Wang, Mingxun and Carver, Jeremy J and Phelan, Vanessa V and Sanchez, Laura M and Garg, Neha and Peng, Yao and Nguyen, Don Duy and Watrous, Jeramie and Kapono, Clifford A and Luzzatto-Knaan, Tal and Porto, Carla and Bouslimani, Amina and Melnik, Alexey V and Meehan, Michael J and Liu, Wei-Ting and Crüsemann, Max and Boudreau, Paul D and Esquenazi, Eduardo and Sandoval-Calderón, Mario and Kersten, Roland D and Pace, Laura A and Quinn, Robert A and Duncan, Katherine R and Hsu, Cheng-Chih and Floros, Dimitrios J and Gavilan, Ronnie G and Kleigrewe, Karin and Northen, Trent and Dutton, Rachel J and Parrot, Delphine and Carlson, Erin E and Aigle, Bertrand and Michelsen, Charlotte F and Jelsbak, Lars and Sohlenkamp, Christian and Pevzner, Pavel and Edlund, Anna and McLean, Jeffrey and Piel, Jörn and Murphy, Brian T and Gerwick, Lena and Liaw, Chih-Chuang and Yang, Yu-Liang and Humpf, Hans-Ulrich and Maansson, Maria and Keyzers, Robert A and Sims, Amy C and Johnson, Andrew R and Sidebottom, Ashley M and Sedio, Brian E and Klitgaard, Andreas and Larson, Charles B and Boya P, Cristopher A and Torres-Mendoza, Daniel and Gonzalez, David J and Silva, Denise B and Marques, Lucas M and Demarque, Daniel P and Pociute, Egle and O'Neill, Ellis C and Briand, Enora and Helfrich, Eric J N and Granatosky, Eve A and Glukhov, Evgenia and Ryffel, Florian and Houson, Hailey and Mohimani, Hosein and Kharbush, Jenan J and Zeng, Yi and Vorholt, Julia A and Kurita, Kenji L and Charusanti, Pep and McPhail, Kerry L and Nielsen, Kristian Fog and Vuong, Lisa and Elfeki, Maryam and Traxler, Matthew F and Engene, Niclas and Koyama, Nobuhiro and Vining, Oliver B and Baric, Ralph and Silva, Ricardo R and Mascuch, Samantha J and Tomasi, Sophie and Jenkins, Stefan and Macherla, Venkat and Hoffman, Thomas and Agarwal, Vinayak and Williams, Philip G and Dai, Jingqui and

Sharing and community curation of mass spectra by GNPS

Mingxun Wang[1,2], Jeremy J Carver[1,2], Vanessa V Phelan[3], Laura M Sanchez[3], Neha Garg[3], Yao Peng[4], Don Duy Nguyen[4], Jeramie Watrous[3], Clifford A Kapono[4], Tal Luzzatto-Knaan[3], Carla Porto[3], Amina Bouslimani[3], Alexey V Melnik[3], Michael J Meehan[3], Wei-Ting Liu[5], Max Crüsemann [6], Paul D Boudreau[6], Eduardo Esquenazi[7], Mario Sandoval-Calderón[8], Roland D Kersten[9], Laura A Pace[3], Robert A Quinn[10], Katherine R Duncan[11,6], Cheng-Chih Hsu[4], Dimitrios J Floros[4], Ronnie G Gavilan[12], Karin Kleigrewe[6], Trent Northen[13], Rachel J Dutton[14], Delphine Parrot[15], Erin E Carlson[16], Bertrand Aigle[17], Charlotte F Michelsen[18], Lars Jelsbak[18], Christian Sohlenkamp[8], Pavel Pevzner[2,1], Anna Edlund[19,20], Jeffrey McLean[21,20], Jörn Piel[22], Brian T Murphy[23], Lena Gerwick[6], Chih-Chuang Liaw[24], Yu-Liang Yang[25], Hans-Ulrich Humpf[26], Maria Maansson[18], Robert A Keyzers[27], Amy C Sims[28], Andrew R. Johnson[29], Ashley M Sidebottom[29], Brian E Sedio[30,12], Andreas Klitgaard[18], Charles B Larson[6,31], Christopher A Boya P.[12], Daniel Torres-Mendoza[12], David J Gonzalez[31,3], Denise B Silva[32,33], Lucas M Marques[32], Daniel P Demarque[32], Egle Pociute[7], Ellis C O'Neill[6], Enora Briand[6,34], Eric J. N. Helfrich[22], Eve A Granatosky[35], Evgenia Glukhov[6], Florian Ryffel[22], Hailey Houson[7], Hosein Mohimani[2], Jenan J Kharbush[6], Yi Zeng[4], Julia A Vorholt[22], Kenji L Kurita[36], Pep Charusanti[37], Kerry L McPhail[38], Kristian Fog Nielsen[18], Lisa Vuong[7], Maryam Elfeki[23], Matthew F Traxler[39], Niclas Engene[40], Nobuhiro Koyama[3], Oliver B Vining[38], Ralph Baric[28], Ricardo R Silva[32], Samantha J Mascuch[6], Sophie Tomasi[15], Stefan Jenkins[13], Venkat Macherla[7], Thomas Hoffman[41], Vinayak Agarwal[42], Philip G Williams[43], Jingqui Dai[43], Ram Neupane[43], Joshua Gurr[43], Andrés M. C. Rodríguez[32], Anne Lamsa[44], Chen Zhang[45], Kathleen Dorrestein[3], Brendan M Duggan[3], Jehad Almaliti[3], Pierre-Marie Allard[46], Prasad Phapale[47], Louis-Felix Nothias[48], Theodore Alexandrov[47], Marc Litaudon[48], Jean-Luc Wolfender[46], Jennifer E Kyle[49], Thomas O Metz[49], Tyler Peryea[50], Dac-Trung Nguyen[50], Danielle VanLeer[50], Paul Shinn[50], Ajit Jadhav[50], Rolf Müller[41], Katrina M Waters[49], Wenyuan Shi[20], Xueting Liu[51], Lixin Zhang[51], Rob Knight[52], Paul R Jensen[6], Bernhard O Palsson[37], Kit Pogliano[44], Roger G Linington[36], Marcelino Gutiérrez[12], Norberto P Lopes[32], William H Gerwick[3,6], Bradley S Moore[3,6,42], Pieter C Dorrestein[3,6,31], Nuno Bandeira[2,3,31]

[1]Computer Science and Engineering, UC San Diego, La Jolla, United States
[2]Center for Computational Mass Spectrometry, UC San Diego, La Jolla, United States
[3]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, UC San Diego, La Jolla, United States
[4]Department of Chemistry and Biochemistry, UC San Diego, La Jolla, United States
[5]Department of Microbiology and Immunology, Stanford University, Palo Alto, United States
[6]Center for Marine Biotechnology and Biomedicine, Scripps Institute of Oceanography, UC San Diego, La Jolla, United States
[7]Sirenas Marine Discovery, San Diego, United States
[8]Centro de Ciencias Genómicas, Universidad Nacional Autonoma de Mexico, Cuernavaca, Mexico
[9]Salk Institute, Salk Institute, La Jolla, United States
[10]Biology Department, San Diego State University, San Diego, United States
[11]Scottish Association for Marine Science, Scottish Marine Institute, Oban, United Kingdom

[12]Center for Drug Discovery and Biodiversity, INDICASAT, City of Knowledge, Panama
[13]Genome Dynamics, Lawrence Berkeley National Laboratory, Berkeley, United States
[14]FAS Center for Systems Biology, Harvard, Cambridge, United States
[15]Produits naturels – Synthèses – Chimie Médicinale, University of Rennes 1, Rennes Cedex, France
[16]Chemistry, University of Minnesota, Minneapolis, United States
[17]Dynamique des Génomes et Adaptation Microbienne, University of Lorraine, Vandœuvre-lès-Nancy, France
[18]Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark
[19]Microbial and Environmental Genomics, J. Craig Venter Institute, La Jolla, United States
[20]School of Dentistry, UC Los Angeles, Los Angeles, United States
[21]Department of Periodontics, University of Washington, Seattle, United States
[22]Institute of Microbiology, ETH Zurich, Zurich, Switzerland
[23]Department of Medicinal Chemistry and Pharmacognosy, University of Illinois Chicago, Chicago, United States
[24]Department of Marine Biotechnology and Resources, National Sun Yat-sen University, Kaohsiung, Taiwan
[25]Agricultural Biotechnology Research Center, Academia Sinica, Taipei, Taiwan
[26]Institute of Food Chemistry, University of Münster, Münster , Germany
[27]School of Chemical & Physical Sciences, and Centre for Biodiscovery, Victoria University of Wellington, Wellington, New Zealand
[28]Gillings School of Global Public Health, Department of Epidemiology, UNC Chapel Hill, Chapel Hill, United States
[29]Department of Chemistry, Indiana University, Bloomington, United States
[30]Smithsonian Tropical Research Institute, Ancón, Panama
[31]Skaggs School of Pharmacy and Pharmaceutical Sciences, UC San Diego, La Jolla, United States
[32]School of Pharmaceutical Sciences of Ribeirao Preto, University of São Paulo, São Paulo, Brazil
[33]Centro de Ciencias Biologicas e da Saude, Universidade Fderal de Mato Grosso do Sul, Campo Grande, Brazil
[34]UMR CNRS 6553 ECOBIO, University of Rennes 1, Rennes Cedex, France
[35]Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, United States
[36]PBSci-Chemistry & Biochemistry Department, UC Santa Cruz, Santa Cruz, United States
[37]Department of Bioengineering, UC San Diego, La Jolla, United States
[38]Department of Pharmaceutical Sciences, College of Pharmacy, Oregon State University, Corvallis, United States
[39]Department of Plant and Microbial Biology, UC Berkeley, Berkeley, United States
[40]Department of Biological Sciences, Florida International University, Miami, United States
[41]Department of Pharmaceutical Biotechnology, Helmholtz Institute for Pharmaceutical Research Saarland, Saarbrücken, Germany

[42]Center for Oceans and Human Health, Scripps Institute of Oceanography, UC San Diego, La Jolla, United States
[43]Department of Chemistry, University of Hawaii at Manoa, Honolulu, United States
[44]Division of Biological Sciences, UC San Diego, La Jolla, United States
[45]Department of Nanoengineering, UC San Diego, La Jolla, United States
[46]School of Pharmaceutical Sciences, University of Geneva, Geneva, Switzerland
[47]Structural and Computational Biology, European Molecular Biology Laboratory, Heidelberg, Germany
[48]Institut de Chimie des Substances Naturelles, CNRS-ICSN, UPR 2301, Labex CEBA, University of Paris-Saclay, Gif-sur-Yvette, France
[49]Biological Sciences, Pacific Northwest National Laboratory, Richland, United States
[50]National Center for Advancing Translational Sciences, National Institute of Health, Rockville, United States
[51]Institute of Microbiology, Chinese Academy of Sciences, Beijing, China
[52]Department of Pediatrics, UC San Diego, La Jolla, United States

These authors contributed equally to this work
Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, and Yao Peng

Affiliations

Computer Science and Engineering, UC San Diego, La Jolla, United States
Mingxun Wang, Jeremy J Carver, Pavel Pevzner

Center for Computational Mass Spectrometry, UC San Diego, La Jolla, United States
Mingxun Wang, Jeremy J Carver, Pavel Pevzner, Hosein Mohimani, Nuno Bandeira

Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, UC San Diego, La Jolla, United States
Vanessa V Phelan, Laura M Sanchez, Neha Garg, Jeramie Watrous, Tal Luzzatto-Knaan, Carla Porto, Amina Bouslimani, Alexey V Melnik, Michael J Meehan, Laura A Pace, David J Gonzalez, Nobuhiro Koyama, Kathleen Dorrestein, Brendan M Duggan, Jehad Almaliti, William H Gerwick, Bradley S Moore, Pieter C Dorrestein, Nuno Bandeira

Department of Chemistry and Biochemistry, UC San Diego, La Jolla, United States
Yao Peng, Don Duy Nguyen, Clifford A Kapono, Cheng-Chih Hsu, Dimitrios J Floros, Yi Zeng

Department of Microbiology and Immunology, Stanford University, Palo Alto, United States
Wei-Ting Liu

Center for Marine Biotechnology and Biomedicine, Scripps Institute of Oceanography, UC San Diego, La Jolla, United States
Max Crüsemann , Paul D Boudreau, Katherine R Duncan, Karin Kleigrewe, Lena Gerwick, Charles B Larson, Ellis C O'Neill, Enora Briand, Evgenia Glukhov, Jenan J Kharbush, Samantha J Mascuch, Paul R Jensen, William H Gerwick, Bradley S Moore, Pieter C Dorrestein

Sirenas Marine Discovery, San Diego, United States
Eduardo Esquenazi, Egle Pociute, Hailey Houson, Lisa Vuong, Venkat Macherla

Centro de Ciencias Genómicas, Universidad Nacional Autonoma de Mexico, Cuernavaca, Mexico
Mario Sandoval-Calderón, Christian Sohlenkamp

Salk Institute, Salk Institute, La Jolla, United States
Roland D Kersten

Biology Department, San Diego State University, San Diego, United States
Robert A Quinn

Scottish Association for Marine Science, Scottish Marine Institute, Oban, United Kingdom
Katherine R Duncan

Center for Drug Discovery and Biodiversity, INDICASAT, City of Knowledge, Panama
Ronnie G Gavilan, Brian E Sedio, Christopher A Boya P., Daniel Torres-Mendoza, Marcelino Gutiérrez

Genome Dynamics, Lawrence Berkeley National Laboratory, Berkeley, United States
Trent Northen, Stefan Jenkins

FAS Center for Systems Biology, Harvard, Cambridge, United States
Rachel J Dutton

Produits naturels – Synthèses – Chimie Médicinale, University of Rennes 1, Rennes Cedex, France
Delphine Parrot, Sophie Tomasi

Chemistry, University of Minnesota, Minneapolis, United States
Erin E Carlson

Dynamique des Génomes et Adaptation Microbienne, University of Lorraine, Vandœuvre-lès-Nancy, France
Bertrand Aigle

Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark
Charlotte F Michelsen, Lars Jelsbak, Maria Maansson, Andreas Klitgaard, Kristian Fog Nielsen

Microbial and Environmental Genomics, J. Craig Venter Institute, La Jolla, United States
Anna Edlund

School of Dentistry, UC Los Angeles, Los Angeles, United States
Anna Edlund, Jeffrey McLean, Wenyuan Shi

Department of Periodontics, University of Washington, Seattle, United States
Jeffrey McLean

Institute of Microbiology, ETH Zurich, Zurich, Switzerland
Jörn Piel, Eric J. N. Helfrich, Florian Ryffel, Julia A Vorholt

Department of Medicinal Chemistry and Pharmacognosy, University of Illinois Chicago, Chicago, United States
Brian T Murphy, Maryam Elfeki

Department of Marine Biotechnology and Resources, National Sun Yat-sen University, Kaohsiung, Taiwan
Chih-Chuang Liaw

Agricultural Biotechnology Research Center, Academia Sinica, Taipei, Taiwan
Yu-Liang Yang

Institute of Food Chemistry, University of Münster, Münster , Germany
Hans-Ulrich Humpf

School of Chemical & Physical Sciences, and Centre for Biodiscovery, Victoria University of Wellington, Wellington, New Zealand
Robert A Keyzers

Gillings School of Global Public Health, Department of Epidemiology, UNC Chapel Hill, Chapel Hill, United States
Amy C Sims, Ralph Baric

Department of Chemistry, Indiana University, Bloomington, United States
Andrew R. Johnson, Ashley M Sidebottom

Smithsonian Tropical Research Institute, Ancón, Panama
Brian E Sedio

Skaggs School of Pharmacy and Pharmaceutical Sciences, UC San Diego, La Jolla, United States
Charles B Larson, David J Gonzalez, Pieter C Dorrestein, Nuno Bandeira

School of Pharmaceutical Sciences of Ribeirao Preto, University of São Paulo, São Paulo, Brazil
Denise B Silva, Lucas M Marques, Daniel P Demarque, Ricardo R Silva, Andrés M. C. Rodríguez, Norberto P Lopes

Centro de Ciencias Biologicas e da Saude, Universidade Fderal de Mato Grosso do Sul, Campo Grande, Brazil
Denise B Silva

UMR CNRS 6553 ECOBIO, University of Rennes 1, Rennes Cedex, France
Enora Briand

Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, United States
Eve A Granatosky

PBSci-Chemistry & Biochemistry Department, UC Santa Cruz, Santa Cruz, United States

Kenji L Kurita, Roger G Linington

Department of Bioengineering, UC San Diego, La Jolla, United States
Pep Charusanti, Bernhard O Palsson

Department of Pharmaceutical Sciences, College of Pharmacy, Oregon State University, Corvallis, United States
Kerry L McPhail, Oliver B Vining

Department of Plant and Microbial Biology, UC Berkeley, Berkeley, United States
Matthew F Traxler

Department of Biological Sciences, Florida International University, Miami, United States
Niclas Engene

Department of Pharmaceutical Biotechnology, Helmholtz Institute for Pharmaceutical Research Saarland, Saarbrücken, Germany
Thomas Hoffman, Rolf Müller

Center for Oceans and Human Health, Scripps Institute of Oceanography, UC San Diego, La Jolla, United States
Vinayak Agarwal, Bradley S Moore

Department of Chemistry, University of Hawaii at Manoa, Honolulu, United States
Philip G Williams, Jingqui Dai, Ram Neupane, Joshua Gurr

Division of Biological Sciences, UC San Diego, La Jolla, United States
Anne Lamsa, Kit Pogliano

Department of Nanoengineering, UC San Diego, La Jolla, United States
Chen Zhang

School of Pharmaceutical Sciences, University of Geneva, Geneva, Switzerland
Pierre-Marie Allard, Jean-Luc Wolfender

Structural and Computational Biology, European Molecular Biology Laboratory, Heidelberg, Germany
Prasad Phapale, Theodore Alexandrov

Institut de Chimie des Substances Naturelles, CNRS-ICSN, UPR 2301, Labex CEBA, University of Paris-Saclay, Gif-sur-Yvette, France
Louis-Felix Nothias, Marc Litaudon

Biological Sciences, Pacific Northwest National Laboratory, Richland, United States
Jennifer E Kyle, Thomas O Metz, Katrina M Waters

National Center for Advancing Translational Sciences, National Institute of Health, Rockville, United States
Tyler Peryea, Dac-Trung Nguyen, Danielle VanLeer, Paul Shinn, Ajit Jadhav

Institute of Microbiology, Chinese Academy of Sciences, Beijing, China
Xueting Liu, Lixin Zhang

Department of Pediatrics, UC San Diego, La Jolla, United States
Rob Knight

Correspondence to
Pieter Dorrestein (pdorrestein@ucsd.edu) or Nuno Bandeira (bandeira@ucsd.edu).

Abstract

Realizing the potential of the diverse chemistries of natural products in biotechnology and medicine has been limited by manual analysis of experimental data through mining mass spectrometry knowledge solely captured in literature. While mass spectrometry techniques have proven well-suited for high-throughput analyses of natural products, there is no infrastructure for researchers to systematically share knowledge or analyze data. We present Global Natural Products Social molecular networking (GNPS, http://gnps.ucsd.edu), an open-access knowledge base for sharing, analysis, and community curation of raw, processed, and identified tandem mass (MS/MS) spectrometry data. GNPS further organizes, curates, and freely redistributes community-wide reference MS libraries, as well as provides a data-driven social networking infrastructure. Finally, GNPS introduces the concept of living data through crowdsourced curation of reference libraries and continuous reanalysis of public data.

Introduction

Natural products (NPs) from marine and terrestrial environments, including their inhabiting microorganisms, plants, animals, and humans, are routinely analyzed using mass spectrometry. However a single mass spectrometry experiment can collect thousands of MS/MS spectra in minutes[1] and individual projects can acquire millions of spectra. These datasets are too large for manual analysis. Further, comprehensive software and proper computational infrastructure are not readily available and only low-throughput sharing of either raw or annotated spectra is feasible, even among members of the same lab. The potentially useful information in MS/MS datasets can thus remain buried in papers, laboratory notebooks, and private databases, hindering retrieval, mining, and sharing of data and knowledge. Although there are several NP databases — Dictionary of Natural Products[2], AntiBase[3] and MarinLit[4] — that assist in dereplication (identification of known compounds), these resources are not freely available and do not process mass spectrometry data. Conversely, mass spectrometry databases including Massbank[5], Metlin[6], mzCloud[7], and ReSpect[8] host MS/MS spectra but limit data analyses to several individual spectra or a few LC-MS files. While Metlin and mzCloud provide a spectrum search function, unfortunately, their libraries are not freely available.

Global genomics and proteomics research has been facilitated by the development of integral resources such as the National Center for Biotechnology Information (NCBI) and UniProt KnowledgeBase (UniProtKB), which provide robust platforms for data sharing and knowledge dissemination[9,10]. Recognizing the need for an analogous community platform to effectively share and analyze natural products MS data, we present the Global Natural Products Social Molecular Networking (GNPS, available at gnps.ucsd.edu). GNPS is a data-driven platform for the storage, analysis, and knowledge dissemination of MS/MS spectra that enables community sharing of raw spectra, continuous annotation of deposited data, and collaborative curation of reference spectra (referred to as spectral libraries) and experimental data (organized as datasets).

GNPS provides the ability to analyze a dataset and to compare it to all publically available data. By building on the large scale computational infrastructure of the University of California San Diego (UCSD) Center for Computational Mass Spectrometry (CCMS), GNPS provides public dataset deposition/retrieval through the *Mass* Spectrometry *I*nteractive *V*irtual *E*nvironment (MassIVE) data repository. The GNPS analysis infrastructure further enables online dereplication[6,11–13], automated molecular networking analysis[14–21], and crowdsourced MS/MS spectrum curation. Each dataset added to the GNPS repository is automatically reanalyzed in the next monthly cycle of continuous identification (see **Living Data by Continuous Analysis** below). Each of these tens of millions of spectra in GNPS datasets is matched to reference spectral libraries to annotate molecules and discover putative analogs (**Fig. 1a)**. From January 2014 to November 2015, GNPS has grown to serve 9,267 users from 100 countries **(Fig. 1b)**, with 42,486 analysis sessions that have processed more than 93 million spectra as molecular networks from a quarter million LC-MS runs. Searches against a combined catalog of over 221,000 MS/MS reference library spectra from 18,163 compounds (**Supplementary Table 1**) are possible, and GNPS has matched almost one hundred million MS/MS spectra in all public and private search jobs using an estimated 84,000 compute hours.

GNPS Spectral Libraries

GNPS spectral libraries enable dereplication, variable dereplication (approximate matches to spectra of related molecules), and identification of spectra in molecular networks. GNPS has collected available MS/MS spectral libraries relevant to NPs (which also include other metabolites and molecules), including MassBank[5], ReSpect[8] and NIST[22] (**Table 1, Fig. 2a, and Supplementary Table 1**). Altogether, these third party libraries total 212,230 MS/MS spectra representing 12,694 unique compounds (**Fig. 2b**). While this combined collection of reference spectra, provides a starting point for dereplication, only 1.01% of all spectra public GNPS datasets has been matched to this collection, indicating insufficient chemical space coverage.

Although the NP community is working to populate this “missing” chemical space, there is no way to report discoveries of new chemistries in an easily verifiable and reusable format. To begin addressing this pressing need, GNPS offers both newly-acquired reference spectra (GNPS-

Collections) as well as a crowdsourced library of community-contributed reference spectra (GNPS-Community). GNPS-Collections includes NPs and pharmacologically active compounds totaling 6,629 MS/MS spectra of 4,243 compounds (**Fig 2b, Supplementary Table 1, Supplementary Note 1,2, and Supplementary Table 2**). The GNPS-Community library has grown to include 2,224 MS/MS spectra of 1,325 compounds from 55 worldwide contributors. While the total number of MS/MS spectra in GNPS libraries is only 4% of the MS/MS spectra in third party libraries, GNPS libraries contribute matches of MS/MS spectra at a scale disproportionate to their size (**Fig. 2c**). The GNPS libraries account for 29% of the unique compound matches and 59% of the MS/MS matches in public (88% of public+private) data. This indicates that the GNPS libraries contain compounds that are complementary to the chemical space represented in other libraries (**Fig. 2c,d**). Moreover, in difference from third party libraries, spectra submitted to GNPS-Community libraries become immediately searchable by the whole community, so such submissions seamlessly transfer knowledge between laboratories (**Fig. 1a**) – a process that is akin to the addition of genome annotations contributed to GenBank[9].

In order to create a robust library, it is important for submissions to be peer-reviewed and, if necessary, annotations corrected or updated as appropriate. Reference spectra submitted to the GNPS-Community library are categorized by the estimated reliability of the proposed submissions. Gold reference spectra must be derived from structurally characterized synthetic or purified compounds and can only be submitted by approved users. Approval is given to contributors who have undergone training. Training is initiated by contacting the corresponding authors or CCMS administrators. Silver reference spectra need to be supported by an associated publication, while Bronze reference spectra are all remaining putative annotations (**Supplementary Table 3**). This type of division of spectra is reminiscent of RefSeq/TPA/GenBank[9,23] (genomics) and Swiss-Prot/TrEMBL/UniProt[24,25] (proteomics), allowing for varying tradeoffs between comprehensiveness and reliability of annotations defined as Gold, Silver, and Bronze (**Fig 2e**).

To enable refinements or corrections of annotations, GNPS allows for community-driven, iterative re-annotation of reference MS/MS spectra in a wiki-like fashion, to progressively improve the library and converge towards consensus annotation of all MS/MS spectra of interest. This is a process similar to the iterative annotation of the human genome (e.g., see series of papers on NCBI GenBank[9]). To date, 563 annotation revisions have been made (**Supplementary Table 4**), most of which added metadata to library spectra or refined compound names. The history of each annotation is retained so that users can discuss the proper annotation and address disagreements via comment threads.

Dereplication using GNPS

High throughput dereplication of NP MS/MS data is implemented in GNPS by querying newly acquired MS/MS spectra against all the accumulated reference spectra in GNPS spectral libraries (**Fig. 3a**). To date, more than 93 million MS/MS spectra from various instruments (including Orbitrap, Ion Trap, qTof, and FT-ICR) have been searched at GNPS, yielding putative dereplication

matches of 7.7 million spectra to 15,477 compounds. In the second stage of dereplication, GNPS goes beyond re-identification by utilizing variable dereplication - a modification-tolerant spectral library search that is mediated by a spectral alignment algorithm. Variable dereplication enables the detection of significant matches to either putative analogs of known compounds (e.g., differing by one modification or substitution of a chemical group) or compounds belonging to the same general class of molecules (**Fig. 3b**).Variable dereplication is not available through any other computational platform. For example, GNPS variable dereplication has detected compounds with different levels of glycosylation on various substrates. As MS/MS fragmentation preferentially results in peaks from glycan fragments, it is possible to detect sets of compounds with related glycans even when the substrates to which the glycans are attached are themselves unrelated[26]. To date, 3,891 putative analogs have been identified in public data using GNPS variable dereplication (**Supplementary Table 5**). These 3,891 putative analogs include several unique molecules that could be user-curated and added to GNPS reference libraries (see **Molecular Explorer** below on accessing and annotating putative analogs).

To assess the reliability of the MS/MS matches found by GNPS dereplication, GNPS users can rate the quality of matches returned by automated GNPS reanalysis (see below). These ratings are 4 star (correct), 3 star (likely correct, e.g. could also be isomers with similar fragmentation patterns), 2 star (unable to confirm the annotation due to limited information) and 1 star (incorrect) (**Supplementary Table 6**). So far, of the 3,608 matches that have been rated, 139 (3.9%) matches were given 1 or 2 stars (insufficient information (2.9%) or incorrect (1%)) by user ratings. These percentages are consistent with the false discovery rates estimated using spectral library searches of benchmark LC-MS datasets with compound standards (**Supplementary Note 3 and Supplementary Fig. 1,2**). Furthermore, these 3,608 match ratings were associated with 2,041 library spectra, therefore the average rating of a library spectrum can offer insight into the reliability of its reference annotation, not unlike Yelp ratings for restaurants. Incorrect matches can arise through either spurious high-scoring matches to library spectra or incorrect annotations for library spectra. Of the 2,041 library spectra with match ratings, 72 (3.5%) spectra had average ratings below 2.5 stars. These percentage ratings were further broken down by spectral library (**Fig. 2e**). We found that for GNPS-Collection and GNPS-Community libraries, only 29 out of 1746 (1.7%) of the rated library spectra had average ratings below 2.5 stars. These ratings demonstrate that the perceived reliability of GNPS spectral libraries compares favorably with established community resources such as NIST and Massbank, with 10.5% and 20.1% of the ratings were below 2.5 stars respectively, and reinforces confidence that the community curation process is, and will continue to be, a success. Thus, the key advantages of searching using GNPS are that one can run simple or variable dereplication against all publicly accessible reference spectra, where community-rated matches can be used to improve the quality of the reference libraries and matching algorithms. None of these dereplication capabilities are possible with existing published resources.

Molecular Networking

Molecular networks are visual displays of the chemical space present in mass spectrometry experiments. GNPS can be used for molecular networking[14–21,27,28], a spectral correlation and visualization approach that can detect sets of spectra from related molecules (so-called spectral networks[29]) even when the spectra themselves are not matched to any known compounds **(Fig. 3a)**. Spectral alignment[15,27] detects similar spectra from structurally related molecules, assuming these molecules fragment in similar ways reflected in their MS/MS patterns (**Fig. 3b**), analogous to the detection of related protein or nucleotide sequences by sequence alignment. GNPS is currently the only public infrastructure that enables molecular networking. The visualization of molecular networks in GNPS represents each spectrum as a node and spectrum-to-spectrum alignments as edges (connections), between nodes. Nodes can be supplemented with metadata including dereplication matches or information that is provided by the user, such as abundance, origin of product, biochemical activity, hydrophobicity, etc., which can be reflected in a node's size or color. It is possible to visualize the map of related molecules as a molecular network[21,30–33] (**Supplementary Fig. 3**) both online at GNPS (**Fig. 3c**) or exported for analysis in Cytoscape[31]. Molecular networking analyses of 272 public datasets (**Fig. 4a**) from a diverse range of samples reveals that on average 35.2% of all unidentified nodes are significantly matched to other spectra of related molecules within a cosine score of 0.8 (increasing to 44.7% of all nodes in more exploratory networks with a cosine score of 0.65). This indicates that a large fraction of all unidentified spectra could be identifiable if their or their neighboring nodes' reference spectra were available in the reference spectral libraries.

Living Data by Continuous Analysis

Funding agencies and publishers have called for raw scientific data, including mass spectrometry data, and analysis methods to be made publically available where possible. Consistent with this aim, GNPS datasets usually comprise the full set of mass spectrometry files produced during a NP research project or the full set of spectra analyzed for a peer-reviewed publication (**Supplementary Note 4**). While it is potentially advantageous to the community for all data to be made public, GNPS user data can remain private until users explicitly choose to make it public (private data is also analyzable and privately sharable, with >93 million spectra in >250,000 private LC/MS runs already searched using GNPS). GNPS has the largest collection of publicly accessible natural product and metabolomics MS/MS datasets and is the only infrastructure where public data sets can be reanalyzed together and compared to each other(**Table 1**). To date, GNPS has made 272 public GNPS datasets openly available which are comprised of more than 30,000 mass spectrometry runs with approximately 84 million MS/MS spectra. In common with other public repositories[34,35], GNPS datasets can be downloaded. However, data availability on its own does not serve to enable data reuse. GNPS is unique among MS repositories by enabling continuous identification: the periodic and automated re-analysis of all public datasets (**Supplementary Note 5,6 and Supplementary Table 7,8**). This continuous re-analysis, which incorporates molecular networking and dereplication tools, implements a 'virtuous cycle' as illustrated in **Figure 1a**. Because GNPS spectral libraries are constantly growing due to community contributions and continued generation of reference spectra, the number of matches made by successive re-

analyses of public datasets has already grown and is expected to continue to grow over time (**Fig. 4b**). GNPS users are periodically updated with alerts of new search results.

For example, a *Streptomyces roseosporeus* project (MSV000078577) was deposited April 8, 2014. At first, only 7 MS/MS spectra were matched. However as of July 14, 2015 36 spectral matches have been made to GNPS libraries. Overall, the total number of compounds matched to GNPS datasets increased more than tenfold, while the number of matched MS/MS spectra in GNPS datasets increased more than twenty-fold in 2015 (**Fig. 4b**). GNPS users can also subscribe to specific datasets of interest, rather like 'following' people on Twitter. When new matches are made, changed, or revoked, all subscribers are notified of new information by an email summarizing changes in identification. From April 2014 to July 2015, 45 updates were initiated by CCMS and automatically sent to subscribers (**Supplementary Fig. 4**). Update emails have led to substantially more views per dataset, compared to non-GNPS datasets (192 proteomics datasets) deposited in MassIVE. Continuous identification not only keeps a single dataset 'alive', it can create connections between datasets and users over time. Similarities between datasets could form the basis of a data-mediated social network of users with potentially related research interests despite seemingly disparate research fields, rather like the "People You May Know" feature on LinkedIn. On average each GNPS user already has 5 suggested collaborators (**Supplementary Fig. 5**).

Molecular Explorer

Molecular Explorer is a new feature that can only be implemented on 'living data' repositories and thus exists only in GNPS. Molecular Explorer allows users to find all datasets and putative analogs that have ever been observed for a given molecule of interest. We anticipate this can guide the discovery of previously unknown analogs of existing antibiotics. Public NP data contains more than one hundred unidentified putative analogs of antibiotics such as valinomycin, actinomycin, etamycin, hormaomycin, stendomycin, daptomycin, erythromycin, napsamycin, clindamycin, arylomycin, and rifamycin, highlighting a clear potential to generate leads to discover structurally related antibiotics though the application of GNPS (**Supplementary Fig. 6,Supplementary Table 5, and Supplementary Note 7**).

To demonstrate this principle we searched for an analog of stenothricin, a broad spectrum antibiotic produced by *S. roseosporus* with a unique biological response profile[36,37] (**Supplementary Fig. 7**). MS/MS data from *S. roseosporus* and *Streptomyces* sp. DSM5940 extracts (MSV000079204) were analyzed by molecular networking and dereplication in GNPS (**Supplementary Note 8 and Supplementary Fig. 8**). Nodes corresponding to the stenothricin[37] from *S. roseosporus* were identified in the molecular network. In addition, a small sub-network corresponding to spectra from *Streptomyces* sp. DSM5940 (**Fig. 5a**) included 14 nodes that were 41 Da smaller than nodes already known to be stenothricin analogs. This sub-network seemed to indicate that *Streptomyces* sp. DSM5940 produces a set of 5 abundant analogs of stenothricin which we named stenothricin-GNPS 1-5 (**Supplementary Table 9**). To our knowledge, a chemical entity that is related to stenothricin with a mass shift of -41 Da has not been described in any

database or in the literature. The most abundant analog, stenothricin-GNPS 2 (*m/z* 1105) was purified and the MS/MS spectra manually compared to MS/MS spectra produced from stenothricin D. This confirmed structural similarity (**Fig. 5b,c Supplementary Fig. 9**). Differential 2D NMR (**Supplementary Fig. 10-14, Supplementary Table 10, and Supplementary Note 9**), Marfey's analysis[38] (**Supplementary Fig. 15**), and genome mining (**Supplementary Fig. 16,17, Supplementary Table 11, and Supplementary Note 10**) all support that the -41 Da mass shift is due to a lysine to serine substitution.

The structural comparison between stenothricin D and stenothricin-GNPS has identified a potential role for the lysine residue of stenothricin D in biological function. Stenothricin-GNPS was subjected to fluorescence microscopy based bacterial cytological profiling[39,40] (**Fig. 5d**). Unlike stenothricin D, stenothricin-GNPS is only active against *Escherichia coli lptD* cells, which are defective in the essential outer membrane protein LptD (**Supplementary Fig. 18 and Supplementary Note 11**). Although both stenothricin D and stenothricin-GNPS increased membrane permeability of bacterial cells within two hours, stenothricin-GNPS did not have the membrane solubilization function of stenothricin D (**Fig. 5d**), indicating that the activity of stenothricin D is altered by the presence of a lysine residue that is absent from stenothricin-GNPS**.** Several published applications of molecular networking and MS/MS based dereplication using GNPS have been reported while the infrastructure has been under development. Specifically, GNPS has enabled the discovery of natural products including colibactin[41–45], characterization of biosynthetic pathways[46,47], understanding of the chemistry of ecological interactions[28,48–52], and development of metabolomics bioinformatics methods[53]. The application of GNPS workflows to such diverse research areas demonstrates the utility of GNPS to broad interdisciplinary science.

Conclusion

GNPS aims to expand our understanding of nature's chemical diversity by supporting community-wide identification of compounds that have important roles in ecology, medicine, and biotechnology. To this end, GNPS delivers a community-centric knowledge space in which NP data is shared, analyzed and annotated by researchers, groups of scientists, and laboratories worldwide. The synergy implemented by GNPS creates a cycle of annotation, drawing users back to curate community data, and a cycle of knowledge, by providing reference spectral libraries, public datasets, and continuous dereplication. GNPS thus provides the NP community with an open, free, and community-curated analysis platform for iterative and collaborative annotation of NP mass spectrometry data.

The living data enabled by the GNPS platform will mediate connections between researchers and has the potential to transform data networks into social networks. Of 1,272 compound identifications obtained by continuous identification with the GNPS-Community library, 1,063 (83.6%) were made using reference spectra that were not uploaded by the submitter - in other words, the vast majority of identifications were enabled by other community members. This reuse of knowledge and data is inline with other community-wide curation efforts including Wikipedia and

crowd-sourced dictionaries. Since their initial deposition, 59% of datasets have an increased number of identifications, with the average dataset more than doubling the number of identifications since submission (**Supplementary Fig. 19**). GNPS enables facile sharing of individual analyses (**Supplementary Fig. 20**) and uses molecular networks to reveal connections between datasets from different laboratories and biological sources that would otherwise remain disconnected. To date, 3,145 analysis jobs have included files shared between GNPS users, encompassing 548 unique pairs of individuals' collaborations. GNPS recasts public datasets as "conversation starters" in a data-mediated social network. Continuous identification means that GNPS transforms data networks into social networks and continuous updates draws users back to GNPS for re-analysis, bringing data to life.

While we have described only one simple application of GNPS to identify an analog of stenothricin, the community has already begun to utilize GNPS to expedite natural product analysis[28,41,43,45,46,50,52]. Further it is expected that the user base of GNPS will expand to the many fields that utilize MS/MS data, including the study of the metabolome, exposome, the chemistry of the human habitat, drug discovery, microbiome, immunology, food industry, agricultural industry, stratification of patients in clinical trials, clinical adsorption/metabolism, and ocean science to name a few, resulting in different GNPS workflows[42,44,47,51,53]. As previously shown in genomics[9] and protein structure analysis[54] , the models of global collaboration and social cooperation which are present in GNPS could empower scientific communities to collectively translate big data into shared, reusable knowledge and profoundly influence the way we explore molecules using mass spectrometry.

Acknowledgements: This work was partially supported by National Institution of Health (NIH) Grants 5P41GM103484-07, GM094802, AI095125, GM097509, S10RR029121, UL1RR031980, GM085770, U01TW0007401, U01AI12316-01; NB was also partially supported as an Alfred P Sloan Fellow. In addition, this work was supported by the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health, Department of Health and Human Services, under Contract Number HHSN272200800060C. VVP is supported by the NIH Grant K01 GM103809. LMS is supported by National Institutes of Health IRACDA K12 GM068524 award. TLK is supported by the United States - Israel Binational Agricultural Research and Development Fund Vaadia-BARD No. FI-494-13. CP is supported by Science without Borders Program from CNPq. AMCR is supported by São Paulo Research Foundation (FAPESP) grant#2014/01651-8, 2012/18031-7. KK was supported by a fellowship within the Postdoc-Programme of the German Academic Exchange Service (DAAD). MC was supported by a Deutsche Forschungsgemeinschaft (DFG) postdoctoral fellowship. EB is supported by a Marie Curie IOF Fellowship within the 7th European Community Framework Program (FP7-PEOPLE-2011-IOF, grant number 301244-CYANOMIC). CCL was supported by a grant from Ministry of Science and Technology of Taiwan (MOST103-2628-B-110-001-MY3). PC and BOP were supported by the Novo Nordisk Foundation. Lixin Zhang and Xueting Liu are supported by National Program on Key Basic Research Project (2013BC734000) and the National Natural Science Foundation of China (81102369 and

31125002). DP is supported by INSA grant, Rennes. RRS is supported by FAPESP grant#2014/01884-2. DPD is supported by FAPESP grant#2014/18052-0. LMMM is supported by FAPESP grant#2013/16496-5. DBS is supported by FAPESP grant#2012/18031-7. NPL is supported by FAPESP(2014/50265-3), CAPES/PNPD, CNPq-PQ 480 306385/2011-2 and CNPq-INCT_if. EAG is supported by the Notre Dame Chemistry-Biochemistry-Biology Interface (CBBI) program and NIH T32 GM075762. WS and JSM are supported by grants from the National Institutes of Health 1R01DE023810-01 and 1R01GM095373. AE is supported by grant from National Institute of Health K99DE024543. CFM and LJ are supported by the Villum Foundation VKR023113, the Augustinus Foundation 13-4656 and the Aase & Ejnar Danielsens Foundation 10-001120. MSC was supported by UC MEXUS-CONACYT Collaborative Grant CN-12-552. MFT was supported by NIH grant 1F32GM089044. Contributions by BES were supported by NSF grant DEB 1010816 and a Smithsonian Institution Grand Challenges Award. EJNH and JP are supported by the DFG (Forschergruppe 854) and by SNF grant IZLSZ3_149025. KFN and AK are supported by the Danish Council for Independent Research, Technology, and Production Sciences (09-064967) and the Agilent Thought Leader Program. ACS and RSB were supported by NIH/NIAID U19-AI106772. BTM and ME were supported under Department of Defense grant #W81XWH-13-1-0171. Contributions by OBV and KLM were supported by Oregon Sea Grant NA10OAR4170059/R/BT-48, and NIH 5R21AI085540 and U01TW006634-06. EEC, ASM and ARJ were supported by an NSF CAREER Award, a Pew Biomedical Scholar Award (EEC), a Sloan Research Fellow Award (EEC), the Research Corporation for Science Advancement (Cottrell Scholar Award; EEC) and an Indiana University Quantitative Chemical Biology trainee fellowship (ARJ). MM was supported by the Danish Research Council for Technology and Production Science with Sapere Aude (116262). PMA was supported by FNS for fellowship on Subside (200020_146200).

We thank Valerie Paul, Rich Taylor, Lihini Aluwihare, Forest Rohwer, Benjamin Pullman, Jinshu Fang, Martin Overgaard, Michael Katze, Richard D. Smith, Sarkis K Mazmanian, William Fenical, Eduardo Macagno, Xuesong He, and Cajetan Neubauer for feedback and support for their lab personnel to contribute to the work. We thank Bertold Gust and co-workers at the University of Tuebingen for assisting us to obtain *Streptomyces* sp. DSM5940.

Author contributions:

Design and oversight of the project: PCD and NB

Algorithms: MW and NB

Web-site: MW, JC

In-house library acquisition and analysis: VVP, LMS, NG

User curated library acquisition and analysis: ACS, AE, JSM, WS, WTL, MJM, VVP, LLM, NG, RAQ, AB, CP, TLK, AMCR, AM, MC, KRD, KK, ECO, BSM, EB, EG, DDN, SJM, PDB, XL, LZ, HUH, CFM, LJ, DP, ST, EAG, MSC, CS, KLK, PMA, RGL, RSB, PRJ, MFT, SJ, BES, LMMM, DPD, DBS, NPL, JP, EJNH, AK, RAK, JEK, TOM, PGW, JD, RN, JG, BA, OBV, KLM, EEC, ASM, ARJ, RDK, JJK, KMW, CCH, MM, CCL, YLY

Sample preparation, data generation and web-site beta testing: AE, WTL, MJM, VVP, LMS, NG, RAQ, AB, CP, TLK, AMCR, AM, DF, MC, JC, NB, PCD, ECO, EB, EG, DDN, SJM, PDB, XL, LZ, CZ, CFM, RRS, EAG, MSC, CS, DP, ST, PMA, RGL, BES, LMMM, JP, EJNH, DTM, CABP, ME, BTM, OBV, KLM, EEC, ASM, ARJ, KRD
GNPS Documentation: MW, VVP, LMS, CK, DDN, RRS, LAP
Genome sequencing, assembly and targeted amplification: YP, PC, RG, MG, BOP, LG
Stenothricin GNPS data analysis: WTL, VVP, LMS, YP, PCD
NMR acquisition and analysis: BMD, PDB, LMS
Marfey's analysis: YP, PDB
Microbiology: YP, ACS, RSB
Peptidogenomics analysis: YP, RDK, PCD
Fluorescence Microscopy: YP, AL, KP
Writing of the paper: MW, VVP, LMS, NG, RK, PCD, and NB

Competing Financial Interests

NB has an equity interest in Digital Proteomics, LLC, a company that may potentially benefit from the research results; Digital Proteomics LLC was not involved in any aspects of this research. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict-of-interest policies.
EE, EP, HH, LV, and VM are employees of Sirenas MD
PCD is on the advisory board for Sirenas MD
TA is the Scientific Director of SCiLS GmbH

References

1. Bouslimani, A., Sanchez, L. M., Garg, N. & Dorrestein, P. C. Mass spectrometry of natural products: current, emerging and future technologies. *Nat. Prod. Rep.* **31,** 718–29 (2014).
2. In. *Dict. Nat. Prod.* (2013).
3. Laatsch, H. AntiBase A data base for rapid structural determination of microbial natural products, and annual updates. (2008).
4. Blunt, J. & Munro, M. MarinLit. *A database Lit. Mar. Nat. Prod. use a macintosh Comput. Prep. Maint. by Mar. Chem. Gr. (Department Chem. Univ. Canterbury Canterbury, New Zealand)* (2003).
5. H, H. *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. (2010).
6. Smith, C. A. *et al.* METLIN: a metabolite mass spectral database. *Ther. Drug Monit.* **27,** 747–751 (2005).
7. mzCloud. mzCloud. at <https://www.mzcloud.org/>

8. Y, S. *et al.* RIKEN tandem mass spectral database (ReSpect) for phytochemicals: a plant-specific MS/MS-based data resource and database. (2012).

9. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41,** (2013).

10. Magrane, M. & Consortium, U. P. UniProt Knowledgebase: A hub of integrated protein data. *Database* **2011,** (2011).

11. Lang, G. *et al.* Evolving trends in the dereplication of natural product extracts: New methodology for rapid, small-scale investigation of natural product extracts. *J. Nat. Prod.* **71,** 1595–1599 (2008).

12. Ito, T. & Masubuchi, M. Dereplication of microbial extracts and related analytical technologies. *J. Antibiot. (Tokyo).* **67,** 353–60 (2014).

13. Little, J. L., Williams, A. J., Pshenichnov, A. & Tkachenko, V. Identification of 'known unknowns' utilizing accurate mass data and chemspider. *J. Am. Soc. Mass Spectrom.* **23,** 179–185 (2012).

14. Moree, W. J. *et al.* Interkingdom metabolic transformations captured by microbial imaging mass spectrometry. *Proceedings of the National Academy of Sciences* **109,** 13811–13816 (2012).

15. Watrous, J. *et al.* From the Cover: PNAS Plus: Mass spectral molecular networking of living microbial colonies. *Proceedings of the National Academy of Sciences* **109,** E1743–E1752 (2012).

16. Nguyen, D. D. *et al.* MS/MS networking guided analysis of molecule and gene cluster families. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E2611–20 (2013).

17. Sidebottom, A. M., Johnson, A. R., Karty, J. A., Trader, D. J. & Carlson, E. E. Integrated metabolomics approach facilitates discovery of an unpredicted natural product suite from Streptomyces coelicolor M145. *ACS Chem. Biol.* **8,** 2009–2016 (2013).

18. 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).

19. Wilson, M. C. *et al.* An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506,** 58–62 (2014).

20. Engel, P., Vizcaino, M. I. & Crawford, J. M. Gut symbionts from distinct hosts exhibit genotoxic activity via divergent colibactin biosynthesis pathways. *Appl. Environ. Microbiol.* **81,** 1502–1512 (2015).

21. Yang, J. Y. *et al.* Molecular networking as a dereplication strategy. *J. Nat. Prod.* **76,** 1686–1699 (2013).

22. The National Institute of Standards and Technology. NIST. at <http://www.nist.gov/srd/nist1a.cfm>

23. Pruitt, K. D., Tatusova, T., Brown, G. R. & Maglott, D. R. NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Res.* **40,** (2012).

24. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28,** 45–48 (2000).

25. Bairoch, A. *et al.* The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33,** (2005).

26. Kersten, R. D. *et al.* Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E4407–16 (2013).

27. Guthals, A., Watrous, J. D., Dorrestein, P. C. & Bandeira, N. The spectral networks paradigm in high throughput mass spectrometry. *Molecular BioSystems* **8,** 2535 (2012).

28. Mascuch, S. J. *et al.* Direct detection of fungal siderophores on bats with white-nose syndrome via fluorescence microscopy-guided ambient ionization mass spectrometry. *PLoS One* **10,** e0119668 (2015).

29. Bandeira, N., Tsur, D., Frank, A. & Pevzner, P. Protein identification by spectral networks analysis. (2007).

30. 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,** 105–12 (2014).

31. Shannon, P. *et al.* Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).

32. Kildgaard, S. *et al.* Accurate dereplication of bioactive secondary metabolites from marine-derived fungi by UHPLC-DAD-QTOFMS and a MS/HRMS library. *Mar. Drugs* **12,** 3681–3705 (2014).

33. Matsuda, F. *et al.* AtMetExpress development: a phytochemical atlas of Arabidopsis development. *Plant Physiol.* **152,** 566–578 (2010).

34. Haug, K. *et al.* MetaboLights - An open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **41,** (2013).

35. Martens, L. *et al.* PRIDE: The proteomics identifications database. *Proteomics* **5,** 3537–3545 (2005).

36. A, H., H, K., WA, K., H, Z. & HJ, Z. [Metabolic products of microorganisms. 134. Stenothricin, a new inhibitor of the bacterial cell wall synthesis (author's transl)]. (1975).

37. Liu, W.-T. *et al.* MS/MS-based networking and peptidogenomics guided genome mining revealed the stenothricin gene cluster in Streptomyces roseosporus. *J. Antibiot. (Tokyo).* **67,** 99–104 (2014).

38. Marfey, P. Determination of D-amino acids. II. Use of a bifunctional reagent, 1,5-difluoro-2,4-dinitrobenzene. *Carlsberg Res. Commun.* **49,** 591–596 (1984).

39. Nonejuie, P., Burkart, M., Pogliano, K. & Pogliano, J. Bacterial cytological profiling rapidly identifies the cellular pathways targeted by antibacterial molecules. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 16169–74 (2013).

40. Lamsa, A., Liu, W. T., Dorrestein, P. C. & Pogliano, K. The Bacillus subtilis cannibalism toxin SDP collapses the proton motive force and induces autolysis. *Mol. Microbiol.* **84,** 486–500 (2012).

41. Purves, K. *et al.* Using Molecular Networking for Microbial Secondary Metabolite Bioprospecting. *Metabolites* **6,** 2 (2016).

42. Bertin, M. J. *et al.* Spongosine production by a Vibrio harveyi strain associated with the sponge Tectitethya crypta. *J. Nat. Prod.* **78,** 493–9 (2015).

43. Boudreau, P. D. *et al.* Expanding the Described Metabolome of the Marine Cyanobacterium Moorea producens JHB through Orthogonal Natural Products Workflows. *PLoS One* **10,** e0133297 (2015).

44. Kleigrewe, K. *et al.* Combining Mass Spectrometric Metabolic Profiling with Genomic Analysis: A Powerful Approach for Discovering Natural Products from Cyanobacteria. *J. Nat. Prod.* **78,** 1671–82 (2015).

45. Duncan, K. R. *et al.* Molecular networking and pattern-based genome mining improves discovery of biosynthetic gene clusters and their products from Salinispora species. *Chem. Biol.* **22,** 460–71 (2015).

46. Vizcaino, M. I. & Crawford, J. M. The colibactin warhead crosslinks DNA. *Nat. Chem.* **7,** 411–7 (2015).

47. Klitgaard, A., Nielsen, J. B., Frandsen, R. J. N., Andersen, M. R. & Nielsen, K. F. Combining Stable Isotope Labeling and Molecular Networking for Biosynthetic Pathway Characterization. *Anal. Chem.* **87,** 6520–6526 (2015).

48. Anderton, C. R., Chu, R. K., Tolić, N., Creissen, A. & Paša-Tolić, L. Utilizing a Robotic Sprayer for High Lateral and Mass Resolution MALDI FT-ICR MSI of Microbial Cultures. *J. Am. Soc. Mass Spectrom.* (2016). doi:10.1007/s13361-015-1324-6

49. Liaimer, A. *et al.* Nostopeptolide plays a governing role during cellular differentiation of the symbiotic cyanobacterium Nostoc punctiforme. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 1862–7 (2015).

50. Liu, Y. *et al.* Diversity of Aquatic Pseudomonas Species and Their Activity against the Fish Pathogenic Oomycete Saprolegnia. *PLoS One* **10,** e0136241 (2015).

51. He, X. *et al.* Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 244–9 (2015).

52. Cha, J.-Y. *et al.* Microbial and biochemical basis of a Fusarium wilt-suppressive soil. *ISME J.* **10,** 119–29 (2016).

53. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 12580–5 (2015).

54. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

55. Frewen, B. & MacCoss, M. J. Using BiblioSpec for creating and searching tandem MS peptide libraries. *Curr. Protoc. Bioinformatics* **Chapter 13,** Unit 13.7 (2007).

56. 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).

57. Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7,** 655–667 (2007).

58. Frank, A. M. *et al.* Clustering millions of tandem mass spectra. *J. Proteome Res.* **7,** 113–122 (2008).

59. Shirling, E. B. & Gottlieb, D. Methods for characterization of Streptomyces species. *International Journal of Systematic Bacteriology* **16,** 313–340 (1966).

60. Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* **16,** 944–945 (2000).

61. Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28,** 464–469 (2012).

62. Carver, T. *et al.* Artemis and ACT: Viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24,** 2672–2676 (2008).

63. Röttig, M. *et al.* NRPSpredictor2 - A web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39,** (2011).

64. Rausch, C., Hoof, I., Weber, T., Wohlleben, W. & Huson, D. H. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol. Biol.* **7,** 78 (2007).

65. Thompson, J. D., Gibson, T. J. & Higgins, D. G. Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protoc. Bioinformatics* **Chapter 2,** Unit 2.3 (2002).

66. Liu, N. J. L., Dutton, R. J. & Pogliano, K. Evidence that the SpoIIIE DNA translocase participates in membrane fusion during cytokinesis and engulfment. *Mol. Microbiol.* **59,** 1097–1113 (2006).

Figure 1 – GNPS Overview. (a) Representation of interactions between the natural product community, GNPS spectral libraries, and GNPS datasets. At present 221,083 MS/MS spectra from 18,163 unique compounds are used for the search at GNPS. These include both 3rd party libraries such as MassBank, ReSpect, and NIST, as well as, spectral libraries created for GNPS (GNPS-Collections) and spectra from the natural product community (GNPS-Community). GNPS spectral libraries grow through user contributions of new identifications of MS/MS spectra. To date, 55 community members have contributed 8,853 MS/MS spectra from 5,568 unique compounds (30.5% of the unique compounds available). In addition, on-going curation efforts have already yielded 563 annotation updates for library spectra. The utility of these libraries is to dereplicate compounds (recognition previously characterized and studied known compounds), in both public and private data. This dereplication process is performed on all public datasets and results are automatically reported, thus enabling users to query for all datasets/organisms/conditions that a particular molecule occurred. Automatic reanalysis of all public data creates a virtuous cycle where new contributions to libraries see immediate impact in the form of matches to all public data. Combined with molecular networking (**Fig. 3**), this automatic analysis empowers community members to identify novel analogs that can then be added to GNPS spectral libraries. (b) GNPS as an analysis platform has grown to serve a global user base including 9,200+ users from 100 countries.

Figure 2 – GNPS spectral libraries. (a) The various computational resources of the metabolomics and natural products community are categorized into two main categories: i) Reference collections (red dots) of MS/MS spectral libraries and ii) Data Repositories (blue dots)designed to publicly share raw mass spectrometry data associated with research projects. Reference collection resources are contributors and aggregators of reference MS/MS spectra, some of which also include data analysis tools, e.g. online multi-spectrum MS/MS search (magnifying glass icon). Several resources have aggregated MS/MS spectra from various reference collections so that the analysis tools at a respective resource can leverage more of the community efforts to annotate data (red and blue arrows). GNPS has imported all freely available reference collections (>221,000 MS/MS spectra) and makes these available for online analysis. GNPS and several other resources provide both reference MS/MS spectra and data in an open and free manner to the public (pink caps). (b) Comparison of spectral library sizes of available libraries (MassBank, ReSpect, and NIST) and GNPS libraries; GNPS-Collections includes newly acquired spectra from synthetic or purified compounds and GNPS-Community includes all community-contributed spectra. (c) Searching all public GNPS datasets revealed that Massbank/ReSpect/NIST libraries matched to 1,217 unique compounds, with GNPS libraries increasing unique compound matches by 41% (corresponding to 29% of total unique matches) with an accompanying 4% increase in spectral library size. Overall, GNPS libraries increase the total number of spectra matched in public datasets by 144% (59% of total public MS/MS matches) and spectra matches across all GNPS public and private data by 767% (88% of all MS/MS matches). (d) The distribution of precursor masses in all GNPS public datasets is shown in gray and

compared to the precursor mass distributions of Massbank, ReSpect, NIST, and GNPS libraries. Though GNPS libraries have a combined size that is significantly smaller than MassBank/ReSpect/NIST, GNPS libraries have a stronger emphasis on molecules in the higher *m/z* range and thus complement the emphasis on lower precursor mass molecules in existing libraries. (e) The quality of spectrum matches obtained by searching against the available spectral libraries is assessed with user ratings (1 to 4 stars see **Supplementary Table 6**) of continuous identification results. The high quality of GNPS library spectra is illustrated by user ratings of 2.5+ stars for 98%+ of GNPS library matches, which compares favorably to the 90% mark for NIST matches, whose high marks demonstrate how important these 3rd party libraries still are to the GNPS platform. We note that the lower mark for NIST matches does not suggest lower quality spectra, as it is more likely explained by its higher emphasis on lower precursor mass molecules with spectra that have fewer peaks and are generally harder to match.

Figure 3 – Molecular Network Creation and Visualization. (a) Molecular networks are constructed from the alignment of MS/MS spectra to one another. Edges connecting nodes (MS/MS spectra) are defined by a modified cosine scoring scheme determines the similarity of two MS/MS spectra with scores ranging from 0 (totally dissimilar) to 1 (completely identical). MS/MS spectra are also searched against GNPS Spectral Libraries, seeding putative nodes matches in the molecular networks. Networks are visualized online in-browser or exported for third party visualization software such as Cytoscape[31]. (b) An example alignment between three MS/MS spectra of compounds with structural modifications that are captured by modification tolerant spectral matching utilized in variable dereplication and molecular networking. (c) In-browser molecular network visualization enables users to interactively explore molecular networks without requiring any external software. To date, over 11,000 molecular networks have been analyzed using this feature. Within this interface, (i) users are able to define cohorts of input data and correspondingly, nodes within the network are represented as pie charts to visualize spectral count differences for each molecule across cohorts. (ii) Node labels indicate matches made to GNPS spectral libraries, with additional information displayed with mouseovers. These matches provide users a starting point to annotate unidentified MS/MS spectra within the network. (iii) To facilitate identification of unknowns, users can display MS/MS spectra in the right panels by clicking on the nodes in the network, giving direct interactive access to the underlying MS/MS peak data. Furthermore, alignments between spectra are visualized between spectra in the top right and bottom right panels in order to gain insight as to what underlying characteristics of the molecule could elicit fragmentation perturbations.

Figure 4 – "Living data" in GNPS through crowdsourcing of molecular annotations. (a) A global snapshot of the state of MS/MS matching of public natural product datasets available at GNPS using molecular networking and library search tools. Identified molecules (1.9% of the data) are MS/MS spectrum matches to library spectra with a cosine greater than 0.7. Putative Analog

Molecules (another 1.9% of the data) are MS/MS spectra that are not identified by library search but rather are immediate neighbors of identified MS/MS spectra in molecular networks. Identified Networks (9.9% of the data) are connected components within a molecular network that have at least one spectrum match to library spectra. Unidentified Networks (25.2% of the data) are molecular networks where none of the spectra match to library spectra; these networks potentially represent compound classes that have not yet been characterized. Exploratory Networks (an additional 20.1% of the data) are unidentified connected components in molecular networks with more relaxed parameters (**Supplementary Table 12**). Thus, 55.3% of the MS/MS spectra at least have one related MS/MS spectrum in spectral networks, with 44.7% having none. In this 44.7% of the data, each MS/MS spectrum has been observed in two separate instances and should not constitute noise. Altogether, this analysis indicates that the vast amount chemical space captured by mass spectrometry remains unexplored. (b) In the past year, there has been significant growth in the GNPS spectral libraries, driving growth in the match rates of all public data. The number of unique compounds matched in the public data has increased 10x; the number of total spectra matched has increased 22x; and the average match rate has increased 3x. It is expected that identification rates will continue to grow with further contributions from the community to the GNPS-Community spectral library.

Figure 5 - GNPS enabled discovery of a new chemical entity. a) The stenothricin molecular family identified during analysis of a molecular network between chemical extracts of *S. roseosporus* NRRL 15998 (Green) and *Streptomyces* sp. DSM5940 (Blue). This analysis indicates that *Streptomyces* sp. DSM5940 produces a structurally similar compound to stenothricin with a -41 Da m/z difference. An enlarged version of the network can be found in the supporting information. b) Based on preliminary structural analysis, stenothricin-GNPS, the -41 Da new chemical entity, is proposed to be due to a Lys to Ser substitution. c) Comparison of the MS/MS of stenothricin D with its -41 Da analog stenothricin-GNPS 2. d) Although structurally related, stenothricin and stenothricin-GNPS have different effects on *E. coli* as visualized using fluorescence microscopy. Red is the membrane stain FM4-64, blue is the membrane permeable DNA stain DAPI, green is the membrane impermeable DNA stain SYTOX green. SYTOX green only stains DNA when the cell membrane is damaged. The scale bar represents 2 μm.

	Summary	*Data repository*	*Reference collections*	*Open online data analysis*	*Pubmed*
GNPS	Natural products and metabolomics crowdsourced analysis infrastructure with public reference libraries, public data repository and living data	Yes, living data with automated reanalysis, minimal required metadata (220 w/MS2, 274 total)	Yes, open access, crowdsourced curation	Can search any number of files, analog searches and molecular networking (G,J,E,NA,R,H,N)	
Reference Collections					
MassBank Japan	The first public large scale database for metabolomics reference spectra.		Yes, open access	Can search up to one file at a time (J)	20623627
MassBank Europe	European counterpart of massbank japan. This public reference spectral library is under construction to include draft structures.		Yes, open access	Can search up to one file at a time (J,E)	
MassBank North America	North American public spectral library warehouse and distribution database.		Yes, open access	Can search up to one file at a time (G,J,NA,R,H)	
ReSpect	Public reference library for plant metabolites.		Yes, open access	Can search single spectrum (R)	22867903
HMDB	Public reference library for human metabolites.		Yes, open access	Can search single spectrum (H)	17202168
XCMS-online/Metlin	Reference library for metabolomics. Can be searched but the library is commercial and not available for public redistribution.	Yes, no reanalysis (10 w/MS2, 23 total)	Yes, not freely available	Can search any number of files up to 25Gb (Mt)	16404815
NIST/EPA/NIH	Reference libraries for metabolomics. Accessible through purchase but not available for redistribution.		Yes, not freely available		
mzCloud	A metabolomics search engine and reference library. The library is not available to the scientific community.		Yes, not freely available		
Data Repositories					
Metabolights	Public data repository for metabolomics data, library capabilities under construction.	Yes, no reanalysis, experimental metadata (13 w/MS2, 131 total)	Aggregator only		23109552
Metabolomics workbench	Public data repository for metabolomics data.	Yes, no reanalysis, extensive metadata required (9 w/open format MS2, 196 total)	Aggregator only		26467476

Table 1 - Metabolomics and Natural Products MS/MS Computational Resources Overview – The various computational resources available to the MS/MS-based metabolomics and natural product communities. For each resource a short summary is provided along with the URL and PubMed identifier for the associated publication. High level core functionality is also listed for each resource. Data repository – denotes whether a resource is designed to publicly share projects data with the community or between different research groups. Total number of MS/MS datasets and total datasets are shown in parenthesis. Reference collection of MS/MS spectra – indicates

whether resources contribute new MS/MS reference spectra to spectral libraries (rather than redistributing them); mode of access to download the MS/MS reference spectra is clarified. Online analysis utilizing MS/MS reference spectra available at each resource, with emphasis on batch capabilities; the MS/MS spectral libraries available for searches at each resource are highlighted with the following notation: GNPS libraries (G), MassBank JP libraries (J), MassBank EU libraries (E), MassBank of North America libraries (NA), HMDB libraries (H), ReSpect libraries (R), NIST libraries (N), Metlin libraries (Mt), mzCloud libraries (Mz).

Methods

Spectral Library Searching

Input MS/MS spectra (i.e., query spectra) are considered matched to library spectra if they meet the following criteria: same precursor charge state, precursor *m/z* is within a user defined Thompson tolerance, share a minimum number of matched peaks, and exceed a user-defined minimum spectral match score. Exact spectral matches between library and query spectra are scored with a normalized dot product[55–57]. The matching of peaks between two spectra is formulated as a maximum bipartite matching problem[15] where peaks from the library and query spectra are represented as nodes with edges connecting library and query peaks. Edges connect peaks that are within a user defined fragment mass tolerance. The bipartite match of library to query peaks that maximizes the normalized dot product is selected. The highest scoring library match for each query spectrum is reported. Estimated false discovery rates of the exact spectral library search are shown in **Supplementary Note 3**. Parameters of the search can be found in **Supplementary Table 13**.

Variable Dereplication

Variable dereplication utilizes a modification tolerant spectral library search. Similar to exact spectral matches, except additional edges are added to the bipartite matching between library and query peaks which differ by a δ (as determined by their precursor mass difference δ) +/- the user defined fragment mass tolerance.

Molecular Network Construction

Molecular networks can be constructed from any collection of MS/MS spectra. First, all MS/MS spectra are clustered with MSCluster[58] such that MS/MS spectra found to be identical are merged into a consensus spectrum. Consensus spectra are then matched against each other using the modification tolerant spectral matching scheme[15]. All spectrum-to-spectrum matches that exceed a user defined minimum match score are retained. MS/MS spectra are then represented as nodes in a graph and significant matches between spectra are represented as edges. Further, edges in the graph are only retained if the two nodes, A and B, connected by a given edge satisfy the following properties: i) B must be in the top K highest scoring neighbors of A and ii) A must be in the top K highest scoring neighbors of B. All other edges are removed.

GNPS Collections – Sample Preparation

The NIH Prestwick Phytochemical Library, NIH Natural Product Library, and NIH Small Molecule Pharmacologically Active Library compounds were received as stock solutions of pure compounds (10 mM in DMSO). They were reformatted by 1 µL of each compound into 89 µL of methanol into

96 well plates with 11 distinct compounds in each well. They were further diluted 100-fold for a final 1 µM concentration.
The NIH Clinical Collections and FDA Library part 2 were received as stock solutions of pure compounds (10 mM in DMSO). They were diluted to final concentration of 1 µM in 50:50 methanol:water and formatted onto 96 well plates with 10 compounds per well.

GNPS Collections – LC MS/MS Acquisition

LC-MS/MS acquisition for all in house generated libraries was performed using a Bruker Daltonics Maxis qTOF mass spectrometer equipped with a standard electrospray ionization source (ESI). The mass spectrometer was tuned by infusion of Tuning Mix ES‐TOF (Agilent Technologies) at a 3 µL/min flow rate. For accurate mass measurements, lock mass internal calibration used a wick saturated with hexakis (1H,1H,3H‐tetrafluoropropoxy) phosphazene ions (Synquest Laboratories, *m/z* 922.0098) located within the source. Samples were introduced by a Thermo Scientific UltraMate 3000 Dionex UPLC using a 20 µL injection volume. A Phenomenex Kinetex 2.6 µm C18 column (2.1 mm × 50 mm) was used. Compounds from NIH Prestwick Phytochemical Library, NIH Natural Product Library, and NIH Small Molecule Pharmacologically Active Library were separated using a seven minute linear water‐acetonitrile gradient (from 98:2 to 2:98 water:acetonitrile) containing 0.1% formic acid. Compounds from NIH Clinical Collections and FDA Library part 2 Library employed a step gradient for chromatographic separation [5% solvent B (2:98 water:acetonitrile) containing 0.1% formic acid for 1.5 min, a step gradient of 5% B-50% B in 0.5 min, held at 50% B for 2 min, a second step of 50% B-100% B in 6 min, held at 100% B for 0.5 min, 100%-5 % B in 0.5 min and kept at 5% B for 0.5 min]. The flow rate was 0.5 mL/min. The mass spectrometer was operated in data dependent positive ion mode; automatically switching between full scan MS and MS/MS acquisitions. Full scan MS spectra (*m/z* 50 – 1500) were acquired in the TOF and the top ten most intense ions in a particular scan were fragmented using collision induced dissociation (CID) utilizing stepping.

GNPS Collections – Spectral Library Creation

All raw data were centroided and converted to 32-bit uncompressed mzXML file using Bruker Data Analysis. A script was developed to select all possible MS/MS spectra in each LC-MS/MS run that could correspond to a compound present in the sample. For each compound, we calculated the theoretical mass M from its chemical composition and searched for the M+H, M+2H, M+K, and M+Na adducts. Putative identifications included all MS/MS spectra whose precursor *m/z* had a ppm error <50 compared to the theoretical mass of each possible precursor *m/z*; all tandem MS/MS spectra with an MS1 precursor intensity of <1E4 were ignored. All candidate identifications were manually inspected and the most abundant representative spectrum for each compound was added to the corresponding library at the gold or bronze level based upon an expert evaluation of the spectrum quality. The best MS/MS spectrum per compound as added to the GNPS-Collections library without filtering or alteration from the mzXML files.

GNPS-Community Contributed Spectral Library Processing and Control

User contributed library spectra are not filtered or altered in any way from the user submission. MS/MS spectra are extracted from the submitted data and are made available in the GNPS libraries. The list and description of metadata fields can be found in GNPS online documentation. To preserve provenance information, the full input file is also retained and made available for download for each library spectrum (e.g. link). Different levels of reference spectra submissions are enforced with access restrictions on a per user basis. The description of each of the quality levels: Gold, Silver and Bronze and be found in **Supplementary Table 3**. While any MS/MS spectrum can be Bronze quality level in the GNPS libraries, Silver contributions require peer-reviewed publication of the MS/MS spectra, and Gold contributions require MS/MS spectra to be of synthetics or purified compounds with complete structural characterization.

Materials and Strains

Streptomyces sp. DSM5940, obtained from Eberhard-Karls-Universität Tübingen, Germany, was originally isolated from a soil sample collected from the Andaman Islands, India. *Streptomyces roseosporus* NRRL 15998 was acquired from the Broad Institute, MIT/Harvard, MA, USA, whose parent strain *S. roseosporus* NRRL 11379 was isolated from soil from Mount Ararat in Turkey. All media components were purchased from Sigma-Aldrich. Organic solvents were purchased from JT Baker at the highest purity.

Streptomyces sp. DSM5940 and *S. roseosporus* Metabolite Extraction

S. roseosporus and *Streptomyces. sp.* DSM5940 were inoculated by 4 parallel streaks onto individual ISP2 agar plates[59]. After incubating for 10 d at 28 °C, the agar was sliced into small pieces and put into a 50 mL centrifuge tube containing 1:1 water:*n*-butanol and shaken at 225 rpm for 12 h. The *n*-butanol layer was collected via transfer pipette, centrifuged, and dried with *in vacuo*.

Streptomyces sp. DSM5940 and *S. roseosporus* MS/MS Acquisition

MS/MS spectra for crude extracts of *S. roseosporus* and *Streptomyces* sp. DSM were collected as previously described[37]. Briefly, MS/MS spectra were collected using direct infusion using an Advion nanomate-electrospray robot and capillary liquid chromatography using a manually pulled 10 cm silica capillary packed with C18 reverse phase resin. Samples were introduced for capillary LC using a Surveyor system using a 10mL injection (10 ng/μL in 10% ACN). Metabolites were separated using a time variant gradient [(minutes, % of solvent B): (20, 5), (30, 60), (75, 95) where solvent A is water with 0.1% AcOH and B is ACN with 0.1% AcOH] using a 200mL flowrate (1% to instrument source with 1.8kV source voltage). Both methods utilized detection by a Thermo Finnigan LTQ/FT-ICR mass spectrometer. The mass spectrometer was operated in data dependent positive ion mode; automatically switching between full scan high resolution FT MS and

low resolution LTQ MS/MS acquisitions. Full scan MS spectra were acquired in the FT and the top six most intense ions in a particular scan were fragmented using collision induced dissociation (CID) at a constant collision energy of 35eV, an activation Q of 0.25, and an activation time of 50 to 80 ms. RAW files were converted to .mzXML using ReAdW.

Molecular Networking Parameters

A molecular network was created at GNPS data from the *S. roseosporus* and *Streptomyces* sp. DSM5940 MS/MS data. The specific job is browse-able online (link). Full parameters can be found in **Supplementary Table 14**.

Stenothricin-GNPS extraction and purification

400 ISP2 agar plates were inoculated with spore suspension of *Streptomyces sp.* DSM5940 strain and incubated for 10 d at 30 °C. The agar was sliced into small pieces and extracted twice with 1:1 water:*n*-butanol for 12 h at 28 °C and 225 rpm in two 2.8 L Fernbach flasks. Agar pieces were removed by filtration. The resultant filtrate was centrifuged and the *n*-butanol layer was collected, dried and resuspended in 1 mL methanol. The extract was fractionated using a Sephadex LH20 column utilizing a methanol mobile phase at a flow rate of 0.5 mL/min. Each fraction was analyzed by dried droplet MALDI-TOF MS for the *m/z* values corresponding to stenothricin-GNPS. For this analysis, 1 mL of each fraction was mixed 1:1 with a saturated solution of Universal MALDI matrix (Sigma-Aldrich) in 78 % acetonitrile containing 0.1 % TFA and spotted on a Bruker MSP 96 anchor plate. The sample was dried and analyzed by either a Microflex or Autoflex MALDI-TOF MS (Bruker Daltonics). Mass spectra were obtained using the FlexControl software and a single spot acquisition of 80 shots. MALDI-TOF MS data was analyzed by FlexAnalysis software. Fractions containing *m/z* values putatively assigned to stenothricin-GNPS were combined and further purified by a two-step reversed-phase HPLC procedure (Solvent A: water with 0.1% TFA; Solvent B: ACN with 0.1% TFA). Initial HPLC analysis (SUPELCO C18, 5 µm, 100 Å, 250 x 10.0 mm) utilized a linear gradient from 50% to 75% solvent B in 35 min at flow rate 2 mL/min. Fractions containing target peptide *m/z* values as detected by MALDI-TOF MS were collected, combined, and evaporated. Subsequent HPLC analysis (Thermo, Syncronis Phenyl HPLC, 5 µm, 150 x 4.6 mm) used an isocratic elution with 35% solvent B. Purified stenothricin-GNPS 2 (*m/z* 1091) and 3 (*m/z* 1105) were lyophilized and stored at -80 °C.

Stenothricin-GNPS NMR

50 µg stenothricin-GNPS 2 was dissolved in 30 µL of CD_3OD for NMR acquisition. ^{1}H-NMR spectra were recorded on Bruker Avance III 600 MHz NMR with 1.7 mm Micro-CryoProbe at 298 K, with standard pulse sequences provided by Bruker. The NMR spectrum was overlayed with the NMR spectrum from stenothricin D and analyzed using the MestReNova software[37].

Genome sequencing and de novo assembly *Streptomyces sp.* DSM5940

Streptomyces sp. DSM5940 genome was subjected to partial genome sequencing by Ion Torrent and Illumina MiSeq with paired end sequencing. The resulting contigs were assembled by Geneious 5.1.1 using the *S. roseosporus* 15998 genome sequence as template. Sequences have been deposited in NCBI with accession number assignment pending.

Sequence definition of the gene cluster in *Streptomyces* sp. DSM5940

To identify the Strenothricin-GNPS gene cluster, the *Streptomyces* sp. DSM5940 genome was annotated using Artemis[60,61]. Non-ribosomal peptide synthesis (NRPS) biosynthetic gene clusters were manually assigned using the Artemis Comparison Tool (an "all-against-all" BLAST (NCBI) comparison of proteins within the database)[62]. The adenylation domains of each NRPS gene cluster were further assessed using NRPSpredictor2[63,64]. The predicted 10 amino acid codes for each A-domain within the NRPS gene clusters was manually compared to those predicted for the putative stenothricin gene cluster from *S. roseosporus*[37]. The gene cluster with highest A-domain similarity was putatively identified as the stenothricin-GNPS gene cluster. Full sequence alignment of both the stenothicin-GNPS and stenothricin using ClustalW2 confirmed high sequence identity and similarity[65].

Phylogenetic Analysis of C-domains

To determine whether the stenothricin and stenothricin-GNPS gene clusters code for similar amino acid stereochemistry, the condensation domain (C-domain) sequences in the putative stenothricin-GNPS and stenothricin gene clusters were aligned with a subset of C-domain sequences representing the six C-domain families (heterocyclization, epimerization, dual condensation/epimerization (dual), condensation of L amino acids to L amino acids (L to L), and condensation of D amino acids to L amino acids (D to L), and starter) using ClustalW2[65].

Fluorescence Microscopy

A pre-culture of *E. coli* lptD cells (NR698) was grown to saturation, then diluted 1:100 into 20 mL LB. Flasks were incubated at 30˚C until an OD_{600} of 0.2 was reached. Cultures were then mixed with the appropriate amount of compound. Compounds were used at the following final concentrations: 1% MeOH, 0.5% DMSO, 20 µg/mL stenothricin D, 40 µg/mL stenothricin-GNPS 2/3. 15 µL of treated cells were transferred into a 1.7 mL tube and incubated at 30˚C in a roller. Samples were collected for imaging at 2 hours. 6 µL of cells were added to 1.5 µL of dye mix (30 µg/mL FM 4-64, 2.5 µM SYTOX green and 1.2 µg/mL DAPI) prepared in 1X T-base, and immobilized on an agarose pad (20% LB, 1.2% agarose) prior to microscopy. All microscopy was performed on an Applied Precision Spectris microscope as previous described[66] Images were deconvolved using softWoRx V 5.5.1 and the medial focal plane shown. The SYTOX green images were normalized within **Figure 5d** based on intensity and exposure length relative to the treatment with the highest fluorescence intensity.

Code availability

Source code and license is available at the CCMS software tools webpage.

a

18,163 Compounds
221,083 Spectra

GNPS Spectral Libraries

Knowledge Extraction into new library spectra and metadata

Automated Dereplication

Natural Product Scientific Community

Library Annotation/Curation

55 Contributors
563 Annotation Revisions

Dataset Uploads

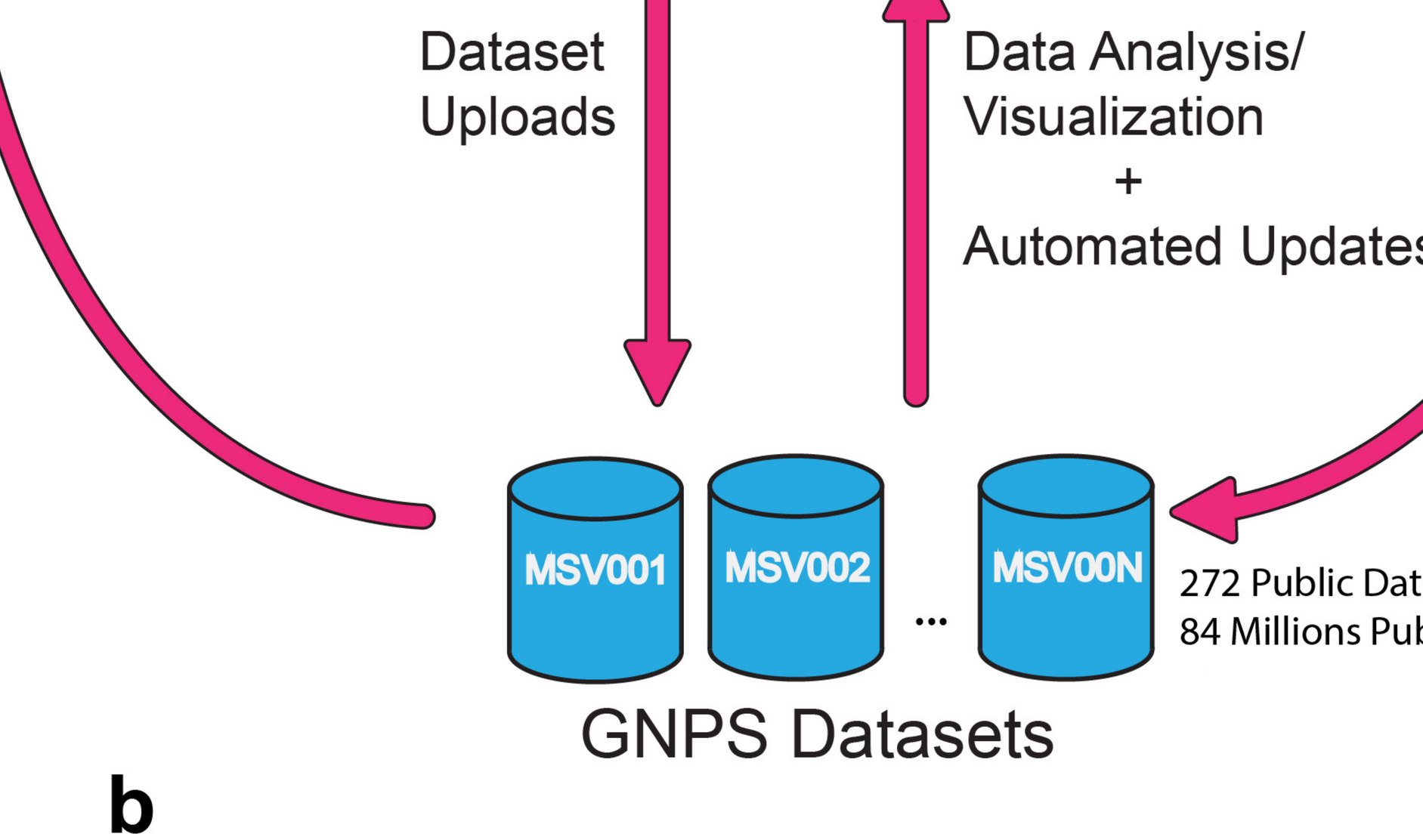
Data Analysis/ Visualization + Automated Updates

MSV001 MSV002 ... MSV00N

272 Public Datasets
84 Millions Public MS/MS Spectra

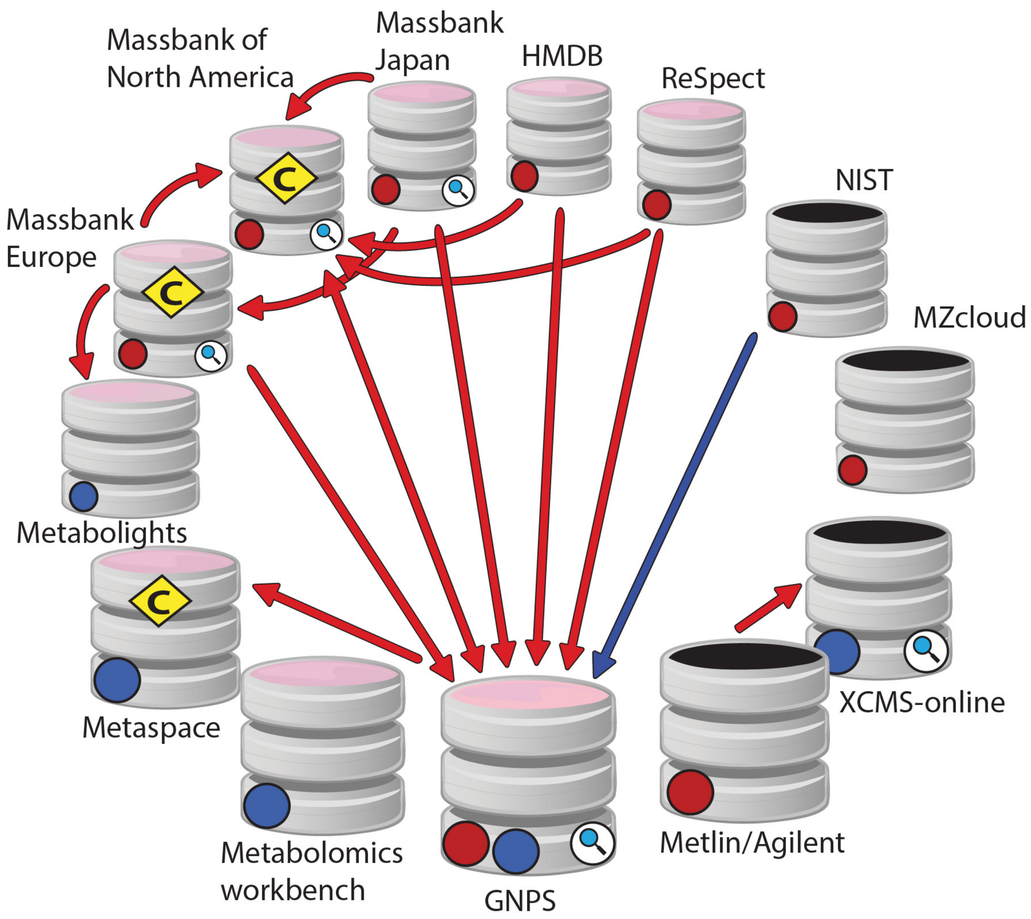
GNPS Datasets

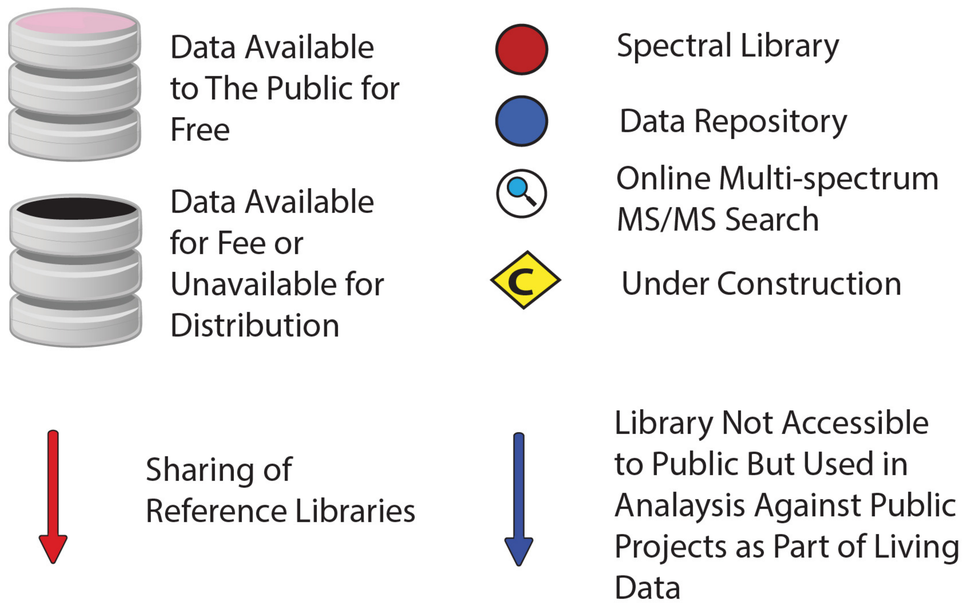
b



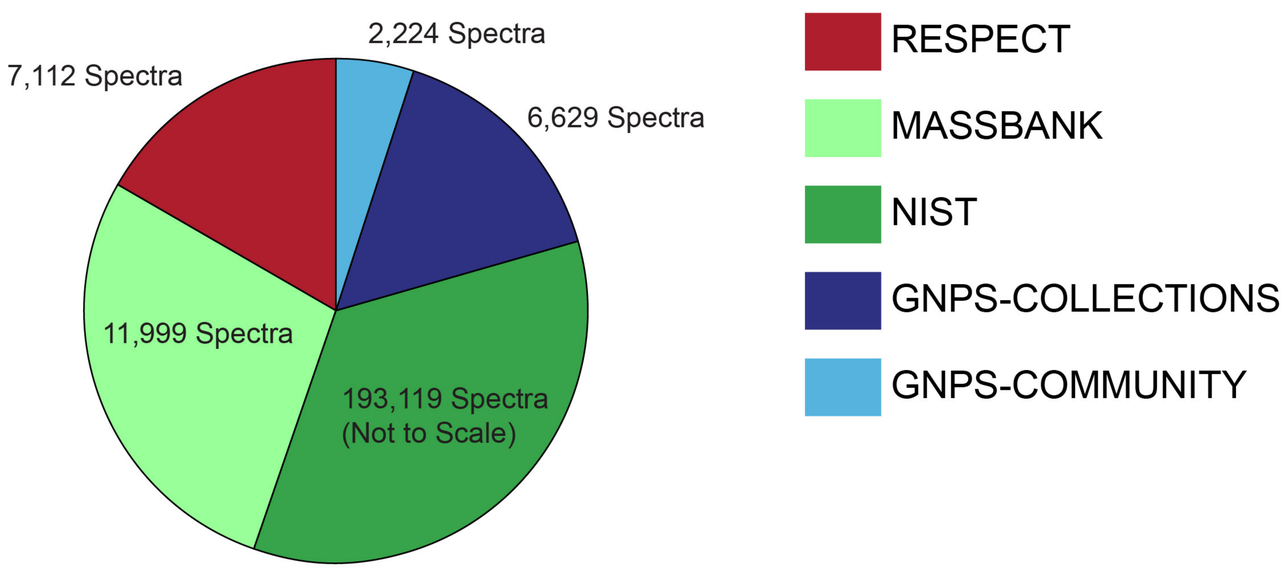
100 Countries
9,267 Users
42,486 Analysis Session
257,975 LC/MS Runs Analyzed
93 Million Spectra Searched

a

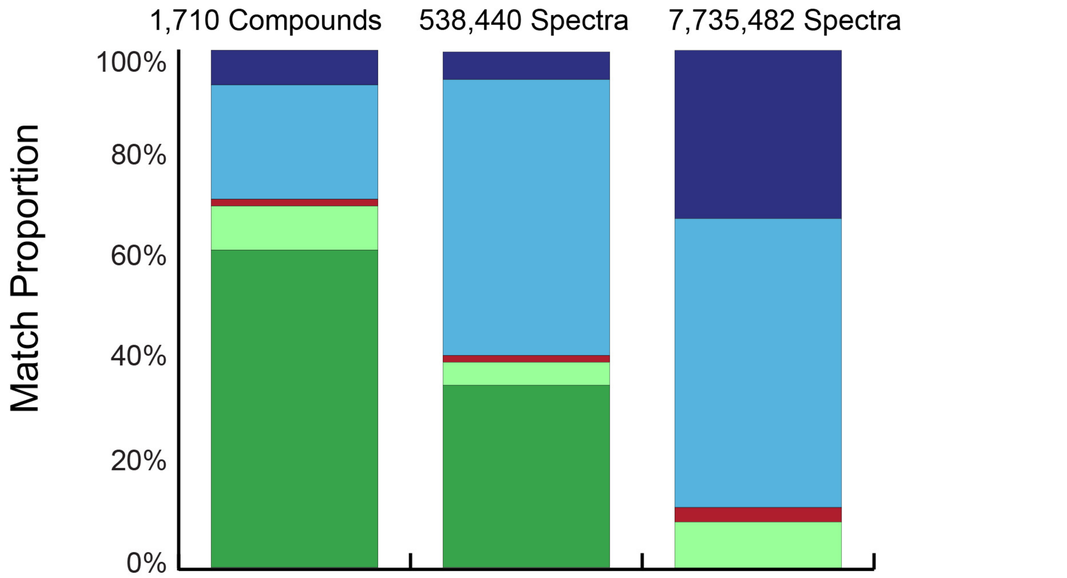

Massbank of North America
Massbank Japan
HMDB
ReSpect
NIST
MZcloud
Massbank Europe
Metabolights
Metaspace
XCMS-online
Metabolomics workbench
GNPS
Metlin/Agilent


Data Available to The Public for Free
Data Available for Fee or Unavailable for Distribution
Sharing of Reference Libraries
Spectral Library
Data Repository
Online Multi-spectrum MS/MS Search
Under Construction
Library Not Accessible to Public But Used in Analaysis Against Public Projects as Part of Living Data

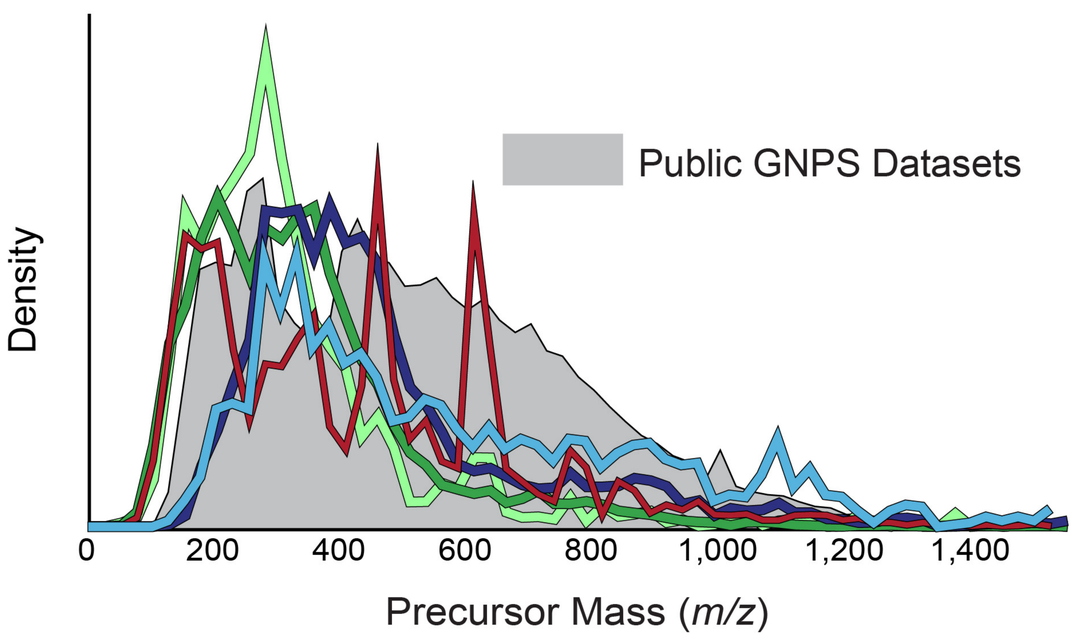
b


Current Spectral Library Size
2,224 Spectra
7,112 Spectra
6,629 Spectra
11,999 Spectra
193,119 Spectra (Not to Scale)
RESPECT
MASSBANK
NIST
GNPS-COLLECTIONS
GNPS-COMMUNITY

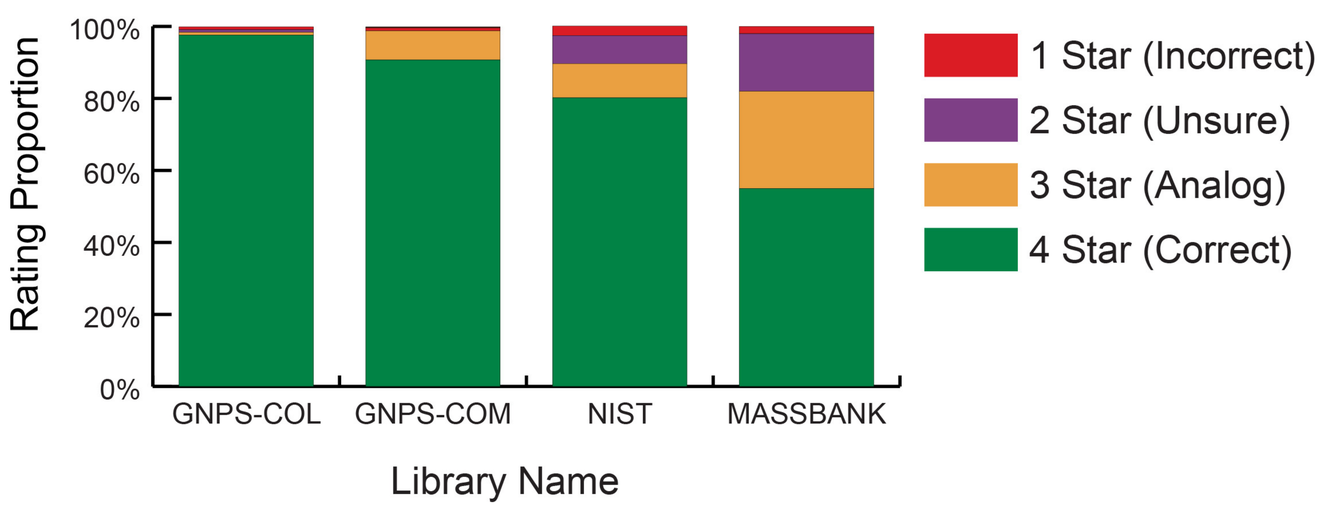
c

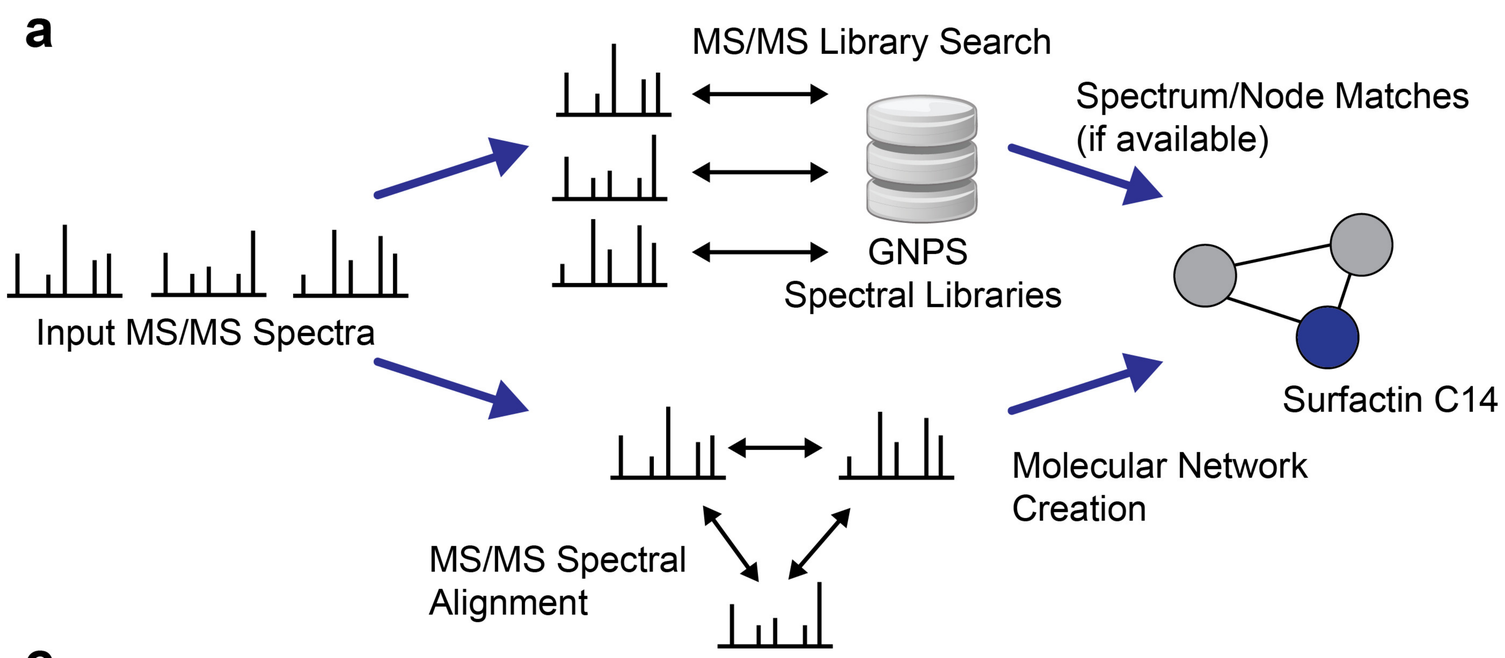

Fraction of Matches per Spectral Library
1,710 Compounds
538,440 Spectra
7,735,482 Spectra
Match Proportion
100%
80%
60%
40%
20%
0%
Number of unique compounds matched in GNPS datasets
Number of spectra matched in GNPS datasets
Number of spectra matched in all GNPS searches

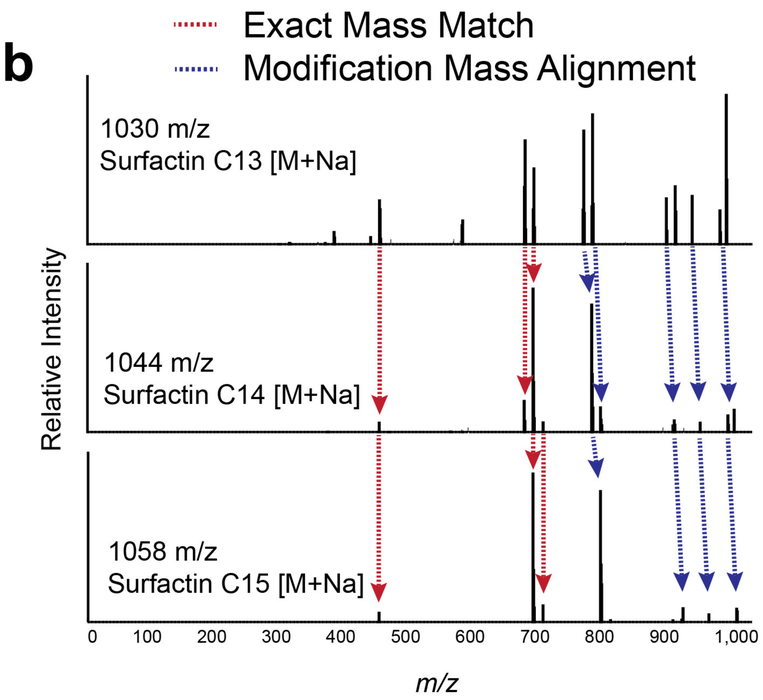
d

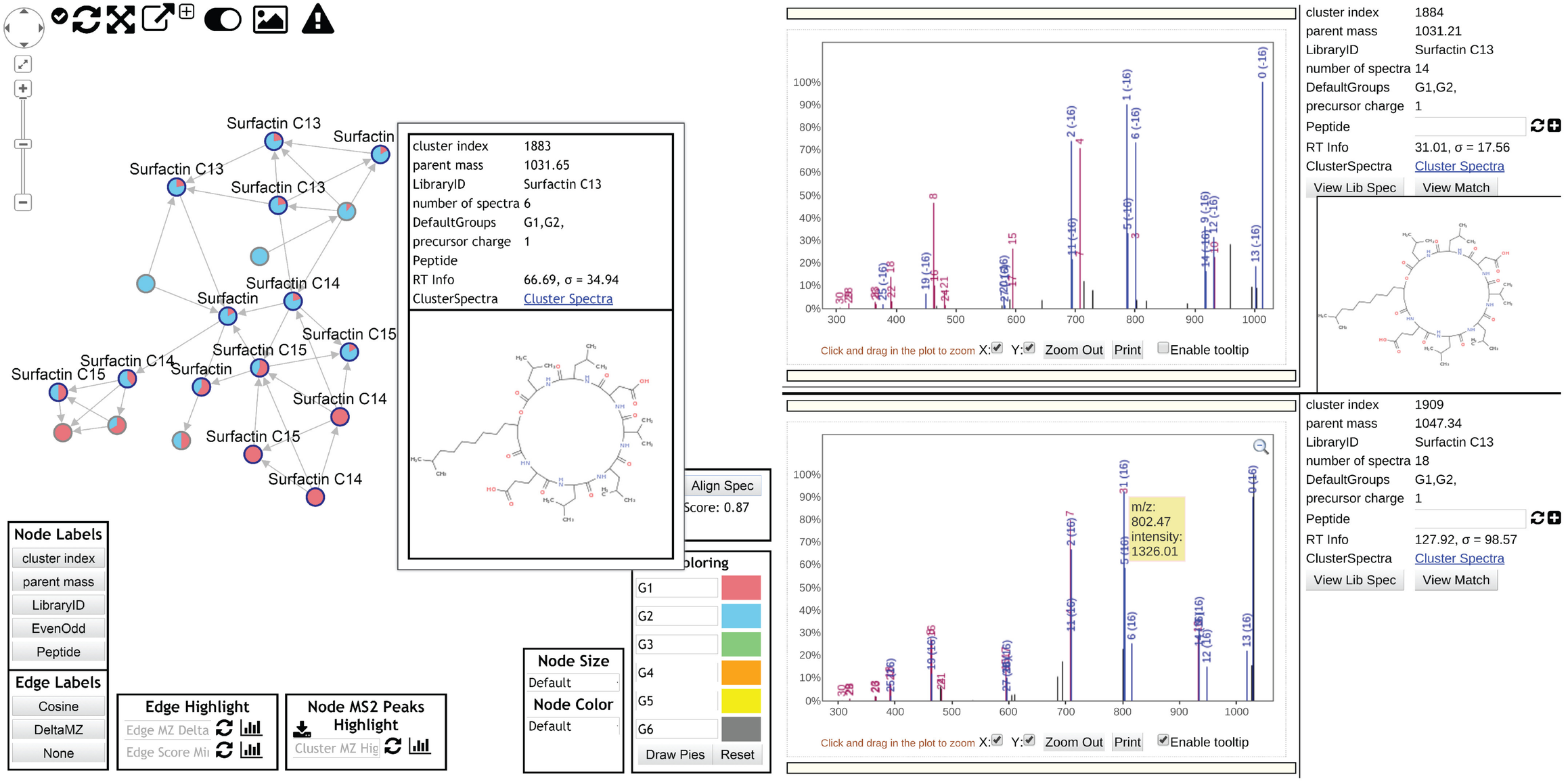

Precursor Mass Distribution
Public GNPS Datasets
Density
0
200
400
600
800
1,000
1,200
1,400
Precursor Mass (m/z)

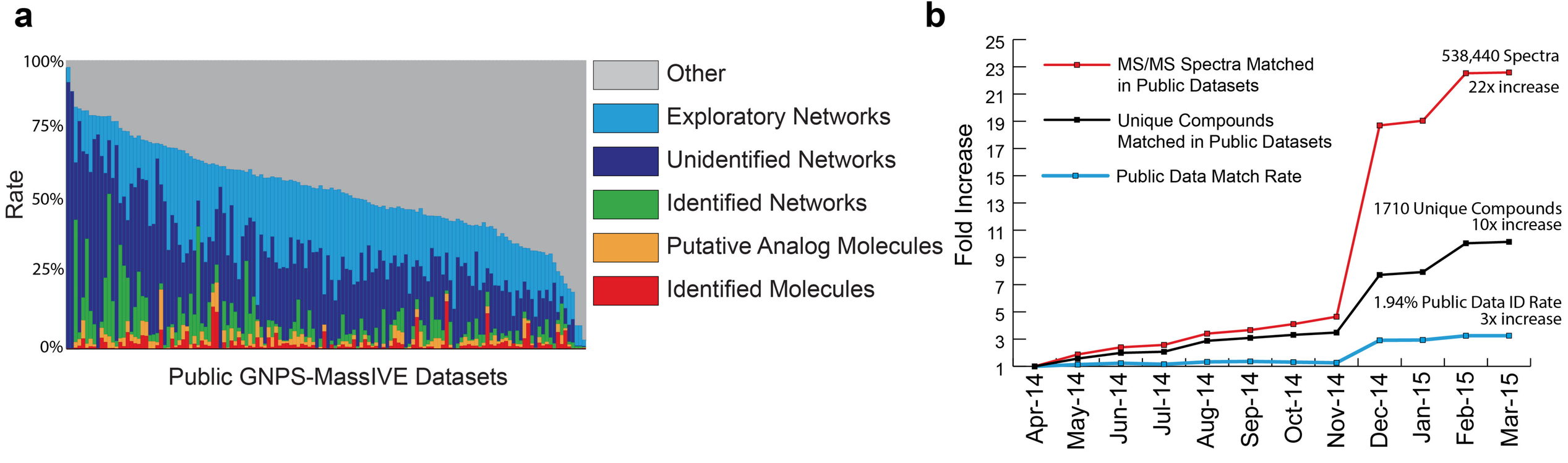
e


User Ratings of the Quality of Spectral Matches
Rating Proportion
100%
80%
60%
40%
20%
0%
GNPS-COL
GNPS-COM
NIST
MASSBANK
Library Name
1 Star (Incorrect)
2 Star (Unsure)
3 Star (Analog)
4 Star (Correct)

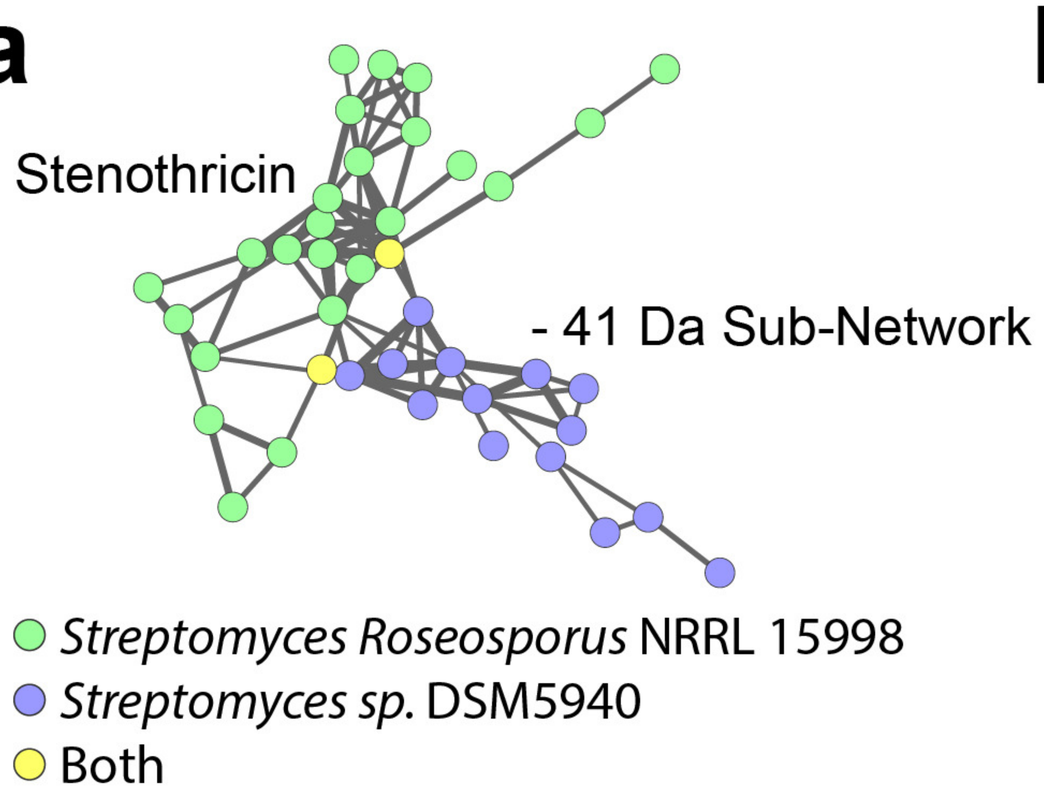

a
MS/MS Library Search
Spectrum/Node Matches
(if available)
GNPS
Spectral Libraries
Input MS/MS Spectra
Surfactin C14
Molecular Network
Creation
MS/MS Spectral
Alignment


b
Exact Mass Match
Modification Mass Alignment
1030 m/z
Surfactin C13 [M+Na]
1044 m/z
Surfactin C14 [M+Na]
1058 m/z
Surfactin C15 [M+Na]
Relative Intensity
0
100
200
300
400
500
600
700
800
900
1,000
m/z

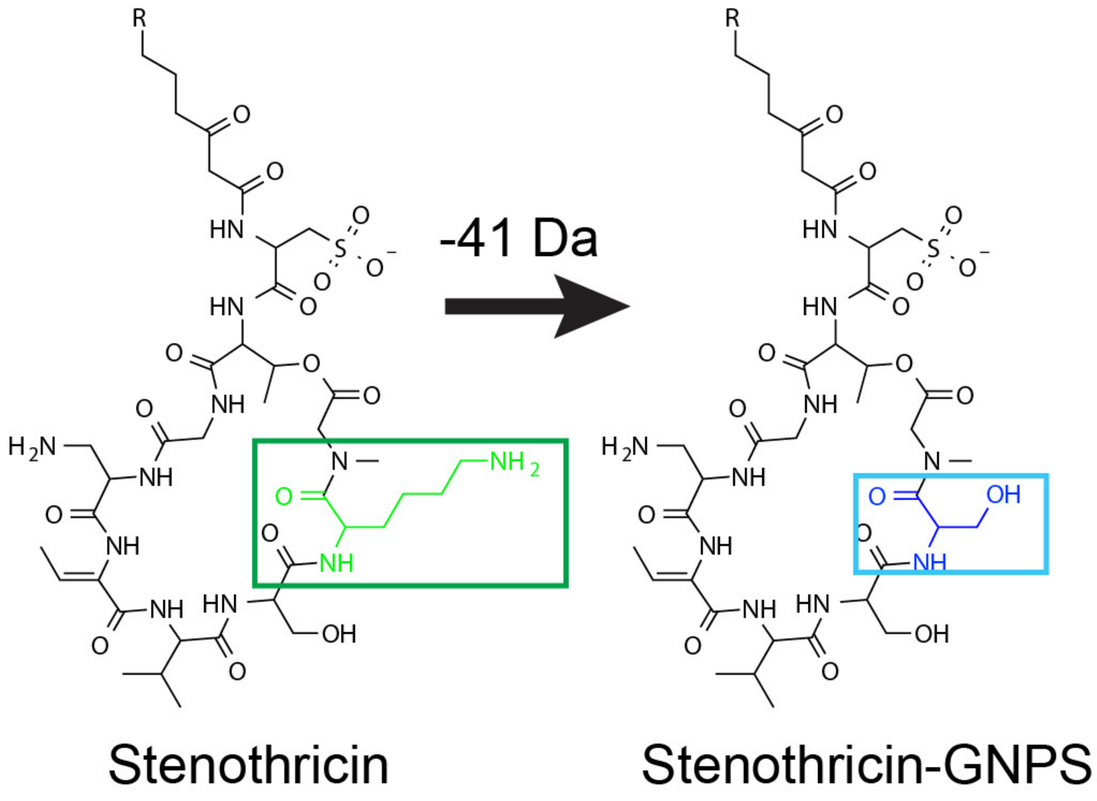

c
Surfactin C13
Surfactin
Surfactin C14
Surfactin C15
cluster index 1883
parent mass 1031.65
LibraryID Surfactin C13
number of spectra 6
DefaultGroups G1,G2,
precursor charge 1
Peptide
RT Info 66.69, σ = 34.94
ClusterSpectra Cluster Spectra
Align Spec
Score: 0.87
Node Labels
cluster index
parent mass
LibraryID
EvenOdd
Peptide
Edge Labels
Cosine
DeltaMZ
None
Edge Highlight
Edge MZ Delta
Edge Score Mi
Node MS2 Peaks Highlight
Cluster MZ Hig
Node Size
Default
Node Color
Default
G1
G2
G3
G4
G5
G6
Draw Pies
Reset
Click and drag in the plot to zoom X: Y: Zoom Out Print Enable tooltip
cluster index 1884
parent mass 1031.21
LibraryID Surfactin C13
number of spectra 14
DefaultGroups G1,G2,
precursor charge 1
Peptide
RT Info 31.01, σ = 17.56
ClusterSpectra Cluster Spectra
View Lib Spec
View Match
cluster index 1909
parent mass 1047.34
LibraryID Surfactin C13
number of spectra 18
DefaultGroups G1,G2,
precursor charge 1
Peptide
RT Info 127.92, σ = 98.57
ClusterSpectra Cluster Spectra
View Lib Spec
View Match
m/z:
802.47
intensity:
1326.01


a
100%
75%
50%
25%
0%
Rate
Public GNPS-MassIVE Datasets
Other
Exploratory Networks
Unidentified Networks
Identified Networks
Putative Analog Molecules
Identified Molecules
b
Fold Increase
25
23
21
19
17
15
13
11
9
7
5
3
1
MS/MS Spectra Matched
in Public Datasets
Unique Compounds
Matched in Public Datasets
Public Data Match Rate
538,440 Spectra
22x increase
1710 Unique Compounds
10x increase
1.94% Public Data ID Rate
3x increase
Apr-14
May-14
Jun-14
Jul-14
Aug-14
Sep-14
Oct-14
Nov-14
Dec-14
Jan-15
Feb-15
Mar-15

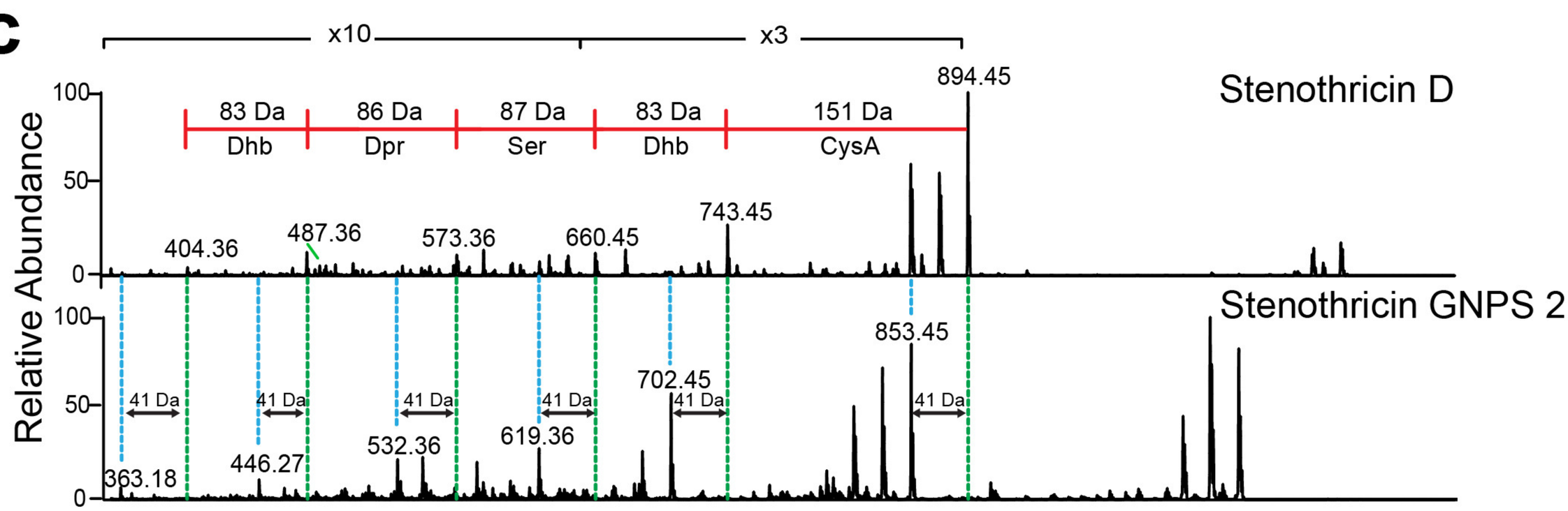
a


Stenothricin
- 41 Da Sub-Network
Streptomyces Roseosporus NRRL 15998
Streptomyces sp. DSM5940
Both

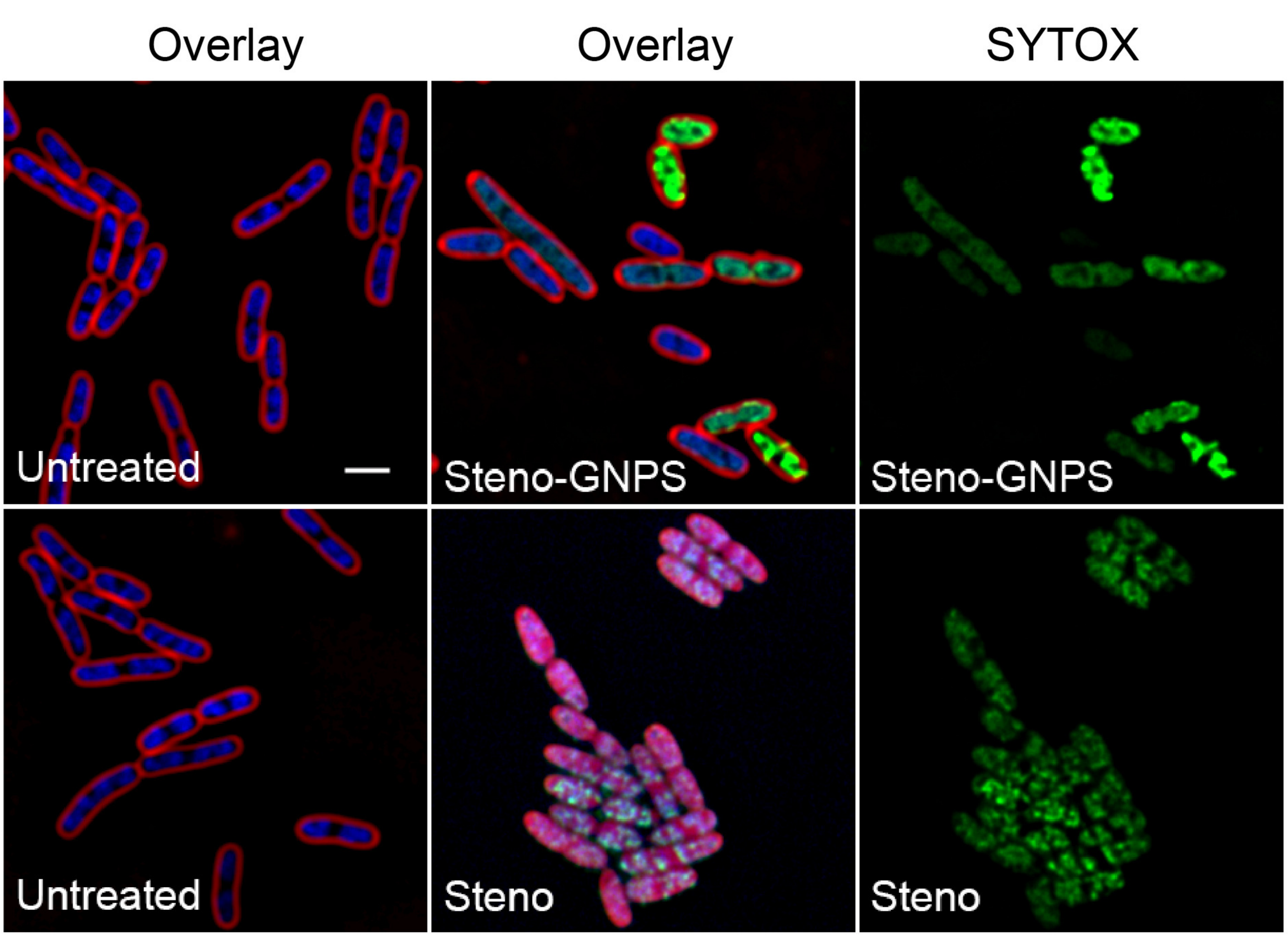
b


-41 Da
Stenothricin
Stenothricin-GNPS

c


x10
x3
Stenothricin D
Stenothricin GNPS 2
Relative Abundance
83 Da
Dhb
86 Da
Dpr
87 Da
Ser
83 Da
Dhb
151 Da
CysA
894.45
743.45
487.36
404.36
573.36
660.45
853.45
702.45
619.36
532.36
446.27
363.18
41 Da

d


Overlay
Overlay
SYTOX
Untreated
Steno-GNPS
Steno-GNPS
Untreated
Steno
Steno