

*MS-CleanR tutorial:
Peak list cleaning, data concatenation and peak annotation
18/04/2020*

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**.

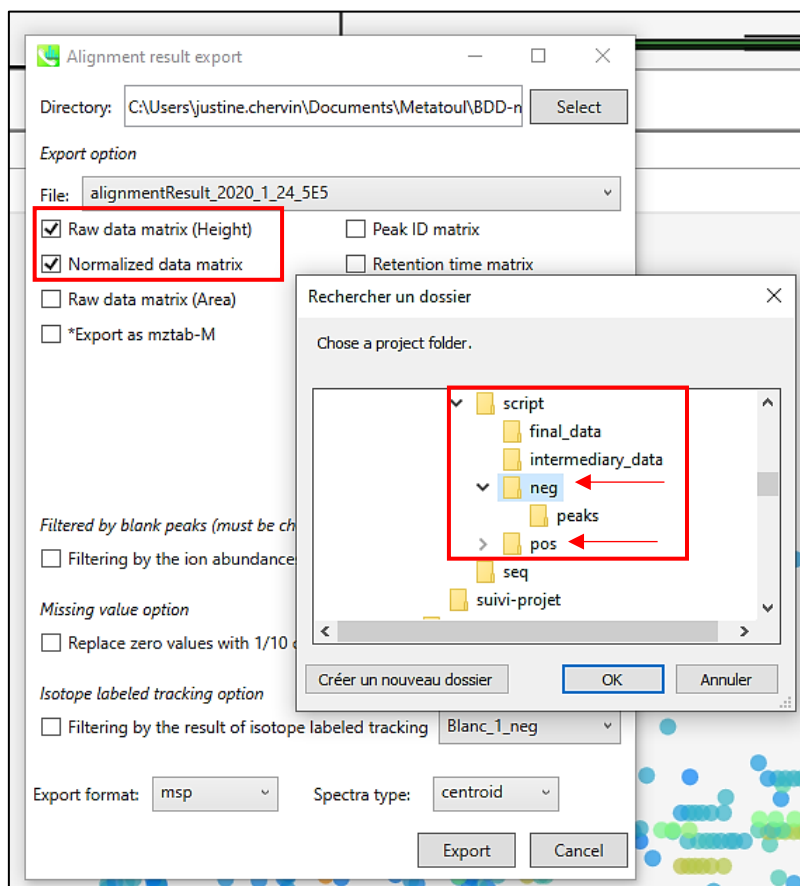
File property setting								
File name	File type	Class ID	Batch	Analytical order	Injection volume (μL)	Y variable	Included	
BLANC-M-POSN	Blank	blank	1	7	1	0	<input checked="" type="checkbox"/>	
BLANC-P-POSN	Blank	blank	1	8	1	0	<input checked="" type="checkbox"/>	
BLANC-Q-POSN	Blank	blank	1	9	1	0	<input 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:

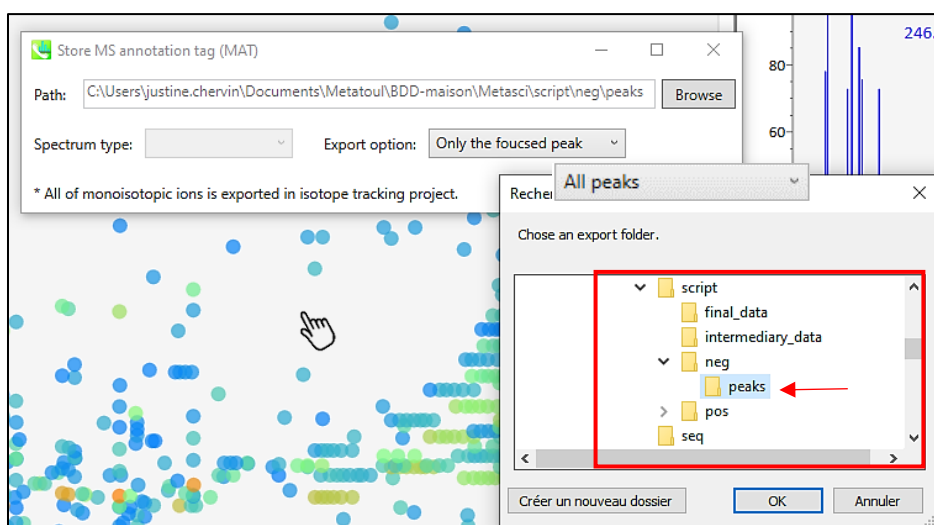
- **Normalized data** by Total ion chromatogram (TIC) or another normalization method

- 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



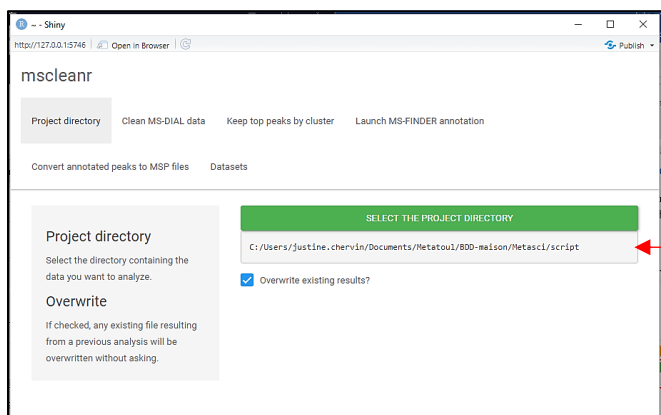
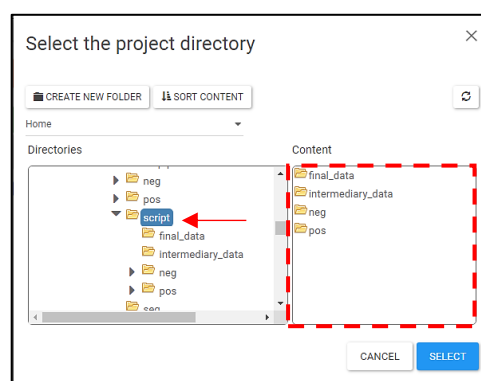
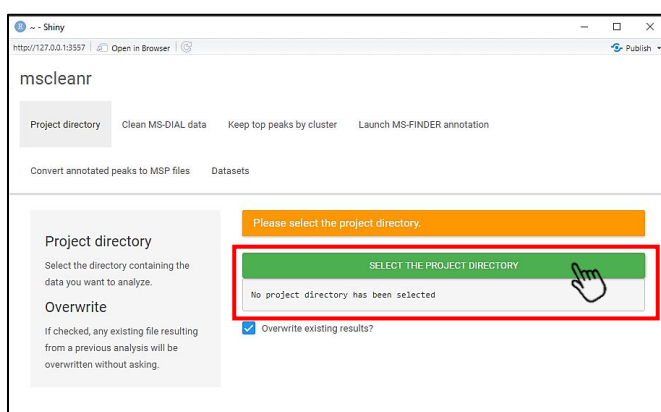
Select the MScleanR package in **Rstudio** and open the shiny interface using the following command:

- 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...*”.

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”.

Annotations on the left side of the screenshot:

- 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

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_MS-DIAL	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

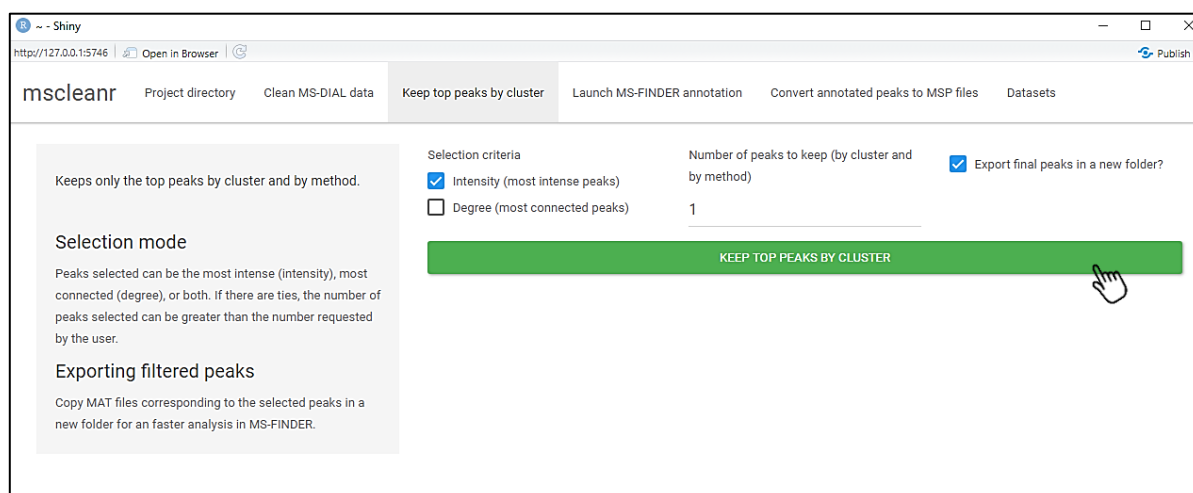
Select number of retained peaks per cluster

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.



The screenshot shows the mscleanr Shiny application interface. The browser address bar indicates the URL is http://127.0.0.1:5746. The application has several tabs: Project directory, Clean MS-DIAL data, Keep top peaks by cluster (active), Launch MS-FINDER annotation, Convert annotated peaks to MSP files, and Datasets. The 'Keep top peaks by cluster' tab contains the following elements:

- 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)**
A text input field containing the number '1'.
- ☒ Export final peaks in a new folder?
- A large green button labeled **KEEP TOP PEAKS BY CLUSTER** with a hand cursor icon pointing to it.

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”.

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)

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



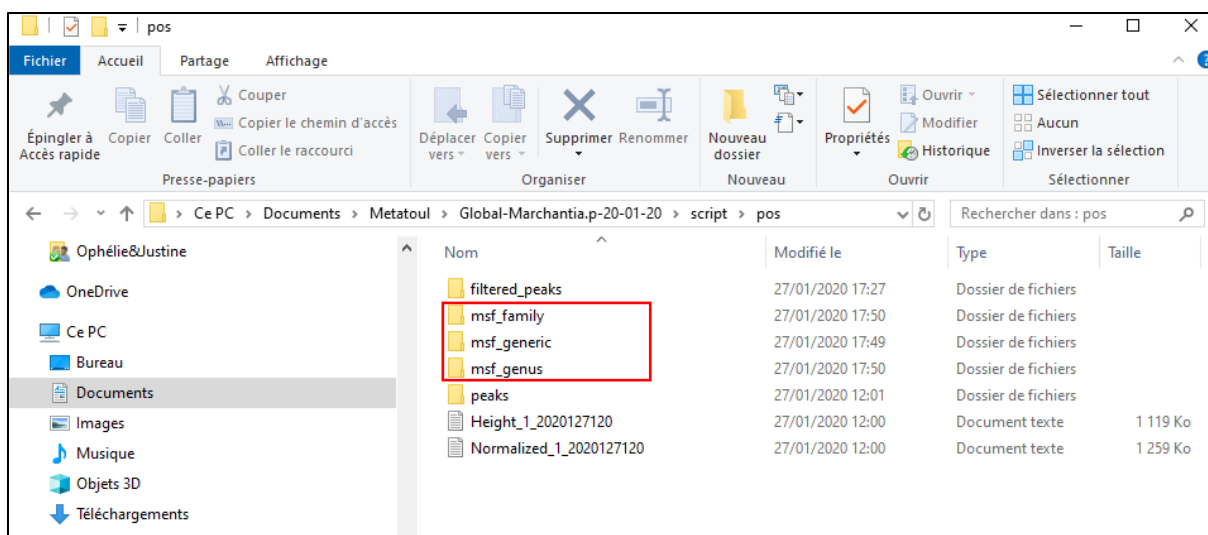
Interrogation of MS-FINDER

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?

Compound levels
The list of compound levels to consider, in the given order (from more important to least important).

Biosource levels
The list of biosource levels to consider, in the given order (from more important to least important). They must correspond to the folders containing MS-FINDER files in your project directory. The level 'generic' is always added as the last biosource level considered.

Levels scores
A list of levels names and their corresponding multiplier to adapt final annotation scores.

(A) Compound levels

Rank	Compound.level
1	1a
2	1b

(B) Biosource levels

Rank	Biosource.level
1	genus
2	family
3	generic

(C) Levels scores

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/Metatou/Global-Marchantia.p-20-01-20/script/fina
*** Treating C:/Users/justine.chervin/Documents/Metatou/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

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V		
1	annotation_result	annotation_warmir	source	Alignment	IC	Average	Mz	Adduct	level	Formula	Structure	Total score	Final score	Title	MS1 count	MS/MS count	PRECURSOR	TYPE	Theoretical	mass	error	Formula	score	Ontology
2	Unknown compound	neg	108	1.346	316.7895	[M-H] ⁻																		
3	Unknown compound	neg	173	6.558	427.03708	[M-H] ⁻																		
4	Unknown compound	neg	56	1.303	242.05179	[M-H] ⁻																		
5	Unknown compound	pos	10	3.463	111.00761	[M+H] ⁺																		
6	Unknown compound	pos	105	7.532	179.04666	[M+H] ⁺																		
7	Unknown compound	pos	424	6.896	401.07579	[M+H] ⁺																		
8	Unknown compound	pos	159	21.543	206.61177	[M+H] ⁺																		
9	Unknown compound	pos	202	1.402	236.14941	[M+H] ⁺																		
25	Simple ID	neg	11	2.818	128.03514	[M-H] ⁻				genus	25H7N03	351-5-carboxy-4,5-dihydro	7.0089	28.0356	(IRQA) PYROGLUTAM	3	13	128.0351	[M-H] ⁻	129.0425931	0.0002166	4.495	Alpha amino acids and derivatives	
26	Simple ID	neg	111	2.014	323.02853	[M-H] ⁻				genus	26H13N2O9P	5-(10S,3R,4S,5R)-3,4-dihydro	7.3562	28.4248	(IRQA) URIDINE MO	3	37	323.0285	[M-H] ⁻	324.058866	9.02E-05	4.314	Pentose phosphates	
27	Simple ID	neg	128	7.555	345.11874	[M-H] ⁻				genus	27H22O9	3,4,5-Trimethoxyphenyl g	5.7246	5.7246	Unknown	3	26	345.1187	[M-H] ⁻	346.1263823	0.0004058	2.949	Phenolic glycosides	
28	Simple ID	neg	13	3.904	129.01901	[M-H] ⁻				genus	28H6O4	itaconic acid	6.4729	6.4729	(IRQA) ITACONATE	3	5	129.019	[M-H] ⁻	130.0266087	0.0003322	3.471	Branched fatty acids	
29	Simple ID	neg	132	1.286	353.06966	[M-H] ⁻				genus	29H18O9	6-Hydroxy-2-methyl-7-[(2Z	5.8621	11.7242	Unknown	3	72	353.0697	[M-H] ⁻	354.0950822	-0.001894	3.344	Phenolic glycosides	
30	Simple ID	neg	140	21.553	367.0425	[M-H] ⁻				genus	30H23N2O13	17-O-Acetylserine	3.9909	3.9909	Unknown	3	12	367.0425	[M-H] ⁻	368.099928	-0.001584	2.205	Alkaline-seropine alkaloids	
31	Simple ID	neg	148	6.837	380.15628	[M-H] ⁻				genus	31H23N5O6	2R,3S,4S,5R,6S)-2-Hydrox	6.5026	26.0104	Unknown	3	62	380.1563	[M-H] ⁻	381.1648335	0.001257	3.908	Fatty acyl glycosides of mono- and disaccharides	
32	Simple ID	neg	155	13.242	392.20822	[M-H] ⁻				genus	32H31N6	Pulchellamine G(+)-Pulch	6.5573	6.5573	Unknown	3	60	392.2082	[M-H] ⁻	393.2151377	-0.000339	3.752	Guaianolides and derivatives	
33	Simple ID	neg	164	11.837	408.20385	[M-H] ⁻				genus	33H31N7O7	Isorutin C(+)-Isorutin C	5.7896	5.7896	Unknown	3	86	408.2039	[M-H] ⁻	409.2100323	-0.00124	3.514	Oxepenes	
34	Simple ID	neg	171	18.496	423.16013	[M-H] ⁻				genus	34H23N2O8	2-(11S,3R,4R,9aS)-1-Hydr	5.3463	5.3463	Unknown	3	77	423.1601	[M-H] ⁻	424.166076	-0.000569	3.345	Sulfonamides	
35	Simple ID	neg	174	11.272	435.1293	[M-H] ⁻				genus	35H24O10	4-(3-(2,4-dihydroxy-6-[(2S	6.1769	24.7076	Unknown	3	72	435.1293	[M-H] ⁻	436.136947	0.0003705	3.631	Flavonoid O-glycosides	
36	Simple ID	neg	175	17.228	437.13907	[M-H] ⁻				genus	36H22O5	Psychotriol A, 6'-Hydroxy	6.7501	13.5002	Unknown	3	78	437.1391	[M-H] ⁻	438.1467238	0.0003474	3.688	Diarylethers	
37	Simple ID	neg	176	18.368	437.17554	[M-H] ⁻				genus	37H24O4	Marchantin C, 3'-Me ethe	6.0043	24.0172	Unknown	3	81	437.1755	[M-H] ⁻	438.1833093	0.0003329	3.277	Lignans, neolignans and related compounds	
38	Simple ID	neg	18	1.722	133.01407	[M-H] ⁻				genus	38H6O5	L-malate	7.4137	29.6548	(IRQA) MALATE	3	10	133.0141	[M-H] ⁻	134.0215233	0.0001468	4.245	Beta hydroxy acids and derivatives	
39	Simple ID	neg	163	8.576	407.13458	[M-H] ⁻				genus	39H24O9	2S,3R,4R,5S,6R)-2-(2-(2-3	6.663	13.326	Unknown	3	64	407.1346	[M-H] ⁻	408.1420232	0.0001559	3.776	Stilbene glycosides	
40	Simple ID	neg	189	12.962	449.10275	[M-H] ⁻				genus	40H18O6	2,2,3,3,7,7'-Hexahydroxy-	5.9568	21.9872	Unknown	3	30	449.1028	[M-H] ⁻	450.1103383	0.0002618	3.418	Phenanthrois	
41	Simple ID	neg	19	1.364	135.0298	[M-H] ⁻				genus	41H8O5	Erythronic acid	7.0979	7.0979	Unknown	3	27	135.0298	[M-H] ⁻	136.0371794	9.69E-05	3.559	Sugar acids and derivatives	
42	Simple ID	neg	190	12.714	449.10278	[M-H] ⁻				genus	42H18O6	2,2,3,3,7,7'-Hexahydroxy-	5.0015	20.006	Unknown	3	45	449.1028	[M-H] ⁻	450.1103383	0.0002618	3.261	Phenanthrois	
43	Simple ID	neg	191	17.84	452.27817	[M-H] ⁻				genus	43H44N7P	LysopEIO(0.16.0)	6.2888	6.2888	Unknown	3	30	452.2782	[M-H] ⁻	453.2855394	6.29E-05	3.643	2-acyl-sn-glycero-3-phosphoethanolamines	
44	Double ID	pos	465	10.368	483.08771	[M+H] ⁺				genus	44H31O12	2R,3R,4S,5S,6S)-6-[(2-(3,4-	7.4323	29.7292	Unknown	3	54	483.0877	[M+H] ⁺	482.079626	-0.000598	4.628	Flavonoid-7-O-glucuronides	
45	Double ID	pos	466	10.005	485.08774	[M+H] ⁺				genus	45H31O12	2R,3R,4S,5S,6S)-6-[(2-(3,4-	7.5668	30.2672	Unknown	3	39	483.0877	[M+H] ⁺	482.079626	-0.000598	4.61	Flavonoid-7-O-glucuronides	
46	Double ID	pos	468	17.402	471.18042	[M+H] ⁺				genus	46H30N2O8S	Dacarpamine	5.4081	5.4081	Unknown	3	237	471.1804	[M+H] ⁺	470.1722869	-0.000837	3.534	Methionine and derivatives	
47	Double ID	neg	207	10.179	473.20294	[M-H] ⁻				genus	47H34N21S	UNPD79762	5.6695	5.6695	Unknown	3	61	473.2029	[M-H] ⁻	474.2101119	-6.45E-05	3.305	Terpene glycosides	
48	Double ID	pos	517	9.931	639.11963	[M+H] ⁺				genus	48H36O18	2R,3R,4S,5R,6S)-6-[(2-(3-4-	7.244	28.676	Unknown	3	47	639.1196	[M+H] ⁺	638.1133014	-0.00041	4.641	Flavonoid-7-O-glucuronides	
49	Double ID	pos	181	17.298	439.15469	[M+H] ⁺				genus	49H24O5	Marchantin C, 12-Hydroxy	6.3413	25.3652	Unknown	3	32	439.1547	[M+H] ⁺	440.1623739	0.0003974	3.493	Lignans, neolignans and related compounds	
50	Double ID	pos	436	17.098	425.17465	[M+H] ⁺				genus	50H28N2O6S	2-(11S,3R,4R,9aS)-1-Hydr	5.8585	5.8585	Unknown	3	115	425.1747	[M+H] ⁺	424.1668076	-0.000616	3.618	Sulfonamides	
51	Double ID	pos	450	17.064	441.16974	[M+H] ⁺				genus	51H24O5	Marchantin C, 2-Hydroxy	6.2661	24.9444	Unknown	3	194	441.1697	[M+H] ⁺	440.1623739	-4.97E-05	3.5	Lignans, neolignans and related compounds	
52	Double ID	pos	45	7.123	203.08241	[M-H] ⁻				genus	52H12N2O2	L-Tryptophan	8.0971	8.0971	(IRQA) TRYPTOPHAN	3	41	203.0824	[M-H] ⁻	204.0898776	0.0002012	4.082	Indolyl carboxylic acids and derivatives	

Export peaks as .msp files

In the fifth tab “**Convert annotated peaks to MSP files**”, you will be able to create two .msp files named “peaks-neg.msp” and “peaks-pos.msp” in the folder “final_data”. All peaks can be converted, or user can choose a scoring threshold based on multiplied MSfinder score. One metadata file per ionization mode is also created containing annotation results and average peak area of each class.

Project directory

Clean MS-DIAL data

Keep top peaks by cluster

Launch MS-FINDER annotation

Convert annotated peaks to MSP files

Datasets

Convert the final CSV file post annotations to MSP format.

Minimum score

Minimum annotation score needed to export peaks to the MSP files.

☒ Export all peaks to MSP files?

CONVERT PEAKS TO MSP FILES

188 peaks to convert in MSP.

Peaks converted, see MSP files in C:/Users/justine.chervin/Documents/Metatoul/Global-Marchantia.p-2

These two files could then be imported in MetGem software or GNPS facility to create mass spectral similarity networks.

11