

*MS-CleanR tutorial:
Peak list cleaning, data concatenation and peak annotation
07/04/2020*

Justine Chervin and Guillaume Marti

guillaume.marti@univ-tlse3.fr
justine.chervin@lrsv.ups-tlse.fr

Prerequisite :

Software installation

Downloading

MS-DIAL version up to 4.16:

http://prime.psc.riken.jp/Metabolomics_Software/MS-DIAL/index2.html

MS-FINDER version up to 3.30:

http://prime.psc.riken.jp/Metabolomics_Software/MS-FINDER/index2.html

R version up to 3.6.1 : <https://cran.r-project.org/>

R studio: <https://rstudio.com/products/rstudio/>

Installation

- In **R**, copy and paste the following command to update R version if necessary

```
if(!require(installr)) {  
install.packages("installr"); require(installr)}  
updateR()
```

- In **R studio**, update all your packages with the command

```
SetRepositories()
```

Select 1 and 2 for CRAN and BIOCONDUCTOR packages
Select the command Update on the right windows in the Package part

- Install MS-cleanR by copying and pasting the command :

```
devtools::install_github("eMetaboHUB/MS-CleanR")
```

MS-CleanR workflow

Within your project directory, create one subfolder for each ionization mode namely “pos” and “neg”. In each of this new directory, create another subfolder named “peaks”.

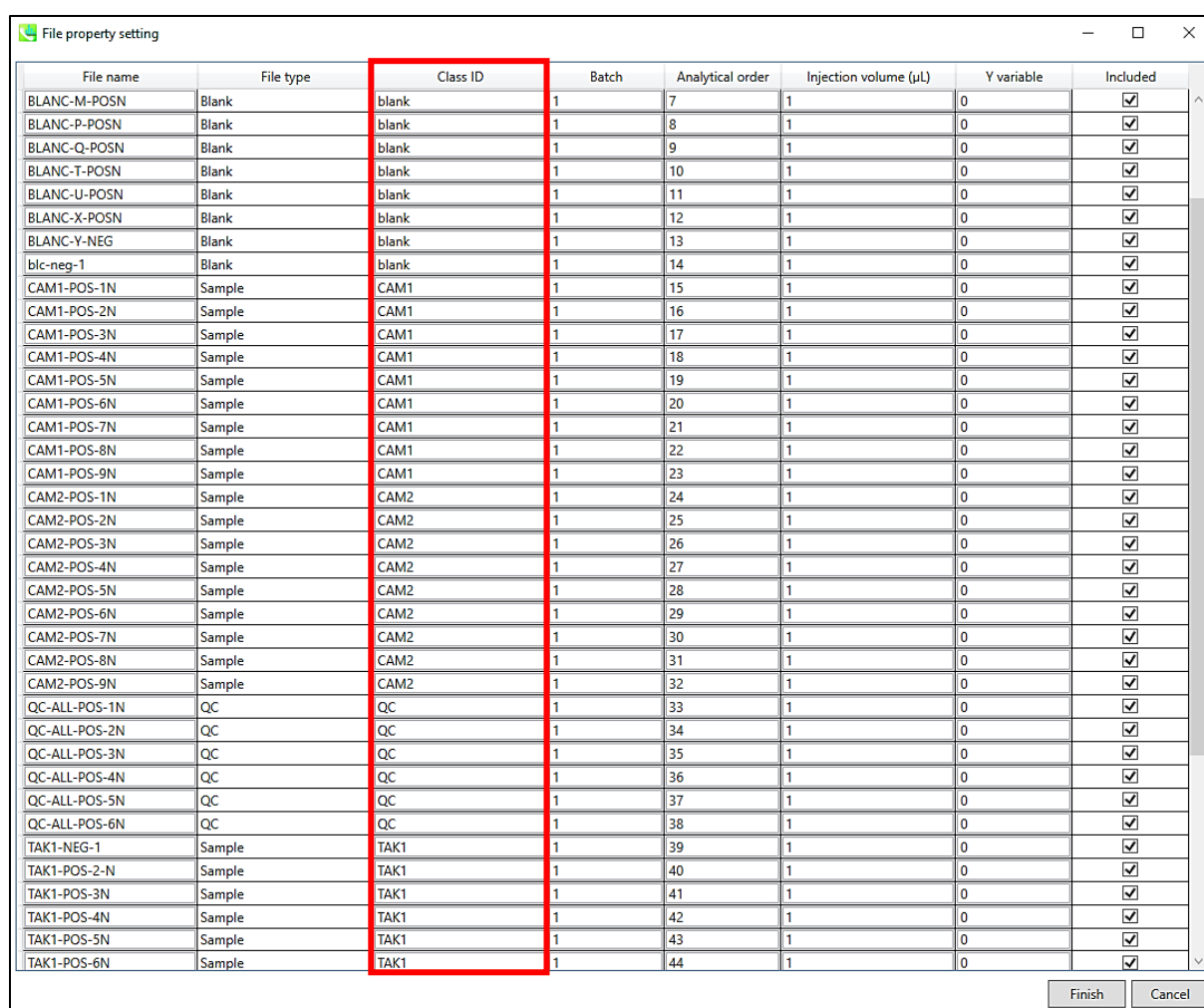
Optional: Only one ionization mode can be treated by MS-CleanR

Process the data with MS-DIAL

Process data with MS-DIAL in either pos or neg mode or both according to the tutorial <https://mtbinfo-team.github.io/mtbinfo.github.io/>

Important notices:

- A) During data importation, it is important to note the type (Blank, QC or Sample) and class of every sample in **Class ID column and File Type** (blank, sample class, QC)
- B) Be careful to have the **same number of samples** between pos and neg mode and in the **same order**.



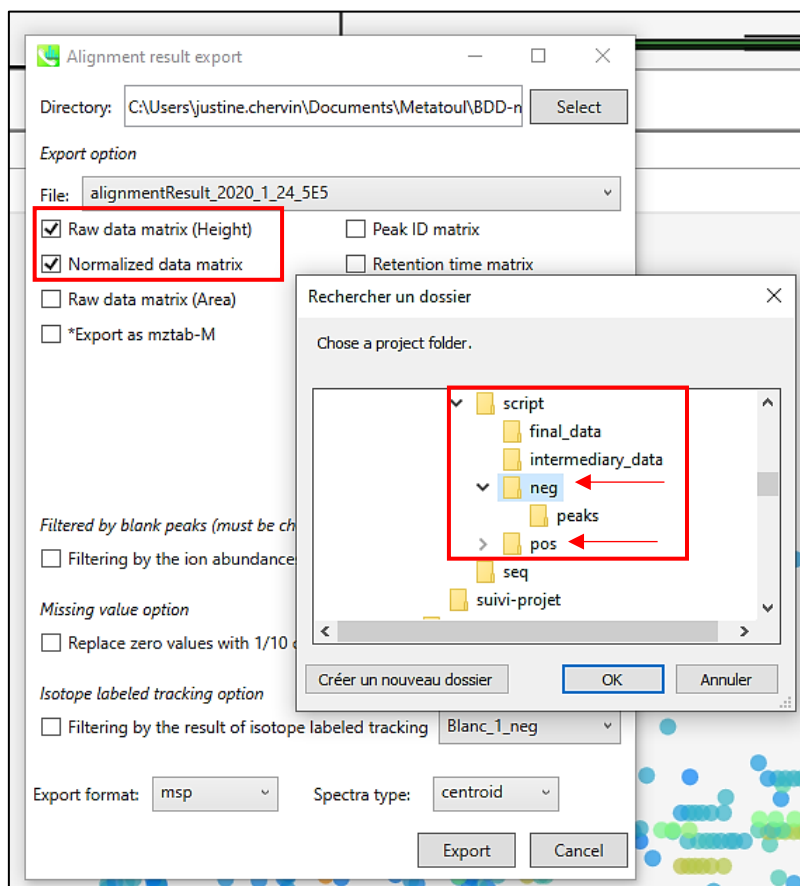
File name	File type	Class ID	Batch	Analytical order	Injection volume (μL)	Y variable	Included
BLANC-M-POSN	Blank	blank	1	7	1	0	<input checked="" type="checkbox"/>
BLANC-P-POSN	Blank	blank	1	8	1	0	<input checked="" type="checkbox"/>
BLANC-Q-POSN	Blank	blank	1	9	1	0	<input checked="" type="checkbox"/>
BLANC-T-POSN	Blank	blank	1	10	1	0	<input checked="" type="checkbox"/>
BLANC-U-POSN	Blank	blank	1	11	1	0	<input checked="" type="checkbox"/>
BLANC-X-POSN	Blank	blank	1	12	1	0	<input checked="" type="checkbox"/>
BLANC-Y-NEG	Blank	blank	1	13	1	0	<input checked="" type="checkbox"/>
blc-neg-1	Blank	blank	1	14	1	0	<input checked="" type="checkbox"/>
CAM1-POS-1N	Sample	CAM1	1	15	1	0	<input checked="" type="checkbox"/>
CAM1-POS-2N	Sample	CAM1	1	16	1	0	<input checked="" type="checkbox"/>
CAM1-POS-3N	Sample	CAM1	1	17	1	0	<input checked="" type="checkbox"/>
CAM1-POS-4N	Sample	CAM1	1	18	1	0	<input checked="" type="checkbox"/>
CAM1-POS-5N	Sample	CAM1	1	19	1	0	<input checked="" type="checkbox"/>
CAM1-POS-6N	Sample	CAM1	1	20	1	0	<input checked="" type="checkbox"/>
CAM1-POS-7N	Sample	CAM1	1	21	1	0	<input checked="" type="checkbox"/>
CAM1-POS-8N	Sample	CAM1	1	22	1	0	<input checked="" type="checkbox"/>
CAM1-POS-9N	Sample	CAM1	1	23	1	0	<input checked="" type="checkbox"/>
CAM2-POS-1N	Sample	CAM2	1	24	1	0	<input checked="" type="checkbox"/>
CAM2-POS-2N	Sample	CAM2	1	25	1	0	<input checked="" type="checkbox"/>
CAM2-POS-3N	Sample	CAM2	1	26	1	0	<input checked="" type="checkbox"/>
CAM2-POS-4N	Sample	CAM2	1	27	1	0	<input checked="" type="checkbox"/>
CAM2-POS-5N	Sample	CAM2	1	28	1	0	<input checked="" type="checkbox"/>
CAM2-POS-6N	Sample	CAM2	1	29	1	0	<input checked="" type="checkbox"/>
CAM2-POS-7N	Sample	CAM2	1	30	1	0	<input checked="" type="checkbox"/>
CAM2-POS-8N	Sample	CAM2	1	31	1	0	<input checked="" type="checkbox"/>
CAM2-POS-9N	Sample	CAM2	1	32	1	0	<input checked="" type="checkbox"/>
QC-ALL-POS-1N	QC	QC	1	33	1	0	<input checked="" type="checkbox"/>
QC-ALL-POS-2N	QC	QC	1	34	1	0	<input checked="" type="checkbox"/>
QC-ALL-POS-3N	QC	QC	1	35	1	0	<input checked="" type="checkbox"/>
QC-ALL-POS-4N	QC	QC	1	36	1	0	<input checked="" type="checkbox"/>
QC-ALL-POS-5N	QC	QC	1	37	1	0	<input checked="" type="checkbox"/>
QC-ALL-POS-6N	QC	QC	1	38	1	0	<input checked="" type="checkbox"/>
TAK1-NEG-1	Sample	TAK1	1	39	1	0	<input checked="" type="checkbox"/>
TAK1-POS-2-N	Sample	TAK1	1	40	1	0	<input checked="" type="checkbox"/>
TAK1-POS-3N	Sample	TAK1	1	41	1	0	<input checked="" type="checkbox"/>
TAK1-POS-4N	Sample	TAK1	1	42	1	0	<input checked="" type="checkbox"/>
TAK1-POS-5N	Sample	TAK1	1	43	1	0	<input checked="" type="checkbox"/>
TAK1-POS-6N	Sample	TAK1	1	44	1	0	<input checked="" type="checkbox"/>

Export peak list

After alignment process:

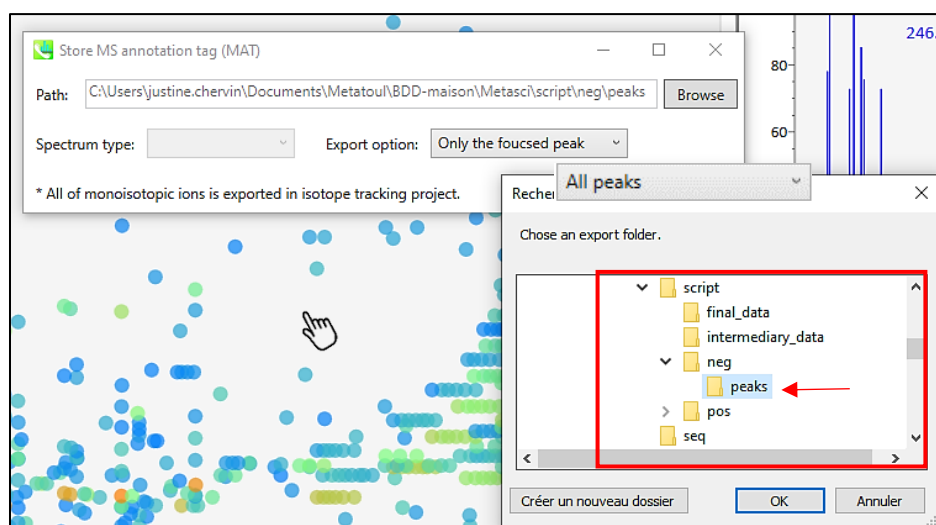
- **Normalized data** by Total ion chromatogram (TIC) or another normalization method

- Export alignment results: both **Raw data matrix (Height)** and **Normalized data matrix** respectively in previously created folders named “pos” and “neg”.



Export all peaks

By clicking on one feature dot, export « **all peaks** » to the “peaks” directory respectively created in “pos” and “neg” folders.



Open the shiny interface of MS-CleanR



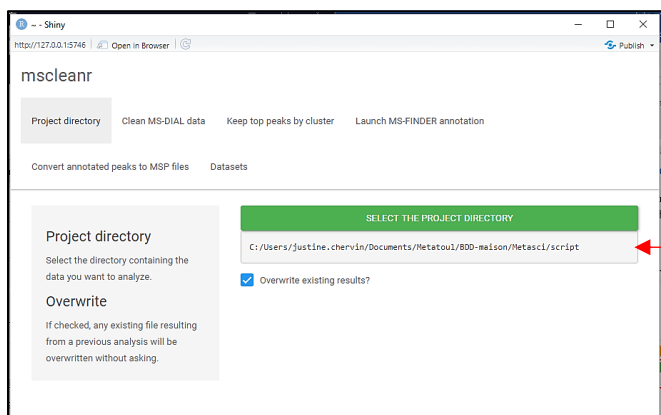
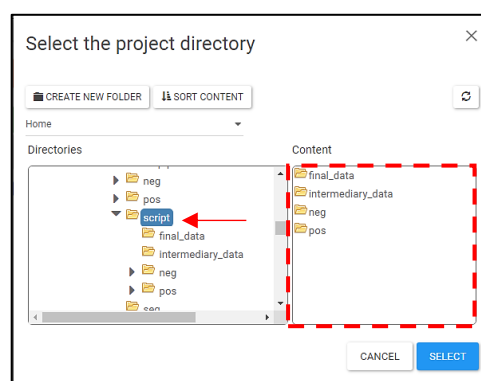
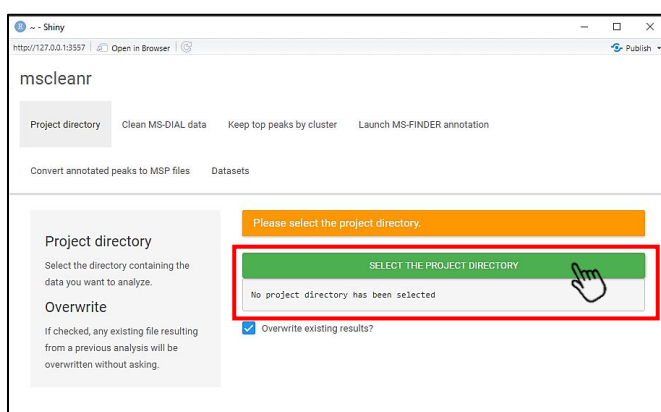
Select the MScleanR package in **Rstudio** and open the shiny interface using the following command:

- Note that if you encounter some issues, try to open the Shiny interface in internet browser.
- Sometimes Windows block file writing, close the shiny or R studio and run it again to solve the problem.

```
runGUI()
```

Select the project directory

First step is to define the project directory on the first tab called “**Project Directory**” by clicking on the green rectangle “Select the project directory” and by selecting the parent folder containing “pos” and “neg” folders.



When your project directory is selected, it is written in the grey rectangle.

Define your parameter of filtration and Clean your data

In the second tab called “**Clean MS-DIAL data**” various parameters can be personalized to filter your data. You can decide to select any filter according to your goal and experimental design.

Command	Description
Blank ratio	Subtract blank peaks to samples based on the indicated “ Minimum blank ratio ” by default at 0.8. This operation is done on the Height files between Blanks and QCs.

Incorrect Mass	Delete all peaks with a mass defect in X.8 and X.9 which appear to be artifacts.
Relative standard Deviation (RSD)	Filter based on the Maximum RSD value set at 30 by default. The RSD is calculated on each defined class. If RSD of one feature is under the defined value for all class, it is removed from the peak list.
Relative Mass Defect (RMD)¹	RMD is calculated in ppm as ((mass defect/measured monoisotopic mass) × 10e6) Analysis of natural products from the DNP shows that 95 % of RMD are comprised between 50 and 3000 (values by default).
Delete ghost peaks	Delete variables with <i>m/z</i> values corresponding to blank peaks but with a different RT in samples.
Maximum mass difference	<i>m/z</i> value tolerance set by default to 0.005 for Pearson correlation and pos/neg merging
Maximum retention time difference	RT value tolerance set by default to 0.025 (absolute value) for Pearson and pos/neg merging
Use Pearson correlation to compute clusters?	Extend MS-DIAL clusters with Pearson correlation. Minimum correlation and maximum p-value are respectively set by default to 0.8 and 0.05

Once your parameters are fixed, click on the green rectangle named “*Clean MS-DIAL data*”. A green window appears with the writing “*Cleaning data...*”.

The screenshot shows the mscleanr web application interface. The 'Clean MS-DIAL data' tab is active. On the left, there are sections for 'Filters' (with checkboxes for Blank ratio, Incorrect Mass, Relative Standard Deviation, and Relative Mass Defect), 'Deltas' (retention time and mass differences), 'Clusterisation options' (Pearson correlation), and '(Optional) Reference files'. The main area contains input fields for: Minimum blank ratio (0.8), Maximum RSD (30), Minimum RMD (50), Maximum RMD (3000), Maximum mass difference (0.005), Maximum retention time difference (0.025), Use Pearson correlation to compute clusters? (checked), Minimum correlation (0.8), and Maximum p-value (0.05). At the bottom, there is a green button labeled 'CLEAN MS-DIAL DATA' with a hand cursor icon pointing to it.

Cleaning data...

During the cleaning:

- Clusters are formed based on MS-DIAL “post curation column”, Pearson correlation, links such as adducts, neutral losses, dimers, ...;
- Adducts are corrected based on previous found links;
- Pos and Neg clusters are concatenated if relational links are found (adducts mass difference)

¹ Ekanayaka EA, Celiz MD, Jones AD. Relative mass defect filtering of mass spectra: a path to discovery of plant specialized metabolites. *Plant Physiol.* 2015;167(4):1221–1232. doi:10.1104/pp.114.251165

- Once the cleaning is done, one new folder is created named “intermediary_data”. Different information is obtained at the bottom of the index “Clean MS-DIAL data”.

Delete previous results if necessary

Number of final peaks

Number of MS-DIAL links

Number of MS-DIAL identification

Adduct / neutral loss relations

Adduct correction if necessary

CLEAN MS-DIAL DATA

```

/!\ Deleting C:/Users/justine.chervin/Documents/Metatoul/BDD-maison/Metasci/script/final_data
/!\ Deleting C:/Users/justine.chervin/Documents/Metatoul/BDD-maison/Metasci/script/intermediary_data
*** Treating C:/Users/justine.chervin/Documents/Metatoul/BDD-maison/Metasci/script ***
MSDial peaks after filtering: 621 positive, 266 negative, 0 NA, 887 total
MSDial links: 1978
73 peaks identified by MSDial
Correlation links found in pos : 1939
Correlation links found in neg : 323
Clusters detected with MSDial data: 180
Using package neutral losses for positive mode
Adducts/Neutral losses detection (cluster pos 91 )
Adducts/Neutral losses detection (cluster pos 92 )
Adducts/Neutral losses detection (cluster pos 93 )
Adducts/Neutral losses detection (cluster pos 94 )
Adducts/Neutral losses detection (cluster pos 95 )
Adducts/Neutral losses detection (cluster pos 96 )
Adducts/Neutral losses detection (cluster pos 97 )
Adducts/Neutral losses detection (cluster pos 98 )
Adducts/Neutral losses detection (cluster pos 99 )
Adducts/Neutral losses detection (cluster pos 100 )
Adducts/Neutral losses detection (cluster pos 101 )

```

At this step, several files are created in the folder “intermediary_data”.

Files	Description
Adducts_massdiff_filtered	Reference file for mass difference between regular adducts
Adducts_massdiff_total	Reference file for mass difference between all possible adducts
Adducts_detected_by_MSdIAL	Reference file for adduct ponderation of regular adducts found by MS-DIAL
Adducts_filtered.graphml	A graph to display feature clusters based on adducts links
Adducts_final_selection	Final adducts resulting from MSdial and modified after pos/neg concatenation
Adducts_initial.graphml	A graph to display feature clusters based on MSdial data
Annotated_MS-peaks-MSDial	List of annotated peaks based on the database (msp file) imported in MS-DIAL
Deleted_blank_ghosts	List of peaks deleted with “delete ghost peaks”
Deleted_blanks	List of peaks deleted with the filter “blank ratio”
Deleted_mz	List of peaks deleted with the filter “incorrect mass”
Deleted_rmd	List of peaks deleted with the filter “RMD”
Deleted_rsd	List of peaks deleted with the filter “RSD”
Links_clusters_final	List of correlation (adduct, neutral loss, msdial) between peaks in neg and pos
Links_post_selection	Feature links after adduct prioritization process
Links_pre_selection	Feature links with all adducts possibilities
MS_peaks-clusters.graphml	A graph of final clusters (MS-DIAL + Pearson)
MS_peaks-clusters_final	List of final clusters (MS-DIAL + Pearson) in both pos and neg ionization
MS_peaks-clusters_msdialog	List of MS-DIAL clusters in both pos and neg ionization
parameters	List of parameter used for the cleaning
samples	List of samples with indication of sample name, class, file type, script class and column name

Select number of retained peaks per cluster

In the third tab “**Keep top peaks by cluster**” you can select the number of features you want to keep in each cluster.

This step is based on the hypothesis that in one cluster, only one unique metabolite is present. The other variables used to come from feature degeneration. Generally, this metabolite appears to be the **most intense** and/or **the most connected within the graph** (adducts, neutral loss, dimers...).

You can then choose to select as many peaks as you want and either the most intense(s) by clicking “**Intensity**”, the most connected by clicking “**Degree**” or both.

We advise to select both criteria and keep 2 top peaks by cluster for further MS-finder request.

Shiny

http://127.0.0.1:5746 | Open in Browser | Publish

mscleanr | Project directory | Clean MS-DIAL data | **Keep top peaks by cluster** | Launch MS-FINDER annotation | Convert annotated peaks to MSP files | Datasets

Keeps only the top peaks by cluster and by method.

Selection mode

Peaks selected can be the most intense (intensity), most connected (degree), or both. If there are ties, the number of peaks selected can be greater than the number requested by the user.

Exporting filtered peaks

Copy MAT files corresponding to the selected peaks in a new folder for an faster analysis in MS-FINDER.

Selection criteria

☒ Intensity (most intense peaks)

☐ Degree (most connected peaks)

Number of peaks to keep (by cluster and by method)

1

☒ Export final peaks in a new folder?

KEEP TOP PEAKS BY CLUSTER

Keeping only selected peaks...

At this step, a new folder is created in both “pos” and “neg” folders named “**filtered peaks**”. All .MAT files corresponding to kept peaks are copied from “peaks” folder and pasted in this new folder “filtered peaks”.

msccleanr

Project directory Clean MS-DIAL data **Keep top peaks by cluster** Launch MS-FINDER annotation Convert annotated peaks to MSP files

Datasets

Keeps only the top peaks by cluster and by method.

Selection mode

Peaks selected can be the most intense (intensity), most connected (degree), or both. If there are ties, the number of peaks selected can be greater than the number requested by the user.

Exporting filtered peaks

Copy MAT files corresponding to the selected peaks in a new folder for a faster analysis in MS-FINDER.

Selection criteria

☒ Intensity (most intense peaks)

☒ Degree (most connected peaks)

Number of peaks to keep (by cluster and by method)

2

☒ Export final peaks in a new folder?

KEEP TOP PEAKS BY CLUSTER

```

/!\ Deleting C:/Users/justine.chervin/Documents/Metatoul/Global-Marchantia.p-20-01-26
/!\ Deleting C:/Users/justine.chervin/Documents/Metatoul/Global-Marchantia.p-20-01-26
Filtering on both ( 2 peaks by cluster and by method)
MSDial peaks after peaks filtering: 186 positive, 115 negative, 0 NA, 301 total
Adduct modification in mat file for peak pos 120
Adduct modification in mat file for peak pos 214
Adduct modification in mat file for peak pos 266
Adduct modification in mat file for peak pos 322
Adduct modification in mat file for peak pos 268
Adduct modification in mat file for peak pos 328
Adduct modification in mat file for peak pos 434
Adduct modification in mat file for peak pos 441
Adduct modification in mat file for peak pos 444
Adduct modification in mat file for peak pos 456
Adduct modification in mat file for peak pos 465
Adduct modification in mat file for peak pos 466

```

Number of kept peaks

Modification of adduct annotation directly in .MAT file for further MS-FINDER annotation



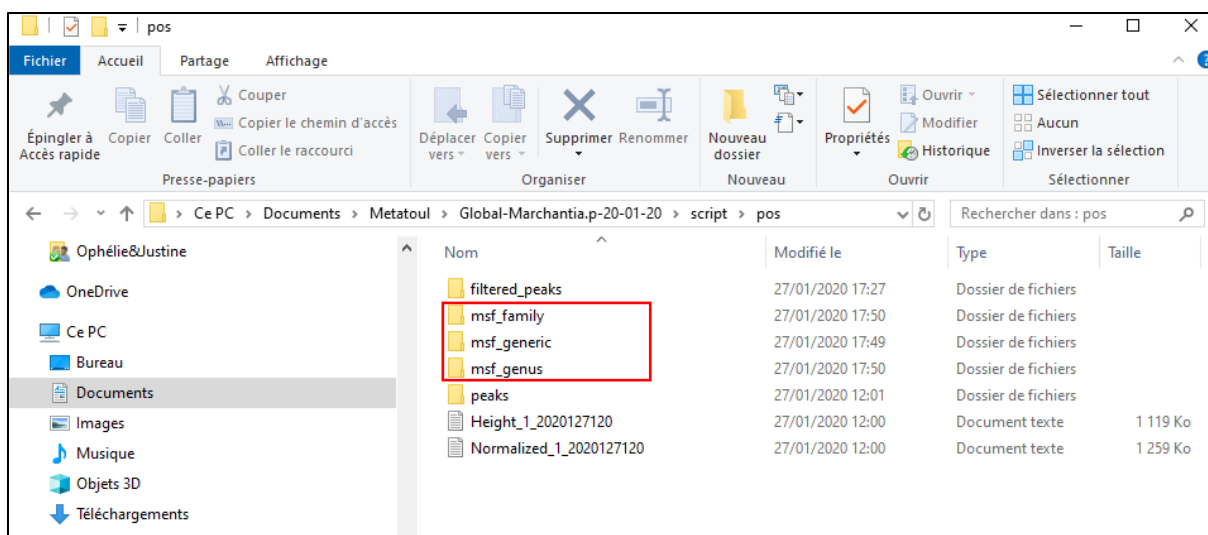
Interrogation of MS-FINDER

From « **filtered peaks** » folder, interrogate MS-FINDER based on several databases of your choice (for example plant genus, plant family, generic databases from MS-FINDER, ...).

Optional: Add a “Compound_level” column within your in-house database for MS-FINDER. This level will be used for annotation ranking in the next step.

The most **important thing** to do is to create respectively in “pos” and “neg” directories, new folders named “**msf_X**” (for example msf_genus) which correspond to the name of each database used for feature annotation. The msf_generic is mandatory and correspond to internal database in MS-Finder.

For each database used, export “structure” and “formula” as a single file in the corresponding folder.



Launch MS-FINDER annotation

Once all your MS-FINDER interrogations are done and your folder “msf_X” filled with “**structure**” and “**formula**” files, go to the fourth tab called “**Launch MS-FINDER annotation**”.

This step will merge feature annotation to the dataset based either only on the score of MS-FINDER or on the prioritization of the different databases, used to indicate the more pertinent annotation.

☐ Select the best annotation for each peak based only on MSFINDER scores?

Indicate the compound levels in your annotation files, separated by commas (leave blank if none).

Rank	Compound.level
1	1a
2	1b

Indicate the biosource levels in your annotation process, separated by commas.

Rank	Biosource.level
1	genus
2	family
3	generic

Indicate the scores multipliers associated to your compound or biosource levels, separated by commas (leave blank if none).

Level	Multiplier
genus	2.00
family	1.50
generic	1.00

LAUNCH MS-FINDER ANNOTATION

This option is used to report the identification with the best MS-FINDER score

This option is used when you want to prioritize some databases.

In (A) you have to indicate the compound level within your database

In (B) you have to order your database

In (C) you can dedicate to your database levels a multiplier to calculate new scores from MS-FINDER ones.

Annotating peaks with MS-FINDER data...

Project directory Clean MS-DIAL data Keep top peaks by cluster **Launch MS-FINDER annotation** Convert annotated peaks to MSP files Datasets

Annotates peaks based on files extracted from MSFinder.

☐ Select the best annotation for each peak based only on MSFINDER scores?

Indicate the compound levels in your annotation files, separated by commas (leave blank if none).

1a,1b

Indicate the biosource levels in your annotation process, separated by commas.

genus,family,generic

Indicate the scores multipliers associated to your compound or biosource levels, separated by commas (leave blank if none).

1a:2,b:1.5,genus:2,family:1.5,generic:1

Rank	Compound.level
1	1a
2	1b

Rank	Biosource.level
1	genus
2	family
3	generic

Level	Multiplier
1a	2.00
b	1.50
genus	2.00
family	1.50
generic	1.00

LAUNCH MS-FINDER ANNOTATION

```

/!\ Level b present in scores but not in biosource or compound levels.
/!\ Deleting C:/Users/justine.chervin/Documents/Metatou1/Global-Marchantia.p-20-01-20/script/fina
*** Treating C:/Users/justine.chervin/Documents/Metatou1/Global-Marchantia.p-20-01-20/script ***
Annotating with 2 compound levels ( 1a, 1b ) and 3 biosource levels ( genus, family, generic ).
*** Annotating clusters with [M+H]+ / [M-H]- couples ***
Annotating cluster 41
Annotating cluster 110
Annotating cluster 124
Annotating cluster 146

```

Summary of compounds and biosource levels used

Paste of annotation in the final peak list

Two files are created in the “final-data” folder:

- **Annotated MS peaks cleaned** = the final peak list with annotation from MS-FINDER
- **Annotated MS peaks normalized** = the final peak list renormalized based on total peak area

The final peak list looks like as follow. Different information are available such as:

- The average m/z value;
- The average RT value;
- The annotation based on MS-FINDER interrogation on the “**Structure**” column with the associated **Total score** of MS-FINDER and **Final score** calculated from the indicated multipliers.
- The source of the annotation in the “**level**” column;
- The ontology of the compound; ...

The variable are also identified as:

- Unknown compound = variable with no annotation
- Simple ID = based on a single feature in pos or neg mode
- Double ID =based on same annotation retrieve in pos and neg mode

