Mass spectrometry searches using MASST

To the Editor — We introduce a web-enabled mass spectrometry (MS) search engine, named Mass Spectrometry Search Tool (MASST; https://masst.ucsd.edu). By enabling searches of all small-molecule tandem MS (MS/MS) data in public metabolomics repositories, we posit that MASST will unlock these resources for clinical, environmental and natural product applications.

Introduced in 1990, a tool for discovering related protein or gene sequences named Basic Local Alignment Search Tool (BLAST) enabled researchers to query entire public sequence data repositories through a web interface (WebBLAST; https://blast.ncbi.nlm.nih.gov/Blast.cgi)[1]. WebBLAST is one of the most widely cited and used bioinformatics tools because it permits any researcher to answer simple questions, such as 'is a protein or DNA sequence common or rare?'. In the early days of public gene and protein databases, metadata, which include descriptions of sample, population or technical details, were limited. No deposition standards existed, except for the Short Read Archive and European Nucleotide Archive, which include experimental details for sequencing, instrumental details and sample description, such as the source of a sample. The current status of much MS data in the public domain is reminiscent of the DNA databanks of the 1990s. To increase usage and unlock the potential of openly available MS resources, we set out to build an infrastructure to enable WebBLAST for MS.

Algorithms developed for MS data, including molecular networking[2] and fragmentation trees[3], enable similarity searches against reference libraries of known molecules, whereas powerful metabolomics analysis software infrastructures, such as MS-DIAL[4], MetaboAnalyst[5], XCMS Online[6] and HMDB[7], focus on annotation of MS/MS spectra, or finding statistical relationships between molecular features. However, none of the existing tools enable searching a single MS/MS spectrum for identical or analogous MS/MS spectra against public data in repositories, including unknown molecules. Finding specific MS/MS spectra of interest, including unannotated spectra or structural analogs, in public repositories of metabolomics MS data and natural product MS data, is not possible. Deposition of untargeted MS data in the public domain is experiencing rapid growth. In March 2017, 910 metabolomics datasets were available[8]; by January 2019, there were >2,000 downloadable metabolomics datasets (about half of these datasets contain MS/MS data)[9]. Despite the availability of metabolomics and natural product data, including environmental and clinical MS datasets, public small-molecule MS data are hardly reused[10]. Now that there is a huge amount of small-molecule untargeted MS datasets publicly available (~1,100 untargeted datasets and ~110,000,000 spectra in ~150,000 files as of December 11, 2018), we felt that the time was right to develop MASST, to enable reuse of these MS data.

MASST comprises a web-based system to search the public data repository part of the GNPS/MassIVE knowledge base[11] and an analysis infrastructure for a single MS/MS spectrum. The developments required for MASST searches included converting deposited public data to a uniform open format[12] (irrespective of instrument type and original data format), the ability to trace the file from which each MS/MS spectrum originated, and a reporting system that shows all identical or similar MS/MS spectra found in public data along with their associated metadata. MASST development has been possible for two main reasons: first, adoption of universal, non-vendor-specific MS data formats has increased, which means that multiple publicly available datasets have been converted to the same data format[13], and second, the recently developed ability to connect all public data in GNPS/MassIVE and connect each MS/MS spectrum to its metadata entries had not been developed yet.

A MASST report also includes matches to any reference spectra in public MS/MS spectral libraries, if the matches are within the user-specified search parameters. Libraries include GNPS user-contributed spectra[11], GNPS libraries[11], all three MassBanks[14] (https://massbank.eu/MassBank/, https://mona.fiehnlab.ucdavis.edu/), ReSpect[15], MIADB/Beniddir[16], Sumner/Bruker, CASMI[17], PNNL lipids[18], Sirenas/Gates, EMBL MCF and several other libraries, listed at https://gnps.ucsd.edu/ProteoSAFe/libraries.jsp. Visualization of the MASST matches uses a mirror view (Fig. 1).

MASST can search against various repositories, including GNPS/MassIVE[11], Metabolomics Workbench[19], MetaboLights[20] or the non-redundant (nr) MS/MS library of all unique MS/MS spectra from all three repositories combined. MASST searching using multiple repositories was enabled by converting data uploaded to the Metabolomics Workbench and MetaboLights repositories to the same open MS format in the GNPS/MassIVE data storage environment. Instructions on how to upload to GNPS/MassIVE can be found at https://ccms-ucsd.github.io/GNPSDocumentation/datasets/.

All public data in GNPS/MassIVE becomes MASST-searchable. MASST searches output results according to user-defined search parameters. The report returns the origin of the matched MS/MS spectrum with respect to the dataset and file information and any metadata associated with the file (Fig. 1). Datasets and files can be tagged with sample or spectral information by the community of MASST users, and this information then becomes part of the metadata reported back in future MASST searches. We also curated ~34,000 additional MS files with ~340,000 tags, mostly from human-associated samples, but also from microbes, food and indoor and outdoor environments, to provide a good foundation for MASST searches.

Metadata can be associated with MS/MS spectra in the GNPS/MassIVE upload portal at the dataset level, file level or single annotated spectrum level. Examples of metadata include instrument type, phylogeny (according to the National Center for Biotechnology Information (NCBI) taxonomy) and keywords at the dataset level; phylogeny, sample type, age, sex, body site (defined using the Uberon anatomy ontology[21]) and disease[22] at the file level; and source, biological activity and structural class information at the single annotated spectrum level. In addition, GNPS/MassIVE is compatible with metadata formats from other software tools (e.g., QIIME2 and Qiita), which are used to analyze microbiome data and have a controlled vocabulary that can be imported[23,24]. Any sample information uploaded to GNPS/MassIVE from another repository (e.g., from MetaboLights and Metabolomics workbench) is also included in the MASST report.

At present, there is only limited metadata at the dataset and file level, but the metadata in the public domain can provide insights into the types of MS/MS signals being analyzed (Box 1 contains examples of usage). Although the amount and quality of metadata is increasing[25], datasets do not always have detailed metadata. To allay this problem, re-annotation of metadata as knowledge increases, while retaining provenance of all changes, is possible in GNPS.[11] If insufficient metadata are available for interpretation of a public dataset search results, the original depositors of the public

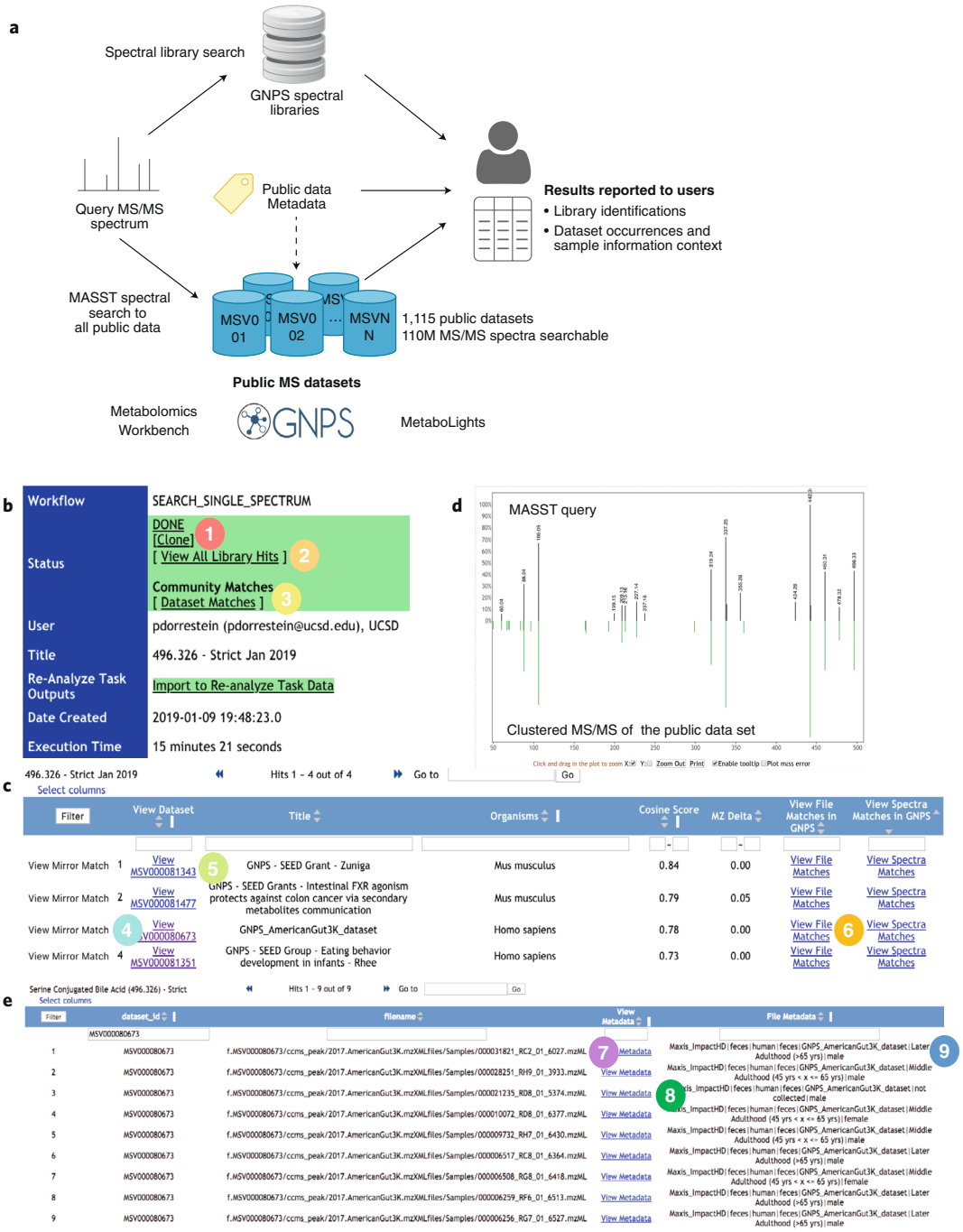

Fig. 1 | MASST search, reporting and match visualization. a, Overview of MASST query procedure. MASST queries MS/MS spectra against all public metabolomics data, including spectra deposited in GNPS, Metabolomics Workbench and MetaboLights. Combining these matches with sample information provides users with a report containing MS/MS compound annotation and MS/MS sample information (metadata). Once a MASST search is completed at https://proteosafe-extensions.ucsd.edu/masst/, the results can be found in the user's job tab or using a link provided through email. **b**, The opening page is shown. There are two options (circle numbers 2 and 3) for inspecting the data and additional options for cloning a job (1). Clicking (2) will reveal all MS/MS spectral matches within the user defined settings. There can be none, one, or more than one match for a given input spectrum. Clicking (3) will reveal all data sets that contain an MS/MS spectrum that has a match to the input spectrum and any associated metadata. **c**,**d**, Clicking on 'View Mirror Match' (4) in **c** shows the mirror match between the input spectrum and the merged MS/MS spectrum (**d**), enabling manual inspection of a match; 'View MSV0000…..' (5) brings the user to the data set: all uploaded information associated with this data set can be found or is linked in this location. Clicking (6) opens the file information window and tabulated metadata. **e**, Circle (7) shows the files where MS/MS matches are found. Circle (8) links to full sample information for the file. Circle (9) displays the abbreviated (and filterable) sample information associated with the files. If no sample information has been uploaded with the original data, then this field will be blank. The MASST_GNPS job link for this search to enable the reader to navigate the same results is available at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=bac3d3788e704af59e4a15a5146e4d6b.

Box 1 | Ten applications and questions addressed with MASST

MASST can be used in several ways for research questions, which are detailed here.

1. Are specific molecular features detected via MS in one clinical cohort also observed elsewhere? Human studies of individuals with disease versus healthy cohorts are confounded by different exercise, diet and medications in the cohort of individuals with disease versus healthy cohort. Using MASST on an MS/MS feature found to be differentiated with non-alcoholic fatty acid liver disease (NAFLD) in humans revealed the same MS/MS could be found in other liver disease studies. The video describing this MASST search is at https://youtu.be/sHHlVTCoQJY. An expanded description of the MASST and discovery of the new bile acid is available in the Supplementary Note.

2. Can findings about a molecule identified in model organism studies be translated to humans? One major expected application will be the translation of molecular information from animal models to humans. Using MASST on the MS/MS data of a MS feature that was differentiated in a mouse model infected with lymphocytic choriomeningitis virus Armstrong resulted in the discovery that a new molecule, cholylserine, is also found in human studies. Details of this MASST job are at https://youtu.be/SExVUrD56-s and in the Supplementary Note.

3. Can MASST be used to reveal the presence and distribution of environmental toxins? In this example, domoic acid — the neurotoxin poison that became famous through the novel *The Birds* by Daphne du Maurier and a film from Alfred Hitchcock as it caused seagulls to attack humans — is found in seven different environmental public datasets, including San Diego, Narragansett Bay and Hawaii. A description of the results of this MASST analysis job is at https://youtu.be/vm6UkYwDGn4 and in the Supplementary Note.

4. In what datasets can we find a published MS/MS spectrum? A published MS/MS spectrum was searched. A description of MASST using the MS/MS spectrum of 3-hydroxyhexadecanoyl glycine and 3-hydroxypentadecanoyl lysine, both *N*-acyl lipids, suggests that these molecules have a very wide ecological distribution and is available at https://youtu.be/8W2BCxtszIA and in the Supplementary Note.

5. Are specific natural products observed in cultured microbes also observed in non-laboratory settings? An example using orfamides revealed four datasets that contained this molecular ion, including field-collected *Trachymyrmex septentrionalis* fungus gardens. More detail is available at https://youtu.be/4Zb5gZIabBU and in the Supplementary Note.

6. Where do we find agricultural fungicides in the environment? Is there evidence that people may be in contact with these fungicides? A description of a MASST search with the MS/MS spectrum of azoxystrobin, a fungicide, is available at https://youtu.be/hGemmjdeOY0 and in the Supplementary Note.

7. Are known toxins from food found in/on people? A MASST search with the MS/MS spectrum of the mycotoxin roquefortine C revealed that it was found in human stool (infants and adults). This MASST search is described in more detail at https://youtu.be/04RSsOY0oGM and in the Supplementary Note.

8. Can we use approximate matches to a natural product to find datasets that may contain analogs? A search for staurosporine derivatives among the public datasets with MASST took less than 15 min, and suggests that there are still yet-to-be-discovered reservoirs of unique staurosporine derivatives, as shown at https://www.youtube.com/watch?v=Yu-ytgjDPeU&t=4s and described in the Supplementary Note.

9. Can MASST be used to track sunscreens in human and environmental samples? A MASST search of the MS/MS spectra of two active ingredients of sunscreen — avobenzone and octocrylene — revealed, as expected, their presence in many human skin datasets, personal objects, meat for human consumption, corals and even in coral reef in remote areas such as Moorea. This MASST analysis job is described at https://youtu.be/Sjv00dpMSQ8 and in the Supplementary Note.

10. Can we find evidence of exposure to opioids in public data? By searching the MS/MS of methadone and cocaine using MASST, we found a matching MS/MS spectrum in five datasets. A description with this MASST analysis job is available at https://youtu.be/9hTsXJ6l1Is and in the Supplementary Note.

data can be contacted. We expect this feature in MASST to foster collaborations worldwide.

MASST can be accessed at https://proteosafe-extensions.ucsd.edu/masst/ by copying and pasting the MS/MS spectrum peak list reported as mass-to-charge ratio (*m/z*) and intensity separated by a space for each fragment ion (also known as product ion), which can also be extracted from the open MS formats (e.g., .mzML, .mzXML and .MGF). Finally, MASST can be accessed as part of a GNPS data analysis. Manual entry at https://proteosafe-extensions.ucsd.edu/masst/ provides researchers with the ability to enter data from theoretical spectra, or spectra from published papers or supporting information, without needing access to the original experimental data. In GNPS users can launch a MASST search using links provided in classic and feature-based molecular networking output created within the GNPS infrastructure[11], which automatically redirects to the MASST search page with prepopulated spectral data by clicking a simple MASST spectrum button. The MS/MS spectrum provided via the MASST website or as a link-out from a GNPS search is then searched against all public data with user defined parameters of minimum number of ions to match, precursor (parent) and product (fragment) ion tolerances, and analog similarity searches based on non-identical precursor masses[2]. An instruction video for running MASST jobs is available at https://youtu.be/4yBKomKzEKU. MASST searches retrieve all associated sample information (dataset and files) that match the MS/MS input spectrum query. A typical search takes about 10–20 min. Multiple search queries are placed in a queue for parallel execution as resources become available.

To promote data analysis reproducibility, the results of every job are stored in each user's space and can be found under the 'Jobs' tab accessible through the banner in the GNPS browser (http://gnps.ucsd.edu). Only MASST jobs run while logged in on GNPS will be retained. Search parameters are also retained with each job and constitute a provenance record that can be provided as hyperlinks to share with others (e.g., collaborators) or in publications. These jobs can be shared, cloned and rerun with or without alterations of the input parameters (examples of links to jobs are shown in Box 1). This feature could enable new matches to be made when relevant public data are uploaded. The matches of MS/MS spectra among datasets are the equivalent to level two (putative annotation based on spectral library similarity) or three (putatively characterized compound

class based on spectral similarity to known compounds of a chemical class) according to the 2007 metabolomics standards initiative[26]. Similar to short sequence read searches, MASST searches will not distinguish chemicals that have nearly identical fragmentation patterns, such as isomeric compounds, which would require an authentic standard and the use of an orthogonal property (such as the retention time). In cases when a MASST search returns no matches, it is possible that either there are no matching data or that MS/MS matches are possible but fall outside the specified search parameters. MASST should be used with these caveats in mind.

MASST, like WebBLAST, will likely find broad application. Uses of MASST might include translation of in vitro or in vivo data from model organisms to humans, or broad ecological questions. Box 1 contains ten example uses to highlight the types of discoveries possible with access, via MASST, to the entire body of public MS/MS data. These examples are illustrative, and we expect the user community to find multiple, innovative ways to use MASST.

Data availability

All data used for testing and validating MASST are deposited in GNPS/MassIVE. MASST is a web-based application that is embedded in GNPS, which is a community service in which all public data are public. All data underlying figures present in the Supplementary Note are included as Supplementary Data 1 and 2. We cannot provide server installation, software engineers or administrator support for individual installations of MASST. The MASST platform is built as a workflow on top of the web repository workflow platform ProteoSAFe (https://github.com/CCMS-UCSD/ProteoSAFe). Each step of the MASST query is written in Python. Web rendering of the results is displayed by ProteoSAFe in the browser.

Code availability

For those who wish to build out MASST and recruit their own programmers, software engineers and system administrators, we have deposited the code at github (https://github.com/CCMS-UCSD/GNPS_Workflows/tree/master/search_single_spectrum). The standalone MASST query interface is written in Python and Flask with a web front end written in HTML and JavaScript. It is open source (https://github.com/mwang87/GNPS_MASST) and released under an LGPL-3 license. ❐

Mingxun Wang[1,2], Alan K. Jarmusch[1], Fernando Vargas[1,3], Alexander A. Aksenov[1], Julia M. Gauglitz[1], Kelly Weldon[1,4], Daniel Petras[1], Ricardo da Silva[1], Robert Quinn[1,5], Alexey V. Melnik[1], Justin J. J. van der Hooft[1,6], Andrés Mauricio Caraballo-Rodríguez[1], Louis Felix Nothias[1], Christine M. Aceves[1], Morgan Panitchpakdi[1], Elizabeth Brown[1], Francesca Di Ottavio[7], Nicole Sikora[1], Emmanuel O. Elijah[1], Lara Labarta-Bajo[3], Emily C. Gentry[1], Shabnam Shalapour[8], Kathleen E. Kyle[9], Sara P. Puckett[10], Jeramie D. Watrous[11], Carolina S. Carpenter[4], Amina Bouslimani[1], Madeleine Ernst[1], Austin D. Swafford[4], Elina I. Zúñiga[3], Marcy J. Balunas[10], Jonathan L. Klassen[9], Rohit Loomba[4,12], Rob Knight[4,13,14], Nuno Bandeira[4,14,15] and Pieter C. Dorrestein[1,4,8,13]*

[1]*Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA.* [2]*Ometa Labs LLC, San Diego, CA, USA.* [3]*Division of Biological Sciences, University of California San Diego, La Jolla, CA, USA.* [4]*Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, USA.* [5]*Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA.* [6]*Bioinformatics Group, Wageningen University, Wageningen, The Netherlands.* [7]*Faculty of Bioscience and Technology for Food, Agriculture, and Environment, University of Teramo, Teramo, TE, Italy.* [8]*Department of Pharmacology, School of Medicine, University of California San Diego, La Jolla, CA, USA.* [9]*Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA.* [10]*Division of Medicinal Chemistry, Department of Pharmaceutical Sciences, University of Connecticut, Storrs, CT, USA.* [11]*Department of Medicine, University of California San Diego, San Diego, California, USA.* [12]*Division of Gastroenterology, University of California San Diego, La Jolla, CA, USA.* [13]*Department of Pediatrics, University of California San Diego, La Jolla, CA, USA.* [14]*Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA.* [15]*Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA.*
**e-mail: pdorrestein@health.ucsd.edu*

Published online: 1 January 2020
https://doi.org/10.1038/s41587-019-0375-9

References

1. Altschul, S. F. et al. *J. Mol. Biol.* **215**, 403–410 (1990).
2. Watrous, J. et al. *Proc. Natl Acad. Sci. USA* **109**, 1743–1752 (2012).
3. Rasche, F. *Anal. Chem.* **83**, 1243–1251 (2011).
4. Lai, Z. et al. *Nat. Methods* **15**, 53–56 (2018).
5. Chong, J. et al. *Nucleic Acids Res.* **46**, W486–W494 (2018).
6. Tautenhahn, R. et al. *Anal. Chem.* **84**, 5035–5039 (2012).
7. Wishart, D. S. et al. *Nucleic Acids Res.* **46**, D608–D617 (2018).
8. Aksenov, A. A. et al. *Nat. Rev. Chem.* **1**, 0054 (2017).
9. Perez-Riverol, Y. et al. *Nat. Biotechnol.* **35**, 406–409 (2017).
10. Rocca-Serra, P. et al. *Metabolomics* **12**, 14 (2016).
11. Wang, M. et al. *Nat. Biotechnol.* **34**, 828–837 (2016).
12. Kirchner, M. et al. *J. Proteome Res.* **9**, 2762–2763 (2010).
13. Kessner, D. et al. *Bioinformatics* **24**, 2534–2536 (2008).
14. Horai, H. et al. *J. Mass Spectrom.* **45**, 703–714 (2010).
15. Sawada, Y. et al. *Phytochemistry* **82**, 38–45 (2012).
16. Otogo N'Nang, E. et al. *Org. Lett.* **20**, 6596–6600 (2018).
17. Schymanski, E. L. et al. *Metabolites* **3**, 517–538 (2013).
18. Kyle, J. E. et al. *Bioinformatics* **33**, 1744–1746 (2017).
19. Haug, K. et al. *Nucleic Acids Res.* **41**, D781–D786 (2013).
20. Sud, M. et al. *Nucleic Acids Res.* **44**, D463–D470 (2016).
21. Mungall, C. J. et al. *Genome Biol.* **13**, R5 (2012).
22. Schriml, L. M. et al. *Nucleic Acids Res.* **47**, D955–D962 (2019).
23. Bolyen, E. et al. *Nat. Biotechnol.* **37**, 852–857 (2019).
24. Gonzalez, A. et al. *Nat. Methods* **15**, 796–798 (2018).
25. Jarmusch, A.K. et al. Preprint at *bioRxiv* https://doi.org/10.1101/750471 (2019).
26. Sumner, L. W. et al. *Metabolomics* **3**, 211–221 (2007).

Acknowledgements

Conversion of data from different repositories was supported by R03 CA211211 on reuse of metabolomics data. The development of a user-friendly interface was in part supported by Gordon and Betty Moore Foundation through grant GBMF7622. The UC San Diego Center for Microbiome Innovation supported the campus wide SEED grant awards for data collection that enabled the development of much of this infrastructure. A.K.J. thanks the American Society for Mass Spectrometry for the 2018 Postdoctoral Career Development Award. We acknowledge C. O'Donovan and K. Haug for help with navigating the MetaboLights data repository. J.V.D.H. was supported by a ASDI eScience grant (ASDI.2017.030) from the Netherlands eScience Center (NLeSC). E.I.Z. and L.L.-B. were supported by NIH grants AI081923 and AI113923. A.M.C.R., K.E.K., S.P.P., J.L.K., M.J.B. and P.C.D. were supported by NSF grant IOS-1656475. A.B. was supported by National Institute of Justice Award 2015-DN-BX-K047. F.V. was supported by the Department of Navy, Office of Naval Research Multidisciplinary University Research Initiative (MURI) Award, award number N00014-15-1-2809. D.P. was supported by the German Research Foundation (DFG) with grant PE 2600/1. Additional support for data acquisition and data storage was provided by P41 GM103484 Center for Computational Mass Spectrometry, Instrument support though NIH S10RR029121. R.L. is supported by NIH grants R01DK106419, 5P42ES010337 and 5UL1TR001442, and NIH K01DK116917 to J.D.W. The development of the web interface and harmonization with Qiita was in part supported by the Sloan Foundation.

Author contributions

P.C.D. and M.W. came up with the concept of MASST. M.W. and N.B. performed the engineering to enable MASST. M.W., A.V.M., A.K.J., J.J.J.v.d.H., J.M.G., M.P., E.O.E., K.W., C.M.A., F.D.O., E.B., A.B., R.Q., M.C., N.S. and S.S. curated metadata. F.V., J.M.G., L.L.-B., K.W., E.B., A.A.A., R.Q., M.C. and C.S.C. generated data for the manuscript. E.C.G. synthesized the bile acids. P.C.D., M.W., D.P., J.D.W., M.J., L.F.N., J.M.G., E.I.Z., L.L.-B., K.E.K., S.P.P., A.M.C.R., A.V.M., F.V., K.W., A.A.A. and S.S. performed experiments and/or analysis for Box 1. P.C.D., D.P., L.F.N., J.J.J.v.d.H., J.M.G., A.A.A., A.M.C.R., F.V., K.W., A.B., F.D.O., M.E. and R.d.S. tested the MASST infrastructure and downloaded public data. P.C.D., N.B., E.I.Z., R.L., R.K., A.D.S., M.J.B. and J.L.K. provided supervision and funding for the project. P.C.D., A.K.J., D.P., J.J.v.d.H., M.E., J.M.G., A.A.A., A.M.C.R., R.K., J.L.K., L.F.N., N.B. and M.W. wrote and edited the manuscript.

Competing interests

M.W. is the founder of Ometa Labs LLC and consults for Sirenas, and P.C.D. is on the scientific advisory board of Sirenas.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0375-9.