Reproducible molecular networking of untargeted mass spectrometry data using GNPS

Nature protocols

Aron, Allegra T.; Gentry, Emily C.; McPhail, Kerry L.; Nothias, Louis Félix; Nothias-Esposito, Mélissa et al

https://doi.org/10.1038/s41596-020-0317-5

nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-020-0317-5

Check for updates

Reproducible molecular networking of untargeted mass spectrometry data using GNPS

Allegra T. Aron[1,25], Emily C. Gentry[1,25], Kerry L. McPhail[2,25], Louis-Félix Nothias[1], Mélissa Nothias-Esposito[1], Amina Bouslimani[1], Daniel Petras[1,3], Julia M. Gauglitz[1], Nicole Sikora[1], Fernando Vargas[1,4], Justin J. J. van der Hooft[5], Madeleine Ernst[1], Kyo Bin Kang[6], Christine M. Aceves[1], Andrés Mauricio Caraballo-Rodríguez[1], Irina Koester[1,3], Kelly C. Weldon[1,7], Samuel Bertrand[8,9], Catherine Roullier[6,9], Kunyang Sun[1], Richard M. Tehan[2], Cristopher A. Boya P.[10,11], Martin H. Christian[10], Marcelino Gutiérrez[10], Aldo Moreno Ulloa[12], Javier Andres Tejeda Mora[12], Randy Mojica-Flores[10,13], Johant Lakey-Beitia[10], Victor Vásquez-Chaves[14], Yilue Zhang[15], Angela I. Calderón[15], Nicole Tayler[10,11], Robert A. Keyzers[16], Fidele Tugizimana[17,18], Nombuso Ndlovu[17], Alexander A. Aksenov[1], Alan K. Jarmusch[1], Robin Schmid[19], Andrew W. Truman[20], Nuno Bandeira[21]✉, Mingxun Wang[1]✉ and Pieter C. Dorrestein[1,22,23,24]✉

Global Natural Product Social Molecular Networking (GNPS) is an interactive online small molecule–focused tandem mass spectrometry (MS^2) data curation and analysis infrastructure. It is intended to provide as much chemical insight as possible into an untargeted MS^2 dataset and to connect this chemical insight to the user's underlying biological questions. This can be performed within one liquid chromatography (LC)-MS^2 experiment or at the repository scale. GNPS-MassIVE is a public data repository for untargeted MS^2 data with sample information (metadata) and annotated MS^2 spectra. These publicly accessible data can be annotated and updated with the GNPS infrastructure keeping a continuous record of all changes. This knowledge is disseminated across all public data; it is a living dataset. Molecular networking—one of the main analysis tools used within the GNPS platform—creates a structured data table that reflects the molecular diversity captured in tandem mass spectrometry experiments by computing the relationships of the MS^2 spectra as spectral similarity. This protocol provides step-by-step instructions for creating reproducible, high-quality molecular networks. For training purposes, the reader is led through a 90- to 120-min procedure that starts by recalling an example public dataset and its sample information and proceeds to creating and interpreting a molecular network. Each data analysis job can be shared or cloned to disseminate the knowledge gained, thus propagating information that can lead to the discovery of molecules, metabolic pathways, and ecosystem/community interactions.

Introduction

Molecular networking for the analysis of tandem mass spectra of small molecules was introduced in 2012 (ref. [1]) for the analysis of metabolite production from a diverse set of live microbial colonies; this enabled the mapping of the chemical diversity observed in an untargeted mass spectrometry

[1]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA. [2]Department of Pharmaceutical Sciences, College of Pharmacy, Oregon State University, Corvallis, OR, USA. [3]Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA. [4]Division of Biological Sciences, University of California San Diego, La Jolla, CA, USA. [5]Bioinformatics Group, Wageningen University, Wageningen, the Netherlands. [6]College of Pharmacy, Sookmyung Women's University, Seoul, Korea. [7]Center of Microbiome Innovation, University of California San Diego, La Jolla, CA, USA. [8]Groupe Mer, Molécules, Santé-EA 2160, UFR des Sciences Pharmaceutiques et Biologiques, Université de Nantes, Nantes, France. [9]ThalassOMICS Metabolomics Facility, Plateforme Corsaire, Biogenouest, Nantes, France. [10]Centro de Biodiversidad y Descubrimiento de Drogas, Instituto de Investigaciones Científicas y Servicios de Alta Tecnología (INDICASAT AIP), Panama City, Panama. [11]Department of Biotechnology, Acharya Nagarjuna University, Guntur, Nagarjuna Nagar, India. [12]Biomedical Innovation Department, CICESE, Ensenada, Mexico. [13]Departamento de Química, Universidad Autónoma de Chiriquí (UNACHI), David, Chiriquí, Panama. [14]Centro de Investigaciones en Productos Naturales (CIPRONA), Universidad de Costa Rica, San José, Costa Rica. [15]Department of Drug Discovery and Development, Harrison School of Pharmacy, Auburn University, Auburn, AL, USA. [16]School of Chemical & Physical Sciences, Victoria University of Wellington, Wellington, New Zealand. [17]Centre for Plant Metabolomics Research, Department of Biochemistry, University of Johannesburg, Auckland Park, South Africa. [18]International R&D Division, Omnia Group (Pty) Ltd., Johannesburg, South Africa. [19]Institute of Inorganic and Analytical Chemistry, University of Münster, Münster, Germany. [20]Department of Molecular Microbiology, John Innes Centre, Norwich, UK. [21]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA. [22]Center for Computational Mass Spectrometry, University of California, San Diego, La Jolla, CA, USA. [23]Department of Pharmacology, University of California, San Diego, La Jolla, CA, USA. [24]Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA. [25]These authors contributed equally: Allegra T. Aron, Emily C. Gentry, Kerry L. McPhail. ✉e-mail: nbandeira@ucsd.edu; miw023@ucsd.edu; pdorrestein@ucsd.edu

experiment. Upon its introduction, molecular networking was compared to sequencing of environmental DNA to study the microbial communities present in diverse ecosystems[2]. In addition to providing unprecedented systems-level views of the chemical space in various environments, molecular networking has aided the elucidation of the structures of many compounds[3–9].

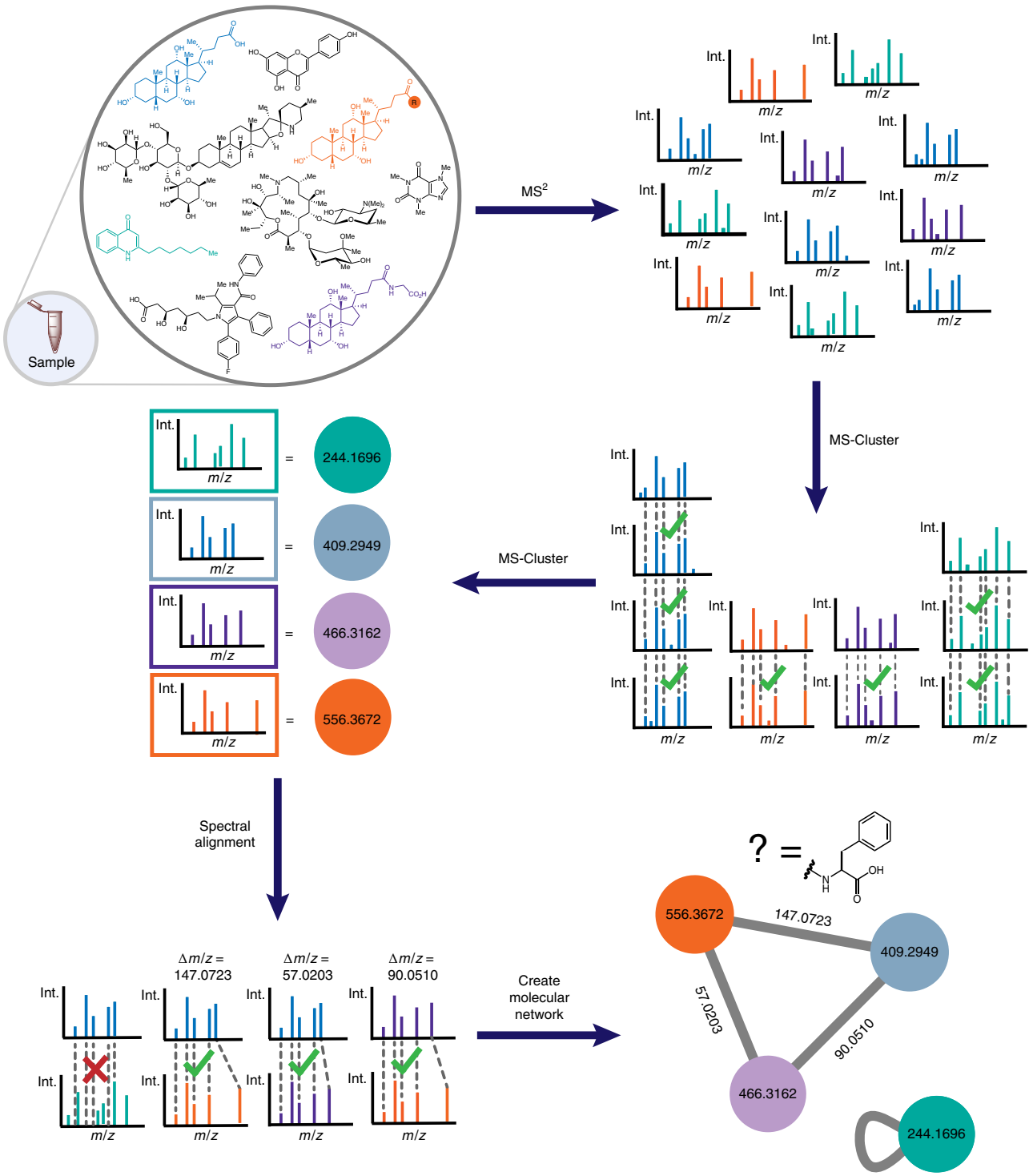
The foundation of molecular networking is pairwise spectral alignment using a modified cosine spectral similarity algorithm originally intended to discover modified forms of peptides and proteins[10]. In a modified spectral similarity search, not only are fragmentation spectra (MS^2) from ions at identical *m/z* values compared, but also MS^2 spectra that are offset by the same *m/z* difference as the precursor ion are compared. By eliminating the amino acid filtering from the original spectral alignment algorithms, it became possible to extend spectral similarity to any set of MS^2 spectra, including those from small molecules and natural products (NPs). When a pairwise spectral similarity search/alignment is performed, each MS^2 spectrum in a given dataset is compared against every other, and a network of MS^2 spectral relations is obtained, from which molecular networks are created (Fig. 1). Molecular networking builds on the fundamental observation that two structurally related molecules share fragment ion patterns when subjected to MS^2 fragmentation methods such as collision induced dissociation (CID). To make the molecular networking algorithm accessible to the scientific community, its script was converted to a web-based platform backed by a supercomputer. This enabled the creation of a community infrastructure supporting both a database and knowledge base around the needs of the community. The result was the GNPS (https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp) community effort that started in 2014 and was published in 2016. The user base has expanded to 49 of 50 states in the United States and worldwide to >150 countries[11]. GNPS is currently widely used by scientists working in industry, academia, and government in the fields of biomedical research, environmental science, ecology, forensics, microbiology, chemistry, and others. This crowd-sourced, community-driven analysis infrastructure not only facilitates data and knowledge storage but also enables knowledge capture, sharing, dissemination, and data-driven social networking while promoting reproducible data analysis. Moreover, GNPS can be accessed on a computer or on any mobile device connected to the Internet, making any public dataset readily accessible for analysis. Although there are many analysis tools available within the GNPS infrastructure, molecular networking is the most frequently used tool. Other tools available on GNPS, such as network annotation propagation (NAP) and 'ili, which enables molecular cartography, are briefly discussed.

To create a molecular network, GNPS first aligns each MS^2 spectrum in a dataset to each of the others and assigns a cosine score to each combination to describe their similarity (Fig. 1). Identical masses are collapsed on the basis of a hierarchical cosine clustering algorithm into a single node or consensus cluster because of the high similarity of their fragment ions. This is accomplished using the MS-Cluster algorithm[12]. Structurally related molecules yield comparable MS^2 spectra due to commonalities in their gas-phase chemistry[13] and are represented by separate nodes that connect within the network via edges. Each consensus spectrum (node) is then queried against spectral library databases to assign it to a putative known molecule within a network.

All mass spectrometry data used in GNPS, both those in the private user workspace and data that are made public, are stored in MassIVE—an interactive virtual environment developed to facilitate

Fig. 1 | Schematic representation of the process for creating a molecular network from tandem mass spectra acquired for metabolites in complex sample mixtures. The colors are used to track how we go from molecules in a sample to nodes in the molecular network. We start by obtaining MS^2 spectra of all ionized molecules in the sample. MS-Cluster first aligns each MS^2 spectrum in a dataset to each of the others. Mass spectra from identical compounds are coalesced using MS-Cluster[12] into a single node or consensus cluster because of the high similarity of their precursor ion and fragment ions. Subsequently a spectral alignment is performed, enabling similarity searches even when the precursor ion masses are not identical. This is accomplished using a modified cosine score, for which all the ions that differ by the mass difference of the two precursor ions are also considered. Structurally related molecules yield comparable MS^2 spectra because of commonalities in their gas-phase chemistry and are represented by separate nodes that connect within the network via edges. Each node is then queried against spectral libraries to assign putative known molecules within a molecular network, and unknowns can be propagated using chemical rationale. For illustration purposes, the blue node with *m/z* 409.2949 is cholate, *m/z* 446.3162 in purple is glycocholic acid (the user would discover this on the basis of MS^2 matches to a reference library), and the orange one is unknown but has a mass shift of 147.0723 Da compared with cholate. This is a typical mass shift of phenylalanine, and thus a prediction can be made that this is a phenylalanine conjugate of cholic acid. The difference between the glycine and phenylalanine conjugate is 90.0510 Da and supports this structural hypothesis. The self-looped teal node (*m/z* 244.1696) is attributed to an unrelated molecule and therefore does not have any structurally related molecule in the sample. Int., intensity.

and encourage the exchange of mass spectrometry data. MassIVE accepts data files (organized as datasets) and facilitates the sharing of datasets with a unique identifier; one can use this unique identifier as an accession number for publications. In addition, public datasets that the user publishes can, by choice of the depositor, have an associated digital object identifier (DOI). Currently, MassIVE is an approved repository for the *Journal of Proteome Research* (https://pubs.acs.org/journal/jprobs) and *Nature* partner journals (https://www.nature.com/sdata/policies/repositories#chem) and is widely used as a repository for other journals[14–23]. GNPS-MassIVE contains more than a thousand public metabolomics datasets. The GNPS knowledge base includes 221,083 reference MS^2 spectra



provided by the GNPS community, spectral libraries generated for GNPS (GNPS-collections), and third-party libraries[11]. Examples include LDB Lichen Database, MIADB Spectral Library[1,24,25], Sumner Spectral Library[26], CASMI Spectral Library[24,27], and MassBank (Japan (http://massbank.jp)[28], EU (https://massbank.eu/MassBank/)[29] and North America (http://mona.fiehnlab.ucdavis.edu/)), a large MS data library that is directly synced with GNPS. There are also tags and sample information (metadata) entries provided by the community in the GNPS knowledge base. Furthermore, all public data are periodically searched against the NIST 2017 spectral library and high-confidence spectral matches are annotated. GNPS-MassIVE now performs >6,000 analysis jobs and has >200,000 page views a month (excluding developers), with the predominant analysis being molecular networking. As a result, GNPS-based analysis has been used for the discovery of hundreds of new molecules in the past few years, ranging from immune regulators to antimicrobials, including antiviral agents and protease inhibitors[9,30–33]. Here, we provide a detailed protocol for generating a publishable and reproducible molecular network from a mass spectrometry dataset. This protocol will take the reader through the following steps: how to upload data, how to make the data public, how to subscribe to public data for living data updates, and how to reproducibly create publishable molecular networks using standardized sample information (metadata) through the GNPS infrastructure (Fig. 1); terminology associated with GNPS molecular networking is defined in Table 1. This protocol was previously made available as a preprint document[34].

Overview of the method

This protocol aims to provide researchers with a basic workflow for reliably and reproducibly creating molecular networks from mass spectrometry data using GNPS. The workflows for GNPS molecular networking described in this protocol can be performed in any laboratory with access to a tandem mass spectrometer, which is usually connected to an HPLC system. The overall procedure consists of five main stages (Fig. 2): (i) collecting MS^2 spectra, (ii) converting instrument-specific raw data files to an open format, (iii) uploading data to the MassIVE public repository, (iv) submitting a job to GNPS, and (v) visualizing the resulting molecular network. We also discuss how to navigate the 'My User' and 'Jobs' options, including how to share the links of a job, how to clone a job (Step 50), and how to subscribe to datasets in the public domain to obtain living data updates (Step 53).

Data collection and processing procedures will vary depending on the instrument available to the user. Although users can modify any procedure to fit their specific goals, this protocol specifies a set of starting parameters for acquiring and converting data with various mass spectrometers, including AB SCIEX, Agilent, Bruker, Shimadzu, Thermo Scientific, and Waters instruments. We also provide a protocol for the conversion of the data from each of these instruments to an open format (.mzXML, .mzML or .mgf) that is usable within the GNPS-MassIVE infrastructure. Once the data are converted to the proper open format, the protocol describes how to upload data files to MassIVE, a public repository that enables community sharing of mass spectrometry data, using either a web browser or an FTP client. The resulting datasets can subsequently be submitted to GNPS for molecular networking analysis, wherein MS^2 spectra are organized in a network according to similarity and compared against a reference database to identify putative known molecules and 'molecular families' in the samples. Finally, visualization and analysis of GNPS-generated molecular networks can be performed either in the web browser itself or in Cytoscape, an open-source software program for visualizing complex networks[35].

Applications of the method

GNPS molecular networking provides the ability to analyze and compare MS^2 spectra in one or more datasets acquired within the scope of a specific study, across datasets from multiple studies, and also to compare those datasets to all publicly available GNPS-MassIVE datasets, including community-curated spectral libraries. In addition, ongoing contributions to spectral libraries and submissions of new public datasets enable continuous identification: the periodic and automated reanalysis of all public datasets. GNPS is being used to network data acquired on a number of different mass spectrometers in a wide variety of exploratory studies, with samples originating from diverse environments and used for various purposes. The environments range from the indoor environment[36–38] to dissolved organic matter in the oceans[39]; from microbes in culture[9,40–43] to mouse[44] or human microbiomes[45,46] or infections[47–49]; from clinical samples[22,50,51] to plants[52], algae[53], sponges[5,54], and corals[55]; as well as a number of other sample types[16,56]. In addition, molecular networking has been applied to NPs discovery from a variety of organisms[57–60], forensics[61], small-molecule

Table 1 | Terminology for GNPS molecular networking

Term	Definition
Annotation	The process of attributing a putative chemical structure to a detected molecule. The level of annotations from spectral matches are considered level 2 or 3, according to the 2007 Metabolomics Standards Initative[120]
Bucket table	A tab-separated table (.tsv file format) downloadable from the GNPS interface that shows per-sample summed precursor ion intensities per MS^2 ion. Pie charts generated in visualization tools are based on intensities in the bucket table
Cluster index	Reference identification number for a MS^2 consensus cluster. In Cytoscape, this identification number is also called 'shared name'
Consensus cluster	A grouping of MS^2 spectra that are considered identical on the basis of the MS-Cluster algorithm[10,12]. Because GNPS brings together approaches from different scientific communities, there are terms, such as 'cluster', that have different meanings. Thus, the context in which the term is used should be considered. The term 'consensus cluster' refers to the grouping of MS^2 spectra into a node and is different from clusters of nodes in molecular networks as visualized in Cytoscape[128,129]
Cosine score	A value that represents the MS^2 spectral similarity between two nodes in a molecular network, where a cosine score of 1 represents identical spectra and a cosine score of 0 denotes no similarity at all. The cosine score takes into account precursor ion, fragment ions, and peak intensities[1]
DDA	Abbreviation for data-dependent acquisition, a method for tandem mass spectrometry data collection in which the most intense MS^1 ions are iteratively selected for MS^2 fragmentation[130]
Dereplication	Rapid identification of previously characterized (known) molecules[131]
Edge	A line connecting nodes that represents related but not identical MS^2 spectra on the basis of a cosine similarity score
Identification	Validation of a molecular assignment using an authentic chemical standard analyzed under the same experimental conditions as the sample containing the unknown compound. Molecular identification requires matching at least multiple physical characteristics, for example, retention time, exact *m/z* value, and MS^2 fragmentation pattern[120,132]
Natural product	A small molecule (generally <2,000 Da) produced by a biological source[133]
m/z	Mass-to-charge ratio, a dimensionless quantity resulting from dividing the mass number of an ion by its charge number[134,135]
Molecular network	A map of all nodes illustrating connectivity that represents the chemical space detected in the experiment
Molecular networking	A computational approach that organizes MS^2 data on the basis of spectral similarity, from which we can infer relationships in chemical structures[1]
MS-Cluster[8]	An algorithm used by GNPS to collapse nearly identical MS^2 spectra with the same precursor ion *m/z* into a single consensus spectrum
MS^1	The precursor ion(s) when detecting the intact molecular ions (*m/z*). MS^1 is the first stage of tandem mass spectrometry, in which compounds can be further fragmented[135,136]. See also tandem MS, MS^2
Node	A consensus cluster of identical MS^2 spectra, or a single MS^2 spectrum if cluster size is 1
Precursor ion (parent ion)	The ionized form of a molecule that is selected for tandem MS fragmentation. In electrospray ionization, the parent ion is a synonym for precursor ion[134,135]
Product ion (fragment ion)	An ion originating from a gas-phase reaction of the precursor ion[13]
Sample information (metadata)	Data that provide basic information about the sample and descriptions to facilitate data analysis and interpretation. Examples of sample information include the identification number, the source and origin of the sample collected, time, age, sex, longitude, latitude, altitude/depth, and date of collection
Small molecule	This protocol considers a molecule with a molecular weight <2,000 Da to be a small molecule
Spectral alignment	An algorithmic approach that aligns related spectra. This is the basis of molecular networking, which relies on the assumption that two structurally related molecules share similarity in their MS^2 spectra[1]
Spectral similarity	The likeness of MS^2 spectra on the basis of all or some of the following: precursor ion, fragment ions, and relative intensities of these peaks. Structurally related molecules tend to exhibit similar fragmentation[13]. In molecular networking, spectral similarity is calculated through a modified cosine score that considers the parent mass differences in the product ions for alignment
Summed ion intensities	Sum of precursor ion intensities in the MS^2 spectra for all ions with the same associated tandem mass spectrum detected by the mass spectrometer.
Tandem MS, MS^2	Abbreviations for tandem mass spectrometry, which defines a technique in which mass-selected ions are subjected to a second mass spectrometric analysis. In the first stage, also referred to as MS^1, precursor ions are formed and detected. In the second stage, also referred to as MS^2, precursor ions are fragmented, resulting in a spectral fingerprint[135,136]

identification[24,62], and biological discovery in hypothesis-driven research[63]. Furthermore, GNPS facilitates large-scale meta-analyses that can compare and potentially link studies from different laboratories by enabling rapid comparisons across multiple public datasets. Finally, to promote data

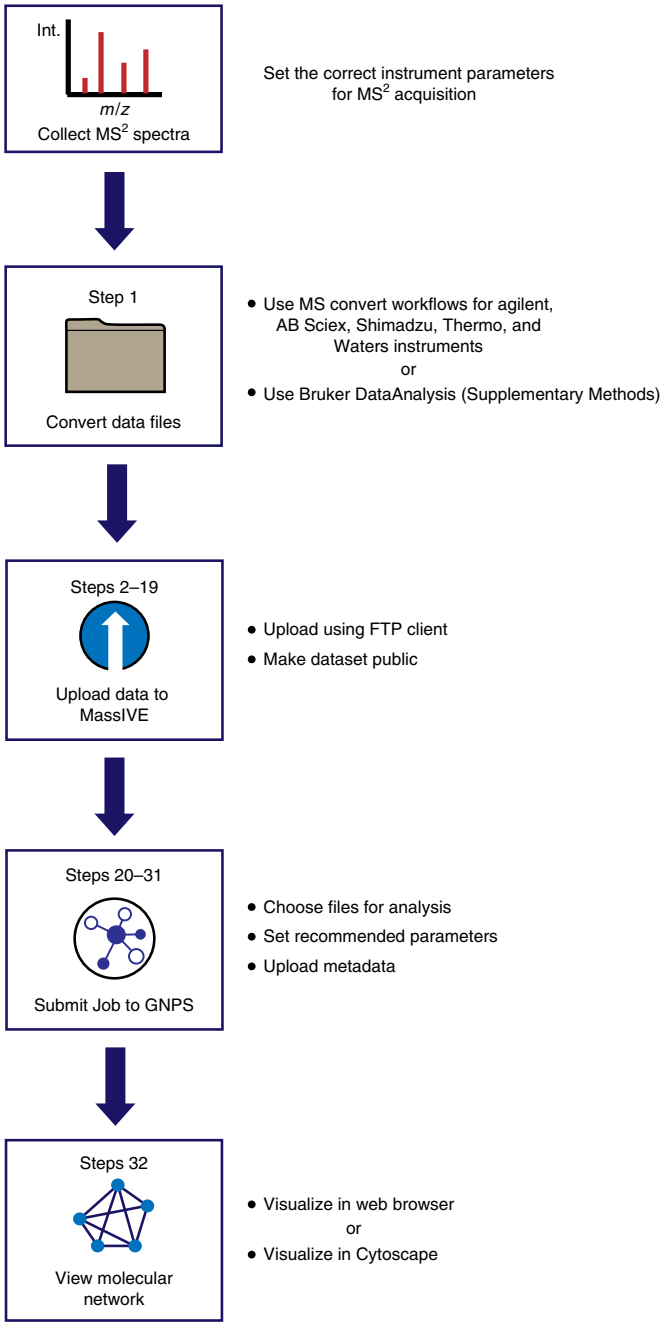


Fig. 2 | Flowchart of the protocol, delineating the workflow through Step 32 (Steps 33–53 address optional analyses, visualizations and sharing of data and molecular networks). The workflow comprises tandem mass spectrometry data acquisition, conversion, upload and networking to visualization. Readers following the example step-by-step instructions can follow these steps to generate a publishable network. Int., intensity.

analysis reproducibility, all analysis jobs are saved together with their parameters, which can be shared or cloned for reanalysis; no other platform provides this service.

Alternative methods

Several aspects of the GNPS-based molecular networking protocol are provided elsewhere but not to our knowledge as a coherent workflow in one package. There are several repositories to which metabolomics data can be uploaded[64–66]. According to the OMICS Discovery Index, the most widely used platforms are GNPS-MassIVE, Metabolomics Workbench[12] and MetaboLights[67,68].

Mass spectral library searching, or comparison of MS^2 spectra of compounds in a sample to reference data to annotate metabolites[69], has been implemented extensively, and successfully, for

decades. Finding analogs of small molecules via variable dereplication[70,71] or hybrid similarity search can be done via modified cosine correlations against spectral libraries. In this concept, originally introduced in 2012 for small molecules, a spectral alignment comparison also considers fragment ions in the alignment that differ by the same mass delta as the two parent ions[1,71,72]. Dereplication- and variable dereplication–based library search is a key part of the GNPS platform and can be performed without performing a molecular networking job[11].

Numerous commercial and non-commercial MS^2 reference databases exist, such as the NIST/EPA/NIH Mass Spectral Library[73]; METLIN[74]; the MassBanks of Japan (http://massbank.jp)[28], EU (https://massbank.eu/MassBank/)[29] and North America (http://mona.fiehnlab.ucdavis.edu/); mzCloud[75,76]; and ReSpect[77], which potentially provides users with access to ~2.4 million MS^2 reference spectra, when GC-MS and LC-MS reference spectra are both considered[64]. Many of these reference databases have an integrated spectral-matching tool for compound identification; these include mzCloud, METLIN/XCMS Online[78,79], Metabox[80], and MassBank. The goal of GNPS is not only to provide a spectral-matching tool, but also to serve as a data storage and knowledge capture/dissemination platform, as well as to provide access to a host of other analysis tools not covered in detail here, such as in silico-based dereplication[81–83], network annotation propagation[84], genome mining tools[85], and MASST (Mass Search Tool) searches.

GNPS is currently the only online platform that provides molecular networking, a computational tool that compares pairs of MS^2 spectra on the basis of their similarities and connects them to MS^2 reference spectral libraries. Molecular networking enables further propagation of annotations through mass spectral relations. MetGem[86] is a standalone software package that can be used for the generation of molecular networks, which works well for smaller datasets; it is not connected to a knowledge base, repository-wide analysis tools or the additional computational resources that GNPS provides.

Expertise needed to implement the protocol

Sampling and sample preparation, including sample extraction, should be performed by a trained analytical chemist, and mass spectrometry data should be acquired by a trained mass spectrometrist. It is imperative that the parameters for mass spectrometry be suitably optimized for the experimental conditions and sample type in order to generate meaningful molecular networks. Important instrument parameters to consider may include precursor isolation window, mass resolution, collision energy, data-dependent acquisition settings (e.g., duty cycle time and dynamic exclusion parameters), and the mass spectrometer has to be properly calibrated before use. Although an expert user will have preferred instrument parameters, recommended data acquisition parameters from major instrument manufacturers are provided in the Supplementary Information for newer mass spectrometry users who aim to create molecular networks in GNPS. Basic knowledge of MS^2 fundamentals, as well as knowledge of sample handling and preparation, is required to further optimize the data analysis parameters appropriate to the instrument used and the experimental design.

Experimental design

After running the molecular networking algorithm, GNPS creates a data table that can be visualized as a network of nodes and edges (Fig. 1) to provide chemical insights—including observed *m/z* values, *m/z* deltas between these values, and similarities between obtained MS^2 spectra—in relation to the metadata (associated sample information) provided by the user. Such data tables can be viewed as networks directly in the GNPS website or exported and manipulated in other data visualization tools and statistical analysis packages. Here we provide step-by-step instructions for molecular networking in GNPS (Supplementary Fig. 4) followed by export of the generated table into the most commonly used third-party network visualization tool (Cytoscape), which supports GNPS outputs. Notably, the information represented in and inferred from a molecular network is dependent on the input, including both the mass spectrometry data[87] and the networking parameters selected.

Reproducibility, blanks, and controls

A well-organized and well-thought-out experimental plan is essential to the successful creation of useful molecular networks, because molecular networks are only as meaningful as the experiment and data from which they originate. This includes providing sample information (metadata) tables and raw data files for the sample set (Box 1); metadata tables aid the creation of molecular networks that have increased interpretative value. To avoid pitfalls associated with large-scale mass spectrometry experiments, for example, batch effects[88], sample carryover and/or contamination[89], and high

Box 1 | Sample information (metadata) collation and input ● Timing typically 1–2 h for a small dataset; up to a few days for large complex metadata entries of large datasets

The inclusion of a metadata (sample information) table is extremely valuable for interpreting the molecular network that is generated using the data[138]. Although a time-consuming step, it is also one of the most valuable steps for interpreting the final molecular network. The more time spent on curating sample information (metadata), the more useful the resulting molecular network will be. The metadata table links the MS files uploaded and selected for molecular networking analysis in GNPS with various attributes of the collated data on the basis of the filename (such as 'Filename.mzXML'). For instance, the metadata table provides the necessary information to visualize the 'origin' of the detected metabolites when 'origin' is one of the attributes used in the metadata table (e.g., column heading: ATTRIBUTE_Origin). A metadata file can be created as follows.

Procedure

1 Prepare the metadata table as a text file (tab separated) with the text editor of choice (e.g., Microsoft Excel, Notepad++ for Windows; gedit for Linux; and TextEdit or TextWrangler for Mac OS) (Supplementary Tables 1 and 2).
- When uploading metadata associated with a GNPS job, specifically formatted column headers are required. The first column header must be 'filename' (no capitals (it is case sensitive) and no unusual characters such as '@', '#', '!', and no tab separations. Note that the filenames must be the filenames of the data (to be) uploaded to GNPS-MassIVE; otherwise, the metadata cannot be linked to the data. We recommend not using any special characters, such as'@', '#', '!', or spaces in any of the metadata fields.
- All the other column headers must begin with the text 'ATTRIBUTE_' before any header description (e.g., 'ATTRIBUTE_Origin') for downstream visualization.

2 For sample information (metadata) to be incorporated into global meta-analyses, the template provided in Supplementary Table 2 should we used and labeled 'gnps_metadata.tsv'.

There are a number of advantages to uploading a metadata table associated with a GNPS job. When the network generated after data processing is subsequently opened in Cytoscape, the nodes of subnetworks can be visualized on the basis of their associated metadata. This can be represented as a pie chart contained within each node. In addition, metadata can be used to color-code categories of samples when visualizing the MS^2-based statistics, such as PCoA, in a browser, using the EMPeror package[137] available in Qiime2[139]. This allows the user to quickly attribute the molecular differences of the samples to certain characteristics found in the metadata. For example, if two distinct groups appear in the PCoA plot, it would then be possible to color all samples of type one blue and all samples of type two red in order to determine if this attribute could be responsible for the separation. However, it is important to understand that PCoA is only visual and does not give any statistical support; a PERMANOVA analysis would have to be performed in order to actually test whether an attribute is responsible for separation. Finally, data sharing is a vital part of modern science because it offers opportunities for collaboration and wider-scope analyses, and transparency promotes reproducibility and thus scientific rigor. Without metadata attached, public data have less value, will not be discovered as easily by others, and will not provide meaningful results with MASST[140] or ReDU[141]. ReDU is a metadata text-based search in GNPS that facilitates reanalysis of and comparisons across all public data files with specifically formatted metadata. Therefore, we encourage protocol users to use the ReDU metadata template (Supplementary Table 2). When no metadata are available, these public data will not be included in MASST or ReDU searches, making the public data less useful to the community. To ensure that the metadata are compatible with existing infrastructure and use identical vocabularies, there is a drag-and-drop validator of the metadata within ReDU. In short, the visibility and value of data increase by improving the amount of metadata that are uploaded. Therefore, uploading metadata associated with the MS data to GNPS promotes a more universal approach to science.

3 In cases in which you want to add a new/external metadata file (tab-delimited text format) to your workspace, under the 'Upload Files' tab select the destination folder for the upload on the left and drag the file for upload to the 'File Drag and Drop' box on the right before following the same actions listed in this step. The online tutorial on metadata formatting, which includes a template file, can be accessed at https://ccms-ucsd.github.io/GNPSDocumentation/networking/#metadata.

Metadata format for 'ili

'ili (ref. [113]) enables molecular cartography, or the detection and mapping of molecules in 2D or 3D, in GNPS; for molecular cartography using 'ili, metadata must contain the following additional information. The spatial coordinates that dictate the spatial distribution of a detected metabolite in a 2D (.png format) or 3D image (.stl format) must be included. In addition to the column 'filename', extra columns containing the following information must be included: 'COORDINATE_x', 'COORDINATE_y', 'COORDINATE_z', 'COORDINATE_radius'. The 'x', 'y', and 'z' correspond to the 3D coordinates, and the radius corresponds to the approximate values of radii of the sampling points. An image viewer can be used to estimate this value; for example, half of the difference between boundaries of a sampling point in a horizontal or vertical dimension can be estimated. Additional information related to 'ili can be obtained from https://github.com/MolecularCartography/ili.

background signal[90], and to maximize reproducibility and signal-to-noise ratio[91], a dataset should include blanks, quality control (QC) samples, and experimental replicates. Dunn et al.[92]. describe an appropriate representative experimental design in detail that includes blanks, QC mixtures, and samples, plus internal standards. Petras et. al.[36] provide an example that illustrates control metrics, including evaluation of QC mixtures and signal deviation of the internal standard.

We recommend preparing control samples using exactly the same protocols and experimental conditions used to prepare the test samples (i.e., the same types of tubes, the same batches of tubes, the same extraction solvent, extraction time, sonication time/power, and so on). These blank samples inform which ions come from the experimental conditions, and they can be subtracted from test sample signals in the molecular networking analysis (see Step 32B(xii)). The requirements for QC associated with a broad assessment of the NP composition of an extract library used in bioactivity screens are different from a detailed clinical study for biomarker discovery. When possible, one should add internal standard(s) to each sample to ensure that the system performs consistently. If the internal standard(s) are not within the user-defined limits of chromatography variability, the sample needs to be either removed from downstream analysis or rerun. This is particularly useful in applications in which

thousands of samples, such as NP extract libraries, are screened. Further, when acquiring data for many samples, especially when multiple batches are used, we advise acquiring data for additional QC samples to monitor batch and plate effects throughout the experiment to assess instrumental variation over time, such as retention-time drift. QC samples can either consist of aliquots from a subset of test samples pooled together (pooled QC) or can be mixtures of molecules specifically defined for quality assurance. For example, it is common to use the last column of a 96-well plate for the QC mixture to ensure that the instrument and chromatography behave in an identical fashion throughout an experiment. Finally, data from experimental replicates, including both technical and biological replicates, should be acquired in a randomized fashion. This is especially important for large-scale population studies to ensure minimized bias. One common problem in metabolomics and LC-MS analysis is sample carryover, which is caused by residual compound(s) from a previous run. One way to reduce this issue is to insert a wash routine between samples, followed by a blank to ensure that no carryover is observed.

Molecular networking parameters

GNPS-based molecular networking parameters—the set of user-defined and/or pre-populated values selected for analysis—can be varied substantially and need to be set appropriately for the acquired dataset, on the basis of the sample (anticipated molecular masses and types of molecules), instrument resolution, and collision energies used for MS acquisition. Networking parameters and basic mass spectrometry/metabolomics terminology are described in detail in Table 2 and the Procedure and should be considered and selected carefully in order to obtain useful networks, which ultimately depend on the quality and quantity of MS^2 spectra.

Limitations and challenges

Because GNPS-based molecular networking uses MS^2 data, it is susceptible to the same challenges encountered in any mass spectrometry data acquisition experiment, such as low signal-to-noise ratio, insufficient separation of analytes, and poor peak shape[93,94]. In addition, classic molecular networking can provide only qualitative information about the experiment because only MS^2 scans are considered in the analysis. Although feature-based molecular networking (Box 2) incorporates MS^1 and chromatographic data, which approximates quantitation, it is still not strictly quantitative. If calibrated quantitative information is needed to answer the scientific question, follow-up experiments should be performed using targeted LC-MS.

One should consider potential issues that accompany metabolomics experiments, such as sample extraction efficiency and reproducibility, as well as unwanted metabolite degradation. Although avoiding degradation or modification of all molecules in a sample is impossible, it is important that all samples for comparison be prepared and analyzed in an identical manner, unless the goal is to understand the effects of sample preparation conditions[95]. Although a few publications describe the impact of storage on the detectable metabolome, these are sample type specific and there is currently no consensus for a 'gold standard'[96–98]. Ultimately, sample preparation is highly dependent on the type of sample collected and includes drying, homogenization, and extraction steps[99]. Although each lab has its own preferences for sample treatment, we strongly advocate for samples to be collected and extracted with solvent as soon as possible. The speed of this is dependent on the experimental environment. For example, samples collected in remote areas, at sea using a small boat, or even in a clinical setting, may be stored for hours or days before they can be extracted, given that some solvents are not easily brought into a clinical setting or used while out at sea. By contrast, samples from a cultured system in a lab or an enzymatic reaction, for example, can be halted in milliseconds using a rapid quench system and can then be extracted in seconds. The choices of solvent and extraction protocol are dictated by the experimentalist's interests and questions. Although there is always overlap among the molecules from even very different extraction protocols, more polar metabolites are extracted with ethanol, methanol, or butanol, whereas more hydrophobic metabolites are extracted with benzene, ethyl acetate, or chloroform[95]. The samples can then be introduced into the mass spectrometer using front-end separation techniques, most often liquid chromatography or ion mobility spectrometry. If mass spectrometry cannot be performed immediately, we recommend completely drying the samples at cryogenic temperatures before storage.

In addition, molecular networking and spectral matching against libraries are challenging when few product ions are available. Although most precursor ions that are observed can be fragmented, some may have too few fragments to reliably network. We advise using caution when looking at spectra with <4 fragment ions, on the basis of false-discovery rate (FDR) estimates of spectral

Table 2 | Parameters for molecular networking in GNPS

Fillable field	Definition	Recommended user input
Advanced network options		
Min Pairs Cos	Minimum cosine score required for an edge to be formed between nodes	Most commonly set to 0.7 when a minimum of 6 ions are matched. When fewer ions are used, it is better to be more stringent and increase this value (e.g., 0.8), but when more ions are required, one can relax this value (e.g., 0.6)[100] Use 0.7 for example MSV000083437
Minimum Matched Fragment Ions	Minimum number of common fragments that must be matched by two nodes for an edge to be formed	This is highly dependent on the experiment. Although 6 is listed as default, a lower value could be used if the user wants to be less restrictive or if the sample largely contains molecules with a small number of fragment ions. The maximum number of significant annotations are found when this value is set to 4 or 5[100] Use 4 for example MSV000083437
Network TopK	Maximum number of neighbor nodes for one single node. The edges between two nodes are kept only if both nodes are within each other's TopK most similar nodes. If this value is set to 10, a single node may be connected to up to 10 other nodes	Default is 10. Adjusting this value enables the network to be more or less stringent. Keeping this value low makes very large networks (many nodes) much easier to visualize Use 10 for example MSV000083437
Minimum Cluster Size	Minimum number of identical MS^2 spectra that are merged by MS-Cluster for the consensus spectrum to be represented as a node	This is a very important parameter because it is a very good filter for quality of spectra. If this is set to 1, then each MS^2 spectrum is compared to all other MS^2 spectra, including MS^2 spectra of noise, thus increasing the computational time and exploding the final molecular network. By requiring more identical spectra to be merged (clustered) before considering the MS^2 spectral alignments, it will ensure that only reproducible and higher-quality data are used in the final molecular network. The default is 2, but if it is a very large dataset (hundreds to thousands of files) one can use 5 or more, whereas for smaller datasets (e.g., 1 or 2 files) it can be set to 1 or 2 Use 4 for example MSV000083437
Run MSCluster	Clusters MS^2 spectra and creates consensus MS^2 spectra using the specified mass tolerance settings	Set to 'yes' for classic molecular networking Set to 'yes' for example MSV000083437
Maximum Connected Component Size (Beta)	Maximum number of nodes that can be connected in a single component (molecular family) of a molecular network. This process iteratively breaks up large 'hairball' networks (of false positives) by removing the lowest-scoring alignments (by cosine score) first, until the resulting pieces fall below the maximum size	Default setting is 100; this value can be set to 0 to allow for an unlimited number of nodes, or a higher setting can be used for larger datasets or for datasets containing many structurally related molecules. Use 100 for example MSV000083437
Metadata File (= sample information file)	File added to the analysis that describes the experimental setup and details to allow for better downstream data visualization, analysis, and interpretation	Add as a .txt file that follows the template and instructions available in the supporting information. Metadata file uploaded is described in Step 13. Example metadata can be found in Supplementary Tables 1 and 2, and a description of how to create a metadata file can be found in Box 3
Group Mapping and Attribute Mapping	Legacy version of metadata file	We advise using the metadata table instead, as described in Box 3
Advanced library search options		
Library Search Min Matched Peaks	Minimum number of shared fragment ions to make a library match	The default value is 6. This is dependent on the aim of the experiment: a lower value may yield more tenuous matches to library spectra, which is suitable for exploratory structure searching; a higher value, selecting for closer matches, facilitates dereplication of putative known compounds. The impact of this parameter is discussed in Scheubert et al.[100] Use 4 for example MSV000083437
Score Threshold	Minimum cosine similarity score to make a library match	The default setting is 0.7. This is dependent on the aim of the experiment: a lower value may yield more tenuous matches to library spectra, which is suitable for exploratory structure searching; a higher value, selecting for closer matches, facilitates dereplication of putative known compounds. Use 0.7 for example MSV000083437
Search Analogs	Matches query spectra against library spectra with a modification-tolerant search within a	Dependent on the user's preferences, selecting 'Do Search' requires more computing time, but the results are more

Table continued

Table 2 (continued)

Fillable field	Definition	Recommended user input
	specified range for mass differences. Precursor ion *m/z* values are allowed to deviate up to a user-defined maximum. Fragment ions that differ by the mass difference of the two parent ions are also considered	exploratory. It allows for dereplication not only of identical molecules, but also of related molecules
Maximum Analog Search Mass Difference	Maximum mass shift allowed between the query spectra and library spectra *m/z* values to make a library match	Use default parameter of 100 Da: the user can increase or decrease the value depending on properties such as anticipated molecular mass shift of related molecules in the samples (e.g., 14 Da for CH_2 for methylations, amino acid substitutions, or different fatty acid chain lengths; 16 Da for oxidation of mass difference between Na^+ and K^+ adducts; 162 Da is a common mass shift for oligosaccharides). The larger this value, the more likely spurious matches will be found
Advanced filtering options		
Filter Below Std Dev	Applied before MS-Cluster. For each MS^2 spectrum, the 25% least intense fragment ions are collected and the standard deviation is calculated, as well as the mean. A minimum peak intensity is calculated as mean + k × s.d., where k is user selectable. All peaks below this threshold are deleted. By default, this filter is inactive (value is set to 0)	Using this filter is not recommended. A default value of 0 should be used so that no filter is applied
Minimum Peak Intensity	All fragment ions in the MS^2 spectrum below this raw intensity will be deleted	This filter is infrequently used. Use a default value of 0 so that no filter is applied, especially if the raw intensities of your data are very low
Filter Precursor Ion Window	All peaks in a ±17 Da range around the precursor ion mass are deleted. This removes the residual precursor ion, which is frequently observed in MS^2 spectra in the comparison of all spectra for molecular networking	Apply this filter, which is the default option
Filter Library	Applies the above precursor ion window filter to the library as well	Apply this filter, which is the default option
Filter Peaks in 50 Da Window	Removes peaks that are not one of the top six most intense within a ±50 Da window	This is commonly turned on. It is dependent on the dataset: if samples contain a large number of low-mass molecules or are complex mixtures containing compounds of low titer, this filtering should be turned off, because it may filter out relevant peaks that could be signals

Box 2 | Feature-based molecular networking

The molecular networking analysis described in the Procedure represents the type of molecular networking that is currently most widely used. This workflow connects clustered MS^2 spectra as nodes on the basis of spectral similarities and makes use of MS^2 data only, even for quantitation. The chromatographic dimension and MS^1 data are not considered in classic molecular networking.

However, in MS-based metabolomics studies, statistical analysis is done predominantly from MS^1-based peak abundances from extracted ion chromatograms (XICs). Those chromatographic peaks with a specific, accurate mass-to-charge ratio are described as features. To bridge this gap between MS^1 abundance and MS^2 qualitative information, there is a workflow to link MS^1 peak areas derived from LC-MS features with MS^2 information from molecular networking[142,143]. This workflow is called feature-based molecular networking (FBMN; https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking/) and can be performed using open-access mass spectrometry processing tools such as MZmine 2[115], XCMS[78], MS-DIAL[144], and OpenMS[114].

In this workflow, feature finding is the computational process of selecting and identifying features at the MS^1 level across multiple samples and must be performed before generating a network. These tools allow the export of a feature table and corresponding MS^2 scans for each feature, which can be submitted for FBMN through GNPS.

matching against libraries[100]. Common methods of ion activation coupled with molecular structure for certain classes of molecules may inherently result in too few fragments for confident molecular networking. In such cases, it is advisable to find alternative fragmentation methods[101] and/or improve the gas-phase reactivity by performing chemistry on the sample before subjecting to LC-MS^2 (ref. [102]).

To annotate unknown molecules, GNPS queries MS^2 spectra against MS^2 data in reference libraries and assigns a cosine score based on their similarity. For the GNPS spectral library, MS^2

spectra are acquired from laboratories around the world using a variety of mass spectrometers and sample preparation protocols. Therefore, mass spectra submitted to GNPS can differ in terms of both quality and content. For instance, MS^2 fragment ions and their intensities can vary greatly between instruments, and even on the same instrument, if the experimental setup is changed[103]. GNPS requires that the instrument and ion source be specified with each reference spectrum submitted, and it is recommended that this be taken into account when assessing the quality of a library hit. Along these lines, annotations of unknown molecules are not all accurate and should be considered putative until confirmed with an authentic chemical standard.

On average, in 2016 when GNPS was published, only 2% of spectra in an untargeted mass spectrometry metabolomics experiment were annotated[104]. Although this percentage has grown to an average of 5–6% of spectra being annotated, a large percentage of MS^2 spectra typically remain unannotated. The structures of these unannotated molecules or 'dark matter'[105] might be known, but their identity is not revealed because no reference spectra exist in library databases against which to compare. To improve annotation rates, in silico tools have been developed to match unknown MS^2 spectra to putative chemical structures[106]. Several of these computational tools, which include MetFrag[107], MetFusion[108], SIRIUS[109,110], CSI:FingerID[111], MS-Finder[112], Network Annotation Propagation (NAP)[84], and Dereplicator[81,82], can be integrated into GNPS molecular networking workflows to provide insight into the annotation; the application of such tools is beyond the immediate scope of the networking protocol presented here.

Materials

▲ **CRITICAL** Materials and methods related to sample preparation and LC-MS analysis are described in the Supplementary Methods.

Software

- MSConvert tool from the ProteoWizard (http://proteowizard.sourceforge.net/downloads.shtml)
- AB SCIEX MS Data Converter (Beta 1.3) is freely available for download from the AB SCIEX website (https://sciex.com/software-support/software-downloads)
- AB SCIEX Analyst Software 1.7 is available for download, trial license use, and purchase from the AB SCIEX website (https://sciex.com/products/software/analyst-software)
- Agilent MassHunter (https://www.agilent.com/en/products/software-informatics/masshunter-suite/masshunter/masshunter-software)
- Bruker DataAnalysis (www.bruker.com/service/support-upgrades/software-downloads/mass-spectrometry.html)
- Shimadzu LabSolutions (https://www.ssi.shimadzu.com/products/liquid-chromatography-mass-spectrometry/lcms-software.html)
- Thermo Scientific Xcalibur (https://www.thermofisher.com/order/catalog/product/OPTON-30965#/OPTON-30965)
- Waters MassLynx MS software (http://www.waters.com/waters/en_US/MassLynx-MS-Software/nav.htm?locale=en_US&cid=513662)
- FTP client (e.g., WinSCP for Windows (https://winscp.net/eng/download.php); Cyberduck for Macintosh (https://cyberduck.io/download/))
- Web browser (Firefox or Google Chrome to access GNPS)
- Cytoscape for data visualization: (https://cytoscape.org/ (current version at the time of writing was 3.7.1; current version at time of publication is 3.7.2))
- Software relevant to optional pipelines (e.g., 2D or 3D Visualization;[113] feature-based molecular networking (Box 2))
- OpenMS TOPPView (https://github.com/OpenMS/OpenMS/releases)[114]
- MZmine2 (https://github.com/mzmine/mzmine2/releases)[115]

Example datasets

▲ **CRITICAL** All LC–MS data used in this paper are publicly available at the GNPS-MassIVE repository under the following accession numbers:

- MSV000083437 (germ-free (GF) and specific-pathogen-free (SPF) mice)
- MSV000083359 (3D cartography of diseased human lung[47])
- MSV000083381 (stenothricin-GNPS analogs[11])

Procedure

▲ **CRITICAL** The data submission and molecular networking workflow can be followed as a tutorial using an untargeted metabolomics dataset for 3D molecular cartography of the mouse duodenum (data not shown; MassIVE dataset MSV000083437). This dataset is a subset of a collection of metabolomes analyzed from organs of GF and SPF mice that led to the discovery of new amide-conjugated bile acids made by bacteria that affect host metabolism via farnesoid X receptor (FXR) agonism. The following procedure will take the reader through submission of dataset MSV000083437 to the molecular networking workflow in GNPS, through the molecular networking workflow in GNPS (including input parameters), and through visualization of the generated network using both in browser and Cytoscape-based visualization (Fig. 3).

▲ **CRITICAL** All steps, albeit in less detail, are also described and continuously updated and maintained in the online GNPS documentation at: https://ccms-ucsd.github.io/GNPSDocumentation/. Documentation should be regularly checked for the most up-to-date information, as well as descriptions of new software releases and features.

Data conversion ● Timing 1 h up to a few days (depending on size of dataset and computer setup)

1 Manually convert the raw data to open file formats prior to uploading to GNPS-MassIVE (Fig. 2, stage 1). The protocol for data conversion should be chosen on the basis of the instrument used for mass spectrometry acquisition. Mass spectrometry files must be converted to an open file format for analysis in GNPS. Open file formats include .mzXML, .mzML, and .mgf formats, with the preferred formats being .mzXML and .mzML, although it is also encouraged to co-submit the raw data to MassIVE. MSConvert can be used for the conversion to a GNPS-compatible format of mass spectrometry data acquired on AB SCIEX, Agilent, Shimadzu (after initial conversion, Supplementary Methods), Thermo Scientific, and Waters instruments. For Bruker files, a separate workflow must be used that applies internal lockmass calibration to the output file. This Bruker workflow is described in more detail in the Supplementary Methods. For AB SCIEX raw files (.wiff files), the data can also be converted into .mzML format using the MS Data Converter (AB SCIEX v.1.3 beta, freely available at https://sciex.com/software-support/software-downloads). There are two options for file conversion using MSConvert: the 'traditional' workflow (option A) and the batch workflow (option B).

(A) **'Traditional' data conversion workflow**

(i) Download the free software MSConvert from ProteoWizard at http://proteowizard.sourceforge.net/download.html. This software is compatible with Windows and Linux operating systems but is not supported for Mac OS. When downloading ProteoWizard, the version of Windows must be specified and .NET Framework 3.5 SP1 and 4.0 must be installed. MSConvert is the recommended software for conversion of data acquired on AB SCIEX, Agilent, Shimadzu, Thermo Scientific, and Waters instruments.

(ii) In the 'Start' menu, select the ProteoWizard folder and open MSConvert.

(iii) To select file(s) for conversion, click 'Browse'; then click 'Add' to add file(s) to the workflow and select a directory for the output.

(iv) To convert the vendor file format to an .mzXML file, select .mzXML under 'Options'; select '32-bit' for binary encoding precision and make sure 'Use zlib compression' is unchecked.

(v) Choose 'Peak Picking' under the 'Filters' heading, and under 'Algorithm', check 'Vendor', then type in 'MS-Levels 1-2', and finally add the filter by clicking 'Add'.
▲ **CRITICAL STEP** Move the peakPicking filter to the top of filter list. The peakPicking filter must be the first filter in the list or the output file will not be centroided.

(vi) Click 'Start' and then check the folder in the 'Output Directory' for the .mzXML files. Multiple software programs can be used to verify that the conversion process has worked properly; these include SeeMS (installed with MSConvert), OpenMS TOPPView, and MZmine2.

(B) **Batch conversion workflow**

(i) There is a simple batch conversion method that includes a complete package that enables Windows users to convert vendor formats to GNPS-compatible formats (.mzXML, .mzML, .mgf). Refer to the Supplementary Methods and https://ccms-ucsd.github.io/GNPSDocumentation/fileconversion/ for additional online data conversion tutorial information.
? TROUBLESHOOTING

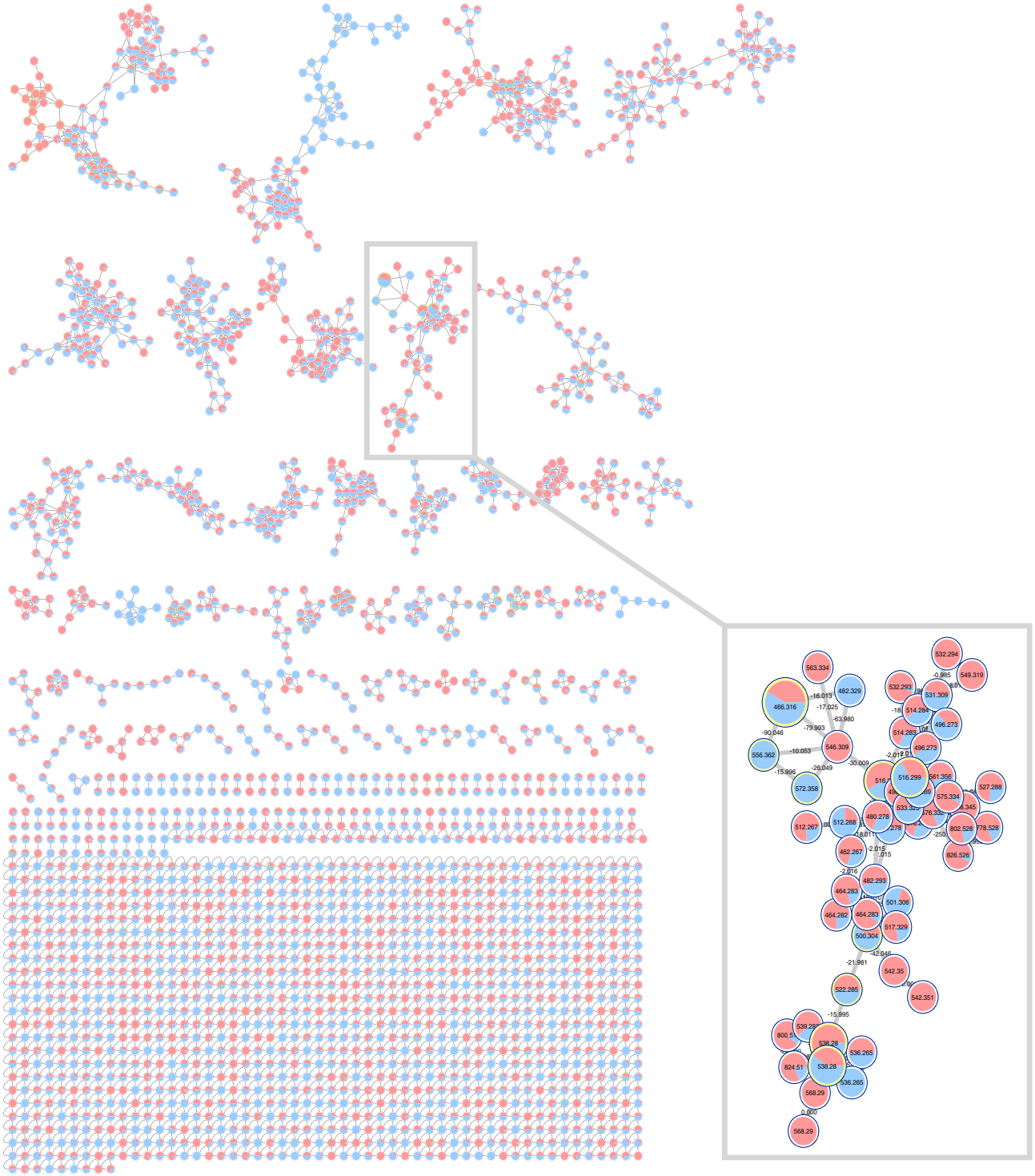
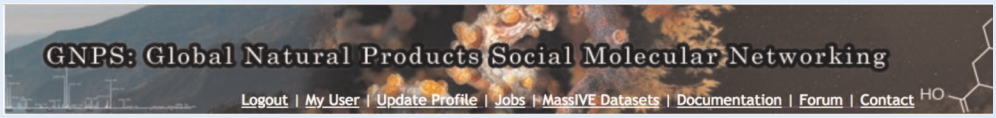


Fig. 3 | Mouse duodenum global molecular network created from MassIVE dataset and visualized in Cytoscape. Protocol users can recreate this network by following the Procedure and using the MassIVE dataset MSV000083437. Pie charts represent relative summed precursor ion intensities per MS^2 spectrum detected within each metadata group: red for germ-free (GF) and blue for specific-pathogen-free (SPF) mice. The box highlights a cluster we examine below in terms of chemical interpretation. Each circle is a pie chart representing the number of MS^2 spectral counts per group (GF versus SPF mice) and each group of nodes is a subnetwork of related molecules.

Data submission to GNPS/MassIVE ● Timing 10 min

▲ **CRITICAL** Create an account with GNPS in order to submit datasets and create workflows, as well as to receive emails about the outcomes (Supplementary Fig. 1). Making a GNPS account automatically sets

Box 3 | Navigating the user workspace portal



At the top of the GNPS website, users will find a banner that allows them to navigate their personal workspace and access additional resources such as the help forum and molecular networking documentation. Within this space, the 'My User' tab provides a way to view all MassIVE datasets and reference spectra deposited by the user, and the 'Jobs' button allows easy access to all jobs submitted by the user through the GNPS and MassIVE interfaces. Clicking on 'MassIVE datasets' allows the user to browse and subscribe (Step 53) to all public MassIVE datasets with GNPS in the title. In addition, this banner is a portal to all resources for help while using GNPS. The 'Documentation' link in the banner takes the user to the GNPS documentation website, which has additional step-by-step instructions and links to supplemental tutorial videos, as well as access to the 'legacy' documentation (from a menu on the right-hand side of the page) that can provide additional information to the user. The 'Forum' button opens a Google groups forum where users can post questions, have discussions, and report potential bugs. The corresponding online tutorial can be. accessed at https://ccms-ucsd.github.io/GNPSDocumentation/quickstart/.

up a MassIVE account that uses the same login and password. To manipulate MS data files in GNPS, they must first be uploaded to MassIVE, which is an online repository for mass spectrometry datasets hosted by the UCSD Center for Computational Mass Spectrometry (CCMS). The user workspace in GNPS/MassIVE provides a personalized location where researchers can curate mass datasets, submit and monitor GNPS workflows, subscribe to datasets that have been made publicly available by others, or clone and reanalyze either their own or other public datasets. More information on subscriptions to data can be found in Step 53.

2 Open a web browser. GNPS is designed to work with Firefox or Google Chrome but also works in Microsoft Edge, Safari, and Opera.

3 Navigate to the GNPS home page by using the following link: https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp

4 Click on 'Register New Account' (right-hand gray box), toward the top center of the page, above the large GNPS logo.

5 Enter a username, name (optional), organization (optional), email, and password (twice for confirmation) in the spaces provided on the new page that loads.

6 Click 'submit'.

7 Sign in to your new GNPS account (http://massive.ucsd.edu/ProteoSAFe/). Once login is successful, navigate to the home page by clicking the user workspace banner (Box 3).

8 Check that your GNPS credentials work for logging in to MassIVE.

Deposit data files into the MassIVE web server using an FTP client ● Timing 30 min

Choose FTP client and log in

9 Choose an FTP client. A number of options are available for free dedicated FTP clients; of these, the following are popular choices that have been tested with MassIVE: WinSCP, CoreFTP (http://www.coreftp.com/), and CoffeeCup Free FTP for Windows (https://www.coffeecup.com/free-ftp/), and Cyberduck or FileZilla (https://filezilla-project.org/) for Macintosh. Additional online tutorial materials on how to submit a dataset to MassIVE can be accessed at: https://ccms-ucsd.github.io/GNPSDocumentation/datasets/#submitting-gnps-massive-datasets. See also Supplementary Fig. 3.
! CAUTION When downloading an FTP client for use, make sure it comes from a trusted source to avoid malware. Data files transferred to MassIVE should be in .mzXML, .mzML, or .mgf formats. The data that are uploaded should not be in a file archive (e.g., .zip, .tar) format. We also recommend that the original vendor raw data files (e.g., .wiff for AB SCIEX, .yep for Agilent, .d for Bruker, .lcd for Shimadzu, .raw for Thermo Scientific) be uploaded together with the open formats as described below.

10 Log in to the FTP server with the host name 'massive.ucsd.edu' using your MassIVE web account username and password in the FTP client program for FTP file transfer. Most FTP clients use this 'Quick Connect' feature. Alternatively, type in the FTP server name, username, and password, and then connect directly.

Run the MassIVE dataset submission workflow on the uploaded files

11 Load the home page for MassIVE from the GNPS home page by scrolling down to the GNPS-MassIVE datasets section and clicking on the 'Deposit dataset' bar in the 'Create Public datasets' block. Alternatively, click on the 'Submit your data' link in the paragraph titled 'Submit Data' on the MassIVE home page. A direct way to deposit the data is to navigate directly to the MassIVE home page (http://massive.ucsd.edu/ProteoSAFe/). This will bring up the Dataset Submission workflow input form, on which there are various numbers of fillable fields under each of the sections described in Steps 12–18.

The reader can follow along (Supplementary Fig. 3) because the mass spectrometry output files for the tutorial example have already been uploaded to MassIVE and can be found in the dataset MSV000083437.

12 In the 'Workflow Selection' section, enter a title for your dataset, noting that GNPS datasets must have a 'GNPS' prefix in the title in order for these GNPS-MassIVE datasets to be visible to GNPS users.

To satisfy this requirement for the dataset that reader will use when following the example, MassIVE dataset MSV000083437 has been titled 'GNPS Example Dataset_GF vs. SPF Mouse Duodenum.'

▲ **CRITICAL STEP** Adding GNPS to the title is therefore absolutely necessary for the dataset to become a part of the community, ensures that the data become 'alive' (Step 53), and enables subscriptions and other analysis features specifically used for the GNPS community (Step 53). If a 'GNPS' tag is not added at the beginning of the title, it will not be part of the GNPS analysis infrastructure. Currently all of MassIVE has almost 11,000 public mass spectrometry datasets (mostly proteomics), ~1,100 of which are also part of GNPS. If GNPS is not added at the beginning, it is possible to go to MassIVE, log in, and edit the title at a later time.

13 In the 'Dataset Metadata' section, enter relevant metadata. To minimize the burden of making datasets for GNPS analysis and to enable as much flexibility in what additional information the user wants to make available, very few metadata fields are absolutely required, although the user is encouraged to provide as much metadata as possible. It should be noted that the datasets that have the most information associated with them are also the datasets that are the most visible to the community. Fields for metadata relevant to the dataset being submitted are listed in Table 3. The first three fields ('Species', 'Instrument', and 'Post-Translational Modifications') are backed by lists of standardized controlled vocabulary (CV) terms, maintained by organizations such as the HUPO Proteomics Standards Initiative[116] and many others, that the user can implement[116,117]. To search these terms, type at least three characters into any of these text boxes, and a drop-down list of supported terms that match your query will be displayed. To select a term, click on it in the drop-down list and it will be added to your dataset. If the term you want is not present in the list, you can type your custom text into the text box and click the adjacent 'Add' button to tag your dataset.

Metadata (sample information) for MassIVE dataset MSV000083437 has been added as shown in Supplementary Fig. 2b.

▲ **CRITICAL STEP** Using the official CV to tag your dataset greatly increases the likelihood that it will be found and processed correctly by any automated software that may interface with the MassIVE repository.

14 Add data files to the 'Dataset File Selection' section. In this section, there are 11 different file types that can be added, and these are organized into three different categories: required, recommended, and optional. Most of these file categories are not strictly required. The only official file requirement for a MassIVE dataset is that at least one file is submitted in either the 'Raw Spectrum Files' or 'Peak List Files' categories. If a submitted dataset does not meet the additional requirements for a 'complete' submission, then it is considered 'partial', which is currently standard for small-molecule datasets that are a part of GNPS (link for the definition of complete submission: http://proteomics.ucsd.edu/service/massive/documentation/submit-data/submission-workflow/#MassIVEDatasetSubmission-SubmissionTypes). Guidelines for which files to upload are summarized in Box 4.

For readers who are following the example, Peak List files were uploaded previously for dataset MSV000083437, as illustrated in Supplementary Fig. 3b, where nine folders (Control, GF1, GF2, GF3, GF4, SPF1, SPF2, SPF3, SPF4) have been added.

15 Check that each spectrum (data) file referenced within a 'Result File' is associated with a file from the 'Peak List Files' category; this is required for a submission to qualify as 'complete'. 'Mapping Spectrum Files to Identification Files' is not necessary for small molecule workflows. This section is where these two types of files are associated with each other as appropriate.

Table 3 | Metadata categories for data upload to MassIVE

Metadata category	Required	Notes	Example dataset MSV000083437
Species	Yes	Enter custom text if the correct species for your dataset is not supported in the list or if your sample is not a specific species (e.g., environmental sample or community of organisms)	*Mus musculus* (house mouse)
Instrument	Yes	Enter custom text if the correct instrument for your dataset is not supported in the list	maXis
Post-Translational Modifications	Yes	For small-molecule metabolomics datasets, the appropriate entry in the drop-down list is: 'PRIDE:0000398, No PTMs are included in the dataset'	No PTMs included in the dataset
Keywords to assign to your dataset	Yes	Your dataset must be tagged with at least one keyword; there is no limit. Keywords are custom text, so you must click the 'Add' button after entering text	mouse duodenum
Principal Investigator	Yes	Needed to identify the lab providing the data	Pieter Dorrestein (pdorrestein@ucsd.edu) UCSD, United States
Description	No	We recommend providing as much detail as possible	Not applicable

Box 4 | Guidelines for files submitted to MassIVE

Recommended for all submissions
- *Raw spectrum files.* Raw mass spectrum files in a nonstandard or instrument-specific format, such as AB SCIEX .wiff files, Agilent .yep files, Shimadzu .lcd files, Bruker .d files, Thermo Scientific .raw files, and Waters .raw files.
- *Peak list files.* Processed mass spectrum files in a standardized format. The following formats are recognized by MassIVE as valid for this category: .mzXML, .mzML, and .mgf. This is the file from which GNPS analysis is enabled.

Strongly encouraged for submissions to improve the ability to interpret the final molecular networks
- *Supplementary files.* All remaining files relevant to this dataset that do not properly fit into any of the other listed file categories. A metadata file (sample information in a tab-delimited text format) with relevant attributes that can be used for visualizing the data in networks should be included here (see Box 1).

Required for 'Complete' Submission
- *Result files.* These are not necessary for small-molecule workflows but can be included and this is encouraged.
- *Spectrum identifications in a standardized format.* The following formats are recognized by MassIVE as valid for this category: mzIdentML[145], mzTab[146], and mzTab-M[147].
- *Search engine files.* The output of any search engine or data analysis tools or pipelines that were used to analyze this dataset, unless provided in a standardized format recognized by the 'Result Files' category (see above).

Optional
- *License files.* These specify how and under what conditions the dataset files can be downloaded and used. Multiple license files can be uploaded, if appropriate. By default, you can simply leave the 'Standard License' checkbox checked and your dataset will be submitted under the default Creative Commons CC0 1.0 Universal license. However, if you wish to provide your own license, then you can uncheck this box and assign your own file to the 'License Files' category.
- *Spectral libraries.* Any custom spectral library files that were searched against in the analysis of this dataset or that were generated using the spectrum files provided in this dataset, if applicable.
- *Methods and protocols.* Any open-format files containing explanations or discussions of the experimental procedures used to obtain or analyze this dataset.

Optional, mostly relevant to peptidomics and proteomics projects
- *Quantification results.* Any data or metadata generated by the analysis software used. Typically applied to the quantification analysis of peptides and proteins.
- *Gel images.* Any gel image files generated in the event that 2D gel electrophoresis has been used as a separation method.
- *Sequence databases.* Any files from protein or other sequence databases that were associated with or searched against in the analysis of this dataset, if applicable (usually .fasta format).

16 Populate the 'Dataset Publication' section. This section has three optional fields:
- 'Enter a Password'—this is used, for example, to share selectively with collaborators and manuscript reviewers.
- 'Share on ProteomeXchange'—this is not applicable to small-molecule workflows; checking the box will submit and announce the dataset via the ProteomeXchange Consortium at the time that

it is made public on MassIVE. The dataset will not appear publicly in either repository until you click the 'Make Public' button on your dataset's status page (see below).

- 'Generate a DOI'—Use this if you want a digital object identifier to be generated and assigned to this dataset. This is encouraged for all public datasets and can be used in publications.

17 Disregard the section titled 'Advanced Global FDR Settings'; this is not applicable to small-molecule workflows. It is currently used for global FDRs across submitted files in proteomics datasets.

18 In the 'Workflow submission' section, enter an email address at which you will receive notifications when workflow jobs are completed.

▲ **CRITICAL STEP** Make your dataset public, as shown in Step 19; this is not automatic and must be done explicitly after submitting data and generating a dataset MSV accession number.

19 Once a dataset is submitted to MassIVE, it will have an MSV accession number and will be a private dataset in the repository, accessible only to the submitter through their personal user interface or via a user-approved password-protected link (e.g., perhaps during a review for publications). To make a dataset public, first select the 'Jobs' tab of the user workspace portal (Box 3) to find the dataset. In the list of all job submissions, MassIVE dataset submissions will appear as 'MASSIVE-COMPLETE' workflows. Click on 'DONE' next to the MassIVE dataset to be made public and choose 'Make Dataset Public'. On the MassIVE website, to enable immediate use of the MassIVE dataset for GNPS workflows, click on the 'Convert Spectra' tab. This converts the uploaded files to .mzML files in a new folder called 'ccms peak'. Otherwise, the uploaded data will be queued for this conversion and will not be immediately available.

The dataset MSV000083437 has been made public, as illustrated in Supplementary Fig. 3; this feature enables any reader to interact with the data and follow along with this workflow.

Molecular networking in GNPS (Supplementary Fig. 4) ● Timing a few minutes to several hours/days (depending on dataset size, user expertise)

Choice of workflow

▲ **CRITICAL** Once MS data files are uploaded as datasets into GNPS-MassIVE, they are available to use for analysis workflows within GNPS. Here, we highlight how to execute the molecular networking workflow. A dataset can be recalled from either private or public domains in MassIVE for networking analysis. Once data files have been added, they will be populated in the 'Basic Options' section of the workflow selection. The user must then input a number of parameters before running the GNPS job in both the 'Basic Options' section and in a number of 'Advanced Options' sections. The advanced parameters are dependent on analysis platform, experimental setup, and conditions for acquisition of mass spectra, and will require the user to understand their ionization methods, fragmentation conditions and energies, mobile and stationary phases, and the fragmentation behavior of molecules of interest. Suggested settings for a variety of platforms are provided in the experimental section (Supplementary Methods). A GNPS job will take ~10 min for small datasets (up to 4 LC/MS files), 1 h for medium datasets (5–400 LC/MS files), and several hours (to days) for larger datasets (400+ LC/MS files).

20 Decide whether you are going to follow the 'molecular networking flow' (Steps 20–31), or whether you will proceed by choosing a Network Parameter Preset. In the 'Networking Parameter Presets' section (which resides directly under the 'Workflow Selection' and 'Workflow Description' sections), one of three options can be selected to set the networking parameters to approximately appropriate values depending on the size of your dataset. Clicking on one of these three options will open a workflow input form in a new tab. The default workflow settings are for 'medium data'. 'Small data' refers to a dataset of up to 4 LC-MS files, 'medium data' corresponds to datasets of 5–400 LC-MS files, and 'large data' is applicable to datasets of >400 LC-MS files (e.g., MSV000083437 is a medium dataset with 113 files in total). Because readers following the tutorial on the dataset MSV000083437 are guided through selection of parameters, no 'Parameter Preset' should be chosen for this example. Network Parameter presets were developed for low-resolution data to ensure a 1% FDR; on the basis of this, readers with high-resolution data or with special requirements are advised to use custom parameters.

▲ **CRITICAL STEP** If the user selects an option in the 'Networking Parameter Presets' section, Steps 21–30 can be bypassed. Because readers following the tutorial on the dataset MSV000083437 are guided through selection of parameters, no parameter preset should be chosen for this example.

Molecular networking workflow

21 Log in to GNPS (refer to Steps 2–8 for information about how to set up an account). The GNPS website banner contains tabs with which to navigate the platform, including tabs with which to

Box 5 | The importance of making your GNPS-MassIVE data public

Many GNPS users do not realize that when they have a dataset with an MSV accession number, their data are not yet public and thus remain in their private space, in accordance with the GNPS-MassIVE philosophy that data depositors should define how much and when they want to share their data in the public domain. Alternatively, upon submission, users can choose to make a dataset entirely available or 'public' to the GNPS community for browsing, commenting, subscribing, and/or downloading. This not only promotes robustness and reproducibility in MS data analysis, but also provides the user with access to the knowledge of the entire community. Indeed, the utility of GNPS for all users increases as more data become public, and the information and knowledge gained by any one user from this free service to the community derives from contributions made by the rest of the GNPS community. Thus, if you are a GNPS user benefiting from community contributions, by making your datasets public (and contributing network annotations, Step 45), you are giving back to the community. All users are encouraged to make their data public as early as possible, which provides the depositor with access to advanced features that are not available for private datasets. These features include being able to subscribe to the dataset, find related datasets, share datasets with collaborators, access living data, and utilize emerging features such as the Mass Spectrometry Search Tool or MASST (the equivalent of BLAST for small molecules[140]). We expect that features will continue to be developed further, thereby continually increasing the value for the end user of both their own and other public datasets.

navigate to MassIVE datasets, help documentation and the GNPS forum, as well as contact information (Supplementary Fig. 1).

22 Upload the desired dataset(s) to MassIVE (Steps 11–18 and Box 4). This step can be skipped if importing existing data files from MassIVE. Readers following the tutorial can omit this step because the GNPS-MassIVE dataset MSV000083437 already exists.

23 From the GNPS splash screen (home page), start a molecular networking job by clicking the 'Create Molecular Network' button (Supplementary Fig. 4a). This will bring up the main workflow input page, which has a number of fillable fields to complete under each of ten sections (Supplementary Fig. 4b).

24 In the 'Workflow Selection' section, enter a descriptive name for the job into the 'Title' field to facilitate retrieval of the workflow upon its completion (Box 5). Readers following the tutorial can type 'GF/SPF Mouse Duodenum Example' into the 'Title' field (Supplementary Fig. 4c).

25 Under 'Basic Options', add the LC-MS files for the molecular networking workflow by choosing the 'Select Input Files' tab next to the 'Spectrum Files (Required)' field. A pop-up window with three tabs will appear: 'Select Input Files', 'Upload Files', and 'Share Files' (Supplementary Fig. 4c). If multiple datasets will be analyzed together, repeat the above procedure with the other MSV numbers to import them into the user space. For readers following the dataset MSV000083437 tutorial, files can be imported by selecting the 'Share Files' tab. In the 'Share Files' window, enter the MassIVE accession number for the dataset (MSV000083437) into the 'Import Data Share' box (Supplementary Fig. 4c). After clicking 'Import', the dataset will appear in your GNPS user workspace and files can be selected for the GNPS networking workflow under the 'Select Input Files' tab as described below.

26 Choose the 'Select Input Files' tab (Supplementary Fig. 4f) to input mass spectrometry files already in your user workspace. From the list of datasets towards the lower left of the window, select all the files you want to analyze by clicking on individual files or an entire folder. For readers following the tutorial, GF1, GF2, GF3, GF4, SPF1, SPF2, SPF3, and SPF4 should be selected from the folder labeled 'peak'.

? TROUBLESHOOTING

27 Click on the 'Spectrum Files G1' button (top of left-hand column list, with green arrow) to mark this folder/files for analysis. Your selection(s) should appear in the 'Selected Spectrum Files G1' folder in the right-hand column of the window. For readers following the tutorial, folders containing data for GF1, GF2, GF3, GF4, SPF1, SPF2, SPF3, and SPF4 should now be under 'Selected Spectrum Files G1' (Supplementary Fig. 4g).

28 Load the associated metadata file (see Box 3 for format) separately into the 'Selected Metadata File' folder. To do this, select the file from your workspace list (often within a MassIVE dataset in the folder labeled as 'other'), click on the 'Metadata File' tab with the green arrow, and check that the file appears in the right-hand 'Selected Metadata File' folder. For readers following the tutorial, '3DMouse_duodenum_metadata.txt" can be selected from the folder labeled 'other' (Supplementary Fig. 4h).

? TROUBLESHOOTING

Table 4 | Absolute mass differences (Da) and associated mass error (parts per-million, p.p.m.) for illustrative *m/z* values

m/z	2.0 Da	0.5 Da	0.1 Da	0.05 Da	0.03 Da	0.025 Da	0.02 Da	0.0175 Da	0.015 Da	0.01 Da	0.0075 Da
200	10,000 p.p.m.	2,500 p.p.m.	500 p.p.m.	250 p.p.m.	150 p.p.m.	250 p.p.m.	100 p.p.m.	87.5 p.p.m.	75 p.p.m.	50 p.p.m.	37.5 p.p.m.
500	4,000 p.p.m.	1,000 p.p.m.	200 p.p.m.	100 p.p.m.	60 p.p.m.	49 p.p.m.	40 p.p.m.	35 p.p.m.	29 p.p.m.	20 p.p.m.	15 p.p.m.
1,000	2,000 p.p.m.	500 p.p.m.	100 p.p.m.	50 p.p.m.	30 p.p.m.	25 p.p.m.	20 p.p.m.	17.5 p.p.m.	15 p.p.m.	10 p.p.m.	7.5 p.p.m.
1,500	1,333 p.p.m.	333 p.p.m.	66 p.p.m.	33 p.p.m.	20 p.p.m.	16 p.p.m.	13 p.p.m.	11.6 p.p.m.	10 p.p.m.	6.6 p.p.m.	5.0 p.p.m.
2,000	1,000 p.p.m.	250 p.p.m.	50 p.p.m.	25 p.p.m.	15 p.p.m.	12.5 p.p.m.	10 p.p.m.	8.75 p.p.m.	7.4 p.p.m.	5.0 p.p.m.	3.75 p.p.m.

29 Once files have been selected, close the popup window by clicking on 'Finish Selection'. Datasets from both your private workspace and the public domain can be recalled using either strategy. For readers following the tutorial, the final data input is shown in Supplementary Fig. 4i.

30 Fill in the 'Precursor Ion Mass Tolerance' (PIMT) and 'Fragment Ion Mass Tolerance' (FIMT) fields in the 'Basic Options' section. Take into consideration the instrument resolution and calibration, as well as the acquisition parameters and the targeted/anticipated molecular masses (Table 4). PIMT: This parameter is used for MS-Cluster[10,12] and spectral library searching, and the value influences the clustering of nearly identical MS^2 spectra via MS-Cluster. FIMT: For each group of MS^2 spectra being considered for clustering (consensus spectrum creation), this value specifies how much fragment ions can be shifted from their expected *m/z* values. The default is ± 2.0 Da for PIMT and ± 0.5 Da for FIMT because the reference libraries also contain spectra from low-resolution instruments (e.g., ion traps of triple quadrupole). These can be adjusted to any appropriate value. For high-resolution instruments, the values commonly used are ±0.01 Da (Orbitrap) and ±0.02 Da (quadrupole time-of-flight (qTOF)) for both PIMT and FIMT. For readers following the tutorial example, the data were acquired on a Bruker MaXis qTOF instrument using ±0.02 Da (Supplementary Fig. 3c).

This 0.02-Da value translates into a maximum error of 40 p.p.m. at *m/z* 500, 20 p.p.m. at *m/z* 1,000 for the precursor ion, and 13 p.p.m. at *m/z* 1,500, which is consistent with the typical *m/z* range for small molecules. These Dalton to parts-per-million conversions are tabulated in Table 4. Peptidic small molecules may be ≥2,000 Da, although multiply charged, and thus PIMT and FIMT values of 0.03 Da should be used.

! CAUTION Although using low-resolution parameters may increase the number of annotations, it will also increase the number of false-positive annotations.

▲ CRITICAL STEP The default parameters recommended above for high-resolution mass spectrometers will not result in comprehensive searches of the spectral libraries generated on low-resolution mass spectrometers, such as ReSpect[77], a large portion of MassBank data[28], and some GNPS community–contributed data; a substantial portion of spectra that were annotated by matching to the NIST Mass Spectral Library with Search Program Data Version: NIST v17 (https://www.nist.gov/srd/nist-standard-reference-database-1a-v17) are also low resolution. In addition the NPs community contributes annotated spectra from a range of different spectrometers that may be high or low resolution (Supplementary Table 3).

31 Click 'Submit' to begin the job. The molecular networking job for the example dataset (MSV000083437) should take about 20 min. The following additional parameters can be also be specified before final submission of the job:

- Complete the remaining fillable fields in 'Advanced Network Options', 'Advanced Library Search Options', and 'Advanced Filtering Options' in the 'Basic Options' section according to the experimental design. Recommendations and values used for the example dataset are provided in Table 2.
- Use the default parameters for 'Advanced GNPS Repository Search Options', 'Advanced Annotation Options', and 'Advanced Output Options'. The option 'Create Cluster Buckets and BioM/PCoA Plots Output' must be enabled in the 'Advanced Output Options' to generate bucket tables and PCoA (interactive principal coordinates analysis) plots from the 'Export' and 'Advanced Views' options on the job status page.
- Finally, enter an email address under 'Workflow Submission' to receive notifications when workflow jobs are completed. Readers following the tutorial should do this to receive notification when the example job is completed.

? TROUBLESHOOTING

Table 5 | Data analysis options

Data analysis option	Description
View all library hits (Supplementary Fig. 5a)	View all spectra with reference database matches and assess the quality of the MS^2 match using the 'View Mirror Match' option. Readers following the tutorial example can view the mirror plot for cholic acid (Supplementary Fig. 5a) to compare experimental spectra with library annotations. Readers can investigate mirror plots for other bile acids because bile acid discovery is the focus of this example
View unique library compounds (Supplementary Fig. 5b)	View all unique spectral matches to the reference database and perform side-by-side comparison between the query spectrum and reference spectrum. Readers following the tutorial can view query and reference spectra for cholic acid (Supplementary Fig. 5b)
View all clusters with IDs (Supplementary Fig. 5c)	View all consensus MS^2 spectra that make up a node
View spectral families (Supplementary Fig. 5d)	List of all spectral families (nodes that are connected to one another); the user can view individual sub-networks using in browser visualization
View EMPeror PCoA plot	Measures the binary Jaccard distance between samples on the basis of the presence/absence of molecular features with associated MS^2 spectra as defined by the mass spectral molecular network. Interactive principal coordinates analysis (PCoA) visualization is enabled through EMPeror[137]

Visualization of the molecular network ● Timing 1 h to a few days/weeks (depending on size and complexity of network)

32 Visualize the molecular networks generated using either direct visualization in browser (option A) or Cytoscape visualization (option B) (Fig. 2, stage 5); these options for data visualization in addition to other data analysis options are tabulated in Table 5. These methods are complementary to one another and the user should choose the preferred visualization strategy on the basis of their data analysis needs. The GNPS in-browser visualization tool is a quick, simple way to begin analyzing data, particularly if the user wants to view and compare MS^2 spectra within the network. However, in-browser visualization allows the user to view only one molecular family (subnetwork) at a time. For more advanced data analysis and formatting options, the user can visualize their network offline using third-party tools. One popular visualization tool for molecular networks generated in GNPS is Cytoscape[35], a program originally introduced by the systems biology community to allow visualization of the complex relationships in biological sequence data. With Cytoscape, one can visualize the chemical space that was detected in the mass spectrometry experiment as a molecular network; the tool provides a way to encode any property of the network (i.e., node label, shape, color, or size, as well as edge label, thickness, and so on) with a metadata category (i.e., cohort, cosine score, compound source). An online tutorial can be accessed at https://ccms-ucsd.github.io/GNPSDocumentation/networking/#online-exploration-of-molecular-networks.

The steps outlined in option B provide the user with a working knowledge of how to configure a network in Cytoscape. Readers following the tutorial example can not only reproduce the same properties described in the steps below to generate a publishable network, but also use this network to specifically focus on the cluster containing bile acids in order to discover novel compounds.

(A) **Molecular network visualization in the GNPS browser**

(i) Access the in-browser data analysis options from the job status page (Fig. 4); several of these are described in Table 5 (see also Supplementary Fig. 5).

The 'View spectral families' option lists each individual molecular family that contributes to the entire molecular network and displays the number of MS^2 spectra and spectral matches to the reference library that contribute to a given subnetwork. This function also allows users to visualize each subnetwork individually in a web browser by selecting the 'Visualize network' link. Once the in browser network is displayed (Fig. 5), the user can immediately distinguish between nodes with library matches (blue circles) and unannotated nodes (gray circles). Edges are represented by gray arrows that point from the low mass spectra to the high mass spectra.

(ii) Perform further data analysis in this online interface as described below:

• *Node Labels*. Nodes can be labeled by their index number given by MS-Cluster (cluster index), parent mass, or library annotation name (LibraryID). In addition, the

Fig. 4 | GNPS job status page.

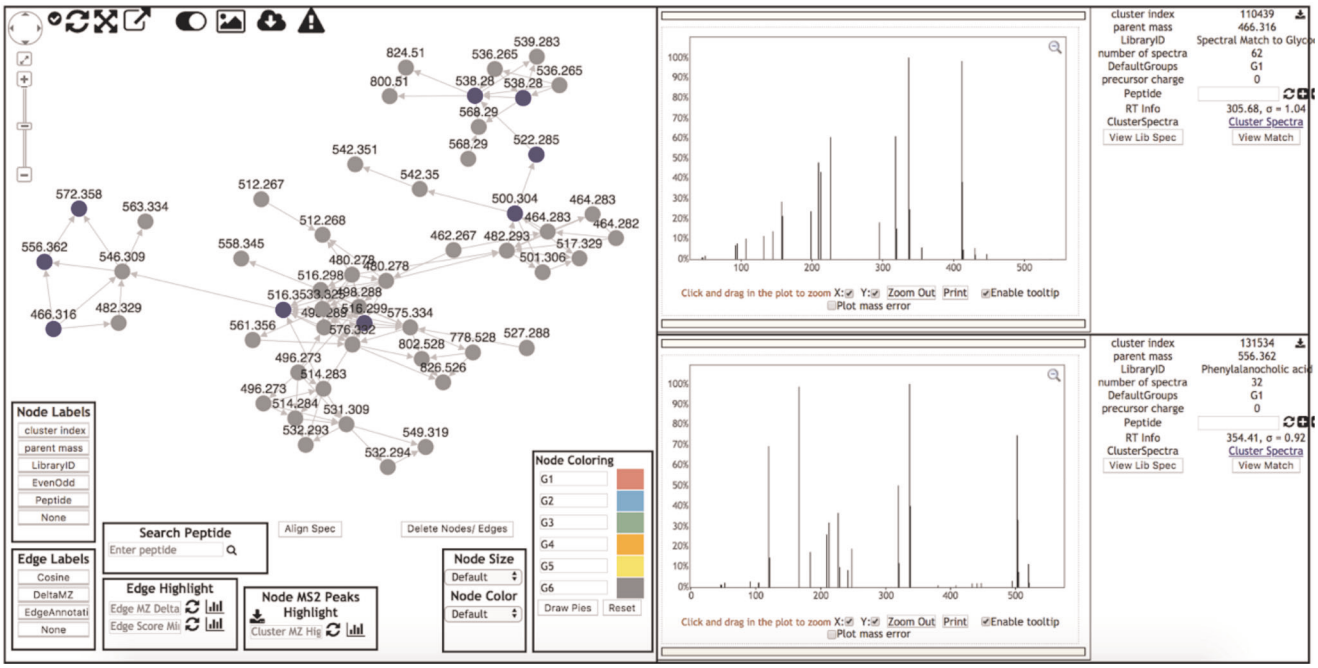


Fig. 5 | Screenshot of the in-browser visualization of the bile acid spectral family from dataset MSV000083437.

node can be labeled by a binary system (EvenOdd) to denote if the parent mass is even (1) or odd (0) to assist in visualizing the nitrogen-rule[118], or with a peptide annotation label ('Peptide'; see also 'Search Peptide' below). If no node label is desired, select 'None'.

- *Node Coloring*. This legacy feature creates pie charts to visualize mapping of metabolites into different groups. However, this option does not use the sample information (metadata) table and will work only if files were inputted into different groups by the user. It is important to note that this is not a quantitative representation of the data because it relies only on MS^2 spectral counts. Rather this feature can be used to understand presence versus absence of compounds in specific groups.
- *Edge Labels*. Edges connecting two nodes can be labeled with either the cosine score (Cosine) or the mass difference between the parent *m/z* values ('DeltaMZ'). If no edge label is desired, select 'None'.
- *Edge Highlight*. Edges by default are represented as arrows pointing from low mass spectra to high mass spectra and can be colored. Users are able to enter a mass difference (*m/z* delta) of their choice in the 'Edge MZ Delta' field, causing those edges to be highlighted in red. Clicking on the graph icon next to 'Edge MZ Delta' opens a new window containing a graph that shows the distribution of all edge *m/z* delta values in the sub-network. Selecting a peak in this 'Network MZ Delta Histogram' highlights the corresponding edges in red. The same function can be performed for 'Edge Score Minimum' to highlight edges that have a cosine score greater than what the user enters.
- *Node Size and Node Color*. The size and color of nodes can be adjusted on the basis of spectral counts, precursor intensity, number of files, parent mass, even/odd mass, or precursor charge.
- *Node MS2 Peaks Highlight*. This option allows users to search the subnetwork for molecules that contain an MS^2 fragment of interest. To perform this query, first click the download button within this box to pull all of the MS^2 spectra into the browser. The desired *m/z* value can then be entered into the field to highlight the nodes comprising spectra that contain the desired product ion. Alternatively, the histogram icon can be selected to visualize all product ions from the MS^2 spectra in the subnetwork.
- *Align Spectra*. This function enables direct comparison between the spectra of two connected nodes at the peak level. To perform this analysis, the user should first select an edge connecting two nodes, which pulls up the spectra for each node in the right display window. Clicking the 'Align Spec' button overlays the spectra, where red peaks represent peaks of the exact same masses shared between the top and bottom spectra and blue peaks denote peaks matching at shifted masses.
- *Search Peptide*. This is a function added to GNPS to support proteomic and peptidomic dataset analysis. If a peptide sequence is found to be associated with the molecular family and was found through automated peptide mining in MassIVE, then the amino acid sequence entered here will be searched.

(B) **Molecular network visualization in Cytoscape**

▲ **CRITICAL** There are a few options for exporting molecular networks for visualization in Cytoscape. Once molecular networks generated from GNPS are imported into Cytoscape, a number of simple commands can be used to make the network generated more informative, visually appealing, and accessible (Supplementary Fig. 6). Documentation on how to use Cytoscape (versions after the 3.7 release) and a Cytoscape community forum are available to assist with troubleshooting and to provide information about the latest plugins (also called Cytoscape Apps): https://cytoscape.org/documentation_users.html and https://cytoscape.org/community.html. An online version of this tutorial is accessible at: https://ccms-ucsd.github.io/GNPSDocumentation/cytoscape/.

(i) To begin using Cytoscape, download the latest version of the software from: https://cytoscape.org/ according to the instructions given at the website (Supplementary Fig. 6a).

(ii) Import a molecular network into Cytoscape for visualization. Once Cytoscape has been downloaded, molecular networks can be imported and visualized using two different strategies. The first option will show a network with no preset layout, whereas the second will show a network with default layout settings. Do one of the following:

- To import data for a network with no layout present, click on 'Download GraphML for Cytoscape' in the GNPS Job Status window (Supplementary Fig. 6b). This will prompt an immediate download of a compressed folder containing the .graphml file of interest;

after uncompressing this folder using a variety of programs, Cytoscape can be opened. The import network button (three nodes connected by edges, Supplementary Fig. 6c) in Cytoscape can be selected, permitting selection of the .graphml file to load the network of interest.

- Click on 'Direct Cytoscape Preview/Download' in the GNPS Job Status window (Supplementary Fig. 6d). This will direct the user to a new window where a preconfigured version of the molecular network will be displayed. In this window, click on 'Download Cytoscape File' to download the file as a Cytoscape session file (.cys file) with the visualization parameters already defined. Cytoscape can then be opened by double-clicking on the downloaded .cys file; this network will come preloaded with GNPS default layout.

Readers following the tutorial can use either strategy to open the completed GNPS job run on dataset MSV000083437.

? TROUBLESHOOTING

(iii) Customize an imported molecular network for viewing. By altering many properties of nodes, edges, and networks, such as colors, sizes, shapes, and labels, the default network can be transformed into a chemically informative molecular network. Readers following the tutorial example are guided through this process below. In the 'Control Panel' window, located on the left side of the screen, the 'Style' and 'Select' tabs offer many options.

To alter a node style, click on the 'Style' tab at the top of the 'Control Panel', then click on the 'Node' tab at the bottom of this window (Supplementary Fig. 6e).

(iv) Change the node labels in Cytoscape by selecting the drop-down arrow next to the 'Label' tab. Readers following the tutorial example can label nodes by selecting 'precursor mass' as column and 'Passthrough Mapping' for mapping type (Supplementary Fig. 6f).

(v) Customize node shapes. Readers following the tutorial example can click directly on the 'Shape' symbol button and select 'Ellipse' or change to another desired shape (Supplementary Fig. 6g). If using 'Ellipse', the shape can be converted into a circle by checking the box labeled 'Lock node width and height' (Supplementary Fig. 6h).

(vi) To change the node color, click on the "Fill Color" drop-down. Under this column, readers following the tutorial example can select the desired value (i.e., 'ATTRIBUTE_host_microbiome') and use this to discriminate groups (i.e., GF versus SPF) from one another. Readers can select 'Discrete Mapping' under the 'Mapping Type' column, which allows for the selection of a color to be associated with each group (Supplementary Fig. 6i).

(vii) Alternatively at the 'Fill Color' option, use the 'Image/Chart 1' tab to visualize the relative ion distribution from each chosen group in the nodes as a pie chart. Readers following the tutorial can perform this type of visualization by clicking on the 'Image/Chart 1' button, selecting the 'Charts' tab, and choosing a chart type (the pie chart is chosen in this example). The spectral count information from groups defined in the metadata file can then be moved from the 'Available columns' to the 'Selected columns' by clicking on the associated button, (Supplementary Fig. 6j) and the user can edit the chart color scheme using the 'Options' tab. In this example, 'Germ free' and 'Specific Pathogen free' can be selected and colored pink and blue, respectively.

(viii) Go to 'Size option, select 'number of spectra' or 'sum(precursor intensity)' as 'Column' and 'Continuous Mapping' as 'Mapping Type' to visualize the variation in the occurrence of each ion across samples (e.g., count of 1 if not zero) as a function of the node size. The opened window allows the user to modify the node size in function of the node metadata column chosen. Begin by setting the values for minimum and maximum node sizes with the button 'Set Min and Max…', and then click 'OK'. Then move the cursor at each extremity. For readers following the tutorial example, set the minimum size at 92 and the maximum size at 362 (Supplementary Fig. 6k).

(ix) Alter the edge style by clicking on the 'Edge' tab at the bottom of the 'Control Panel' (next to the 'Node' tab) (Supplementary Fig. 6l). Readers following the tutorial example can select this tab to make alterations in edge color and width, in addition to other settings.

(x) To change an edge label, click on the 'Label' drop-down arrow then select the desired value. For example, select 'mass_difference' as 'Column' in the 'Passthrough Mapping' mode (Supplementary Fig. 6m).

(xi) Click on the drop-down arrow next to 'Width' to change the edge width. Under the 'select value' tab next to the 'Column' tab, select the desired value used for scaling edges (such as 'cosine_score'). At this point, select 'Continuous Mapping' under 'Mapping Type' (Supplementary Fig. 6n). Select 'cosine_score' in the 'Column' tab and "Continuous Mapping" can be chosen under 'Mapping Type' to easily visualize the approximate cosine score of all edges.

(xii) Subtract the ions from experimental conditions present in the blank sample from the molecular networks, if this is desired. In the 'Table Panel', readers following the tutorial example can go to the column 'GNPSGROUP:blank', select every row with ion occurrence (>0), then click on the right mouse button and 'Select nodes from selected rows' can be chosen (Supplementary Fig. 6o). The selected nodes were automatically highlighted in yellow in the network. Then, right-click to choose 'hide selected nodes and edges' in the selected rows (Supplementary Fig. 6p). However, it is possible to remove the ions from experimental conditions before generating a molecular network by data processing[119].

? TROUBLESHOOTING

(xiii) To separate one or some specific desired network(s), press 'Ctrl' or 'Command' (Windows or MacOS, respectively) while at the same time selecting the network(s) with the mouse. Then, click on the bottom as shown in Supplementary Fig. 6q. Automatically, the subnetwork is created. For going back to the main network, go into the 'Control Panel' by selecting 'Network', and then click on the main network bottom.

(xiv) At this point, readers following the tutorial example have generated a publishable network in Cytoscape from the output of molecular networking in GNPS. This network should look like that shown in Fig. 3. If you are interested, look more closely at the subnetwork containing key bile acids to practice manual propagation of annotations throughout a sub-network (Fig. 3). Style options are described in more detail in the Cytoscape manual: http://manual.cytoscape.org/en/stable/Styles.html.

Assessing the quality of a library hit ● Timing 30 min–days

33 All spectral matches are putative annotations[4] until experimentally validated. Spectral matches from molecular networking analysis are annotations at level 2 (compounds that have been putatively annotated, e.g., no reference standards) or 3 (compounds that can be putatively assigned to a chemical class on the basis of physicochemical properties and/or spectral similarity) before validation with chemical standards. For level 1 annotation, the molecules would have to be isolated and structures elucidated or confirmed with other techniques such as NMR or X-ray analysis, or matching MS^2 and retention times, together with co-analysis with pure standards, ideally under more than one chromatographic condition. All non-annotated molecules in a molecular network are level 4, unless they are part of a molecular family containing a library match. Consult the definitions of these levels as agreed on by the 2007 Metabolomics Initiative[120] and subsequently refined by the Compound Identification work group of the Metabolomics Society at the 2017 annual meeting of the Metabolomics Society[121].

34 To judge the quality of a match, consider the mass accuracy of the reference spectra (resolution and calibration of the instrument) as compared with that of the experimental spectra. The sample type, experimental setup, and associated sample information (metadata) should also be taken into account when judging the accuracy of the matches. Notably, MS^2 spectra typically cannot differentiate regio- or stereo-isomers, and additional experiments, including comparison with standards, are required to assign the absolute structure.

35 To decrease the impact of this variation, subject all spectra, when compared, to a square root conversion. This decreases the high-intensity ions and increases the low-intensity ions.

36 Furthermore, to address variability in data quality and source of the reference spectra, access the GNPS ranking system for submitted reference spectra to enable filtering of the reference library, either before performing molecular networking or afterward, which is the default approach.

37 Similarly, consider the instrument on which the reference data were collected. This is done after doing the analysis in GNPS using post-molecular networking filtering capabilities.

38 Consider the quality of the reference spectra. 'Gold' reference spectra can only be submitted by approved users and must originate from fully characterized synthetic or purified compounds.

This is the same gold standard by which other metabolomics reference libraries such as NIST17[73], METLIN[74] mzCloud (https://www.mzcloud.org/)[122], and WeizMass (https://www.weizmann.ac.il/LS_CoreFacilities/weizmass-spectral-library-high-confidence-metabolite-identification)[123] libraries are curated. Gold-level spectra comprise 83% of the MS^2 spectra provided to GNPS as libraries. A 'silver' rating signifies that the spectrum was submitted with an associated publication. However, GNPS also curates crowd-sourced knowledge from users in the community. All remaining reference spectra provided by the user community receive a 'bronze' rating to denote that the annotation is contributed by users including partial or putative annotations. The annotation within GNPS can be made directly from the data and thus relies on the expertise of the experimentalist, and purification of the molecules is not required. This gives access to a curated reference database that is crowd-sourced and does not rely on commercially available standards. For example, most NPs from microbes, food, and plants are not commercially available, but partial annotations and thus crowd-sourced knowledge capture provides a resource of information that is inaccessible any other way. The only other resource that currently accepts putative and partial annotations is MassBank EU (https://massbank.eu/MassBank/). Examples of useful but partial annotations include modifications of molecules, such as oxidation of a molecule in which the site of oxidation is unknown[124] and thus a SMILES or InChI cannot be drawn, but the partial annotation provides valuable insight to the end user. Additional partial annotations would include adduct clusters—such as sodium formate clusters or polymeric substances, including oligosaccharides—commonly detected in mass spectrometry where a structure cannot be drawn, but the cluster information is useful knowledge for the community when performing an untargeted LC-MS^2 experiment.

39 Consider the corresponding cosine score, which is calculated in the program and takes into account the number of matching fragment ions and differences in peak intensities, and parent mass accuracy to assess the quality of annotation. An empirical cutoff for cosine scoring of 0.7 with 6 MS^2 ions matching is the default setting in GNPS. On average, this gives rise to 91% accurate annotations and ~1% incorrect annotations, with the remainder being attributed to possible isomers (4%) or having not enough information by the user to judge (4%)[100]. However, use of a target decoy–based method to estimate confidence measures of annotations and FDRs in large-scale metabolomics experiments revealed that the annotation quality is dataset dependent, as well as dependent on analysis settings such as number of ions that are required to match. The general trend was that when few MS^2 ions are required to match, a much higher cosine is required and fewer matches will be obtained at the same FDR compared to when more MS^2 ions are required to match the reference spectra. When more ions are matched, the cosine score can be lowered. There is a dataset-dependent optimum for the maximum number of spectral library matches at a specific FDR that is typically ~4–6 minimum matched peaks[100]. Although the confidence in the spectral matches increases when more MS^2 fragment peaks are required, there are fewer spectra that have a larger number of ions, resulting in a diminished number of annotations, especially for low-molecular-weight compounds.

Propagating annotations through manual interpretation of the networks ● Timing days–weeks

▲ **CRITICAL** A molecular network can be very useful in manually propagating annotations—using the information from one annotated node to inform the annotation of nearby, non-annotated nodes—through manual interpretation of networks in parallel with raw MS^2 spectra. Manual annotation can be performed by looking at mass differences (deltas) in the molecular network and assigning the source of these deltas, that is, charge retention fragmentations such as retro-Diels–Alder reactions or McLafferty rearrangements and charge migration fragmentations such as simple inductive cleavages or α- or β-eliminations[125]. The novel bile acids found in the mouse duodenum provide an example of the utility of manual interpretation of networks (Supplementary Fig. 7b). One can use the mass deltas between unknown nodes and neighboring library hits to determine new structures. In the above example, three unknown nodes were determined to be novel bile acids conjugated with phenylalanine, leucine, and tyrosine on the basis of their mass deltas with respect to glycocholic or glycomuricholic acid. A description of how manual propagation of annotations can be performed in the context of the example is given below:

40 Search nodes or edge metadata (e.g., 'shared name') using Cytoscape's toolbar. Readers following the tutorial example can enter "glycocholic acid" with the quotation marks. The nodes of interest at

m/z 466.316 that match glycocholic acid in the GNPS library are automatically selected and highlighted in yellow in the network (Supplementary Fig 6g).

? TROUBLESHOOTING

41 Manually propagate annotation based on mass shifts. In Supplementary Fig. 7a, glycocholic acid connects to a node with *m/z* 556.363. On the basis of the mass shift of 90.047, the unknown node can be manually annotated as glycocholic acid conjugated with phenylalanine. Analogously, nodes with *m/z* 572.358 and 522.379 could be manually annotated as glycocholic acid conjugated with tyrosine and leucine, respectively, accounting for mass shifts of 106.042 and 56.063 Da.

42 Use the 'select' function to assist in finding annotated nodes within the network with an *m/z* error from 0 to 10 p.p.m. between precursor ions. This tool is available in the 'Control Panel' at the 'Select' tab and can be used to create a selection of nodes and/or edges on the basis of their metadata and/or network topology. Readers following the tutorial example can click on the '+' button and choose 'MZErrorPPM' as column filter and move the cursor from 0 to 10, and then click on 'Apply' (Supplementary Fig. 7b). These nodes are automatically selected and highlighted in yellow in the network.

43 Use advanced computational tools for automated annotation propagation, such as the Network Annotation Propagation (NAP) tool[84], or perform manual annotation using the results of Dereplicator[81,82] and MS2LDA[126], which can be accessed through GNPS at https://gnps.ucsd.edu/ProteoSAFe/static/gnps-theoretical.jsp.

Capturing information by adding reference spectra from your data ● Timing 10 min–days

▲ CRITICAL Once an MS^2 spectrum has been fully annotated, it can be added as a reference spectrum to GNPS. Because the GNPS library database is crowd-sourced, users are encouraged to submit spectral annotations because knowledge they have is captured through these annotations of reference spectra and is reusable by others. This enables the creation of reference spectra from MS^2 spectra in the dataset without needing to purify the molecule. The assumption is made that the people who collected the data are experts with regard to their samples and thus are in the best position to curate. In addition, if the same user or lab then uploads another related dataset, and it contains the same molecule, it will be automatically annotated.

44 To upload a single reference spectrum, click on 'View All Clusters With IDs' in the Job Status page, then select the cluster desired for annotation from the 'ClusterIdx' column.

45 Select the 'AnnotatetoGNPS' button. This button brings up the workflow for annotation, where input files, sample parameters, desired annotation, advanced annotations, and library selections can be added and the job can be submitted.

46 To add a known spectrum to the library from a file uploaded to MassIVE, select 'Contribute' under the 'Add Your Spectrum' heading on the main page, even if molecular networking has not been performed on this file.

47 To upload >50 reference spectra to GNPS, perform a separate batch upload as detailed in the online help documentation at https://ccms-ucsd.github.io/GNPSDocumentation/batchupload/.

All annotations can be refined at a later step, and the provenance of each contribution is retained within the GNPS-MassIVE environment. For example one person may annotate that they think a compound is a lipid, the next person may update and specify that it is a phosphatidylcholine, and a third person may refine this to be 1-oleoyl-2-palmitoyl-phosphatidylcholine, and all of this is logged into the CCMS spectral library for each MS^2 spectrum. To annotate an existing spectrum, select "View All Spectra with Ids" then select "AnnotatetoGNPS" in the resulting table. Alternatively, a reference spectrum can be added by selecting "Add Your Spectrum" in the main GNPS page.

Data sharing and reproducibility of molecular networking ● Timing 1–a few h

▲ CRITICAL Data sharing is essential for the reproducibility of molecular networking analysis; therefore, we encourage all users to make datasets public, share molecular networking job links, and so on.

48 To facilitate dissemination of the findings, refer to both the raw mass spectrometry data and the associated molecular networking jobs in the peer-reviewed articles where the findings are published. To do this, provide the MassIVE accession number (e.g., MSV000083437) and a hyperlink to the GNPS job in the methods or experimental details section of the publication.

Datasets uploaded to MassIVE ideally include all raw and peak-picked mass spectrometry data and associated sample information (metadata). GNPS records all data inputs, transformations, mathematical operations on, and analyses of the data, providing a historical record of the data and its origins. This data provenance promotes reproducibility and ultimately quality of the data and its annotations.

49 Deposit reference MS^2 spectra for all newly discovered NPs into the GNPS reference library (described above) and share the unique CCMSLIB identifier of the MS^2 spectra in the NP characterization data[3] upon publication. These are then included in the spectra used for dereplication, the identification of known substances.

Cloning a job ● Timing 10 min

50 Once a job's URL address is shared, any GNPS user can clone the job. To do this, follow the provided link and click 'clone' on the Job Status page (Supplementary Fig. 8). Cloning a job allows users to view all parameters and files that were used to create the existing network and easily rerun the molecular networking job with the same (or adjusted) parameters and files. Cloning a GNPS job is an extremely useful tool that promotes reproducibility and scientific rigor. This is a feature many users use to submit multiple molecular networking jobs with modified parameters. If a job has been run in the previous V1 version of GNPS (i.e., it was run using the 'METABOLOMICS-SNETS' workflow), it can be cloned and re-run in v.2 (V2) of GNPS by simply clicking 'Clone Job to Latest Molecular Networking V2 Workflow' on the Job Status page (Supplementary Fig. 8b).

▲ **CRITICAL STEP** Note that if data were imported from your private user workspace and not from within MASSIVE, other users will not have access to the mass data and consequently will not alter the analysis in GNPS.

Accessing an existing dataset on GNPS ● Timing 5 min

51 If a dataset is public, users are able to download all files for reanalysis, including raw data and the sample information table (metadata). To access a MassIVE dataset of interest, select 'MassIVE Datasets' in the GNPS workspace portal (Box 1) and enter the MassIVE accession number or defining keywords into the search bar.

▲ **CRITICAL STEP** Private datasets can be viewed only by the user who uploaded the data and anyone who has a link to the Job Status page. The user can create a password-protected link. When downloading data from a private dataset, you will be prompted to enter a password for that MassIVE dataset ID.

52 Click on the MassIVE accession number highlighted in green to link to the 'MassIVE dataset information page', and select the 'FTP Download' link to download files. Alternatively, this link can be pasted into the quick connect box of an FTP client.

▲ **CRITICAL STEP** If you are accessing private datasets using an FTP client, you will need to enter the MassIVE ID as the username, followed by a password. If the submitter did not specify a password, then it should be accessible using the password 'a'.

Subscribing to a dataset and living data ● Timing 5 min into the future

53 Public datasets remain alive long after publication; for example, they will be searched periodically against the ever-growing annotated GNPS spectral libraries, potentially yielding new putative annotations within those datasets. If you subscribe to a dataset, you will receive email notifications of new identifications that are made within that dataset as well as about other datasets that exhibit chemical similarities to the subscribed dataset. This allows for users to be connected via their research interest to similar datasets. Updates are sent out about once a month and only when there is new information associated with the dataset. To subscribe to a dataset, navigate to the 'MassIVE dataset information' page as described in Steps 11–19 and click 'Subscribe'. This feature changes the way we interact with data. Previously, data were periodically reanalyzed by the submission of new jobs, but in GNPS, data are automatically reanalyzed and updates are sent to the subscribers. Therefore, data may give rise to useful results a few weeks or even a few years later after they are uploaded or they may enable the dissemination of all the knowledge of this dataset to all lab members or collaborators.

? TROUBLESHOOTING

Troubleshooting

Troubleshooting advice can be found in Table 6. We also recommend checking the forum link from the banner in GNPS, where users can post questions to the GNPS community.

Table 6 | Troubleshooting table

Step	Problem	Possible reason and/or solution
All steps	This protocol does not address the issues that the user faces	Check the GNPS forum and post questions
1	Cannot convert Waters .raw files to .mzXML/.mzML files from data acquired in the MS^E mode of Waters mass spectrometers using ProteoWizard	Datasets acquired on a Waters mass spectrometer using the MS^E mode can currently be converted to .mzML only by using the vendor's UNIFI platform. Alternatively, data need to be collected in DDA (data-dependent acquisition) or MS^2 mode, for which data conversion to .mzXML/.mzML is enabled through ProteoWizard.
20–31	GNPS network is much smaller (fewer nodes) than expected	Check that you selected the mzXML peaklist files from the 'ccms_peak' folder of your MassIVE dataset for the GNPS workflow, not the mzXML files generated directly from the raw data files in the 'raw' folder. The value of the minimum cluster size can be reduced. The minimum cosine score can also be decreased to increase the number of edges in the networks
	Molecules known to be structurally similar do not appear to form a cluster	Check consensus spectra for the molecules of interest. It is possible that low-abundance noisy spectra are included, which results in poor consensus. Although most ions that are observed can be fragmented, some ions that are fragmented may have few ions to reliably network; these ions should be avoided. We advise using caution when looking at spectra with less than four fragment ions, on the basis of FDR estimates of spectral matching against libraries[100]. For some classes of compounds that do not fragment efficiently—for example, certain lipids—the MS^2 spectra are not informative enough to build meaningful network
26	Protocol user cannot see own file(s) after drag-and-drop upload to GNPS workspace	Check that the targeted folder is highlighted before dragging and dropping the file
26 & 31	Job fails with the message 'Empty MS/MS'	Check that your data are in a supported file format; check that the submitted files are centroided and have MS^2 data; check that filtering criteria are not too aggressive; check that raw files are not included in the file selection
28 & 31	GNPS job fails because of improper metadata format	The metadata file must be formatted as a tab-separated .txt file
31	Job fails with the message 'spectral library search exceeded memory'	This means that the spectral library search step used too much memory and had to be terminated. This is probably caused by changing the set of spectral libraries used in the search (such as removing the spectra filtering). This issue can potentially be resolved by increasing the maximum cluster size value to reduce the number of searched spectra. It is not recommended to change the set of libraries included, unless you are an advanced user. Remove all libraries except for the default 'speclibs' and rerun
32B(ii)	Network is too large to view in Cytoscape	If a dataset cannot be loaded into Cytoscape, a subnetwork of interest can be opened. Alternatively, larger networks can be opened on a computer with more RAM
32B(xii)	Protocol user does not know how to include/exclude blanks	This is most easily addressed if the blanks are included in the metadata. Then the user can opt to visualize spectra found in blanks using discrete mapping in Cytoscape or another visualization tool
40	Metadata does not sync with Cytoscape	The metadata (sample information) table must be formatted correctly. In particular, check whether the first column is named 'filename', whether all filenames match exactly the files uploaded to GNPS and have '.mzXML' extensions (or another compatible file format), whether each metadata column uses the prefix 'ATTRIBUTE_', and that there are no trailing spaces in any of the headings.

Timing

Step 1, data conversion: 1 h up to a few days (depending on size of dataset and computer setup)
Steps 2–8, creation of a GNPS/MassIVE account: 10 min
Steps 9–19, depositing data files by submitting a dataset: 30 min
Steps 20–31, starting a GNPS job: a few minutes to several hours/days (depending on dataset size, user expertise)

Step 32, visualize and analyze a network: 1 h to a few days/weeks (depending on size and complexity of network)
Steps 33–39, assessing the quality of a library hit : 30 min–days
Steps 40–43, propagating annotations through manual interpretation of the networks: days–weeks
Steps 44–47, capturing information by adding reference spectra from your data: 10 min–days
Steps 48 and 49, data sharing and reproducibility of molecular networking: 1 h–a few hours
Step 50, cloning a job: 10 min
Steps 51 and 52, accessing an existing dataset on GNPS: 5 min
Step 53, subscribing to a dataset and living data: 5 min into the future
Box 1, sample information (metadata) collation and input: typically 1–2 h for a small dataset; up to a few days for large, complex metadata entries of large datasets

Anticipated results

Molecular networking of LC–MS^2 data according to the protocol described herein integrates an associated sample information table (metadata file) with the latest molecular networking workflow, to yield a network (.graphml file) that can be visualized directly in GNPS or imported into Cytoscape. The tutorial example followed throughout the protocol demonstrates how contemporary GNPS molecular networking can be used to discover a new set of conjugated bile acids from the mouse gut microbiome as described in steps 1–32 of ref. [63]. The network produced from the protocol should contain a molecular family of conjugated bile acids that includes a library hit for glycocholic acid (Fig. 6a). This annotation can be propagated to identify new bile acids by converting the mass differences of the edges into structural motifs. For instance, the user can identify the *m/z* 546.309 node as a sulfated cholic acid by using its mass difference of 79.993. This strategy was key in determining the structures for the phenylalanine (*m/z* 556.362) and tyrosine (*m/z* 572.358) conjugated cholic acids. This example also showcases how manual comparison of the MS^2 spectra that make up the conjugated bile acid molecular family can also be critical for structural annotation. For example, spectra of Gly-, Phe-, and Tyr-conjugated cholic acid all contain fragment ions identical in mass to their respective amino acid conjugates (Fig. 6b). Furthermore, the mass difference between the precursor ion and the common peak at *m/z* 337.25, which corresponds to amide bond cleavage,

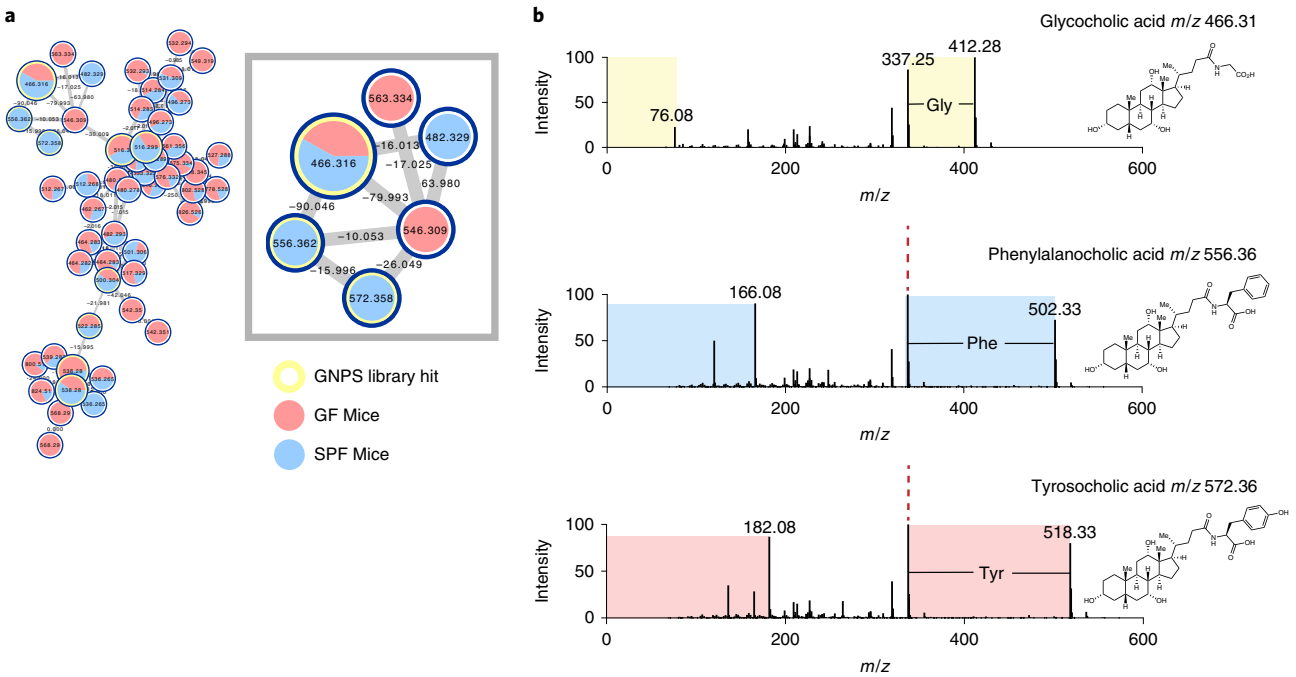


Fig. 6 | Propagation of molecular networking to discover relationships between molecules. a, The molecular family of conjugated bile acids from the duodenum of germ-free (GF) (red) versus specific-pathogen-free (SPF) (blue) mice in the MSV000083437 dataset. As shown in the inset, a library hit for glycocholic acid (Gly; *m/z* 466.316) is present in both GF and SPF mice, whereas the new phenylalanine (Phe; *m/z* 556.362) and tyrosine (Tyr; *m/z* 572.358)-conjugated bile acids are seen only in colonized mice. **b**, Comparison of MS^2 spectra for Gly-, Phe-, and Tyr-conjugated bile acids.

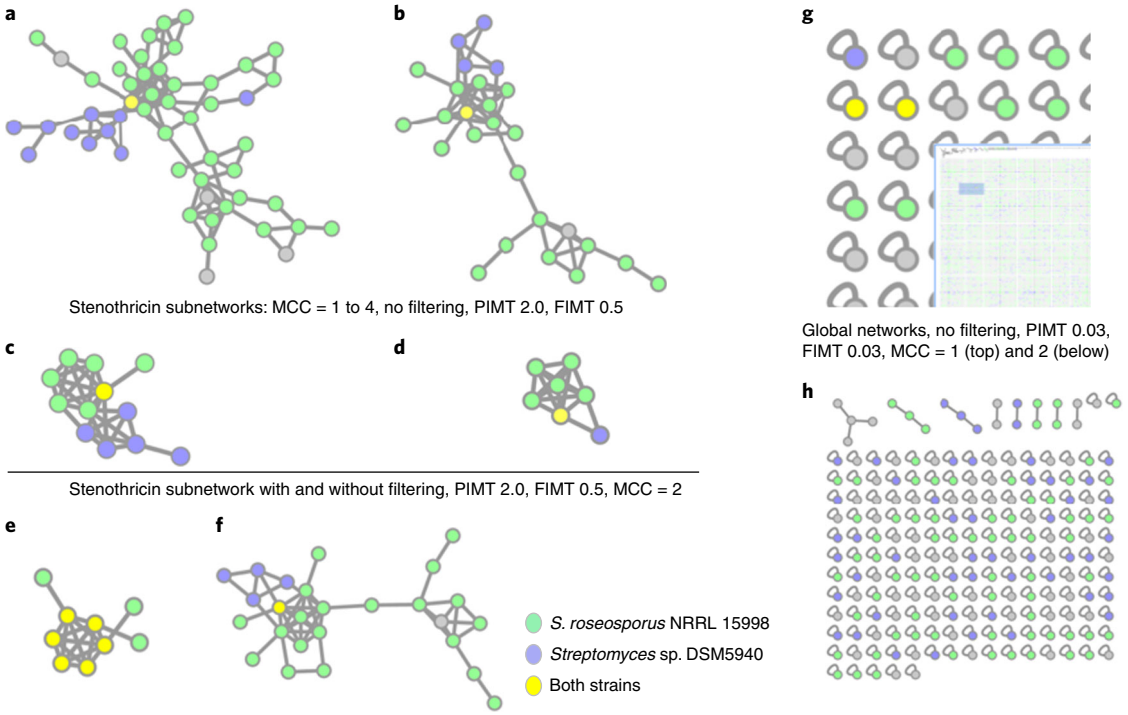


Fig. 7 | Networking of the stenothricin natural product molecular family (MSV000083381) detected in *Streptomyces* sp. DSM5940 (purple nodes), *S. roseosporus* NRRL 15998 (green nodes) or both strains (yellow nodes). Samples were extracted as described previously[11]; briefly, agar was sectioned into 1:1 water/*n*-butanol, shaken for 12 h, and then the organic layer was collected, centrifuged, and dried. Variation in number of nodes and spectra with 'Minimum Consensus Cluster Size' (MCC) yields subnetworks: **a**, MCC = 1, 52 nodes, 169 spectra; **b**, MCC = 2, 29 nodes, 144 spectra; **c**, MCC = 3, 12 nodes, 89 spectra; **d**, MCC = 4, 7 nodes, 73 spectra (no filtering). **e**,**f**, Selecting advanced filtering options results in 9 nodes (**e**), compared to 26 nodes (**f**). **g**,**h**, High-resolution settings for PIMT (0.03) and FIMT (0.03) reduce stenothricin annotations with MCC = 1 providing two stenothricin nodes of 7,642 total (**g**), and (**h**) MCC = 2 giving no stenothricin annotations and only 192 nodes (**h**). FIMT, fragment ion mass tolerance; PIMT, parent ion mass tolerance.

matches the exact mass of the conjugated amino acid. In addition to the conjugated bile acids, the user can also find hits for cholic acid and deoxycholic acid in the network. These compounds are present only in colonized mice, because microbes deconjugate tauro- and glyco-conjugated bile acids in the duodenum.

In addition to the tutorial example, which highlights how molecular networking can be used for the discovery of new endogenous metabolites related to human health, two more examples are presented from published studies[11,47]. One highlights the use of molecular networking in NP discovery and the other integrates metabolomic and microbiome data into 3D maps. The molecular networking workflow in GNPS continues to be updated and additional reference library entries are continually added by the GNPS community, which may result in some new network annotations since the original publication. The current reference libraries used (curated in speclibs, https://gnps.ucsd.edu/ProteoSAFe/libraries.jsp, December 2018) are listed in the supporting information (Supplementary Table 3).

To illustrate the utility of GNPS in revealing the extent of suites of related NPs, the discovery of new stenothricins-GNPS 1-5 from *Streptomyces* strains reported in Wang et al.[11] is revisited here. The dataset MSV000083381 comprises MS^2 data for *n*-butanol and methanol extracts from each of *Streptomyces* sp. DSM5940 and *Streptomyces roseosporus* NRRL 15998 cultures grown on solid agar, together with a metadata table that links each of the four MS^2 data files with the originating *Streptomyces* strain. In reproducing the observation of a distinct subnetwork comprising the MS^2 data from *Streptomyces* sp. DSM5940 connected to known *S. roseosporus* stenothricin analogs, we highlight the effect of minimum consensus cluster size, PIMT and FIMT settings, and advanced filtering options (Fig. 7). Importantly, the choice of low-resolution settings for PIMT (2.0) and FIMT (0.5) to facilitate library searching enables annotation of multiple stenothricin analogs in an expansive subnetwork, which is otherwise lost with more stringent mass tolerance settings of 0.03. Minimum consensus cluster size also has a pronounced effect on the range of stenothricin analogs detected. As is common for many NP molecular families, a few major stenothricin analogs are probably

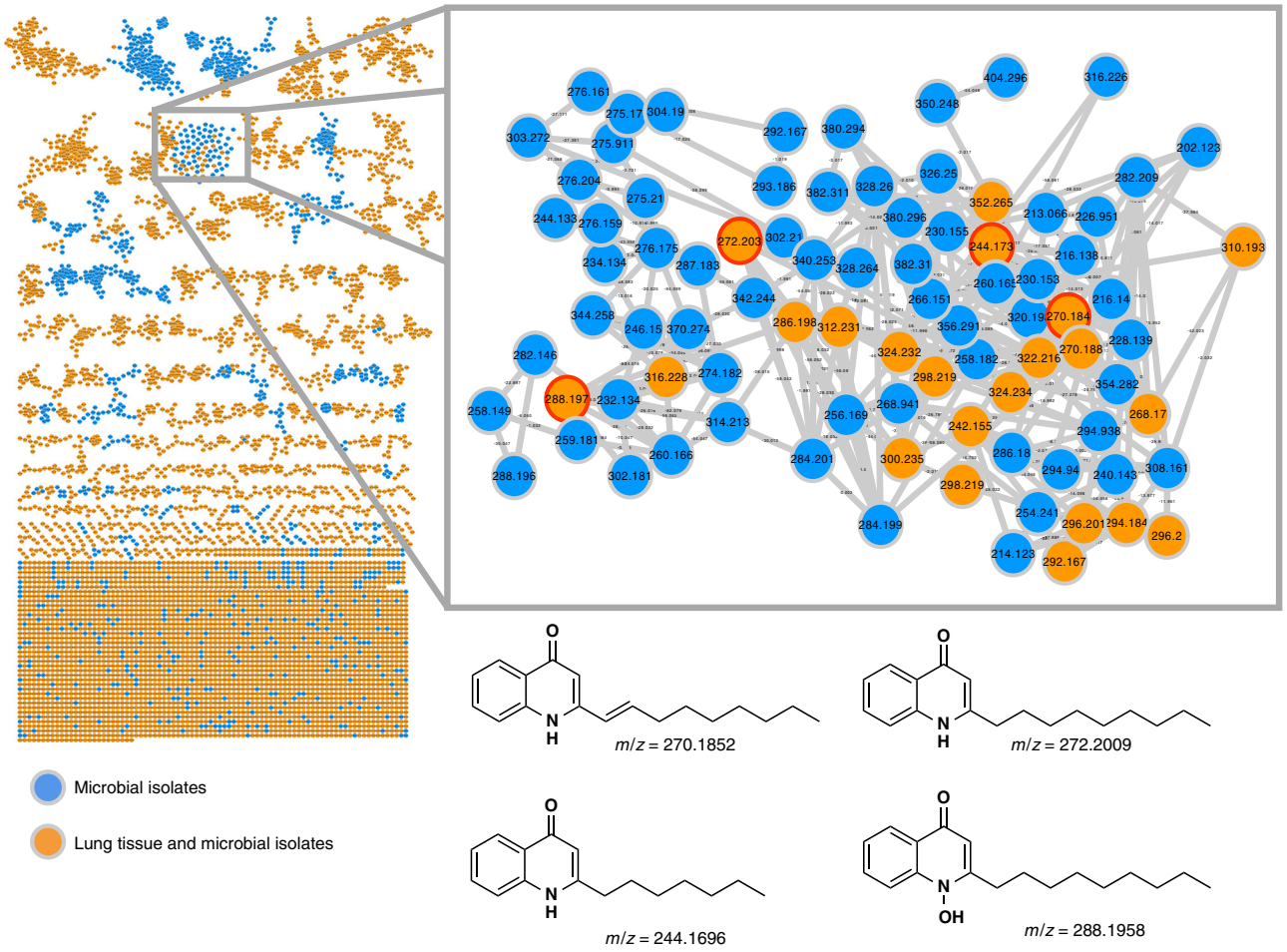


Fig. 8 | Molecular family (a subnetwork) of quinolones detected in lung tissue extracts and cultured *Pseudomonas* isolates created from MassIVE dataset MSV000083359. Lung tissue extracts were prepared using a 1:1:1 chloroform/methanol/water extraction followed by a 2:2:1 ethyl acetate/methanol/water extraction; extracts were combined and then dried[47]. Microbial isolates were prepared from sectioned tissues swabbed with sterile, prepared swabs; then isolates were cultured from swabs and metabolite extraction was performed as with lung tissue extracts[47]. 2-heptyl-4-quinolone (HHQ), 2-nonyl-4-quinolone (NHQ) and its unsaturated derivative (NHQ-C9:1 db), and 2-nonyl-4-quinolone-N-oxide (NQNO) were found in lung tissue and are highlighted by a red node border. Lung tissue extracts (orange nodes); cultured *Pseudomonas* isolates (cyan nodes).

accompanied by numerous minor stenothricins, for which the MS^2 spectra generated readily fall below the threshold for representation as a node. The distinct clustering of stenothricins from *Streptomyces* sp. DSM5940 in Fig. 7a is because the parent ion *m/z* values for these nodes are 41 Da less than the corresponding values for the known *S. roseosporus* stenothricin compounds, consistent with the substitution of serine for lysine in stenothricin-GNPS 1-5[11].

To further illustrate that molecular networking in GNPS can be used for a diverse range of applications, we highlight that molecular networking can be used to visualize quinolones produced by *Pseudomonas* isolated from a patient lung[47]. Figure 8 reproduces the previous analysis (MSV000083359), where the orange nodes represent quinolones detected in both lung tissue extracts and cultured microbial isolates, whereas cyan nodes represent those only detected in cultured microbial isolates.

With a network in hand, there are a number of data analysis tools and experimental validation steps that can be performed. As discussed in Steps 41–43, to legitimize a library annotation beyond inspecting mirror plots, the user should verify the molecular formula and identify associated adducts using MS^1 data. In addition, rationalization based on biological source is recommended. Ideally, an annotation is authenticated by comparison with a known standard compound or isolation and full characterization. In the example followed throughout the protocol, the molecular structures of the new conjugated bile acids from the mouse duodenum were confirmed by comparison with synthetic standards. For more complex structures such as those in the stenothricin example[11] (Fig. 6), the most abundant analog, stenothricin-GNPS 2, was purified for acquisition. The structure was assigned from

1D and 2D NMR data, Marfey's analysis[127], and manual comparison of the MS^2 spectra with MS^2 spectra for previously reported stenothricin D. Genome mining further supported the conclusion that the −41 Da mass shift observed for stenothricin-GNPS 1-5 is due to a Lys to Ser substitution. For nodes that are not annotated, the in silico Dereplicator may predict peptidic NPs, whereas NAP (Network Annotation Propagation) can use annotated nodes to predict related metabolites. Molecular formulas can be generated by using additional tools, one of which is SIRIUS[110]. This software uses MS^2 features to arrive at the best molecular formula for the precursor MS^1 ion and works best for smaller molecules (<600 Da).

In the example of the human lung colonized by *Pseudomonas* bacteria (Fig. 8; ref. [47]), the authors use spatial mapping to visualize annotated molecules on an exploded lung and then correlate the distribution of molecules to microbiome maps generated from 16S rRNA gene amplicon sequencing. This study shows how molecular networking can be used to elucidate spatial variation in chemical profiles and how this can be correlated with microbial makeup using 3D maps. Statistical analyses of microbiome sequence data were performed in QIIME2; a number of additional statistical tools were used as well. Ongoing developments in GNPS include the integration of some of these statistical analysis tools into GNPS. Ultimately, it is envisioned that streamlined integration of pre- and post-networking tools with the GNPS platform will facilitate both creation and mining of molecular networks.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All LC–MS data used in this paper are publicly available at the GNPS-MassIVE repository under the following accession numbers.

MSV000083437 (GF and SPF mice, data not shown)
MSV000083359 (3D cartography of diseased human lung[47])
MSV000083381 (stenothricin-GNPS analogs[11])

References

1. Watrous, J. et al. Mass spectral molecular networking of living microbial colonies. *Proc. Natl Acad. Sci. USA* **109**, E1743–E1752 (2012).
2. Traxler, M. F. & Kolter, R. A massively spectacular view of the chemical lives of microbes. *Proc. Natl Acad. Sci. USA* **109**, 10128–10129 (2012).
3. Fox Ramos, A. E., Evanno, L., Poupon, E., Champy, P. & Beniddir, M. A. Natural products targeting strategies involving molecular networking: different manners, one goal. *Nat. Prod. Rep.* **36**, 960–980 (2019).
4. Teta, R. et al. A joint molecular networking study of a *Smenospongia* sponge and a cyanobacterial bloom revealed new antiproliferative chlorinated polyketides. *Org. Chem. Front* **6**, 1762–1774 (2019).
5. Kalinski, J. J. et al. Molecular networking reveals two distinct chemotypes in pyrroloiminoquinone-producing *Tsitsikamma favus* sponges. *Mar. Drugs* **17**, 60 (2019).
6. Raheem, D. J., Tawfike, A. F., Abdelmohsen, U. R., Edrada-Ebel, R. & Fitzsimmons-Thoss, V. Application of metabolomics and molecular networking in investigating the chemical profile and antitrypanosomal activity of British bluebells (*Hyacinthoides non-scripta*). *Sci. Rep.* **9**, 2547 (2019).
7. Trautman, E. P., Healy, A. R., Shine, E. E., Herzon, S. B. & Crawford, J. M. Domain-targeted metabolomics delineates the heterocycle assembly steps of colibactin biosynthesis. *J. Am. Chem. Soc.* **139**, 4195–4201 (2017).
8. Vizcaino, M. I., Engel, P., Trautman, E. & Crawford, J. M. Comparative metabolomics and structural characterizations illuminate colibactin pathway-dependent small molecules. *J. Am. Chem. Soc.* **136**, 9244–9247 (2014).
9. Nguyen, D. D. et al. Indexing the *Pseudomonas* specialized metabolome enabled the discovery of poaeamide B and the bananamides. *Nat. Microbiol.* **2**, 16197 (2016).
10. Frank, A. M. et al. Clustering millions of tandem mass spectra. *J. Proteome Res.* **7**, 113–122 (2008).
11. Wang, M. et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
12. Frank, A. M. et al. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nat. Methods* **8**, 587–591 (2011).
13. De Vijlder, T. et al. A tutorial in small molecule identification via electrospray ionization-mass spectrometry: the practical art of structural elucidation. *Mass Spectrom. Rev.* **37**, 607–629 (2018).
14. Artyukhin, A. B. et al. Metabolomic "dark matter" dependent on peroxisomal β-oxidation in *Caenorhabditis elegans*. *J. Am. Chem. Soc.* **140**, 2841–2852 (2018).

15. Edwards, E. D., Woolly, E. F., McLellan, R. M. & Keyzers, R. A. Non-detection of honeybee hive contamination following *Vespula* wasp baiting with protein containing fipronil. *PLoS One* **13**, e0206385 (2018).
16. Hoffmann, T. et al. Correlating chemical diversity with taxonomic distance for discovery of natural products in myxobacteria. *Nat. Commun.* **9**, 803 (2018).
17. Leipoldt, F. et al. Warhead biosynthesis and the origin of structural diversity in hydroxamate metalloproteinase inhibitors. *Nat. Commun.* **8**, 1965 (2017).
18. Kang, K. B., Gao, M., Kim, G. J., Choi, H. & Sung, S. H. Rhamnellosides A and B, omega-phenylpentaene fatty acid amide diglycosides from the fruits of *Rhamnella franguloides*. *Molecules* **23**, 752 (2018).
19. Remy, S. et al. Structurally diverse diterpenoids from *Sandwithia guyanensis*. *J. Nat. Prod.* **81**, 901–912 (2018).
20. Riewe, D., Wiebach, J. & Altmann, T. Structure annotation and quantification of wheat seed oxidized lipids by high-resolution LC-MS/MS. *Plant Physiol.* **175**, 600–618 (2017).
21. Senges, C. H. R. et al. The secreted metabolome of *Streptomyces chartreusis* and implications for bacterial chemistry. *Proc. Natl Acad. Sci. USA* **115**, 2490–2495 (2018).
22. van der Hooft, J. J. J. et al. Unsupervised discovery and comparison of structural families across multiple samples in untargeted metabolomics. *Anal. Chem.* **89**, 7569–7577 (2017).
23. Wolff, H. & Bode, H. B. The benzodiazepine-like natural product tilivalline is produced by the entomopathogenic bacterium *Xenorhabdus eapokensis*. *PLoS One* **13**, e0194297 (2018).
24. Schymanski, E. L. et al. Critical assessment of small molecule identification 2016: automated methods. *J. Cheminf.* **9**, 22 (2017).
25. Beniddir, M. MTBLS142: collected tandem mass spectrometry data on monoterpene indole alkaloids from natural product chemistry research. *MetaboLights* https://www.ebi.ac.uk/metabolights/MTBLS142 (2018).
26. Lei, Z. et al. Construction of an ultrahigh pressure liquid chromatography-tandem mass spectral library of plant natural products and comparative spectral analyses. *Anal. Chem.* **87**, 7373–7381 (2015).
27. Nikolic, D., Jones, M., Sumner, L. & Dunn, W. CASMI 2014: challenges, solutions and results. *Curr. Metab.* **5**, 5–17 (2017).
28. Horai, H. et al. MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).
29. Stravs, M. A., Schymanski, E. L., Singer, H. P. & Hollender, J. Automatic recalibration and processing of tandem mass spectra using formula annotation. *J. Mass Spectrom.* **48**, 89–99 (2013).
30. von Eckardstein, L. et al. Total synthesis and biological assessment of novel albicidins discovered by mass spectrometric networking. *Chemistry* **23**, 15316–15321 (2017).
31. Vizcaino, M. I. & Crawford, J. M. The colibactin warhead crosslinks DNA. *Nat. Chem.* **7**, 411–417 (2015).
32. Saleh, H. et al. Deuterium-labeled precursor feeding reveals a new pABA-containing meroterpenoid from the mango pathogen *Xanthomonas citri* pv. *mangiferaeindicae*. *J. Nat. Prod.* **79**, 1532–1537 (2016).
33. Fox Ramos, A. E. et al. Collected mass spectrometry data on monoterpene indole alkaloids from natural product chemistry research. *Sci. Data* **6**, 15 (2019).
34. Aron, A. T. et al. Reproducible Molecular networking of untargeted mass spectrometry data using GNPS. Preprint at https://doi.org/10.26434/chemrxiv.9333212.v1 (2019).
35. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
36. Petras, D. et al. Mass spectrometry-based visualization of molecules associated with human habitats. *Anal. Chem.* **88**, 10775–10784 (2016).
37. Kapono, C. A. et al. Creating a 3D microbial and chemical snapshot of a human habitat. *Sci. Rep.* **8**, 3669 (2018).
38. Adams, R. I. et al. Microbes and associated soluble and volatile chemicals on periodically wet household surfaces. *Microbiome* **5**, 128 (2017).
39. Petras, D. et al. High-resolution liquid chromatography tandem mass spectrometry enables large scale molecular characterization of dissolved organic matter. *Front. Mar. Sci.* **4**, 405 (2017).
40. Trautman, E. P. & Crawford, J. M. Linking biosynthetic gene clusters to their metabolites via pathway-targeted molecular networking. *Curr. Top. Med. Chem.* **16**, 1705–1716 (2016).
41. Luzzatto-Knaan, T., Melnik, A. V. & Dorrestein, P. C. Mass spectrometry uncovers the role of surfactin as an interspecies recruitment factor. *ACS Chem. Biol.* **14**, 459–467 (2019).
42. Machushynets, N. V., Wu, C., Elsayed, S. S., Hankemeier, T. & van Wezel, G. P. Discovery of novel glycerolated quinazolinones from *Streptomyces* sp. MBT27. *J. Ind. Microbiol. Biotechnol.* **46**, 483–492 (2019).
43. Yao, L. et al. Discovery of novel xylosides in co-culture of basidiomycetes *Trametes versicolor* and *Ganoderma applanatum* by integrated metabolomics and bioinformatics. *Sci. Rep.* **6**, 33237 (2016).
44. Tripathi, A. et al. Intermittent hypoxia and hypercapnia, a hallmark of obstructive sleep apnea, alters the gut microbiome and metabolome. *mSystems* **3**, e00020-18 (2018).
45. Smits, S. A. et al. Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. *Science* **357**, 802–806 (2017).
46. McDonald, D. et al. American Gut: an open platform for citizen science microbiome research. *mSystems* **3**, e00031-18 (2018).
47. Garg, N. et al. Three-dimensional microbiome and metabolome cartography of a diseased human lung. *Cell Host Microbe* **22**, 705–716 e704 (2017).

48. Edlund, A. et al. Metabolic fingerprints from the human oral microbiome reveal a vast knowledge gap of secreted small peptidic molecules. *mSystems* **2**, e00058-17 (2017).
49. McCall, L. I. et al. Mass spectrometry-based chemical cartography of a cardiac parasitic infection. *Anal. Chem.* **89**, 10414–10421 (2017).
50. Watrous, J. D. et al. Directed non-targeted mass spectrometry and chemical networking for discovery of eicosanoids and related oxylipins. *Cell Chem. Biol.* **26**, 433–442.e4 (2019).
51. Allard, S., Allard, P. M., Morel, I. & Gicquel, T. Application of a molecular networking approach for clinical and forensic toxicology exemplified in three cases involving 3-MeO-PCP, doxylamine, and chlormequat. *Drug Test. Anal.* **11**, 669–677 (2018).
52. Ernst, M. et al. Assessing specialized metabolite diversity in the cosmopolitan plant genus *Euphorbia* L. *Front. Plant Sci.* **10**, 846 (2019).
53. Philippus, A. C. et al. Molecular networking prospection and characterization of terpenoids and C15-acetogenins in Brazilian seaweed extracts. *RSC Adv.* **8**, 29654–29661 (2018).
54. Li, F., Janussen, D., Peifer, C., Perez-Victoria, I. & Tasdemir, D. Targeted isolation of tsitsikammamines from the Antarctic deep-sea sponge *Latrunculia biformis* by molecular networking and anticancer activity. *Mar. Drugs* **16**, 268 (2018).
55. Hartmann, A. C. et al. Meta-mass shift chemical profiling of metabolomes from coral reefs. *Proc. Natl Acad. Sci. USA* **114**, 11685–11690 (2017).
56. Tobias, N. J. et al. Natural product diversity associated with the nematode symbionts *Photorhabdus* and *Xenorhabdus*. *Nat. Microbiol.* **2**, 1676–1685 (2017).
57. Nothias, L. F. et al. Bioactivity-based molecular networking for the discovery of drug leads in natural product bioassay-guided fractionation. *J. Nat. Prod.* **81**, 758–767 (2018).
58. Zou, Y. et al. Computationally assisted discovery and assignment of a highly strained and PANC-1 selective alkaloid from Alaska's deep ocean. *J. Am. Chem. Soc.* **141**, 4338–4344 (2019).
59. Parkinson, E. I. et al. Discovery of the tyrobetaine natural products and their biosynthetic gene cluster via metabologenomics. *ACS Chem. Biol.* **13**, 1029–1037 (2018).
60. Naman, C. B. et al. Integrating molecular networking and biological assays to target the isolation of a cytotoxic cyclic octapeptide, samoamide A, from an American Samoan marine cyanobacterium. *J. Nat. Prod.* **80**, 625–633 (2017).
61. Bouslimani, A. et al. Lifestyle chemistries from phones for individual profiling. *Proc. Natl Acad. Sci. USA* **113**, E7645–E7654 (2016).
62. Fox Ramos, A. E. et al. CANPA: computer-assisted natural products anticipation. *Anal. Chem.* **91**, 11247–11252 (2019).
63. Quinn, R. A. et al. Niche partitioning of a pathogenic microbiome driven by chemical gradients. *Sci. Adv.* **4**, eaau1908 (2018).
64. Aksenov, A. A., da Silva, R., Knight, R., Lopes, N. P. & Dorrestein, P. C. Global chemical analysis of biology by mass spectrometry. *Nat. Rev. Chem.* **1**, 0054 (2017).
65. Tsugawa, H. Advances in computational metabolomics and databases deepen the understanding of metabolisms. *Curr. Opin. Biotechnol.* **54**, 10–17 (2018).
66. Johnson, S. R. & Lange, B. M. Open-access metabolomics databases for natural product research: present capabilities and future potential. *Front. Bioeng. Biotechnol.* **3**, 22 (2015).
67. Haug, K. et al. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **41**, D781–D786 (2013).
68. Perez-Riverol, Y. et al. Discovering and linking public omics data sets using the Omics Discovery Index. *Nat. Biotechnol.* **35**, 406–409 (2017).
69. Stein, S. E. & Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **5**, 859–866 (1994).
70. Mohimani, H. & Pevzner, P. A. Dereplication, sequencing and identification of peptidic natural products: from genome mining to peptidogenomics to spectal networks. *Nat. Prod. Rep.* **33**, 73–86 (2016).
71. Yang, J. Y. et al. Molecular networking as a dereplication strategy. *J. Nat. Prod.* **76**, 1686–1699 (2013).
72. Moorthy, A. S., Wallace, W. E., Kearsley, A. J., Tchekhovskoi, D. V. & Stein, S. E. Combining fragment-ion and neutral-loss matching during mass spectral library searching: a new general purpose algorithm applicable to illicit drug identification. *Anal. Chem.* **89**, 13261–13268 (2017).
73. Klinman, J. P. The multi-functional topa-quinone copper amine oxidases. *Biochim. Biophys. Acta* **1637**, 131–137 (2003).
74. Guijas, C. et al. METLIN: a technology platform for identifying knowns and unknowns. *Anal. Chem.* **90**, 3156–3164 (2018).
75. Reddi, A. R. & Culotta, V. C. SOD1 integrates signals from oxygen and glucose to repress respiration. *Cell* **152**, 224–235 (2013).
76. Sheldon, M. T., Mistrik, R. & Croley, T. R. Determination of ion structures in structurally related compounds using precursor ion fingerprinting. *J. Am. Soc. Mass Spectrom.* **20**, 370–376 (2009).
77. Sawada, Y. et al. RIKEN tandem mass spectral database (ReSpect) for phytochemicals: a plant-specific MS/MS-based data resource and database. *Phytochemistry* **82**, 38–45 (2012).
78. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **78**, 779–787 (2006).

79. Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. XCMS Online: a web-based platform to process untargeted metabolomic data. *Anal. Chem.* **84**, 5035–5039 (2012).
80. Wanichthanarak, K., Fan, S., Grapov, D., Barupal, D. K. & Fiehn, O. Metabox: a toolbox for metabolomic data analysis, interpretation and integrative exploration. *PLoS ONE* **12**, e0171046 (2017).
81. Mohimani, H. et al. Dereplication of microbial metabolites through database search of mass spectra. *Nat. Comm.* **9**, 4035 (2018).
82. Mohimani, H. et al. Dereplication of peptidic natural products through database search of mass spectra. *Nat. Chem. Biol.* **13**, 30–37 (2017).
83. Gurevich, A. et al. Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nat. Microbiol.* **3**, 319–327 (2018).
84. da Silva, R. R. et al. Propagating annotations of molecular networks using in silico fragmentation. *PLoS Comput. Biol.* **14**, e1006089 (2018).
85. Mohimani, H. et al. Automated genome mining of ribosomal peptide natural products. *ACS Chem. Biol.* **9**, 1545–1551 (2014).
86. Olivon, F. et al. MetGem software for the generation of molecular networks based on the t-SNE algorithm. *Anal. Chem.* **90**, 13900–13908 (2018).
87. Olivon, F., Roussi, F., Litaudon, M. & Touboul, D. Optimized experimental workflow for tandem mass spectrometry molecular networking in metabolomics. *Anal. Bioanal. Chem.* **409**, 5767–5778 (2017).
88. Wehrens, R. et al. Improved batch correction in untargeted MS-based metabolomics. *Metabolomics* **12**, 88 (2016).
89. Koal, T. & Deigner, H. P. Challenges in mass spectrometry based targeted metabolomics. *Curr. Mol. Med.* **10**, 216–226 (2010).
90. Bylda, C., Thiele, R., Kobold, U. & Volmer, D. A. Recent advances in sample preparation techniques to overcome difficulties encountered during quantitative analysis of small molecules from biofluids using LC-MS/MS. *Analyst* **139**, 2265–2276 (2014).
91. Vuckovic, D. Current trends and challenges in sample preparation for global metabolomics using liquid chromatography-mass spectrometry. *Anal. Bioanal. Chem.* **403**, 1523–1548 (2012).
92. Dunn, W. B. et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* **6**, 1060 (2011).
93. Taylor, P. J. Matrix effects: the Achilles heel of quantitative high-performance liquid chromatography-electrospray-tandem mass spectrometry. *Clin. Biochem.* **38**, 328–334 (2005).
94. Annesley, T. M. Ion suppression in mass spectrometry. *Clin. Chem.* **49**, 1041–1044 (2003).
95. Crüsemann, M. et al. Prioritizing natural product diversity in a collection of 146 bacterial strains based on growth and extraction protocols. *J. Nat. Prod.* **80**, 588–597 (2017).
96. Wandro, S., Carmody, L., Gallagher, T., LiPuma, J. J. & Whiteson, K. Making it last: storage time and temperature have differential impacts on metabolite profiles of airway samples from cystic fibrosis patients. *mSystems* **2**, e00100-17 (2017).
97. Zhao, J., Evans, C. R., Carmody, L. A. & LiPuma, J. J. Impact of storage conditions on metabolite profiles of sputum samples from persons with cystic fibrosis. *J. Cyst. Fibros.* **14**, 468–473 (2015).
98. Hirayama, A. et al. Effects of processing and storage conditions on charged metabolomic profiles in blood. *Electrophoresis* **36**, 2148–2155 (2015).
99. Mushtaq, M. Y., Choi, Y. H., Verpoorte, R. & Wilson, E. G. Extraction for metabolomics: access to the metabolome. *Phytochem. Anal.* **25**, 291–306 (2014).
100. Scheubert, K. et al. Significance estimation for large scale metabolomics annotations by spectral matching. *Nat. Commun.* **8**, 1494 (2017).
101. Sleno, L. & Volmer, D. A. Ion activation methods for tandem mass spectrometry. *J. Mass. Spectrom.* **39**, 1091–1112 (2004).
102. Tang, Z. & Guengerich, F. P. Dansylation of unactivated alcohols for improved mass spectral sensitivity and application to analysis of cytochrome P450 oxidation products in tissue extracts. *Anal. Chem.* **82**, 7706–7712 (2010).
103. Bazsó, F. L. et al. Quantitative comparison of tandem mass spectra obtained on various instruments. *J. Am. So. c. Mass Spectrom.* **27**, 1357–1365 (2016).
104. Bowen, B. P. & Northen, T. R. Dealing with the unknown: metabolomics and metabolite atlases. *J. Am. Soc. Mass Spectrom.* **21**, 1471–1476 (2010).
105. da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proc. Natl Acad. Sci. USA* **112**, 12549–12550 (2015).
106. Blaženović, I., Kind, T., Ji, J. & Fiehn, O. Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites* **8**, 31 (2018).
107. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J. Cheminf.* **8**, 3 (2016).
108. Gerlich, M. & Neumann, S. MetFusion: integration of compound identification strategies. *J. Mass Spectrom.* **48**, 291–298 (2013).
109. Böcker, S., Letzel, M. C., Liptak, Z. & Pervukhin, A. SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics* **25**, 218–224 (2009).
110. Dührkop, K. et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).

111. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Bocker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl Acad. Sci. USA* **112**, 12580–12585 (2015).
112. Tsugawa, H. et al. Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. *Anal. Chem.* **88**, 7946–7958 (2016).
113. Protsyuk, I. et al. 3D molecular cartography using LC-MS facilitated by Optimus and 'ili software. *Nat. Protoc.* **13**, 134–154 (2018).
114. Röst, H. L. et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* **13**, 741–748 (2016).
115. Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinforma.* **11**, 395 (2010).
116. Deutsch, E. W. et al. Proteomics Standards Initiative: fifteen years of progress and future work. *J. Proteome Res.* **16**, 4288–4298 (2017).
117. Brooksbank, C., Cameron, G. & Thornton, J. The European Bioinformatics Institute's data resources. *Nucleic Acids Re. s.* **38**, D17–D25 (2010).
118. McLafferty, F. W. & Tureček, F. *Interpretation of Mass Spectra* 4th edn (University Science Books, 1993).
119. Cleary, J. L., Luu, G. T., Pierce, E. C., Dutton, R. J. & Sanchez, L. M. BLANKA: an algorithm for blank subtraction in mass spectrometry of complex biological samples. *J. Am. Soc. Mass Spectrom.* **30**, 1426–1434 (2019).
120. Sumner, L. W. et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **3**, 211–221 (2007).
121. Viant, M. R., Kurland, I. J., Jones, M. R. & Dunn, W. B. How close are we to complete annotation of metabolomes? *Curr. Opin. Chem. Biol.* **36**, 64–69 (2017).
122. Wang, J., Peake, D. A., Mistrik, R., Huang, Y. & Araujo, G. D. *A Platform to Identify Endogenous Metabolites Using a Novel High Performance Orbitrap MS and the mzCloud Library*. http://www.unitylabservices.eu/content/dam/tfs/ATG/CMD/CMD%20Documents/posters/PN-ASMS13-a-platform-to-identify-endogenous-metabolites-using-a-novel-high-performance-orbitrap-and-the-mzcloud-library-E.pdf (Thermo Scientific, 2013).
123. Shahaf, N. et al. The WEIZMASS spectral library for high-confidence metabolite identification. *Nat. Commun.* **7**, 12423 (2016).
124. Schymanski, E. L. et al. Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ. Sci. Technol.* **48**, 2097–2098 (2014).
125. Demarque, D. P., Crotti, A. E. M., Vessecchi, R., Lopes, J. L. C. & Lopes, N. P. Fragmentation reactions using electrospray ionization mass spectrometry: an important tool for the structural elucidation and characterization of synthetic and natural products. *Nat. Prod. Rep.* **33**, 432–455 (2016).
126. van der Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E. V. & Rogers, S. Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl Acad. Sci. USA* **113**, 13738–13743 (2016).
127. Marfey, P. Determination of D-amino acids .2. Use of a bifunctional reagent, 1,5-difluoro-2,4-dinitrobenzene. *Carlsberg Res. Commun.* **49**, 591–596 (1984).
128. Su, G., Morris, J. H., Demchak, B. & Bader, G. D. Biological network exploration with Cytoscape 3. *Curr. Protoc. Bioinforma.* **47**, 8 13 11–24 (2014).
129. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431–432 (2011).
130. Sandhu, C. et al. Evaluation of data-dependent versus targeted shotgun proteomic approaches for monitoring transcription factor expression in breast cancer. *J. Proteome Res.* **7**, 1529–1541 (2008).
131. Hubert, J., Nuzillard, J.-M. & Renault, J.-H. Dereplication strategies in natural product research: how many tools and methodologies behind the same concept? *Phytochem. Rev.* **16**, 55–95 (2017).
132. Rochat, B. Proposed confidence scale and ID score in the identification of known-unknown compounds using high resolution MS data. *J. Am. Soc. Mass Spectrom.* **28**, 709–723 (2017).
133. All natural. *Nat. Chem. Biol.* **3**, 351 (2007).
134. IUPAC (International Union of Pure and Applied Chemistry). *Compendium of Chemical Terminology—The "Gold Book"* (eds McNaught, A. D. & Wilkinson, A.) (Blackwell Scientific Publications, 1997).
135. McLafferty, F. W. Tandem mass spectrometry. *Science* **214**, 280–287 (1981).
136. Gross, J. H. *Mass Spectrometry: A Textbook* 415-478 (Springer, 2011).
137. Vazquez-Baeza, Y., Pirrung, M., Gonzalez, A. & Knight, R. EMPeror: a tool for visualizing high-throughput microbial community data. *GigaScience* **2**, 16 (2013).
138. McDonald, D. et al. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience* **1**, 7 (2012).
139. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
140. Wang, M. et al. Mass spectrometry searches using MASST. *Nat. Biotechnol.* **38**, 23–26 (2020).
141. Jarmusch, A. K. *et al.* Repository-scale co- and re-analysis of tandem mass spectrometry data. Preprint at https://www.biorxiv.org/content/10.1101/750471v1 (2019).
142. Olivon, F., Grelier, G., Roussi, F., Litaudon, M. & Touboul, D. MZmine 2 data-preprocessing to enhance molecular networking reliability. *Anal. Chem.* **89**, 7836–7840 (2017).
143. Winnikoff, J. R., Glukhov, E., Watrous, J., Dorrestein, P. C. & Gerwick, W. H. Quantitative molecular networking to profile marine cyanobacterial metabolomes. *J. Antibiot. (Tokyo)* **67**, 105–112 (2014).

144. Tsugawa, H. et al. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* **12**, 523–526 (2015).
145. Jones, A. R. et al. The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol. Cell. Proteom.* **11**, M111.014381 (2012).
146. Griss, J. et al. The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol. Cell. Proteom.* **13**, 2765–2775 (2014).
147. Hoffmann, N. et al. mzTab-M: a data standard for sharing quantitative results in mass spectrometry metabolomics. *Anal. Chem.* **91**, 3302–3310 (2019).

Acknowledgements

We acknowledge funding from the following: National Research System (SNI) of SENACYT Panama (C.A.B.P., M.H.C., J.L.-B., and M. G.); the Gordon and Betty Moore Foundation (P.C.D., N.B., and K.L.M.); the National Institutes of Health (GM122016-01; K.L.M.); the National Science Foundation (DEB1354944; R.M.T.) and (IOS-1656481; A.M.C.R and P.C.D). A.K.J. acknowledges an American Society for Mass Spectrometry 2018 Postdoctoral Career Development Award. D.P. was supported through the Deutsche Forschungsgemeinschaft (DFG; PE 2600/1). F.T. and N.N. acknowledge Shimadzu South Africa (Pty) Ltd for the support and training. We are grateful for grant R03 CA211211 (P.C.D.) on reuse of metabolomics data and grant P41 GM103484 (P.C.D., N.B.) to the Center for Computational Mass Spectrometry, as well as instrument support though NIH S10RR029121 (P.C.D.). A.I.C. and Y.Z. were supported through an Auburn University Presidential Award for Interdisciplinary Research (PAIR).

Author contributions

Design and oversight of the project: P.C.D., M.W., N.B. Instrument acquisition parameters: A.T.A., E.C.G., R.A.K., K.L.M., R.M.T., K.B. K., S.B., C.R., A.W.T., F.T., N.N., A.K.J., A.M.U. Data conversion and upload: K.L.M., E.C.G., A.T.A., J.J.J.v.d.H., M.E. GNPS documentation: M.W., L.F.N., E.C.G., A.T.A., K.L.M., J.J.J.v.d.H., M.E., M.N.-E. Cytoscape documentation: M.N.-E., F.V., K.C.W., I.K., A.M. C.-R. Metadata curation: J.M.G., C.M.A., F.V., A.M.C.-R. Mass spectra annotations: D.P., R.S., M.E. Theoretical tools and advanced features, statistical analysis: L.F.N., A.A.A. Supplementary information: A.T.A., N.S., E.C.G., K.L.M., M.E. Testing the workflows described and improving the descriptions: Y.Z., A.I.C., A.B., K.S., N.T., A.M.U., J.A.T.M., M.H.C., C.A.B.P., M.G., V.V.-C., J.L.-B., R.M.-F., M.E.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41596-020-0317-5.

Correspondence and requests for materials should be addressed to N.B. or M.W. or P.C.D.

Peer review information *Nature Protocols* thanks Vinayak Agarwal, Mehdi Beniddir, Alfonso Mangoni and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 5 June 2019; Accepted: 3 March 2020;
Published online: 13 May 2020

Related links

Key reference(s) using this protocol

Vermeeren, P., Sun, X. & Bickelhaupt, F.M. *Sci. Rep.* **8**, 10729 (2018): https://doi.org/10.1038/s41598-018-28998-3

Sun, X., Soini, T. M., Poater, J., Hamlin, T. A. & Bickelhaupt, F. M. *J. Comput. Chem.* **40**, 2227–2233 (2019): https://doi.org/10.1002/jcc.25871

Key data used in this protocol

Quinn, R. A. et al. *Sci. Adv.* **4**, eaau1908 (2018): https://doi.org/10.1126/sciadv.aau1908

Wang, M. et al. *Nat. Biotechnol.* **34**, 828–837 (2016): https://doi.org/10.1038/nbt.3597

Garg, N. et al. *Cell Host Microbe* **22**, 705–716.e4 (2017): https://doi.org/10.1016/j.chom.2017.10.001

Preprint version of this protocol

Aron, A. T. et al. Preprint at https://doi.org/10.26434/chemrxiv.9333212.v1 (2019)

nature research

Corresponding author(s): Professor Pieter Dorrestein

Last updated by author(s): October 11, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data was collected in the original study using commercial Bruker software (DataAnalysis). No new data was collected for the tutorial example.
Data analysis	Global Natural Products Social Molecular Networking platform (GNPS, https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash2.jsp) was used for data analysis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data provided can be accessed through the accession code (accession code starting with MSV).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size used for the main study detailed in this protocols paper is only a subset of those samples collected for the original study. Three samples from each cohort were chosen for the example; this is a suitable size because statistics were not done for the tutorial protocol example.
Data exclusions	Data was excluded for the tutorial example, as only a subset of samples collected were relevant for the tutorial example.
Replication	The reproducibility of the tutorial examples was verified by each co-first author, who independently ran data analysis on the data collected for the original study.
Randomization	Samples in the original study were allocated into groups based on their origin - germ-free (GF) mice were grouped and specific pathogen free (SPF) were grouped.
Blinding	The tutorial example performs data analysis on samples run previously; data analysis for the tutorial example was not performed blindly because knowledge of sample grouping and information is required.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Germ-free (GF) C57Bl/6J mice and conventionally-colonized specific pathogen free (SPF) mice (C57Bl/6J) were used in the original study; samples were acquired from 8-week-old female mice.
Wild animals	No wild animals were used.
Field-collected samples	No field-collected samples were used.
Ethics oversight	Animal experiments performed for the original study were approved by the California Institute of 55 Technology's Institutional Animal Care and Use Committee (IACUC).

Note that full information on the approval of the study protocol must also be provided in the manuscript.