The Journal of Antibiotics (2014) 67, 105–112
© 2014 Japan Antibiotics Research Association All rights reserved 0021-8820/14
www.nature.com/ja

ORIGINAL ARTICLE

Quantitative molecular networking to profile marine cyanobacterial metabolomes

Jacob R Winnikoff[1], Evgenia Glukhov[1], Jeramie Watrous[2], Pieter C Dorrestein[2,3] and William H Gerwick[1,3]

Untargeted liquid chromatography-MS (LC-MS) is used to rapidly profile crude natural product (NP) extracts; however, the quantity of data produced can become difficult to manage. Molecular networking based on MS/MS data visualizes these complex data sets to aid their initial interpretation. Here, we developed an additional visualization step for the molecular networking workflow to provide relative and absolute quantitation of a specific compound in an extract. The new visualization also facilitates combination of several metabolomes into one network, and so was applied to an MS/MS data set from 20 crude extracts of cultured marine cyanobacteria. The resultant network illustrates the high chemical diversity present among marine cyanobacteria. It is also a powerful tool for locating producers of specific metabolites. In order to dereplicate and identify culture-based sources of known compounds, we added MS/MS data from 60 pure NPs and NP analogs to the 20-strain network. This dereplicated six metabolites directly and offered structural information on up to 30 more. Most notably, our visualization technique allowed us to identify and quantitatively compare several producers of the bioactive and biosynthetically intriguing lipopeptide malyngamide C. Our most prolific producer, a Panamanian strain of *Okeania hirsuta* (PAB10FEB10-01), was found to produce at least 0.024 mg of malyngamide C per mg biomass (2.4%, w/dw) and is now undergoing genome sequencing to access the corresponding biosynthetic machinery.
The Journal of Antibiotics (2014) **67,** 105–112; doi:10.1038/ja.2013.120; published online 27 November 2013

Keywords: biosynthesis; cyanobacteria; liquid chromatography-MS; metabolome; molecular network; natural products; quantitation

INTRODUCTION

For millennia, naturally occurring metabolites have been used to treat human disease and improve health. Perhaps the most familiar natural products (NPs) in modern medicine comprise the wealth of antibiotics produced by fungi and actinomycete bacteria, although in fact 50% of all small-molecule drugs clinically approved in the United States and Europe between 1981 and 2010 have been derived from or inspired by NPs.[1] In recent decades, advances in synthetic chemistry and genetics have evolved to modify, refine and produce NPs in order to improve their pharmaceutical properties and resulting clinical efficacy.

The global ocean biome represents an enormous reservoir of biochemical diversity that has become increasingly accessible owing to SCUBA diving and other technologies that aid sub-tidal marine collections. Marine cyanobacteria are an especially promising source of NPs because of their lack of physical defenses, ancient evolutionary origins, high genetic mutation rates relative to eukaryotes and capacity for horizontal gene transfer.[2,3] These factors combine to result in diverse secondary metabolomes, with components ranging from simple hydrocarbons to halogenated macrolides and complex peptides.[4] Because marine cyanobacteria produce many distinct types of NPs, a need has arisen for more powerful and efficient methods by which to profile their metabolomes.

Molecular networking is an analytical method well suited to this task.[5] Molecular networks are visual representations of mass-spectral data sets generated using a vector-based similarity score for tandem mass spectra.[6] Initially, MS/MS spectra that are nearly identical are combined and averaged to form 'consensus' spectra, each of which is represented graphically in the network as a circular 'node.' The consensus spectra are then compared pairwise, and their corresponding nodes in the network are linked with edges based on their structural similarity. Figure 1 illustrates a traditional molecular network representing the ionizable metabolome of one organism. The high scalablility of molecular networking in terms of number of MS/MS data sets that can be included and captured in one visual network has proven useful in dereplication efforts and in seeking analogs within desired molecular classes,[6] motivating the networking of multiple metabolomes. Larger networks featuring more organisms promise faster dereplication and unique insights yielded by metabolomic comparison.[6,7] Although molecular networking has no

[1]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA; [2]Department of Chemistry and Biochemistry, University of California San Diego, La Jolla, CA, USA and [3]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA
Correspondence: Professor WH Gerwick, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, 9500 Gilman Drive MC 0212, La Jolla, California 92093, USA.
E-mail: wgerwick@ucsd.edu
Received 31 August 2013; revised 8 October 2013; accepted 11 October 2013; published online 27 November 2013

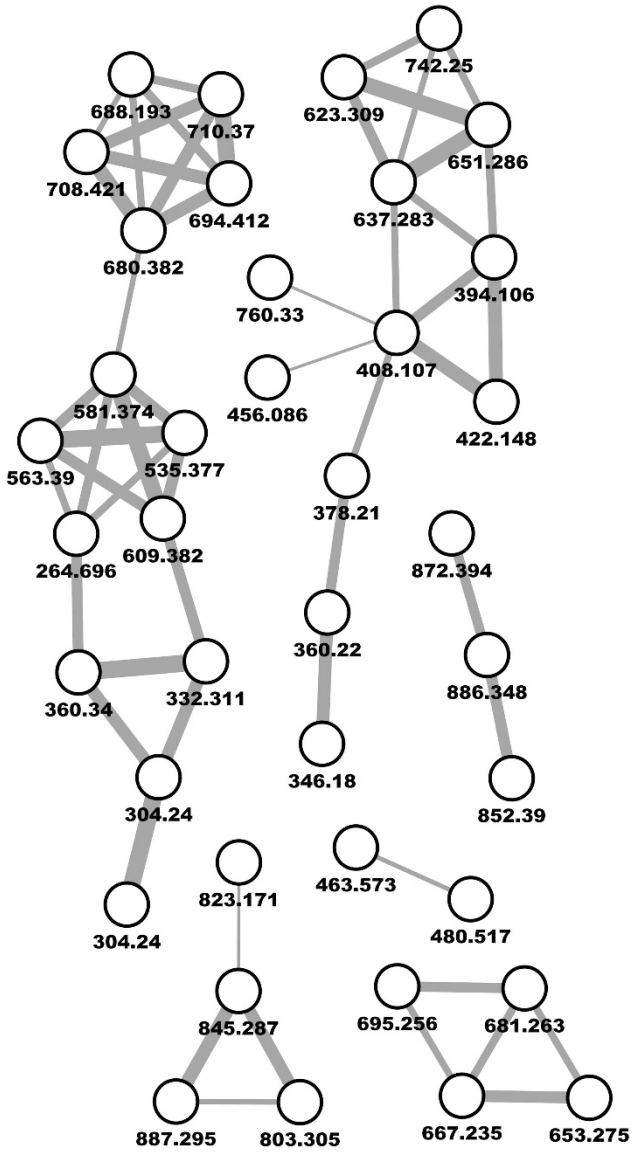

Figure 1 Simple molecular network from a cultured black *Moorea* sp. (collection PAL15AUG08-1). Nodes are labeled with parent *m/z* ratio and edge thickness is mapped to cosine similarity score.

known limit for data volume,[6] existing methods for visualizing multi-metabolomic networks can become challenging, especially with respect to where each MS signal comes from and its abundance.

Herein, we describe the design and application of a new addition to molecular networking software. The new script, termed TOrTE (Tandem-MS Origin Tracing Engine), offers a way to quickly and intuitively visualize the extract origins and, if applicable, pure compound 'seed' matches for each compound in a molecular network. Furthermore, it includes an algorithm to compare the quantities of each networked compound in all of the extracts and pure compound seeds in which it is found. Operating with or without this quantitative feature, the TOrTE can create large, multi-metabolome molecular networks with a wealth of intuitively visualized data (Figure 2[8–15]). Such networks offer deep metabolomic insight and applications in chemotaxonomy, expression analysis, study of biosynthetic mechanisms and microbial culture management.

We applied the newly expanded capacity of molecular networks to search our cyanobacterial culture collection for strains producing known metabolites of interest and their analogs. This effort was rewarded with the discovery of several malyngamide C producers. Malyngamide C and its corresponding acetate are structurally intriguing lipopeptides with antifungal and cancer cell cytotoxic properties ($LC_{50} = 3.1$ and 2.0 μM for NCI-H460 cells).[16] Studies of the biosynthesis of other cyanobacterial metabolites[17] suggest that they are likely assembled via a hybrid polyketide synthase (PKS) and non-ribosomal peptide synthase pathway. Analysis of the likely sequence of reactions leading to malyngamide C further suggests a potentially novel mechanism for carbocyclic ring formation coincident with off-loading from the terminal PKS module (Supplementary Figure S3), but producer DNA has been unavailable to date. Thus, locating a malyngamide C producer in the Scripps collection of marine cyanobacterial cultures was a high priority.[16] TOrTE-based molecular networking facilitated the search for a malyngamide C producer and revealed sources of several other secondary metabolites in our culture collection. In so doing, the new technique demonstrated great potential to streamline and make more fruitful NP drug discovery and chemical biology programs.

RESULTS

Visualizing strain origin of NPs with qualitative color coding

When molecular networks are expanded beyond a single metabolome, visual schemes like that in Figure 1 can obscure valuable data such as organism(s) of origin for each compound and relative quantities of metabolites. Color-coding nodes according to origin solves this issue, and multiple color-coding approaches exist to serve distinct needs.[5–7] Thus, the TOrTE was adapted accordingly. Qualitative color coding as seen in Figure 3a can be useful in simply ascertaining production of a compound by a certain strain. It may also be necessary when quantitatively valid data are unavailable, or when mathematical corrections have not yet been applied to data obtained under different LC-MS/MS protocols. Qualitative color coding by origin has been used in molecular networks, but not typically with multicolored nodes.[5–7] Because of consistency, pie charts visualized using one color per strain or seed provide an inherent ease-of-use advantage for many studies over systems in which unique colors correspond to particular combinations of strains. The one-to-one color mapping provided by the TOrTE aided in some of the metabolomic insights discussed below.

Relative and absolute quantitation

The version of TOrTE used to visualize the network shown in Figures 3 and 4 calculates extracted ion chromatogram (XIC) areas as described in METHODS, which are subsequently mapped to the angles used to create node pie charts. XIC area, however, was not the sole quantitative measure with which we experimented. Early versions of the TOrTE were designed to take full advantage of molecular networks' compound differentiation ability[5] by calculating ion abundances for each node based solely on the scans used to form that node's consensus spectrum. Unfortunately, data produced by this method were poorly replicable, with standard deviation of chromatogram area between injections reaching 120% of the mean. This error revealed that when used with the above-mentioned LC-MS/MS method, the approach used to create consensus spectra was not quantitatively accurate. Although internodal connections and consensus contents remained similar across networks produced from identical samples, numbers of consensus-forming scans varied widely and therefore skewed quantitative values. The algorithm described in Figure 4, which uses a precursor mass filter to control integration,

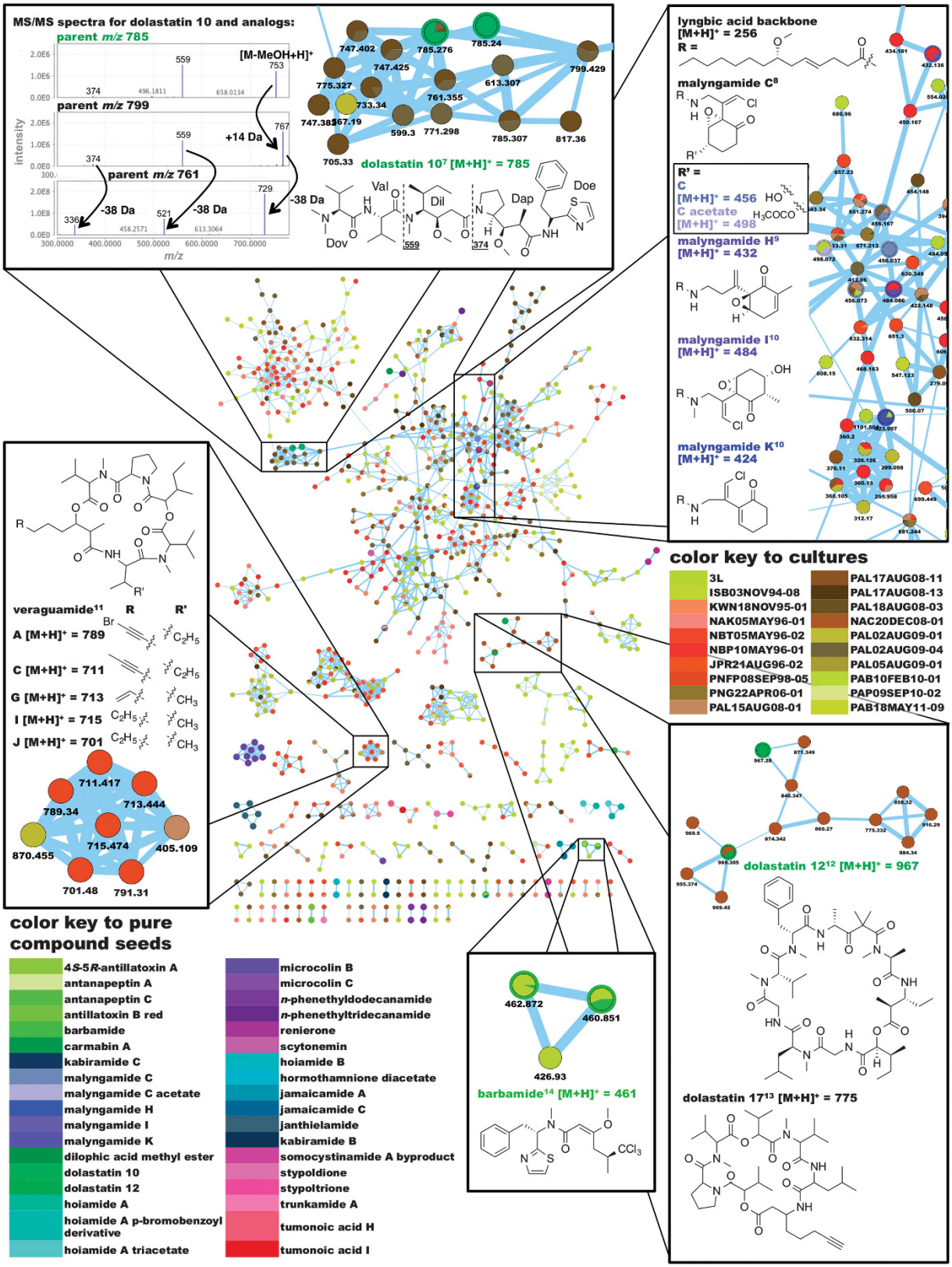

Figure 2 Network of 20 cultured strains seeded with 60 pure natural products (NPs) and NP analogs. Nodes are quantitatively color coded using the process illustrated in Figure 4. Nodes that formed consensus with pure compound seeds are circled with the color corresponding to that seed in the color key, and also contain a seed sector in the interior portion of the node to facilitate visual estimates of compound concentration in various extracts. Structures are shown for molecules in selected clusters, with probable stereochemistry depicted for seed-matching compounds. Dov, Dolavaline; Dil, Dolaisoleucine; Dap, Dolaproine; Doe, Dolaphenine.

yielded much more consistent data between replicate LC-MS injections, with standard deviations not exceeding 53.4% of the mean ($n = 5$, see Supplementary Figure S4 for data used in this variability analysis). Standard deviation of malyngamide C abundance in the pure standard was 11.5% of the mean.

Figures 2–4 display products of the TOrTE run in quantitative mode, and visualize the relative quantities of distinct compounds ionized in each sample (for example, the quantities of malyngamide C from PAL15AUG08-01 and PAB10FEB10-01 in the pie chart, Figure 4). When all of the data in the network is collected using a uniform LC-MS/MS protocol, major producers of a given compound can be easily differentiated from minor ones. Mathematical corrections (see METHODS) can be devised if data collected using different protocols needs to be compared quantitatively. Automated relative quantitation and pure compound seeding can also be used synergistically to provide absolute quantitation of select metabolites. Associated calculations are given in METHODS and malyngamide C is offered in Supplementary Information S2 as a specific example.

TOrTE-based molecular networking as a metabolomic screening technique

The mass-spectral data set used in this investigation was obtained via LC-MS/MS analysis of 20 cultured strains of tropical marine cyanobacteria from which lipid extracts could be obtained (see Table 1). An advantage of using molecular networking to profile cyanobacterial metabolomes is that the technique is sensitive enough to use with MS data made from unfractionated 'crude' extracts. In Figure 1, 35 distinct compounds resolved from one 0.020 ml injection of a crude cyanobacterial lipid extract. By directly analyzing crude extracts, analysis time was significantly reduced and the natural proportions between the constituents of an extract as they elute and enter into the ion source were maintained. The latter point enables very reasonable comparative quantitation.

Figure 2 represents the entire 20-strain set of crude extracts wherein nodes are color coded according to cyanobacterial strain. If multiple strains were found to produce the same compound, then that

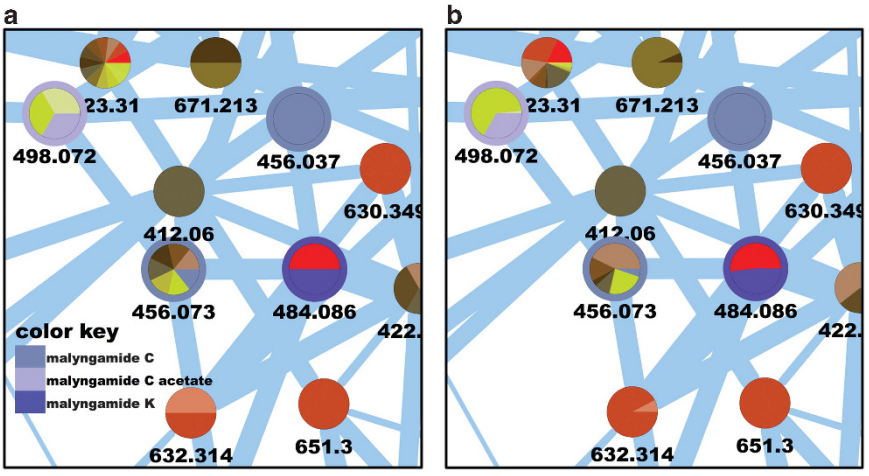

Figure 3 A section of the seeded molecular network in Figure 2 containing malyngamides and malyngamide-like compounds, shown with qualitative (**a**) and quantitative color coding (**b**) as implemented with the Tandem-MS Origin Tracing Engine. Different color-coding modes do not alter network topology. Qualitative color coding is less computationally intensive than its quantitative counterpart, and creates pie charts with equal-sized sectors. This visualization can be useful because it identifies all of the extracts (and thus producer strains) yielding a given MS signal irrespective of signal amplitude.

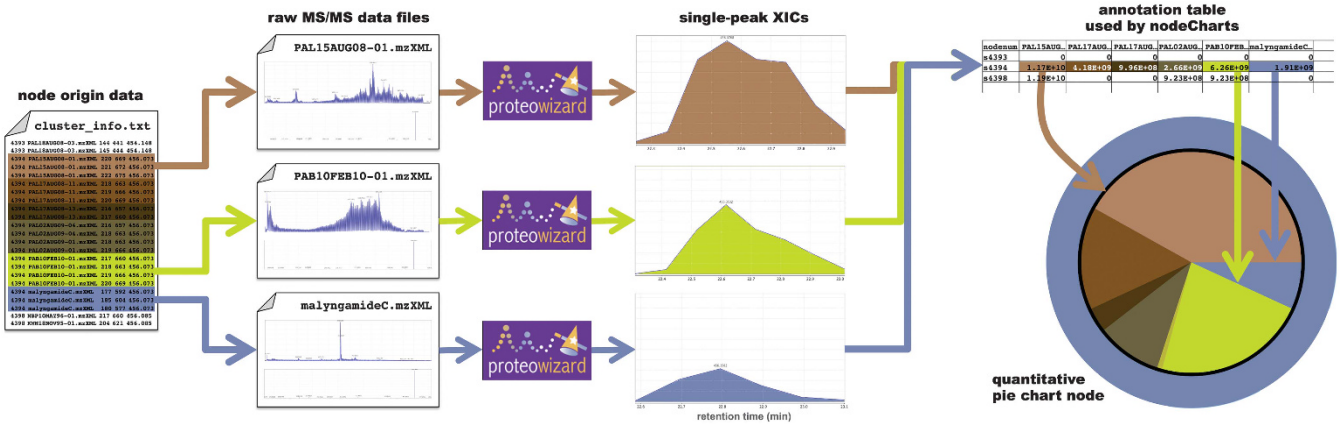

Figure 4 Flowchart illustrating quantitation algorithm used by the Tandem-MS Origin Tracing Engine (TOrTE). The TOrTE uses the cluster information file generated by the Spectral Networks scripts to identify the MS/MS data files giving rise to each node. When running in qualitative mode, these data are entered directly into a large Boolean matrix, which Cytoscape uses to generate pie charts. When running in quantitative mode, the TOrTE loads the MS/MS data files referenced in the cluster information file, extracts the chromatogram peaks giving rise to each node with ProteoWizard, and then calculates the areas for each peak. These area values are entered into a matrix and mapped to pie chart sectors (and in the case of pure seeds, to colored borders) in Cytoscape.

Table 1 Source organisms for crude extracts processed in the quantitative network shown in Figure 2.

SIO collection code[a]	*Taxonomic/field ID*	*Collection site*
3L	*Moorea producens* (GU727199)[b]	Curaçao
ISB03NOV94-08	*Phormidium* (KC207938)[b]	Indonesia
KWN18NOV95-01	*Moorea* sp.	Key West, FL
NAK05MAY96-01	*Schizothrix* sp.	Curaçao
NBT05MAY96-02	'Green mossy filamentous' cyanobacterium	Curaçao
NBP10MAY96-01	*Moorea producens* (NYS)[b]	Curaçao
JPR21AUG96-02	*Moorea producens*	Jamaica
PNFP08SEP98-05	'Dark green tuft' cyanobacterium	Papua New Guinea
PNG22APR06-01	*Moorea producens* (FJ356669-70)[b]	Papua New Guinea
PAL15AUG08-01	black *Moorea* sp. (NYS)[b]	Palmyra Atoll
PAL17AUG08-11	*Moorea*, *Hormothamnion*, Red Macroalga mixture	Palmyra Atoll
PAL17AUG08-13	Chrysophyte/cryptophyte	Palmyra Atoll
PAL18AUG08-03	Green/brown *Schizothrix* sp.	Palmyra Atoll
NAC20DEC08-01	*Moorea* sp.	Curaçao
PAL02AUG09-01	Brown, thin *Moorea* sp.	Palmyra Atoll
PAL02AUG09-04	Brown *Schizothrix* sp.	Palmyra Atoll
PAL05AUG09-01	*Symploca* sp. and *Moorea* sp.	Palmyra Atoll
PAB10FEB10-01	*Okeania hirsuta* (NYS)[b]	Panama
PAP09SEP10-02	*Symploca* sp.	Panama
PAB18MAY11-09	'Pink mat' cyanobacterium	Panama

Abbreviations: NYS, not yet submitted; SIO, Scripps Institution of Oceanography.
[a]See Supplementary Information for guide to collection codes.
[b]Confirmed by 16S phylogeny (GenBank accession code in parenthesis). Other IDs are morphological.

compound's node is represented with a multicolored pie chart. The network also contains MS data from 36 of the 60 pure NP and NP analog 'seed' compounds initially added (see Supplementary Table S6). These 36 were the compounds that formed consensus spectra or correlating edges with compounds in the crude extracts. For easy identification, nodes that matched (formed consensus spectra) with seeds were marked by a color-coded outer border in addition to a seed sector in the node's pie chart.

A number of new insights were gleaned by processing 20 metabolomes into a single large network. Not only were common metabolites instantly visible but also entire clusters of nodes, representing families of structurally related molecules,[7] were found to overlap. The cluster surrounding dolastatin 10,[8] Figure 2 inset, is an example in which 5 of 13 clustered molecules are found in both PAL18AUG08-03 and PAL02AUG09-04 in roughly equal amounts. Indeed, this observation supports the field identifications of these two collections as *Schizothrix*, as this genus is known to produce several dolastatin-type compounds, some of which are of interest as antimalarial agents.[18]

The established technique of seeding a molecular network with pure compounds[6] was found to be especially powerful in the network of the 20 cyanobacterial metabolomes. Although some compounds, such as the veraguamides[12] (Figure 2 inset), could be dereplicated without seeds using fragmentation data from the literature,[12] the seeds added to the network in Figure 2 accelerated and confirmed network annotation by identifying in a single step six known compounds and 30 analogs of known compounds. For example, the structure of an analog of barbamide[15] was suggested from a three-node cluster (Figure 2, inset). The single non-seed-matching node represents a compound of 34 Da lighter than barbamide whose molecular ion isotope pattern was consistent with the dichloro-analog dechlorobarbamide.[19] In another example, dolastatin 10 was linked to partially identifiable analogs: to the right of the seed-matching node was a compound with a 'CH_2' added to either the Dolavaline or Val residues. The presence of a fragment with $m/z = 559$ for both dolastatin 10 and the analog (spectra shown in Figure 2) indicated that structure of the three other residues was likely conserved. Below the dolastatin 10 seed another analog resolved with $m/z = 761$; it also bore an additional CH_2 on Dolavaline or Val, but showed a fragment for the rest of the molecule that was 38 Da lighter, likely representing a modification of the Dolaproine and Dolaphenine residues. Absolute quantitation of these two dolastatin analogs was impossible because of a lack of seeds, but if their ionization efficiencies are comparable to that of dolastatin 10, then the relatively large areas of their eluted LC-MS peaks suggest they may be present in isolable quantities.

Application of TOrTE-based networking

One of the known compounds dereplicated in our network of 20 cyanobacterial metabolomes was the structurally intriguing lipopeptide malyngamide C. Using the full capabilities of TOrTE-based molecular networking, we were able to locate the malyngamide C producers in our culture collection, determine their yields in terms of both extract mass and dry weight (dw), and thus select our most prolific producer per unit dw for DNA extraction and sequencing. This process began with the quantitatively color coded, seeded network in Figure 2. As shown in the inset, a node with parent $m/z = 456$ exhibited a consensus between six of the extracts and the seed for malyngamide C. Furthermore, the mass spectra associated with this node displayed a monochlorinated signature, and the node was located within a more complex cluster featuring parent masses consistent with malyngamides C acetate, I, J, K, L, S and T, along with the precursor fatty acid 'lyngbic acid.'[9–11] Four of these surrounding nodes were also seed-matched, permitting us to confidently dereplicate malyngamide C. By simple inspection of the node pie chart for malyngamide C, we were able to discern our most prolific producer of the compound by extract mass: PAL15AUG08-01, a black *Moorea* sp.

At this point, relative quantitation and the pure compound seed were used together to determine absolute concentrations of malyngamide C. In Figures 3 and 4, each pure seed is mapped as a colored sector as well as an identically colored outer border. Because the amount of malyngamide C injected to obtain the seed data was known (3.3 μg), the ratio (6.26:1.91) of the malyngamide C sectors between PAL15AUG08-01 and the seed sample were used to calculate the amount of malyngamide C in the crude extract injection (0.10 mg of malyngamide C per mg of extract = 10% (w/w); see METHODS). Multiplied by total extract mass over tissue dw, this was equivalent to 1.9×10^{-3} mg mg^{-1} biomass = 0.19% (w/dw). Repeating these calculations led us to another producing culture (PAB10FEB10-01, *Okeania hirsuta*[20]) that produced 2.4% malyngamide C (w/dw), about 10 times more than PAL15AUG08-01 (see Supplementary Information for calculations comparing strains PAL15AUG08-1 and PAB10FEB10-01 and Supplementary Figure S5 confirming that the seed was within the linear dynamic range). Thus, PAB10FEB10-01 was scaled up and its DNA extracted for genome sequencing using Ilumina methods. The malyngamide C gene cluster and biochemical characterization of the unusual off-loading/carbocyclization reaction will be reported in due course.

DISCUSSION

The NP sciences continue to evolve in response to the growing number of described metabolites, perceptions of rediscovery of known compounds (dereplication), tremendous advances in knowledge of the biosynthesis of some major secondary metabolite classes, and exponential increases in the speed and economy of whole-genome or metagenome sequencing. These advances have increased the need for new methods of analysis of complex NP metabolomes, and MS has emerged as an ideal tool in this regard. A consequence, however, is an ever-increasing expansion of primary data, and thus a significant bottleneck has become data analysis. New methods are needed to improve automated data analysis and visualization, which can then permit perception, appreciation and utilization of these large data sets to direct further NP efforts in the most efficient ways possible. The molecular networking algorithm previously described[5] is an analysis platform upon which new sub-routines may be added with relative ease, thus allowing further improvements as described herein.

As with any automated analysis platform, care must be taken to maintain consistency of results with the raw data. In the course of this investigation, we identified two main caveats in our methodology, and here suggest measures to mitigate them. Often, compounds with the same parent mass and very similar MS/MS fragmentation spectra resolve as multiple adjacent nodes. This is especially noticeable when a subset of such nodes is seed-matched, as in the cases of malyngamide C, dolastatin 10 and dolastatin 12 (see Figures 3 and 4). Such nodes may represent (a) mass spectrometry data itself (e.g., the differences of a low concentration vs high concentration of a molecule; although the main ions are the same, some ions will be missing in the spectrum of a molecule at low concentration) that result in a consensus score that bins the data separately or (b) co-eluting isomers with subtly different fragmentation spectra. The first possibility can be accentuated by the combination of data from different LC-MS experiments, thus underscoring the benefit of a standardized LC-MS method to the generation of any molecular network. When the network of cultured strains in Figure 2 was processed with pure compound seeds made using a Kinetex HPLC column (see METHODS), clusters of multiple nodes with apparent masses of malyngamide C and dolastatin 12 resolved, but only one was seed-matched within each of these clusters. Because the scans used to form each of the nodes' consensus spectra were consecutive, it was difficult to tell whether the compounds lacking seed consensus were identical to the seed-matching ones. For diagnostic purposes, the network was reprocessed with five replicate malyngamide C samples made using the same column and gradient as the crudes but at a concentration and injection volume equivalent to that of the other pure compounds. Having been collected under the same protocol, the spectra produced by these runs consistently matched consensus spectra formed by malyngamide C from the crude extracts. Rather than two or more consensus spectra deviating as more spectra are averaged, a phenomenon that occurs as a consequence of the computational method, all nodes with $m/z = 456$ that previously shared an edge with malyngamide C converged to match the malyngamide C pure seed. The dolastatin 12 seed was not re-run and its inset in Figure 2 illustrates an unresolved instance of this issue. The second interpretation of apparently identical neighboring nodes is that they represent true isomeric compounds. This possibility demonstrates a point of caution unique to the TOrTE algorithm. If isomers co-elute within the time window specified for chromatogram integration, their ion intensities are combined, even if the isomers resolve as separate nodes. This issue is due to the XIC integration method used (see METHODS) and is best resolved by developing an HPLC method to separate the isomers in question. However, it should be noted that this theoretical shortcoming was not encountered in the present study.

Although it is generally appreciated that the metabolomes of marine cyanobacteria are rich in NPs, efforts to date have focused primarily on those that are present in large quantity or show biological activity in the relatively few assays used so far in screening campaigns. Thus, a full appreciation of marine cyanobacterial metabolomes is lacking, although recent expansion of the number and phylogenetic coverage of sequenced genomes has given some insight on this issue.[21–23] In the current study, the metabolomes of 20 strains of cultured cyanobacteria were profiled by LC-MS/MS, and the resultant data combined into a molecular network along with the MS/MS data for 60 pure and structurally defined NPs and analogs. The facility with which such an amalgamation of data can be accomplished is a notable strength of molecular networking, and with addition of the TOrTE, multi-metabolomic networks can now be quantified and intuitively visualized. Compounds and even compound families of related chemical structure from different cyanobacteria are connected to one another in a network, oftentimes giving insight into the nature of the chemotype.[22] Pure compound seeds, whose dereplicative power is increased many-fold in a multi-metabolomic network, can enormously strengthen node annotation by possessing a correlating edge or forming a consensus node, as shown with the malyngamide, barbamide and dolastatin 10 seeds. All these features of TOrTE-based molecular networks contribute to a much more comprehensive concept of cyanobacterial metabolomics. They are likewise practical, with numerous applications in contemporary NP research.

The practicality of quantitative molecular networking was confirmed in this study by identification of malyngamide C producers within our marine cyanobacterial culture collection. The importance of this discovery follows from a theoretical analysis of the biosynthesis of the malyngamide compound class and recognition that the terminal step might involve a novel off-loading from a modular type I PKS enzyme manifold. As shown in Supplementary Figure S3, it is conceivable that the enolic form of an enzyme-tethered δ-keto intermediate is involved in coincident carbocyclic ring formation and off-loading from the enzyme. Although there is some precedence for carbon–carbon bond formation during PKS off-loading,[24] to our knowledge this would represent the first occasion wherein a carbocyclic ring is directly formed. Because a defined protocol was employed in the analysis of the 20 cyanobacterial extracts and equal masses of extract were introduced for each species, the chromatogram areas for specific compounds (nodes) occurring in multiple extracts were comparable, and allowed relative quantitation. This knowledge was applied to our investigation of malyngamide C, revealing that the PAL15AUG08-01 strain contained approximately two times more of this compound by crude extract mass than PAB10FEB10-01. Moreover, a known quantity of pure malyngamide C was analyzed by LC-MS/MS to produce the pure seed. The ratio of the ion counts for this seed and malyngamide C appearing in the extracts provided a basis for absolute quantitation. We calculated 0.10 mg of metabolite per mg of crude extract (10%, w/w) in PAL15AUG08-01, but only 0.19% (w/dw) as compared with 2.4% (w/dw) for PAB10FEB10-01.

In conclusion, LC-MS has become a preferred method for profiling crude NP extracts and derived fractions with its assets being: (a) minimal sample preparation, (b) speed, (c) robustness and (d) high information content. At the same time, it suffers from an excess of data that is laborious to analyze on a peak-by-peak basis. Hence, new automated routines are needed to provide initial interpretations by showing relationships between the data in a visually

clear manner, especially when evaluating multi-metabolomic data sets for patterns. In the current survey of the extracts of 20 cultured cyanobacteria, we used a seeded molecular network to dereplicate a variety of NPs. Relatively simple additions to the molecular networking workflow made the dereplication process more visually intuitive, and yielded a wealth of additional metabolomic information. This information included relative quantities of malyngamide C in the extracts of different cyanobacterial cultures, and with the authentic seed compound present in defined quantity, allowed us to determine absolute quantities as well. Results obtained using TOrTE-based molecular networks not only offer a more complete picture of cyanobacterial metabolomics, but also enabled us to identify our highest-yielding malyngamide C producers, one of whose genomes we are now sequencing as part of a biosynthetic investigation.

In continuing this work, we aspire to produce a high content molecular universe of marine cyanobacterial metabolites by profiling hundreds of crude extracts and pure compounds in our library. Thus, we will release TOrTE as a tool for the NP research community in the near future. With its qualitative color-coding feature, metabolomes can be compared with improved understanding of an organisms' evolution and chemotaxonomy. Quantitative color-coding visualization offers myriad applications, including chemical ecology studies, expression analyses, biosynthetic investigations and culture management tasks, such as optimization of growth conditions, selection of producer organisms and scaling of cultures to produce desired quantities of a metabolite. A cyanobacterial molecular universe, using a robust algorithm for consensus generation and integrated with databases such as AntiMarin,[25] will allow dereplication of known compounds, identification of promising analogs, and discovery of novel molecules within days of obtaining samples from the field.

METHODS

Cyanobacteria were hand-collected in various tropical waters (see Table 1) at depths from 0.3–15 m with the aid of snorkel or SCUBA. Chemistry samples were preserved in 1:1 seawater/EtOH and frozen at −20 °C. Live samples were brought back to the laboratory in vented tissue culture flasks with 0.2 μm-filtered seawater and subsequently cultured in SWBG-11 media with 35 g l^{-1} Instant Ocean (United Pet Group, Cincinnati, OH, USA). The cultures were kept at 28 °C in a 16 h light/8 h dark cycle with a light intensity of ~7 μmol photon s^{-1} m^{-2} provided by 40 W cool white fluorescent lights. To produce crude extracts, cyanobacterial tissue samples were extracted up to five times in 2:1 CH_2Cl_2/MeOH, then dried *in vacuo*, resuspended at 10 mg ml^{-1} in pure CH_2Cl_2 and run through a 0.2-μm glass fiber syringe filter to eliminate particulates.

The filtrates were dried again, resuspended at 10 mg ml^{-1} in MeOH, and 0.020 ml of each were injected (no-waste mode) into a reverse-phase HPLC system using a Phenomenex Prodigy C_{18} column (3 μm × 100 mm × 4.60 mm) with a gradient of 30–100% acetonitrile (ACN) in water with 0.1% formic acid over 20 min, followed by a 10-min isocratic period at 100% ACN. Total solvent flow was held at 0.50 ml min^{-1}. Pure marine NP and NP analog samples, referred to as 'seeds,' were prepared similarly using >85% pure samples (by NMR analysis) from the in-house pure compound library diluted to 0.33 mg ml^{-1} in MeOH. An aliquot of 0.010 ml of each was injected into a Phenomenex Kinetex C_{18} column (5 μm × 100 mm × 4.60 mm) and subjected to a gradient of 30–99% ACN in 0.1% formic acid over 17 min, followed by an 3-min isocratic period at 99% ACN. Total solvent flow was held at 0.70 ml min^{-1}. All solvents were purchased as LC-MS grade.

The HPLC eluate was electrospray ionized (35 eV) and analyzed for positive ions using a Thermo-Finnigan LCQ Advantage ion trap mass spectrometer (Thermo-Finnigan, San Jose, CA, USA). MS/MS spectra were obtained in a data-dependent manner using collision induced dissociation (CID) at 35 eV.[26] LC-MS data files were converted from Thermo RAW to mzXML format using msconvert from the ProteoWizard suite (v3.0.4743).[27] Seed files were processed using an msconvert data filter that retained only MS/MS scans with precursor masses within 1 Da of the *m/z* ratio of the seed metabolite's most common ion (see Supplementary Table S6). All mzXML files were used to generate a molecular network with the Spectral Networks script suite.[5] A minimum cosine similarity score of 0.95 and a parent mass tolerance of ± 0.3 Da was specified for consensus spectrum generation, and nodes were networked with minimum cosine similarity score of 0.6, 6 matching peaks minimum and 10 connections per node maximum.

Once the basic molecular network had been generated, the new TOrTE utility was applied to the data set (Figure 4). TOrTE uses a list of consensus spectra generated with the network to ascertain the source mzXML files for each of the spectra contributing to each consensus spectrum. If operating in quantitative mode, the program opens each of these mzXML files and searches within it for the most intense precursor peak to a consensus-forming spectrum. The program then uses ProteoWizard to generate an XIC spanning a time window about the most intense precursor scan with width manually specified according to the typical elution profile. In this investigation, 48 s was used. The *m/z* range of the XIC is also a specified value about the node's parent mass. ± 1 Da was used here. Quantitation is achieved by calculating area under the XIC trace. This is done automatically for each file contributing to every node and the area values are stored in a large annotation table with one row for each node and a column for each MS/MS data file. When TOrTE runs in qualitative mode, each cell of the annotation table holds a Boolean value: 1 if a compound is present in a data file or 0 if it is not.

The network was loaded into Cytoscape 2.8.3 and both qualitative and quantitative TOrTE annotation tables were added as sets of node attributes. The nodeCharts plugin for Cytoscape[28,29] was used to visualize TOrTE output data with pie charts. See Figures 3 and 4 for the resultant images.

After the pie charts were used to identify PAL15AUG08-01 and PAB10-FEB10-01 as principal malyngamide C producers, the absolute concentrations of malyngamide C in the extract and tissue of these strains were calculated as follows:

Variable definitions:

$[s]_i \equiv$ seed, concentration injected (mg ml^{-1})
$[c]_i \equiv$ crude extract, concentration injected (mg ml^{-1})
$V_s \equiv$ seed, volume injected (ml)
$V_c \equiv$ crude extract, volume injected (ml)
$s_i \equiv$ seed, mass injected (mg)
$c_i \equiv$ crude extract, mass injected (mg)
$A_m \equiv$ metabolite, area under LC-MS peak from crude extract (ion-min)
$A_s \equiv$ seed, area under LC-MS peak from pure sample (ion-min)
$m_i \equiv$ metabolite, mass injected (mg)
$[m]_c \equiv$ metabolite, concentration in crude extract (mg mg^{-1})
$c_{tot} \equiv$ crude extract, total mass obtained (mg)
$m_{tot} \equiv$ metabolite, total mass obtained (mg)
$dw \equiv$ extracted cyanobacterial tissue, dw mass (mg)
$[m]_{dw} \equiv$ metabolite, concentration in non-water biomass (mg mg^{-1})

To obtain masses of injected material for seed and crude samples:

$$s_i = [s]_i \cdot V_s$$

$$c_i = [c]_i \cdot V_c$$

To obtain concentration of metabolite in crude extract [mg metabolite per mg crude]:

$$m_i = \frac{A_m}{A_s} \cdot S_i$$

$$[m]_c = \frac{m_i}{c_i}$$

To normalize the above value in terms of dry tissue mass [mg metabolite per mg dw]:

$$[m]_{dw} = \frac{m_{tot}}{dw + c_{tot}} = [m]_c \cdot \frac{c_{tot}}{dw + c_{tot}}$$

Crude extract mass was added to dw of the extracted tissue in order to represent total non-water biomass of the cyanobacterial tissue sample.

ACKNOWLEDGEMENTS

This work was supported by NIH CA100851, TW006634 and NS053398 (WG), and GM097509 and AI095125 (PCD). We thank T Byrum and S Desfor for help with maintaining the cyanobacterial cultures. We thank the countries of Curaçao, Indonesia, Jamaica, Papua New Guinea and Panama for permission to collect environmental samples of marine cyanobacteria reported in this study.

1 Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.* **75,** 311–335 (2012).
2 Jones, A. C., Gu, L., Sorrels, C. M., Sherman, D. H. & Gerwick, W. H. New tricks from ancient algae: natural products biosynthesis in marine cyanobacteria. *Curr. Opin. Chem. Biol.* **13,** 216–223 (2009).
3 Leikoski, N. *et al.* Genome mining expands the chemical diversity of the cyanobactin family to include highly modified linear peptides. *Chem. Biol.* **20,** 1033–1043 (2013).
4 Choi, H., Pereira, A. R. & Gerwick, W. H. *In Handbook of Marine Natural Products* (eds Fattorusso, E., Gerwick, W. H. & Taglialatela-Scafati, O.) 55–152 (Springer, Netherlands, Dordrecht, Netherlands, 2012).
5 Watrous, J. *et al.* Mass spectral molecular networking of living microbial colonies. *Proc. Natl Acad. Sci. USA* **109,** E1743–E1752 (2012).
6 Yang, Jane Y. *et al.* Molecular networking as a dereplication strategy. *J. Nat. Prod.* **76,** 1686–1699 (2013).
7 Nguyen, D. D. *et al.* MS/MS networking guided analysis of molecule and gene cluster families. *Proc. Natl Acad. Sci. USA* **110,** E2611–E2620 (2013).
8 Pettit, G. R. *et al.* The isolation and structure of a remarkable marine animal antineoplastic constituent: dolastatin 10. *J. Am. Chem. Soc.* **109,** 6883–6885 (1987).
9 Ainslie, R. D., Barchi, J. J., Kuniyoshi, M., Moore, R. E. & Mynderse, J. S. Structure of malyngamide C. *J. Org. Chem.* **50,** 2859–2862 (1985).
10 Orjala, J., Nagle, D. & Gerwick, W. H. Malyngamide H, an ichthyotoxic amide possessing a new carbon skeleton from the Caribbean cyanobacterium Lyngbya majuscula. *J. Nat. Prod.* **58,** 764–768 (1995).
11 Wu, M., Milligan, K. E. & Gerwick, W. H. Three new malyngamides from the marine cyanobacterium Lyngbya majuscula. *Tetrahedron* **53,** 15983–15990 (1997).
12 Mevers, E. *et al.* Cytotoxic veraguamides, alkynyl bromide-containing cyclic depsipeptides from the marine cyanobacterium cf. Oscillatoria margaritifera. *J. Nat. Prod.* **74,** 928–936 (2011).
13 Harrigan, G. G. *et al.* Isolation, structure determination, and biological activity of dolastatin 12 and lyngbyastatin 1 from Lyngbya majuscula/Schizothrix calcicola cyanobacterial assemblages. *J. Nat. Prod.* **61,** 1221–1225 (1998).
14 Pettit, G. R., Xu, J. P., Hogan, F. & Cerny, R. L. Isolation and structure of dolastatin 17. *Heterocycles* **47,** 491–496 (1998).
15 Orjala, J. & Gerwick, W. H. Barbamide, a chlorinated metabolite with molluscicidal activity from the Caribbean cyanobacterium Lyngbya majuscula. *J. Nat. Prod.* **59,** 427–430 (1996).
16 Gross, H., McPhail, K. L., Goeger, D. E., Valeriote, F. A. & Gerwick, W. H. Two cytotoxic stereoisomers of malyngamide C, 8-epi-malyngamide C and 8-O-acetyl-8-epi-malyngamide C, from the marine cyanobacterium Lyngbya majuscula. *Phytochem* **71,** 1729–1735 (2010).
17 Balunas, M. J. *et al.* Coibacins A-D, anti-leishmanial marine cyanobacterial polyketides with intriguing biosynthetic origins. *Org. Lett.* **14,** 3878–3881 (2012).
18 Linington, R. G. *et al.* Antimalarial peptides from marine cyanobacteria: isolation and structural elucidation of gallinamide A. *J. Nat. Prod.* **72,** 14–17 (2009).
19 Flatt, P. M. *et al.* Characterization of the initial enzymatic steps of barbamide biosynthesis. *J. Nat. Prod.* **69,** 938–944 (2006).
20 Engene, N. *et al.* Five chemically rich species of tropical marine cyanobacteria of the genus *Okeania* gen. nov. (Oscillatoriales, Cyanoprokaryota). *J. Phycol.* (e-pub ahead of print 21 October 2013; doi:10.1111/jpy.12115).
21 Engene, N., Coates, R. C. & Gerwick, W. H. 16S rRNA gene heterogeneity in the filamentous marine cyanobacterial genus Lyngbya. *J. Phycol.* **46,** 591–601 (2010).
22 Engene, N., Gunasekera, S. P., Gerwick, W. H. & Paul, V. J. Phylogenetic inferences reveal a large extent of novel biodiversity in chemically rich tropical marine cyanobacteria. *App. Environ. Microbiol.* **79,** 1882–1888 (2013).
23 Swan, B. K. *et al.* Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc. Natl Acad. Sci. USA* **110,** 11463–11468 (2013).
24 Du, L. *et al.* Biosynthesis of sphinganine-analog mycotoxins. *J. Ind. Microbiol. Biotechnol.* **35,** 455–464 (2008).
25 Blunt, J. W., Munro, M. H. G. & Laatsch, H. (eds) *AntiMarin Database* (University of Canterbury, Christchurch, New Zealand and University of Göttingen, Göttingen, Germany, 2013).
26 Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422,** 198–207 (2003).
27 Kessner, D., Chambers, M., Burke, R., Agus, D. & Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **24,** 2534–2536 (2008).
28 Satio, R. *et al.* A travel guide to Cytoscape plugins. *Nat. Methods* **9,** 1069–1076 (2012).
29 Morris, J. (2011) nodeCharts (Version 0.94) [Software]. Available from http://chianti.ucsd.edu/svn/csplugins/trunk/ucsf/scooter/nodeCharts/

Supplementary Information accompanies the paper on The Journal of Antibiotics website (http://www.nature.com/ja)