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™). *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 *ProteoWizard*.
   There is a useful guide provided by XMCS about this topic here.
2. Next, you can use either XCMS online 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 pre-processing 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.

ID, m/z and retention time

QC samples

Supplementary data

id	MZ	Time	Sample01	Sample02	Sample03	Sample04	QC01	QC02	QC03	Solvent01	Solvent02	Solvent03	P
1168	115,920779	54,22	195762,57	168440,135	154238,645	191020,794	148878,391	155089,491	151265,18	166062,599	153432,002	154007,044	
1373	115,920782	57,56	248655,678	226313,542	197471,54	238602,034	204201,258	205385,673	200953,784	216265,518	209199,443	208235,13	
1762	115,920783	53,46	253928,68	212646,425	321043,892	244475,066	188509,207	194239,736	184863,121	210534,395	191803,205	203465,204	
1575	115,920786	52,87	32365,8032	30393,3103	28023,441	32747,8867	24014,7097	26802,603	31244,942	27736,1693	22252,2715	499698,375	
1570	115,920790	53,88	21490,0845	17259,7829	16044,4964	17364,2428	15487,8822	17515,3611	25221,9247	27607,9787	0	0	
1506	115,920794	52,16	34717,7191	27725,0135	15598,7771	33286,1086	30096,0324	24073,5488	23272,8596	27071,7395	25618,7036	25554,7392	
2021	115,920804	52,49	485377,972	421876,158	366696,955	466845,53	348313,228	371576,242	346877,002	421182,509	360399,057	387851,223	
2407	116,972646	0,72	2464926,7	2481862,15	2664817,09	972060,473	1547709,74	2271739,15	3032681,12	2473818,58	2377016,93	286004,904	
2199	118,969652	0,71	997248,964	948800,616	989684,599	438348,822	634402,625	839506,457	1302943,82	812540,27	746562,432	1053228,8	
1705	127,001392	58,41	369692,208	315653,459	330578,224	301972,897	397863,981	336565,402	366226,203	260020	297460,569	376977,332	
2331	141,016867	59,83	494203,063	4809742,96	494822,915	12736996,9	4896892,4	4854510,45	490770,681	15012712	14951917,2	4898541,03	
1566	141,016919	10,62	6006636,92	5484131,97	10760164,9	4076188,53	5943578,4	5947459,34	2814129,01	5830907,27	5491059,34	6072639,94	
1136	141,016922	13,55	14942178,9	7846032,77	10025559,2	14653036,4	8735436,94	8438929,69	395045,589	14865971,2	19506890,3	9489477,14	
1747	141,016934	44,58	18995173,8	22444572,1	20492832,2	14513285,1	20846685,6	20369230,3	21346678,4	21028016,2	18087950,3	20695142,4	
2572	141,016939	12,51	14313785,4	9657464,63	10565508	10938345,5	10775594,6	10395359,4	6066435,44	6090582,07	11163453,9	11156645	
1779	141,016941	6,74	66375302	48305994,7	36060899,6	52088211	52942909,9	42157122,6	31313297,2	47895457,6	82229253,8	52108574	
1090	141,016944	36,97	6474560,4	18137136,3	18004754,3	10324037,2	19343977,9	18110071,5	29520064	54052674,2	25148780,8	19060243	
1513	141,016947	56,06	1460861,84	13749229	36969078,1	44174434	14513178,5	13784025,4	30763635,8	13954191,3	13282288,6	13905117,3	

Sample replicates

Solvent samples

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.