Protocol for Community-created Public MS/MS Reference Library Within the GNPS Infrastructure

Fernando Vargas[1,2,†], Kelly C. Weldon[1,4,†], Nicole Sikora[1,†], Mingxun Wang[1], Zheng Zhang[1], Emily C. Gentry[1], Morgan W. Panitchpakdi[1], Mauricio Caraballo[1], Pieter C. Dorrestein[1,3,4]*, and Alan K. Jarmusch[1]*

Affiliations

1. Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences and Collaborative Mass Spectrometry Innovation Center, University of California, San Diego, La Jolla, CA 92093, United States of America
2. Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093, United States of America
3. Department of Pediatrics, University of California, San Diego, La Jolla, CA 92093, United States of America
4. Center for Microbiome Innovation, University of California, San Diego, La Jolla, CA 92093, United States of America

† contributed equally for this project
* corresponding authors

Contributions

Software and Workflow - ZZ and MW
Instrument Methods - AKJ, FV
Data Generation - FV, KW
Protocol Documentation - MC, AKJ, FV, KW, NS
Testing and Evaluation - AKJ, ECG, FV, KW, MP, MW, NS
Manuscript - PCD, AKJ, FV, KW, NS
Concept - AKJ, FV, MW
Mentorship and Support - PCD

Abstract

Rationale

A major hurdle in identifying chemicals in mass spectrometry experiments is the availability of MS/MS reference spectra in public databases. Currently, scientists purchase databases or use public databases such as GNPS. The MSMS-Chooser workflow empowers the creation of MS/MS reference spectra directly in the GNPS infrastructure.

Methods

An MSMS-Chooser sample template was completed with the required information and sequence tables were generated programmatically. Standards in methanol-water (1:1) solution (1 µM) were placed into wells individually. An LC-MS/MS system using data-dependent acquisition in positive and negative modes was used. Species that may be generated under typical ESI conditions are chosen. The MS/MS spectra and MSMS-Chooser sample template were subsequently uploaded to MSMS-Chooser in GNPS for automatic MS/MS spectral annotation.

Results

Data acquisition quickly and effectively collected MS/MS spectra. MSMS-Chooser was able to accurately annotate 99.2% of the manually validated MS/MS scans that were generated from the chemical standards. The output of MSMS-Chooser includes a table ready for inclusion in the GNPS library (after inspection) as well as the ability to directly launch searches via MASST. Altogether, the data acquisition, processing, and upload to GNPS took ~2 hours for our proof-of-concept results.

Conclusions

The MSMS-Chooser workflow enables the rapid data acquisition, analysis, and annotation of chemical standards, and uploads the MS/MS spectra to community-driven GNPS. MSMS-Chooser democratizes the creation of MS/MS reference spectra in GNPS which will improve annotation and strengthen the tools which use the annotation information.

1. Introduction

Identifying the chemicals detected in untargeted metabolomics experiments is challenging. Mass spectrometry (MS) is one method by which untargeted metabolomics can be performed. MS detects the mass of ions, *i.e.* mass-to-charge (*m/z*), as well as their abundance. Tandem MS (*aka* MS/MS), provides information about the molecular structure of chemicals. MS/MS, specifically product ion scans, contain interpretable patterns of product ions from which structural information is derived. Manual interpretation of MS/MS spectra is impracticable given the large amount of data generated in untargeted metabolomics experiments.

One solution to the identification challenge is to match MS/MS spectra with that of MS/MS reference spectra to provide an annotation. MS/MS matches to reference library spectra are considered level 2 or level 3 annotations based on the Metabolomics Standards Initiative (MSI).[1] Reference libraries are time-consuming, costly to generate and often focused on a particular type of chemical class or the research interests of a laboratory or institute. Curatr, a web-based application for library generation, is one solution that improves the economy of effort.[2] The Global Natural Products Social Molecular Networking (GNPS) platform[3] offers a community-built MS/MS reference library which is part of the suite of libraries available for use, *e.g.* MassBank (Japan, European Union, and North America),[4,5] ReSpect,[6] MIADB, CASMI,[7] PNNL lipids,[8] Sirenas/Gates, and EMBL MCF. [2] While the option to contribute MS/MS spectra to the GNPS spectral library has existed since its inception, it remains a time-consuming process to select reference MS/MS and provide the required information for association with the entry (*e.g.* InChi, SMILES, CAS number, adduct, charge, and instrument).

To address the limitations in building public MS/MS spectral libraries, we have created a workflow directly into GNPS (dubbed MSMS-Chooser), borrowing concepts in Curatr to empower the GNPS community. The protocol details how to collect and contribute MS/MS spectra to the GNPS MS/MS reference library, including instructions for sample preparation information, a template for data entry, and web-enable data processing workflow in GNPS. To illustrate the utility of MSMS-Chooser, we generated MS/MS spectra from chemical standards in a 96-well plate format quickly, accurately, and with minimal manual interpretation.

2. Experimental

2.1 MSMS-Chooser Template

Chemical standards used to test the MSMS-Chooser workflow (**Figure 1A**) were purchased from Sigma-Aldrich (SKU: MSMLS-1EA). The MSMS-Chooser template (**Table S1**) was completed using the information provided with the purchased chemical standards. The MSMS-Chooser workflow uses the information in the MSMS-Chooser template to calculate the monoisotopic mass [M] and subsequently calculate the monoisotopic *m/z* of adducts. The current set of adducts considered are $[M+H]^+$, $[2M+H]^+$, $[M+Na]^+$, $[2M+Na]^+$, and $[M-(H_2O)_n+H]^+$, $[M-H]^-$, $[2M-H]^-$, $[2M-2H+Na]^-$; additional adducts can be added and retrospectively determined in data deposited in MassIVE, an MS data repository (massive.ucsd.edu). Additional information provided in the MSMS-Chooser template is used to automatically populate the fields required to upload MS/MS reference spectra to the GNPS library, such as the InChi and SMILES which are used to represent the chemical structure.

2.2 Sample Preparation and Data Acquisition

Chemical standards were resuspended to a final concentration of 1 µM using a resuspension solvent of methanol-water (1:1) and transferred to a 96-well plate (**Figure 1A**). An ultra-high performance liquid chromatography (UHPLC) system (Vanquish, Thermo) coupled to an orbitrap mass spectrometer (QExactive, Thermo) was used (**Figure 1B**). The MSMS-Chooser template automatically populates a sequence table compatible with the Q-Exactive (future updates will include additional instrument specific sequence table templates). The data acquisition time was set to 0.3 min for each injection. Separate positive and negative ionization mode methods were used (*i.e.* one injection in positive mode and one injection in negative mode); the major differences in the methods are the ionization source parameters. 1 µL was injected from each well into a flow of solvent. The flow rate was set to 0.500 mL min^{-1} with a composition of 0.1% formic acid in water and 0.1 % formic acid in methanol (1:1). More detailed information on the instrument method and method files can be found in the MSMS-Chooser documentation (https://ccms-ucsd.github.io/GNPSDocumentation/msmschooser/).

2.3 Data Processing and Upload

Thermo's proprietary MS file format (.raw) were converted to the open-source format (.mzXML), using the ProteoWizard tool MSConvert.[9] All MS files (.raw and .mzXML files) and the MSMS-Chooser template were uploaded to MassIVE. It is recommended that all MSMS-Chooser datasets uploaded to MassIVE be prepended with the following text, "GNPS - MSMS-Chooser -". Upon completion of the upload, the MassIVE accession was made public. Classyfire was used to generate chemical class information for the chemical standards by querying the SMILES.[10]

2.4 MSMS-Chooser

The MSMS-Chooser (v1.0) workflow, **Figure 1C**, is accessed on GNPS via the following link: https://gnps.ucsd.edu/ProteoSAFe/index.jsp?params=%7B%22workflow%22:%22MSMS-CHOOSER%22%7D with the source code available at https://github.com/CCMS-UCSD/GNPS_Workflows. The .mzXML/.mzML files (positive and negative mode) and the MSMS-Chooser template were selected into the "spectrum file" folder and "annotation table" folder respectively. Upon completion of the job, the resulting file output of MSMS-Chooser should be checked manually for accuracy, and poor quality or incorrect spectra should be deleted (row-wise) from the results table. The MSMS-Chooser output was then used to generate the associated MS/MS library spectra which were subsequently uploaded to the GNPS MS/MS spectral library (**Figure 1D**). Detailed instructions can be found in the MSMS-Chooser documentation (https://ccms-ucsd.github.io/GNPSDocumentation/msmschooser/). The MSMS-Chooser result page and resulting table are available via the following link https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=8454490b1ecc49ab85e1cece2f2f944c. A MASST search[11] was performed on an illustrative compound, biotin, from the MSMS-Chooser results page (**Figure S1**). The illustrative MASST job can be accessed using the following link: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=39cd886540c147e9a8a618b275e5f541.

2.5 Data Availability

QExactive tune files and instrument methods for positive and negative mode can be found on MassIVE via the following accession number: MSV000084286. The data (.raw and .mzXML) acquired for the Mass Spectrometry Metabolite Library (MSMLS) from Sigma-Aldrich and the associated MSMS-Chooser template can be found on MassIVE via the following accession number: MSV000084072. The MSMLS MSMS-Chooser job can be accessed via the following link:
https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=73da384ea02a4e8ca3edd82649e540c3.
The MSMLS library can be accessed via the following link: https://gnps.ucsd.edu/ProteoSAFe/gnpslibrary.jsp?library=GNPS-MSMLS.

3. Results and Discussion

Proof-of-concept results were obtained using the MSMS-Chooser workflow on a 96-well plate containing 87 chemical standards. Positive and negative mode MS/MS spectra were collected using two separate injections into a flow of solvent. The method was 0.3 min in length providing enough time for the chemical standard to be detected and the line purged of any remaining material, as depicted in the chromatograms of two illustrative chemicals in positive and negative mode (**Figure 2A** and **Figure 2D**, respectively). The flow injection method was used for throughput data acquisition (~2 hrs for a 96-well plate), opposed to performing chromatographic separation; however, any method for MS/MS data acquisition is compatible with the MSMS-Chooser workflow.

The data obtained were processed using the MSMS-Chooser workflow in GNPS to identify predesignated adducts calculated from the theoretical monoisotopic *m/z* within a mass error of 10 ppm. The resulting file from MSMS-Chooser indicates the scan number of the MS/MS spectrum and associated adduct for each chemical standard if detected. Multiple adducts were detected for some of the chemical standards which are formed during the electrospray ionization process under the experimental conditions (*e.g.* salinity, pH, and concentration). For example, the analysis of dethiobiotin in the positive ionization mode resulted in the detection of the protonated ($[M+H]^+$), sodiated ($[M+Na]^+$), and singly dehydrated species ($[M-H_2O+H]^+$), **Fig 2B**, and their corresponding MS/MS spectra selected by MSMS-Chooser (**Fig 2C**). Similarly, *N*-acetylleucine was detected as a deprotonated ($[M-H]^-$) ion and as a $[2M-2H+Na]^-$ ion in the negative ionization mode (**Fig 2D**), and MSMS-Chooser selected the corresponding MS/MS spectra (**Fig 2F**).

63 of the 87 chemical standards were detected upon manual inspection of the dataset. From the chemical standards detected, multiple hours of manual inspection found 73 MS/MS scans in positive mode and 53 MS/MS scans in negative mode (**Table 1**). MSMS-Chooser was able to correctly identify a total of 70 MS/MS scans in positive mode and 50 MS/MS scans in negative mode in minutes without user input. MSMS-Chooser correctly selected 99.2% of MS/MS spectra, excluding five MS/MS spectra not selected due to their mass error exceeding the 10 ppm tolerance. The adducts detected from this illustrative set of chemical standards were tabulated and plotted in **Figure 3A**. The predominant adducts in positive mode were $[M+H]^+$ and $[M+Na]^+$; and $[M-H]^-$ in negative mode. The high levels of sodiated adducts observed, **Figure 3B**, is likely related to the number of sugars (*i.e.* Organooxygen compounds) included in the test dataset.

The MSMS-Chooser output provides users the ability to query their MS/MS spectra against all publicly available tandem MS datasets in GNPS via the tool MASST. The MASST results will indicate the datasets in which the query spectra had been previously detected. **Figure S1** depicts the MASST result of biotin, one of the chemical standards run in the MSMS-Chooser proof of concept dataset. Biotin was shown to be found in 16 GNPS datasets, comprised of microbial, humans, and food sample types.

4. Conclusion

The MSMS-Chooser workflow provides users with a web-based platform and instrument protocols for the rapid and systematic generation of MS/MS reference spectra. In our proof-of-concept results, MSMS-Chooser correctly identified 120 MS/MS spectra from 87 chemical standards in the positive and negative mode in minutes and output a file ready for the GNPS MS/MS spectral library. MSMS-Chooser was 99.2% accurate in selecting MS/MS spectra compared to manual evaluation (inspection of results is recommended). One strength of the workflow is the consideration of all possible adducts. Not all chemicals have the same predilection to ionization via one adduct (*e.g.* protonation or deprotonation). MSMS-Chooser assists in building MS/MS spectral libraries by considering multiple adducts that may be encountered using different electrospray ionization conditions. Any method of generating MS/MS (specifically product ion spectra) via different collisional activation methods is compatible with the GNPS spectral library. The current protocols do not capture collision energy but it is possible to do so; however, such information is not currently utilized in GNPS and therefore has limited utility. The variability in MS/MS spectra is best captured via the contribution of spectra of the same chemical collected on multiple instruments and multiple conditions. Contribution of all MS/MS spectra is highly encouraged.

MSMS-Chooser will aid the community of scientists who use GNPS (or the associated GNPS knowledge base) by increasing the number of MS/MS reference spectra in the public domain, improving annotation rates in untargeted metabolomics experiments. We demonstrated the ability to readily determine the location in which the chemical standards MS/MS spectrum selected via MSMS-Chooser were found using MASST.[11] Any additional MS/MS spectra uploaded to the GNPS MS/MS spectral library will be included in the periodic living data analysis in GNPS. Data submitters will be emailed when new chemical annotations are found in their MassIVE deposited data. Further, additional MS/MS library spectra will extend the depth of annotation in ReDU,[12] a recently developed web-enabled tool to reuse public MS/MS data. Lastly, the providence of the original data and MSMS-Chooser template is retained in MassIVE and available for re-analysis in a systematic and automated fashion.

Acknowledgments

The authors would like to thank Julia M. Gauglitz, Ph.D. and Daniel Petras Ph.D. for insightful discussion. AKJ thanks the American Society for Mass Spectrometry for the Postdoctoral Career Development Award. The authors acknowledge funding by the Office of Naval Research Multidisciplinary University Research Initiative Award, (No. N00014-15-1-2809) and the Center for Microbiome Innovation. MC and PCD were supported by NSF grant IOS-1656475.

References

(1) Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W.-M.; Fiehn, O.; Goodacre, R.; Griffin, J. L.; et al. Proposed Minimum Reporting Standards for Chemical Analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **2007**, *3* (3), 211–221. https://doi.org/10.1007/s11306-007-0082-2.

(2) Palmer, A.; Phapale, P.; Fay, D.; Alexandrov, T. Curatr: A Web Application for Creating, Curating and Sharing a Mass Spectral Library. *Bioinformatics* **2018**, *34* (8), 1436–1438. https://doi.org/10.1093/bioinformatics/btx786.

(3) Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; et al. Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking. *Nature Biotechnology*. Nature Publishing Group September 8, 2016, pp 828–837. https://doi.org/10.1038/nbt.3597.

(4) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; et al. MassBank: A Public Repository for Sharing Mass Spectral Data for Life Sciences. *J. Mass Spectrom.* **2010**, *45* (7), 703–714. https://doi.org/10.1002/jms.1777.

(5) The MassBank consortium https://github.com/MassBank (accessed Nov 10, 2019).

(6) Sawada, Y.; Nakabayashi, R.; Yamada, Y.; Suzuki, M.; Sato, M.; Sakata, A.; Akiyama, K.; Sakurai, T.; Matsuda, F.; Aoki, T.; et al. RIKEN Tandem Mass Spectral Database (ReSpect) for Phytochemicals: A Plant-Specific MS/MS-Based Data Resource and Database. *Phytochemistry* **2012**, *82*, 38–45. https://doi.org/10.1016/j.phytochem.2012.07.007.

(7) Schymanski, E.; Neumann, S. The Critical Assessment of Small Molecule Identification (CASMI): Challenges and Solutions. *Metabolites* **2013**, *3* (3), 517–538. https://doi.org/10.3390/metabo3030517.

(8) Kyle, J. E.; Crowell, K. L.; Casey, C. P.; Fujimoto, G. M.; Kim, S.; Dautel, S. E.; Smith, R. D.; Payne, S. H.; Metz, T. O. LIQUID: An-Open Source Software for Identifying Lipids in LC-MS/MS-Based Lipidomics Data. *Bioinformatics* **2017**, *33* (11), 1744–1746. https://doi.org/10.1093/bioinformatics/btx046.

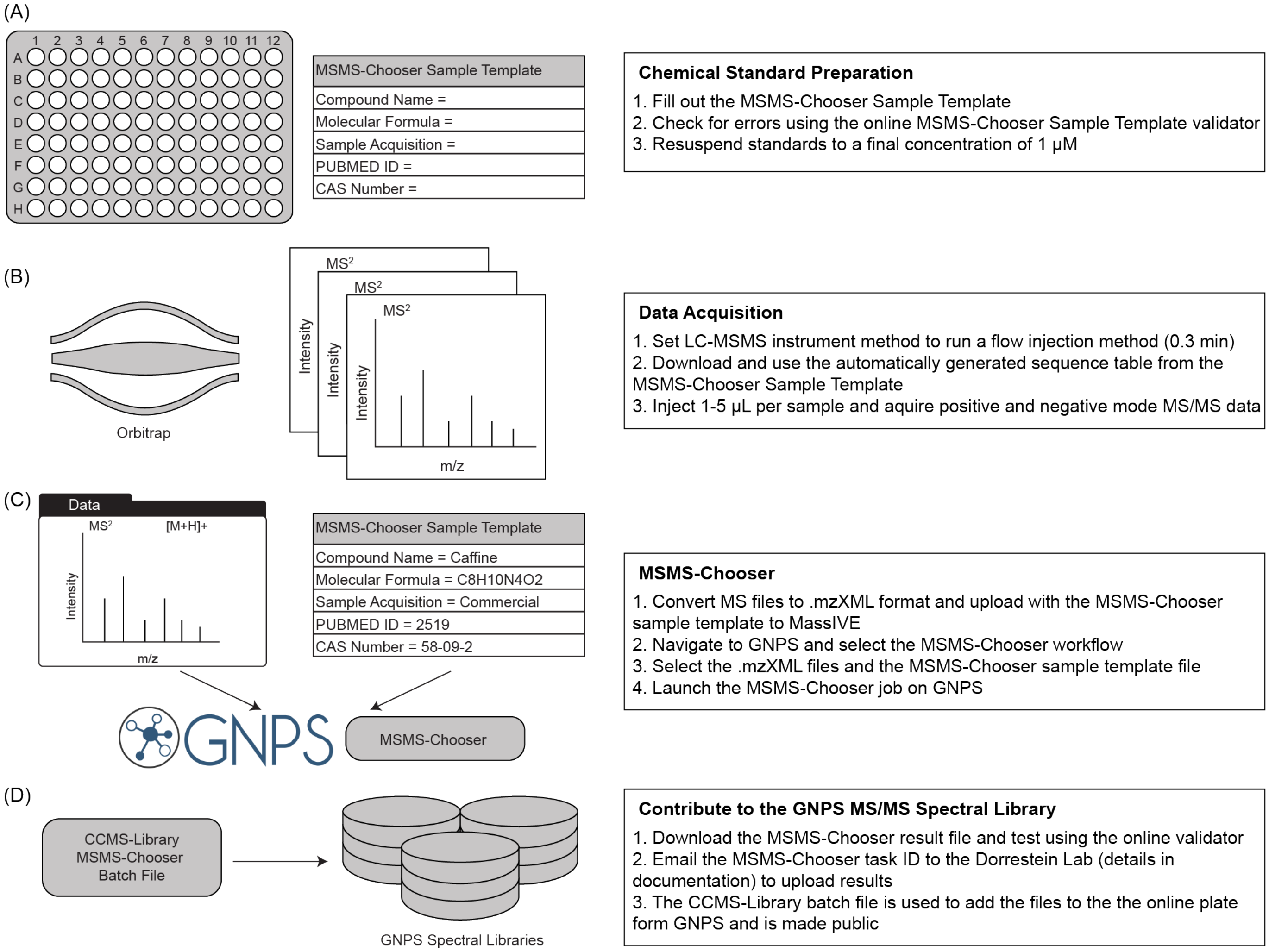
(9) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: Open Source Software for Rapid Proteomics Tools Development. *Bioinformatics* **2008**, *24* (21), 2534–2536. https://doi.org/10.1093/bioinformatics/btn323.

(10) Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; et al. ClassyFire: Automated Chemical Classification with a Comprehensive, Computable Taxonomy. *J. Cheminform.* **2016**, *8* (1), 1–20. https://doi.org/10.1186/s13321-016-0174-y.

(11) Wang, M.; Jarmusch, A. K.; Vargas, F.; Aksenov, A. A.; Gauglitz, J. M.; Weldon, K.; Petras, D.; Silva, R. da; Quinn, R.; Melnik, A. V.; et al. MASST: A Web-Based Basic Mass Spectrometry Search Tool for Molecules to Search Public Data. *bioRxiv* **2019**, 591016. https://doi.org/10.1101/591016.

(12) Jarmusch, A. K.; Wang, M.; Aceves, C. M.; Advani, R. S.; Aguire, S.; Aksenov, A. A.; Aleti, G.; Aron, A. T.; Bauermeister, A.; Bolleddu, S.; et al. Repository-Scale Co- and Re-Analysis of Tandem Mass Spectrometry Data. *bioRxiv* **2019**, 750471. https://doi.org/10.1101/750471.

Figures and Tables



MSMS-Chooser Sample Template

Compound Name =	
Molecular Formula =	
Sample Acquisition =	
PUBMED ID =	
CAS Number =	

Chemical Standard Preparation

1. Fill out the MSMS-Chooser Sample Template
2. Check for errors using the online MSMS-Chooser Sample Template validator
3. Resuspend standards to a final concentration of 1 μM

Data Acquisition

1. Set LC-MSMS instrument method to run a flow injection method (0.3 min)
2. Download and use the automatically generated sequence table from the MSMS-Chooser Sample Template
3. Inject 1-5 μL per sample and aquire positive and negative mode MS/MS data

MSMS-Chooser Sample Template

Compound Name = Caffine	
Molecular Formula = C8H10N4O2	
Sample Acquisition = Commercial	
PUBMED ID = 2519	
CAS Number = 58-09-2	

MSMS-Chooser

1. Convert MS files to .mzXML format and upload with the MSMS-Chooser sample template to MassIVE
2. Navigate to GNPS and select the MSMS-Chooser workflow
3. Select the .mzXML files and the MSMS-Chooser sample template file
4. Launch the MSMS-Chooser job on GNPS

Contribute to the GNPS MS/MS Spectral Library

1. Download the MSMS-Chooser result file and test using the online validator
2. Email the MSMS-Chooser task ID to the Dorrestein Lab (details in documentation) to upload results
3. The CCMS-Library batch file is used to add the files to the the online plate form GNPS and is made public

Figure 1. Overview of the MSMS-Chooser workflow and protocol.

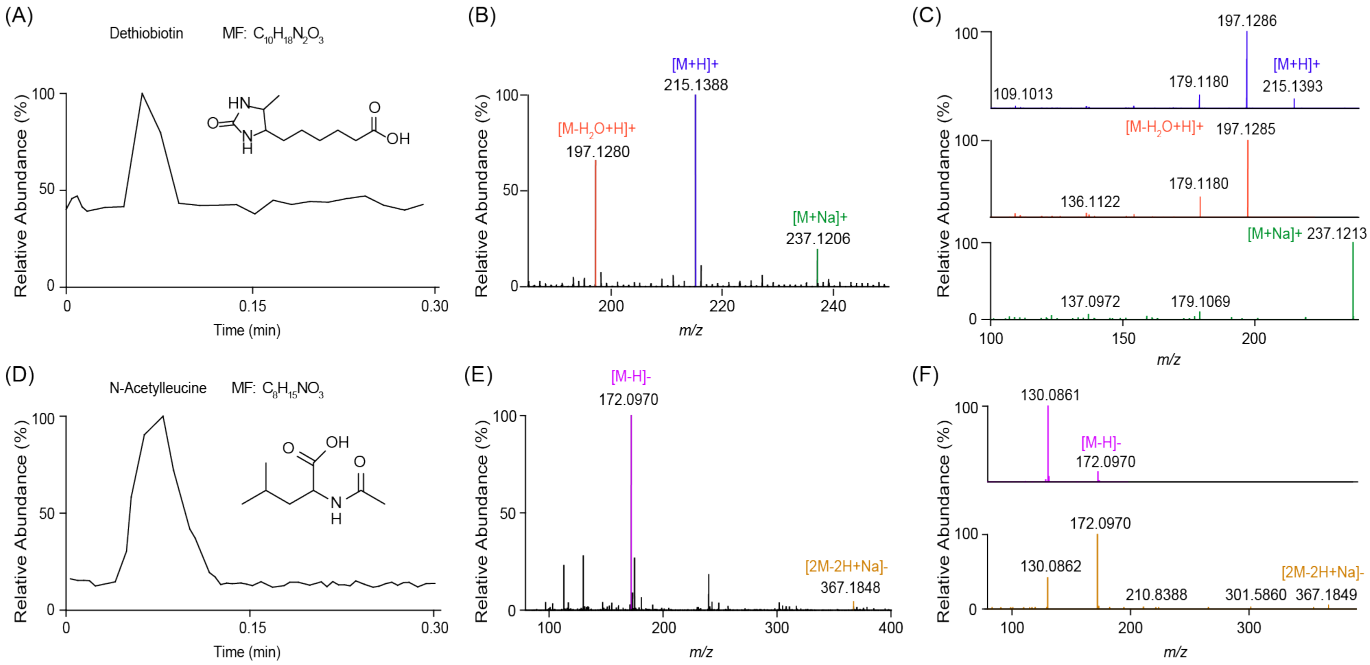


Figure 2. (A) Illustrative base peak chromatogram, (B) average MS spectrum, and (C) MS/MS spectra selected by MSMS-Chooser for dethiobiotin in the positive ionization mode. Peaks in the MS are colored corresponding to their MS/MS spectra. (D) Illustrative base peak chromatogram, (E) average MS spectrum, and (F) MS/MS spectra selected by MSMS-Chooser for N-acetylleucine in the negative ionization mode. Peaks in the MS are colored corresponding to their MS/MS spectra.

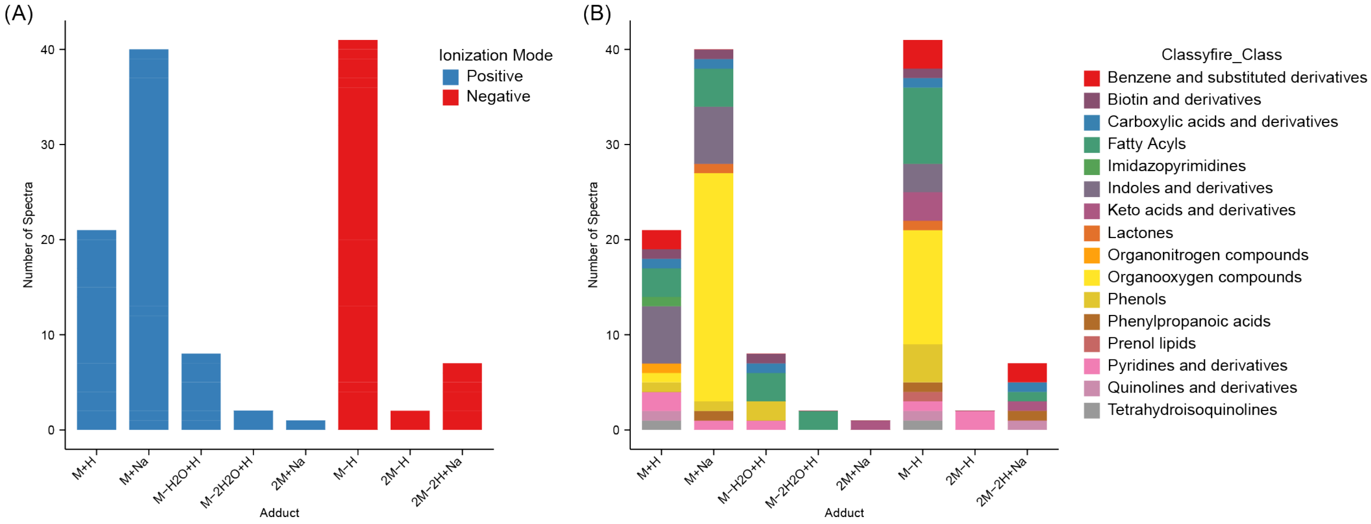


Figure 3. Adducts versus the number of MS/MS spectra selected by MSMS-Chooser in the proof-of-concept set of chemical standards, (A) colored by ionization mode and (B) colored by chemical classification, class level, obtained via Classyfire.

Table 1. Tabulated results comparing MSMS-Chooser and manual inspection of the data.

Ionization Mode	Number of MS/MS spectra, Manual Inspection	Number of MS/MS spectra, MS-MS Chooser[a]	MS-MS Chooser Accuracy (%)	MS-MS Chooser Accuracy (%)[b]
Positive	73	70	95.9	98.6
Negative	53	50	94.3	100
Total	126	120	95.2	99.2

[a] MSMS-Chooser correctly identified 70 MS/MS spectra and incorrectly identified one MS/MS spectra in positive mode upon manual inspection which as removed for subsequent calculations.

[b] Manual inspection found 2 positive mode and 3 negative mode spectra that were not chosen by MSMS-Chooser. These spectra were excluded in the accuracy calculation as their mass error was greater than the 10 ppm mass error limit (an adjustable parameter) used in MSMS-Chooser.

Supplemental Information

Protocol for Community-created Public MS/MS Reference Library Within the GNPS Infrastructure

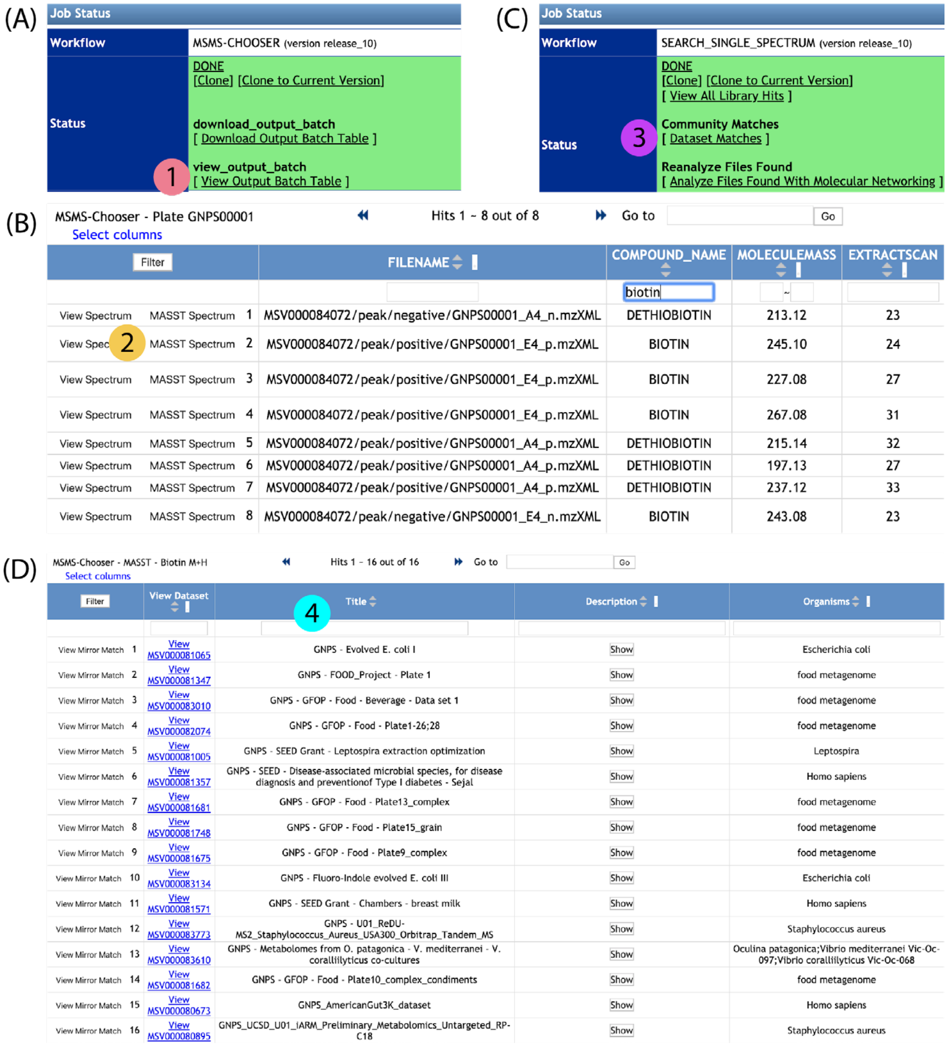
Fernando Vargas[1,2,†], Kelly C. Weldon[1,4,†], Nicole Sikora[1,†], Mingxun Wang[1], Zheng Zhang[1], Emily C. Gentry[1], Morgan W. Panitchpakdi[1], Mauricio Caraballo[1], Pieter C. Dorrestein[1,3,4]*, and Alan K. Jarmusch[1]*

Affiliations

1. Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences and Collaborative Mass Spectrometry Innovation Center, University of California, San Diego, La Jolla, CA 92093, United States of America
2. Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093, United States of America
3. Department of Pediatrics, University of California, San Diego, La Jolla, CA 92093, United States of America
4. Center for Microbiome Innovation, University of California, San Diego, La Jolla, CA 92093, United States of America

[†] contributed equally for this project

* corresponding authors



(A)

Job Status	
Workflow	MSMS-CHOOSER (version release_10)
Status	DONE [Clone] [Clone to Current Version] download_output_batch [Download Output Batch Table] view_output_batch [View Output Batch Table]

(C)

Job Status	
Workflow	SEARCH_SINGLE_SPECTRUM (version release_10)
Status	DONE [Clone] [Clone to Current Version] [View All Library Hits] Community Matches [Dataset Matches] Reanalyze Files Found [Analyze Files Found With Molecular Networking]

(B) MSMS-Chooser - Plate GNPS00001 Hits 1 - 8 out of 8 Go to Go
Select columns

Filter			FILENAME	COMPOUND_NAME	MOLECULEMASS	EXTRACTSCAN
				biotin	~	
View Spectrum	MASST Spectrum	1	MSV000084072/peak/negative/GNPS00001_A4_n.mzXML	DETHIOBIOTIN	213.12	23
View Spec	MASST Spectrum	2	MSV000084072/peak/positive/GNPS00001_E4_p.mzXML	BIOTIN	245.10	24
View Spectrum	MASST Spectrum	3	MSV000084072/peak/positive/GNPS00001_E4_p.mzXML	BIOTIN	227.08	27
View Spectrum	MASST Spectrum	4	MSV000084072/peak/positive/GNPS00001_E4_p.mzXML	BIOTIN	267.08	31
View Spectrum	MASST Spectrum	5	MSV000084072/peak/positive/GNPS00001_A4_p.mzXML	DETHIOBIOTIN	215.14	32
View Spectrum	MASST Spectrum	6	MSV000084072/peak/positive/GNPS00001_A4_p.mzXML	DETHIOBIOTIN	197.13	27
View Spectrum	MASST Spectrum	7	MSV000084072/peak/positive/GNPS00001_A4_p.mzXML	DETHIOBIOTIN	237.12	33
View Spectrum	MASST Spectrum	8	MSV000084072/peak/negative/GNPS00001_E4_n.mzXML	BIOTIN	243.08	23

(D) MSMS-Chooser - MASST - Biotin M+H Hits 1 - 16 out of 16 Go to Go
Select columns

Filter		View Dataset	Title	Description	Organisms
View Mirror Match	1	View MSV000081065	GNPS - Evolved E. coli I	Show	Escherichia coli
View Mirror Match	2	View MSV000081347	GNPS - FOOD_Project - Plate 1	Show	food metagenome
View Mirror Match	3	View MSV000083010	GNPS - GFOP - Food - Beverage - Data set 1	Show	food metagenome
View Mirror Match	4	View MSV000082074	GNPS - GFOP - Food - Plate1-26;28	Show	food metagenome
View Mirror Match	5	View MSV000081005	GNPS - SEED Grant - Leptospira extraction optimization	Show	Leptospira
View Mirror Match	6	View MSV000081357	GNPS - SEED - Disease-associated microbial species, for disease diagnosis and preventionof Type I diabetes - Sejal	Show	Homo sapiens
View Mirror Match	7	View MSV000081681	GNPS - GFOP - Food - Plate13_complex	Show	food metagenome
View Mirror Match	8	View MSV000081748	GNPS - GFOP - Food - Plate15_grain	Show	food metagenome
View Mirror Match	9	View MSV000081675	GNPS - GFOP - Food - Plate9_complex	Show	food metagenome
View Mirror Match	10	View MSV000083134	GNPS - Fluoro-Indole evolved E. coli III	Show	Escherichia coli
View Mirror Match	11	View MSV000081571	GNPS - SEED Grant - Chambers - breast milk	Show	Homo sapiens
View Mirror Match	12	View MSV000083773	GNPS - U01_ReDU-MS2_Staphylococcus_Aureus_USA300_Orbitrap_Tandem_MS	Show	Staphylococcus aureus
View Mirror Match	13	View MSV000083610	GNPS - Metabolomes from O. patagonica - V. mediterranei - V. coralliilyticus co-cultures	Show	Oculina patagonica;Vibrio mediterranei Vic-Oc-097;Vibrio coralliilyticus Vic-Oc-068
View Mirror Match	14	View MSV000081682	GNPS - GFOP - Food - Plate10_complex_condiments	Show	food metagenome
View Mirror Match	15	View MSV000080673	GNPS_AmericanGut3K_dataset	Show	Homo sapiens
View Mirror Match	16	View MSV000080895	GNPS_UCSD_U01_iARM_Preliminary_Metabolomics_Untargeted_RP-C18	Show	Staphylococcus aureus

Figure S1. Overview of launching a MASST search from the MSMS-Chooser workflow results page. (A) MSMS-Chooser results page, (1) click on the “View Output Batch Table”. (B) View output batch page, (2) clicking on the “MASST Spectrum” button will launch a MASST search (*e.g.* biotin). (C) MASST search results page (https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=39cd886540c147e9a8a618b275e5f541), (3) click on the “Dataset Matches”. (D) Datasets in which biotin (an example) was detected using MASST, (4) the title and dataset accession link can be used to further investigate the data.

Table S1. The minimum chemical information required for the MSMS-Chooser workflow which is found in the MSMS-Chooser template. Caffeine, urobilin, and cocaine are used as examples.

Project ID	Date Template Created	96 Well Plate Position	Compound Name	Molecular Formula	SMILES	INCHI	Acquisition	PubMED ID	CAS
2019_dorrestein	20190420	A1	Caffeine	C8H10N402	CN1C=NC2=C1C(=O)N(C(=O)N2C)C	InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3	Commercial	2519	58-08-2
2019_dorrestein	20190420	A2	Urobilin	C33H42N4O6	CCC1=C(C(=O)NC1CC2=C(C(=C(N2)C=C3C(=C(C(=N3)CC4C(=C(C(=O)N4)CC)C)C)CCC(=O)O)CCC(=O)O)C)C	InChI=1S/C33H42N4O6/c1-7-20-19(6)32(42)37-27(20)14-25-18(5)23(10-12-31(40)41)29(35-25)15-28-22(9-11-30(38)39)17(4)24(34-28)13-26-16(3)21(8-2)33(43)36-26/h15,26-27,35H,7-14H2,1-6H3,(H,36,43)(H,37,42)(H,38,39)(H,40,41)/b28-15-/t26-,27-/m0/s1	Commercial	6433298	1856-98-0
2019_dorrestein	20190420	A3	Cocaine	C17H21NO4	CN1C2CCC1C(C(C2)OC(=O)C3=CC=CC=C3)C(=O)OC	InChI=1S/C17H21NO4/c1-18-12-8-9-13(18)15(17(20)21-2)14(10-12)22-16(19)11-6-4-3-5-7-11/h3-7,12-15H,8-10H2,1-2H3/t12-,13+,14-,15+/m0/s1	Commercial	446220	50-36-2