

## Supplementary 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<sup>1-3</sup>. The matching of peaks between two spectra is formulated as a maximum bipartite matching problem<sup>4</sup> 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 8**. Source code can be found at the CCMS [github page](#) and also included in this manuscript as **Supplementary Source Code**.

### 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  $\delta$  (as determined by their precursor mass difference  $\delta$ ) +/- 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<sup>5</sup> 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<sup>4</sup>. 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. Source code can be found at the CCMS [github page](#) and also included in this manuscript as **Supplementary Source Code**.

## **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  $\mu$ L of each compound into 89  $\mu$ L of methanol into 96 well plates with 11 distinct compounds in each well. They were further diluted 100-fold for a final 1  $\mu$ M 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  $\mu$ M 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  $\mu$ L/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  $\mu$ L injection volume. A Phenomenex Kinetex 2.6  $\mu$ m C18 column (2.1 mm  $\times$  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 was 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<sup>6</sup>. 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<sup>7</sup>. 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 11**.

### **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  $\mu$ m, 100  $\text{\AA}$ , 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, Synchronis Phenyl HPLC, 5  $\mu$ m, 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  $\mu$ g stenothricin-GNPS 2 was dissolved in 30  $\mu$ L of CD<sub>3</sub>OD for NMR acquisition. <sup>1</sup>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 overlaid with the NMR spectrum from stenothricin D and analyzed using the MestReNova software<sup>7</sup>.

### **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 the accession numbers KX356656, KX356657, and KX356658.

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

To identify the Stenothricin-GNPS gene cluster, the *Streptomyces* sp. DSM5940 genome was annotated using Artemis<sup>8,9</sup>. 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)<sup>10</sup>. NRPSpredictor2 further assessed the adenylation domains of each NRPS gene cluster<sup>11,12</sup>. 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*<sup>7</sup>. The gene cluster with highest A-domain similarity was putatively identified as the stenothricin-GNPS

gene cluster. Full sequence alignment of both the stenothricin-GNPS and stenothricin using ClustalW2 confirmed high sequence identity and similarity<sup>13</sup>.

### 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<sup>13</sup>.

### Fluorescence Microscopy

A pre-culture of *E. coli* IptD cells (NR698) was grown to saturation, then diluted 1:100 into 20 mL LB. Flasks were incubated at 30°C until an OD<sub>600</sub> 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 previously described<sup>14</sup>. 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.

<b>Library Name</b>	<b>#Spectra</b>	<b>#Compounds</b>	<b>Data Generators</b>
GNPS-Community	2,224	1,325	Community
GNPS-Collections (FDA Approved Library from Selleck Chem Pt 1)	2,389	297	Sirenas MD
GNPS-Collections (FDA Approved Library from Selleck Chem Pt 2)	656	535	Dorrestein Lab
GNPS-Collections (NIH Clinical Collection 1)	377	329	Dorrestein Lab
GNPS-Collections (NIH Clinical Collection 2)	195	164	Dorrestein Lab
GNPS-Collections (Natural Products in NIH Small Molecule Repository)	1,268	1,255	Dorrestein Lab
GNPS-Collections (Pharmacologically Active Compounds in the NIH Small Mol Rep)	1,460	1,398	Dorrestein Lab
GNPS-Collections (Prestwick Phytochemical Library)	143	140	Dorrestein Lab
GNPS-Collections(Faulkner Legacy Library)	127	125	Sirenas MD
MassBank ESI MS/MS Library (Feb 2016)	24,545	5,992	Various
ReSpect MS/MS Library (July 2014)	7,112	1,986	Various
HMDB (Feb 2015)	2,235	747	Various
NIST 2014 ESI MS/MS Library	193,119	8,351	NIST
<b>Total</b>	<b>235,850</b>	<b>22,644</b>	

**Supplementary Table 1 - GNPS Libraries Summary.** Current spectral libraries available at GNPS, including compound libraries run exclusively for GNPS, community contributed spectra, and 3<sup>rd</sup> party libraries. Note: Original numbers created for the manuscript are inconsistent with this table as updated numbers were generated 3/21/2016.

<b>Library Name</b>	<b>Dataset Link</b>
GNPS-Collections (FDA Approved Library from Selleck Chem Pt 2)	<a href="#">MSV000078567</a>
GNPS-Collections (NIH Clinical Collection 1)	<a href="#">MSV000078567</a>
GNPS-Collections (NIH Clinical Collection 2)	<a href="#">MSV000078567</a>
GNPS-Collections (Natural Products in the NIH Small Molecule Repository)	<a href="#">MSV000078708</a>
GNPS-Collections (Pharmacologically Active Compounds in the NIH Small Mol Rep)	<a href="#">MSV000078710</a>
GNPS-Collections (Prestwick Phytochemical Library)	<a href="#">MSV000078711</a>

**Supplementary Table 2 - GNPS-Collections Raw Dataset links.** Links to locations to download mzXML MS/MS data that was used to populate major portions of the GNPS-Collections library.



Quality Level	Description	Spectra Count
Gold	Synthetic or purified, Complete structural characterization with NMR, crystallography or other standard methods as defined in the publication guidelines for Journal of Natural Products. Requires administrative approval.	4308
Silver	Isolated or lysate/crude, Published data showing presence of molecule in the sample. Requires administrative approval. Recommended attendance of GNPS curator workshop.	446
Bronze	Any other putative, complete, or partial annotation	4085

**Supplementary Table 3 – Library Quality Levels Description** The general public is granted access to add to the Bronze quality levels. However, special permissions are required to have access to add to Silver and Gold level MS/MS spectra. The oversight of these permissions falls under the administrators of GNPS, and currently is managed by Pieter Dorrestein.

Number of Annotation Revisions	Number of Library Spectra
1	351
2	46
3	19
4	7
5	4
6	0
7	1
8	1

**Supplementary Table 4 – Number of Annotation Revisions** – This table summarizes the total number of spectra with a given number of annotation revisions. There are a total of 563 annotation revisions for 429 GNPS library spectra.

**Supplementary Table 5 - Library Analogs Tables. See separate xlsx file.**

Rating	Description
1 Star	Incorrect Identification
2 Star	Not enough information, fragmentation is not sufficient to tell
3 Star	Putative Analog Identification, possibly not exactly correct isomer
4 Star	Correct Identification

**Supplementary Table 6 - Rating Description Table:** Description of each of the ratings for matches made in continuous identification.

Parameter	Value
MS/MS Fragment Ion Tolerance	0.5
Precursor <i>m/z</i> tolerance	2.0
Remove Peaks around Precursor Peak	1
Filter for top 6 most intense peaks in every +/- 50Da window in MS/MS spectrum	1
Minimum MS/MS Peak Intensity	0.0
Perform Spectrum Filtering Operations on the Library	1
Perform Analog Library Search	0
Minimum matching peaks in library search	6
Minimum library search match score	0.7

**Supplementary Table 7 - Library Validation Parameters:** Spectral library search parameters that were used to search the constructed spectral libraries against the NIH Natural Product and NIH Small Molecule Pharmacologically Active Libraries.

Parameter	Value
MS/MS Fragment Ion Tolerance	0.5
Precursor <i>m/z</i> tolerance	2.0
Filter for top 6 most intense peaks in every +/- 50Da window in MS/MS spectrum	1
Remove Peaks around Precursor Peak	1
Minimum MS/MS Peak Intensity	50.0
Run MSCluster	on
Minimum Consensus Cluster Size	2
Minimum Matched Peaks in Network Edge	6
Minimum MS/MS cosine score in Network Edge	0.65
Number of Neighbors to Retain in Network	10
Maximum Connected Component Size	0 (infinite)

**Supplementary Table 8 - Exploratory Networks Parameters:** Parameters for Molecular Networking for exploratory networks creation.

Parameter	Value
MS/MS Fragment Ion Tolerance	0.5
Precursor <i>m/z</i> tolerance	2.0
Filter for top 6 most intense peaks in every +/- 50Da window in MS/MS spectrum	1
Remove Peaks around Precursor Peak	1
Minimum MS/MS Peak Intensity	50.0
Run MSCluster	on
Minimum Consensus Cluster Size	2
Minimum Matched Peaks in Network Edge	6
Minimum MS/MS cosine score in Network Edge	0.8
Number of Neighbors to Retain in Network	10
Maximum Connected Component Size	0 (Infinite)

**Supplementary Table 9 - Continuous Networking Parameters:** Parameters for Molecular Networking for continuous identification data.

Parameter	Value
MS/MS Fragment Ion Tolerance	0.5
Precursor <i>m/z</i> tolerance	1.5
Remove Peaks around Precursor Peak	1
Filter for top 6 most intense peaks in every +/- 50Da window in MS/MS spectrum	1
Minimum MS/MS Peak Intensity	50.0
Perform Spectrum Filtering Operations on the Library	1
Perform Analog Library Search	0
Minimum matching peaks in library search	6
Minimum library search match score	0.7

**Supplementary Table 10 - Continuous Dereplication Parameters:** Parameters for spectral library search that occurs periodic for continuous identification.



Parameter	Value
MS/MS Fragment Ion Tolerance	0.5
Precursor m/z tolerance	2.0
Filter for top 6 most intense peaks in every +/- 50Da window in MS/MS spectrum	0
Remove Peaks around Precursor Peak	0
Minimum MS/MS Peak Intensity	0
Run MSCluster	On
Minimum Consensus Cluster Size	1
Minimum Matched Peaks in Network Edge	6
Minimum MS/MS cosine score in Network Edge	0.65
Number of Neighbors to Retain in Network	10
Maximum Connected Component Size	100
Perform Spectrum Filtering Operations on the Library	0
Perform Analog Library Search	0
Minimum Matched Peaks in Library Search	3
Minimum library search match score	0.7

**Supplementary Table 11 - Molecular Networking parameters for *S. roseosporus* and *Streptomyces sp.* DSM5940 MS/MS data.**

Compound Name	<i>m/z</i>	Characterization Notes
<b>Stenothricin-GNPS 1</b>	1077	Putative Structure
<b>Stenothricin-GNPS 2</b>	1091	NMR Confirmed
<b>Stenothricin-GNPS 3</b>	1105	Putative Structure
<b>Stenothricin-GNPS 4</b>	1119	Putative Structure
<b>Stenothricin-GNPS 5</b>	1133	Putative Structure
<b>Stenothricin B</b>	1118	Putative Structure
<b>Stenothricin D</b>	1132	NMR Confirmed
<b>Stenothricin E</b>	1146	Putative Structure
<b>Stenothricin G</b>	1160	Putative Structure
<b>Stenothricin H</b>	1174	Putative Structure

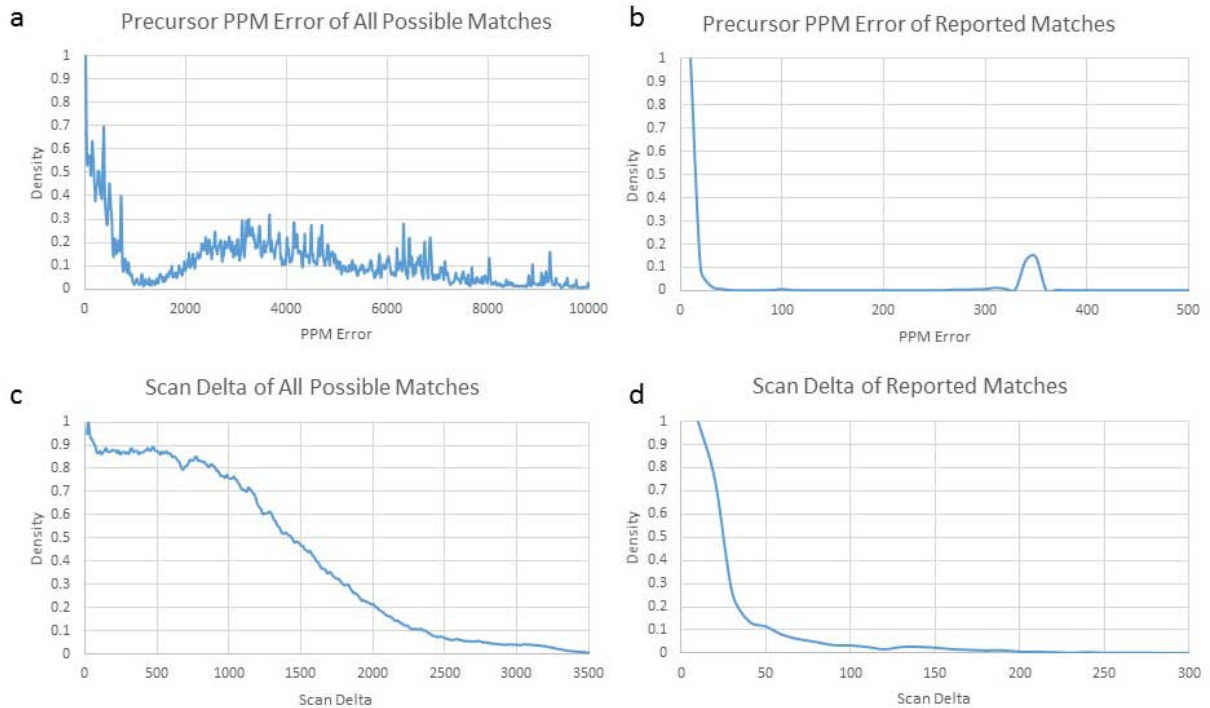
**Supplementary Table 12 - Stenothricin and Stenothricin-GNPS analogs.**

Residue	Position	$\delta_C$ , type	$\delta_H$	COSY	TOCSY	HMBC
N-Me-Gly-9	1	168.96, C	–			
	2a	52.98, CH <sub>2</sub>	4.80	2b		1, 3'
	2b	52.98, CH <sub>2</sub>	3.93,	2a		1, 3'
	3	36.77, CH <sub>3</sub>	3.18,			4
	3' rotomer	35.92, CH <sub>3</sub>	2.96,			2, 4
Ser-8	4	172.64, C	–			
	5	51.27, CH	5.17	6a/b	6a/b	4, 6a/b
	5' rotomer	53.17, CH	5.05	6'	6'	4, 6'
	6a	63.10, CH <sub>2</sub>	3.70,	5	5	4,5
	6b	63.24, CH <sub>2</sub>	3.81	5	5	4
Ser-7	6' rotomer	63.62, CH <sub>2</sub>	3.75,	5'	5'	
	7	172.62				
Ser-7	8	56.28, CH	4.84			
	9	64.02, CH <sub>2</sub>	3.56	8	8	7
	10	173.19, C	–			
Val-6	11	60.00, CH	4.38	12, 13/14	12, 13/14	10, 12, 13/14, 15
	11' rotomer	60.26, CH	4.34	12', 13/14	12', 13/14	10, 12', 13/14, 15
	12	32.99, CH	2.11	13/14	11, 13/14	11, 13/14
	12' rotomer	32.63, CH	2.16	11', 13/14	11', 13/14	11', 13/14
	13	19.43, CH <sub>3</sub>	1.01	12, 12'	11, 11', 12, 12'	11, 11', 12, 12', 14
	14	19.43, CH <sub>3</sub>	1.01	12, 12'	11, 11', 12, 12'	11, 11', 12, 12', 13
Dhb-5	15	165.43, C	–			
	16	130.24, C	–			
	17	134.82, CH	6.83	15	15	15, 16, 18
	18	13.24, CH <sub>3</sub>	1.83	17	17	15, 16, 17, 19
Dpr-4	19	170.23	–			
	20	52.32, CH	4.81	21a/b	21a/b	22
	21a	40.48, CH <sub>2</sub>	3.68	20, 21b	20, 21b	19, 20
	21b	40.48, CH <sub>2</sub>	3.48	20, 21a	20, 21a	19, 20
Ser-3	22	172.42, C	–			
	23	58.5, CH	4.41	24	24	22, 24
	23' rotomer	57.13, CH	4.54	24'	24'	22, 24'
	24	61.96, CH <sub>2</sub>	3.99	23	23	23
	24' rotomer	62.69, CH <sub>2</sub>	3.96	23'	23'	
Thr-2	25	173.92	–			
	26	52.40, CH	4.71		27', 28'	25
	27	72.76, CH	5.46	26, 28	26, 28	1
	27' rotomer	73.13, CH	5.20	26, 28'	26, 28'	28'
	28	23.01, CH <sub>3</sub>	0.93	27	26, 27	27
	28' rotomer	16.04, CH <sub>3</sub>	1.04	27'	26, 27'	
Cys-1	29	172.66, C	–			
	30	50.66, CH	5.58	31a/b	31a/b	
	30' rotomer	50.60, CH	5.74	31'a	31'a/b	
	31a	52.44, CH <sub>2</sub>	3.23	30, 31b	30, 31b	29, 30, 30'
	31'a rotomer	52.55, CH <sub>2</sub>	3.20	30', 31'b	30', 31'b	29, 30, 30'
	31b	52.44, CH <sub>2</sub>	3.72	30, 31a	30, 31a	29
	31'b rotomer	52.35, CH <sub>2</sub>	3.76	30', 31'a	30', 31'a	29

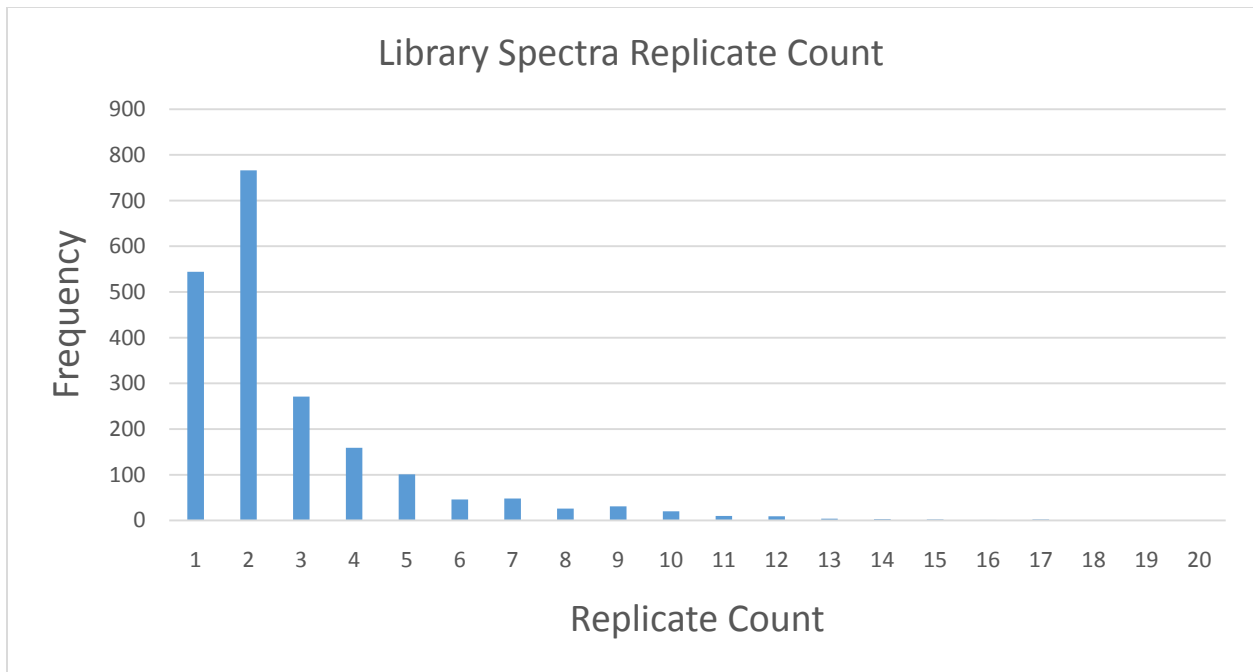
**Supplementary Table 13 - 2D NMR spectroscopic data for amino acids of stenothricin-GNPS 2 in CD<sub>3</sub>OD.** <sup>13</sup>C and <sup>1</sup>H chemical shifts were determined by HSQC and HMBC spectra.

Gene	Size [aa]	Predicted function	Protein homolog	Accession number
StenA	514	Multidrug resistance efflux pump	major facilitator superfamily permease [Streptomyces roseosporus NRRL 11379]	ZP_04707162.1
StenB	189	TetR-family transcriptional regulator	TetR family transcriptional regulator [Streptomyces roseosporus NRRL 15998]	ZP_06582833.1
StenC	721	helicase	helicase [Streptomyces globisporus C-1027] (93/89)	ZP_11377726.1
StenD	334	Oxidoreductase	oxidoreductase [Streptomyces roseosporus NRRL 15998]	ZP_06582835.1
StenE	475	argininosuccinate lyase (ArgH)	argininosuccinate lyase [Streptomyces roseosporus NRRL 15998]	ZP_04707166.1
StenF	398	argininosuccinate synthase (ArgG)	argininosuccinate synthase [Streptomyces roseosporus NRRL 11379]	ZP_04707167.1
StenG		Gap in Sequencing Data		
StenH	236	secreted protein/L,D-transpeptidase	secreted protein [Streptomyces roseosporus NRRL 15998]	ZP_06582839.1
StenI	178	arginine repressor (ArgR)	arginine repressor [Streptomyces roseosporus NRRL 11379]	ZP_04707170.1
StenJ	403	N2-acetyl-L-ornithine:2-oxoglutarate aminotransferase (ArgD)	acetylornithine aminotransferase [Streptomyces globisporus C-1027]	ZP_11377720
StenK	314	N-acetylglutamate kinase (ArgB)	acetylglutamate kinase [Streptomyces roseosporus NRRL 11379]	ZP_04707172
StenL	384	N2-acetyl-L-ornithine:L-glutamate N-acetyltransferase (ArgJ)	bifunctional ornithine acetyltransferase/N-acetylglutamate synthase protein [Streptomyces globisporus C-1027]	ZP_11377718
StenM	342	N-acetyl-gamma-glutamylphosphate reductase (ArgC)	N-acetyl-gamma-glutamyl-phosphate reductase [Streptomyces globisporus C-1027]	ZP_11377717.1
StenN		Gap in Sequencing Data		
StenO	351	ornithine cyclodeaminase/2,3-diaminopropionate biosynthesis protein (SbnB)	ornithine cyclodeaminase [Streptomyces roseosporus NRRL 15998]	ZP_06582846.1
StenP	1198	NRPS	non-ribosomal peptide synthetase [Streptomyces roseosporus NRRL 15998]	ZP_06582847.1
StenQ	271	type II thioesterase	conserved hypothetical protein [Streptomyces roseosporus NRRL 15998]	ZP_06582848.1
StenR	75	MbtH-like protein	hypothetical protein SrosN1_04335 [Streptomyces roseosporus NRRL 11379]	ZP_04707179.1
StenS	6082	NRPS	non-ribosomal peptide synthetase [Streptomyces roseosporus NRRL 15998]	ZP_06582851.1
StenT	3833	NRPS	non-ribosomal peptide synthetase [Streptomyces roseosporus NRRL 15998]	ZP_06582853.1
StenU	431	cysteate synthase	L-threonine synthase [Streptomyces roseosporus NRRL 11379]	ZP_04707187.1

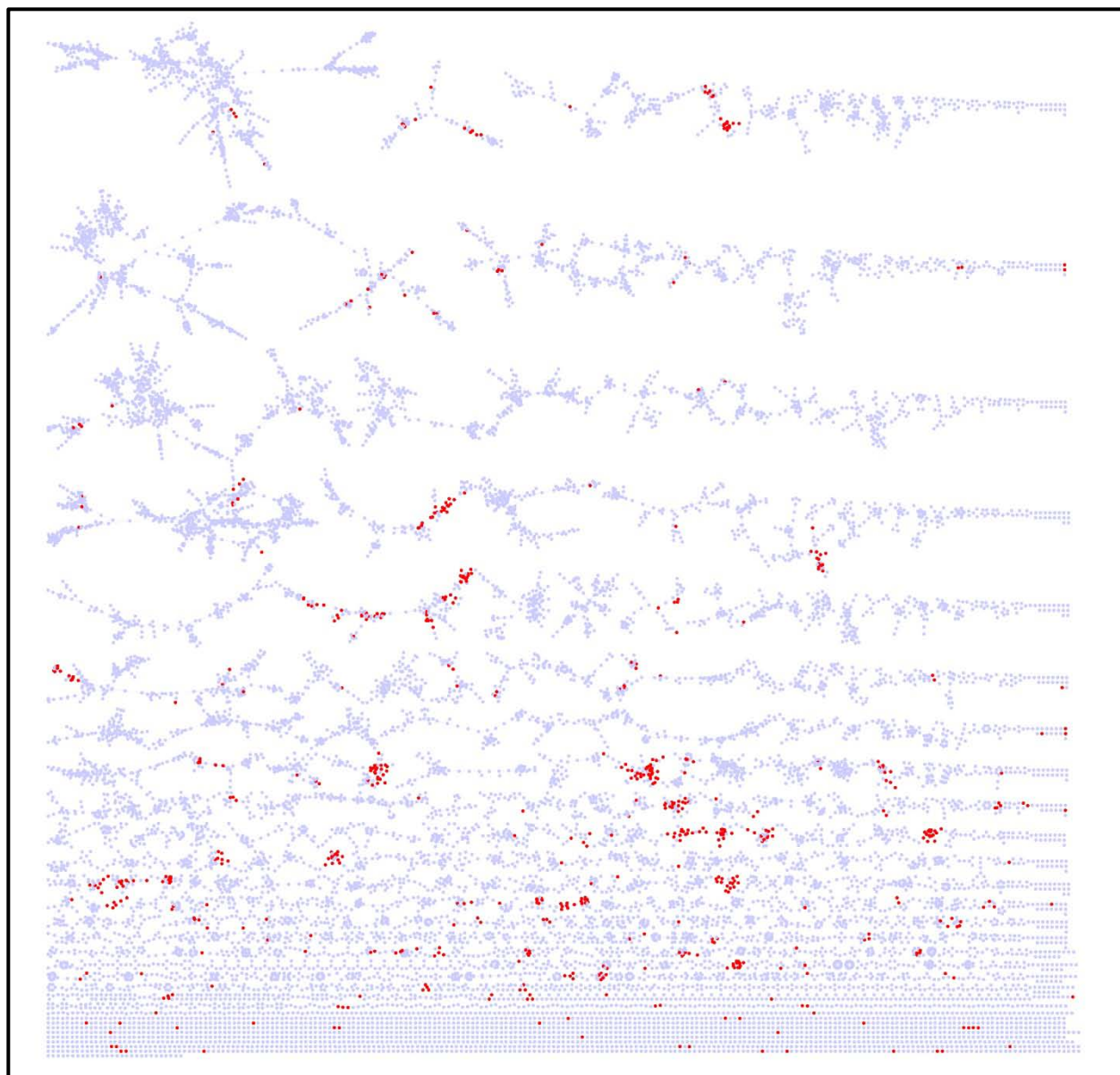
**Supplementary Table 14 - Annotation of genes in the stenothricin-GNPS biosynthetic cluster and predicted function.**



**Supplementary Figure 1 – Spectral Library Search Match Assessment** (a) The PPM error of all possible matches from the NIH Natural Product Library and NIH Pharmacologically Active Library to the query datasets MSV000078708 and MSV000078710. The spread of possible PPM errors for all possible matches illustrates the large possible space of erroneous matches as measured by PPM error. (b) The PPM errors of all reported matches. There is a slight bump around 350PPM as all LC-MS/MS runs from one of the plates was miscalibrated. The total number of non-identity matches that within 50 PPM error is 5,520 matches out of a 5,927 matches. (c) The scan deltas of all possible matches between the query spectra and library spectra. The large spread of possible scan number deltas illustrates large possible space of erroneous matches as measured by scan number deltas. (d) A histogram of scan number delta between query and library reported matches by library search. 5,789 of all 5,927 matches reported fell within 200 scans of the library spectrum.




**Supplementary Figure 2 – Replicate Count Histogram within Reported Matches:** A histogram of the number of replicates for each library spectrum. These do not include identity matches or incorrect matches.




**Supplementary Figure 3 –Chemical Space Visualization with Molecular Networking.** MS/MS data from a human skin metabolomics datasets (Bouslimani et al. 2015, Massive dataset MSV000078556) was analyzed with molecular networking at GNPS. Consensus MS/MS spectra matched to a GNPS library spectrum are highlighted as red nodes, unmatched in grey.

[GNPS] Updated Continuous Identification Results for MassIVE  
Datasets 07-03-2015



 ccms@ucsd.edu

Jul 3 



to me 

MassIVE Dataset MSV000078556 GNPS - Topobiographical molecular analysis of human skin  
Reporting 0.7712082% more IDs and 3 more IDs  
0.0% different IDs and 0 different IDs  
0.0% deleted IDs and 0 deleted IDs  
with 389 total

Results are available at:

<http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=706907b06199422bb979f7fe169006ec>

The dataset is available at:

[http://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=6b9dcff3899e4d5f89f0daf9489a3a5e&view=group\\_all\\_annotations](http://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=6b9dcff3899e4d5f89f0daf9489a3a5e&view=group_all_annotations)

---

MassIVE Dataset MSV000078566 GNPS-Pseudomonas\_Isolates\_HumanLung\_CF  
Reporting 2.857143% more IDs and 1 more IDs  
0.0% different IDs and 0 different IDs  
0.0% deleted IDs and 0 deleted IDs  
with 35 total

Results are available at:

<http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2ffe425d47d04e85b31b49404ed499cc>

The dataset is available at:

[http://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=ca854d9c399745379142fde622ef7ab1&view=group\\_all\\_annotations](http://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=ca854d9c399745379142fde622ef7ab1&view=group_all_annotations)

**Supplementary Figure 4 – Example Continuous Identification Digest Email** – Digest email reporting new identifications per dataset. For each dataset, the spectral match changes are reported as well as total match count, links to results, and links to each dataset.



### (a) Related Datasets

Filter	Dataset	Score	Description
<input type="checkbox"/> checked only			
<input type="checkbox"/> 1	MSV000079146	14	<a href="#">GNPS Papua New Guinea pH gradient ARMS</a>
<input type="checkbox"/> 2	MSV000078598	10	<a href="#">GNPS Southern Line Islands Coral Interactions Metabolome</a>
<input type="checkbox"/> 3	MSV000079104	8	<a href="#">GNPS YW CF Sputum Samples Exacerbation Longitudinal</a>
<input type="checkbox"/> 4	MSV000078922	8	<a href="#">GNPS UPLC May 2014 CF Sputum</a>
<input type="checkbox"/> 5	MSV000078565	8	<a href="#">GNPS-Lung_cysticFibrosis_tandemMS</a>
<input type="checkbox"/> 6	MSV000079105	8	<a href="#">GNPS CF sinusitis samples</a>
<input type="checkbox"/> 7	MSV000079091	7	<a href="#">GNPS coral-algae laboratory interaction experiments</a>
<input type="checkbox"/> 8	MSV000078992	7	<a href="#">GNPS CF sputum bacterial isolates</a>
<input type="checkbox"/> 9	MSV000078830	7	<a href="#">GNPS - P. aeruginosa 14 transposon library ethyl acetate/methanol extractions</a>
<input type="checkbox"/> 10	MSV000078812	6	<a href="#">GNPS Moorea Reef Laboratory Metabolite Interactions UPLC MS/MS</a>

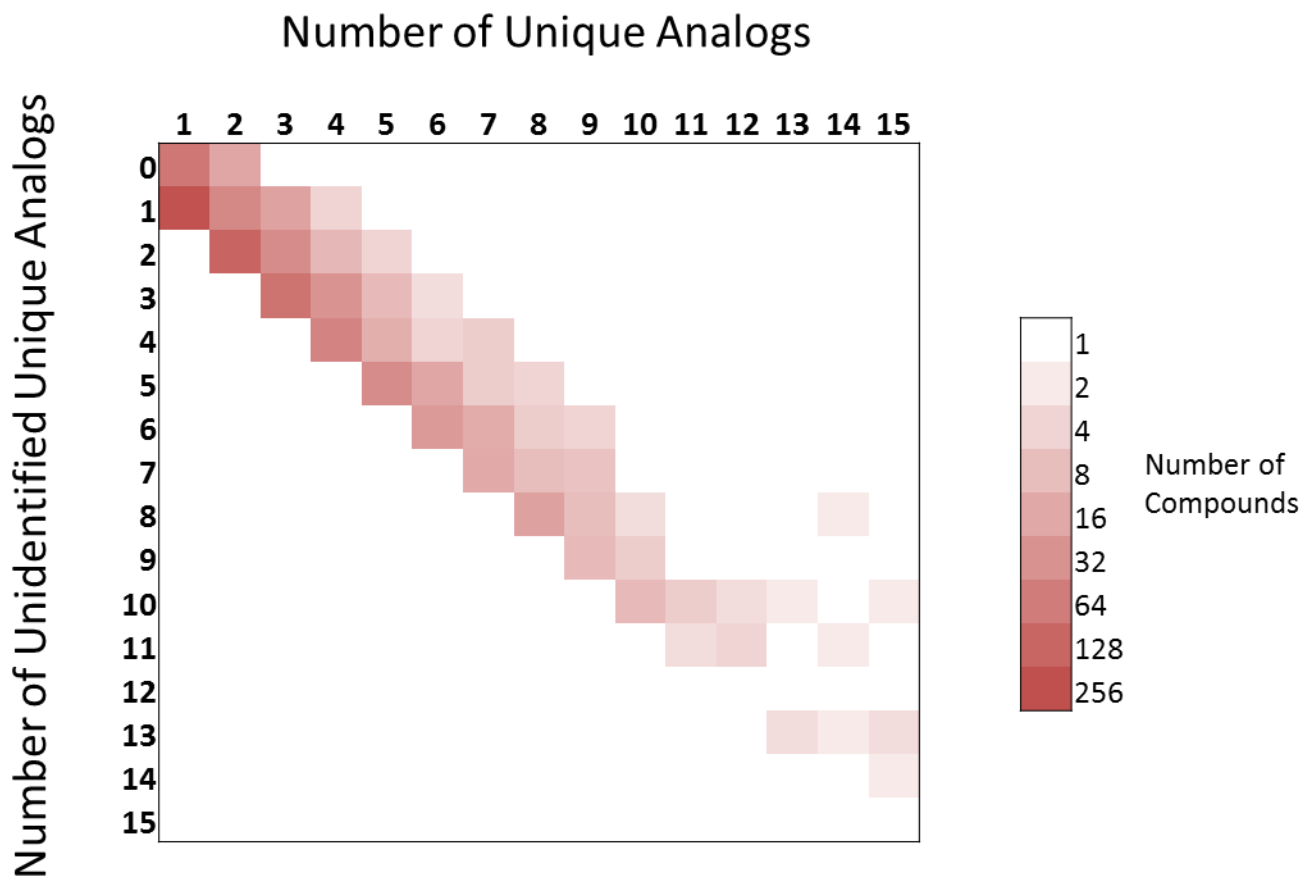
### (b)

MassIVE Dataset Information	
Title	GNPS - GNPS Paper Novel Stenothricin Analog Supporting Information.
Description	Mass spectrometry data, NMR, and sequencing data in support of the discovery of stenothricin-GNPS. Secondary metabolite production of Streptomyces Roseosporus was compared against Streptomyces DSM5940.
MassIVE Accession	MSV000079204
Principal Investigators	Pieter Dorrestein
Username	<a href="#">mwang87</a>
Contact Email	<a href="mailto:miv023@ucsd.edu">miv023@ucsd.edu</a>
Species	Streptomyces

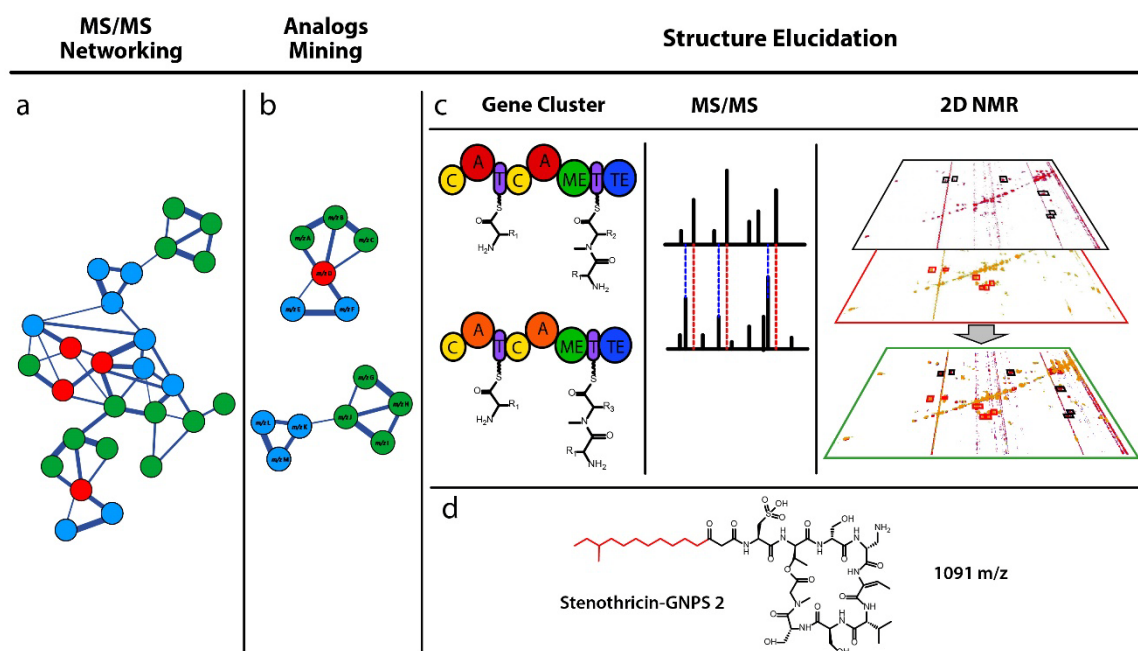
### (c)

Latest Library Spectrum Information	
Spectrum ID	CCMSLIB00000077202
Compound Name	C18_Sphingosine
PI	Dorrestein
Data Collector	Amina
CAS Number	<a href="#">123-78-4</a>
Original Submitter	<a href="#">aboulimani</a>
Original Submitter Email	<a href="mailto:aboulimani@ucsd.edu">aboulimani@ucsd.edu</a>
Most Recent Revisor	<a href="#">aboulimani</a>
Most Recent Revisor Email	<a href="mailto:aboulimani@ucsd.edu">aboulimani@ucsd.edu</a>
Library Quality	Silver Spectrum
Smiles	CCCCCCCCCCCC=CC(C(CO)N)O

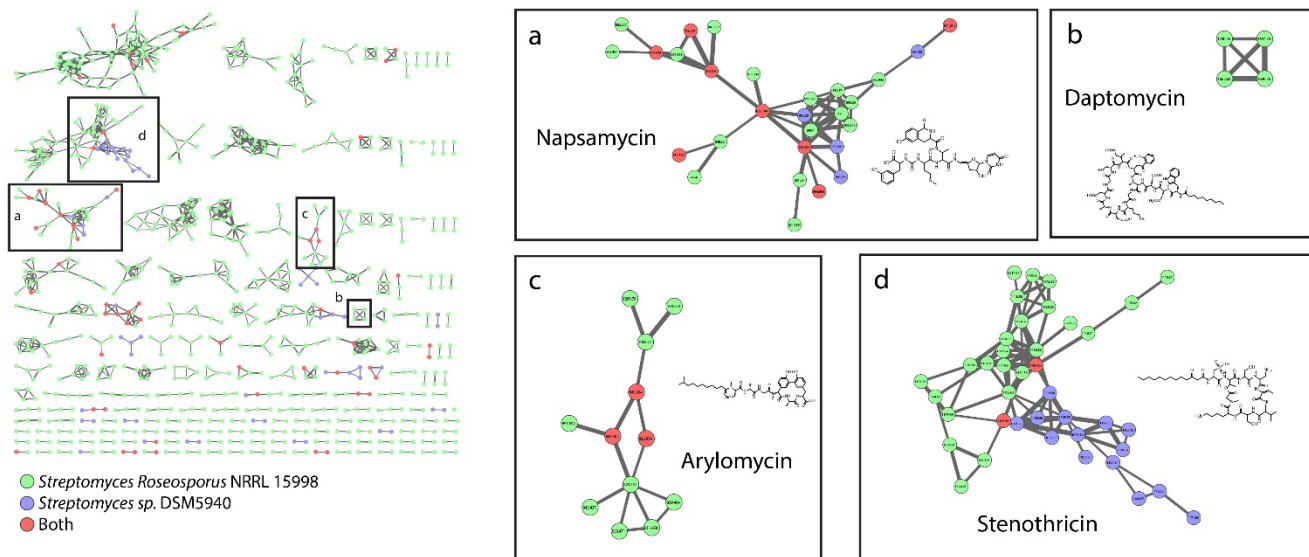
**Supplementary Figure 5 - GNPS Collaboration Initiation.** (a) For each public dataset, GNPS provides a list of related datasets. This allows users to find related datasets to their own in order to spark collaboration. In selecting a related dataset, users reach a dataset page (b). From here more detailed metadata including an email link are available to initiate contact. (c) Further, identifications made via continuous identification link to spectrum library pages and enables users to initiate contact with the annotators of these library spectra.



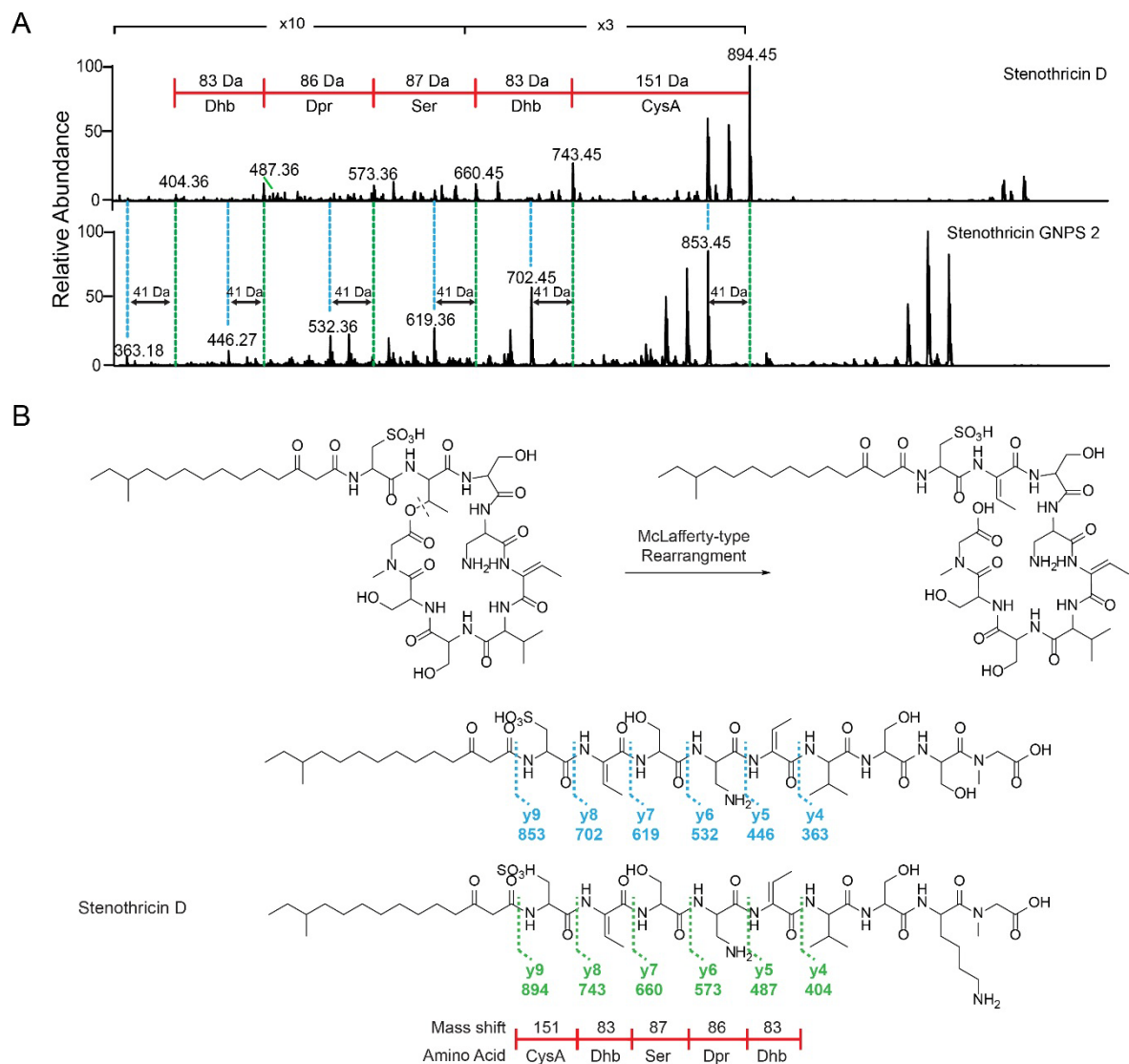
**Supplementary Figure 6 - Molecular Explorer Summary Histogram.** Molecular explorer highlights the number and locations of related MS/MS spectra of known compounds, indicating they are related (defined as analogs). Analogs are direct neighbors of matched compounds in molecular networks. For each library compound we tallied the number of unique mass analogs. Here we illustrate exactly how much diversity of known compounds is found in all public GNPS datasets and how much of that diversity is actually captured in the GNPS spectral libraries. For each library spectrum matched to public data all unique mass analogs were counted (X axis) and all unidentified analogs were counted (Y axis); the relationship between these is shown as a 2D heatmap. Since most of the density of the plot is near the diagonal, it indicates that the majority of the putative analogs of known compounds in the data remained unidentified.



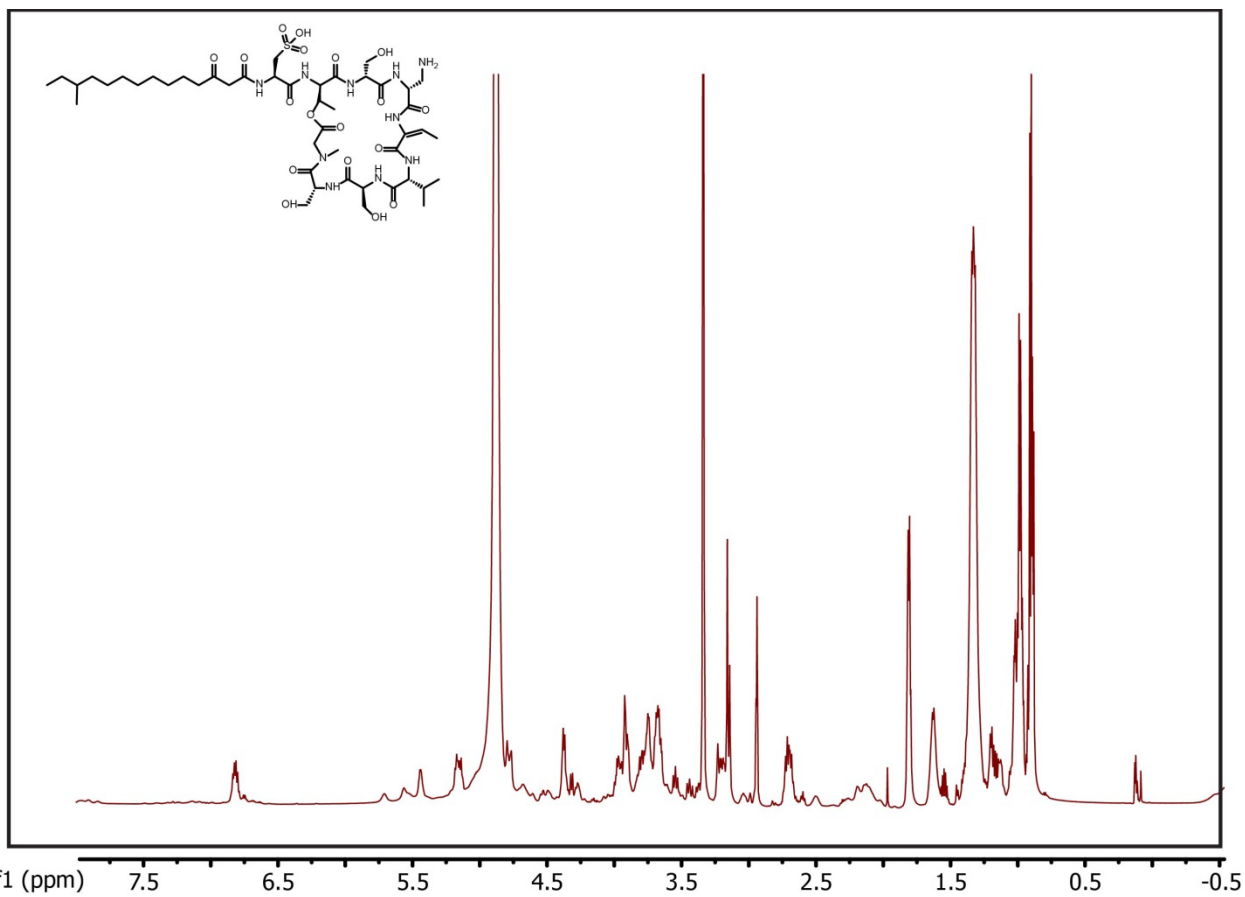
**Supplementary Figure 7 - GNPS based MS/MS networking guided workflow.** For identification of novel analogues from microbes without genome sequences, the steps are as follows: (a) Generation of molecular network from specialized metabolites produced by both sequenced and not sequenced organisms. (b) Dereplication of known molecules and detection of novel analogues. (c) Structure elucidation by differential analysis of MS/MS data, 2D NMR data and gene cluster annotation. (d) Characterization of stenothricin-GNPS.



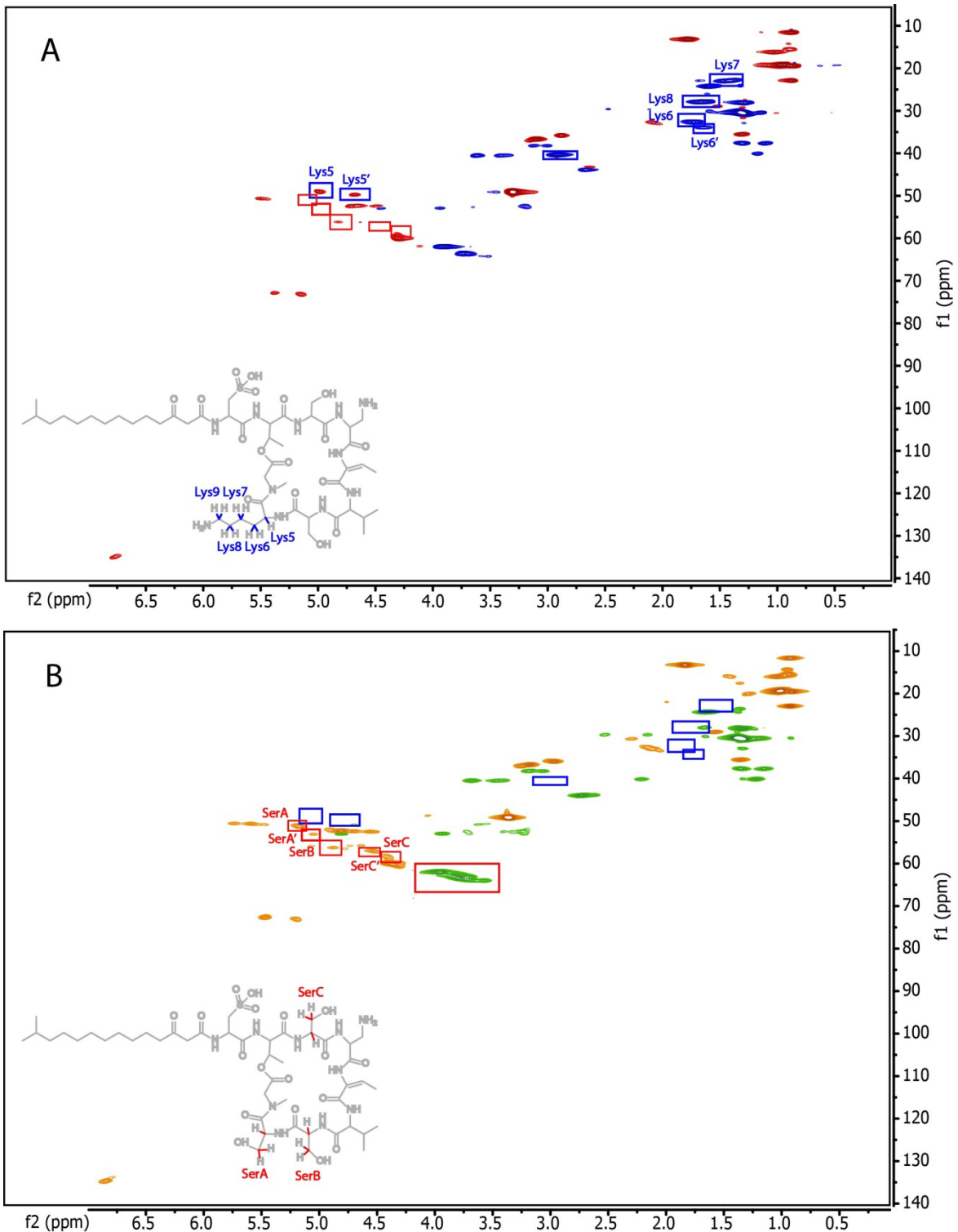
**Supplementary Figure 8 - Molecular network of *Streptomyces roseosporus* and *Streptomyces sp.* DSM5940.** Four molecular families were identified in the molecular network (a) Napsamycin, (b) Daptomycin, (c) Arylomycin, (d) Stenothricin.



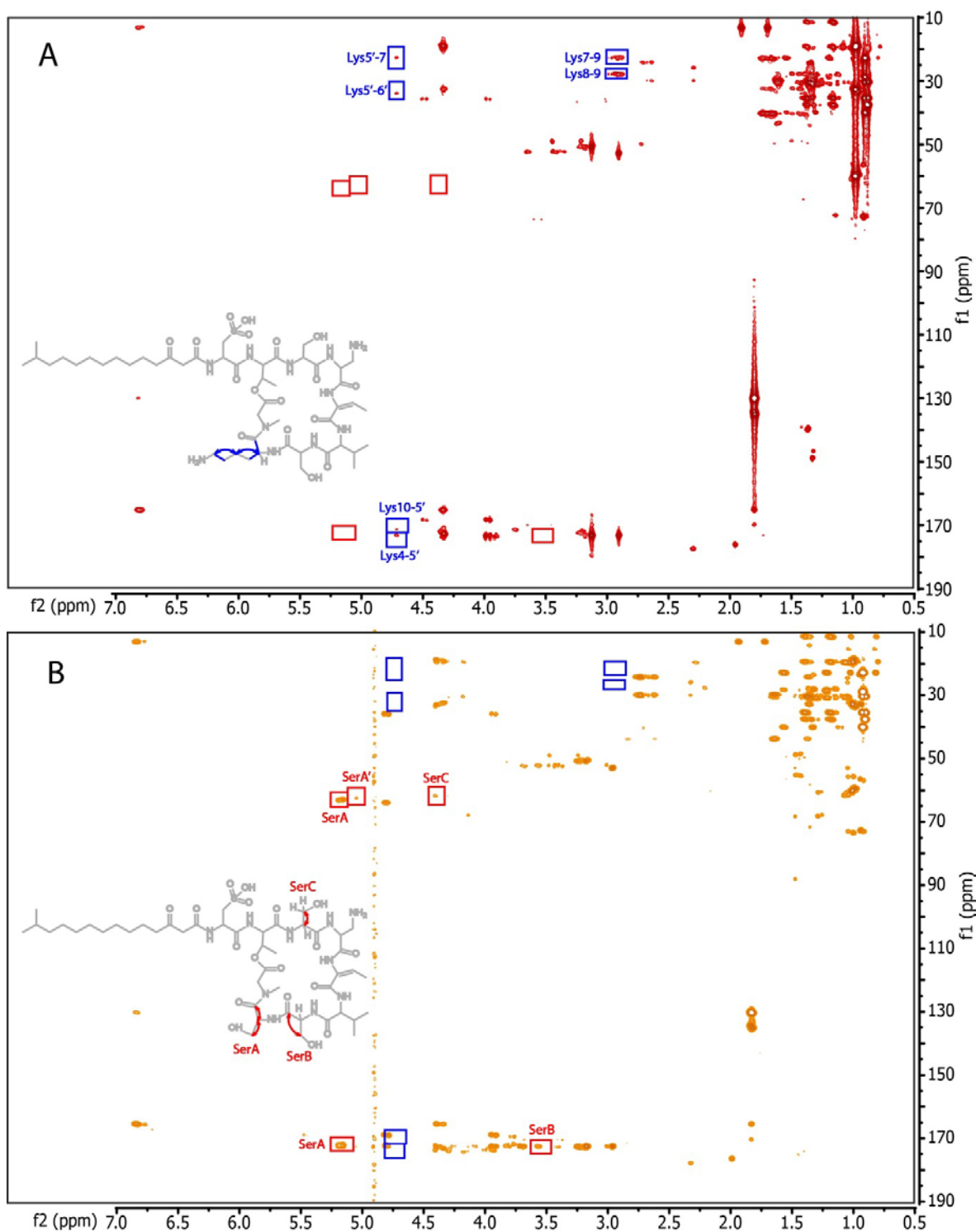
**Supplementary Figure 9 - Comparison of stenothricin D and stenothricin-GNPS 2.** (A) Alignment of MS/MS spectra of stenothricin D (*S. roseosporus*) and stenothricin-GNPS 2 (*Streptomyces. sp.* DSM5940) (B) Identification of a common peptide sequence tag CysA-Dhb-Ser-Dpr-Dhb in the stenothricin and stenothricin-GNPS core structure.



**Supplementary Figure 10 -  $^1\text{H}$  NMR spectrum of stenothricin-GNPS (600 MHz,  $\text{CD}_3\text{OD}$ ).**

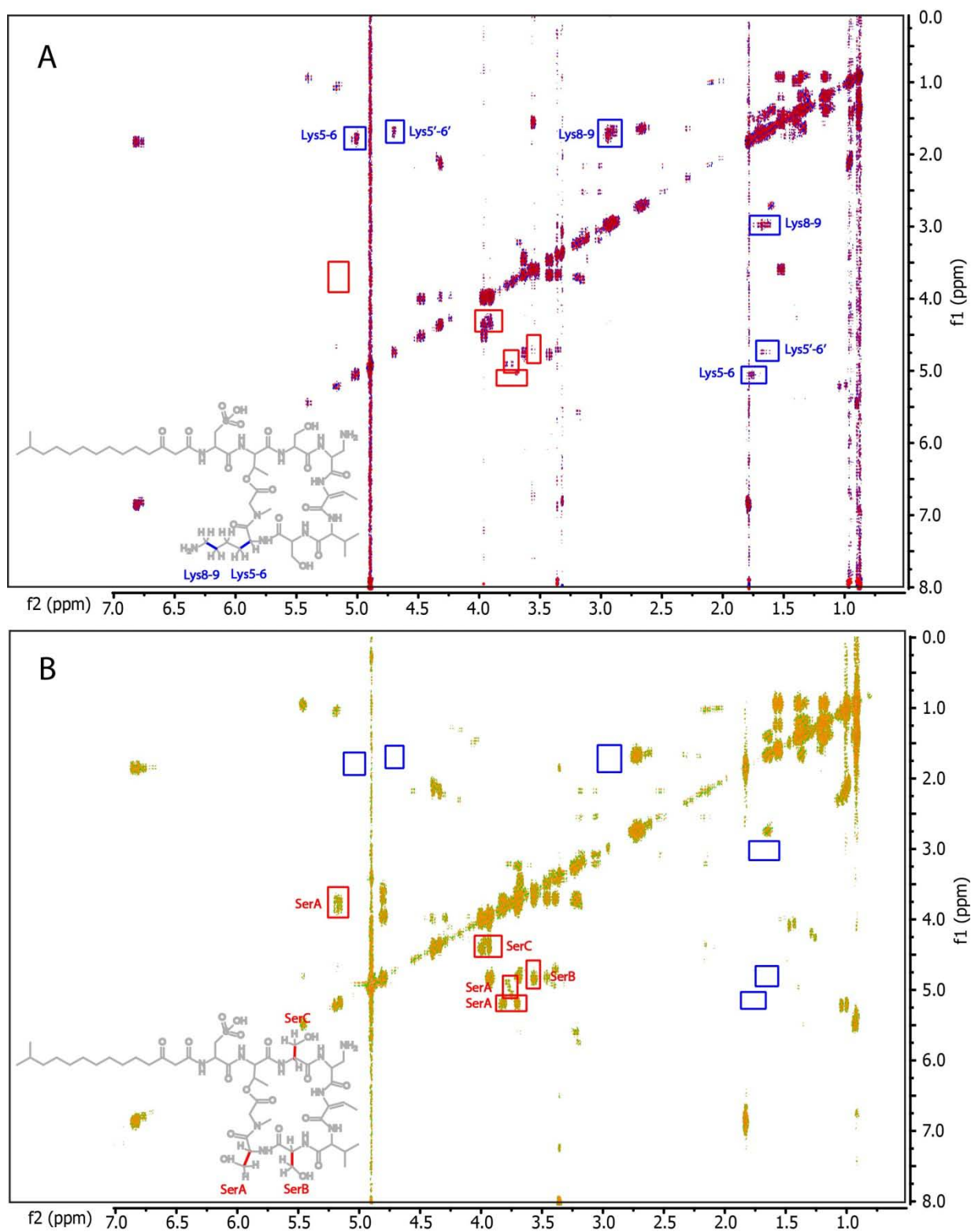


**Supplementary Figure 11 - HSQC based differential analysis of stenothricin D and stenothricin-GNPS 2 (600 MHz, CD<sub>3</sub>OD).** (A) Stenothricin D spectrum. (B) Stenothricin-GNPS 2 spectrum. Differences in signals are boxed, representing the disappearance of lysine (blue) and the existence of extra serine (red) in stenothricin-GNPS 2.

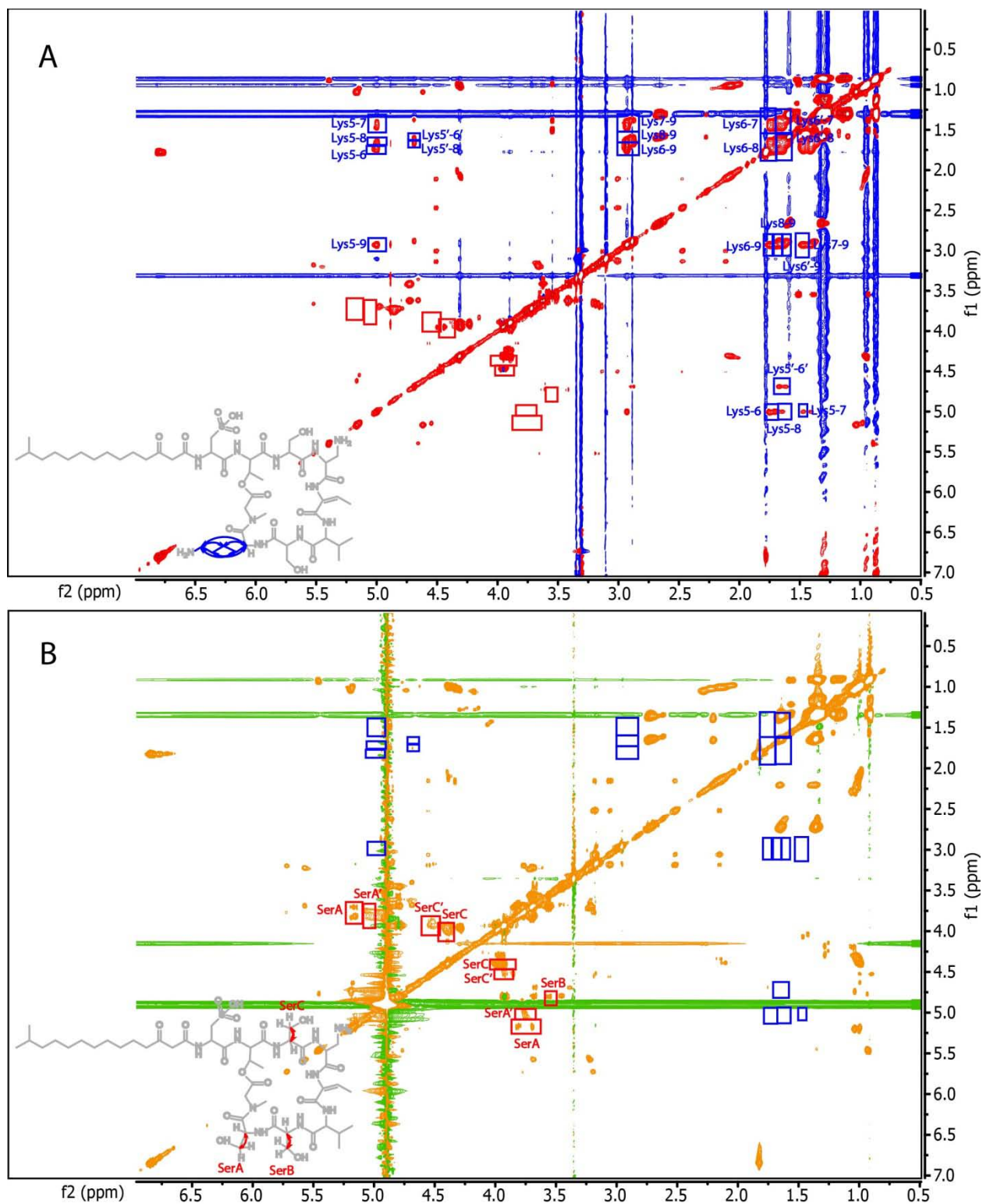


**Supplementary Figure 12 - HMBC based differential analysis of stenothricin D and stenothricin-GNPS 2 (600 MHz, CD<sub>3</sub>OD).** (A) Stenothricin D spectrum. (B) Stenothricin-GNPS 2 spectrum. Differences in signals are boxed, representing the disappearance of lysine 5-10 correlations (blue) and the existence of extra serine (red) in stenothricin-GNPS 2.

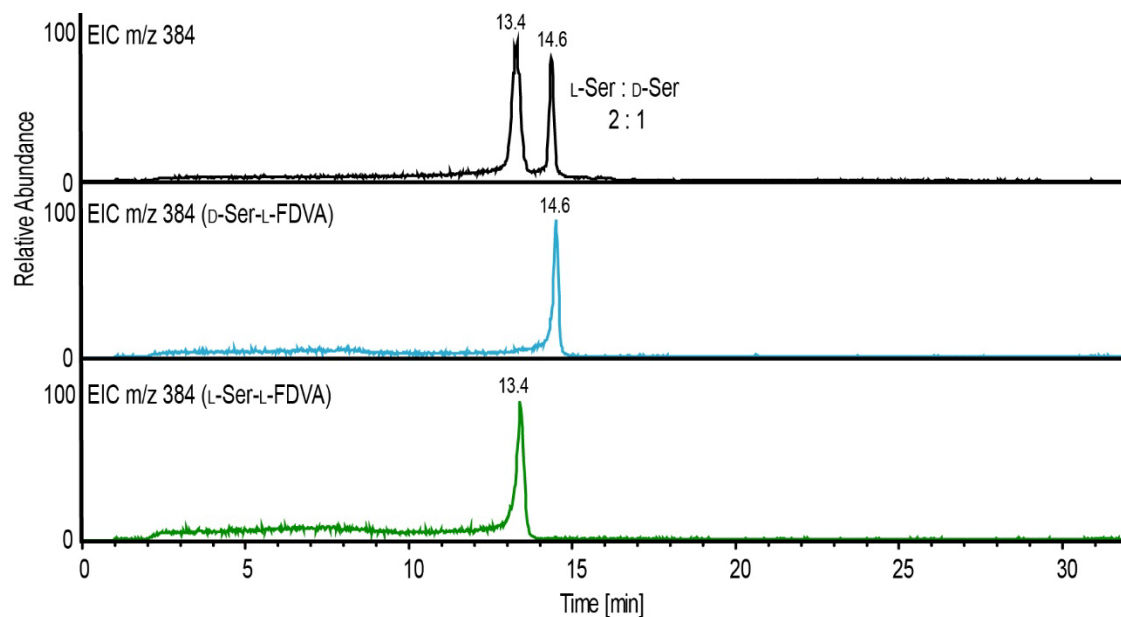




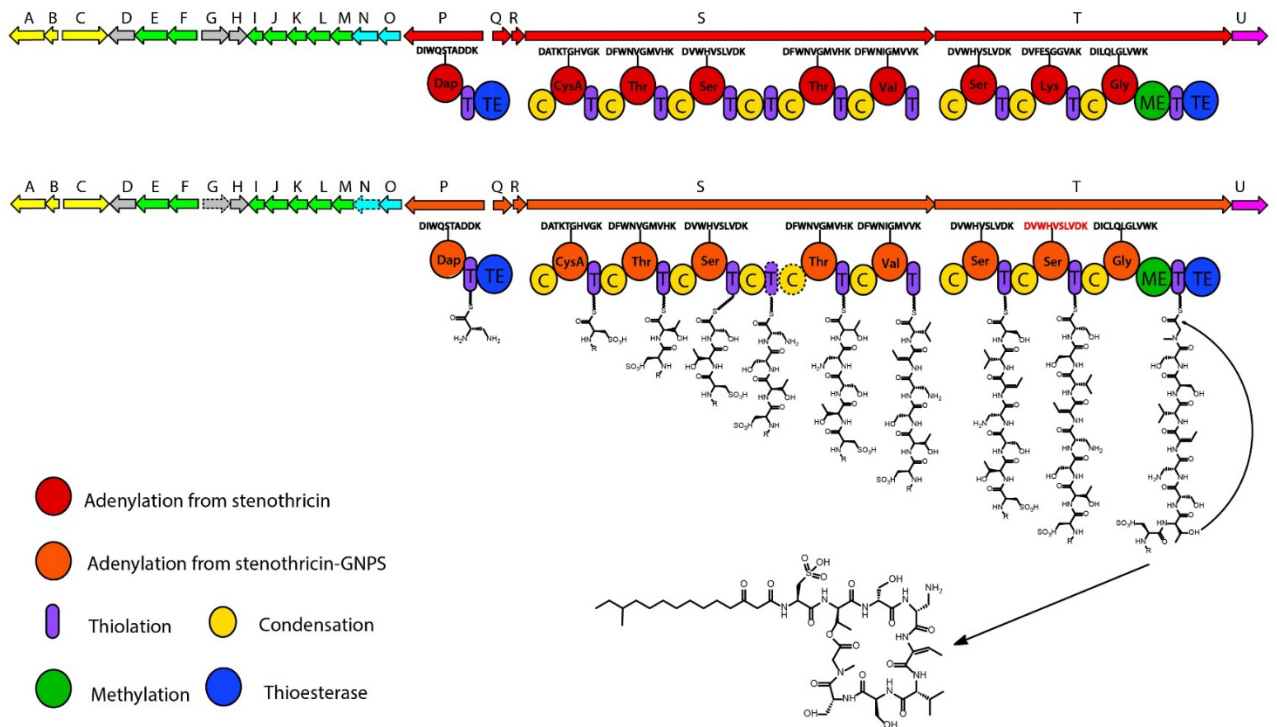
**Supplementary Figure 13 - dqfCOSY based differential analysis of stenothricin D and stenothricin-GNPS 2 (600 MHz, CD<sub>3</sub>OD).** (A) Stenothricin D spectrum. (B) Stenothricin-GNPS 2 spectrum. Differences in signals are boxed.



**Supplementary Figure 14 - TOCSY based differential analysis of stenothricin D and stenothricin-GNPS 2 (600 MHz, CD<sub>3</sub>OD). (A) Stenothricin D spectrum. (B) Stenothricin-GNPS 2 spectrum. Differences in signals are boxed.**

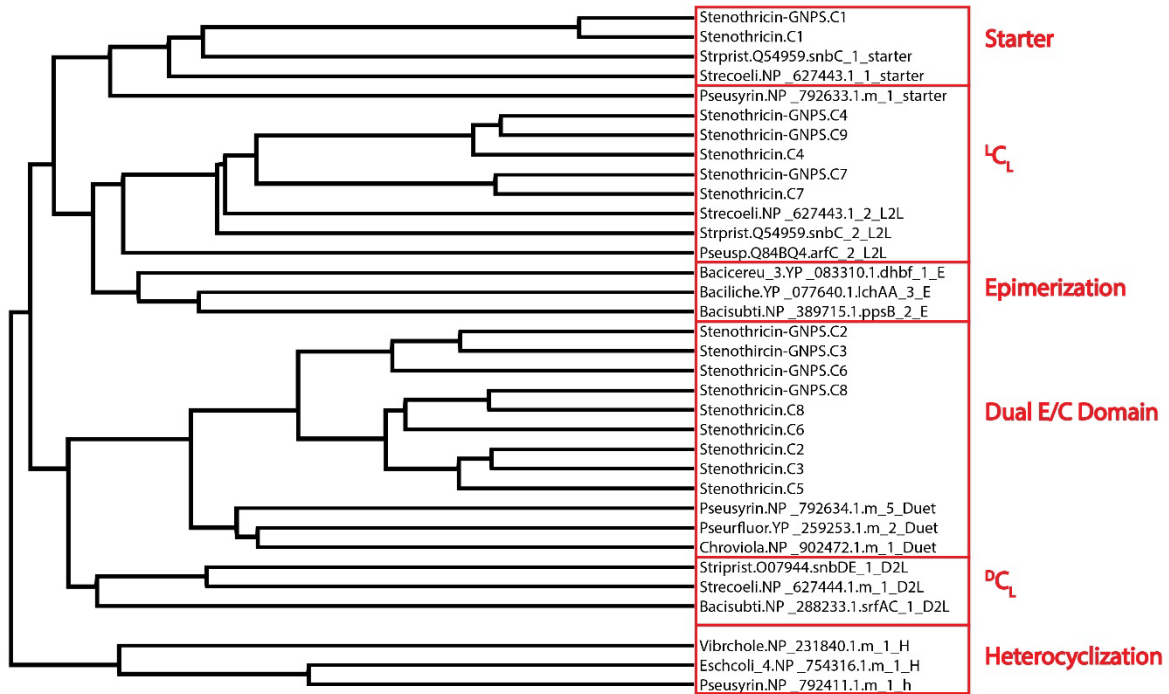


**Supplementary Figure 15 - Marfey's Analysis of stenothricin-GNPS 2.** Marfey's analysis suggests the presence of a third serine in stenothricin-GNPS 2, supporting 2D NMR data that the 41Da shift between the stenothricin molecular family and the stenothricin-GNPS subfamily is a lysine to serine substitution. Retention time alignment suggests that the third serine is in the L configuration.

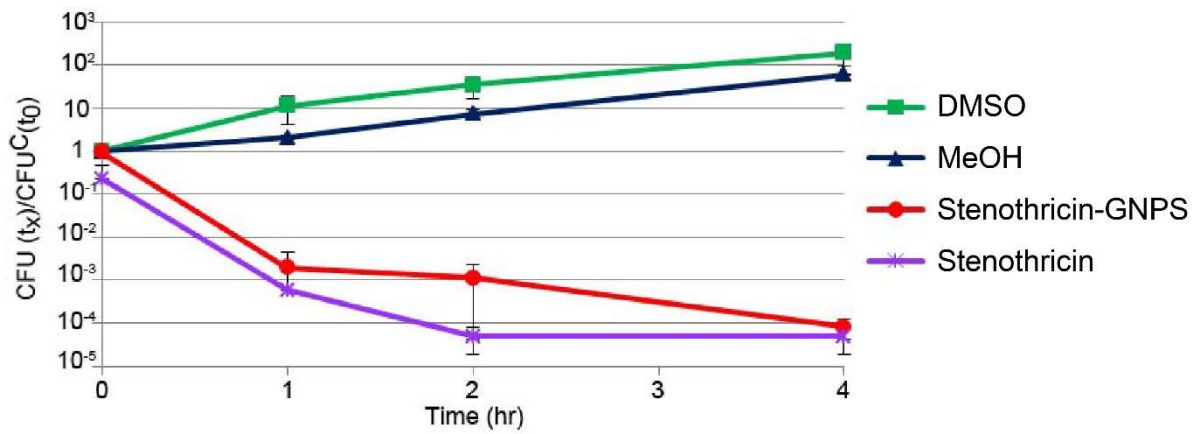


\* Dash lines represent gaps in genes

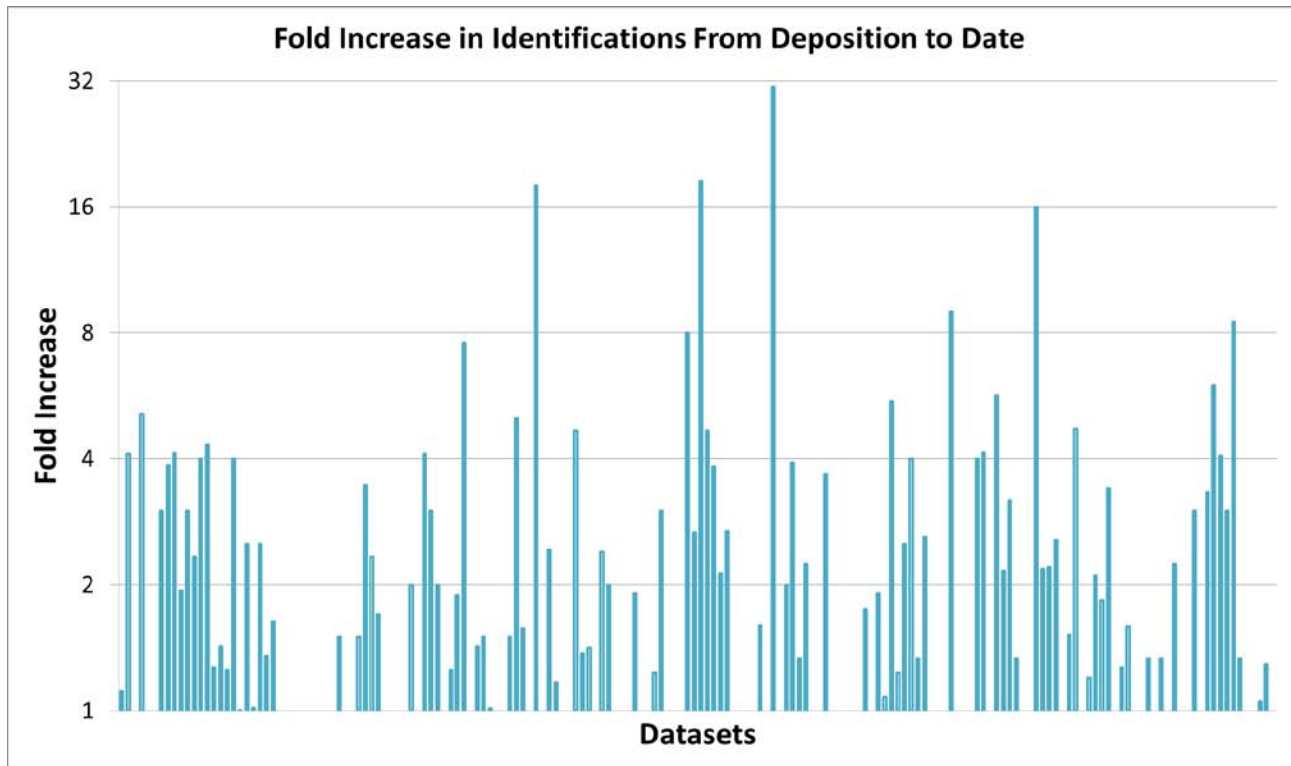
**Supplementary Figure 16 - Comparison of the stenothricin and stenothricin-GNPS candidate gene clusters.** Alignment of stenothricin and stenothricin-GNPS candidate biosynthetic gene clusters and proposed biosynthetic pathways suggest high sequence homology with the key lysine to serine substitution due to the specificity of the eighth A-domain



**Supplementary Figure 17 - Phylogenetic analysis of C-domains from stenothricin-GNPS NRPS gene cluster.** Amino acid configuration can be inferred from phylogenetic clustering of C-domains. The C-domains of the stenothricin-GNPS gene cluster clustered with the C-domains of the stenothricin gene cluster suggesting similar amino acid stereochemistry.



**Supplementary Figure 18 - Viability Over Time of *E. coli* cells treated with stenothricin-GNPS and stenothricin.** Stenothricin-GNPS treatment of *E. coli* *lptD* cells led to decreased viability at 40  $\mu\text{g}/\text{mL}$  similar to that of stenothricin D at 20  $\mu\text{g}/\text{mL}$



**Supplementary Figure 19 - Dataset fold increase in identifications by continuous identification.** The fold increase of identifications on a per dataset basis from the time of deposition to date. 59% of these datasets increased their identifications, averaging a 143% increase in identifications since deposition.



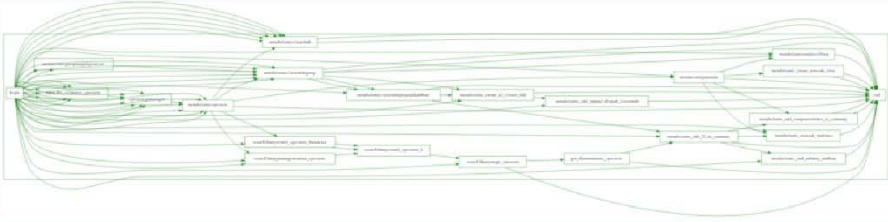
gnps.ucsd.edu/ProteoSAFe/status.jsp?task=82a65552624c442093e357a75707e585 **Shareable URL**

GNPS: Global Natural Products Social Molecular Networking

MassIVE Datasets | Documentation | Forum | Contact

Back to main page

**Job Status**

<b>Workflow</b>	METABOLOMICS-SNETS
<b>Status</b>	<p><b>DONE</b></p> <p><a href="#">[Clone]</a> <b>Rerun Analysis</b></p> <p><a href="#">[ View All Library Hits ]</a>   <a href="#">View All Clusters With IDs (Beta)</a>   <a href="#">View All Compounds</a> ]</p> <p><b>Legacy Views</b></p> <p><a href="#">[ View All Clusters With IDs ]</a></p> <p><b>Experimental Views</b></p> <p><a href="#">[ Reanalyze Cluster Spectra ]</a>   <a href="#">Networking Parameters and Written Network Description</a>   <a href="#">View Raw Spectra</a> ]</p> <p><b>Auxiliary Views</b></p> <p><a href="#">[ View Network, Node Centric ]</a>   <a href="#">View Network Pairs</a>   <a href="#">Networking Statistics</a> ]</p> <p><b>Network Visualizations</b></p> <p><a href="#">[ View Spectral Families (In Browser Network Visualizer) ]</a>   <a href="#">Network Summarizing Graphs</a> ]</p> <p><b>Export</b></p> <p><a href="#">[ Download Clustered Data ]</a>   <a href="#">Download Cytoscape Data</a>   <a href="#">Make Public Dataset</a> ]</p>
<b>User</b>	mwang87 (miw023@ucsd.edu), UCSD CSE Student
<b>Title</b>	GNPS Paper Stenothricin Network Recreation Trial 6, 0.65 Cosine top 10, 100 component size limited data as per yao recommendation, added daptomycin files
<b>Date Created</b>	2015-07-10 09:08:09.0
<b>Execution Time</b>	7 minutes 35 seconds
<b>Progress</b>	

**Supplementary Figure 20 - Workflow Status Page** – Example of a status page for GNPS analysis job. URLs are sharable to collaborators. Analysis jobs can also be rerun (by cloning) in order to facilitate reproducibility and transparency in methods.



## Supplementary Notes

### Note 1. Browsing Available Libraries

All available libraries at GNPS, including freely available 3<sup>rd</sup> party libraries can be browsed and or downloaded [here](#). Users are able to see continually updated snapshots of each of the respective libraries described in **Supplementary Table 1**. User instructions, available at [gnps.ucsd.edu](http://gnps.ucsd.edu) under documentation, for the Library Spectrum View are be found under the appropriate sections.

### Note 2. GNPS Library Addition Procedure (Collections and Community)

To facilitate library growth, GNPS enables users to contribute to the GNPS-Community spectral library from both private and public datasets. Reference spectra can be uploaded as a set with a batch upload process, one spectrum at a time, or even directly from the output of an analysis tool at GNPS with a single click “Add To Library” button. The metadata fields that are required for user submissions and descriptions can be found at [gnps.ucsd.edu](http://gnps.ucsd.edu) under documentation under the Library Contribution section.

All provenance information is automatically maintained so that proper credit can be given in follow-up analysis. Even with Gold, Silver, and Bronze quality levels in place, it is possible that i) some submissions will be tentative (especially in the Bronze category), ii) some spectra will not yet have the highest possible signal (e.g., low abundance NPs), or iii) metadata might be incomplete at the time of submission (e.g. tentative structures). Therefore a wiki-style Update Annotation workflow is available, providing a path for annotations of library spectra to become more complete. This revision approach enables annotation updates for example from “a hypothetical sugar containing natural product” to a much more specific “Erythromycin” when new information becomes available and can also be used to fix incorrect or imprecise annotations. Also, missing metadata, such as complete structural information, can be added to further complete annotation of library spectra (e.g. [Mycophenolic Acid](#)). To date 376 annotations been updated. Further, to reach a community consensus on certain annotations, a data centric exchange may need to occur. GNPS promotes these conversations by allowing comments on annotations with additional supporting information attachments (e.g. [stenothricin-GNPS 2](#)).

### Note 3. Library Search False Discovery Rate Estimation

With library search being an integral part of molecular networking, dereplication, and continuous identification, it is important to assess the false discovery rates of the library search available at GNPS.

The massive datasets MSV000078708 and MSV000078710 are the full LC-MS/MS runs for to create the NIH Natural Product and NIH Small Molecule Pharmacological Libraries. They total 1,327,215 MS/MS query spectra. The dynamic exclusion settings were to fragment a precursor at most three times unless the MS1 peak intensity increased 2x since the last

MS2 acquisition. This ensured that there are multiple MS2 spectra per precursor. Further details can be found in **Methods**. These datasets were searched against the NIH Natural Product and NIH Small Molecule Pharmacological Libraries. The parameters of the search can be found in **Supplementary Table 7** and the analysis job can be found [here](#).

The matches returned were categorized into the following: identity match, consistent correct match, inconsistent correct match, and incorrect match. Identity matches were matches between the library spectrum and the identical spectrum in the data that was extracted to create the library spectrum. Consistent Correct matches are between library spectra and query spectra where the query spectrum occurred in the same LC-MS/MS run as the library spectrum. Inconsistent Correct matches are between library spectra and query spectra such that the ppm error of the match is <50ppm and are within 200 scan numbers of each other, but the library and query did not come from the same LC-MS/MS file. These could have resulted from contamination and impurities in compound library. All other matches are considered incorrect.

The 50ppm tolerance and 200 scan thresholds were chosen based on the empirical match data. We took all library spectra and matched them to all MS/MS query spectra with a 2 Da precursor tolerance. The top scoring match for each query spectrum was reported if the match score exceeded 0.7. The distribution of all possible ppm precursor  $m/z$  errors are shown in **Supplementary Figure 1a**. For all of these matches that were reported, the ppm precursor  $m/z$  error histogram is shown in **Supplementary Figure 1b**. With the possible ppm errors of reported matches ranging beyond 10,000, the reported matches having almost all PPMs below 50 is in indication the scoring scheme is discriminating between true and false matches.

Further, scan number deltas for all possible matches are shown in **Supplementary Figure 1c**, which again shows a high variance in the possible deltas that could occur. However, of the reported matches, the scan number delta converges significantly as shown in **Supplementary Figure 1d** with nearly all of the identifications occurring under a scan delta of 200 (~5% of the chromatographic time).

The spectral library search at GNPS returned 8,355 matches, of which 2,428 were Identity matches, 4,557 were Consistent Correct matches. 1,138 were Inconsistent Correct matches, and 232 were incorrect matches. We calculated the false discovery rate to be 3.9% as the number of incorrect matches divided by the total number of correct identifications with the identity matches removed.

Further, to show the sensitivity of the search, the library included a total of 2,716 MS/MS spectra. Of which, 2,428 library spectra matched to the identical spectra used to create the library. Thus, 288 of the library spectra were not matched because they did not meet the minimum number of peaks required to be considered identifiable (6 peaks). Of the 2,428 library spectra that were deemed identifiable, 2,045 library spectra had at least one replicate identification (**Supplementary Fig. 2**). This allows us to estimate the sensitivity of the

search at 84% by dividing the total number of library spectra that had at least one correct non-identity match (2,045 spectra) by the number of library spectra that had an identity match (2,428 spectra).

#### **Note 4. GNPS Dataset Creation and Access**

Entire project data sets, irrespective of size and number of data files, can be uploaded, managed, analyzed, and shared with the entire community. Each of these public dataset is given a unique identifier that can be referenced and linked in journal articles. A variety of open file formats (mzXML, mzML, and MGF) as well as several proprietary file formats can be uploaded to the system. Public datasets will be made available in their original submission format and in an open format that is compatible with all GNPS tools shortly after dataset publication. This process usually takes 48 hours and availability status is shown on the dataset page under the heading “Analyze Data”. Since GNPS datasets will be public and compatible with GNPS tools, public datasets can be re-analyzed online. While metadata such as PI and username are required, metadata such as protocols and similar details can be provided but are not mandatory (though can be updated in the future). For further documentation regarding dataset creation, dataset browsing (including downloading data, online re-analysis of data, and viewing continuous identification results), see the appropriate section in the GNPS online documentation (available at [gnps.ucsd.edu](http://gnps.ucsd.edu) under documentation).

#### **Note 5. Public Data Molecular Networks**

All public NP datasets were clustered<sup>5</sup>, producing consensus spectra as described in **Methods – Molecular Network Construction**. The molecular networking parameters can be found in **Supplementary Table 9**. Exploratory molecular networks were also created with parameters found in **Supplementary Table 8**. For each dataset, the Identified Molecules is the number of consensus spectra matched to a library spectrum by continuous identification. The Putative Analog Molecules is the number of consensus spectra that had an edge in the molecular network to a library matched consensus spectrum. The Identified Networks is the number of consensus spectra that were a part of connected components (node count > 1) in the molecular network that had at least one consensus spectrum match to a library spectrum. The Unidentified Networks is number of consensus spectra that were a part of connected components (node count > 1) in the molecular network that did not have any matches to library spectra. The Exploratory Networks is the number of consensus spectra that were a part of connected components (node count > 1) in the exploratory molecular network (**Fig. 4a**). The overall public data identification rate was Identified Molecules summed across all datasets divided the total number of consensus spectra across all datasets (**Fig. 4b**).

#### **Note 6. Continuous Identification Procedure**

Each single public GNPS dataset is run through the standard Molecular Networking workflow. Parameters can be found in **Supplementary Table 9**. The clustered set of spectra are then searched against the GNPS-Collections, GNPS-Community, MassBank<sup>15</sup>, ReSpect<sup>16</sup>, and NIST<sup>17</sup> libraries with the molecular library search workflow (see search parameters at **Supplementary Table 10**). The most current search results are then compared with the previous continuous identifications on a per dataset basis. Any changes in matches (New, Different, or Deleted) are reported. Matches for previously unidentified spectra are defined as New. Changes in matches of spectra due to annotation updates or a better library match are defined as Different. Spectra that previously had a match but currently do not due to removal of library spectra or updates to library search scoring functions are defined as Deleted. These results are reported in two different ways: i) all are shown on the GNPS Dataset page under the Continuous Identification section (e.g. [link](#)) ii) users subscribing to each dataset are emailed a summary of changes in identifications (Each user receives only one email with all summaries to avoid spamming). An email is only when a change or new entry is obtained for a dataset to which a user is subscribed to.

#### **Note 7. Molecular Explorer Creation and Presentation**

All Continuous Identification matches are aggregated and grouped by matched compound name. The GNPS [Molecular Explorer](#) presents all matched library compounds and their respective occurrences in the public data. Thus, it is possible to easily ask the question in which public datasets contain a certain molecule. Further, since GNPS also constructs molecular networks for public datasets, the molecular explorer also presents occurrences of putative analogs of library compounds. **Supplementary Figure 6** and **Supplementary Table 5** highlight the amount of diversity of analogs captured by the molecular explorer as well as the fraction of this diversity that is already captured by reference spectra in GNPS spectral libraries. For further instructions, see the Molecular Explorer section GNPS online documentation.

#### **Note 8. Views from GNPS users**

*Nature Biotechnology* asked independent researchers for feedback on different aspects of GNPS. Their views are presented below.

##### **Is this interface different from others available?**

Respondent 1 (anonymous). The interface is different from others that are available. We laud the efforts of trying to combine as many mass spectra databases as possible and to provide analytical tools to help you home in on significant aspects for your spectra. Our main concern about the user interface is the complexity—it's a bit difficult to navigate/use but it can likely be learned once you become familiar with the layout of the site and the intent of each page.

Respondent 2 (Bo Li and Ashley Kretsch). This is the first interface of its kind that I have worked with, but I have limited experience in metabolomics and molecular networking before GNPS.

**Comment by GNPS:** There are many experiments possible with GNPS and therefore the complexity of the analysis depends on the complexity of each experiment. For example, while dereplication of a few LC-MS files can be done using an in-browser drag-and-drop interface, more complex network visualizations may require uploading metadata files, transferring of large mass spectrometry files using an FTP client, and exporting files from GNPS for offline visualization. To tackle this complexity, GNPS's workflows have detailed step-by-step written instructions and online instructional videos (linked to through "Documentation" on the banner at GNPS). In addition, the GNPS forum facilitates the answering of more detailed questions and assists with hands-on troubleshooting where both GNPS administrators as well as the community can provide feedback.

### **Does it offer unique and compelling features that mean you want to continue to use it?**

Respondent 1. The major compelling feature of this tool is the network analysis of your spectra relative to all known spectra. This is an idea whose time has come that we hope will be useful. For our work, it is unlikely we will need/use this approach as our work is almost always driven by genetics. Where I imagine this will be useful is more classic "grind-n-find" or crude extract approaches where you could upload data without any idea of what's present, and if you're lucky, be given a clue about what is in your extract. The potential utility of this tool would be to guide an investigator to look at the extract that provided that anomalous peak.

Respondent 2. Having the networking available on the website platform allows for fast and efficient evaluation of compound clusters without having to upload and annotate in cytoscape. In addition, this is a great tool to be able to compare MSMS spectra directly between two linked compounds. This has been especially useful when working with novel compounds, where the structure for one might be unknown.

**Comment by GNPS:** The analysis capabilities provided by GNPS enable a shift in thinking about one molecule in isolation to thinking about relationships between all molecules in a sample, collection of samples or even shared data from several samples. GNPS can be used to organize entire culture collections and this data can be shared only by one lab or person, or made accessible to the whole community. We foresee developing automated tools for genetic manipulation and performing mass spectrometry screens on hundreds of thousands of samples that can then be analyzed in GNPS. Currently there is no other computational infrastructure capable of molecular analysis at such scales.

### **How straightforward is it to upload spectra and run tests on your data?**

Respondent 1. We didn't try this.

Respondent 2. The tutorials are easy to follow, once you have uploaded and analyzed data it is easy to repeat the process. Any questions I have had are also answered in a very timely manner via e-mail.

**Comment by GNPS:** We have included links to the appropriate documentation/videos for running a user's first analysis on the data analysis page.

**Would you consider using this interface to deposit all spectra/experiments as you work?**

Respondent 1. Probably not. The main issue would be concerns about the privacy of the data. If this was a community-agreed upon repository for post-publication deposition, we would consider it.

Respondent 2. My work deals mostly with comparative metabolomics and structure identification, so my data might not be useful for the metabolite community as a whole compared to some of these broader databases. I like the feature where you can pick what data is accessible to the whole group and what is private to the user.

**Comment by GNPS:** All GNPS user data is considered private until users explicitly decide to make it public through GNPS workflows designed for data sharing. But while private data can always be manually re-searched using frequently updated GNPS libraries, making data public has the advantage that it becomes part of the living data space where knowledge from continuous identification is shared and disseminated to all subscribers.

**Do you think there are problems with existing databases for MS?**

Respondent 1. For us they work fine, but we usually have 1) genetic information and 2) purified compounds. For this case the existing databases are fine.

Respondent 2. We find that with our smaller molecules, our library hits can be unrelated to our compounds despite a high cosine score. This might be due to the low number of fragments (i.e., for compounds less than 200).

**Comment by GNPS:** Having MS/MS of pure compounds in GNPS is valuable for the community. Such data enable non-natural product scientists to search data. For example, a microbiome person may be aiming to understand the biology of a soil community or a gut community and find matches to purified standards. Regarding matches to small molecules, this is unfortunately a known inherent limitation of the chemical properties of small molecules in the gas phase—there will indeed be fewer fragment ions with small molecules and these are indeed more challenging to match than larger molecules that generate more fragment ions. We recommend a minimum of 6 MS/MS fragment ions to match in addition to the parent mass and we show in supplementary materials (**Supplementary Note 3**) how these settings result in very low estimated false discovery rates. While GNPS-based analysis is powerful, we always advise validation of results with additional methods.

### **Are the social features appealing (live data?).**

Respondent 1. The main issue for us with the live data idea is that the social aspects are only useful if there is some agreement about data use. The hope for the GNPS platform is a Google-like scaling effect where as you get more data, you get better at making predictions. This virtuous cycle is virtuous only if you can benefit from it, so the social problem is that whoever controls this data set has an advantage, potentially at the expense of the depositor.

Respondent 2. I think it is very helpful for ongoing experiments. As more knowledge is generated, this is carried over to your data, much like a system update. As the field continues to expand and GNPS is more widely used, this will enhance the overall experience of the platform.

**Comment by GNPS:** All GNPS public data, spectral libraries, data analysis workflows and continuous identification results are publically, immediately and freely available to all users, not just GNPS administrators. If no one had shared gene annotations that were carefully curated in a community-wide platform, BLAST would not be very informative. We further note that in the same way that the openness of NCBI repositories (e.g., GenBank) and algorithms (e.g., BLAST) have not resulted in 'unfair advantages' for NIH intramural researchers. We also expect that the openness of GNPS data and algorithms will enable researchers equally across the community, regardless of their affiliation or geographical location (a trend that is already supported by the GNPS community including users from 100+ countries, several of which have already published independently of GNPS administrators). Community members have benefited from the openness of data at GNPS, with 83.6% of identifications made in public data by matching to reference spectra uploaded to GNPS by another member of the community.

### **Is the openness an enabling and attractive feature?**

Respondent 1. Not really. While openness is a good trend we believe the primary beneficiaries of the openness are almost always those who are running the platform.

Respondent 2. Yes, it influences a sense of collaboration in the metabolomics field. For instance, we have been investigating secondary metabolites in *Burkholderia cenocepacia*. By looking at our data in comparison to samples taken from the CF lung, we can explore how *B. cenocepacia* is involved in infection at a chemical level

**Comment by GNPS:** Sharing data is becoming a requirement of funding agencies because it benefits the whole research community. For example, Genbank is an open repository that has become all but indispensable in genomics research. GNPS aims to meet this pressing need in the natural product field.

### **Are there creative aspects of GNPS that you feel deserve highlighting?**

Respondent 1. The clustering idea is a great idea. Or rather, the idea of applying computational techniques to large scale data sets to direct experimental research is a first-rate idea. As the data accrues, it may be true that the predictive power of these spectra get

better and better. So we would say that the creativity is actually in keeping the idea simple (clustering spectra with similarly occurring peaks). This will hopefully allow the scaling of data to the point that old data begins to provide useful clues that would have been difficult to achieve at smaller scales. Respondent 2. The web interface, especially the networks, where clicking between two compounds shows a side by side comparison, points can be labeled by library hits or  $m/z$  values, size can be proportional to intensity, and color coded by what groups this compound is included in. It allows easy transition to view all of the available features of each compound.

### **Note 9. Stenothricin-GNPS - Molecular Characterization Workflow**

Previous genomic investigations showed that *Streptomyces* sp. DSM5940, a known producer of napsamycin analogs, has 95-99% identity to *S. roseosporus* NRRL 15998 at the protein level of obtained sequences<sup>18</sup>. Due to the high sequence homology between the two strains, we hypothesized that *Streptomyces* sp. DSM5940 may have the genetic capacity to produce similar molecules to *S. roseosporus* including daptomycin, arylomycin, and stenothricin analogs<sup>19</sup>. To determine if this was indeed the case, we utilized a GNPS based molecular networking workflow (**Supplementary Fig. 7**). First, we generated a molecular network with MS/MS data from crude extracts of *Streptomyces* sp. DSM5940 and *S. roseosporus*. Dereplication of known metabolites such as those in the napsamycin, daptomycin, arylomycin, and stenothricin molecular families allows for the detection of novel analogues. Once a molecule was prioritized, structure elucidation and biosynthetic analysis was undertaken, leading to the characterization of previously unreported metabolites.

A molecular network of *Streptomyces* sp. DSM5940 with *S. roseosporus* was generated using GNPS (parameters in **Supplementary Table 13**) from *n*-butanol and methanol extracts from cultures grown on solid agar (**Supplementary Fig. 8**). We identified four molecular networks corresponding to analogs of napsamycin, daptomycin, arylomycin, and stenothricin. Interestingly, daptomycin was only produced by *S. roseosporus* (**Supplementary Fig. 8b**) indicating that either *Streptomyces* sp. DSM5940 does not have the daptomycin gene cluster or daptomycin was not produced under our culture conditions. For the napsamycin and arylomycin molecular networks, we detected previously reported metabolites (**Supplementary Fig. 8a,c**). Of particular interest to us was the stenothricin molecular network. The MS/MS data from *Streptomyces* sp. DSM5940 created a distinct sub-network connected to known stenothricin analogs produced by *S. roseosporus* (**Supplementary Fig. 8d**). The precursor  $m/z$  values for the sub-network were 41 Da less than corresponding precursor values for known stenothricin compounds in the network. We named these five analogues stenothricin-GNPS 1-5 (**Supplementary Table 12**). The molecular networking job can be found [here](#).

### **Note 10. Stenothricin GNPS 2 Structure Elucidation**



We manually examined the MS/MS fragmentation data of the stenothricin-GNPS sub-network to verify their relatedness to the stenothricin molecules (**Supplementary Fig. 9**). Members of both the stenothricin molecular network and stenothricin-GNPS sub-network had the same MS/MS sequence tag resulting from a McLafferty-type rearrangement corresponding to the amino acid sequence CysA-Dhb-Ser-Dpr-Dhb. To identify the location of the 41 Da mass difference, a comparative NMR approach was utilized<sup>20</sup>. Stenothricin GNPS 2 was isolated from *Streptomyces* sp. DSM5940 and 2-dimensional NMR was acquired including HSQC, HMBC, dqfCOSY, and TOCSY spectra. These spectra were then compared with the original data of stenothricin D using an overlay designed to highlight differences in the NMR data (**Supplementary Fig. 10-14, Supplementary Table 13**). Observed differences in the NMR spectra of stenothricin GNPS 2 compared to stenothricin D included a lack of Lys related signals, as well as a third Ser spin system suggesting that the stenothricin-GNPS sub-network is the result of a Ser substitution for Lys (**Fig. 5b**). This was further supported by targeted Marfey's analysis<sup>21,22</sup> where three serine residues were detected (**Supplementary Fig. 15**).

#### **Note 11. Stenothricin GNPS *in silico* Biosynthetic Analysis**

To further corroborate the Lys to Ser substitution, we sequenced the *Streptomyces* sp. DSM5940 genome and performed comparative analyses of the stenothricin gene cluster families between *Streptomyces* sp. DSM5940 and *S. roseosporus* NRRL 15998 (**Supplementary Fig. 16**)<sup>11</sup>. Stenothricin is a putative non-ribosomally encoded peptide natural product. Comparison of the adenylation domain (A-domain) 10 amino acid specificity codes<sup>23,24</sup> between the two gene clusters revealed identical specificity with the exception of the eighth A-domain. In *S. roseosporus*, the eighth A-domain has a 10 amino acid code (DVFESGGVAK) consistent with Lys activation. The corresponding A-domain in the *Streptomyces* sp. DSM5940 has a Ser activating specificity code (DVWHVSLVDK)<sup>12</sup>. Phylogenetic analysis of the condensation domains (C-domains) of both gene clusters suggests that stenothricin GNPS has the same stereochemistry as the stenothricin compounds from *S. roseosporus* (**Supplementary Fig. 17**).

#### **Note 12. Bioactivity of Stenothricin GNPS**

Stenothricin D produced by *S. roseosporus* has been previously shown to have unique antimicrobial activity against both Gram-positive and -negative bacteria in a detergent like manner determined by fluorescence microscopy based cytological profiling<sup>25</sup>. Visualization of *E. coli* *lptD* and *Bacillus subtilis* treated with stenothricin D revealed a uniform staining of membrane throughout the cells with the membrane stain FM 4-64. The identification and purification of stenothricin-GNPS sub-network members (five analogs of varying length of the lipid chain, see **Supplementary Table 12**) allowed for structure activity relationship (SAR) comparison of the Lys to Ser substitution. *B. subtilis* and *E. coli* *lptD* were treated with stenothricin-GNPS 2/3 (**Supplementary Fig. 18**). Even at 40 µg/mL, stenothricin-GNPS had no effect on *B. subtilis*. Stenothricin-GNPS treatment of *E. coli* *lptD* cells led to

decreased viability at 40 µg/mL similar to that of stenothricin D at 20 µg/mL and led to cell permeability as visualized using Sytox green. However, stenothricin-GNPS did not show the same uniform membrane staining unique to stenothricin D. This strongly suggests that the Lys residue of stenothricin D is critical for its antimicrobial mechanism of action.

## References

1. Frewen, B. & MacCoss, M. J. Using BiblioSpec for creating and searching tandem MS peptide libraries. *Curr. Protoc. Bioinformatics* **Chapter 13**, Unit 13.7 (2007).
2. 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).
3. Lam, H. *et al.* Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655–667 (2007).
4. 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).
5. Frank, A. M. *et al.* Clustering millions of tandem mass spectra. *J. Proteome Res.* **7**, 113–122 (2008).
6. Shirling, E. B. & Gottlieb, D. Methods for characterization of *Streptomyces* species. *International Journal of Systematic Bacteriology* **16**, 313–340 (1966).
7. 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).
8. Rutherford, K. *et al.* Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944–945 (2000).
9. 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).
10. Carver, T. *et al.* Artemis and ACT: Viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24**, 2672–2676 (2008).
11. Röttig, M. *et al.* NRPSpredictor2 - A web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39**, (2011).
12. 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).
13. 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).
14. 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).
  15. H, H. *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. (2010).
  16. Y, S. *et al.* RIKEN tandem mass spectral database (ReSpect) for phytochemicals: a plant-specific MS/MS-based data resource and database. (2012).
  17. The National Institute of Standards and Technology. NIST. at <<http://www.nist.gov/srd/nist1a.cfm>>
  18. Kaysser, L. *et al.* Identification of a Napsamycin Biosynthesis Gene Cluster by Genome Mining. *ChemBioChem* **12**, 477–487 (2011).
  19. Höltzel, A. *et al.* Arylomycins A and B, new biaryl-bridged lipopeptide antibiotics produced by *Streptomyces* sp. Tü 6075. II. Structure elucidation. *J. Antibiot. (Tokyo)*. **55**, 571–577 (2002).
  20. Forseth, R. R. *et al.* Identification of cryptic products of the gliotoxin gene cluster using NMR-based comparative metabolomics and a model for gliotoxin biosynthesis. *J. Am. Chem. Soc.* **133**, 9678–9681 (2011).
  21. Bhushan, R. & Brückner, H. Marfey's reagent for chiral amino acid analysis: A review. *Amino Acids* **27**, 231–247 (2004).
  22. Boudreau, P. D., Byrum, T., Liu, W. T., Dorrestein, P. C. & Gerwick, W. H. Viequeamide a, a cytotoxic member of the kulolide superfamily of cyclic depsipeptides from a marine button cyanobacterium. *J. Nat. Prod.* **75**, 1560–1570 (2012).
  23. Stachelhaus, T., Mootz, H. D. & Marahiel, M. A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6**, 493–505 (1999).
  24. Challis, G. L., Ravel, J. & Townsend, C. A. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.* **7**, 211–224 (2000).
  25. 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).