MSProfileR tool - Tutorial

Table of Contents

Introduction

The MSProfileR is a web application to analyse mass spectra profiles, to assess their quality, to detect their peaks and to classify spectra by clustering tool based on their profiles.

To launch the MSProfileR tool on Windows, Linux or macOS, the user must apply the following instructions mentioned in the README file on the following github link https://github.com/Almeras-Lionel/MSProfileR

- install R (https://cran.biotools.fr/bin/windows/base/R-4.3.1-win.exe) without changing the proposed directory;

- Install the development version from GitHub, by typing on rstudio console environment:

```
> install.packages("devtools")
> devtools::install_github("Almeras-Lionel/MSProfileR")
```

-Install the vignette:

>devtools::install_github("Almeras-Lionel/MSProfileR", upgrade = "never")

-Install the “MSProfileR” package on R:

```
> library(MSProfileR)
```

- Launch the app:

```
> runMSProfileR()
```

The interface app will be accessible for users, after these steps and a tutorial is available on the github package page.

This application is based on the workflow presented in Figure 1:

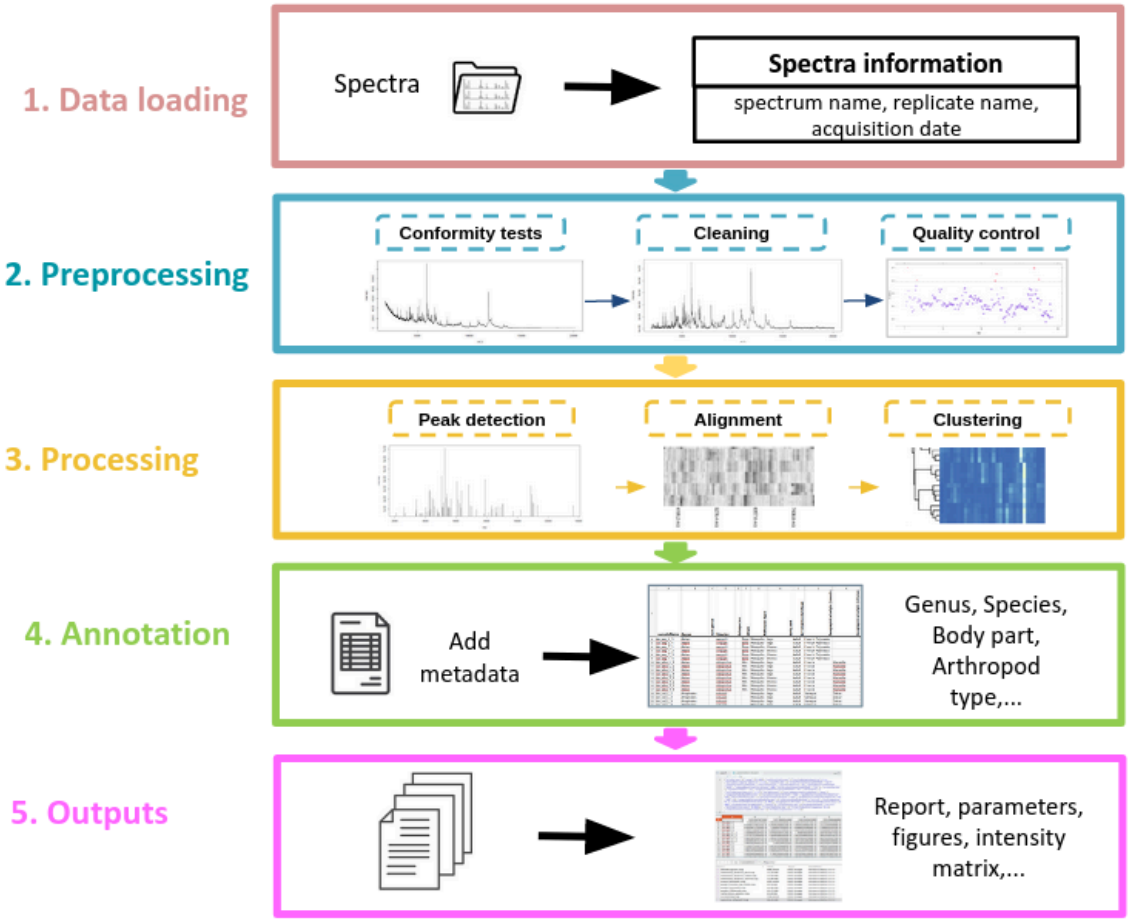

Figure 1. MSProfileR workflow. The user interface is structured in 5 sections: 1) Data loading, 2) Pre-processing, 3) Processing, 4) Annotation and 5) Outputs.

1. Data loading

The data loading part compounds two modules, the spectra loading from the user device and the uploading of setting parameters applied in a previous analysis. The uploading of setting parameters is optional (Figure 2).

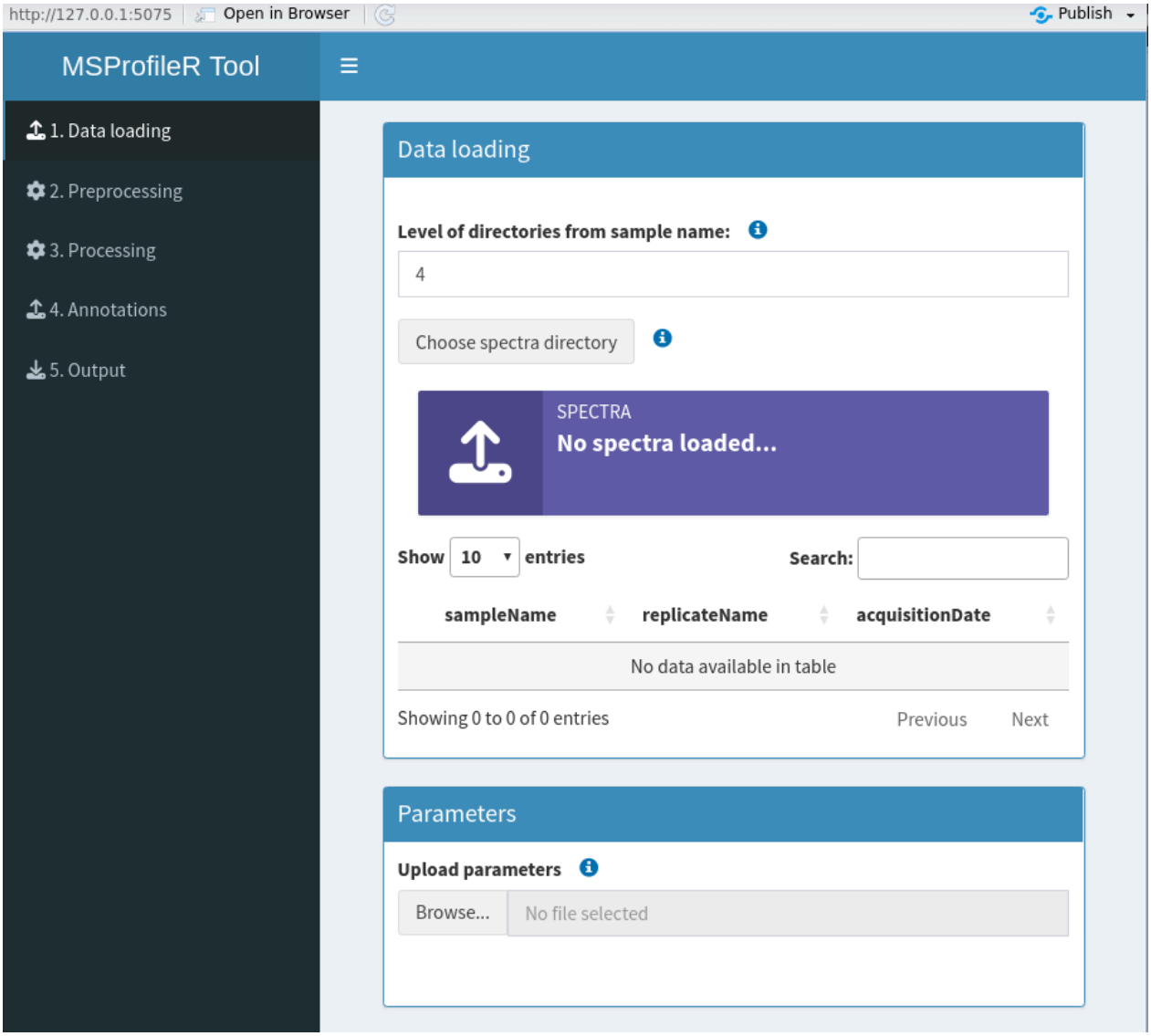

Figure 2. *Screen capture of the data loading page*

1.1 Module of the MS spectra loading

The first step consists of uploading MS spectra from the dataset. The user must select the path of the folder containing the spectra files to analyse in format of Bruker Daltonics flex series mass spectrometer (fid and acqu files) (Figure 3).

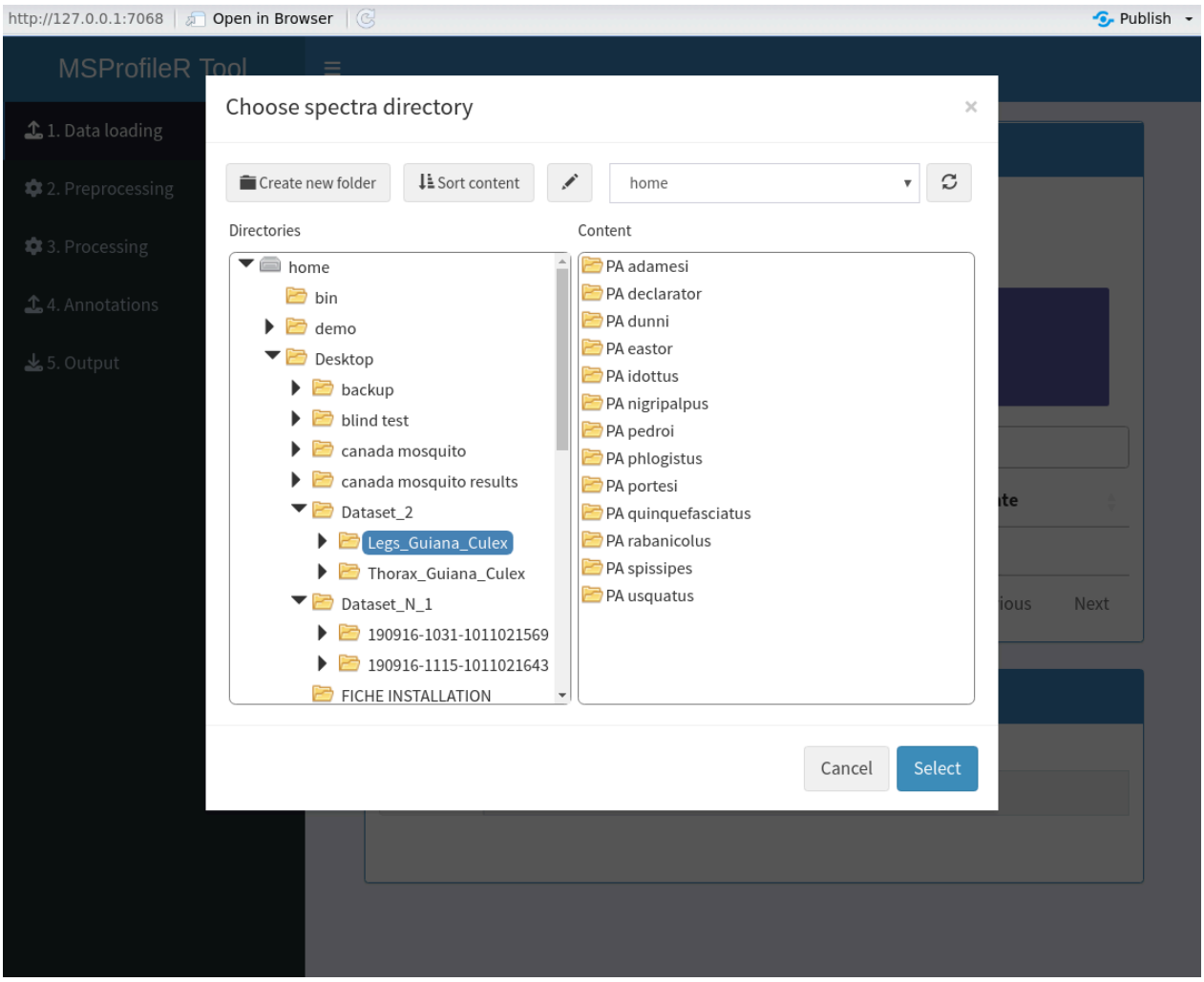

Figure 3. *Screen capture of the selected folder to import*

NB: For the moment, the MS-profileR tool works only with mass spectra files from Bruker Daltonics. The next updates will make it compatible with mass spectra files from other manufacturers notably Shimadzu/Vitek MS. The aim is to allow scientists working with spectrometers from different manufacturers to use the same tool for analysing mass spectra profiles.

During the importation of spectra, the data stored in each Bruker spectrum file are retrieved and summarized in the table including the sampleName, the replicateName, and the acquisitionDate.

NB: According to the version of the Bruker Daltonics acquisition software, it was noticed that the number of folders generated between the sampleName and the fid or acqu files could vary. Until now, we observed four or five folder levels. This path is the primordial key to download the spectra correctly, in terms of extracting the spectra information in a table presented on the interface, such as the sample and replicate names to import them in the sampleName and the replicateName columns, respectively. In this way, to translate appropriately the path organisation, a box titled "Level of directories from sample name" is available. By default, a level of four folders was set. Once this level is set, the user can load the spectra by clicking on "chose spectra directory" and control that the sampleName, replicateName and the acquisitionDate are correctly classified, in other way, the user can modify this level according to its dataset structure.

1.2 Module of parameter loading

The user has the option to upload register parameters set from the analysis of a previous dataset (Figure 4). The parameters were saved in a Json file. These upload parameters are then applied for the entire MSProfileR process. This option allows to the user the analysis of the current dataset with the same parameters. Nevertheless, the user keep the hand to modify all the parameters and methods used if necessary.

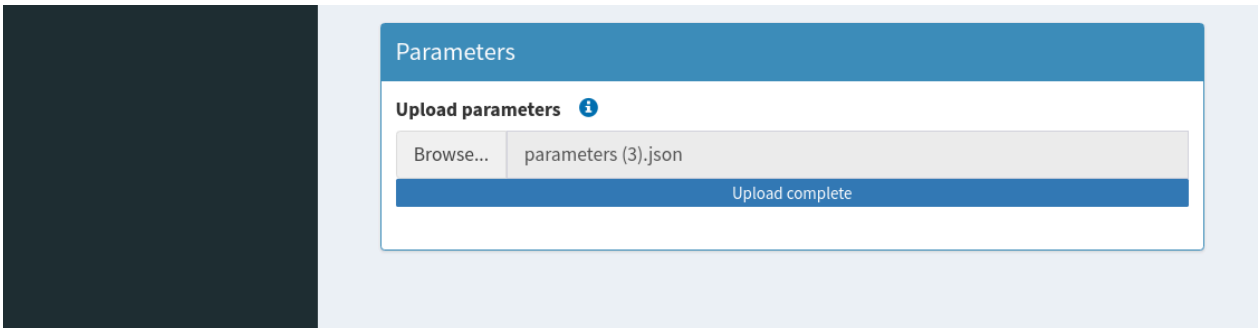

Figure 4. *Screen capture of the parameters uploading*

2. Pre-processing

The loaded spectra are preprocessed in four successive modules, 1) the trimming and conformity tests, 2) the cleaning, 3) the quality control and 4) the averaging of spectra.

2.1 Module of trimming and conformity tests

Although the same m/z acquisition scale was applied to MALDI-TOF-MS apparatus among different series of samples analysed, variations of m/z upper and lower limits occurred. To facilitate spectra comparison, the trimming step, which delimits the inferior and superior borders of the spectra dataset by selecting kDa values (Figure 5). The purpose of this step is to ensure whether the lengths (i.e. ranges of m/z values) of spectra are homogeneous.

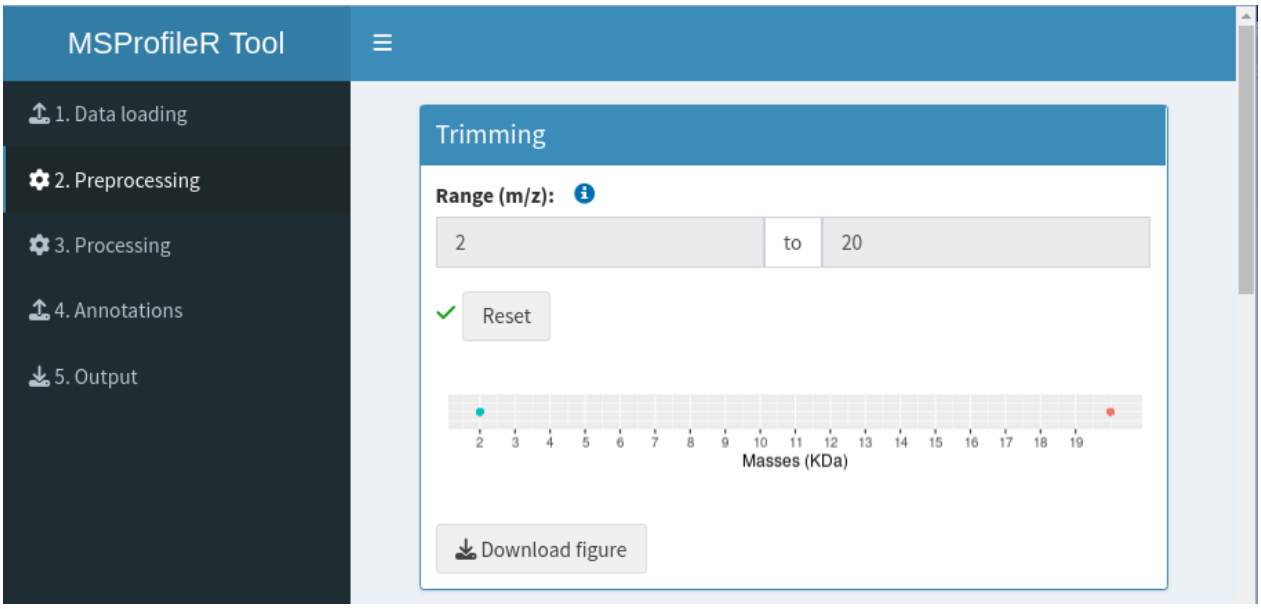

Figure 5. *Screen capture of the trimming page*

The conformity tests checked the compliance of the spectra by assessing three criteria (Figure 6):

1. Are there any empty spectra? (no data)
2. Are there any spectra with irregular m/z values? (uninterrupted values)
3. Do length (i.e. m/z values range) of spectra differ? (similar number of values).

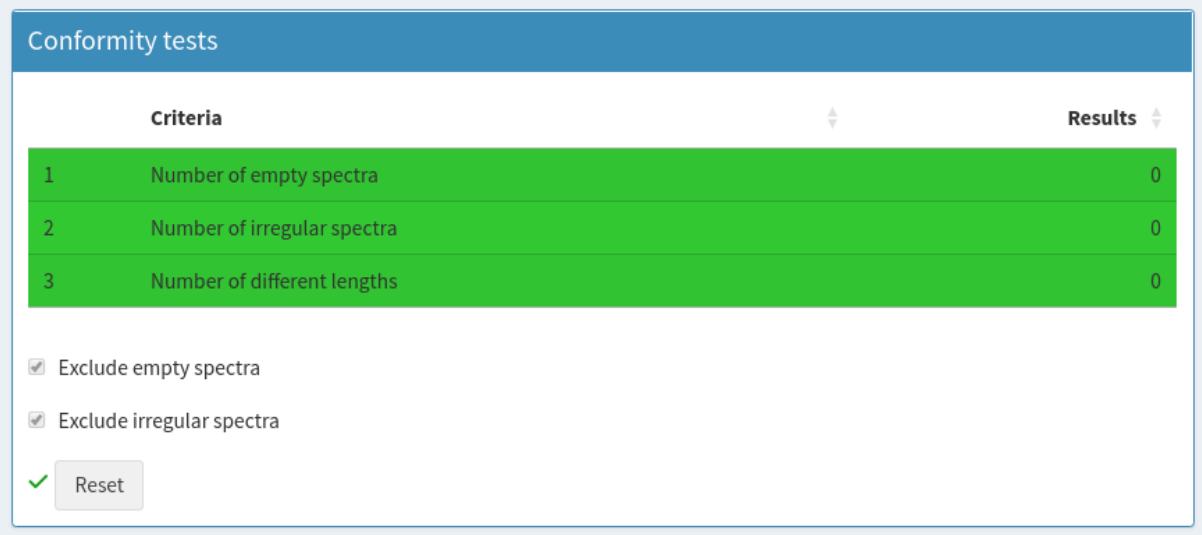

Figure 6. *Screen capture of the conformity tests page*

If all spectra meet the criteria, the rows are coloured in green. On the other hand if the criteria are unmet, the corresponding rows are coloured in grey and the respective number of spectra was indicated. All non-compliant spectra are excluded by default, but they can be conserved by unticking respective boxes.

2.2 Module of spectra cleaning

MS spectra which passed the conformity tests are then subjected to a succession of transformations in order to homogenise the data (Figure 7): 1) intensity transformation, 2) smoothing, 3) baseline correction and 4) normalisation.

(1) For intensity transformation, the square root transformation (sqrt) is used by default for variance stabilisation and to improve the graphical visualisation of spectra, by reducing the flattening effect exerted by the high-intensity peaks on the low-intensity peaks. Three logarithmic transformations are also available to stabilize the intensity variance of spectra profiles by reducing the fluctuation of high and low intensity peaks. (2) The spectra are smoothed using the Savitzky-Golay-Filter algorithm or moving average methods to reduce spectra noise. (3) Four methods (SNIP, TopHat, Convex hull or Median) are available to correct the baseline of each spectrum. (4) To calibrate intensities, spectra normalization was done using one of the three methods provided (TIC, PQN or Median). All these transformations are performed through the mass spectra processing functions (transformIntensity, smoothIntensity, removeBaseline, calibrateIntensity) of the MALDIquant and MALDIrppa packages.

For each step in this module, the user can select either method offered by MSProfileR tool to clean the dataset. In addition, the user can visualise the transformation carried out of each profile of spectra by the plots on the interface and exporting them on his device.

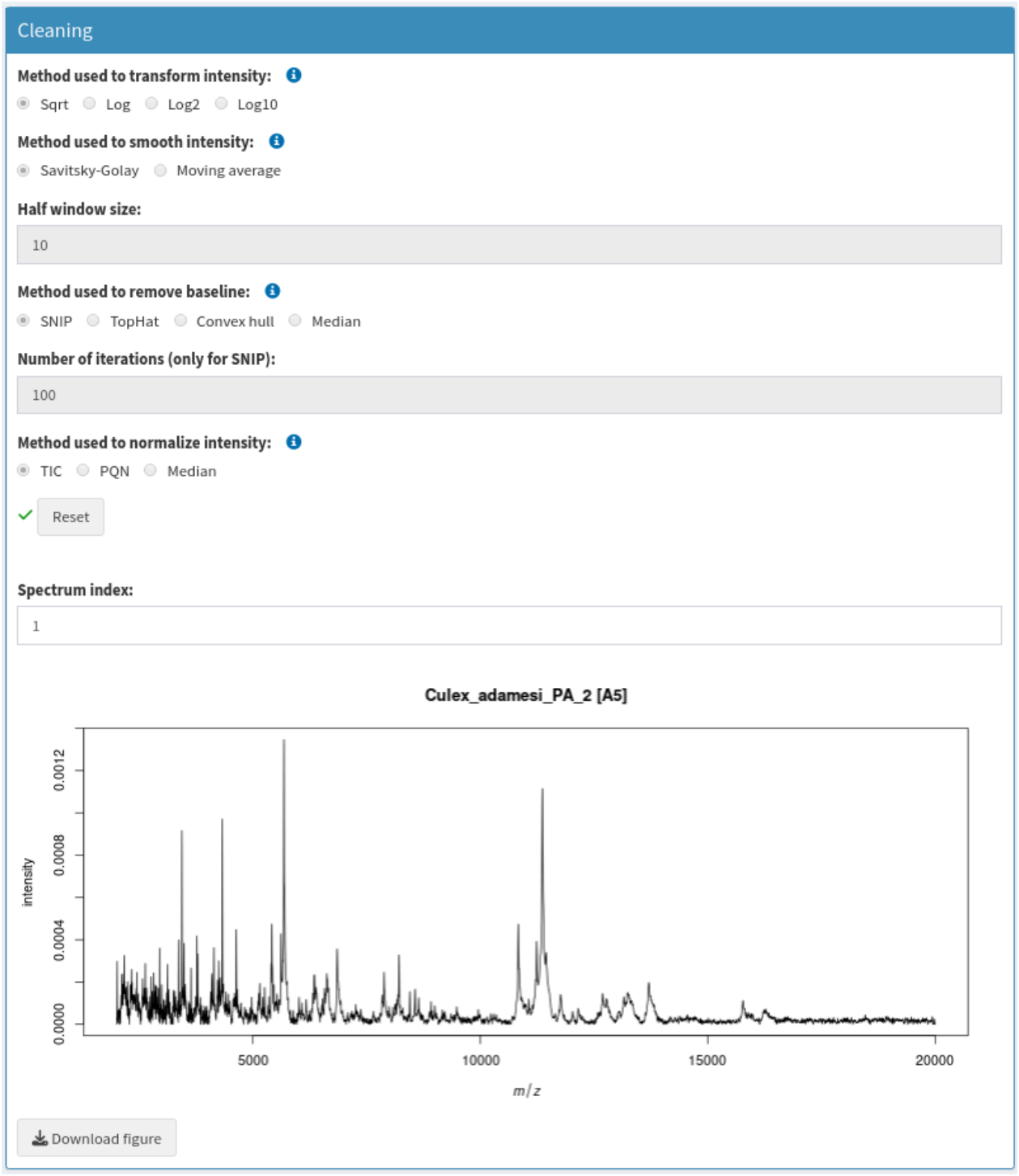

Figure 7. *Screen capture of the cleaning page*

2.3 Module of quality control

The compliant spectra are screened to detect the low-quality MS spectra. Generally, low-quality spectra are characterised by rippled and indented profiles due to low signal to noise intensities which could be absorbed by the background noise. For each spectrum, an atypical score is computed (Figure 8). The calculation of this score is based on robust estimators (Q, MAD) of the derivative mass spectra and median of the intensities. Five methods for estimating the tolerance limit are available (RC, boxplot, ad.boxplot, Hampel, ESD). The results of the screening are illustrated through a plot as above which represents the atypical scores of the spectra indexed according to the order they were loaded. The spectra with a score outside the cut-off value are coloured in red and regarded as outlier spectra. Using the side panel on the left, the user can modify the parameters of screening such as the estimator used in the calculation of the atypical scores or the threshold and method from which the cut-off is estimated. In this way, the user can observe in real time the changes in the screening results.

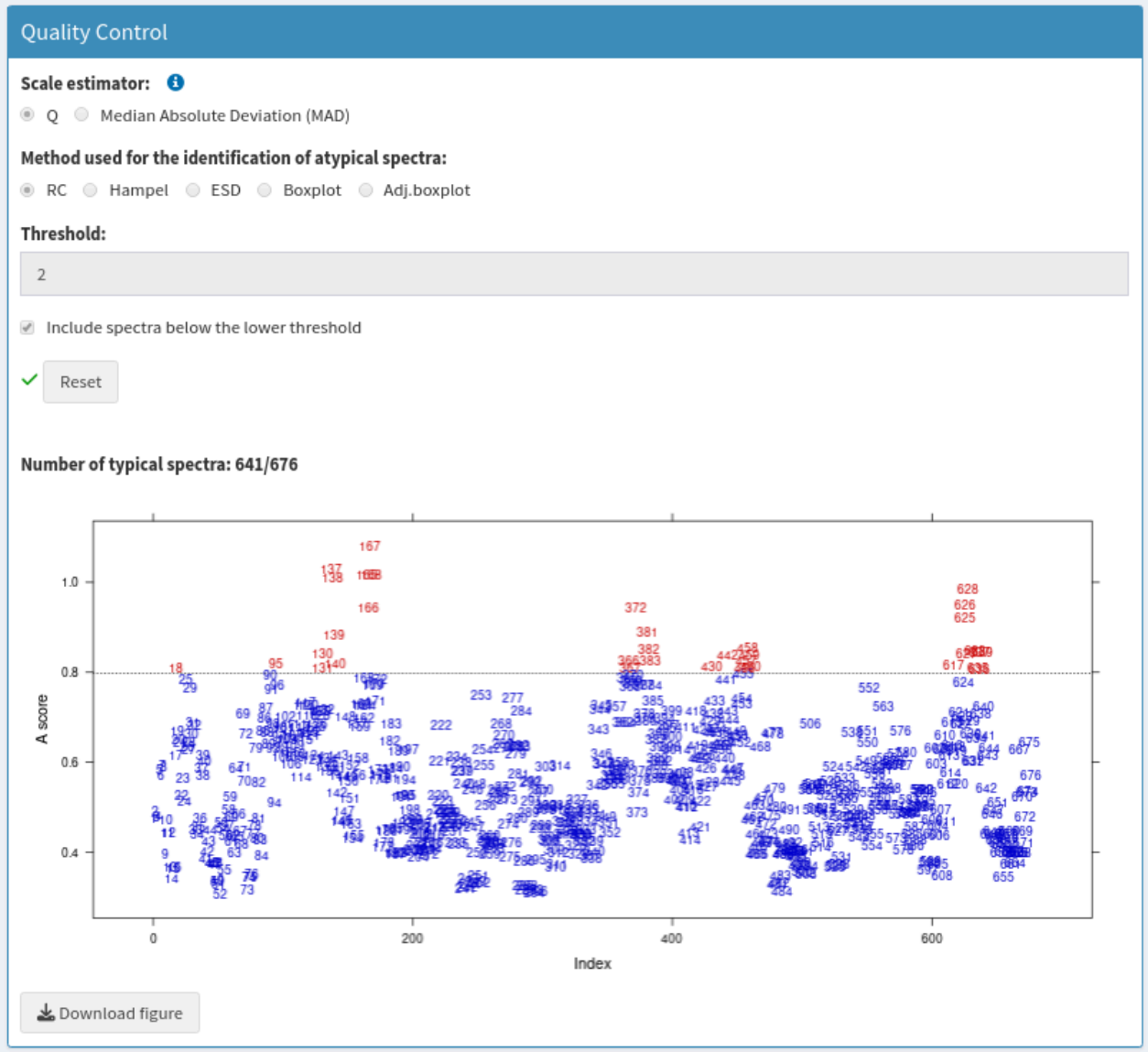

Figure 8. *Screen capture of the spectra screening step*

By default, the spectra detected as outliers (coloured in red on the plot) are automatically rejected for the rest of the analysis. But the user can inspect the spectra profiles and decide definitively which spectra to keep or to reject for the rest of the analysis.

In the selection panel (Figure 9), spectra are indexed as outliers and the typical spectra, in the boxes “atypical spectra” and “selected spectra”, respectively, according to their A score value compared to threshold. Using the listing, numbers could be selected for plot visualisation of respective spectra. With the arrow boxes, the user can move selected spectra from one list to another.

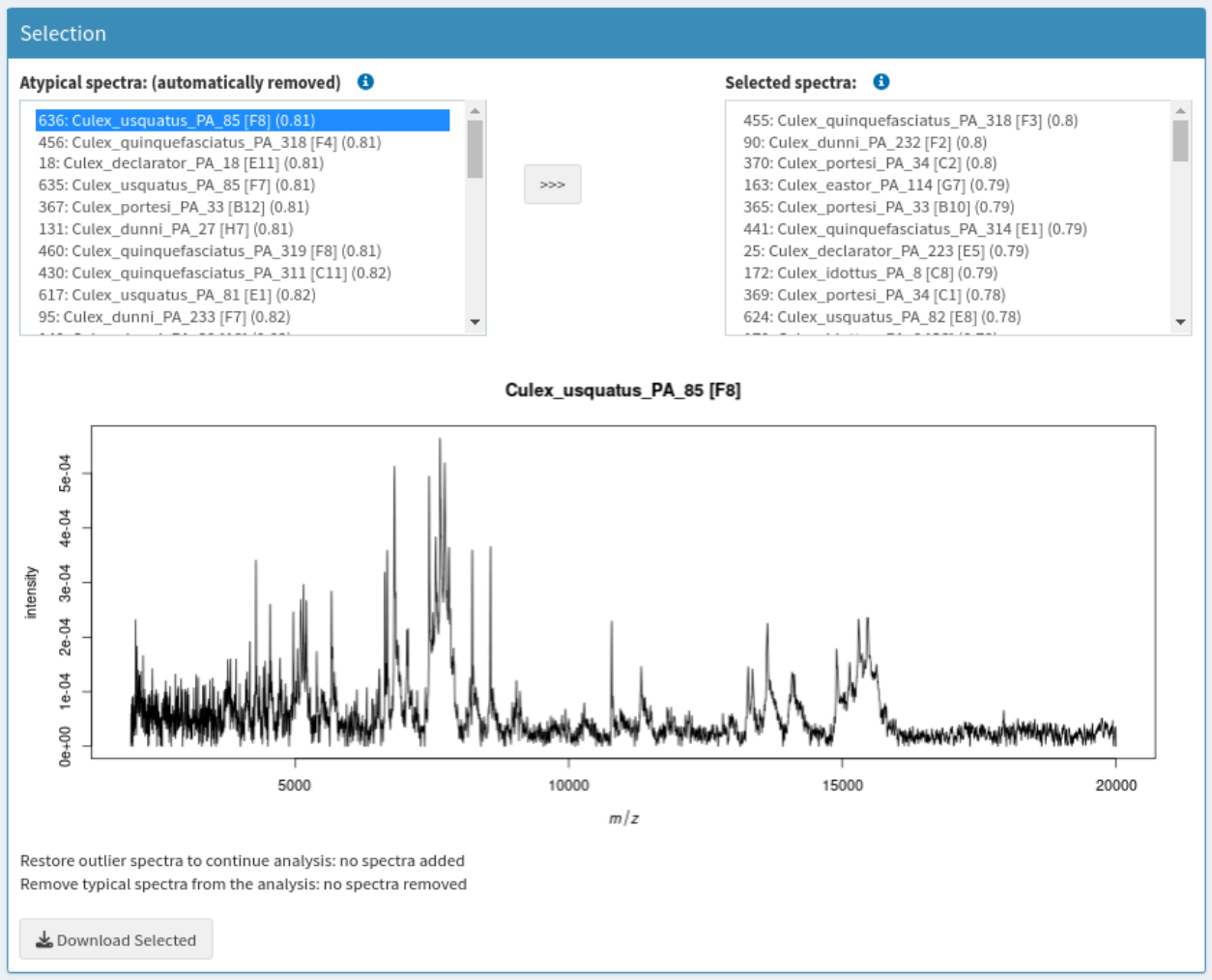

Figure 9. *Screen capture of the graphical interface dedicated to visualisation of spectra*

2.4 Module of averaging

To avoid analysing technical replicates from the same sample, average spectra are computed at the end of this step from annotated spectra which successfully passed the quality control steps. The user can choose one among three methods (mean, median and sum) to generate the future representative averaged spectra which is illustrated in the attached table below (Figure 10). The sampleName of the averaged spectra and the spot position of replicates are indicated in the table.

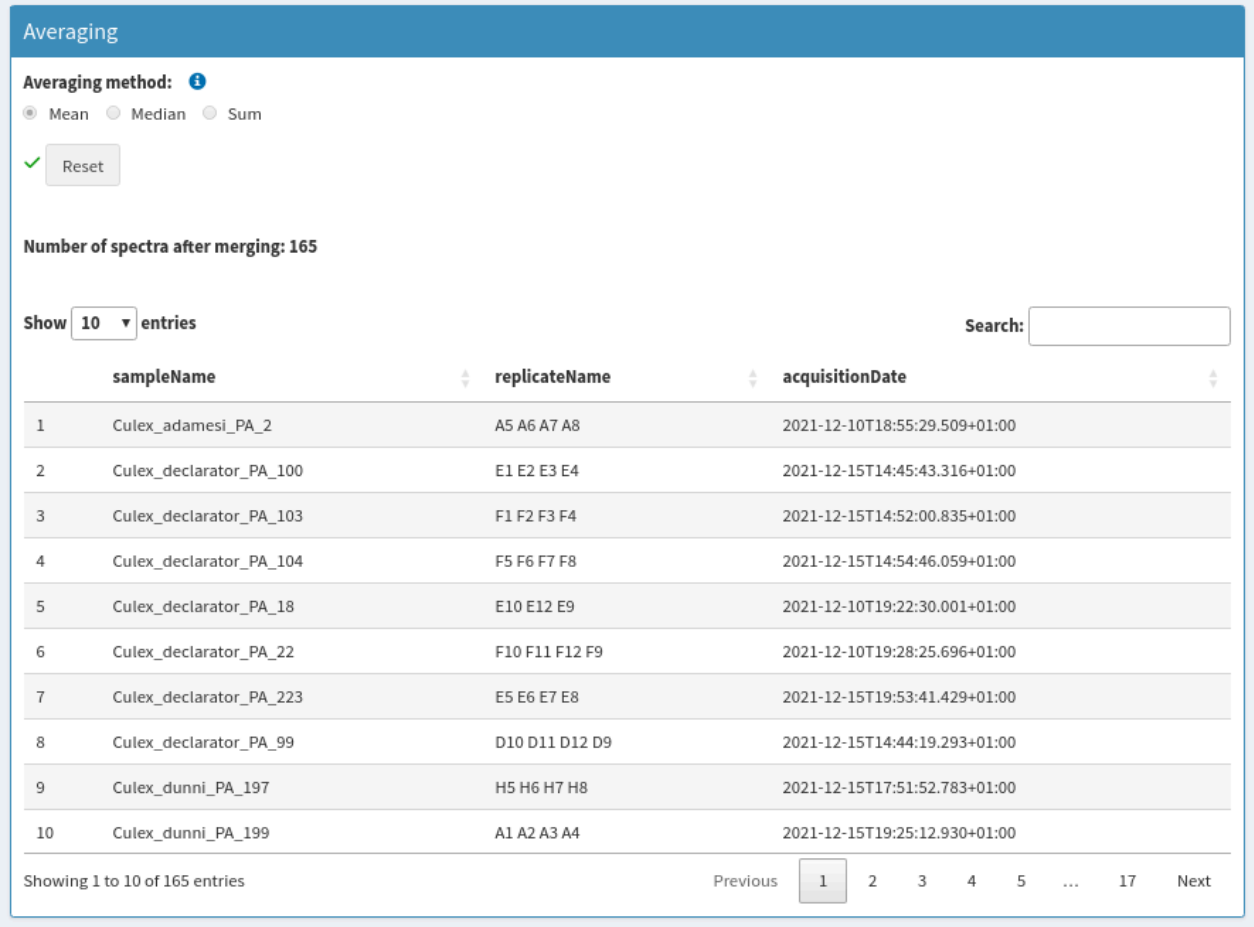

Figure 10. *Screen capture of the spectra averaging*

3. Processing

At the end of the preprocessing part, spectra profiles can be compared on the basis of the detected peaks through the following modules of processing and can be visualised by the graphical representations.

3.1 Module of peak detection

The first module of the processing is the detection of peaks for each spectrum (Figure 11). The user can visualise the distribution of the number of peaks detected per averaged spectra according to the Signal to Noise Ratio (SNR). A boxplot illustrated the peak number distribution and the standard deviation for SNR values ranging from 2 to 7 in front of the interface. Based on this boxplot, a slider allows to the user to select the optimal value of SNR for peak detection. A visualization of the peak list detected per spectra is possible on the plot. Numbers on the plot order detected peaks according to their intensity from the higher to the lower.

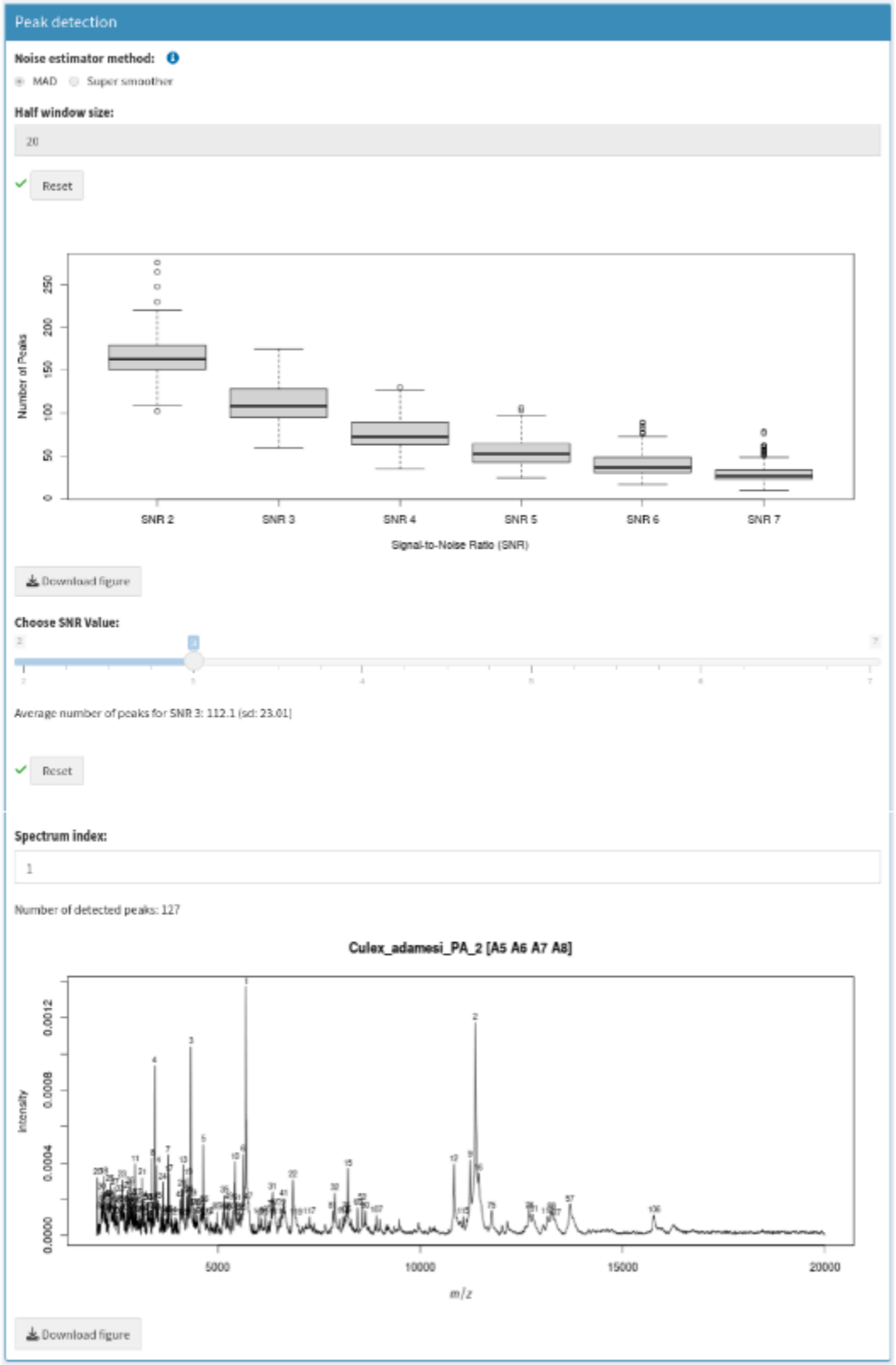

Figure 11. *Screen capture of the peak detection panel*

3.2 Module of spectrum alignment

The alignment module includes four successive steps, 1) the reference peak detection, 2) the warping (correction phase), 3) the binning and 4) the filtering.

(1) To realize an alignment, the determination of reference peaks among the spectra dataset is compulsory (Figure 12). This list is obtained by adjusting two parameters, the minimum frequency (minimum of peak occurrence) and the alignment tolerance of reference peaks. The number of the reference peaks is indicated as well as their m/z distribution available on a plot in the interface. An adjustment of the two parameters remains possible by the user to decrease or increase the number of the reference peaks.

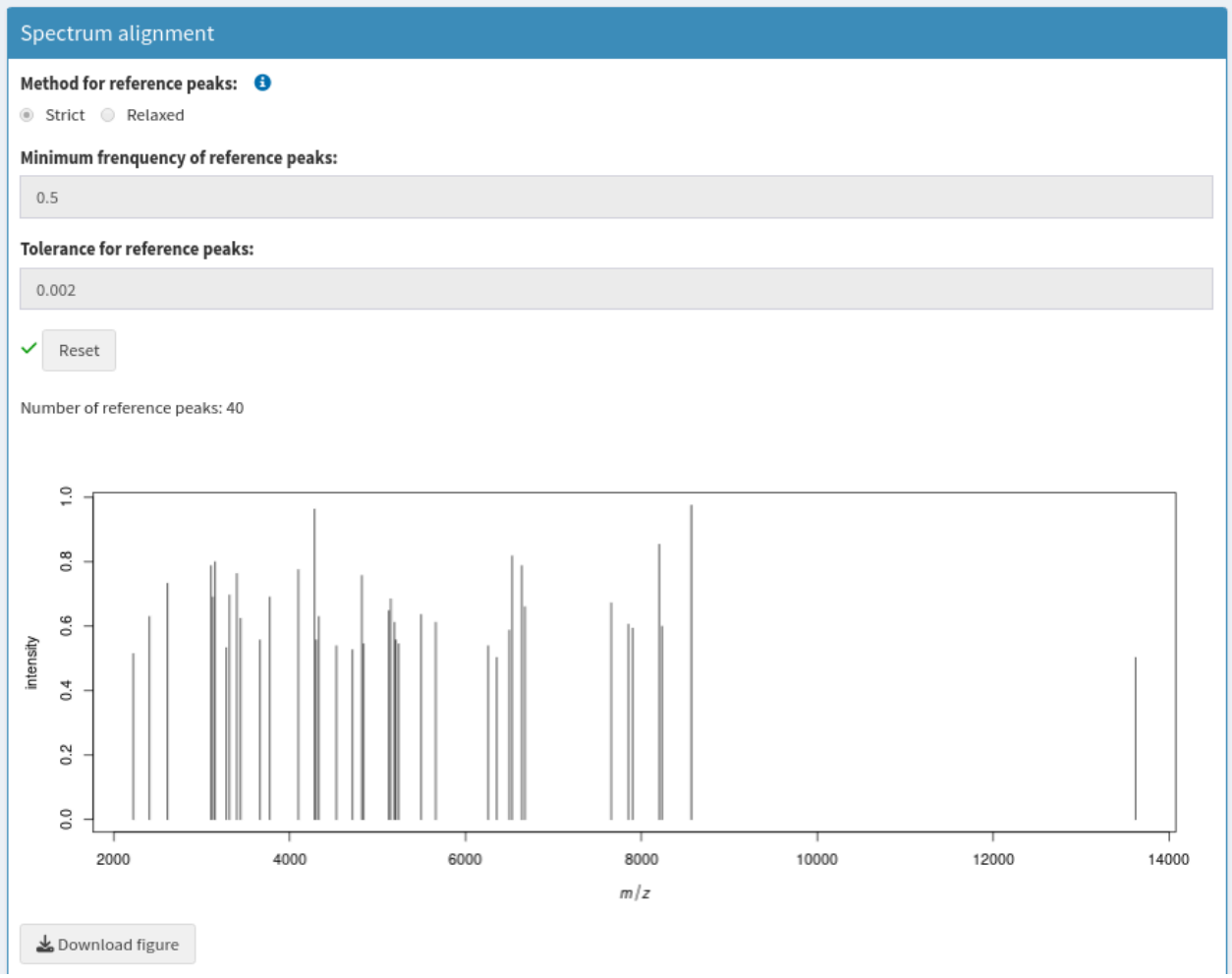

Figure 12. *Screen capture of the reference peak detection panel*

(2) The next alignment step consists in the applying of a warping method among four available (i.e., lowess, linear, quadratic and cubic). A gelview on grey scale representing the entire averaged, illustrates the results of warping step (Figure 13).

Figure 13. *Screen capture of the warping panel*

(3) For the binning step, the user can control the tolerance to consider whether the near peaks in the m/z range are the same one or not, with choosing one of the two methods: strict or relaxed (Figure 14).

Another parameter called Minimum frequency not related to the peak detection is also implemented in the step of filtering (Figure 15). It is used as a visualisation option for the graphical plot. It allows to display peaks with occurrence frequency greater than the indicated percentage (set at 20% per default). For example, by setting this parameter to 100%, uniquely peaks common to all spectra will be selected for the next step. The chosen minimum frequency defines the number of filtered peaks that will serve as input for the clustering module.

The results of the binning and filtering steps are illustrated by a gelview and the number of peaks included at each step was also indicated.

Figure 14. *Screen capture of the binning panel*

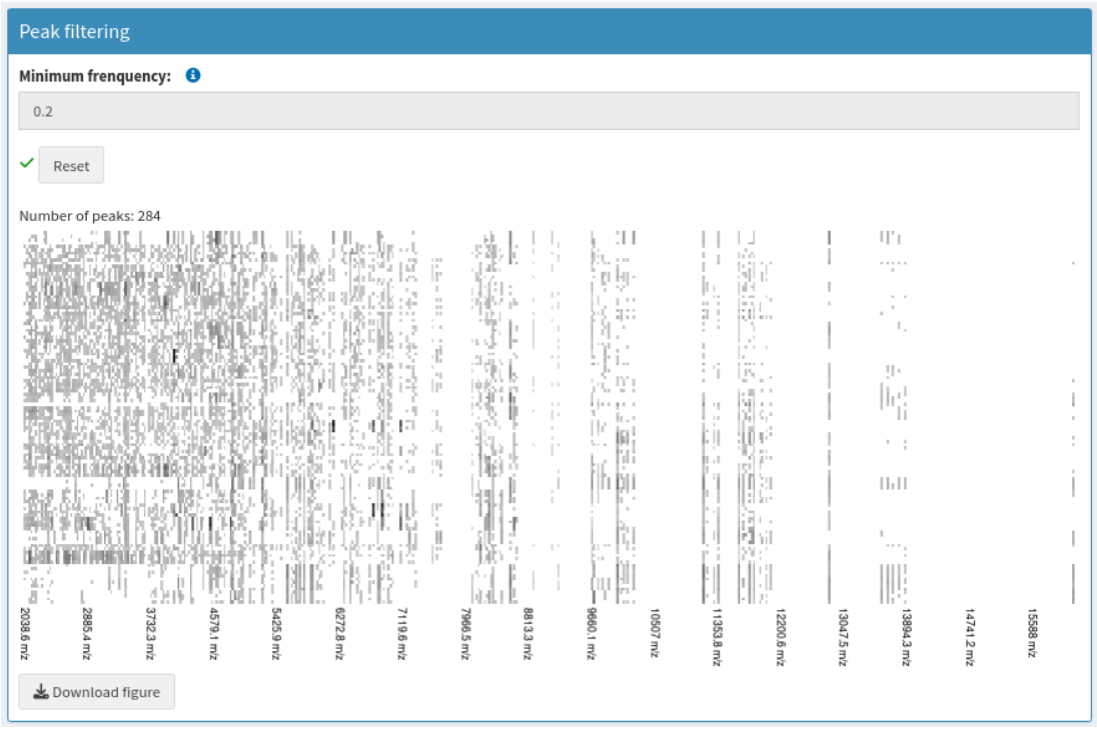

Figure 15. *Screen capture of the filtering panel*

3.2 Module of spectra clustering

Based on the results of the filtering panel, an intensity matrix is generated and a classification of averaged spectra is operated by the clustered heatmap (Figure 16). To avoid missing values (NA values), the respective intensities from each averaged spectra were retrieved and filled into the intensity matrix. This action allows to obtain a more realistic view of the profiles and will be helpful for future statistical analyses required complete data (no missing values). A checkbox is available to deactivate this filling. In this case, when peaks are missing, NA values are inserted into the intensity matrix and are indicated in grey. The results of spectra clustering are downloadable, providing uniquely the dendrogram (Download dendrogram) or three clustering heatmap with distinct dimensions. The "Download matrix (8x4)" plot a clustering heatmap without the sample names with a low resolution; the "Download matrix (16x8)" plot a clustering heatmap with sample names (for large dataset, an overwriting of sample names could occurred); the "Download matrix (32x16)" plot a clustering heatmap with sample names of high resolution preventing the risk of sample names overwriting also for large dataset.

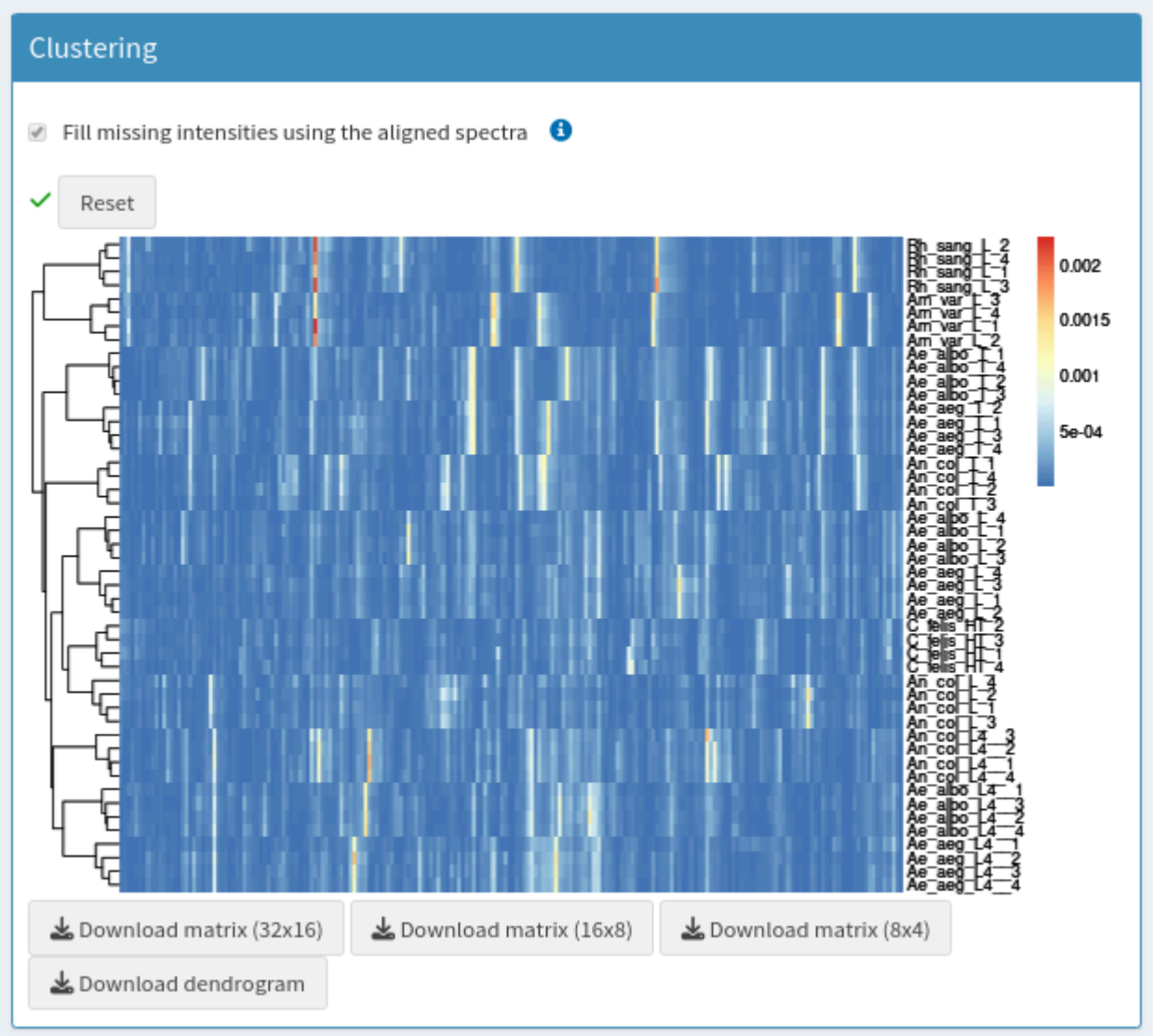

Figure 16. *Screen capture of the clustering module*

4. Annotations

This tool was initially developed to analyse mass spectra profiles of arthropods. However, it is possible to import MS profiles from diverse origins and to include annotations. This module is specifically created to annotate spectra, taking into account some relevant information such as the species, genus, arthropod type, body part from which the sample was extracted (e.g. thorax or legs), geographical origin, developmental stage, storing mode, collection date or else the protocol applied for sample preparation.

For spectra annotation, an empty Xlsx table can be downloaded (Figure 17) containing only the name of the averaged spectra under the primordial column headed sampleName (Figure 18). The user can then fill this table with desired supplementary information in the next columns. The annotated table can then be uploaded using the panel "upload an annotation table" (Figure 19). At the end of this step, the tool generates a summarising table to verify if all the spectra are annotated or not, this table is an indicator of the correspondence of the spectra with their annotations (Figure 20). This table contains four lines serving to verify the correspondence between the annotation table and averaged spectra list. The first two lines control whether the number of annotations to add and the number of averaged spectra obtained from the MSProfileR process are equal. The last two lines check the number of annotations without spectra and inversely the number of spectra without annotation allowing to control a correct pairing of annotation and spectra.

NB: The importation of annotated spectra is not mandatory. This tool is also usable for analysing any mass spectra profiles without annotations.

Figure 17. *Screen capture of the empty annotation table downloading button*

	A	B	C	D	E	F	G
1	sampleName						
2	Ae_aeg_L_1						
3	Ae_aeg_L_2						
4	Ae_aeg_L_3						
5	Ae_aeg_L_4						
6	Ae_aeg_T_1						
7	Ae_aeg_T_2						
8	Ae_aeg_T_3						
9	Ae_aeg_T_4						
10	Ae_albo_L_1						
11	Ae_albo_L_2						
12	Ae_albo_L_3						
13	Ae_albo_L_4						
14	Ae_albo_T_1						
15	Ae_albo_T_2						
16	Ae_albo_T_3						
17	Ae_albo_T_4						
18	An_col_L_1						
19	An_col_L_2						
20	An_col_L_3						
21	An_col_L_4						

Figure 18. *Screen capture of the empty annotation table generated by the tool*

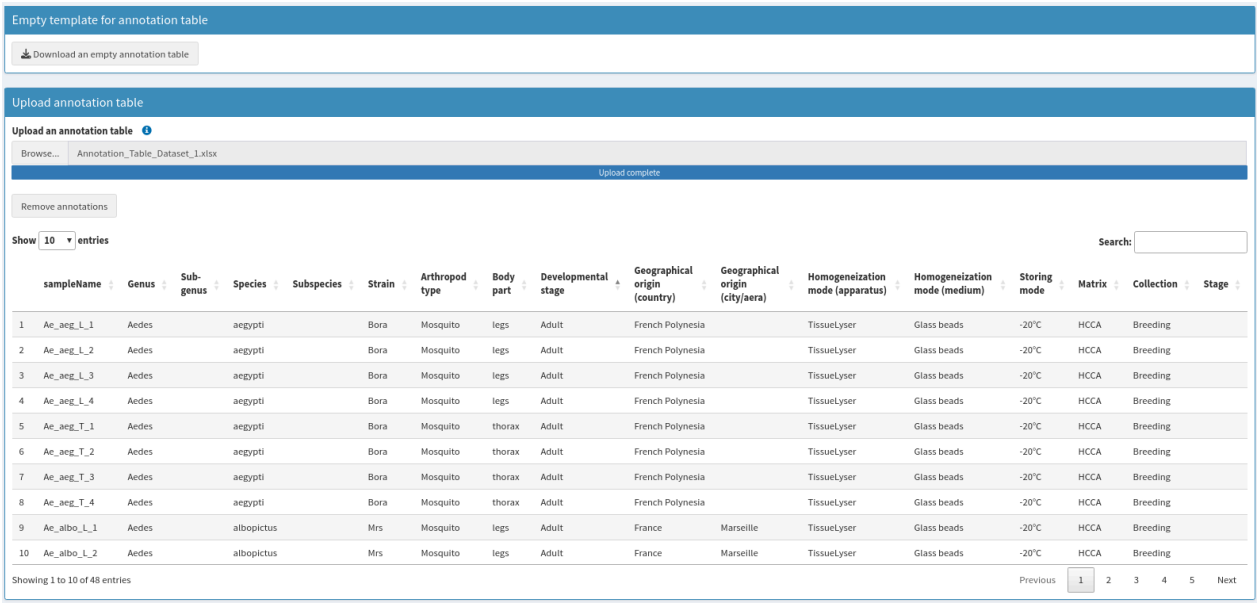

Figure 19. *Screen capture of the filled annotation table loaded by the tool*

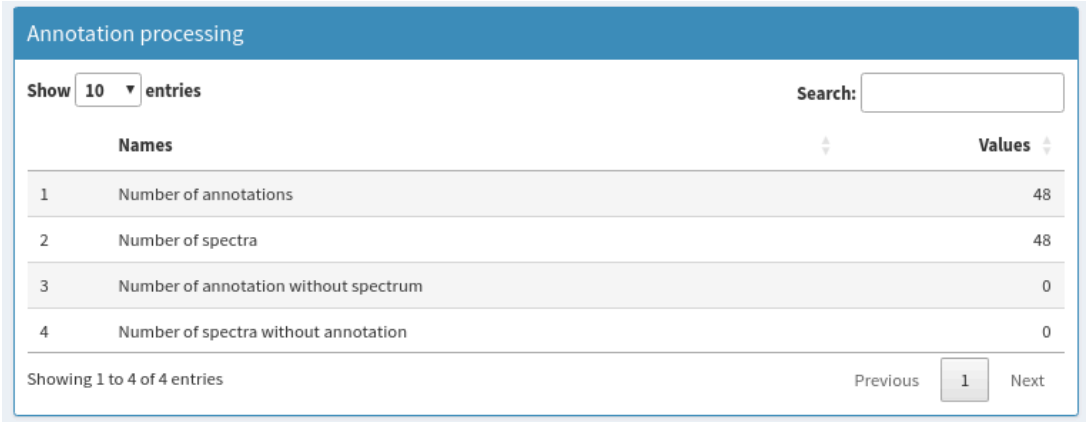

Figure 20. *Screen capture of the confirmation of spectra annotation*

4. Outputs

The last part of the MSProfileR tool consists of documenting the whole done process of spectra analysis on the one hand by the report, the list of figures and of generating some modules serving to construct a future analysis with the application on the other hand, which are the parameters serving as a model to process a new dataset at the same previous factors, the intensity matrix serving to the reference database construction and the HDF5 file serving to store the averaged spectra, the annotation as well as the intensity matrix to perform the statistical analysis and the machine learning (Figure 21).

To export the data in this module just click on the download buttons.

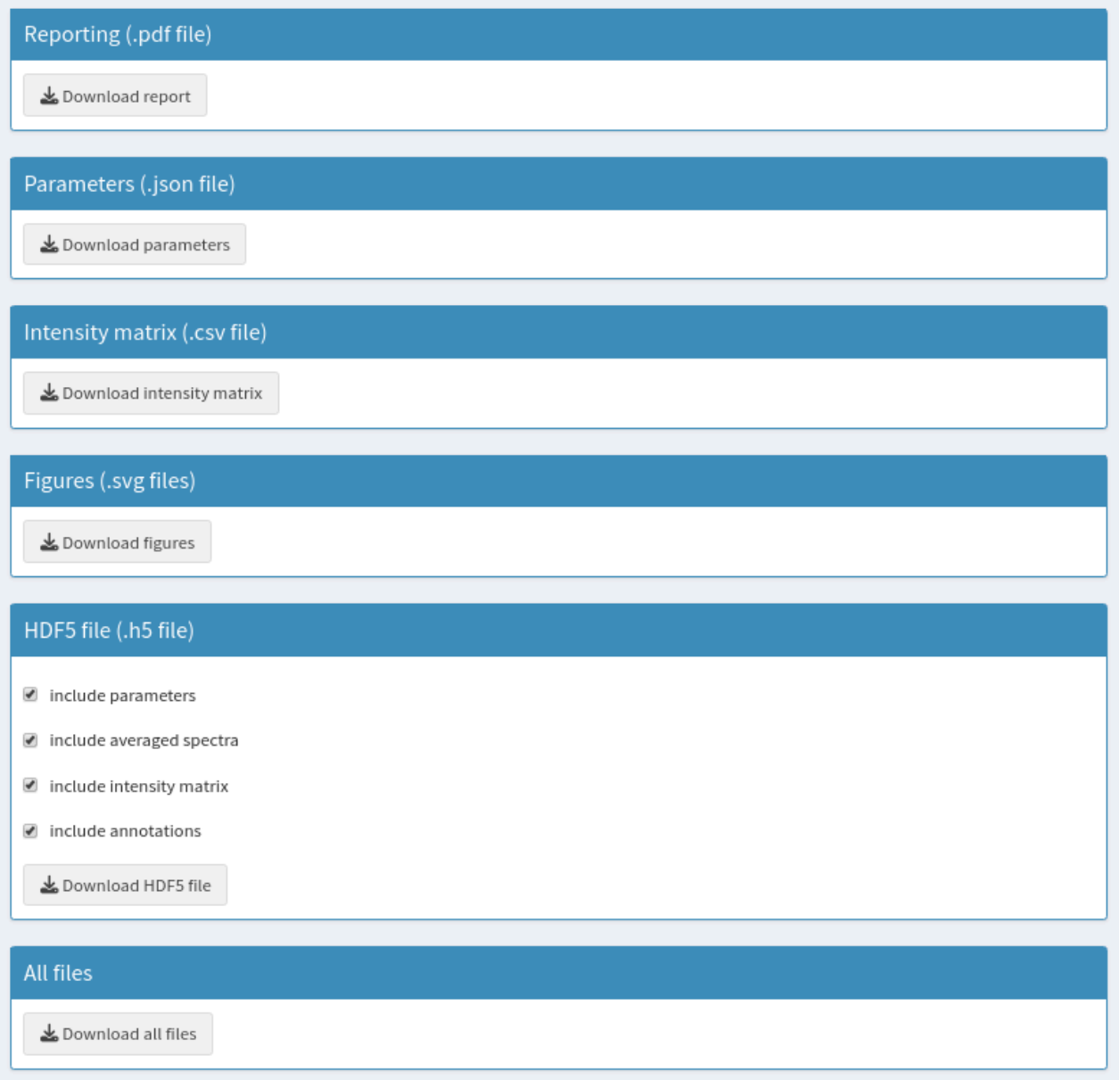

Figure 21. *Screen capture of the outputs page*