

*MS-CleanR tutorial:
Peak list cleaning, data concatenation and peak annotation
25/05/2020
Justine Chervin and Guillaume Marti*

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

Table des matières

| | |
|--|----|
| Requirements: | 1 |
| 1.1. Downloading | 1 |
| 1.2. Installation | 2 |
| MS-CleanR workflow | 3 |
| 1. Process the data with MS-DIAL | 3 |
| 1.3. Export peak list | 3 |
| 1.4. Export all peaks | 4 |
| 2. Open the shiny interface of MS-CleanR | 4 |
| 2.1. Select the project directory | 5 |
| 2.2. Define your parameter of filtration and Clean your data | 5 |
| 2.3. Select number of retained peaks per cluster | 9 |
| 2.4. Interrogation of MS-FINDER | 10 |
| 2.5. Launch MS-FINDER annotation | 11 |
| 2.6. Export peaks as .msp files | 13 |

Requirements:

Software installation

1.1. 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/>

1.2. 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”.

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

1. Process the data with MS-DIAL



Process data with MS-DIAL in either pos or neg mode or both according to the tutorial found within MS-CleanR GitHub page or more detailed tutorial found here: <https://mtbinfo-team.github.io/mtbinfo.github.io/>

Importantes notices :

- 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)
- If both ionization mode have been used, 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"/> |

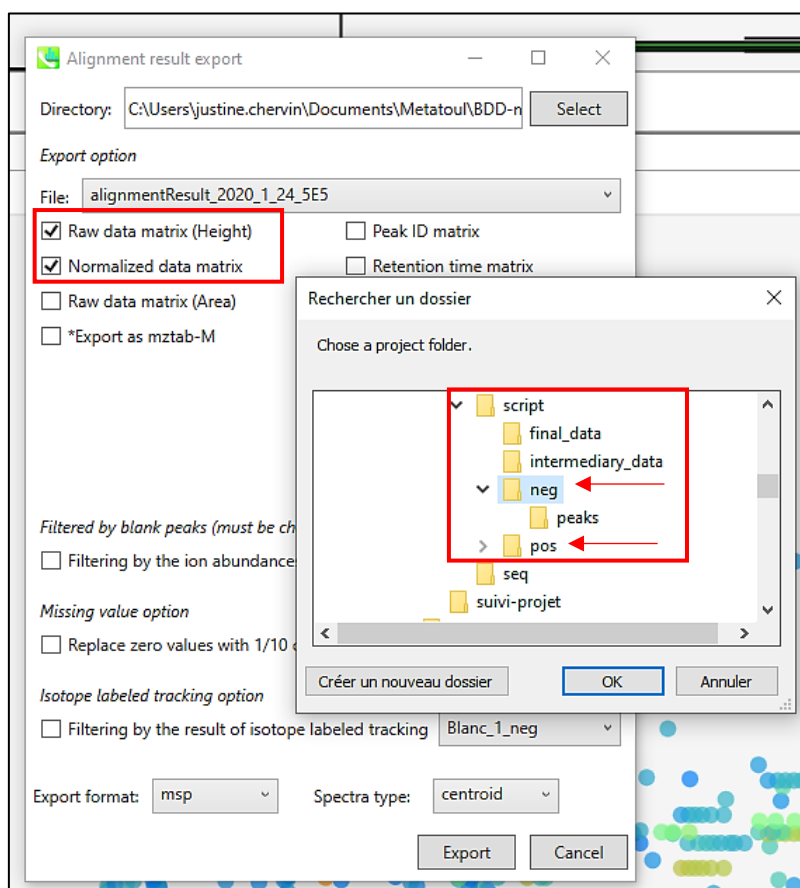
Finish Cancel

1.3. Export peak list

After alignment process:

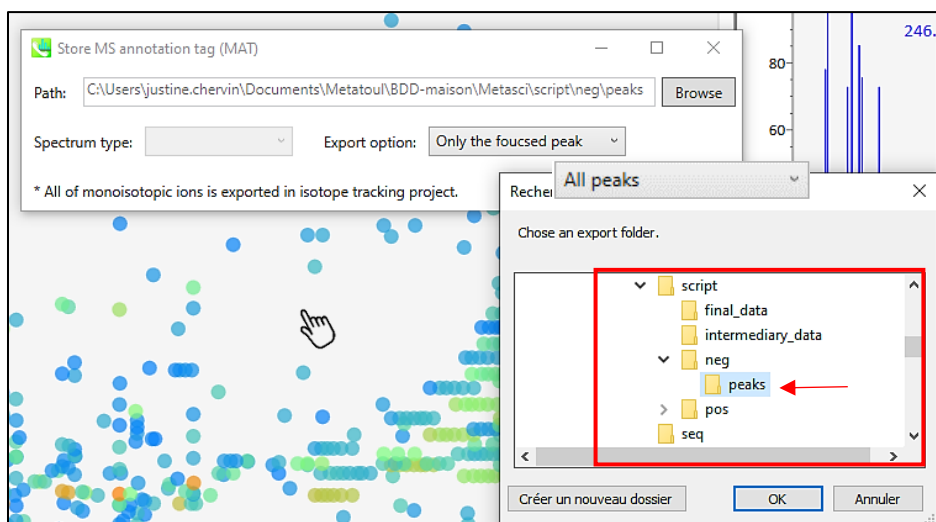
- Normalized data by Total ion chromatogram (TIC) or another normalization method

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



1.4. Export all peaks

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



2. Open the shiny interface of MS-CleanR

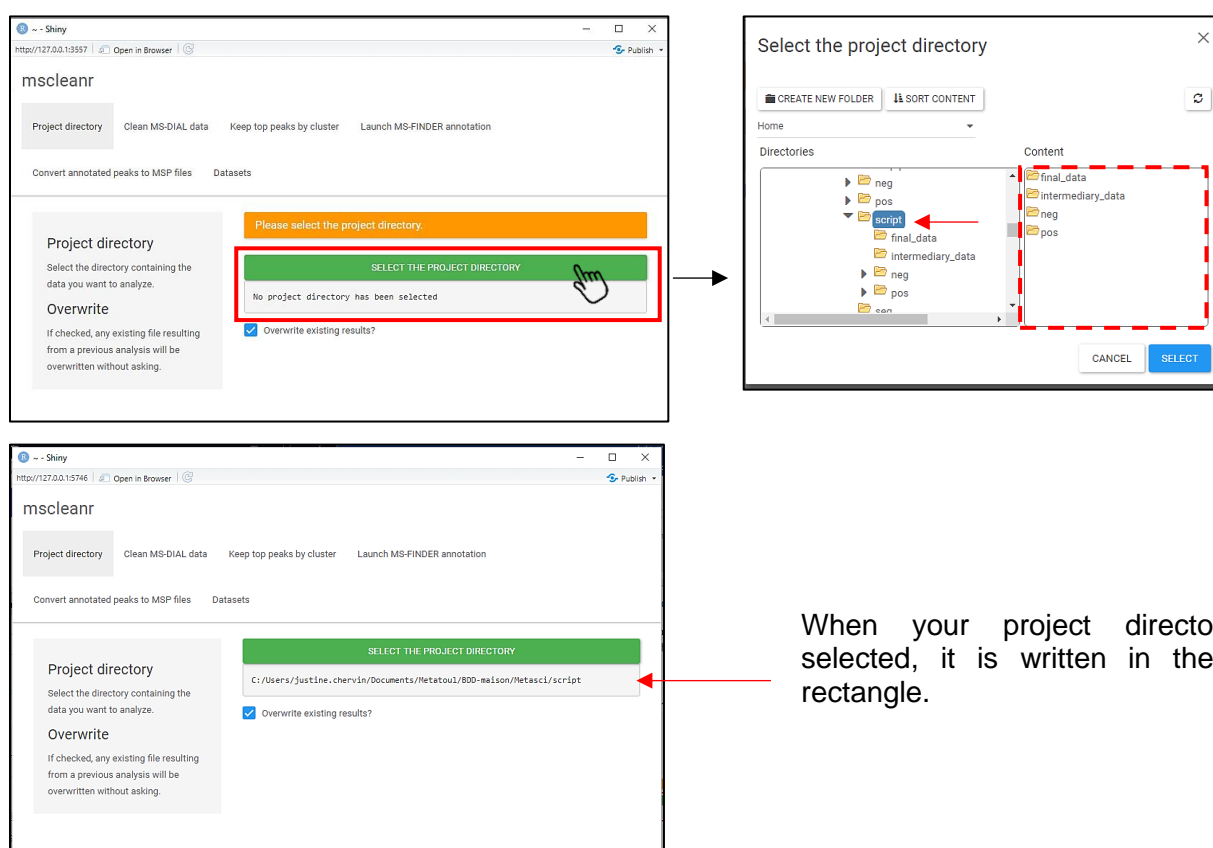
Select the MSCleanR package in **Rstudio** and open the shiny interface using the following command:

- 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()

2.1. 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.

2.2. 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. Tips: Carefully inspect your raw data between blanks and QCs to set a proper value. |
| Incorrect Mass | Delete all peaks with a mass defect in X.8 and X.9 which appear to be artifacts. Tips: If working on C,H,N,O compounds only, this option can be used. |

| | |
|---|---|
| Relative standard Deviation (RSD) | <p>Filter based on the Maximum RSD value set at 30 by default.</p> <p>The RSD is calculated on each defined class.</p> <p>If RSD of one feature is under the defined value for all class, it is removed from the peak list.</p> <p>Tips: To set a proper value, we advise to superpose QC chromatograms and inspect major and minor peaks all along the chromatogram.</p> |
| Relative Mass Defect (RMD)¹ | <p>RMD is calculated in ppm as ((mass defect/measured monoisotopic mass) × 10e6)</p> <p>Tips: Analysis of natural products from the DNP shows that 95 % of RMD are comprised between 50 and 3000 (values by default).</p> |
| Delete ghost peaks | <p>Delete variables with <i>m/z</i> values corresponding to mass drift of blank peaks.</p> |
| Maximum mass difference | <p><i>m/z</i> value tolerance set by default to 0.005 (in Da) for Pearson correlation and pos/neg merging. This parameter depends of MS instrument used.</p> <p>Tips: A value tow times above threshold used in MS-DIAL is a good starting point</p> |
| Maximum retention time difference | <p>RT value tolerance set by default to 0.025 (absolute value) for Pearson and pos/neg merging. This parameter will depend of chromatographic condition.</p> <p>Tips: A value tow times above threshold used in MS-DIAL is a good starting point</p> |
| Use Pearson correlation to compute clusters? | <p>Extend MS-DIAL-PCE clusters with Pearson correlation.</p> <p>Minimum correlation and maximum p-value are respectively set by default to 0.8 and $p \leq 0.05$</p> <p>Tips: This function increase cluster size and consequently reduce the number of feature in the final annotated peak list.</p> |

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

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

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

Combine positive and negative files from MS-DIAL and filter peaks according to user parameters.

Filters

Check which filters you want to use to clean your MS data.

Deltas

Indicates the acceptable retention time and mass differences to consider that peaks are related.

Clusterisation options

You can choose to use the Pearson correlation between peaks as a supplementary data used during clusterisation

(Optional) Reference files

Optionally, you can use your own files for adducts and neutral losses. See the documentation for more information.

What filters to use?

☒ Blank ratio

☒ Incorrect Mass

☒ Relative Standard Deviation

☒ Relative Mass Defect

☒ Delete ghost peaks?

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?

Minimum correlation: 0.8

Maximum p-value: 0.05

You can optionally import personal reference files for adducts and neutral losses. By default, data displayed in the Datasets tab will be used.

☐ Use personal reference files?

CLEAN MS-DIAL DATA

Cleaning data...

During the cleaning:

- Clusters are formed based on MS-DIAL-PCE algorithm, Pearson correlation, links such as adducts, neutral losses, dimers, ...;
- Adducts are corrected based on previous found links;
- If both modes were acquired, Pos and Neg clusters are concatenated if relational links are found (adducts mass difference)
- 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".

Shiny

http://127.0.0.1:4872 | Open in Browser

Check which filters you want to use to clean your ms data.

Deltas

Indicates the acceptable retention time and mass differences to consider that peaks are related.

Clusterisation options

You can choose to use the Pearson correlation between peaks as a supplementary data used during clusterisation

(Optional) Reference files

Optionally, you can use your own files for adducts and neutral losses. See the documentation for more information.

Maximum mass difference
0.005

Maximum retention time difference
0.025

☒ Use Pearson correlation to compute clusters? Minimum correlation
0.8

Maximum p-value
0.05

You can optionally import personal reference files for adducts and neutral losses. By default, data displayed in the Datasets tab will be used.

☐ Use personal reference files?

DELETE PREVIOUS RESULTS IF NECESSARY

Number of final peaks

Number of MS-DIAL links

Adduct / neutral loss relations

CLEAN MS-DIAL DATA

```

/!\ Deleting C:/Users/justine.chervin/Documents/Metatoul/BD0-maison/Metasci/script/final_data
/!\ Deleting C:/Users/justine.chervin/Documents/Metatoul/BD0-maison/Metasci/script/intermediary_data
*** Treating C:/Users/justine.chervin/Documents/Metatoul/BD0-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_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_msdl | List of MS-DIAL clusters in both pos and neg ionization |
| samples | List of samples with indication of sample name, class, file type, script class and column name |

Possible errors during this step:

Error: undefined column selected

→mismatch between *class names* in “pos” and “neg” mode

Error: Can't process data without blanks.

→Blank samples not defined in “File type”

2.3. 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.

Tips: 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/or “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 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

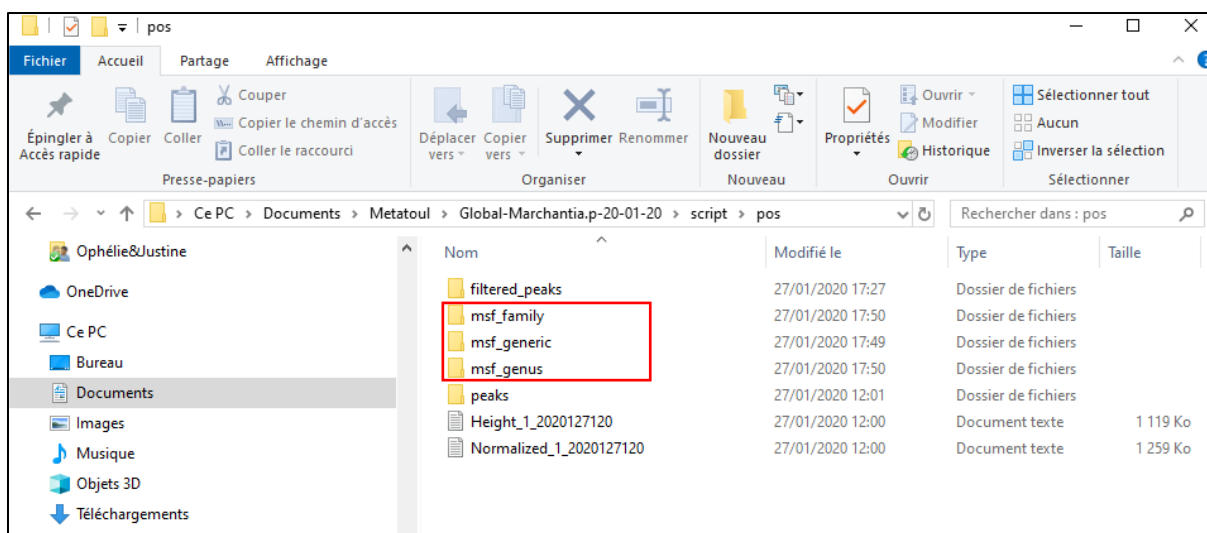
2.4. Interrogation of MS-FINDER

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 (X) 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. (In MS-FINDER menu: “Export→Export Batch results”)



2.5. 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.

msccleanr

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

Datasets

☒ 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).

1a,1b (A)

| Rank | Compound.level |
|------|----------------|
| 1 | 1a |
| 2 | 1b |

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.

genus, family, generic (B)

| Rank | Biosource.level |
|------|-----------------|
| 1 | genus |
| 2 | family |
| 3 | generic |

Levels scores

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

genus:2,family:1.5,generic:1|

| 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 rank 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.

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.

☐ 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-28-01-20/script/fina
*** Treating C:/Users/justine.chervin/Documents/Metatou/Global-Marchantia.p-28-01-20/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 features 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

| J | 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_warnir | source | Alignment | Average Rt. min | Average Mz | Adduct type | level | formula | Structure | Total score | Final score | Title | MS1 count | MSMS count | PRECURSOR | PRECURSOR TYPE | Theoretical mass | Mass error | Formula score | Ontology |
| 2 | Unknown compound | neg | 108 | 1.346 | 316.7885 | [M-H] ⁻ | | | | | | | | | | | | | | | |
| 3 | Unknown compound | neg | 173 | 6.558 | 427.0378 | [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 | 434 | 6.896 | 401.07574 | [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 | 5H7N03 | 15(1)-5-carboxy-4,5-dihydro | 7.0089 | 28.0356 | (IRAO) PYROGLUTAM | 3 | 13 | 128.0351 | [M-H] ⁻ | 129.0425931 | 0.0002166 | 4.495 | Alpha amino acids and derivatives |
| 36 | Simple ID | neg | 111 | 2.014 | 323.02853 | [M-H] ⁻ | | genus | 9H13N2O9P | 5-(12S,3R,4S,5R)-3,4-dihyd | 7.3562 | 28.4246 | (IRAO) URIDINE MO | 3 | 37 | 323.0285 | [M-H] ⁻ | 324.058666 | 9.02E-05 | 4.314 | Pentose phosphates |
| 37 | Simple ID | neg | 128 | 7.55 | 345.11874 | [M-H] ⁻ | | generic | 15H2O29 | 3,4,5-Trimethoxyphenyl g | 5.7246 | 5.7246 | | 3 | 26 | 345.1187 | [M-H] ⁻ | 346.126323 | 0.0004058 | 2.949 | Phenolic glycosides |
| 38 | Simple ID | neg | 13 | 3.904 | 129.01901 | [M-H] ⁻ | | generic | 5H6O4 | itaconic acid | 6.4729 | 6.4729 | (IRAO) ITACONATE | 3 | 5 | 129.019 | [M-H] ⁻ | 130.0266087 | 0.0003322 | 3.471 | Branched fatty acids |
| 39 | Simple ID | neg | 152 | 1.286 | 353.0696 | [M-H] ⁻ | | genus | 16H18O9 | 6-Hydroxy-2-methyl-7-[(2 | 5.8621 | 11.7242 | Unknown | 3 | 72 | 353.0697 | [M-H] ⁻ | 354.0950822 | -0.001894 | 3.344 | Phenolic glycosides |
| 40 | Simple ID | neg | 140 | 21.553 | 367.0425 | [M-H] ⁻ | | generic | 22H28N2O3 | 17-O-Acetylalpine | 3.9909 | 3.9909 | Unknown | 3 | 12 | 367.0425 | [M-H] ⁻ | 368.099928 | -0.001584 | 2.205 | Alkaline-terpene alkaloids |
| 41 | Simple ID | neg | 148 | 6.837 | 380.15628 | [M-H] ⁻ | | genus | 16H23N5O6 | (2R,3S,4S,5R,6S)-2-Hydrox | 6.5026 | 26.0104 | Unknown | 3 | 62 | 380.1563 | [M-H] ⁻ | 381.164835 | 0.001257 | 3.908 | Fatty acyl glycosides of mono- and disaccharides |
| 42 | Simple ID | neg | 155 | 13.242 | 392.20822 | [M-H] ⁻ | | genus | 21H31N06 | Pulchellamine G(+)-Pulcl | 6.5573 | 6.5573 | Unknown | 3 | 60 | 392.2082 | [M-H] ⁻ | 393.2151377 | -0.000339 | 3.752 | Guaniloides and derivatives |
| 43 | Simple ID | neg | 164 | 11.837 | 408.2038 | [M-H] ⁻ | | generic | 22H31N07 | Isorictin C(+)-Isorictin C | 5.7896 | 5.7896 | Unknown | 3 | 86 | 408.2029 | [M-H] ⁻ | 409.2100323 | -0.00124 | 3.514 | Oxepenes |
| 44 | Simple ID | neg | 171 | 18.496 | 423.16013 | [M-H] ⁻ | | generic | 20H28N2O5 | 2-(13S,3R,4R,9aS)-1-Hydr | 5.3463 | 5.3463 | Unknown | 3 | 77 | 423.1601 | [M-H] ⁻ | 424.1668076 | -0.000569 | 3.345 | Sulfanilides |
| 45 | Simple ID | neg | 174 | 11.272 | 435.1293 | [M-H] ⁻ | | genus | 21H24O10 | 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 |
| 46 | Simple ID | neg | 175 | 17.228 | 437.13907 | [M-H] ⁻ | | genus | 28H12O5 | Psychotriol A, 6'-Hydroxy | 6.7501 | 13.5002 | Unknown | 3 | 78 | 437.1391 | [M-H] ⁻ | 438.1467238 | 0.0003474 | 3.688 | Diarylethers |
| 47 | Simple ID | neg | 176 | 18.368 | 437.17554 | [M-H] ⁻ | | genus | 29H16O4 | 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 |
| 48 | Simple ID | neg | 18 | 1.722 | 133.01407 | [M-H] ⁻ | | genus | 4H6O5 | L-malic acid | 7.4137 | 29.6548 | (IRAO) MALATE | 3 | 10 | 133.0141 | [M-H] ⁻ | 134.0215233 | 0.0001468 | 4.245 | Beta hydroxy acids and derivatives |
| 49 | Simple ID | neg | 163 | 8.576 | 407.13458 | [M-H] ⁻ | | genus | 20H24O9 | (2S,3R,4R,5S,6R)-2-(2-[2-(3 | 6.663 | 13.326 | Unknown | 3 | 64 | 407.1346 | [M-H] ⁻ | 408.1420323 | 0.0001559 | 3.776 | Stribene glycosides |
| 50 | Simple ID | neg | 189 | 12.962 | 449.10275 | [M-H] ⁻ | | genus | 28H18O6 | 2,2',3,3',7,7'-Hexahydroxy- | 5.9968 | 21.9872 | Unknown | 3 | 30 | 449.1028 | [M-H] ⁻ | 450.1103383 | 0.0002618 | 3.418 | Phenanthrois |
| 51 | Simple ID | neg | 19 | 1.344 | 135.0298 | [M-H] ⁻ | | generic | 4H8O5 | Ethynic acid | 7.0979 | 7.0979 | Unknown | 3 | 27 | 135.0298 | [M-H] ⁻ | 136.0371794 | 9.69E-05 | 3.559 | Sugar acids and derivatives |
| 52 | Simple ID | neg | 190 | 12.714 | 449.10278 | [M-H] ⁻ | | genus | 28H18O6 | 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 |
| 53 | Simple ID | neg | 191 | 17.84 | 452.27817 | [M-H] ⁻ | | generic | 21H44N07P | LysopEIO(0/16.0) | 6.2888 | 6.2888 | Unknown | 3 | 30 | 452.2782 | [M-H] ⁻ | 453.2853394 | 6.29E-05 | 3.643 | 2-acyl-sn-glycero-3-phosphoethanolamines |
| 54 | Double ID | pos | 465 | 10.968 | 463.08771 | [M+H] ⁺ | | genus | 22H38O12 | (2R,3R,4S,5S,6S)-6-[(2-(3,4 | 7.4323 | 29.7295 | Unknown | 3 | 54 | 463.0877 | [M+H] ⁺ | 462.079626 | -0.000598 | 4.628 | Flavonoid-7-O-glucuronides |
| 55 | Double ID | pos | 466 | 10.005 | 463.08774 | [M+H] ⁺ | | genus | 21H38O12 | (2R,3R,4S,5S,6S)-6-[(2-(3,4 | 7.5668 | 30.2672 | Unknown | 3 | 39 | 463.0877 | [M+H] ⁺ | 462.079626 | -0.000598 | 4.61 | Flavonoid-7-O-glucuronides |
| 56 | Double ID | pos | 468 | 17.402 | 471.18042 | [M+H] ⁺ | | generic | 21H30N2O8S | Dacarpamine | 5.4081 | 5.4081 | Unknown | 3 | 237 | 471.1804 | [M+H] ⁺ | 470.1722869 | -0.000837 | 3.534 | Methionine and derivatives |
| 57 | Double ID | neg | 207 | 10.179 | 473.20294 | [M-H] ⁻ | | generic | 22H34O21 | UNPD79762 | 5.6695 | 5.6695 | Unknown | 3 | 61 | 473.2029 | [M-H] ⁻ | 474.2101119 | -6.45E-05 | 3.305 | Terpene glycosides |
| 58 | Double ID | pos | 517 | 8.931 | 439.11963 | [M+H] ⁺ | | genus | 22H36O18 | (2R,3R,4S,5R,6S)-6-[(2-(3-[| 7.244 | 28.676 | Unknown | 3 | 47 | 439.1196 | [M+H] ⁺ | 438.1133414 | -0.00041 | 4.641 | Flavonoid-7-O-glucuronides |
| 59 | Double ID | pos | 181 | 17.298 | 439.15469 | [M+H] ⁺ | | genus | 28H24O5 | 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 |
| 60 | Double ID | pos | 436 | 17.098 | 425.17463 | [M+H] ⁺ | | generic | 20H28N2O5 | 2-(13S,3R,4R,9aS)-1-Hydr | 5.8585 | 5.8585 | Unknown | 3 | 115 | 425.1747 | [M+H] ⁺ | 424.1668076 | -0.000616 | 3.618 | Sulfanilides |
| 61 | Double ID | pos | 450 | 17.064 | 441.16974 | [M+H] ⁺ | | genus | 28H24O5 | Marchantin C, 2'-Hydroxy | 6.2661 | 24.9444 | Unknown | 3 | 194 | 441.1697 | [M+H] ⁺ | 440.1623739 | -4.07E-05 | 3.5 | Lignans, neolignans and related compounds |
| 62 | Double ID | pos | 45 | 7.123 | 203.08241 | [M-H] ⁻ | | generic | 11H12N2O2 | L-Tryptophan | 8.0971 | 8.0972 | (IRAO) TRYPTOPHAN | 3 | 41 | 203.0824 | [M-H] ⁻ | 204.0898776 | 0.0002012 | 4.082 | Indolyl carboxylic acids and derivatives |

2.6. Export peaks as .msp files

Shiny

http://127.0.0.1:7131

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

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/Metabolu/Global-Marchantia.p-2

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.

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

13