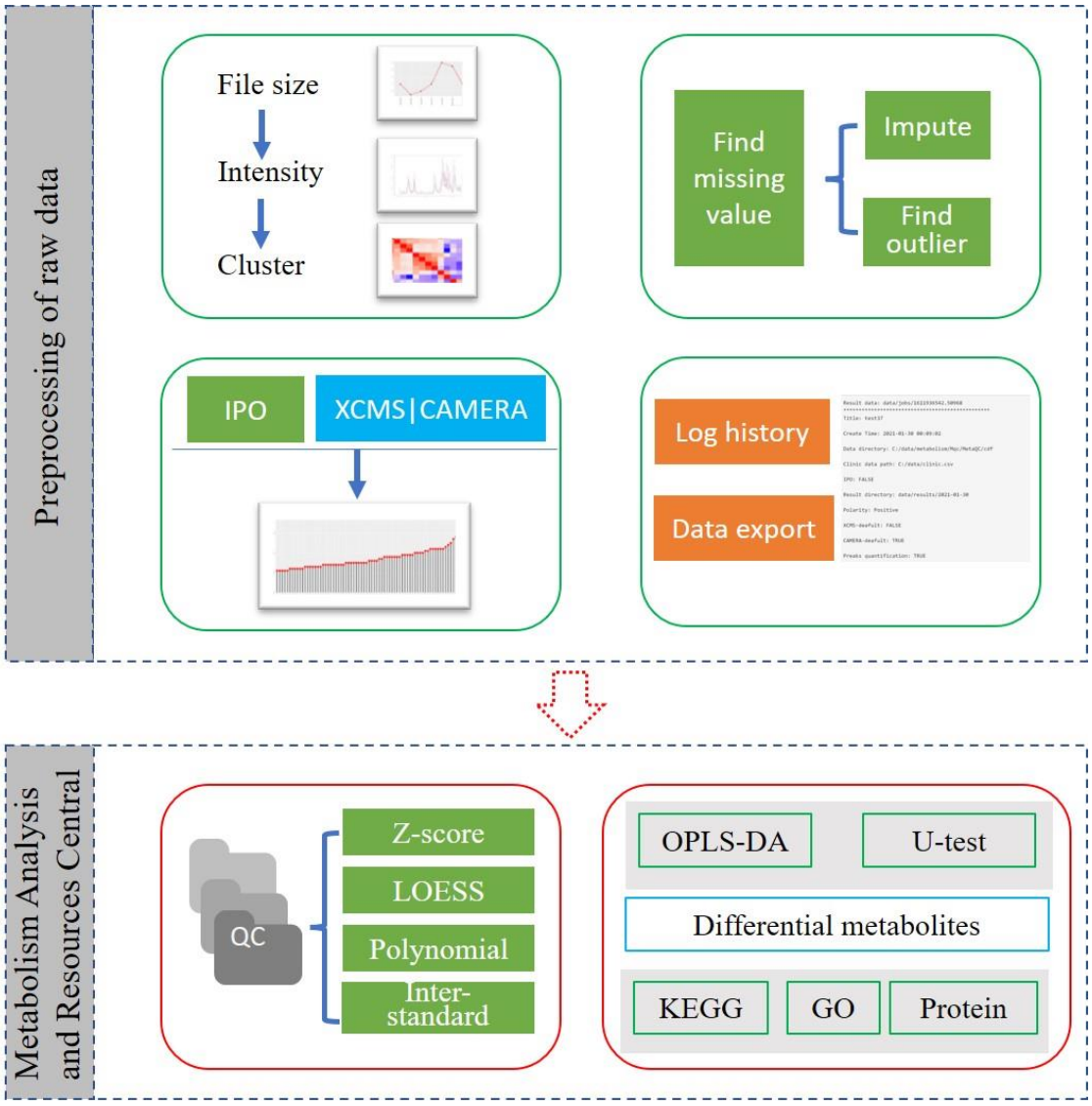
OpenNAU: An open-source platform for Normalizing, Analyzing， and visualizing Untargeted metabolomics data V1.0.0



1. LC-MS Peak Annotation and Identification with MetaQC

Description

This software has four modules, including Overview, Find peaks, Data cleaning, and All jobs (Figure 1)。

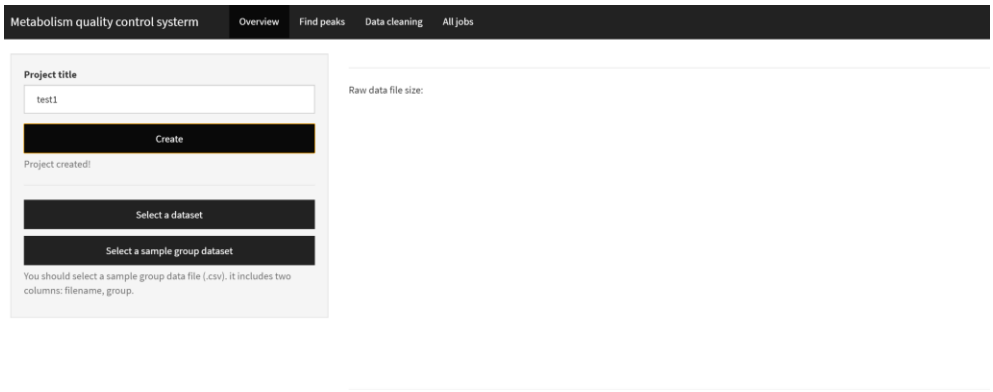


Figure 1. The index page for MetaQC

1. Overview

1) Creating a title for your jobs

2) Select the dataset for analysis. In figure 2, the user can choose the corresponding directory. Of course, if the user uploads MetaQC to the cloud server, the user should upload a dataset to the server and then select the dataset for analysis. All samples should be divided into multigroup based on batch information or custom by users.

Notes (This note determines the later data matching and transformation):

Data files should be named to avoid the following situations:

1) The file name cannot contain "-". Eg: "CRC-01.xml" can be changed to "CRC_01.xml", "CRC01.xml" or "CRC.01.xml".

2) The file or folder name must contain letters, preferably starting with a letter. Eg: "11.xml" should be changed to "X11.xml" or "11X.xml".

3) The file name should be consistent with the sample name in the clinical data (clinic.csv). Eg: "CRC_01.xml" should consist with sample name ("CRC_01") in clinical data.

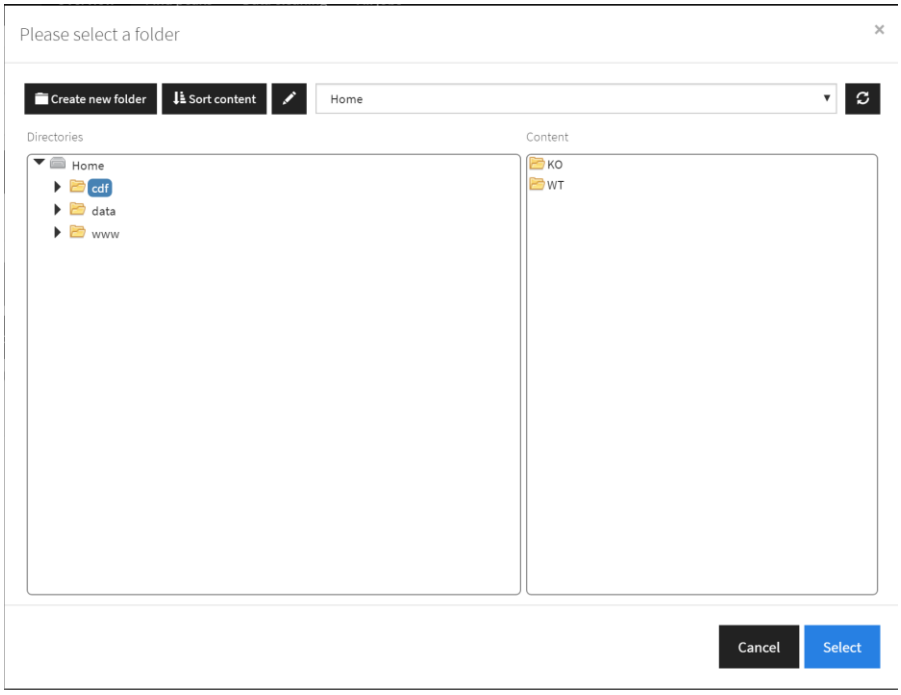


Figure 2. The page for selecting a dataset

3) Showing the file size for raw data. The line plot (Figure 3) will be shown on the right side of this page. In this figure, the user can check the file size for every sample.

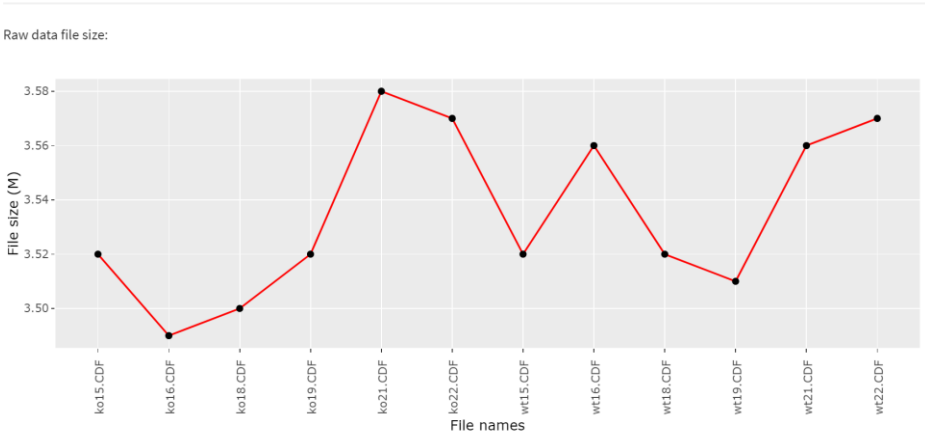


Figure 3. line plot for raw data size

4) Selecting the extended information for a dataset. This file includes two columns (sample name, sample group) and will be used for the sample cluster by heatmap in figure 4.

From Figure 4, we can assess whether the peak intensity distribution of the two groups of samples is consistent.

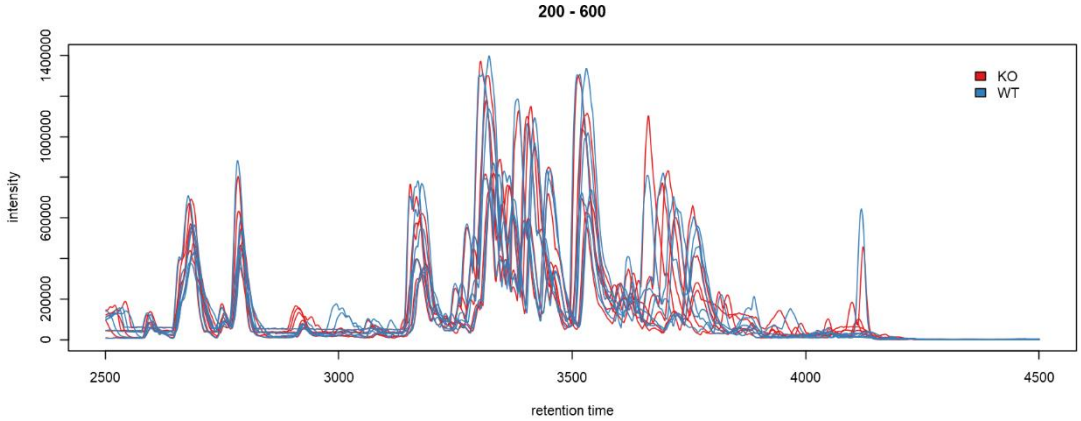


Figure 4. The intensity distribution for peaks

We compute the correlation for peaks and get the R-value for Pearson correlation analysis. Then we show the heatmap for all samples in Figure 5. From figure 5, we can find the batch effect for all samples. If only one or more batches are clustered together, there are differences between batches. If there is no significant clustering between all batches, the sample is not affected by batches.

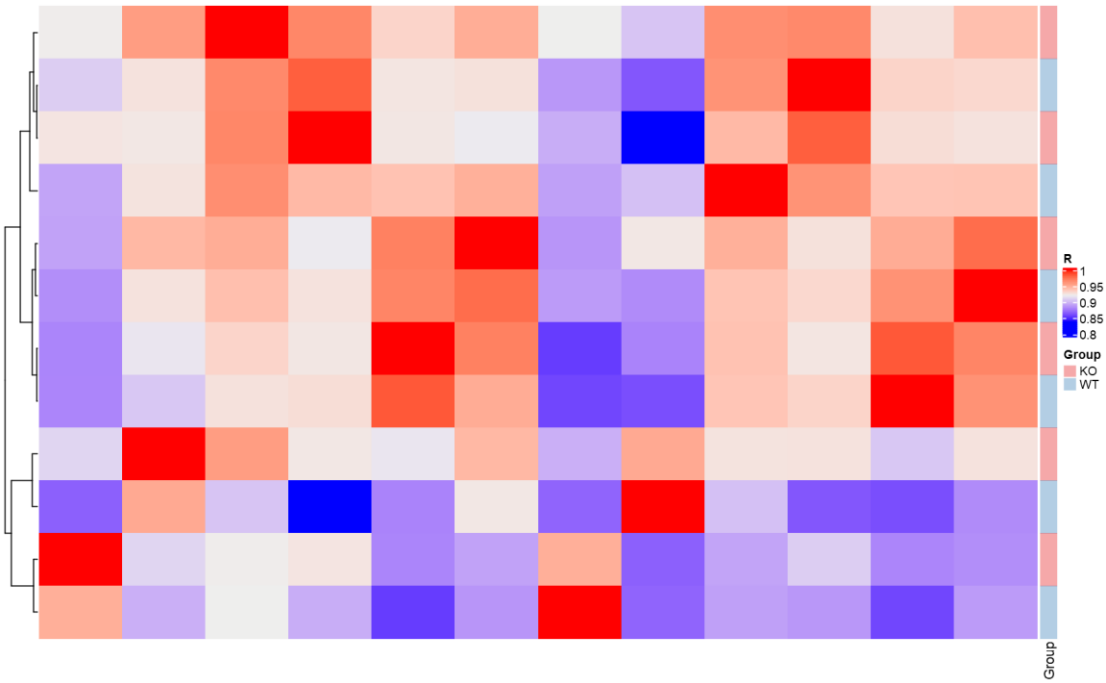


Figure 5. The heatmap for intensity value of all samples. Group replaces the data grouping or batch information.

2. Find peaks

In Figure 5, it shows the page detail for the “Finding peaks” module. In this module, the user can choose differential analysis methods for computing the intensity value.

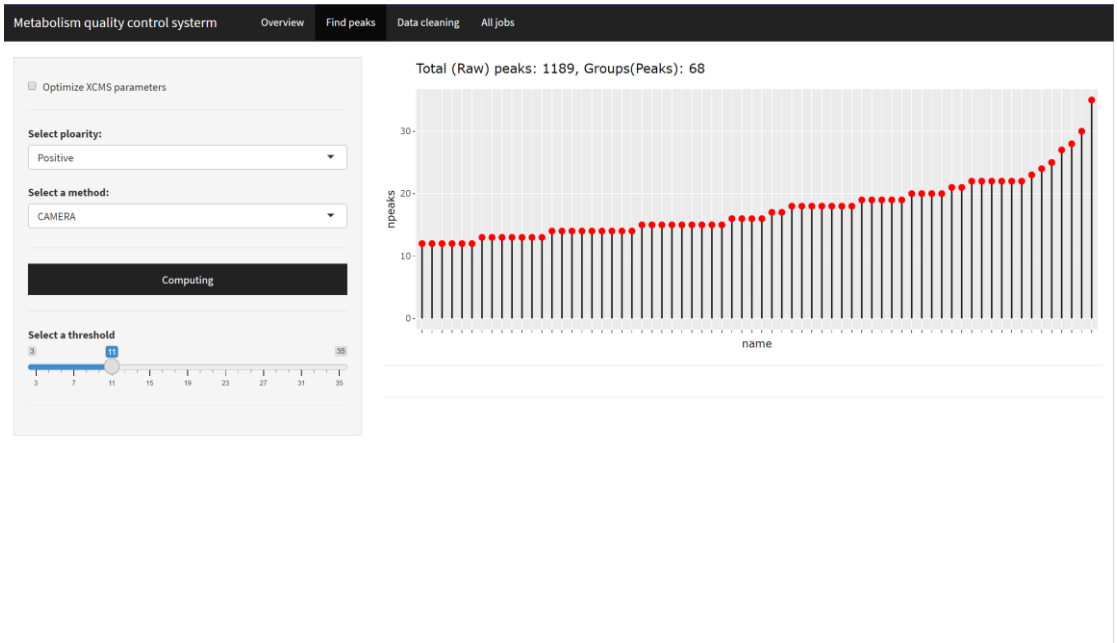


Figure 6. application page of the ‘Find peaks’ module

1) Selecting the analysis methods. In this section, the user can use XCMS or CAERMA software default parameters to analyze the LC-MS data. Users also can use IPO software to optimize the XCMS parameters and then compute the intensity value for every sample (Figure 6).

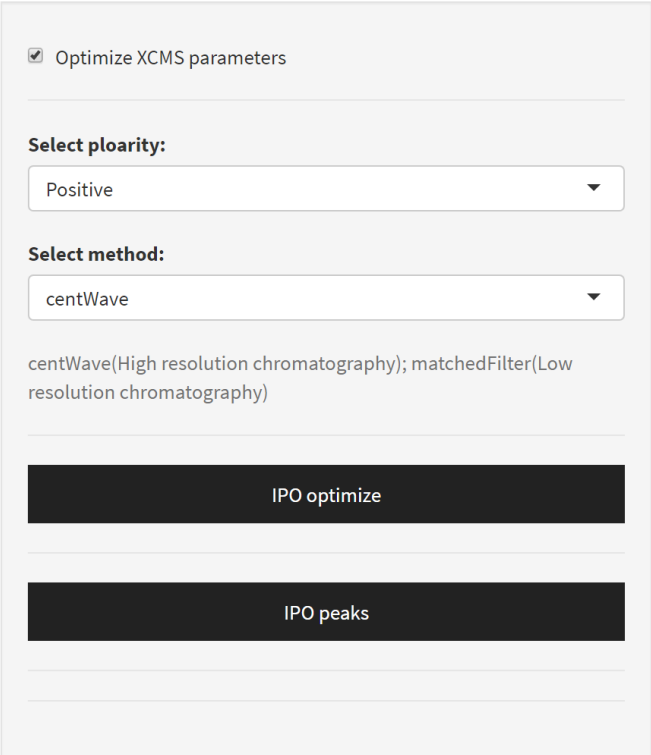


Figure 7. The menu for optimizing XMCS parameters

2) Showing the peaks for every group. In Figure 7, it shows the total peaks and the number of peaks in every group.

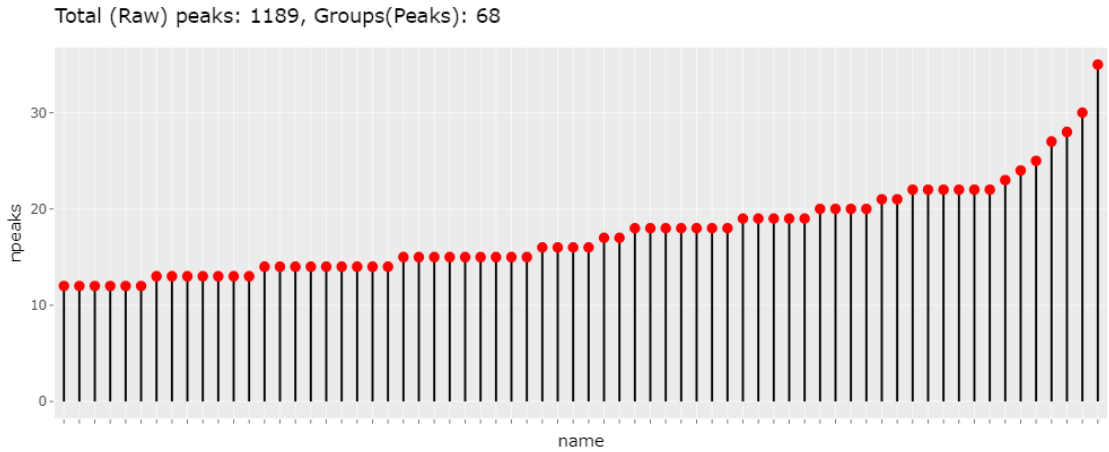


Figure 8. The number of peaks in every group

3. Data cleaning

In this part, the user can clean the dataset by missing samples or peaks value (detail panel in Figure 8).

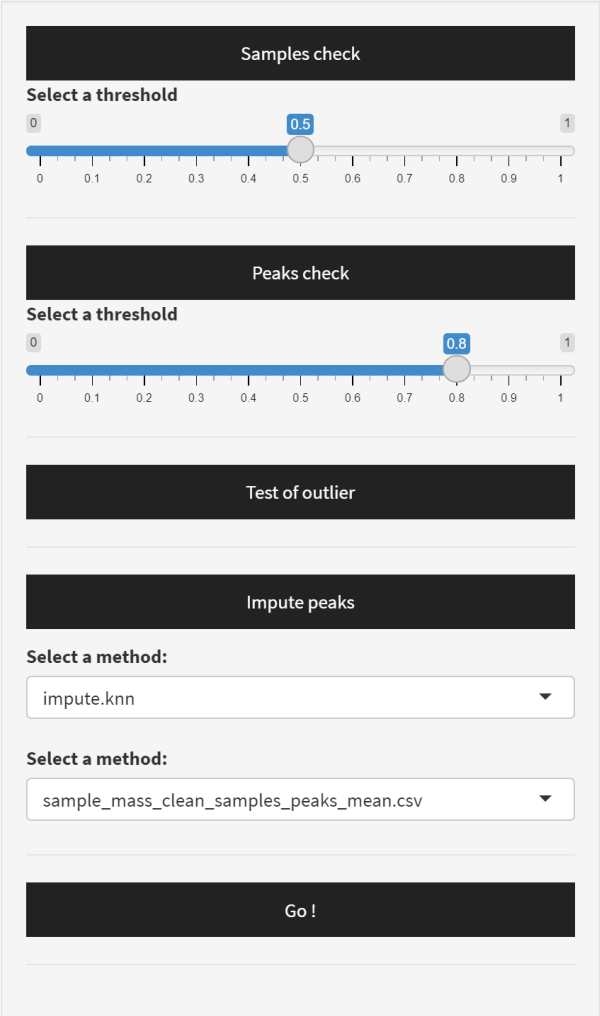


Figure 9. operation panel for data cleaning

1) Cleaning based on samples. In this section, we set the default threshold to 0.5, which means one sample should have more than 50% of total peaks that are not NA or 0 (Figure 9). Of course, if all samples show in the bar plot, it means that all samples satisfied the threshold.

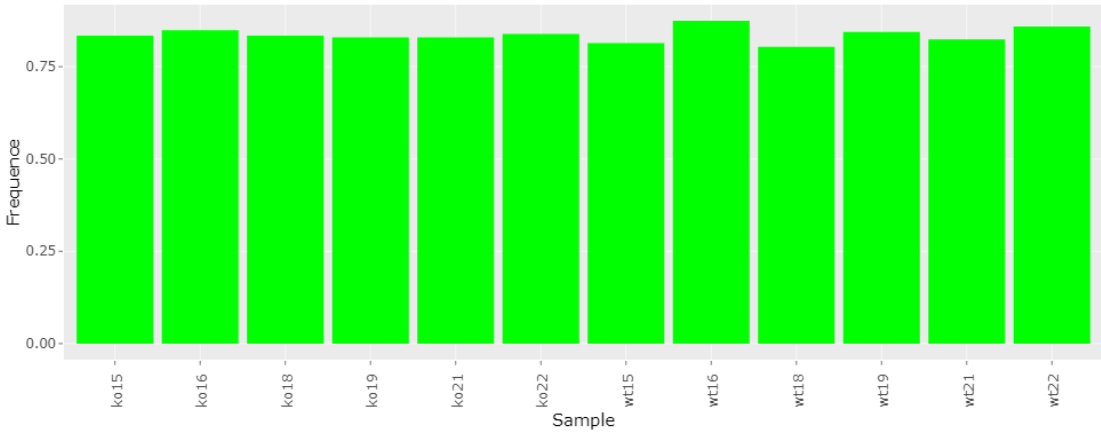


Figure 10. bar plot for cleaning results based on samples

2) Cleaning data by peaks. The default threshold is 0.8 in this section. If all peaks satisfied the threshold, we will show all peaks. The detailed results showed in Figure 10.

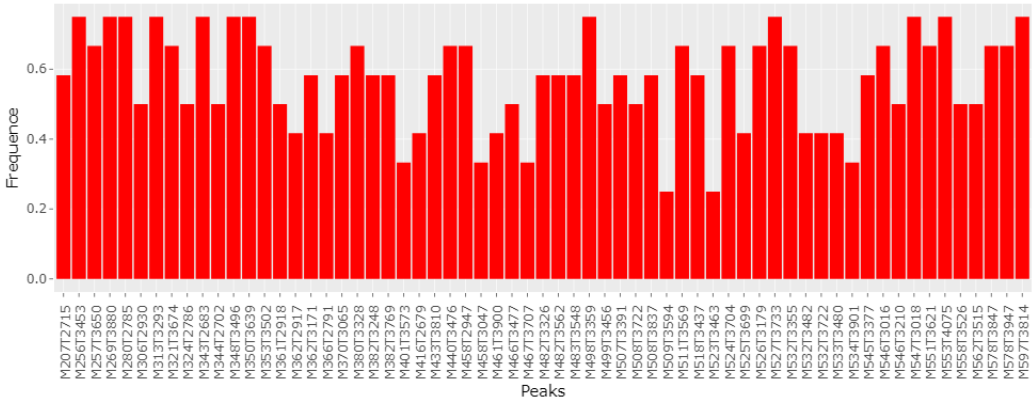


Figure 11. bar plot for results

3) Finding outlier of samples. In this section, the user can use Pcout{the Fast algorithm for identifying multivariate outliers in high-dimensional and/or large datasets, using the algorithm of Filzmoser, Maronna, and Werner (CSDA, 2007)} form package "mvoutlier" to find outlier samples. The detailed results showed in Figure 11.

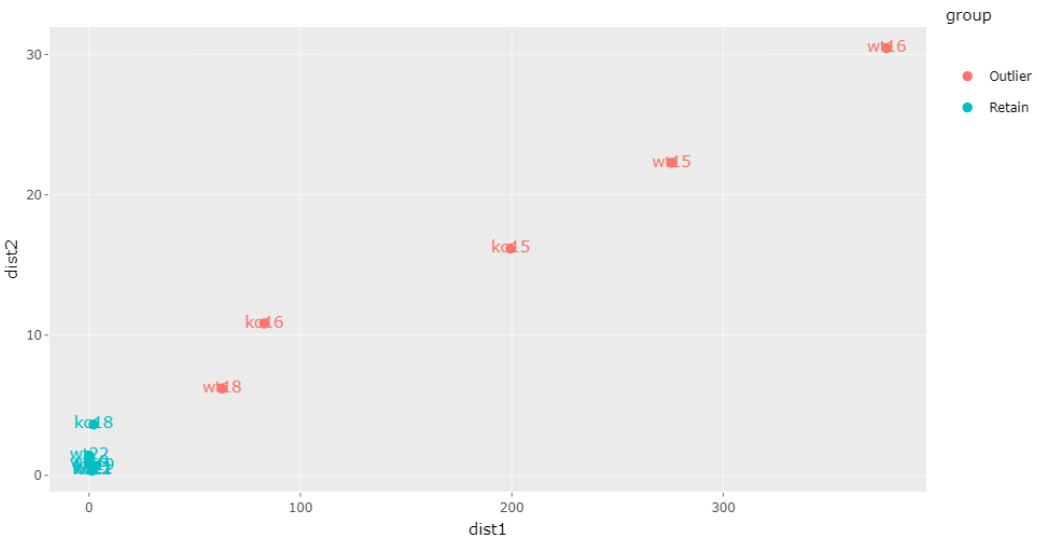


Figure 12. scatter plot for outlier results

4) Imputing the missing value of intensity. In this section, we give the Mean value and SD value to evaluate the imputing results.

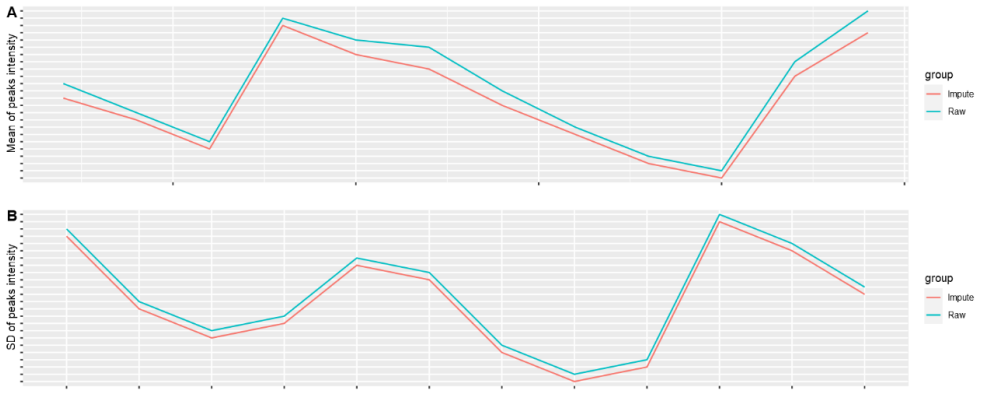


Figure 13. line plot for assessment of results

5) Data download. Users can download the result data at the bottom of the page.

Download imputed results | Download peaks information

4. All jobs

In this module, the users can browse job logs of all analysis tasks by date and transform the data structure to apply to the MetaboAnalyst software. The detailed operation panel showed in Figure 13.

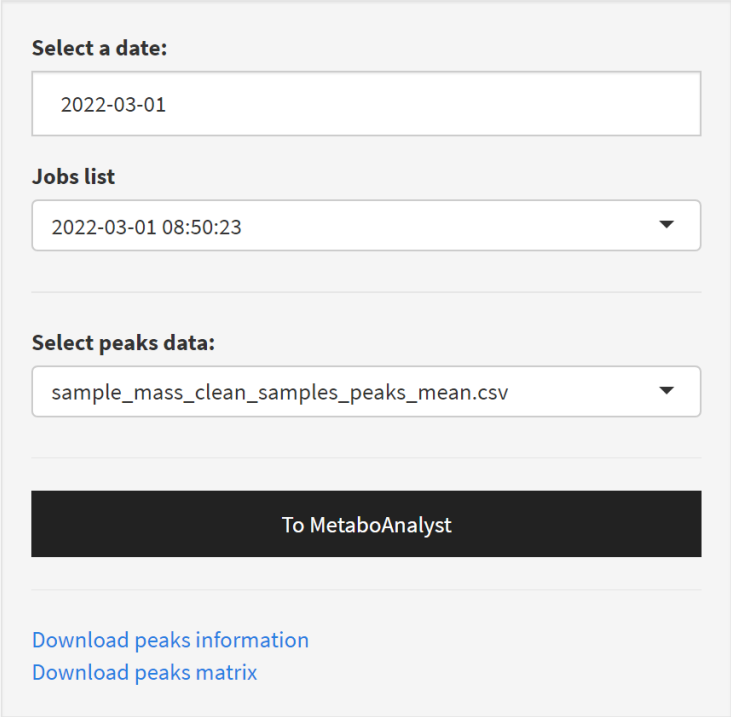
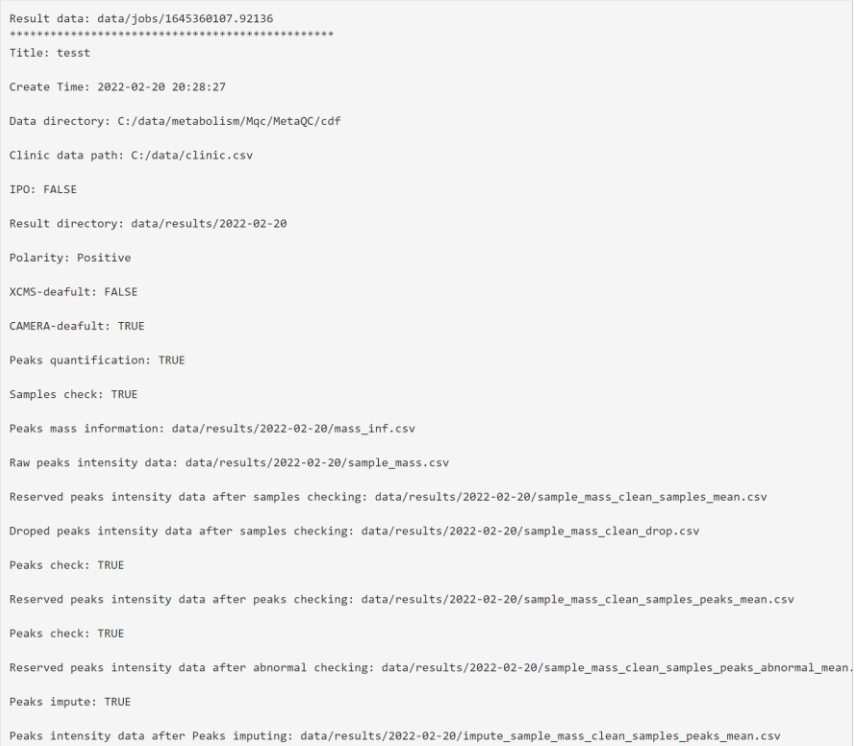


Figure 14. Operation panel

1) Browsing all jobs by date. This page will show the newest job log (Figure 14)



```
Result data: data/jobs/1645360107.92136
**************************************************
Title: tesst

Create Time: 2022-02-20 20:28:27

Data directory: C:/data/metabolism/Mqc/MetaQC/cdf

Clinic data path: C:/data/clinic.csv

IPO: FALSE

Result directory: data/results/2022-02-20

Polarity: Positive

XCMS-deafult: FALSE

CAMERA-deafult: TRUE

Peaks quantification: TRUE

Samples check: TRUE

Peaks mass information: data/results/2022-02-20/mass_inf.csv

Raw peaks intensity data: data/results/2022-02-20/sample_mass.csv

Reserved peaks intensity data after samples checking: data/results/2022-02-20/sample_mass_clean_samples_mean.csv

Droped peaks intensity data after samples checking: data/results/2022-02-20/sample_mass_clean_drop.csv

Peaks check: TRUE

Reserved peaks intensity data after peaks checking: data/results/2022-02-20/sample_mass_clean_samples_peaks_mean.csv

Peaks check: TRUE

Reserved peaks intensity data after abnormal checking: data/results/2022-02-20/sample_mass_clean_samples_peaks_abnormal_mean.

Peaks impute: TRUE

Peaks intensity data after Peaks imputing: data/results/2022-02-20/impute_sample_mass_clean_samples_peaks_mean.csv
```

Figure 15. The log for jobs

2） Transforming peaks data to MetaboAnalyst. In this section, we convert data according to the input requirements of the software (https://www.metaboanalyst.ca/MetaboAnalyst/upload/PeakUploadView.xhtml). See Figure 15 for specific parameter forms.

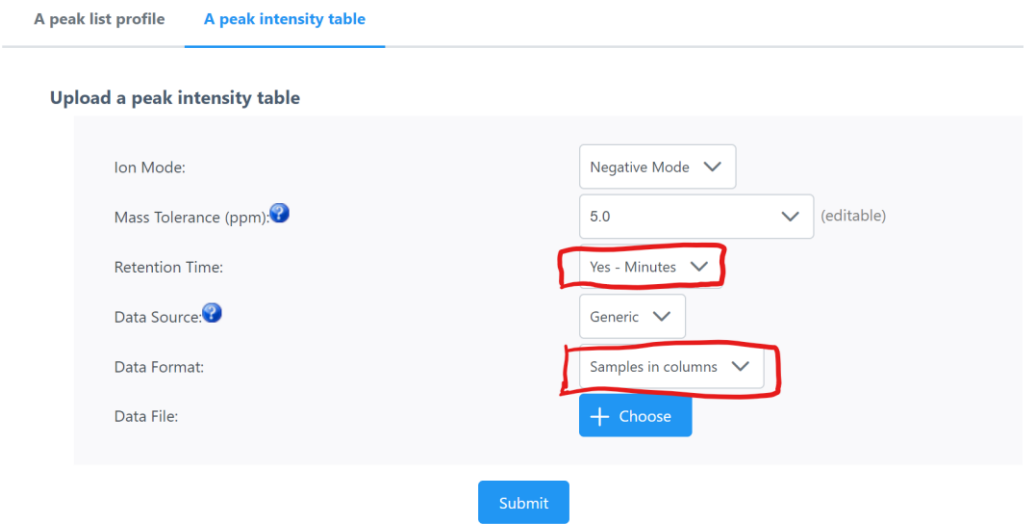


Figure 16. The upload data page of MetaboAnalyst

2. Metabolism Analysis and Resources Central (MARC)

Description

In this section, we constructed the main function modules: Search, Analysis, and Download. This section will be used to analyze the peak intensity data from MetaQC software or others by the data table standard of MARC. The running of MARC needs users to build a web cloud server for this software, and show the main page for users (Figure 16).

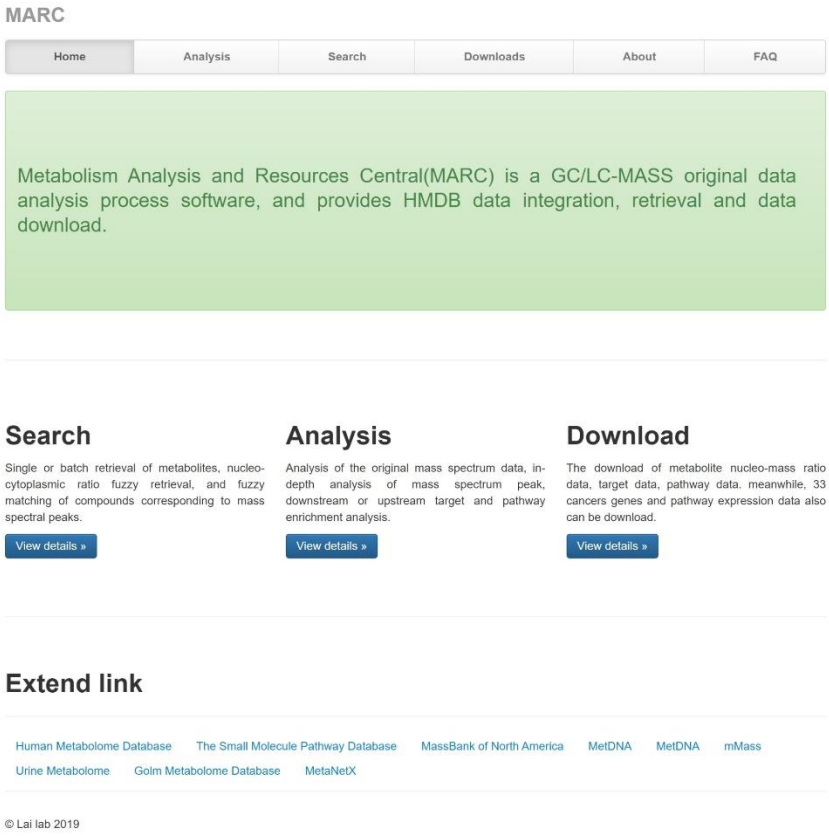


Figure 17. The main web page for MARC

1. Search function

In this module, the user can search the metabolites by m/z and Mass error (unit: ppm). The detailed information about the database is also shown on this page (Figure 17).

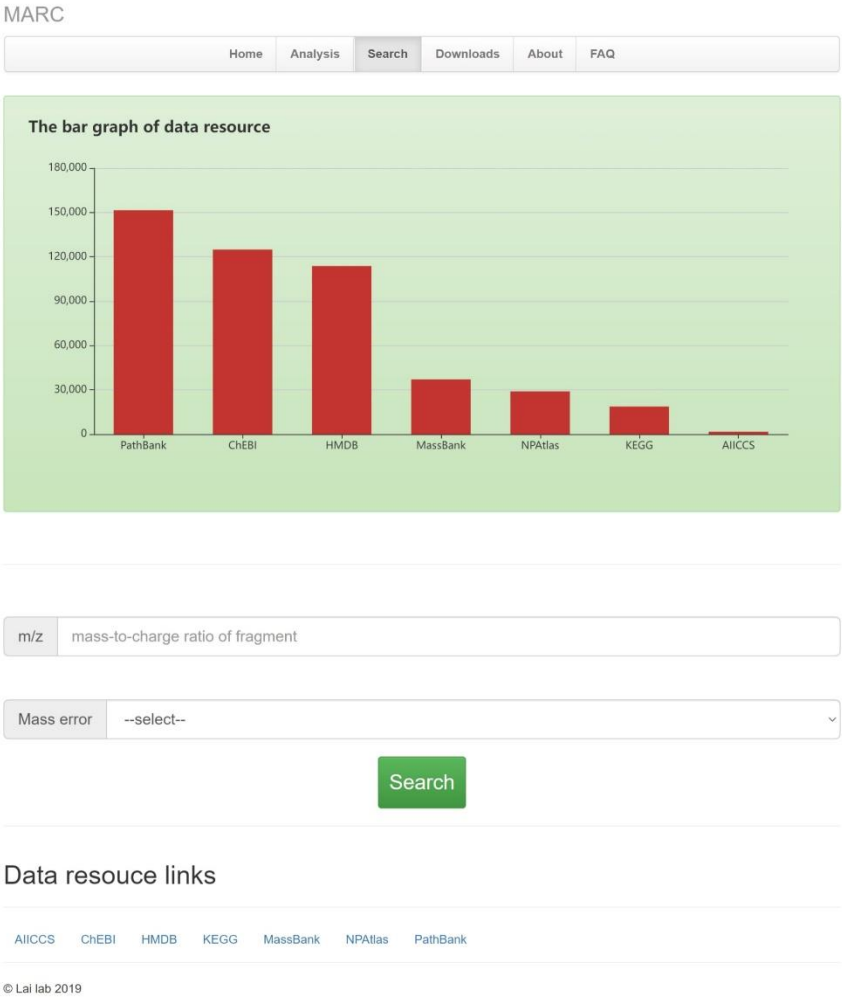


Figure 18. The search module of MARC

1) Showing the database resources by barplot. We collected the metabolites from the above 7 databases and integrated data into our database.

2) Searching metabolites. In this part, users can submit the m/z of target metabolite and mass error, then we will provide the results for searching (Figure 18). Of course, we set a test (300000ppm) for mass error and verify that the software is correctly deployed.

MARC

Home | Analysis | Search | Downloads | About | FAQ

Mass: 45.32122321, **Mass error:** 300000ppm

#ID	MID	CID	Source	Mass	HMDB/KEGG	Link
1	MRC00004313	CHEBI:15862	ChEBI	45.0837	HMDB0013231	Detail
2	MRC00004753	CHEBI:16397	ChEBI	45.04066	HMDB0001536	Detail
3	MRC00005369	CHEBI:17170	ChEBI	45.08372	HMDB0000087	Detail
4	MRC00019892	CHEBI:35468	ChEBI	45.04404	NA	Detail
5	MRC00021931	CHEBI:42241	ChEBI	45.0605	NA	Detail
6	MRC00022123	CHEBI:44730	ChEBI	45.0605	NA	Detail
7	MRC00022820	CHEBI:48431	ChEBI	45.04066	HMDB0001536	Detail
8	MRC00023520	CHEBI:50341	ChEBI	45.0605	NA	Detail
9	MRC00024473	CHEBI:52092	ChEBI	45.0605	NA	Detail
10	MRC00004313	CHEBI:15862	ChEBI	45.0837	HMDB0013231	Detail
11	MRC00004753	CHEBI:16397	ChEBI	45.04066	HMDB0001536	Detail
12	MRC00005369	CHEBI:17170	ChEBI	45.08372	HMDB0000087	Detail
13	MRC00019892	CHEBI:35468	ChEBI	45.04404	NA	Detail
14	MRC00021931	CHEBI:42241	ChEBI	45.0605	NA	Detail
15	MRC00022123	CHEBI:44730	ChEBI	45.0605	NA	Detail
16	MRC00022820	CHEBI:48431	ChEBI	45.04066	HMDB0001536	Detail
17	MRC00023520	CHEBI:50341	ChEBI	45.0605	NA	Detail
18	MRC00024473	CHEBI:52092	ChEBI	45.0605	NA	Detail
19	MRC00005369	HMDB0000087	HMDB	45.05784923	HMDB0000087	Detail
20	MRC00022820	HMDB0001536	HMDB	45.02146372	HMDB0001536	Detail
21	MRC00004313	HMDB0013231	HMDB	45.05784923	HMDB0013231	Detail

Total number: 21. Dowmload

Figure 19. The results from searching

2. Analysis module

In this module, we constructed a normalization and data analysis function. To refine the analysis process, we split it into four separate sections (Figure 19).

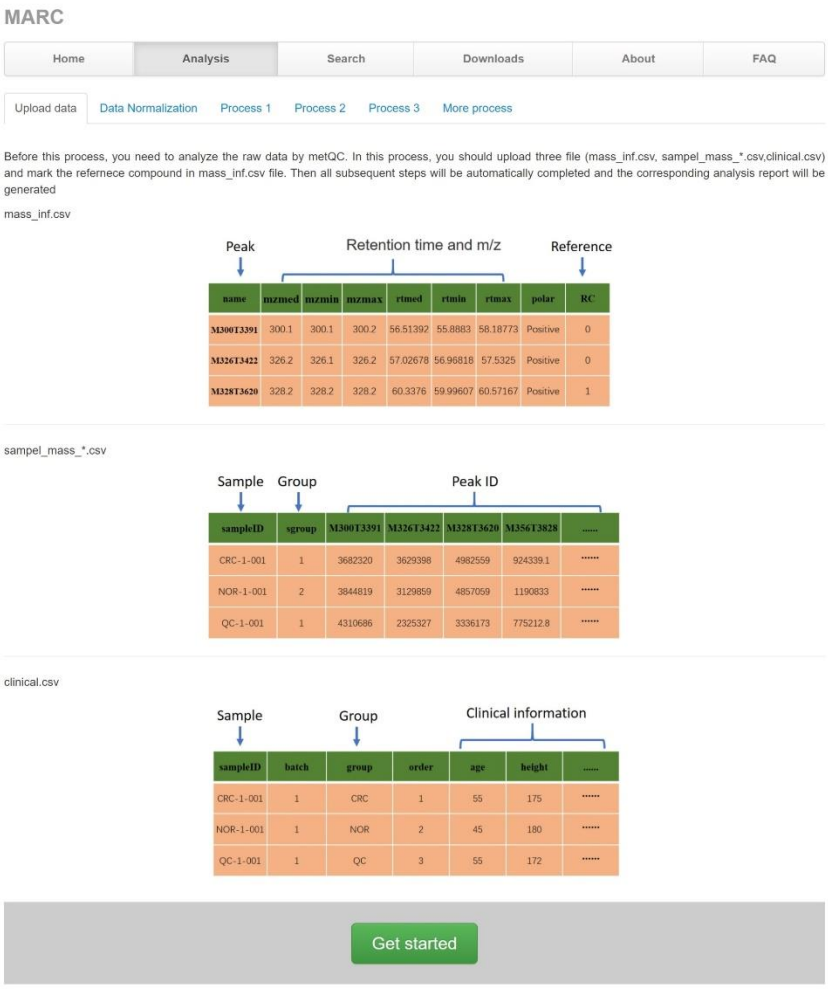


Figure 20. The analysis module of MARC

When users start to analyze the data, users should sign up for a job or browse the job detail (Figure 20).

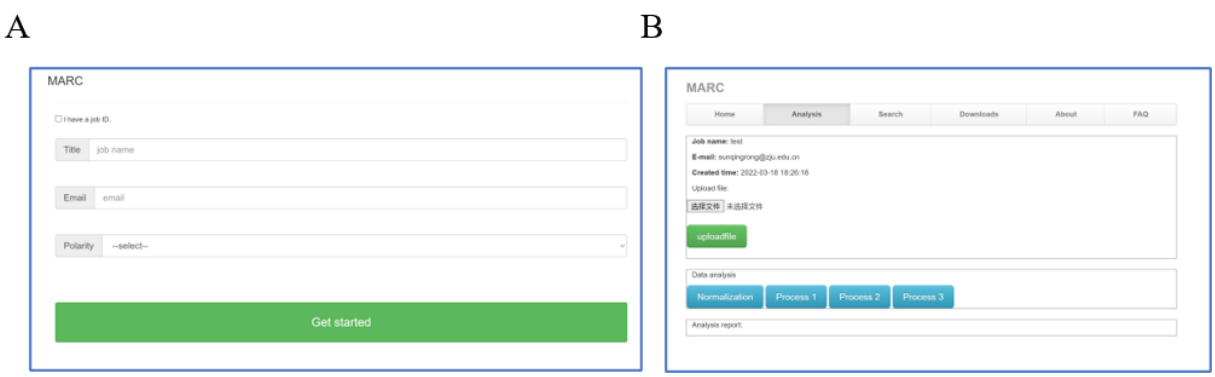


Figure 21. The page for job information of the analysis. A) sign up page, B) job detail page

1) Uploading data.

The data upload process requires three files for finishing all analysis, including mass_inf.csv （Figure 22A）, *_sample_mass_*.csv (Figure 22B), clinical.csv (Figure 22C).

Peak　　Retention time and m/z　　Reference

A.

name	mzmed	mzmin	mzmax	rtmed	rtmin	rtmax	polar	RC
M300T3391	300.1	300.1	300.2	56.51392	55.8883	58.18773	Positive	0
M326T3422	326.2	326.1	326.2	57.02678	56.96818	57.5325	Positive	0
M328T3620	328.2	328.2	328.2	60.3376	59.99607	60.57167	Positive	1

Sample　　Group　　Clinical information

B.

sampleID	batch	group	order	age	height	.....
CRC-1-001	1	CRC	1	55	175	.....
NOR-1-001	1	NOR	2	45	180	.....
QC-1-001	1	QC	3	55	172	.....

Sample　　Group　　Peak ID

C.

sampleID	sgroup	M300T3391	M326T3422	M328T3620	M356T3828	.....
CRC-1-001	CRC	3682320	3629398	4982559	924339.1	.....
NOR-1-001	NOR	3844819	3129859	4857059	1190833	.....
QC-1-001	QC	4310686	2325327	3336173	775212.8	.....

Figure 22. the files structure for uploading to MARC

2) Normalization for data.

3) Data analysis progress.

3. Download the module

This module is just for data download. In this part, the data includes metabolites and pathways information, as shown in Figure 21.

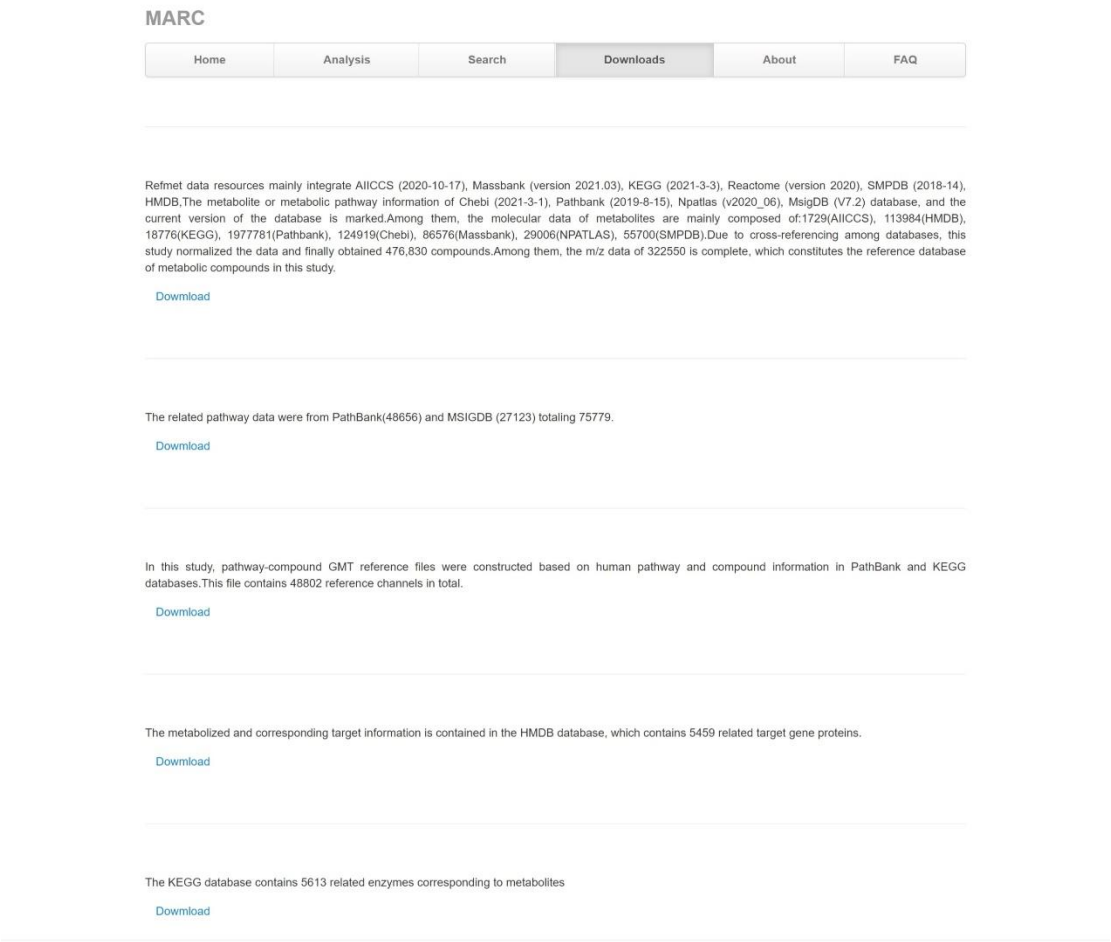


Figure 23. The download module of MARC

Appendix

Table 1. The data table description information of analysis results in MetaQC.

Filename	Description
*_output_xcms.tsv	Storing the peak data extracted from XCMS.
*_output_camera.csv	Storing the peaks data extracted from CAMERA.
IPO_optimize.R	This file saved the IPO optimized parameters list.
IPO_output_camera.csv	Storing the peak data extracted by CAMERA after IPO optimization was saved.
mass_inf.csv	Storing the basic information of ion peak. Includes m/z value, retention time and composition of prediction.
mass_inf_mean.csv	Storing the basic information of ion peak after filtering based on peak missing threshold.
sample_mass.csv	Storing ion peak abundance data. Each line is a sample, each column represents an ion peak, and the grouping information is contained in the second column.
sample_mass_clean_drop.csv	Storing deleted ion peak intensity data filtered based on sample missing thresholds.
sample_mass_clean_samples_mean.csv	Storing the reserved ion peak abundance data filtered based on sample missing thresholds.
sample_mass_clean_samples_peaks_mean.csv	Storing the reserved ion peak abundance data

	filtered based on peaks missing thresholds.
sample_mass_clean_samples_peaks_abnormal_mean.csv	Storing the ion peak abundance data after outlier value screening.
impute_sample_mass_*.csv	Storing the ion peaks data after imputing.
MetaboAnalyst_*.csv	Storing the data was transformed to input for MetaboAnalyst.