Tutorial of validating KGMN unknowns with repository mining

Zhiwei Zhou, Ph.D., 2022.06.08

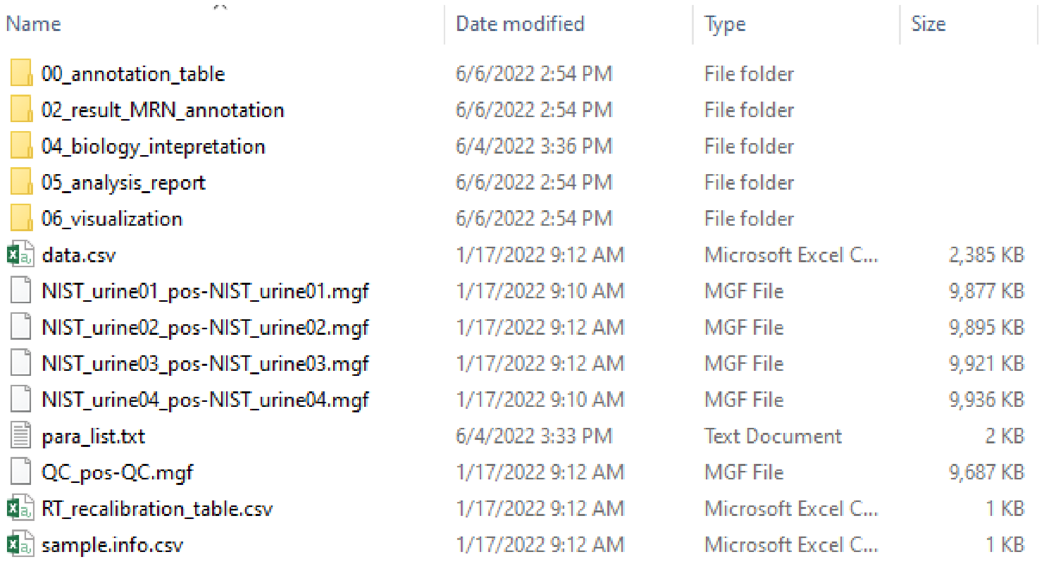
This tutorial aims to help users to select and validate their interesting unknown peaks from KGMN through repository mining. In the manuscript, we mainly used **MASST** to perform repository mining. The MASST[1] is a tool to query spectrum in context of where it occurs against all GNPS data sets. In this tutorial, we focus on demonstrating how to combine KGMN results and MASST. The detail instructions of MASST can be found in **GNPS document**.

The step-by-step instruction has been provided below.

1. Data preparing.

In this workflow, the data files require KGMN (MetDNA2) processed firstly. Here, we utilized NIST human urine data set as example. The data set has been analyzed with KGMN (v1.0.4), and the results can be downloaded **here**.

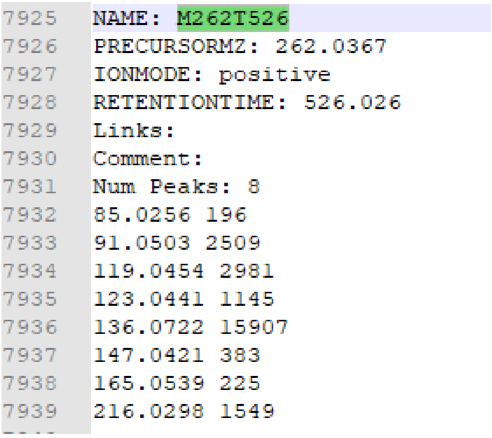
The folders should look like as below:



Name	Date modified	Type	Size
00_annotation_table	6/6/2022 2:54 PM	File folder	
02_result_MRN_annotation	6/6/2022 2:54 PM	File folder	
04_biology_intepretation	6/4/2022 3:36 PM	File folder	
05_analysis_report	6/6/2022 2:54 PM	File folder	
06_visualization	6/6/2022 2:54 PM	File folder	
data.csv	1/17/2022 9:12 AM	Microsoft Excel C...	2,385 KB
NIST_urine01_pos-NIST_urine01.mgf	1/17/2022 9:10 AM	MGF File	9,877 KB
NIST_urine02_pos-NIST_urine02.mgf	1/17/2022 9:12 AM	MGF File	9,895 KB
NIST_urine03_pos-NIST_urine03.mgf	1/17/2022 9:12 AM	MGF File	9,921 KB
NIST_urine04_pos-NIST_urine04.mgf	1/17/2022 9:10 AM	MGF File	9,936 KB
para_list.txt	6/4/2022 3:33 PM	Text Document	2 KB
QC_pos-QC.mgf	1/17/2022 9:12 AM	MGF File	9,687 KB
RT_recalibration_table.csv	1/17/2022 9:12 AM	Microsoft Excel C...	1 KB
sample.info.csv	1/17/2022 9:12 AM	Microsoft Excel C...	1 KB

The users can browser and select interesting known/unknown peaks in the **annotation table “table1_identification.csv”** in the “00_annotation_table” folder. It should be note that the selection of targeted peak is customized.

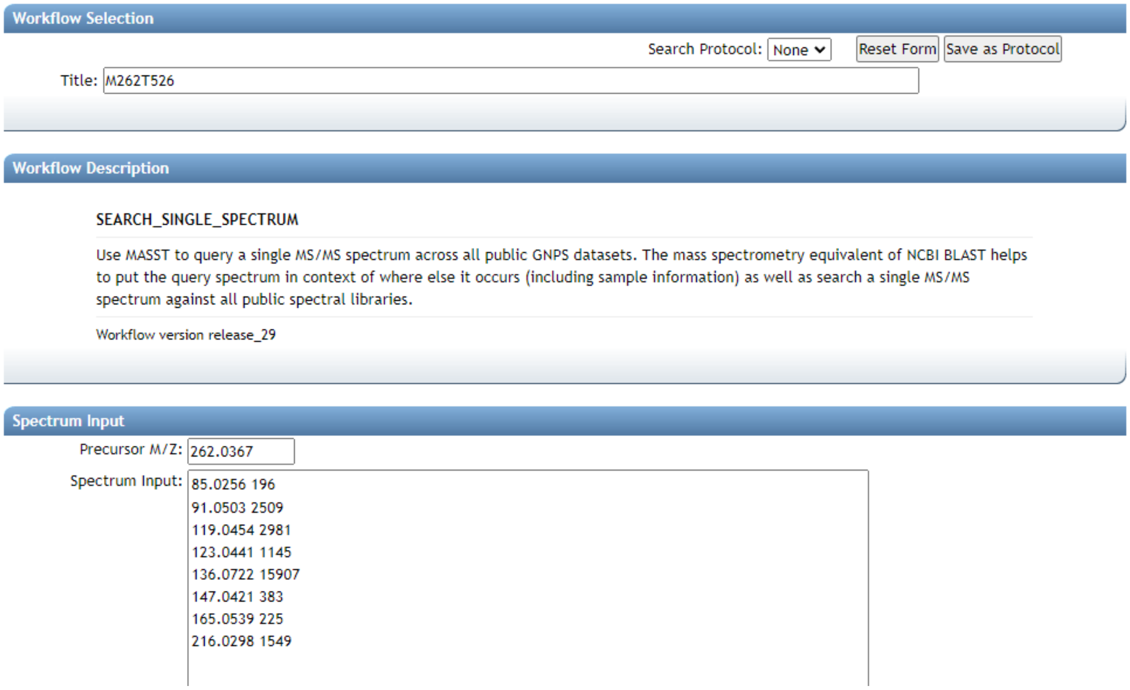
For demonstration, we utilized the unknown peak M262T526 as an example (Figure 5d in manuscript). The MS/MS spectrum of this peak can be found in the **“ms2_data.msp”** in “06_visualization” folder. You can open it with text tool (e.g. Notepad++).



```
NAME: M262T526
PRECURSORMZ: 262.0367
IONMODE: positive
RETENTIONTIME: 526.026
Links:
Comment:
Num Peaks: 8
85.0256 196
91.0503 2509
119.0454 2981
123.0441 1145
136.0722 15907
147.0421 383
165.0539 225
216.0298 1549
```

2. Upload and analysis in MASST.

Users can upload this file to MASST (https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp?redirect=auth) to perform repository mining. The users need to login first. Then, click the "**query spectrum**" button in MASST panel to start the analysis. Copy **related texts from MSP** file to "title", "precursor m/z", "spectrum input" panel in the web server, respectively.



Title: M262T526

SEARCH_SINGLE_SPECTRUM

Use MASST to query a single MS/MS spectrum across all public GNPS datasets. The mass spectrometry equivalent of NCBI BLAST helps to put the query spectrum in context of where else it occurs (including sample information) as well as search a single MS/MS spectrum against all public spectral libraries.

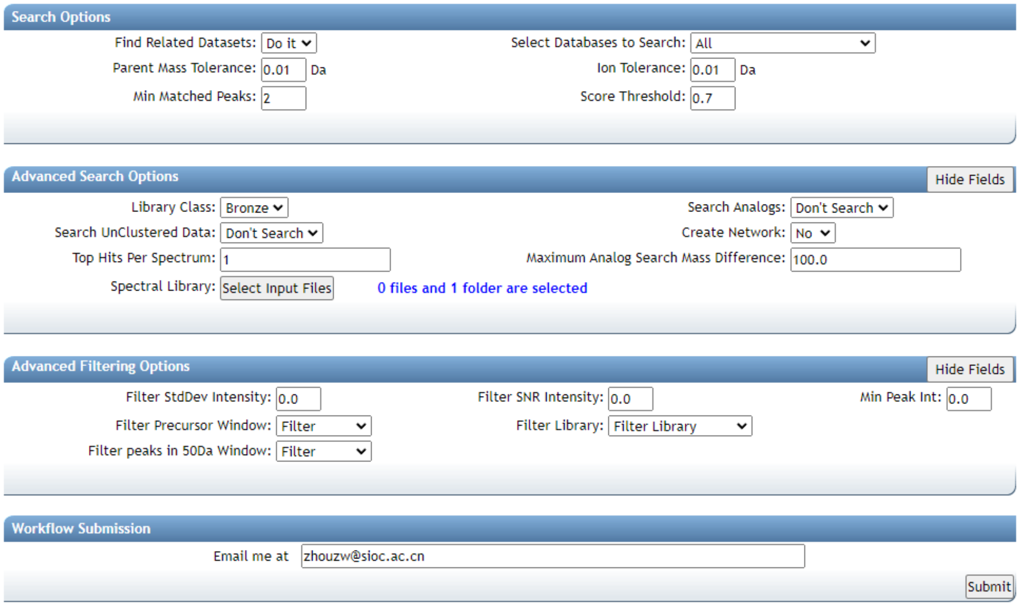
Workflow version release_29

Precursor M/Z: 262.0367

Spectrum Input:

```
85.0256 196
91.0503 2509
119.0454 2981
123.0441 1145
136.0722 15907
147.0421 383
165.0539 225
216.0298 1549
```

Modify the search parameters and click “submit” button. The **used parameters** in KGMN manuscript have been provided below.



Search Options

Find Related Datasets: Do it
Parent Mass Tolerance: 0.01 Da
Min Matched Peaks: 2
Select Databases to Search: All
Ion Tolerance: 0.01 Da
Score Threshold: 0.7

Advanced Search Options

Library Class: Bronze
Search UnClustered Data: Don't Search
Top Hits Per Spectrum: 1
Spectral Library: Select Input Files   0 files and 1 folder are selected
Search Analogs: Don't Search
Create Network: No
Maximum Analog Search Mass Difference: 100.0

Advanced Filtering Options

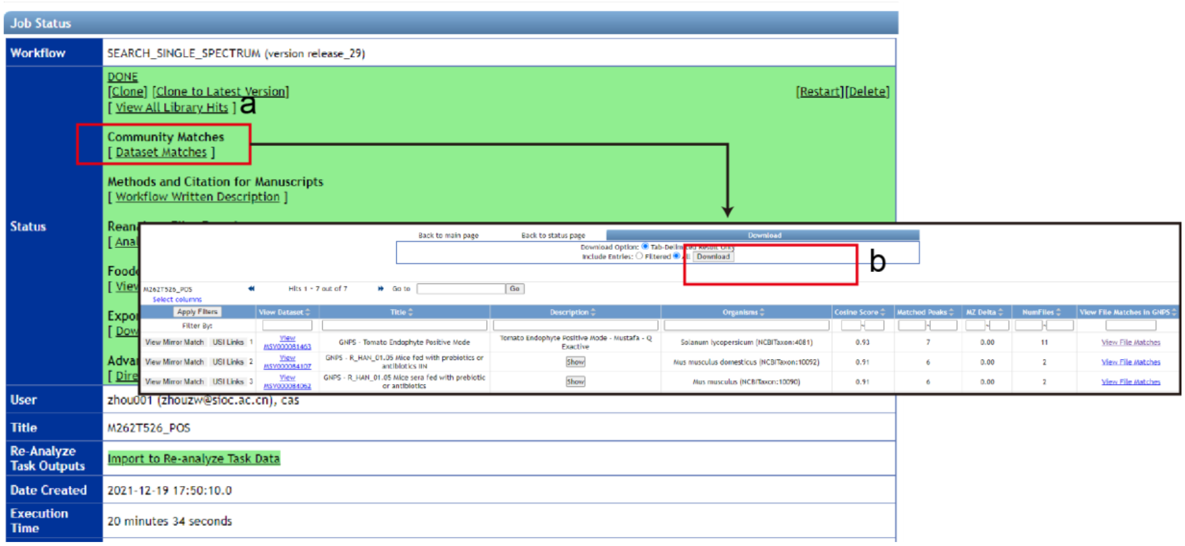
Filter StdDev Intensity: 0.0
Filter Precursor Window: Filter
Filter peaks in 50Da Window: Filter
Filter SNR Intensity: 0.0
Filter Library: Filter Library
Min Peak Int: 0.0

Workflow Submission

Email me at zhouzw@sioc.ac.cn

When the job finished, you will receive an email with a link. You can view and download results in the webserver.

- Matched data set: Dataset Matches → View File Matches → Download



Job Status

Workflow	SEARCH_SINGLE_SPECTRUM (version release_29)
Status	DONE [Clone] [Clone to Latest Version] [Restart][Delete] [View All Library Hits] Community Matches [Dataset Matches] Methods and Citation for Manuscripts [Workflow Written Description] Rean [Anal Food [View Expo [Dow Adva [Dire
User	zhou001 (zhouzw@sioc.ac.cn), cas
Title	M262T526_POS
Re-Analyze Task Outputs	Import to Re-analyze Task Data
Date Created	2021-12-19 17:50:10.0
Execution Time	20 minutes 34 seconds

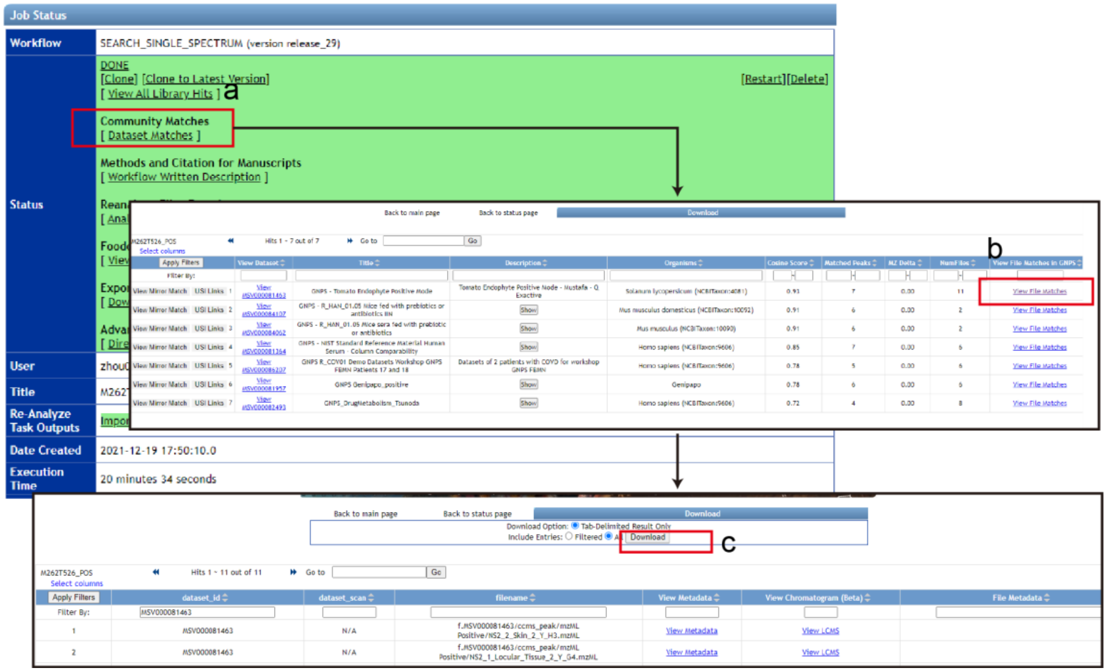
Back to main page   Back to status page

Download
Download Option: ● Tab-Delimited Result Only
Include Entries: ○ Filtered ● All Download

M262T526_POS   Hits 1 - 7 out of 7   Go to   Go
Select columns

Apply Filters		View Dataset	Title	Description	Organisms	Cosine Score	Matched Peaks	MZ Delta	NumFiles	View File Matches in GNPS
Filter By:										
View Mirror Match USI Links	1	View MSV000081463	GNPS - Tomato Endophyte Positive Mode	Tomato Endophyte Positive Mode - Mustafa - Q Exactive	Solanum lycopersicum (NCBITaxon:4081)	0.93	7	0.00	11	View File Matches
View Mirror Match USI Links	2	View MSV000084107	GNPS - R_HAN_01.05 Mice fed with prebiotics or antibiotics IIN	Show	Mus musculus domesticus (NCBITaxon:10092)	0.91	6	0.00	2	View File Matches
View Mirror Match USI Links	3	View MSV000084062	GNPS - R_HAN_01.05 Mice sera fed with prebiotic or antibiotics	Show	Mus musculus (NCBITaxon:10090)	0.91	6	0.00	2	View File Matches

- Matched files: Dataset Matches → View File Matches → Download



Job Status

Workflow	SEARCH_SINGLE_SPECTRUM (version release_29)
Status	DONE [Clone] [Clone to Latest Version] [Restart][Delete] [View All Library Hits] Community Matches [Dataset Matches] Methods and Citation for Manuscripts [Workflow Written Description] Rean [Anal Food [View Expo [Dow Adva [Dire
User	zhou
Title	M262
Re-Analyze Task Outputs	Impo
Date Created	2021-12-19 17:50:10.0
Execution Time	20 minutes 34 seconds

Back to main page Back to status page Download

M262T526_POS Hits 1 - 7 out of 7 Go to Go
Select columns

Apply Filters		View Dataset	Title	Description	Organisms	Cosine Score	Matched Peaks	MZ Delta	NumFiles	View File Matches in GNPS
Filter By:										
View Mirror Match USI Links	1	View MSV000081463	GNPS - Tomato Endophyte Positive Mode	Tomato Endophyte Positive Node - Mustafa - Q Exactive	Solanum lycopersicum (NCBITaxon:4081)	0.93	7	0.00	11	View File Matches
View Mirror Match USI Links	2	View MSV000084107	GNPS - R_HAN_01.05 Mice fed with prebiotics or antibiotics IIN	Show	Mus musculus domesticus (NCBITaxon:10092)	0.91	6	0.00	2	View File Matches
View Mirror Match USI Links	3	View MSV000084062	GNPS - R_HAN_01.05 Mice sera fed with prebiotic or antibiotics	Show	Mus musculus (NCBITaxon:10090)	0.91	6	0.00	2	View File Matches
View Mirror Match USI Links	4	View MSV000081364	GNPS - NIST Standard Reference Material Human Serum - Column Comparability	Show	Homo sapiens (NCBITaxon:9606)	0.85	7	0.00	6	View File Matches
View Mirror Match USI Links	5	View MSV000086207	GNPS R_COV01 Demo Datasets Workshop GNPS FBMN Patients 17 and 18	Datasets of 2 patients with COVD for workshop GNPS FBMN	Homo sapiens (NCBITaxon:9606)	0.78	5	0.00	6	View File Matches
View Mirror Match USI Links	6	View MSV000081957	GNPS Genipapo_positive	Show	Genipapo	0.78	6	0.00	6	View File Matches
View Mirror Match USI Links	7	View MSV000082493	GNPS_DrugMetabolism_Tsunoda	Show	Homo sapiens (NCBITaxon:9606)	0.72	4	0.00	8	View File Matches

Back to main page Back to status page Download
Download Option: ● Tab-Delimited Result Only
Include Entries: ○ Filtered ● All Download

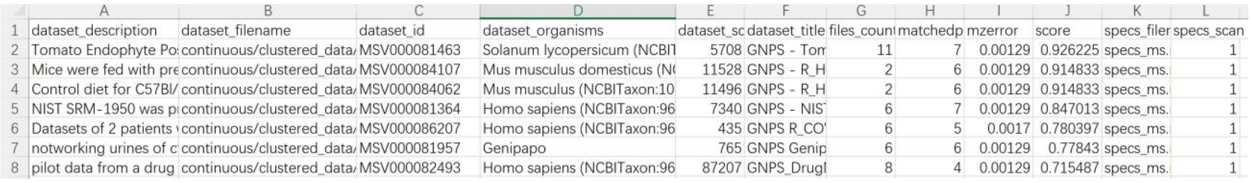
M262T526_POS Hits 1 - 11 out of 11 Go to Go
Select columns

Apply Filters	dataset_id	dataset_scan	filename	View Metadata	View Chromatogram (Beta)	File Metadata
Filter By:	MSV000081463					
1	MSV000081463	N/A	f.MSV000081463/ccms_peak/mzML Positive/NS2_2_Skin_2_Y_H3.mzML	View Metadata	View LCMS	
2	MSV000081463	N/A	f.MSV000081463/ccms_peak/mzML Positive/NS2_1_Locular_Tissue_2_Y_G4.mzML	View Metadata	View LCMS	

3. Result interpretation and visualization.

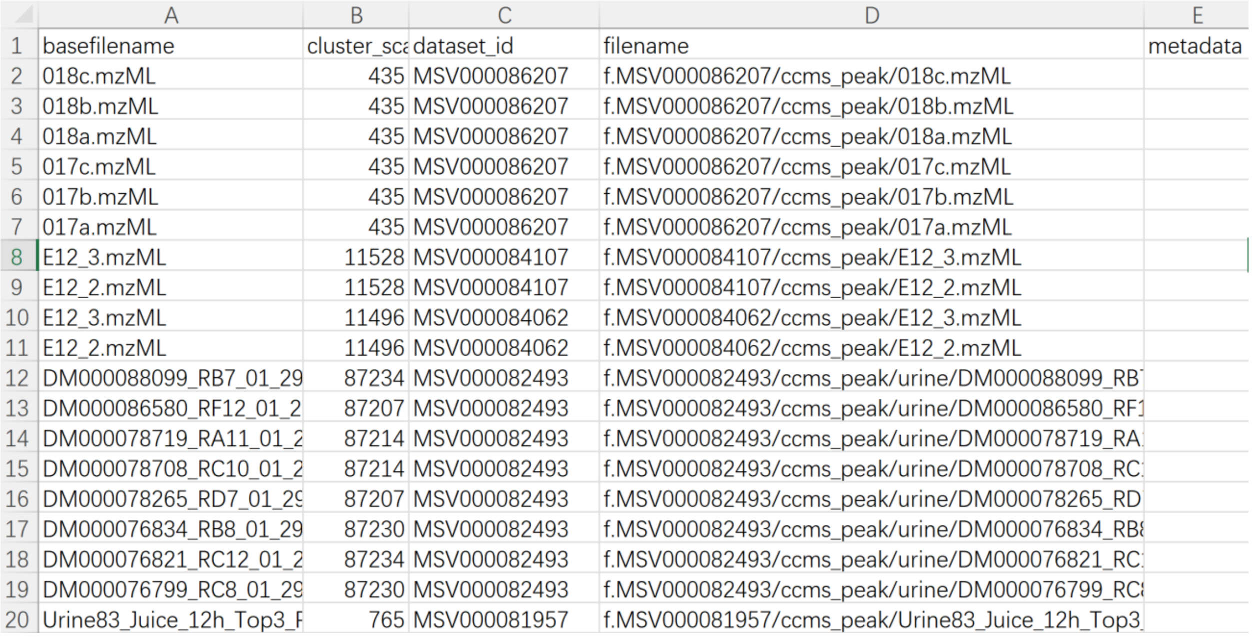
The downloaded results include 2 ZIP files, “view_all_datasets_matched.zip” and “view_all_file_datasets_matched.zip”. The files in packages can be further opened with Microsoft Office Excel or other program tools (e.g. R, Python).

- The table of “view_all_datasets_matched” contains meta information of appeared data sets, like “dataset description”, “dataset id”, “dataset organisms” and “files count”. Furthermore, we can conclude the species and sample information based on the dataset description. For our examples, it was appeared in 7 datasets, and 3 organisms (where genipapo is from human urine actually according to the data set description).



	A	B	C	D	E	F	G	H	I	J	K	L
1	dataset_description	dataset_filename	dataset_id	dataset_organisms	dataset_sc	dataset_title	files_count	matchedp	mzerror	score	specs_filer	specs_scan
2	Tomato Endophyte Po	continuous/clustered_data/	MSV000081463	Solanum lycopersicum (NCBIT	5708	GNPS - Tom	11	7	0.00129	0.926225	specs_ms.	1
3	Mice were fed with pre	continuous/clustered_data/	MSV000084107	Mus musculus domesticus (N(	11528	GNPS - R_H	2	6	0.00129	0.914833	specs_ms.	1
4	Control diet for C57Bl/	continuous/clustered_data/	MSV000084062	Mus musculus (NCBITaxon:10	11496	GNPS - R_H	2	6	0.00129	0.914833	specs_ms.	1
5	NIST SRM-1950 was p	continuous/clustered_data/	MSV000081364	Homo sapiens (NCBITaxon:96	7340	GNPS - NIS	6	7	0.00129	0.847013	specs_ms.	1
6	Datasets of 2 patients	continuous/clustered_data/	MSV000086207	Homo sapiens (NCBITaxon:96	435	GNPS R_CO	6	5	0.0017	0.780397	specs_ms.	1
7	notworking urines of c	continuous/clustered_data/	MSV000081957	Genipapo	765	GNPS Genip	6	6	0.00129	0.77843	specs_ms.	1
8	pilot data from a drug	continuous/clustered_data/	MSV000082493	Homo sapiens (NCBITaxon:96	87207	GNPS_DrugI	8	4	0.00129	0.715487	specs_ms.	1

- The table of "view_all_file_datasets_matched" contains names of matched files. Each file can be viewed online through the filename in GNPS dashboard (https://gnps-lcms.ucsd.edu/), while the files and dataset can be accessed in GNPS datasets (https://gnps.ucsd.edu/ProteoSAFe/datasets.jsp).



	A	B	C	D	E
1	basefilename	cluster_sca	dataset_id	filename	metadata
2	018c.mzML	435	MSV000086207	f.MSV000086207/ccms_peak/018c.mzML	
3	018b.mzML	435	MSV000086207	f.MSV000086207/ccms_peak/018b.mzML	
4	018a.mzML	435	MSV000086207	f.MSV000086207/ccms_peak/018a.mzML	
5	017c.mzML	435	MSV000086207	f.MSV000086207/ccms_peak/017c.mzML	
6	017b.mzML	435	MSV000086207	f.MSV000086207/ccms_peak/017b.mzML	
7	017a.mzML	435	MSV000086207	f.MSV000086207/ccms_peak/017a.mzML	
8	E12_3.mzML	11528	MSV000084107	f.MSV000084107/ccms_peak/E12_3.mzML	
9	E12_2.mzML	11528	MSV000084107	f.MSV000084107/ccms_peak/E12_2.mzML	
10	E12_3.mzML	11496	MSV000084062	f.MSV000084062/ccms_peak/E12_3.mzML	
11	E12_2.mzML	11496	MSV000084062	f.MSV000084062/ccms_peak/E12_2.mzML	
12	DM000088099_RB7_01_29	87234	MSV000082493	f.MSV000082493/ccms_peak/urine/DM000088099_RB	
13	DM000086580_RF12_01_2	87207	MSV000082493	f.MSV000082493/ccms_peak/urine/DM000086580_RF1	
14	DM000078719_RA11_01_2	87214	MSV000082493	f.MSV000082493/ccms_peak/urine/DM000078719_RA	
15	DM000078708_RC10_01_2	87214	MSV000082493	f.MSV000082493/ccms_peak/urine/DM000078708_RC	
16	DM000078265_RD7_01_2	87207	MSV000082493	f.MSV000082493/ccms_peak/urine/DM000078265_RD	
17	DM000076834_RB8_01_29	87230	MSV000082493	f.MSV000082493/ccms_peak/urine/DM000076834_RB	
18	DM000076821_RC12_01_2	87234	MSV000082493	f.MSV000082493/ccms_peak/urine/DM000076821_RC	
19	DM000076799_RC8_01_2	87230	MSV000082493	f.MSV000082493/ccms_peak/urine/DM000076799_RC	
20	Urine83_Juice_12h_Top3_	765	MSV000081957	f.MSV000081957/ccms_peak/Urine83_Juice_12h_Top3_	

With above information, it would be easy to reproduce figures of repository validation. The result of above example can be downloaded **here**.

Reference:

1. Wang, M., Jarmusch, A.K., Vargas, F. et al. Mass spectrometry searches using MASST. Nat Biotechnol 38, 23–26 (2020). https://doi.org/10.1038/s41587-019-0375-9