restrictions must be fulfilled by both negative and positive input CSV files:

- The **first column** must always be the **frame ID**: a number that uniquely identifies each row.
- There must be an **m/z and retention time columns** anywhere but with the same name in both input files. Their name is now part of the parameters (*mzCol* and *rtCol*).
- There must be a **Polarity** column (case-sensitive) anywhere in each file.
- All sample **mean intensity columns** must be *together* but *anywhere* in each file (can be in different order).

MSSearch's input CSV file layout

MSSearch's input CSV file must reflect the following restrictions:

- The **first column** must always be the **frame ID**: a number that uniquely identifies each row.
- There must be an **m/z and retention time columns** anywhere in each file. Their name is now part of the parameters (*mzCol* and *rtCol*).
- There must be a **Polarity** column (case-sensitive) anywhere in each file.