Title

The Moorena bouillonii and Alpheus frontalis Symbiosis: Patterns in Chemical Ecology, Chemogeography, and Genomics

Permalink

https://escholarship.org/uc/item/6cv0t2dw

Author

Leber, Christopher Avery

Publication Date

2020

Supplemental Material

https://escholarship.org/uc/item/6cv0t2dw#supplemental

UNIVERSITY OF CALIFORNIA SAN DIEGO

The *Moorena bouillonii* and *Alpheus frontalis* Symbiosis:

Patterns in Chemical Ecology, Chemogeography, and Genomics

A dissertation submitted in partial satisfaction of the requirements for the degree

Doctor of Philosophy

in

Marine Biology

by

Christopher Avery Leber

Committee in charge:

Professor William H. Gerwick, Chair
Professor Pieter Dorrestein
Professor Paul Jensen
Professor Stuart Sandin
Professor Jennifer Smith

2020

The dissertation of Christopher Avery Leber is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2020

Dedication

This dissertation is dedicated to my wife Grace Leber and our dog Heinz Leber. Thank you both for keeping me smiling and being the best part of every day.

Epigraph

I hope for your help to explore and protect the wild ocean in ways that will restore the health and, in so doing, secure hope for humankind. Health to the ocean means health for us.

– Sylvia Earle

Table of Contents

List of Abbreviations

BGCs: biosynthetic gene clusters

bp: base pairs

COSY: ^{1}H-^{1}H correlated spectroscopy

FBMN: feature-based molecular networking

FGFR: fibroblast growth factor receptor

GC-MS: gas chromatography – mass spectrometry

GNPS: Global Natural Products Social Molecular Networking

HMBC: ^{1}H-^{13}C heteronuclear multiple bond coherence

HPLC: high-performance liquid chromatography

HRESIMS: high resolution electrospray ionization mass spectrometry

HSQC: ^{1}H-^{13}C heteronuclear single quantum coherence

HSQMBC: long-range ^{1}H-^{13}C heteronuclear single quantum multiple bond coherence

kb: kilobase pairs

LC-MS: liquid chromatography – mass spectrometry

LC-MS/MS: liquid chromatography – tandem mass spectrometry

LHDH: latitudinal herbivory-defense hypothesis

LPS: lipopolysaccharide

MALDI: matrix-assisted laser desorption/ionization

Mb: megabase pairs

m/z: mass-to-charge ratio

NMR: nuclear magnetic resonance

NRPS: non-ribosomal peptide synthetase

PCA: principal component analysis

PKS: polyketide synthase

RiPPS: ribosomally synthesized and post-translationally modified peptides

SCUBA: self-contained underwater breathing apparatus

s.d.: standard deviation

s.e.: standard error

SPE: solid phase extraction

STAT: signal transducer and activator of transcription

TMD: transmembrane domain

TOCSY: HSQC-total correlation spectroscopy

TOF: time-of-flight, referring to the method of mass spectrometry

VEGF: vascular endothelial growth factor

List of Supplemental Videos

List of Figures

List of Tables

Acknowledgements

I would like to acknowledge Professor William H. Gerwick for his support as my PhD advisor and the chair of my committee. Thank you for fostering my curiosity, encouraging my explorations, and providing me with opportunities to do fascinating things.

I would also like to acknowledge past and present members of the Gerwick Lab. Thanks for sharing knowledge, challenges, laughs, and beverages with me over these past five years, and cheers to a future of friendship. Special thanks to Dr. rer. nat. Lena Keller and Dr. C. Benjamin Naman.

Chapter 1, in full, has been submitted for publication of the material as it may appear in Leber, Christopher A.; Reyes, Andres Joshua; Biggs, Jason S.; Gerwick, William H. "Cyanobacteria-Shrimp Colonies in the Mariana Islands". The dissertation author was the primary investigator and author of this paper. The dissertation author co-conceived of the work, designed and implemented the surveys and experiments, participated in collections, performed all data analyses, and was the primary author of the work.

Chapter 2, in full, is a reprint of the material as it appears in Leber, Christopher A.; Naman, C. Benjamin; Keller, Lena; Almaliti, Jehad ; Caro-Diaz, Eduardo J. E.; Glukhov, Evgenia; Joseph, Valsamma; Sajeevan, T. P.; Reyes, Andres Joshua; Biggs, Jason S.; Li, Te; Yuan, Ye; He, Shan; Yan, Xiaojun; Gerwick, William H. "Applying a Chemogeographic Strategy for Natural Product Discovery from the Marine Cyanobacterium *Moorena bouillonii*", Marine Drugs, vol. 18, 2020. The dissertation author was the primary investigator and author of this paper. The dissertation author co-conceived of the work, acquired and prepared necessary samples, collected mass spectral and NMR data, wrote code and performed all analyses, led structure elucidation efforts, conducted or otherwise participated in synthetic reactions for determination of configuration, implemented *in silico* antibacterial screening, and was the primary author of this work.

Chapter 3 is coauthored with Naman, C. Benjamin; Aksenov, Alexander A.; Reyes, Andres Joshua; Glukhov, Evgenia; Dorrestein, Pieter C.; Biggs, Jason S.; Gerwick, William H. The dissertation author was the primary author of this chapter. The dissertation author co-conceived of the work, participated in cyanobacteria and shrimp collections, designed and implemented all experiments, collected LC-MS data and prepared samples for comparative GC-MS, conducted data analyses, and was the primary author of this work.

Chapter 4 is coauthored with Caraballo-Rodríguez, Andrés Mauricio; Naman, C. Benjamin; Glukhov, Evgenia; Reyes, Andres Joshua; Joseph, Valsamma; Sajeevan, T. P.; Li, Te; Yuan, Ye; He, Shan; Yan, Xiaojun; Miller, Bailey W.; Thornburg, Christopher C.; McPhail, Kerry L.; Schmidt, Eric W.; Haygood, Margo G.; Dorrestein, Pieter C.; Biggs, Jason S.; Gerwick, William H. The dissertation author was the primary author of this chapter. The dissertation author co-conceived of this work, participated in cyanobacterial collections on Guam and Lakshadweep, acquired and prepared all samples for LC-MS/MS data collection, wrote code and conducted all analyses, and was the primary author of this work.

Chapter 5 is coauthored with Leão, Tiago; Podell, Sheila; Moss, Nathan A.; Whitner, Syrena; Glukhov, Evgenia; Sanders, Jon G.; Reyes, Andres Joshua; Biggs, Jason S.; Humphrey, Gregory; Zhu, Qiyun; Belda-Ferre, Pedro; Allen, Eric E.; Knight, Rob; Gerwick, Lena; Gerwick, William H.. The dissertation author was the primary author of this chapter. The dissertation author co-conceived of the work, performed a portion of the necessary DNA extractions and sample preparations for sequencing, wrote and implemented code for the genome assembly module and the draft genome improvement module, conducted analyses of results, and was the primary author of this work.

In addition, I would like to acknowledge T. Deerinck, M. H. Ellisman, and the National Center for Microscopy and Imaging Research at UCSD for facilitating the acquisition of scanning electron and transmission electron micrographs of woven *Moorena bouillonii*; P. Boudreau (University of Mississippi), L. T. Tan (Nanyang Technological University, Singapore), A. C. Jones (Gordon and Betty Moore Foundation), D. H. Sherman (University of Michigan), D. J. Edwards (California State University, Chico), A. Pereira (Gilead Sciences), N. Moss (Arzeda), L. Simmons (Vectura Group), R. V. Wilson (Biogen), B. Miller (University of Utah), P. Flatt (Western Oregon University), N. Engene (Vattenmyndigheten Västerhavet), and R. C. Coates (Joint BioEnergy Institute) for their contributions to the *M. bouillonii* data aggregation questionnaire; N. Moss previously of Scripps Institution of Oceanography, UCSD and A. Townsend, D. Drake, and T. Mesa of University of Guam Marine Laboratory for assisting with in-field surveys; N. Moss and T. Leão of UCSD for guidance in genomic DNA extraction and 16S rRNA sequencing; A. B. Strum of Florida Atlantic University, Harbor Branch Oceanographic Institute, for culture experiment monitoring support; A. Duenas, J. Miller, J. Peralta, and J. Cummings of University of Guam Marine Laboratory for field and logistical support; S. Becker, M. Miller, J. Ballard, and the Oceanographic Data Facility at Scripps Institution of Oceanography for nutrient analyses; F.M. Brunner and C.P. Kubiak (UCSD Department of Chemistry & Biochemistry) for facilitating acquisition of IR data; Y. Su, and L. Gross for HiRes mass spectrometry support at the UCSD Molecular Mass Spectrometry Facility; B. Duggan and A. Mrse for NMR support; M.P. Christy (Scripps Institution of Oceanography, UCSD) for support with the synthesis of standards; the

government of Sansha city, China, for permission to collect and study several of the marine samples used in this research; sample collection assistance from L. Zhenhua, L. Daning and H. Da of the Xisha Marine Science Comprehensive Experimental Station, South China Sea Institute of Oceanology, Chinese Academy of Sciences; the support of I.S. Bright Singh of the National Centre for Aquatic Animal Health, Cochin University of Science and Technology, India; the Department of Science and Technology, Kavaratti, Lakshadweep Islands, India for the required research permits; Diaa T.A. Youssef for facilitation and the Red Sea Protectorate for permission to collect Red Sea cyanobacteria; and SINU for help with access to cyanobacteria from the Solomon Islands.

Vita

2015	Bachelor of Science, University of California, Los Angeles
2015-2020	Graduate Student Researcher, University of California, San Diego
2016-2018	NIH Training Program, Marine Biotechnology
2018	Master of Science, University of California, San Diego
2020	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Leber CA et al (2020) Applying a Chemogeographic Strategy for Natural Product Discovery from the Marine Cyanobacterium *Moorena bouillonii*. Mar Drugs 18:515.

Leber CA, Reyes AJ, Biggs JS, Gerwick WH (2020) Cyanobacteria-Shrimp Colonies in the Mariana Islands. *Under Review*.

Maloney KN, [et al, including **Leber CA**] (2020) Cryptic Species Account for the Seemingly Idiosyncratic Secondary Metabolism of *Sarcophyton glaucum* Specimens Collected in Palau. J Nat Prod 83:693-705.

Kim HW, [et al, including **Leber CA**] (2020) NPClassifier: A deep neural network-based structural classification tool for natural products. *Under Review*.

Naman CB, **Leber CA**, Gerwick WH (2017) Chapter 5 - Modern Natural Products Drug Discovery and Its Relevance to Biodiversity Conservation. In: Kurtböke I (ed.) *Microbial Resources*. Academic Press, pp 103-120

Petras D, [et al, including **Leber CA**] (2016) Mass Spectrometry-Based Visualization of Molecules Associated with Human Habitats. Anal Chem 88:10775–10784.

FIELDS OF STUDY

Major Field: Marine Chemical Biology

Studies in Marine Natural Products and Chemical Ecology

Professor William H. Gerwick

ABSTRACT OF THE DISSERTATION

The *Moorena bouillonii* and *Alpheus frontalis* Symbiosis:

Patterns in Chemical Ecology, Chemogeography, and Genomics

by

Christopher Avery Leber

Doctor of Philosophy in Marine Biology

University of California San Diego, 2020

Professor William H. Gerwick, Chair

The tropical marine benthic filamentous cyanobacterium *Moorena bouillonii* and the snapping shrimp *Alpheus frontalis* engage in a charismatic symbiosis, in which the shrimp weaves the cyanobacterial filaments into tubes and hollow chambers in and around coral reefs. Pairs of *A. frontalis* live in and feed upon these *M. bouillonii* structures. *M. bouillonii* produces a vast arsenal of biologically active and structurally diverse natural products, suggesting that *A. frontalis* may benefit from an associational defense. It is unclear, however, how *M. bouillonii*

may benefit from this partnership, and what impact the symbiosis may have on the surrounding reef. Also unknown is what role chemistry plays in directing the interactions between the shrimp and cyanobacterium. Furthermore, the breadth of chemodiversity currently known from *M. bouillonii* that originates from a small collection of discrete locations, paired with the broad distribution of the *M. bouillonii – A. frontalis* symbiosis across the Indian Ocean and the western tropical Pacific raises the question of what more chemical diversity there is to be discovered in relation to this association.

Field surveys and photographic documentation were employed to better understand how *M. bouillonii – A. frontalis* colonies are distributed across scleractinian coral reefs in the Mariana Islands, and revealed that colonies tend to inhabit interstices in the reef. Nutrient analysis of paired water samples and culturing experiments suggested that *A. frontalis* may benefit *M. bouillonii* via an enriched nutrient environment. Competing hypotheses on why shrimp are found in association with certain chemotypes of the cyanobacterium propelled a series of behavioral assays and comparative gas chromatography – mass spectrometry experiments to ultimately yield a lead compound for a shrimp attractant waterborne chemical cue produced by *M. bouillonii*. Chemogeographic investigations utilizing liquid chromatography – tandem mass spectrometry facilitated the discovery of a regionally specific family of natural products, while also improving understanding of how *M. bouillonii* chemodiversity is distributed across geographic space and generating paths forward for future discovery. Development of a new bioinformatic tool provided substantial improvements to *M. bouillonii* genome contiguity, allowing for new insights into the concentrations of potential biosynthetic diversity across compound classes.

Introduction

Sinking down below the ocean's surface and opening one's eyes to a tropical coral reef, one is immediately engulfed in a kaleidoscope of color and texture. Fishes of various palettes pulse over a chaotic landscape of hard coral often interrupted by bursts of algae and sprawling sponges. Amidst this captivating display, one might not notice the deep red streaks and blobs winding through the cracks and crevices in the reef. A closer inspection reveals intricate networks of living tubes composed of thread-like filaments of algae (Figure 0.1). Too close of an inspection might result in a startling introduction to the inhabitants and architects of these tubes, a species of snapping shrimp called *Alpheus frontalis* (Figure 0.2). *A. frontalis*, like other species of snapping or pistol shrimp, is equipped with one larger claw and one smaller claw. The particular shape of the larger claw, along with the force with which it can be closed, allows for the generation of a small bubble of vapor, known as a cavitation bubble (Versluis et al 2000). The bubble's formation is followed shortly by a shockwave and loud pop as the bubble implodes. Like other species of snapping shrimp, *A. frontalis* uses its large snapping claw for hunting prey, aggressing against competitors, and defending against predators. Just below the snapping claws, a more delicate set of appendages is used for a much more charming task: weaving.

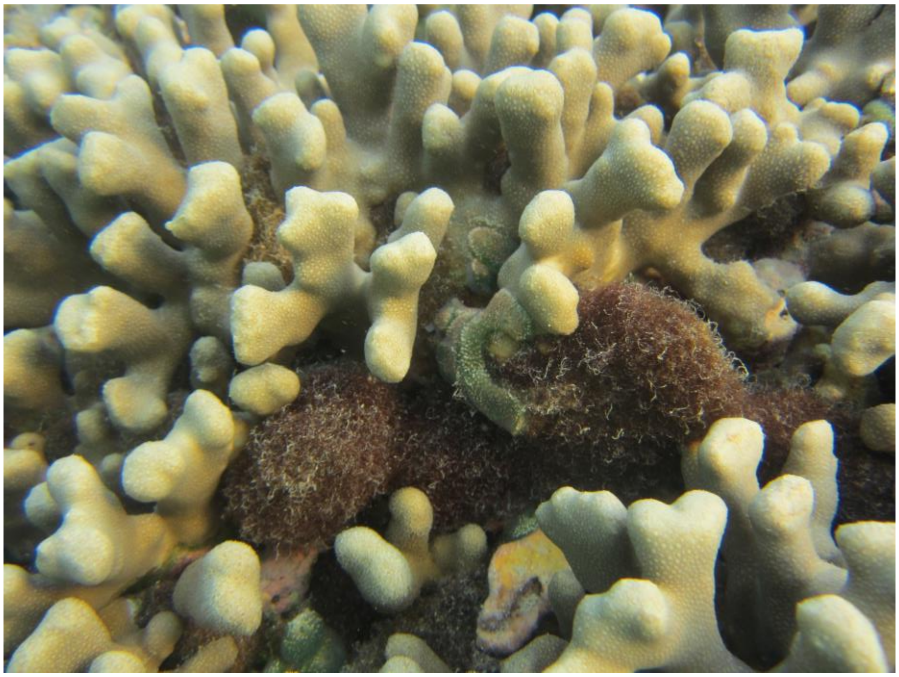


Figure 0.1 – A tube constructed of filaments of the cyanobacterium *Moorena bouillonii*, winding amongst branches of coral.

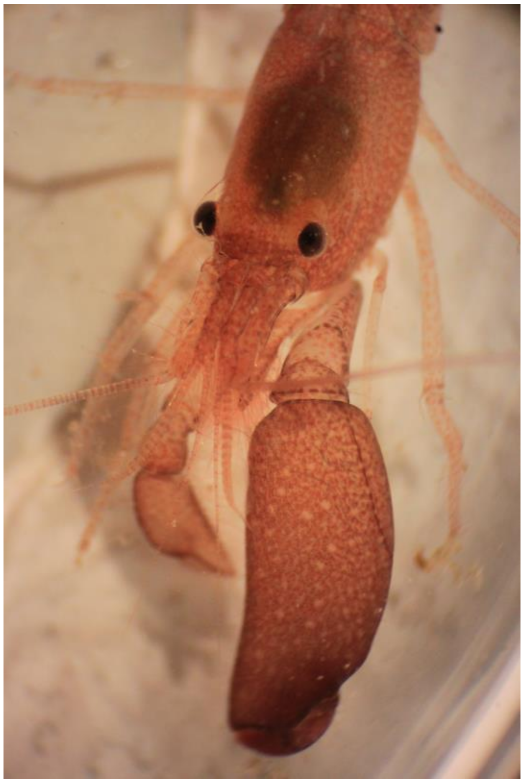


Figure 0.2 – *Alpheus frontalis*, a snapping shrimp that lives in woven tubes of *M. bouillonii*.

Mated pairs of *A. frontalis* take filaments of *Moorena bouillonii*, a blue-green algae (cyanobacterium), and aggregate them in such a way as can only be described as weaving, in order to construct tunnels and chambers that emerge from holes and extend deep into the reef. The filaments are woven tightly together, so tightly that the tubes have the appearance of holding a volume of water, and they are in many cases incredibly clean, as the shrimp meticulously maintain their woven domiciles. The filaments look to have been assembled in a somewhat regular and uniform fashion, but microscopic viewing reveals an overwhelming complex of loops and bundles of cyanobacteria (Figure 0.3). Hidden within the individual filaments of *M. bouillonii*, beyond the visible, is something perhaps more fantastic than the weaving shrimp: a unique bouquet of specialized chemistry, known as natural products or secondary metabolites.

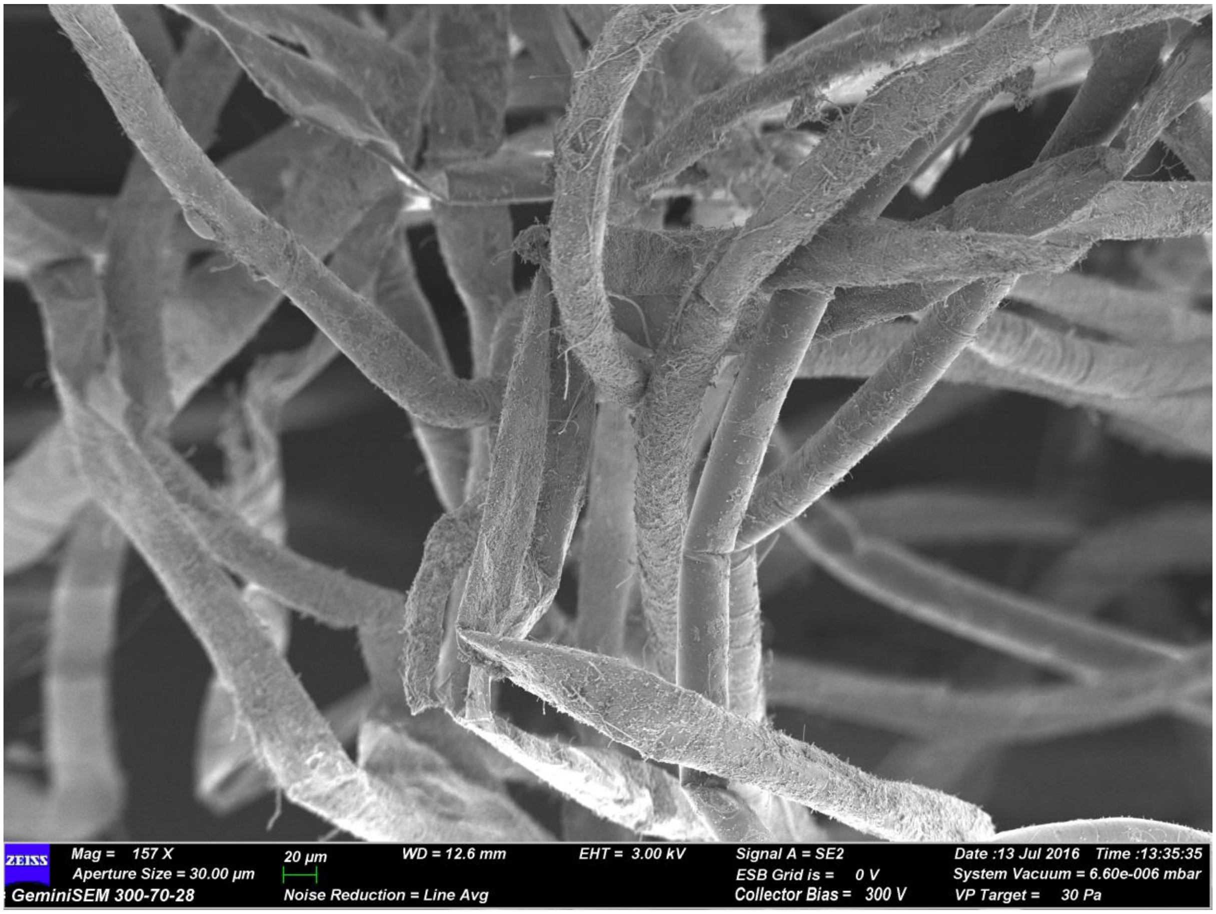


Figure 0.3 – Scanning electron micrograph of a bundle of *M. bouillonii* filaments, as woven by *A. frontalis*.

Natural products are auxiliary compounds produced by living organisms that are not necessary for basic biological functioning but provide benefits to their producers that improve survival and/or reproductive success (Dixon 2001). There are many different roles that the natural products produced by *M. bouillonii* may be filling, from protecting the cyanobacterium (and its shrimp residents) against potential predators, to acting as chemical smoke signals that allow the shrimp to locate their building materials of choice. Some have suggested that these chemicals are facilitating the infection and take-over of coral reefs by *M. bouillonii*, and while evidence for such an assertion is scant, it is true that these compounds are undoubtedly impacting how organisms are interacting with each other on the reef. Far away from the reef, in research laboratories around the world, natural products discovered from *M. bouillonii* have been shown to act upon different signaling pathways and to kill various types of cancer cells, with efforts ongoing to turn one family of cell-killing compounds into therapeutics to treat pancreatic cancer.

0.1: A Brief Early History

The *M. bouillonii – A. frontalis* symbiosis has been a part of collective human knowledge for well over a century. *A. frontalis* was first described in 1837 by Milne-Edwards (Milne-Edwards 1837) after its collection off of the coast of Australia (then known as New Holland), by "Messrs. Quoy and Gaimard." Milne-Edwards, Quoy, and Gaimard were prominent French naturalist-zoologists at the time and were involved with numerous expeditions, and so their names are linked with myriad different species. The initial description of *A. frontalis* was purely anatomical, and so it includes little that significantly distinguishes it from many other snapping shrimp species. *A. frontalis* was made much more distinctive in 1880 (Richters 1880), when two of the shrimp were collected from Mauritius in woven tubes of cyanobacteria that were mis-identified as *Oscillatoria* sp.; this represents the scientific literature's earliest introduction to the *M. bouillonii – A. frontalis* symbiosis.

The weaving behavior exhibited by *A. frontalis* was first described in 1913 by Cowles (Cowles 1913). He recounts cutting open a tube of cyanobacteria in a dish full of seawater, placing a pair of shrimp in the dish, and then observing as the male shrimp used its second thoracic limbs (also referred to as the small chelae) to thread cyanobacterial filaments through the preexisting woven cyanobacterial structure and stitch the tube back up, into which both of the shrimp retreated. He noted the strategy by which the shrimp conducted such a repair, first forming a tube by attaching the edge of the woven cyanobacterial mat at multiple points, and then stitching together any remaining gaps. Cowles also described placing a pair of shrimp in a container with seawater and a rock to hide under, and providing them with shreds of what used to be a cyanobacterial tube. Over the course of some hours, the shrimp had securely attached the cyanobacterial filaments to the underside of the rock and reconstructed their filamentous domicile. A particularly charming sentence from Cowles' manuscript reads:

> "When the alpheus found a hole in the rapidly forming tube, the slender legs came through, caught hold of the filaments of the alga, and manipulated them in much the same manner as a man might the thread with which he darns a hole in his sock; that is, by drawing the edges of the hole together and fastening them."

There is disagreement as to what species of shrimp Cowles had observed. Cowles reported the species as *A. pachychirus*, a congener to *A. frontalis* that is also known to weave cyanobacterial filaments and has been studied in even less detail, but he also references the report of Richters (1880) who indicated his finding of shrimp in algal tubes to be of the species *A. frontalis*. Fishelson (1966) suggests that Cowles misidentified *A. frontalis* as *A. pachychirus,* while Banner and Banner (1982) casts doubt on Fishelson's assessment, but neither provide much substantive evidence to back their claims, and Cowles did not report any anatomical observations about the shrimp that could otherwise be used to definitively identify one species over the other. Cowles does, however, describe and depict the woven cyanobacterial colony as being found under a rock and consisting of a larger chamber from which more narrow tubes of 20 mm diameter extend; this colony description is much more aligned with how *A. frontalis* colonies are characterized (Fishelson 1966; Banner and Banner 1982). The admittedly less well described colonies of *A. pachychirus* are noted by Banner and Banner (1982) as being composed of narrower tubes and occurring "…between the fronds of dead racemous coral." Regardless of which *Alpheus* species Cowles observed in his zoological laboratory at the

University of the Philippines, his manuscript is notable as being the first publication to document that the alpheid shrimp living in algal tubes are, in fact, constructing those tubes with a weaving technique.

In 1966, Fishelson built upon the work of Cowles by observing and conducting experiments upon *A. frontalis* collected from the Red Sea (Fishelson 1966). In addition to corroborating Cowles' descriptions of shrimp weaving, Fishelson expanded understanding of how *A. frontalis* interacts with its cyanobacterial associate in a number of different ways. Cowles had only reported male shrimp as participating in cyanobacterial tube construction, but Fishelson observed both male and female shrimp working together to repair and maintain their cyanobacterial abodes. Similarly, Cowles indicated that shrimp used their second thoracic limbs to manipulate and weave the cyanobacterial filaments, while Fishelson observed the use of these appendages combined with assistance from the second and third maxillipeds (mouth parts) in the construction of the cyanobacterial tubes. Fishelson conducted small-scale experiments in which he separated shrimp from their cyanobacterial shelters and placed them 20 cm away. Shrimp responded by rapidly swimming around and manipulating their antennae until they relocated and reentered their cyanobacterial tubes, on the time scale of 5 minutes. This suggested to Fishelson that the shrimp have the capacity to recognize their cyanobacterial tubes. When not given access to their preferred cyanobacterial building material, but provided *Enteromorpha* sp., shrimp constructed tubes out of the green algal filaments; if the preferred cyanobacterial filaments were reintroduced, they were incorporated into the woven structure. Shrimp were unable to construct tubes out of large expanses of *Ulva* sp., a non-filamentous green alga, but did incorporate small pieces of the alga into structures otherwise woven out of algal filaments. Through observation of shrimp behavior within their tubes and inspection of shrimp excrement, which contained undigested cyanobacterial cells, Fishelson was able to determine that the shrimp not only construct and take shelter within their woven cyanobacterial colonies, they also feed upon them.

Up until 1982, there is no indication in the scientific literature that the woven cyanobacterial home of *A. frontalis* was thought to have function beyond providing shrimp a place to hide, an established territory, and something to feed upon. That is, until Banner and Banner (1982) proposed that the cyanobacterial building material was likely to be toxic, and so may be deterrent to potential predators. Banner and Banner based this hypothesis upon mounting evidence for the toxicity of the marine cyanobacterium *Lyngbya majuscula*, including a 1971 report that indicated it can cause acute dermatitis in humans (Moikeha, Chu, and Berger 1971). The inflammatory response was instigated via subcutaneous injection, which is an unlikely exposure route for many humans who otherwise may encounter the cyanobacterium. Nevertheless, this added to a growing body of evidence that suggested toxic and antibacterial activities associated with *Lyngbya majuscula*, which is one of the identifications that Banner and Banner ascribed to the cyanobacterium used by *A. frontalis*.

Less than a decade later, Hoffmann and Demoulin established this filamentous cyanobacterium to be a new species in the genus *Lyngbya*, giving it the name *L. bouillonii* (1991). While there is no mention of alpheid shrimp in this latter manuscript, the species is described as forming ‘tenacious mats’ that are firmly attached to the benthos, amongst corals and filling holes in the reef. This ‘tenacity’ is likely the result of the cyanobacterial filaments being stitched together by their shrimp associates, and figure 1 in the manuscript displays a tube of cyanobacterial filaments that have unmistakably been woven by shrimp. Hoffmann and

Demoulin offered a foreshadowing of countless hours of scientific labor to come when they suggested that *L. bouillonii* produces a “complex mixture of toxins.”

0.2 Natural Products Chemistry of *Moorena bouillonii*

In the years since Hoffmann and Demoulin first established a (relatively) stable name for the species, samples of *Moorena bouillonii* (Tronholm and Engene 2019), formerly *Lyngbya bouillonii* (Hoffmann and Demoulin 1991), and then *Moorea bouillonii* (Engene et al 2012), have indeed yielded “complex mixtures of toxins,” and numerous other chemical components. As of the writing of this dissertation (Fall 2020), there are 70 new secondary metabolites that have been characterized from *M. bouillonii* extracts (Leber et al 2020). These include the apramides (Luesch et al 2000) and lyngbyapeptins (e.g. Klein et al 1999), families of lipopeptides; the apratoxins (Luesch et al 2001; Luesch et al 2002c; Gutiérrez M et al 2008; Matthew, Schupp, Luesch 2008; Tidgewell et al 2010; Thornburg et al 2013), lyngbyabellins (Matthew et al 2010; Choi et al 2012), and ulongamides (Luesch et al 2002a), families of cyclic depsipeptides; the kakeromamides (Nakamura et al 2018; Sweeney-Jones et al 2020), a pair of cyclic peptides; and the columbamides (Kleigrewe et al 2015; Lopez et al 2017), a family of fatty acid amides (Figure 0.4). These chemical families suggest a tendency in *M. bouillonii* for non-ribosomal peptide synthetase (NRPS) biosynthetic machinery and hybrid polyketide synthase (PKS) -NRPS biosynthetic machinery, with some level of promiscuity that allows for the generation of various analogs. However, this fails to capture the breadth of biosynthetic ability contained within *M. bouillonii*. The only high-quality genome of *M. bouillonii* that has thus far been published (Leão et al 2017) was found to possess 31 biosynthetic gene clusters (BGCs), of which 8 were annotated as NRPS and only 4 were annotated as hybrid PKS-NRPS;

3 BGCs were annotated as PKS, 11 as ribosomally synthesized and post-translationally modified peptides (RiPPs), 3 as terpenes, and 2 as ‘other’. Macrolides such as the lyngbyalosides (Klein et al 1997; Luesch et al 2002b) and the lipid mooreamide A (Mevers et al 2014) are potential end products for some of *M. bouillonii*’s PKS BGCs. However, examples of RiPPs and terpene natural products from *M. bouillonii* are lacking, and so represent still untapped sources of potential chemodiversity that reside in this organism.

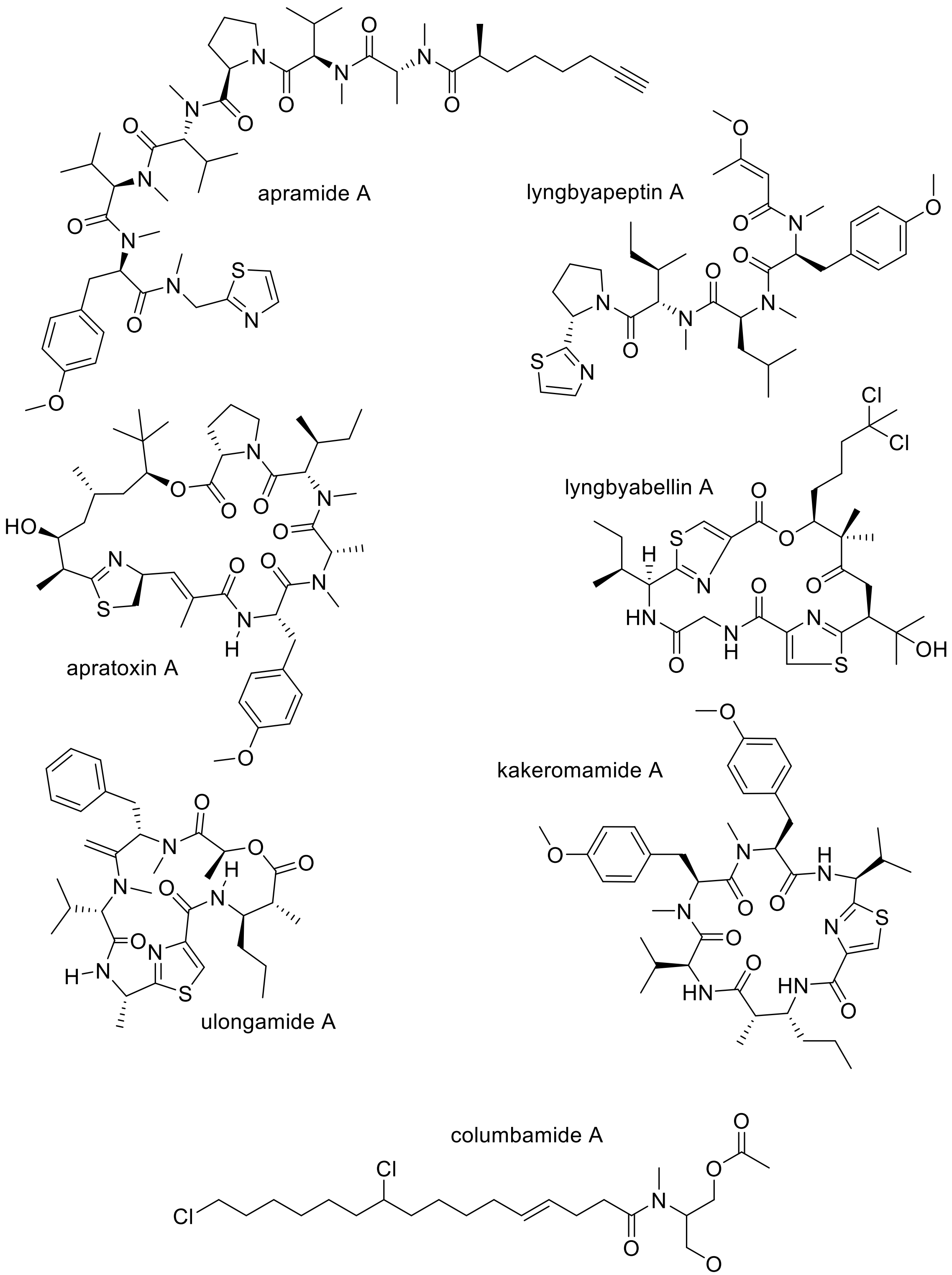


Figure 0.4 – Representative structures of families of natural products produced by *M. bouillonii*: The apramides, the lyngbyapeptins, the apratoxins, the kakeromamides, the ulongamides, and the columbamides.

The diverse set of compounds that have been characterized from *M. bouillonii* is associated with a range of different pharmaceutically relevant biological activities. Two general themes that can be ascribed to the varied biological activities thus far established are cellular signaling and, just as Banner and Banner (1982), and Hoffmann and Demoulin (1991) foretold, cytotoxicity. Alotamide A was found to increase both the concentration and the frequency of calcium (Ca^{2+}) oscillations in murine cerebrocortical neurons (Soria-Mercado et al 2009). Mooreamide A displayed cannabimimetic activity, with greater than 50x binding selectivity for cannabinoid receptor CB(1) versus CB(2) (Mevers et al 2014). Columbamides A and B have also been found to possess potent cannabinoid receptor binding properties, though with much less specificity for CB(1) over CB(2) of only about 2-fold (Kleigrewe et al 2015). Perhaps most intriguing of the signaling activities found to be driven by *M. bouillonii* natural products is the reported ability of kakeromamide A to trigger the differentiation of neural stem cells to astrocytes, while inhibiting differentiation into neurons (Nakamura et al 2018). Regarding toxicity, bouillonamide (Tan, Okino, and Gerwick 2013), lyngbouilloside (Tan, Márquez, and Gerwick 2002), and multiple lyngbyabellins (Matthew et al 2010; Choi et al 2012) are reported to have cytotoxic activities. The most notable of the toxins characterized from *M. bouillonii*, not only due to the potency of their toxicity, but also the selectivity of that toxicity, are the apratoxins.

0.3: Apratoxins and the very real potential for spectacular drugs from the sea

The apratoxins are a family of cytotoxic cyclic depsipeptides currently composed of nine different natural analogs, apratoxins A-H and apratoxin A sulfoxide (Luesch et al 2001; Luesch et al 2002c; Gutiérrez M et al 2008; Matthew, Schupp, Luesch 2008; Tidgewell et al

2010; Thornburg et al 2013). The apratoxin BGC was described in 2011 (Grindberg et al 2011), revealing a type I modular mixed PKS-NRPS framework. Variation between the analogs occurs in various flavors, including substitution of amino acid residues, alterations in methylation and PKS tailoring, and differences in the number of PKS modules (Tidgewell et al 2010; Thornburg et al 2013). Commensurate with this structural diversity, a range of potencies have been recorded in cytotoxicity associated with these different analogs, ranging from apratoxin A sulfoxide's respectable IC_{50} 89.9 nM against human NCI-H460 lung cancer cells to the exquisitely potent low single digit nanomolar IC_{50}'s of apratoxins A, G, and H (Tidgewell et al 2010; Thornburg et al 2013).

Having multiple analogs in a compound family that possess single digit nanomolar IC_{50} cytotoxicity values against a human cancer cell line is bound to capture interest, but it is not sufficient for showcasing promise as a therapeutic, namely due to issues of selectivity. In other words, a compound might be very effective at killing cancer cells, but if it is also effective at killing healthy cells, it is unlikely to provide enough therapeutic benefit to warrant its use as a chemotherapeutic. Indeed, when apratoxin A was first discovered, although it yielded sub-nanomolar IC_{50} values in vitro against two different cancer cell lines, it was found to be ineffective in vivo in treating a breast cancer model in mice (Luesch et al 2001). Additionally, while apratoxin A was observed to have some effect in treating a colon cancer model in mice, it caused issues with weight loss in the test animals (Luesch et al 2001). Histopathological investigations that followed indicated that apratoxin A was localized in the pancreatic and salivary gland cells of exposed mice and resulted in severe apoptotic cell losses in the pancreas (Akare et al 2015; Huang et al 2016). Orthogonal approaches in studying the cellular impacts of apratoxin A revealed a differential cancer cell toxicity profile indicative of a novel mechanism of action. Apratoxin A was found to induce G1-phase cell cycle arrest in sensitive cell lines, and to disrupt fibroblast growth factor receptor (FGFR) signaling via the ultimate inhibition of the phosphorylation of the transcription factor signal transducer and activator of transcription (STAT) 3 (Luesch et al 2006; Ma et al 2006). Taken together, these results suggested the potential to use apratoxin A in combatting pancreatic cancers dependent on STAT3-involved FGFR signaling pathways, but a mechanism of action remained unclear.

Following up on the impacts of apratoxin A on STAT3 and FGFR signaling indicated that apratoxin A disrupts multiple cancer-related signaling cascades, without specifically acting upon relevant receptors (Liu, Law, and Luesch 2009). Instead, apratoxin A was found to prevent the cotranslational translocation of newly synthesized secretory pathway proteins into the endoplasmic reticulum. This generated the hypothesis, that was ultimately confirmed (Paatero et al 2016; Huang et al 2016), that apratoxin A binds to the alpha subunit of Sec61. In binding to Sec61, apratoxin A prevents the transmembrane domain (TMD) of nascent proteins from interacting with the Sec61 lateral gate that would otherwise initiate translocation through the channel into the endoplasmic reticulum.

The establishment of a mechanism of action by which apratoxin A exerts its selective pancreatic toxicity represents a necessary advancement for the earnest development of apratoxin-inspired therapeutics. The unique aspects of apratoxin biology have caused drug-developing labs to focus in on two potential therapeutic uses: the treatment of pancreatic cancer, and the treatment of angiogenesis-related cancers and diseases. Synthetic efforts to design new apratoxin analogs resulted in the generation of a structural hybrid of apratoxins A and E (referred to as apratoxin S4) that was found to be more selective for cancer cells, and less irreversibly toxic, in mice with a human colon cancer model (Chen, Liu, and Luesch 2011).

Additional synthetic iterations have continued to produce therapeutically-improved apratoxins (Chen et al 2014; Cai et al 2017), ultimately producing apratoxin S10, which has been shown to not only disrupt pancreatic cancer cells and their secretions, but also to act upon stromal cells present in the tumor microenvironment (Cai et al 2019). A major challenge in treating pancreatic cancers is the general inability of therapeutic agents to infiltrate the abnormal structural microenvironment within pancreatic tumors, so apratoxin S10's activity against multiple cell types within the tumor tissue is very promising.

In addition to FGFR signaling, apratoxin A's disruption of the secretory pathway via Sec61 has implications for other signaling cascades, including those associated with angiogenesis (the generation of new blood vessels extending from older vessels); one of the signaling cascades involved with angiogenesis is vascular endothelial growth factor (VEGF) signaling (Luesch et al 2006; Cai et al 2017). Apratoxin S10 has been shown to effectively inhibit angiogenesis (Cai et al 2017), suggesting an orthogonal and complementary method by which apratoxin therapeutics could be used to combat cancers that depend on vascularization for proliferation. The antiangiogenic activity associated with apratoxins has been further shown to have therapeutic utility beyond cancer, with apratoxin S4 having been successfully used with ex vivo models and in vivo in selectively treating ocular diseases that involve pathological vascular growth (Qiu et al 2019).

0.4: Ecology

In sharp contrast to the depth and breadth of scholarship that has been orchestrated around the pharmaceutically relevant natural products chemistry of *M. bouillonii*, there has been relatively little research aimed at better understanding the chemical ecology and general ecology of *M. bouillonii* and its symbiont *A. frontalis*. Beyond assays in toxicity and neuromodulatory activities, which indirectly and obliquely suggest how natural products from *M. bouillonii* may impact other organisms in the ecosystem, *M. bouillonii* natural products have also been tested for their antimolluscal (Pereira, McCue, and Gerwick 2010) and quorum modulating (Liang et al 2019) activities. These results were derived from pharmacologically focused models, so while they also may suggest how *M. bouillonii* metabolites are impacting other coral reef organisms, they are not directly ecologically relevant. Despite the common assertion that *M. bouillonii* natural products are acting as feeding deterrents, to date there are no published studies that substantiate this claim. This idea is likely an extrapolation from *M. bouillonii*'s congener *M. producens* (previously *Lyngbya majuscula* and *L. sordida*) (Engene et al 2012).

The feeding deterrent characteristics of *M. producens* have been extensively studied. Unlike *M. bouillonii*, *M. producens* is notable for its bloom formation, and it has been documented to be a part of at least one mixed assemblage algal bloom that resulted in a large fish die-off (Nagle and Paul 1998). Crude extracts of the cyanobacterium have been incorporated into food pellets and offered to a large array of different consumers, provoking feeding deterrence from various fishes (Paul and Pennings 1991; Nagle and Paul 1998; Nagle and Paul 1999; Cruz-Rivera and Paul 2002; Capper et al 2006;), sea urchins (Nagle and Paul 1998; Cruz-Rivera and Paul 2002; Capper et al 2006), amphipods (Capper et al 2006), crabs, (Capper et al 2006) and cephalaspideans (bubble snails) (Capper et al 2006). In several cases, researchers have successfully isolated and characterized individual compounds from these crude extracts to which at least some of the antifeedant activity can be attributed. From a mixed bloom of *Schizothrix calcicola* (another filamentous cyanobacterium) and *M. producens* (reported as *L. majuscula*) found on Ypao beach on Guam, ypaoamide, likely being produced by the *M. producens* filaments, was shown to deter the feeding of juvenile *Siganus spinus* and *S. argenteus* (rabbitfishes), *Scarus schlegeli* (a parrotfish), and *Echinometra mathaei* (a sea urchin) (Nagle, Paul, and Roberts 1996; Nagle and Paul 1998). Pitipeptolide A, isolated from *M. producens* (reported as *L. majuscula*) growing at Piti Bomb Holes, Guam, has been reported with antifeedant properties against *E. mathaei*, *Menaethius monoceros* (a small crab), *Parhyale hawaiensis* (an amphipod) and *Cymadusa imbroglio* (an amphipod) (Cruz-Rivera and Paul 2007). Growths of *M. producens* (reported as *L. majuscula*) from Rizal Beach, Guam yielded compounds malyngamide A, malyngamide B, and malyngolide, all three of which deterred feeding by the fishes *S. schlegeli* and *S. spinus* (Thacker, Nagle, and Paul 1997). Malyngamides A and B, and a combination treatment of majusculamides A and B, have further been shown to have antifeedant activity against *Canthigaster solandri* (pufferfish) and *Leptodius* spp. (crabs) (Pennings, Weiss, and Paul 1996).

Paradoxically, *M. producens* is also well-documented as producing natural products that stimulate feeding, specifically for sea hare species of the genus *Stylocheilus*. Crude extracts of *M. producens* have been shown to stimulate feeding by *S. striatus* (Cruz-Rivera and Paul 2002; Capper et al 2006). Various compounds isolated from *M. producens* have been tested against *S. longicauda*, and while some were shown to be feeding deterrent, or to have no significant effect, barbamide and malyngamide I stimulated feeding, while malyngamide A, malyngamide B, and majusculamides A & B stimulated feeding at lower concentrations and deterred feeding at higher concentrations (Nagle, Camacho, and Paul 1998). *S. longicauda* specializes in grazing upon *M. producens*, growing at much higher rates when able to feed upon its preferred food as compared to other diets, and sequestering the antifeedants malyngamide A and malyngamide

B in its digestive gland (Paul and Pennings 1991; Pennings and Paul 1993). Notably, *S. longicauda* sequesters malyngamide A as is, but transforms malyngamide B to malyngamide B acetate (Paul and Pennings 1991); this transformation negates the antifeedant activity of malyngamide B, making the purpose of the transformation unclear (Paul and Pennings 1991; Pennings, Weiss, and Paul 1996).

While no studies have been reported on *M. bouillonii* natural products as feeding deterrents, *M. bouillonii* was included in two papers by Cruz-Rivera and Paul (2002, 2006) that surveyed the use of algae and cyanobacteria as food and shelter on coral reefs. They found that when faced with the choice between 5 eukaryotic algal varieties, *M. producens*, and *M. bouillonii*, assemblages of fishes from two sites on Guam as well as *E. mathaei* preferred some or all of the eukaryotic algal options over the two cyanobacteria. *S. striatus* preferred *M. producens*, followed by *M. bouillonii*. The preferences against *M. bouillonii* by the fishes and urchins suggest, but are not conclusive of, poor palatability at the very least, if not antifeedant characteristics. This is further bolstered by the preference of *S. striatus* for *M. bouillonii*, perhaps suggesting protective chemistry produced by *M. bouillonii* that is worthy of sequestering. Even as *S. striatus* was quite willing to feed upon *M. bouillonii* in a laboratory setting, it was infrequently (Cruz-Rivera and Paul 2006), or not at all found (Cruz-Rivera and Paul 2002) amongst *M. bouillonii* in the field. This lack of *S. striatus*, along with a general lack of organisms beyond *A. frontalis* found on *M. bouillonii* in the field, suggested to the authors that *A. frontalis* aggressions against intruding organisms may serve to protect *M. bouillonii* from falling prey to other grazers. It must be noted that Fishelson (1966) reported finding unicellular algae, foraminifera, ciliates, and larval and small adult crustaceans alongside *A. frontalis* in the tubes of cyanobacteria that he collected from the Red Sea. This is consistent with my own observations, of frequently finding numerous miniscule bivales, crustaceans, and annelids within woven tubes of *M. bouillonii* inhabited by *A. frontalis*.

The weaving of *M. bouillonii* by *A. frontalis* may seem remarkable, and yet it represents only one of many curious ways by which alpheid shrimp secure shelter and territory. As previously mentioned, the lesser understood *A. pachychirus* is also thought to utilize woven cyanobacterial filaments for its home, but according to Bowers (1970) and Banner & Banner (1982), there are in fact three other *Alpheus* spp. that construct algal shelters: *A. clypeatus* tends to build with the red filamentous alga *Acrochaetium* sp. augmented by other algal varietals, *A. brevipes* also constructs tubes with filamentous red algae, and *A. bucephalus* builds tubes from algae sometimes in combination with sponge materials. Multiple *Synalpheus* spp. (at least 36(!!) just at Carrie Bow Cay, Belize) live within sponges, with some species establishing in large eusocial colonies (Macdonald, Ríos, and Duffy 2006). Other snapping shrimp species are known for burrow-building, digging tunnels in soft substrate that they cohabitate with gobiid fishes (Karplus 1987) or for excavating through hard corals (Kropp 1987). Alpheid shrimp can associate with brittle stars (Marin et al 2005), crinoids (Anker and Marin 2007), tunicates (Fiore and Jutte 2010), and perhaps even sea turtles (Frick et al 2003). The breadth of intricate niches inhabited by alpheid shrimp suggests a clade of organisms that is particularly adept at evolving to take advantage of resources that may be otherwise underutilized. Indeed, *A. heterochelis* has been studied with a particular interest in its second thoracic limbs (Read, McTeague, and Govind 1991). These limbs were found to be incredibly flexible and dexterous, related to their additional segmentations and disproportionate neural tissue. While *A. heterochelis* makes use of these limbs for probing its environment and scavenging food, this dexterity and flexibility undoubtedly contributes to the ability of *A. frontalis* to use its second thoracic limbs to weave.

Despite being generally understudied, particularly in terms of ecology, some researchers have raised concerns about *M. bouillonii* and its snapping shrimp symbiont as an emerging threat to coral reef health (Titlyanov, Yakovleva, and Titlyanov 2007). Perhaps inspired by the toxic nuisance blooms formed by *M. producens*, it has been suggested that *M. bouillonii* natural products play an allelopathic role that allow it to outcompete and dominate healthy corals (Titlyanov, Yakovleva, and Titlyanov 2007). There is no indication from the literature that crude extracts or isolated compounds from *M. bouillonii* have ever been found to be active in coral toxicity assays. The only experiments aimed at addressing such questions that have been reported involved the "draping" of *M. bouillonii* filaments over coral fragments, which did indeed result in negative outcomes for the corals (Titlyanov, Yakovleva, and Titlyanov 2007). It is unclear, however, if such an experimental design is truly reflective of how *M. bouillonii* contacts and interacts with corals in the ecosystem. Regardless of its relevance to true ecosystem dynamics, researchers built upon this work in reporting a bloom of *M. bouillonii* near Okinawa that they characterized as a mortal infection of a gorgonian soft coral population (Yamashiro, Isomura, and Sakai 2014). They indicated *A. frontalis* to be an accomplice that allowed for sustained coral infection by *M. bouillonii*. It must be noted that the assertion of blooming *M. bouillonii* was not substantiated with temporal monitoring data and is inconsistent with previous descriptions of *M. bouillonii* growth habits (e.g. Fishelson 1966).

0.5: Dissertation Outline

The diversity of chemistry being produced, the limited scope by which that chemistry has been studied in an ecological context, and the promise for *M. bouillonii* natural products in playing a therapeutic role for human health, suggest that much more is to be learned about the charismatic *M. bouillonii* and *A. frontalis* duo, the roles that natural products play in their relationships with each other and the reef ecosystems that they inhabit, and how this chemistry can inspire solutions to human health ailments. In the following research chapters, I will:

Chapter 1: explore the growth habits of *M. bouillonii – A. frontalis* colonies and how *M. bouillonii* may be benefitting from its relationship with *A. frontalis*

Chapter 2: leverage the chemogeography of *M. bouillonii* for the discovery of a family of new natural products

Chapter 3: elucidate the mechanism by which *A. frontalis* selects *M. bouillonii* filaments to weave with

Chapter 4: identify and assess geographical distributional patterns in secondary metabolites and secondary metabolite families produced by *M. bouillonii*

Chapter 5: develop bioinformatic strategies to facilitate the study of how *M. bouillonii* chemodiversity is represented genomically.

In the concluding chapter, I will provide perspective on what we have learned about chemical ecology, chemogeography, and genomics of the *M. bouillonii* and *A. frontalis* symbiosis and highlight exciting opportunities for future study.

0.6: References

Akare S, Condon K, Eckley S, Huang KC, Chen Z, Jiang Y, Kotake Y, Van Gessel Y, Hutto D (2015) Apratoxin-A Induces Pancreatic Toxicity. FASEB J 29:612.6. https://doi.org/10.1096/fasebj.29.1_supplement.612.6

Anker A, Marin IN (2007) *Athanas anatidactylus* sp. nov., a New Alpheid Shrimp (Crustacea: Decapoda) Associated with Crinoids in the Tropical Western Pacific. Zool Stud 46:162-167.

Banner DM, Banner AH (1982) The alpheid shrimp of Australia. Rec Aust Mus Suppl 34:359–362

Bowers, RL (1970) The behavioral ecology of *Alpheus clypeatus* Coutiere (Decapoda, Alpheidae). PhD dissertation. University of Hawaii.

Cai W, Chen QY, Dang LH, Luesch H (2017) Apratoxin S10, a Dual Inhibitor of Angiogenesis and Cancer Cell Growth To Treat Highly Vascularized Tumors. ACS Med Chem Lett 8:1007–1012. https://doi.org/10.1021/acsmedchemlett.7b00192

Cai W, Ratnayake R, Gerber MH, Chen QY, Yu Y, Derendorf H, Trevino JG, Luesch H (2019) Development of apratoxin S10 (Apra S10) as an anti-pancreatic cancer agent and its preliminary evaluation in an orthotopic patient-derived xenograft (PDX) model. Invest New Drugs 37:364–374. https://doi.org/10.1007/s10637-018-0647-0

Capper A, Cruz-Rivera E, Paul VJ, Tibbetts IR (2006) Chemical Deterrence of a Marine Cyanobacterium against Sympatric and Non-sympatric Consumers. Hydrobiologia 553:319. https://doi.org/10.1007/s10750-005-1129-x

Chen QY, Liu Y, Cai W, Luesch H (2014) Improved Total Synthesis and Biological Evaluation of Potent Apratoxin S4 Based Anticancer Agents with Differential Stability and Further Enhanced Activity. J Med Chem 57:3011–3029. https://doi.org/10.1021/jm4019965

Chen QY, Liu Y, Luesch H (2011) Systematic Chemical Mutagenesis Identifies a Potent Novel Apratoxin A/E Hybrid with Improved in Vivo Antitumor Activity. ACS Med Chem Lett 2:861-865. https://doi.org/10.1021/ml200176m

Cowles RP (1913) The habits of some tropical Crustacea. Philipp J Sci 8:119-125

Choi H, Mevers E, Byrum T, Valeriote FA, Gerwick WH (2012) Lyngbyabellins K-N from two Palmyra atoll collections of the marine cyanobacterium *Moorea bouillonii*. European J Org Chem 2012:5141–5150. https://doi.org/10.1002/ejoc.201200691

Cruz-Rivera E, Paul VJ (2002) Coral reef benthic cyanobacteria as food and refuge: diversity, chemistry and complex interactions. In: Proceedings of 9th international coral reef symposium, vol 1, pp 515–520

Cruz-Rivera E, Paul VJ (2006) Feeding by coral reef mesograzers: algae or cyanobacteria?. Coral Reefs 25: 617-627. https://doi.org/10.1007/s00338-006-0134-5

Cruz-Rivera E, Paul VJ (2007) Chemical Deterrence of a Cyanobacterial Metabolite against Generalized and Specialized Grazers. J Chem Ecol 33:213–217. https://doi.org/10.1007/s10886-006-9212-y

Dixon RA (2001) Natural products and plant disease resistance. Nature 411:843–847. https://doi.org/10.1038/35081178

Engene N, Rottacker EC, Kaštovský J, Byrum T, Choi H, Ellisman MH, Komárek J, Gerwick WH (2012) *Moorea producens* gen. nov., sp. nov. and *Moorea bouillonii* comb. nov., tropical marine cyanobacteria rich in bioactive secondary metabolites. Int J Syst Evol Microbiol 62:1171–1178. https://doi.org/10.1099/ijs.0.033761-0

Fiore CL, Jutte PC (2010) Characterization of macrofaunal assemblages associated with sponges and tunicates collected off the southeastern United States. Invertebr Biol 129:105–120. https://doi.org/10.1111/j.1744-7410.2010.00184.x

Fishelson L (1966) Observations on the littoral fauna of Israel, V. on the habitat and behaviour of *Alpheus frontalis* H. Milne-Edwards (Decapoda, Alpheidae). Crusteceana 11:98-104. https://doi.org/10.1163/156854066X00496

Frick MG, Mason PA, Williams KL, Andrews K, Gerstung H (2003) Epibionts of Hawksbill Turtles in a Caribbean Nesting Ground: A Potentially Unique Association with Snapping Shrimp (Crustacea: Alpheidae). MTN 99:8-11.

Grindberg RV, Ishoey T, Brinza D, Esquenazi E, Coates RC, Liu W, Gerwick L, Dorrestein PC, Pevzner P, Lasken R, Gerwick WH (2011) Single Cell Genome Amplification Accelerates Identification of the Apratoxin Biosynthetic Pathway from a Complex Microbial Assemblage. PLOS One 6:e18565. https://doi.org/10.1371/journal.pone.0018565

Gutiérrez M, Suyama TL, Engene N, Wingerd JS, Matainaho T, Gerwick WH (2008) Apratoxin D, a Potent Cytotoxic Cyclodepsipeptide from Papua New Guinea Collections of the Marine Cyanobacteria *Lyngbya majuscula* and *Lyngbya sordida*. J Nat Prod 71:1099–1103. https://doi.org/10.1021/np800121a

Hoffmann L, Demoulin V (1991) Marine Cyanophyceae of Papua New Guinea. II. Lyngbya bouillonii Sp. Nov., A remarkable tropical reef-inhabiting blue-green alga. Belj J Bot 124:82-88

Huang KC, Chen Z, Jiang Y, Akare S, Kolber-Simonds D, Condon K, Agoulnik S, Tendyke K, Shen Y, Wu KM, Mathieu S, Choi H, Zhu X, Shimizu H, Kotake Y, Gerwick WH, Uenaka T, Woodall-Jappe M, Nomoto K (2016) Apratoxin A Shows Novel Pancreas-Targeting Activity through the Binding of Sec 61. Mol Cancer Ther 15:1208-1216. https://doi.org/10.1158/1535-7163.MCT-15-0648

Karplus I (1987) The Association between Gobiid Fishes and Burrowing Alpheid Shrimps. In: Barnes H, Barnes M (ed.) Oceanography and Marine Biology: An Annual Review. Aberdeen University Press, Aberdeen, vol. 25, pp 458

Kleigrewe K, Almaliti J, Tian IY, Kinnel RB, Korobeynikov A, Monroe EA, Duggan BM, Di Marzo V, Sherman DH, Dorrestein PC, Gerwick L, Gerwick WH (2015) Combining Mass Spectrometric Metabolic Profiling with Genomic Analysis: A Powerful Approach for Discovering Natural Products from Cyanobacteria. J Nat Prod 78:1671–1682. https://doi.org/10.1021/acs.jnatprod.5b00301

Klein D, Braekman JC, Daloze D, Hoffmann L, Castillo G, Demoulin V (1999) Lyngbyapeptin A, a modified tetrapeptide from *Lyngbya bouillonii* (Cyanophyceae). Tetrahedron Lett 40:695–696. https://doi.org/10.1016/S0040-4039(98)02451-4

Klein D, Braekman JC, Daloze D, Hoffmann L, Demoulin V (1997) Lyngbyaloside, a Novel 2,3,4-Tri-O-methyl-6-deoxy-α-mannopyranoside Macrolide from *Lyngbya bouillonii* (Cyanobacteria). J Nat Prod 60:1057–1059. https://doi.org/10.1021/np9702751

Kropp RK (1987) Descriptions of Some Endolithic Habitats for Snapping Shrimp (Alpheidae) in Micronesia. Bull Mar Sci 41:204-213

Leão T, Castelão G, Korobeynikov A, Monroe EA, Podell S, Glukhov E, Allen EE, Gerwick WH, Gerwick L (2017) Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus *Moorea*. Proc Natl Acad Sci USA 114:3198-3203. https://doi.org/10.1073/pnas.1618556114

Leber CA, Naman CB, Keller L, Almaliti J, Caro-Diaz EJE, Glukhov E, Joseph V, Sajeevan TP, Reyes AJ, Biggs JS, Li T, Yuan Y, He S, Yan X, Gerwick WH (2020) Applying a Chemogeographic Strategy for Natural Product Discovery from the Marine Cyanobacterium *Moorena bouillonii*. Mar Drugs 18:515. https://doi.org/10.3390/md18100515

Liang X, Matthew S, Chen QY, Kwan JC, Paul VJ, Luesch H (2019) Discovery and Total Synthesis of Doscadenamide A: A Quorum Sensing Signaling Molecule from a Marine Cyanobacterium. Org Lett 21:7274–7278. https://doi.org/10.1021/acs.orglett.9b02525

Liu Y, Law BK, Luesch H (2009) Apratoxin A Reversibly Inhibits the Secretory Pathway by Preventing Cotranslational Translocation. Mol Pharmacol 76:91-104. https://doi.org/10.1124/mol.109.056085

Lopez JAV, Petitbois JG, Vairappan CS, Umezawa T, Matsuda F, Okino T (2017) Columbamides D and E: Chlorinated Fatty Acid Amides from the Marine Cyanobacterium *Moorea bouillonii* Collected in Malaysia. Org Lett 19:4231–4234. https://doi.org/10.1021/acs.orglett.7b01869

Luesch H, Chanda SK, Raya RM, DeJesus PD, Orth AP, Walker JR, Belmonte JCI, Schultz PG (2006) A functional genomics approach to the mode of action of apratoxin A. Nat Chem Biol 2:158–167. https://doi.org/10.1038/nchembio769

Luesch H, Williams PG, Yoshida WY, Moore RE, Paul VJ (2002a) Ulongamides A−F, New β-Amino Acid-Containing Cyclodepsipeptides from Palauan Collections of the Marine Cyanobacterium *Lyngbya* sp. J Nat Prod 65:996–1000. https://doi.org/10.1021/np0200461

Luesch H, Yoshida WY, Harrigan GG, Doom JP, Moore RE, Paul VJ (2002b) Lyngbyaloside B, a New Glycoside Macrolide from a Palauan Marine Cyanobacterium, *Lyngbya* sp. J Nat Prod 65:1945–1948. https://doi.org/10.1021/np0202879

Luesch H, Yoshida WY, Moore RE, Paul VJ (2000) Apramides A−G, Novel Lipopeptides from the Marine Cyanobacterium *Lyngbya majuscula*. J Nat Prod 63:1106–1112. https://doi.org/10.1021/np000078t

Luesch H, Yoshida WY, Moore RE, Paul VJ (2002c) New apratoxins of marine cyanobacterial origin from guam and palau. Bioorg Med Chem 10:1973–1978. https://doi.org/10.1016/S0968-0896(02)00014-7

Luesch H, Yoshida WY, Moore RE, Paul VJ, Corbett TH (2001) Total Structure Determination of Apratoxin A, a Potent Novel Cytotoxin from the Marine Cyanobacterium *Lyngbya majuscula*. J Am Chem Soc 123:5418–5423. https://doi.org/10.1021/ja010453j

Ma D, Zou B, Cai G, Hu X, Liu JO (2006) Total Synthesis of the Cyclodepsipeptide Apratoxin A and Its Analogues and Assessment of Their Biological Activities. Chem Eur J 12:7615-7626. https://doi.org/10.1002/chem.200600599

Macdonald III KS, Ríos R, Duffy JE (2006) Biodiversity, host specificity, and dominance by eusocial species among sponge-dwelling alpheid shrimp on the Belize Barrier Reef. Divers Distrib 12:165-178. https://doi.org/10.1111/j.1366-9516.2005.00213.x

Marin IN, Anker A, Britayev TA, Palmer AR (2005) Symbiosis between the Alpheid Shrimp, *Athanas ornithorhynchus* Banner and Banner, 1973 (Crustacea: Decapoda), and the Brittle Star, *Macrophiothrix longipeda* (Lamarck, 1816) (Echinodermata: Ophiuroidea). Zool Stud 44:234-241

Matthew S, Salvador LA, Schupp PJ, Paul VJ, Luesch H (2010) Cytotoxic Halogenated Macrolides and Modified Peptides from the Apratoxin-Producing Marine Cyanobacterium *Lyngbya bouillonii* from Guam. J Nat Prod 73:1544–1552. https://doi.org/10.1021/np1004032

Matthew S, Schupp PJ, Luesch H (2008) Apratoxin E, a Cytotoxic Peptolide from a Guamanian Collection of the Marine Cyanobacterium *Lyngbya bouillonii*. J Nat Prod 71:1113–1116. https://doi.org/10.1021/np700717s

Mevers E, Matainaho T, Allara M, Di Marzo V, Gerwick WH (2014) Mooreamide A: A cannabinomimetic lipid from the marine cyanobacterium *Moorea bouillonii*. Lipids 49:1127–1132. https://doi.org/10.1007/s11745-014-3949-9

Milne-Edwards H (1837) Histoire Naturelle des Crustacés, Comprenant l´Anatomie, la Physiologie et la Classification de ces Animaux. In: Librairie Encyclopédique de Roret (1834-1840). Paris, vol 3, p 638, plates 1–42

Moikeha SN, Chu GW, Berger LR (1971) Dermatitis-Producing Alga *Lyngbya majuscula* Gomont in Hawaii. I. Isolation and Chemical Characterization of the Toxic Factor. J Phycol 7:4-8. https://doi.org/10.1111/j.1529-8817.1971.tb01470.x

Nagle DG, Camacho FT, Paul VJ (1998) Dietary preferences of the opisthobranch mollusc *Stylocheilus longicauda* for secondary metabolites produced by the tropical cyanobacterium *Lyngbya majuscula*. Mar Biol 132:267-273. https://doi.org/10.1007/s002270050392

Nagle DG, Paul VJ (1998) Chemical defense of a marine cyanobacterial bloom. J Exp Mar Biol Ecol 225:29-38. https://doi.org/10.1016/S0022-0981(97)00205-0

Nagle DG, Paul VJ (1999) Production of Secondary Metabolites by Filamentous Tropical Marine Cyanobacteria: Ecological Functions of the Compounds. J Phycol 35:1412–1421. https://doi.org/10.1046/j.1529-8817.1999.3561412.x

Nagle DG, Paul VJ, Roberts MA (1996) Ypaoamide, a new broadly acting feeding deterrent from the marine cyanobacterium Lyngbya majuscula. Tetrahedron Lett 37:6263-6266. https://doi.org/10.1016/0040-4039(96)01391-3

Nakamura F, Maejima H, Kawamura M, Arai D, Okino T, Zhao M, Ye T, Lee J, Chang YT, Fusetani N, Nakao Y (2018) Kakeromamide A, a new cyclic pentapeptide inducing astrocyte differentiation isolated from the marine cyanobacterium *Moorea bouillonii*. Bioorg Med Chem Lett 28:2206–2209. https://doi.org/10.1016/j.bmcl.2018.04.067

Paatero AO, Kellosalo J, Dunyak BM, Almaliti J, Gestwicki JE, Gerwick WH, Taunton J, Paavilainen VO (2016) Apratoxin Kills Cells by Direct Blockade of the Sec61 Protein

Translocation Channel. Cell Chem Biol 22:561-566. https://doi.org/10.1016/j.chembiol.2016.04.008

Paul VJ, Pennings SC (1991) Diet-derived chemical defenses in the sea hare *Stylocheilus longieauda* (Quoy et Gaimard 1824). J Exp Mar Biol Ecol 151:227-243. https://doi.org/10.1016/0022-0981(91)90126-H

Pennings SC, Paul VJ (1993) Sequestration of dietary secondary metabolites by three species of sea hares: location, specificity and dynamics. Mar Biol 117:535–546. https://doi.org/10.1007/BF00349763

Pennings SC, Weiss AM, Paul VJ (1996) Secondary metabolites of the cyanobacterium *Microcoleus lyngbyaceus* and the sea hare *Stylocheilus Iongicauda*: palatability and toxicity. Mar Biol 26:735-743.

Pereira AR, McCue CF, Gerwick WH (2010) Cyanolide A, a Glycosidic Macrolide with Potent Molluscicidal Activity from the Papua New Guinea Cyanobacterium *Lyngbya bouillonii*. J Nat Prod 73:217–220. https://doi.org/10.1021/np9008128

Qiu B, Tan A, Veluchamy AB, Li Y, Murray H, Cheng W, Liu C, Busoy JM, Chen QY, Sistla S, Hunziker W, Cheung CMG, Wong TY, Hong W, Luesch H, Wang X (2019) Apratoxin S4 Inspired by a Marine Natural Product, a New Treatment Option for Ocular Angiogenic Diseases. Invest Ophthalmol Vis Sci 60:3254-3263. https://doi.org/10.1167/iovs.19-26936

Read AT, McTeague JA, Govind CK (1991) Morphology and Behavior of an Unusually Flexible Thoracic Limb in the Snapping Shrimp, *Alpheus heterochelis*. Biol Bull 181:158-168. https://doi.org/10.2307/1542498

Richters F (1880) Decapoda. In: Möbius K, Richters F and von Martens E (eds) Beiträge zur Meeresfauna der Insel Mauritius und der Seychellen. Gutmann'schen Buchhandlung, Berlin, pp 139-178, plates 15-18

Soria-Mercado IE, Pereira A, Cao Z, Murray TF Gerwick WH (2009) Alotamide A, a novel neuropharmacological agent from the marine cyanobacterium *Lyngbya bouillonii*. Org Lett 11:4704–4707. https://doi.org/10.1021/ol901438b

Sweeney-Jones AM, Gagaring K, Antonova-Koch J, Zhou H, Mojib N, Soapi K, Skolnick J, McNamara CW, Kubanek J (2020) Antimalarial Peptide and Polyketide Natural Products from the Fijian Marine Cyanobacterium *Moorea producens*. Mar Drugs 18:167. https://doi.org/10.3390/md18030167

Tan LT, Márquez BL, Gerwick WH (2002) Lyngbouilloside, a novel glycosidic macrolide from the marine cyanobacterium *Lyngbya bouillonii*. J Nat Prod 65:925-928. https://doi.org/10.1021/np010526c

Tan LT, Okino T, Gerwick WH (2013) Bouillonamide: A Mixed Polyketide-Peptide Cytotoxin from the Marine Cyanobacterium *Moorea bouillonii*. Mar Drugs 11:3015–3024. https://doi.org/10.3390/md11083015

Thacker RW, Nagle DG, Paul VJ (1997) Effects of repeated exposures to marine cyanobacterial secondary metabolites on feeding by juvenile rabbitfish and parrotfish. Mar Ecol Prog Ser 147:21-29. https://doi.org/10.3354/meps147021

Thornburg CC, Cowley ES, Sikorska J, Shaala LA, Ishmael JE, Youssef DTA, McPhail KL (2013) Apratoxin H and Apratoxin A Sulfoxide from the Red Sea Cyanobacterium *Moorea producens*. J Nat Prod 76:1781–1788. https://doi.org/10.1021/np4004992

Tidgewell K, Engene N, Byrum T, Media J, Doi T, Valeriote FA, Gerwick WH (2010) Evolved Diversification of a Modular Natural Product Pathway: Apratoxins F and G, Two Cytotoxic Cyclic Depsipeptides from a Palmyra Collection of *Lyngbya bouillonii*. Chembiochem. 11:1458–1466. https://doi.org/10.1002/cbic.201000070

Titlyanov EA, Yakovleva IM, Titlyanov TV (2007) Interaction between benthic algae (*Lyngbya bouillonii*, *Dictyota dichotoma*) and scleractinian coral *Porites lutea* in direct contact. J Exp Mar Bio Ecol 342:282-291. https://doi.org/10.1016/j.jembe.2006.11.007

Tronholm A, Engene N (2019) *Moorena* gen. nov., a valid name for "*Moorea* Engene & al." nom. inval. (Oscillatoriaceae, Cyanobacteria). Notulae Algarum 122:1–2

Versluis M, Schmitz B, von der Heydt A, Lohse D (2000) How Snapping Shrimp Snap: Through Cavitating Bubbles. Science 289:2114-2117. https://doi.org/10.1126/science.289.5487.2114

Yamashiro H, Isomura N, Sakai K (2014) Bloom of cyanobacterium *Moorea bouillonii* on the gorgonian coral *Annella reticulata* in Japan. Sci Rep 4:6032. https://doi.org/10.1038/srep06032

Chapter 1: Cyanobacteria-Shrimp Colonies in the Mariana Islands

1.0: Abstract

Cyanobacteria have multifaceted ecological roles on coral reefs. *Moorena bouillonii*, a chemically rich filamentous cyanobacterium, has been characterized as a pathogenic organism with an unusual ability to overgrow gorgonian corals, but little has been done to study its general growth habits or its unique association with the snapping shrimp *Alpheus frontalis*. Quantitative benthic surveys, and field and photographic observations were utilized to develop a better understanding of the ecology of these species, while growth experiments and nutrient analysis were performed to examine how this cyanobacterium may be benefiting from its shrimp symbiont. Colonies of *M. bouillonii* and *A. frontalis* displayed considerable habitat specificity in terms of occupied substrate. Although found to vary in abundance and density across survey sites and transects, *M. bouillonii* was consistently found to be thriving with *A. frontalis* within interstitial spaces on the reef. Removal of *A. frontalis* from cyanobacterial colonies altered *M. bouillonii* pigmentation, whereas cyanobacteria-shrimp colonies exhibited elevated nutrient levels compared to the surrounding seawater.

1.1: Introduction

Cyanobacteria, formerly known as blue-green algae, represent a wide array of organisms that operate in a diversity of roles in marine ecosystems. In addition to producing oxygen via photosynthesis, some cyanobacteria can also fix nitrogen; these capabilities make cyanobacteria important primary producers in many marine systems, especially in oligotrophic open ocean environments (Carpenter, Subramaniam and Capone 2004). Conversely, cyanobacteria, often empowered by anthropogenic perturbations, can disrupt aquatic ecosystems by forming sustained and deleterious blooms (CyanoHABs) that produce toxins and can cause localized hypoxia (Paerl and Paul 2012). In the context of coral reef ecosystems, a similar dichotomy can be identified. In addition to CyanoHABs occurring in and around coral reefs, cyanobacteria have been implicated in coral diseases, such as black band disease and grey-patch disease (Sweet et al. 2019; Frias-Lopez et al. 2003). Cyanobacteria also contribute positively to coral reef ecosystems, serving as habitat and sources of food (Cruz-Rivera and Paul 2002, 2006), and engaging in specialized symbioses, including with sponges (Schorn et al 2019) and shrimp (Banner and Banner 1982).

One intriguing interaction between a cyanobacterium and a tropical reef organism that was serendipitously observed through drug discovery efforts in the Tropical Western Pacific (Tan, Márquez and Gerwick 2002) is the association between the filamentous cyanobacterium *Moorena bouillonii* (L.Hoffmann & Demoulin) Engene & Tronholm 2019 (Oscillatoriaceae) and a symbiotic snapping shrimp *Alpheus frontalis* H. Milne Edwards 1837 (Alpheidae) (Figure 1.1). *Alpheus frontalis* was first described in 1837 (Milne-Edwards 1837), and reports of its association with *M. bouillonii* date back to as early as 1880 (Richters 1880), when two shrimp were collected from Mauritius in woven tubes of cyanobacteria, which at the time were identified as *Oscillatoria* sp. *Alpheus frontalis* weaves *M. bouillonii* filaments into labyrinthine shelters consisting of interconnected tunnels and chambers (See Supplemental Video 1.S1). *Alpheus frontalis* (often collected with *M. bouillonii*) has been documented in various publications, catalogs of crustacean taxonomy, and collection records, suggesting that this partnership extends from the eastern African coast and the Red Sea, across the Indian Ocean, and throughout the Tropical Western Pacific, bounded by Australia, Southern Japan, and French Polynesia (Richters 1880; Fishelson 1966; Simões, Apel and Jones 2001; Davie and Short 2001; Nomura et al. 1996; Poupin 1998).

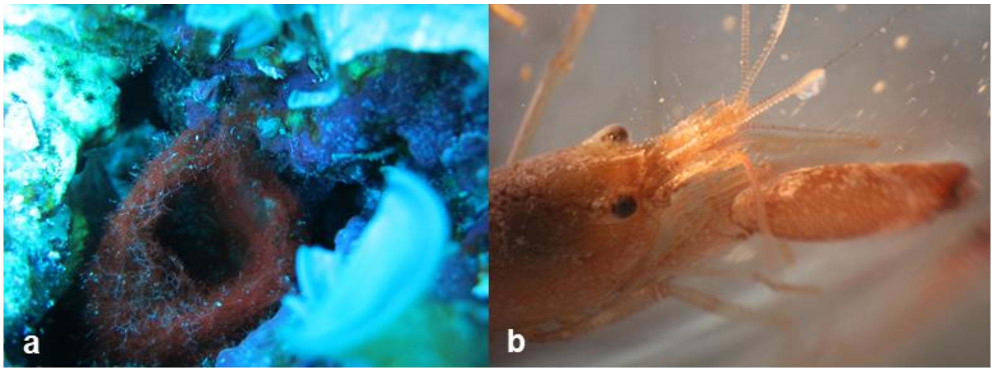


Figure 1.1 - *Moorena bouillonii* and *Alpheus frontalis*; a) A woven tube of the filamentous cyanobacterium *M. bouillonii* and b) its shrimp symbiont *A. frontalis*

While much of the scholarly effort directed towards *A. frontalis* has focused on its taxonomic classification and distribution, research on *M. bouillonii* has largely centered on its rich secondary metabolite chemistry relevant to human health. In fact, the entire genus *Moorena* (previously *Moorea,* formerly a part of *Lyngbya*) has been identified as a prolific source of biologically active compounds (Tronholm and Engene 2019; Leão et al. 2017). Since its initial description in 1991 (Hoffmann and Demoulin 1991), more than 60 unique natural products have been discovered from *M. bouillonii* (according to the MarinLit database [https://pubs.rsc.org/marinlit/] and SciFinder [https://scifinder.cas.org/]), including the apratoxins, a family of exquisitely potent cancer cell cytotoxins (Luesch et al. 2001). In terms of driving ecological interactions among reef organisms, metabolites isolated from the closely related *M. producens* (previously *Lyngbya majuscula*) have been shown to be strong herbivore feeding deterrents (Nagle and Paul 1999); this has led to speculation that *M. bouillonii* may be producing similar anti-predatory chemicals that could also benefit its cohabitants.

Despite the passage of over one hundred years since the discovery of the *M. bouillonii* - *A. frontalis* association, there have been strikingly few ecological investigations of either species. The weaving behavior of *A. frontalis* was first described in 1913 (Cowles 1913). This was built upon in a 1966 report that briefly described the habitat of *M. bouillonii* and *A. frontalis* colonies and detailed how pairs of shrimp interact with *M. bouillonii* in the laboratory to construct shelters (Fishelson 1966). Cruz-Rivera and Paul (2002, 2006) further described the commensal nature of this association in their studies of cyanobacteria as sources of food and shelter on coral reefs, highlighting the specificity of the *M. bouillonii* - *A. frontalis* pairing, articulating the protective behaviors observable in the shrimp, and noting the shrimp's use of the cyanobacteria for both food and shelter. Various other manuscripts also noted these interactions, but the nature of their association and the extensive structures created through the remarkable weaving skills of these shrimp remain to be elucidated.

Without detailed ecological studies, it is difficult to characterize *M. bouillonii* and *A. frontalis* as an emerging threat to coral reef health and resiliency. Experiments measuring effects of the cyanobacteria coming into contact with corals suggest that *M. bouillonii* natural products are toxic to corals and/or their symbionts and may serve as allelopathic agents that provide a competitive advantage against healthy corals (Titlyanov, Yakovleva and Titlyanov 2007). Additionally, localized blooms of *M. bouillonii* near Okinawa were reported as a major driver of mortality for the gorgonian *Annella reticulata* (Yamashiro, Isomura and Sakai 2014). Although somewhat in contrast to the historical knowledge of this species-pair's distribution and habits on coral reefs, these studies curiously suggest a complexity and context dependence for the ecological roles played by *M. bouillonii* and *A. frontalis*.

Herein we describe efforts towards a more thorough and expansive understanding of *M. bouillonii* and its symbiotic association with *A. frontalis* in the context of coral reef systems found within the Mariana Islands. We achieve this by combining historical perspectives along with contemporary field studies and experiments. Moreover, we attempt to clarify the ecological relationships of this symbiotic pair among the stony corals with which they become associated, hypothesizing that the engulfing overgrowth of gorgonians by *M. bouillonii* as reported by Yamashiro et al. (2014) may not represent *M. bouillonii* ecology within all reef ecosystems. We address this hypothesis through *in situ* observations and photography, quantitative surveys of abundance and habitat, and culturing experiments designed to measure how *M. bouillonii* may benefit from its association with *A. frontalis*. With this foundation of knowledge, we hope to expand knowledge regarding the ecology of *M. bouillonii* and *A. frontalis* on coral reefs.

1.2: Methods

1.2.1: Surveys and Underwater Photography

Abundance surveys were conducted in Laulau Bay, Saipan, Commonwealth of the Northern Mariana Islands, USA (15°09'18"N:145°42'20"E; June 17, 2017) (Figure 1.2b), and Finger Reef, Apra Harbor, Guam, USA (13°26'40"N:144°38'11"E; July 1, 2017) (Figure 1.2c). Each consisted of two 50 m transects, haphazardly placed in areas of high *M. bouillonii* abundance. All *M. bouillonii* colonies within 1 m of either side of the transect that displayed morphologies consistent with shrimp association (e.g. displayed the ‘cobweb,’ woven, tube-forming morphology indicative of recent weaving activity by shrimp) were counted. Shrimp were not often visible during surveys but could reliably be detected by coming into gentle contact with surfaces of *M. bouillonii* that display this morphology and waiting for the distinct, and often startling “snap” of a defending shrimp, were recorded (100 m^2 per transect). Colonies were also photographed. *Moorena bouillonii* growing without shrimp, which has previously been reported on Guam (Matthew, Schupp, and Luesch 2008) was occasionally observed near transects at Finger Reef growing beneath overhanging reef structures. These cyanobacterial growths were not included in colony counts as the growth morphology is much more dispersed and without discrete, distinct colonies. Records were kept for each consecutive 20 m^2 area (10 linear meters) in order to capture the local variability and patchiness in colony distribution at these smaller scales. Surveys were conducted by two different SCUBA divers experienced in recognizing *M. bouillonii – A. frontalis* colonies, after which the midpoints between both counts were calculated per 20 m^2 segment. The Laulau Bay abundance surveys were conducted along the 10 m and 12 m depth contours, while the Finger Reef abundance surveys were conducted along the 7 m and 12 m depth contours on the reef slope. Means and sample standard deviations for colonies/m^2 were calculated using each 10 m segment along the 50 m transects as replicates and computed using Microsoft Excel. Substrate surveys were conducted at Finger Reef, Apra Harbor, Guam, USA (13°26'40"N:144°38'11"E; May 31, 2018), Piti Bomb Holes, Guam, USA (13°28'15"N:144°42'04"E; June 1, 2018), and a fringing reef shoreward of Cocos Lagoon in Merizo, Guam, USA (13°16'12"N:144°39'42"E; June 13, 2018) (Figure 1.2c). Each consisted of two to three 50 m transects, haphazardly placed in areas of high *M. bouillonii – A. frontalis* colony abundance. One survey at Finger Reef was conducted on the reef flat, between 1 m and 2 m depth, and the other was conducted on the reef slope around the 9 m depth contour. Surveys at Piti Bomb Holes were conducted between less than 1 m and 2 m on the reef flat near its boundary with the reef slope. The surveys on the Merizo fringing reef were conducted between 1 m and 3 m at the boundary between the reef flat and reef slope. All *M. bouillonii* colonies within 1 m of either side of the transect that displayed morphologies consistent with shrimp association were photographed, providing a record of the substrate occupied. Colonies were attributed substrate types based upon the structure they were contacting; in cases where colonies were contacting multiple substrates, substrate was assigned based on the majority of contacts. Videos were also recorded along each transect to document surrounding benthic habitat and provide additional context for determining substrate occupied by colonies with multiple contacts. Both photos and videos were captured using a Canon PowerShot® ELPH 100 HS set for underwater macro-photography or video mode, respectively. Additional observational photographs were made via SCUBA from various locations on Guam, and via snorkeling in Laulau Bay, Saipan. Observations and photographic documentation were also conducted on reefs around Palmyra Atoll, American Samoa, Papua New Guinea, and Kavaratti in the Lakshadweep Islands.

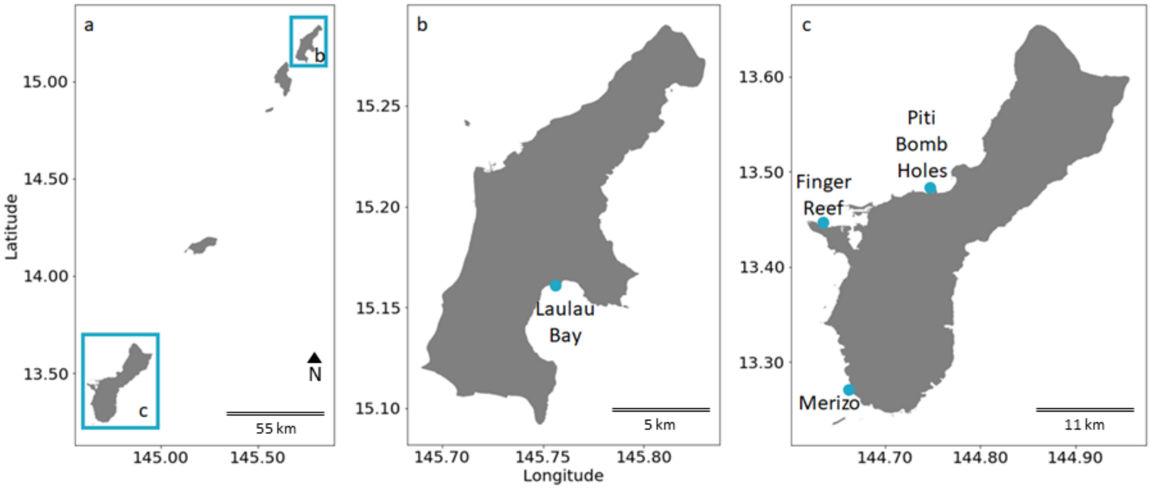


Figure 1.2 - Maps of a) the Mariana Islands, b) Saipan, and c) Guam with marked survey sites. *M. bouillonii – A. frontalis* colony growth form and abundance was surveyed on the reefs in Laulau Bay, Saipan, and Finger Reef, Apra Harbor, Guam, while colony substrate preference was surveyed on Finger Reef, on the reef at Piti Bomb Holes, Guam, and on a fringing reef in Merizo, Guam. Sites are marked on the maps with blue dots. The map was created using a NOAA-produced coastline shapefile (NOAA 2005).

1.2.2: Taxonomic Identifications

Moorena bouillonii – A. frontalis colonies were identified in the field based on their distinctive woven-tube morphology and deep red color (Figure 1.1a). Moreover, snapping shrimp such as *Alpheus frontalis* defend against other organisms that physically disturb their cyanobacterial colony by rapidly closing their major chela, which results in a distinctive popping noise that announces their presence within the woven structure. In most cases, the visual or audible presence of *A. frontalis* provided further morphotaxonomic verification during our surveys. Care was also taken to examine cryptic locations within the reef structure, as the cyanobacteria-shrimp colonies are often nestled within crevices and around the bases of solid substrata. Small sections (1-3 cm^2) of colonies were collected and examined via compound light microscopy to inspect the woven *M. bouillonii* tubes for other species of cyanobacterial filaments. In addition, partial *16S rRNA* gene sequences were obtained from four shrimp-associated and three non-shrimp-associated *M. bouillonii* for final confirmation of field identifications (See Figure 1.S1 and Supplemental Methods in 1.4: Appendix: Supplemental Information).

1.2.3: Culturing Experiment

Alpheus frontalis and *M. bouillonii* growing with (shrimp-associated cyanobacterium) and without *A. frontalis* (non-shrimp-associated cyanobacterium) were collected from Apra Harbor, Guam (with additional shrimp collected near Merizo, Guam) and used to assemble four different growth conditions in reusable plastic food storage containers: shrimp-associated cyanobacterium with a shrimp, shrimp-associated cyanobacterium without a shrimp, non-shrimp-associated cyanobacterium with a shrimp, and non-shrimp-associated cyanobacterium without a shrimp, each with five replicates. Weighted containers were covered with window screen, placed in a blocked arrangement, and submerged in a flow-through raw seawater table. Submersion allowed for water exchange and the deposition of materials for shrimp to scavenge in each container, while the mesh was intended to prevent shrimp escape. The water table was covered with 50% shade cloth to limit light stress. The experiment was conducted over 15 days (June 11, 2016 to June 26, 2016), with photos taken each day to document changes in *M. bouillonii* pigmentation and wet weights of cyanobacterial biomass recorded on the first and last days. Excess water was removed from cyanobacterial biomass prior to wet weight measurements using a salad spinner. For each sample, the salad spinner was pumped 20 times at a rate of 1 pump per second, followed by a 10 second period of unconstrained spinning. Statistical analysis of changes in wet weight were conducted using the statsmodel package for Python 3 to build a two-way ANOVA model. Type of cyanobacteria (shrimp versus non-shrimp) and presence or absence of shrimp were set as independent variables for the two-way

ANOVA, and an additional term was added to account for the blocked experimental design.

1.2.4: Nutrient Analyses

Five paired water samples were collected from Finger Reef, Apra Harbor, Guam on July 3, 2017. Syringes (50 mL) with blunt-tip needles were used to collect water from inside the woven tubes and chambers of five individual shrimp-associated *M. bouillonii* colonies, and from approximately one meter above each colony in the water column. Samples were flash frozen in dry ice and kept frozen until they were submitted for nutrient analyses at the Oceanographic Data Facility, Scripps Institution of Oceanography, San Diego, California. Statistical significance was determined for each nutrient using Student's t-test (paired, two-tailed), administered in Microsoft Excel.

1.3: Results & Discussion

1.3.1: Abundance Surveys

Field abundance surveys of *M. bouillonii – A. frontalis* colony growth in Laulau Bay, Saipan (Figure 1.2b) and at Finger Reef in Apra Harbor, Guam (Figure 1.2c), indicate that colony density varies greatly within and among reefs. The two transects on the reef in Laulau Bay revealed higher densities of *M. bouillonii – A. frontalis* colonies (mean = 4.73, 3.01 colonies per m^2; standard deviation (s.d.) = 0.79, 0.49; n = 5) than the two transects on the reef slope of Finger Reef (mean = 0.47, 1.33 colonies per m^2; s.d. = 0.61, 1.50; n = 5) (See Figure 1.3 and Table 1.S1). In consideration of colony density at a smaller scale, it is of note that the most densely populated transect surveyed in Laulau Bay was composed of the top five most densely populated 10 m increments (5.8, 5.2, 4.5, 4.2, and 3.9 colonies per m^2); however, the sixth most densely populated 10 m segment of reef was surveyed on Finger Reef, with 3.8 colonies per m^2. Every 10 m interval along the Saipan transects recorded densities greater than two *M. bouillonii – A. frontalis* colonies per square meter, and the most densely populated Saipan transect in particular represents an area of consistent high-density *M. bouillonii – A. frontalis* colony growth (mean = 4.73 colonies per m^2; s.d. = 0.79; n = 5). The two transects in Guam, however, ranged from 3.75 to 0.05 colonies per m^2 and from 1.5 to 0.08 colonies per m^2, with the former illustrating wide variation in colony density even within small spatial scales (s.d. = 1.50). Of particular note, while many colonies were observed to be in contact with stony coral, none of the colonies observed during the abundance surveys were found to display morphologies that would suggest coral overgrowth (e.g. cyanobacterial biomass engulfing the extended, external surfaces of corals) (See Figure 1.S2).[1]

[1] While quantification of coral health was beyond the scope of this study, it is worth anecdotally noting that colonies of *M. bouillonii* were observed in the interstices of both healthy and unhealthy corals. Particularly in Laulau Bay, colonies were found in the interstices of coral structures that were heavily overgrown by other species of cyanobacteria and macroalgae. In contrast, many colonies in Apra Harbor were located in association with large, seemingly healthy and thriving colonies of *P. rus*. Removal of *M. bouillonii* from coral interstices could reveal live coral tissue, crustose coralline algae, bare reef, or accumulated sediment and coral rubble. Further study is needed to understand how *M. bouillonii – A. frontalis* colonies originate, and to determine whether colony initiation has localized impacts on the coral tissues in and adjacent to favored reef interstices.

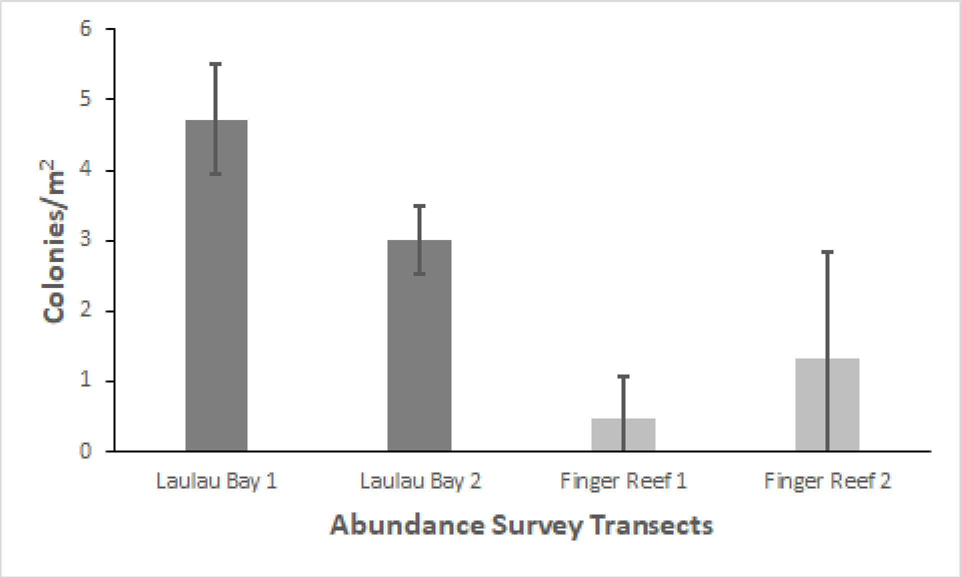


Figure 1.3 - *M. bouillonii – A. frontalis* colony density and abundance. *M. bouillonii – A. frontalis* colonies were found to be consistently abundant and densely occupying the reef in Laulau Bay, while colony density and abundance was more varied on Finger Reef. Note: For each transect, colonies along each linear 10 m (20 m^2) of the 50 m transect were separately counted and recorded and make up the replicates represented by the standard deviation error bars**.**

1.3.2: Substrate Surveys

Substrate surveys conduct ed around the island of Guam at Finger Reef in Apra Harbor, Piti Bomb Holes, and a fringing reef by Merizo (Figure 1.2c) yielded detection of shrimp-associated *M. bouillonii colonies* growing mainly in coral-associated interstitial spaces. Colonies were documented to predominantly grow in association with *Porites rus* and *Porites cylindrica*, with a much smaller number of colonies growing on bare reef substrate and only two colonies found growing with *Porites* cf. *deformis* (See Figure 1.4 and Table 1.S2). Colonies were not found growing on any other substrate types during these surveys. *Porites rus* was the predominant colony substrate for both the reef flat and reef slope transects at Finger Reef, hosting 83% and 91% of documented colonies, respectively. *Porites cylindrica* hosted most of the documented colonies at the Merizo reef, with 78% and 97%. Transects at Piti Bomb Holes were less consistent, with 76% of the colonies along the first transect growing with *P. rus*, while

97% of colonies along transect two and 93% along transect three were growing with *P. cylindrica*. In agreement with the abundance surveys, while colonies were found to be growing amongst live coral substrate, none of the colonies observed during the substrate surveys displayed a morphology consistent with coral overgrowth, and macroscopically, none of the corals appeared to be injured by their associated cyanobacteria-shrimp colonies.

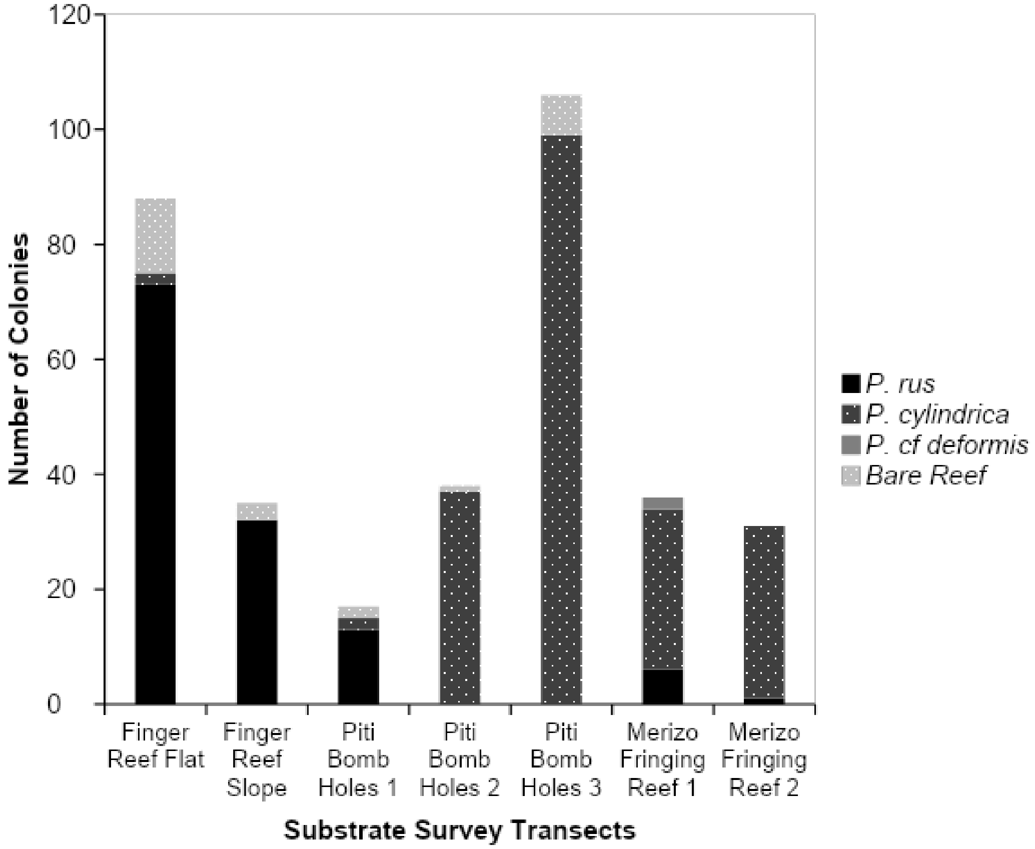


Figure 1.4 - *M. bouillonii – A. frontalis* colony substrate distribution. *M. bouilloni*i - *A. frontalis* colonies were most often found occupying interstitial spaces amongst *Porites rus* and *P. cylindrica* corals and were not found to display growth morphologies indicative of coral overgrowth.

1.3.3: Culturing Experiment

The photographic time series conducted during the culturing experiment revealed generally consistent patterns in pigmentation change for each of the four growth conditions (Figure 1.5). Non-shrimp-associated cyanobacteria began the experiment with a red-orange hue occasionally accented with yellow-green filaments, which quickly developed into a darker, deeper red hue when grown with shrimp, but remained relatively unchanged when grown without. Shrimp-associated cyanobacteria appeared dark red and heavily pigmented at the onset of the experiment. In the absence of shrimp, shrimp-associated cyanobacteria dramatically changed appearance, rapidly shifting from dark red filaments to a bright green, gelatinous mass. By day fifteen, the shrimp-associated cyanobacteria grown without a shrimp appeared to begin to recover, with new filament growth becoming visible extending out of the green mass. Shrimp-associated cyanobacteria grown with shrimp proved to be the least consistent growth condition in this study - two replicates retained dark red pigmentation, two replicates developed patches of green, and one replicate changed to a dark green coloration. This lack of consistency can be explained by the inadvertent removal of the three shrimp living in the three replicates that displayed greening - two died during the course of the experiment (blocks B and D), and one escaped (block A) (See Figure 1.S3, for complete photographic time course).

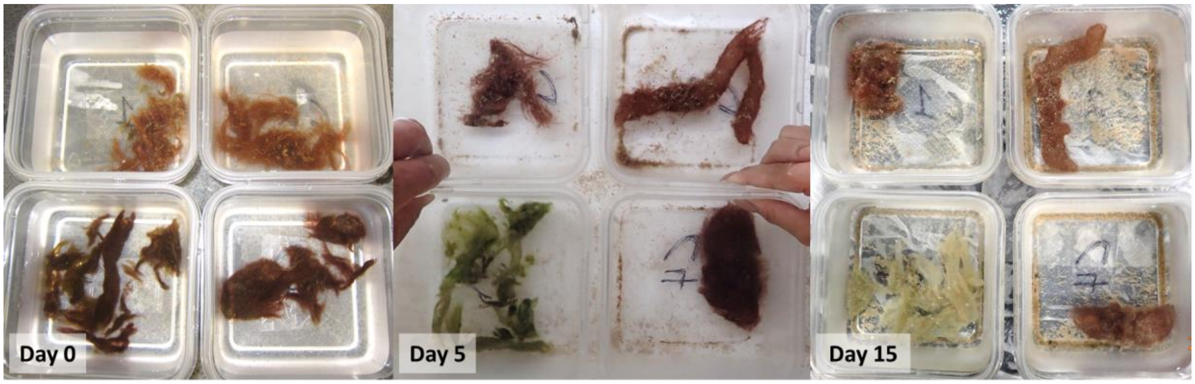


Figure 1.5 - Days 0, 5, and 15 of a representative block (block C) in the photographic time series of the *M. bouillonii* culturing experiment. Representative images from days 0, 5, and 15 of the photo time series of the *M. bouillonii* culturing experiment. Clockwise from top left: non-shrimp-associated cyanobacterium without a shrimp, non-shrimp-associated cyanobacterium with a shrimp, shrimp-associated cyanobacterium with a shrimp, and shrimp-associated cyanobacterium without a shrimp.

Out of the four growth conditions, non-shrimp-associated cyanobacteria grown without a shrimp was the only condition that averaged net positive growth, as measured by change in wet weight (mean = 0.37 g, s.d. = 0.54 g, n = 5) (Figure 1.6, Table 1.S3). Mean change in wet weight was close to zero for shrimp-associated cyanobacteria grown without a shrimp (mean = -0.14 g, s.d. = 0.59 g, n = 5), while both types of cyanobacteria grown with a shrimp decreased in wet weight on average (shrimp-associated: mean = -0.92 g, s.d. = 0.78 g, n = 5; non-shrimp-associated: mean = -0.61 g, s.d. = 0.54 g, n = 5). Two-way ANOVA accounting for the blocked experimental design was administered to test the statistical significance of the effects that cyanobacterial type and shrimp presence had on changes in wet weight. The blocked experimental design (F = 10.71, p-value < 0.01), type of cyanobacteria (F = 7.41, p-value < 0.05), and presence of shrimp (F = 34.14, p-value < 0.01) were all found to have statistically-significant effects on cyanobacterial growth, as measured by changes in wet weight, while the interaction between cyanobacterial type and shrimp presence was not found to be significant (F = 0.43, p-value = 0.53) (Table 1.S4).[2] The significant effect of the blocked experimental design is attributable to the heterogeneity of flow environments in the water table in which the experiment took place; blocks differed in their proximity to inflow and outflow of seawater in the table.

[2] A question could be raised whether there were statistically significant interactions between blocks and either experimental factor (type of cyanobacteria or presence of shrimp). Testing for these potential interactions yielded non-statistically significant results, with p-values of 0.40 and 0.91 for type of cyanobacteria and presence of shrimp, respectively.

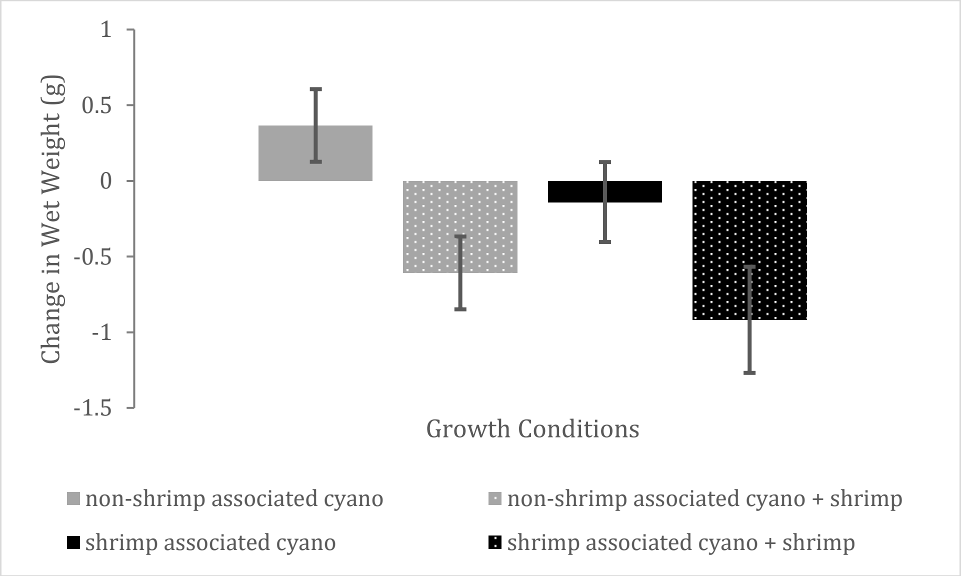


Figure 1.6 - Changes in wet weight from the *M. bouillonii* culturing experiment
Presence of shrimp corresponded to negative changes in wet weight, indicating a loss in cyanobacterial biomass resulting from shrimp feeding. Two-way ANOVA accounting for the blocked experimental design indicated statistically significant effects on changes in wet weight attributable to the type of cyanobacteria (F = 7.41, p-value < 0.05), the presence of shrimp (F = 34.14, p-value < 0.01), and the blocked experimental design (F = 10.71, p-value < 0.01). The interaction between cyanobacterial type and shrimp presence was not found to be significant (F = 0.43, p-value = 0.53). Error bars display standard error.

1.3.4: Comparative Nutrient Analyses

From Finger Reef, Apra Harbor, Guam, five pairs of water samples were collected from inside *M. bouillonii - A. frontalis* colonies and approximately one meter above each colony, were analyzed for their concentrations of silicate, PO_4^{3-}, NO_3^-, NO_2^-, and NH_4^+. For all five measured nutrients, statistically significant differences in concentration between colonies and ambient seawater were detected (Table 1.1, Table 1.S5). Colonies were shown to have significantly higher concentrations of PO_4^{3-}, NO_3^-, NO_2^-, and NH_4^+, while ambient seawater was measured to have higher concentrations of silicate.

Table 1.1 – Summary of paired water sample nutrient analysis data, revealing statistically significant differences in nutrient concentration between the water column and *M. bouillonii* colonies for all nutrients measured. (n=5)

	NO_3^-	PO_4^{3-}	silicate	NO_2^-	NH_4^+
	µmol/L	µmol/L	µmol/L	µmol/L	µmol/L
Water column mean	0.402	0.038	2.3	0.004	0.444
Water column s.e.	0.074	0.006	0.071	0.004	0.039
Colony mean	1.032	0.148	2.1	0.05	0.678
Colony s.e.	0.209	0.025	0.084	0.006	0.108
P-value[a]	0.020	0.006	0.022	0.002	0.040

[a] P-values from Student's T-test (paired, two tailed)

1.3.5: Additional Discussion and Conclusions

Study of *Moorena bouillonii - Alpheus frontalis* colonies across the surveyed reefs in Laulau Bay and Apra Harbor allowed for the identification of three common habits of *M. bouillonii* growth (Figure 1.7). The first growth form is cryptic and recessed, with colonies growing in crevices and extending deep into or under live or dead coral. The second is entrenched, where colonies extend from under horizontal shelves of coral. The third consists of semi-exposed tubes and chambers of *M. bouillonii* winding between and around upright coral structures, particularly the columnar growths of *Porites rus* and the branches of *Porites cylindrica*. It is common for large colonies to display multiple growth forms, which appear to be largely driven by the type of benthic structures available for the secure attachment of *M. bouillonii*, and so are highly plastic. On Finger Reef, *P. rus* is extremely prevalent, and serves as the major substrate for *M. bouillonii – A. frontalis* colonies in this area. Both the entrenched growth form, found extending from under *P. rus* plates, as well as tubes winding around the bases of columnar *P. rus*, were common. In contrast, Laulau Bay is much more varied in its benthic structure, with apparent higher coral species diversity, areas of substantial algal and cyanobacterial growth, and regions of hard substrate devoid of live corals. The cryptic growth form of *M. bouillonii* was particularly common at this site; shrimp-constructed colonies were frequently located in crevices, holes, and other areas not occupied by corals or other algae. Of the close to one thousand colonies that were counted during the abundance surveys and over three hundred fifty colonies documented in the substrate surveys combined over these various reefs, although many colonies were observed in direct contact with corals none were observed with growth morphologies suggesting overgrowth of these substrates. This suggests that *M. bouillonii – A. frontalis* colonies most likely do not interact with stony corals in the same manner as previously reported for the octocoral, *A. reticulata* near Okinawa (Yamashiro, Isomura and Sakai 2014). This was particularly notable in Laulau Bay, where there is substantial overgrowth on the reef by other species of cyanobacteria and macroalgae (Figure 1.8).

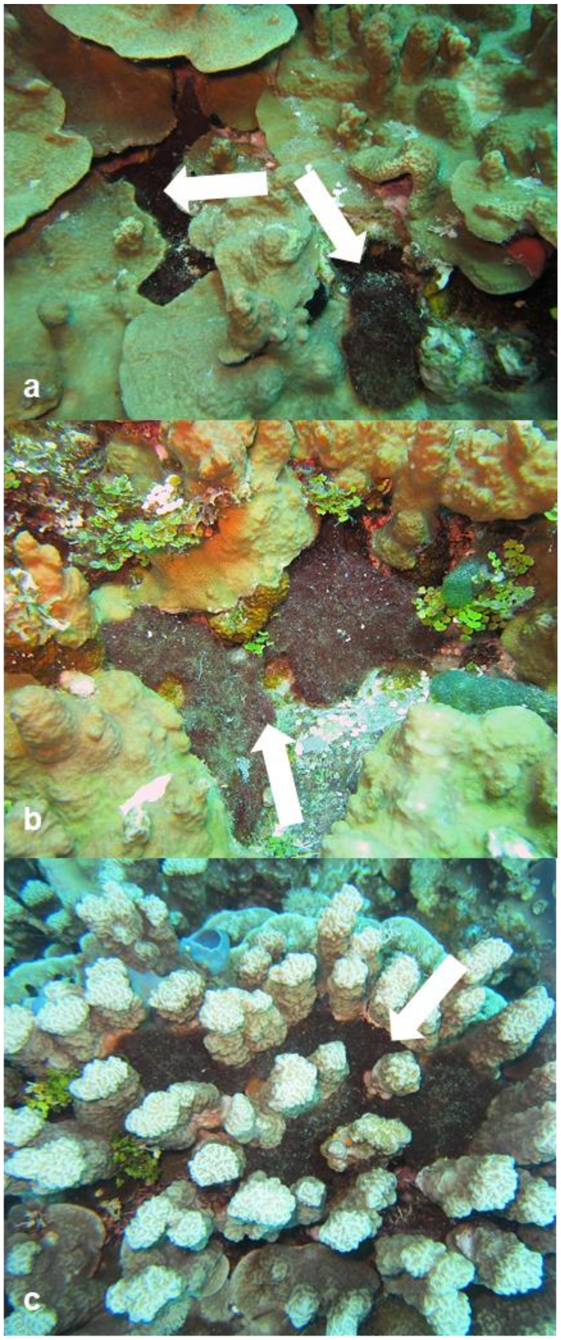


Figure 1.7 - Examples of *M. bouillonii* – *A. frontalis* colony growth habits observed in Saipan and Guam Colonies of *M. bouillonii* and *A. frontalis* take different growth morphologies related to the substrate they are occupying. Colonies grow a) cryptic and recessed in holes and crevices, b) entrenched under shelves of hard substrate, and c) semi-exposed and winding around columnar structures. Colonies are denoted by white arrows.

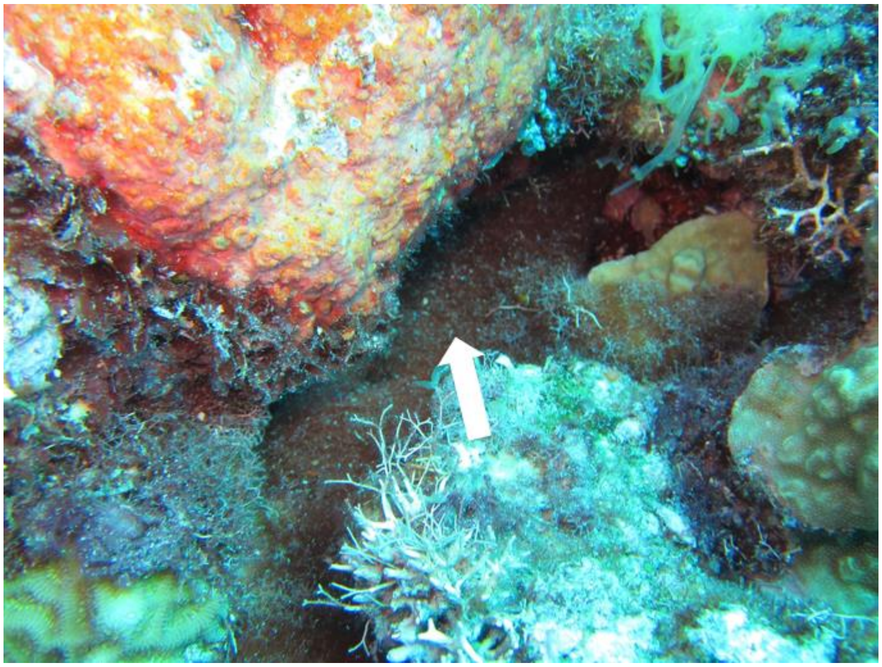


Figure 1.8 - *M. bouillonii – A. frontalis* colony growing in Laulau Bay reef crevice. Amidst other algae and cyanobacteria that are overgrowing parts of the reef in Laulau Bay, *M. bouillonii - A. frontalis* colonies can be found growing in an apparently benign fashion in interstitial and cryptic spaces. Colony is denoted by white arrow.

The three *M. bouillonii – A. frontalis* colony growth forms observed in this study share the common characteristic of occupying interstitial spaces within and around their substrates, rather than growing over outer exposed surfaces. As the most colonized substrates in our study, the stony corals *P. rus* and *P. cylindrica* are morphologically endowed with a high degree of such interstitial features. *P. rus* grows in a plate form, under which *M. bouillonii – A. frontalis* colonies are commonly entrenched, as well as a columnar form, around the bases of which tubes of woven *M. bouillonii* typically wind. Similarly, the branches of *P. cylindrica* provide structures around which tubes of shrimp-associated *M. bouillonii* commonly encircle. The columns of *P.* cf. *deformis*, as well as eroded crevices and columns in old reef structures, provide additional, less frequently inhabited refugia for these cyanobacterial-shrimp colonies.

Two notable substrates that were not documented as being occupied by *M. bouillonii – A. frontalis* colonies were soft corals and macroalgae. Along with *P. rus* and *P. cylindrica*, soft coral species of the genus *Sinularia* have been documented to be prevalent at Piti Bomb Holes (Gochfeld 2010), and these soft corals were noted along the transects at this site. Even though *Sinularia* species have branching structures providing substantial interstices, *M. bouillonii – A. frontalis* colonies were not found growing amongst them. In terms of macroalgae, along transects on the Finger Reef slope, tufts of *M. bouillonii* growing independently of shrimp were observed interspersed with *Halimeda* sp. on the underside of overhanging reef structures. The lack of *M. bouillonii – A. frontalis* colonies growing in this fashion might hint at the process by which these colonies are initiated. Shrimp may be collecting free filaments from the underside of overhangs and then moving the filaments to interstitial spaces to begin weaving, resulting in colonies tending to be found in the cryptic benthic spaces known to be favored by snapping shrimp (Johnson, Everest and Young 1947). However, more studies are needed to understand the origins of *M. bouillonii - A. frontalis* colonies.

When observations, such as the anecdotal data compiled in the supplemental knowledge aggregation questionnaire (see Supplemental Methods and Results in 1.4: Appendix: Supplemental Information), are combined with the gathered field data, a picture of *M. bouiilonii - A. frontalis* colony growth habits emerges that is strikingly similar to that described in historical studies. Both Cowles (1913) and Fishelson (1966) described this species pair as being found under coral rubble and stones, while Hoffmann and Demoulin (1991) assert that *M. bouillonii* habitually grows in the gaps between coral structures and in cavities in the reef, highlighting the cryptic nature of *M. bouillonii – A. frontalis* colony growth, and the tendency of these colonies to limit exposure by occupying interstices. Others have described shrimp-associated *M. bouillonii* growing as mats (Engene, Coates and Gerwick 2010; Engene et al. 2012) and tubes (Tidgewell et al 2010) that were attached to debris such as rocks and wood, or growing between coral structures (Engene, Coates and Gerwick 2010; Matthew, Schupp and Luesch 2008).

The most notable difference between the sites surveyed in the present study and that which was studied in Aka Jima by Yamashiro et al. (2014) is that no colonies were found to be associated with gorgonians or other soft corals in Guam or Saipan, whereas colonies in Aka Jima were reported to display an overgrowth morphology exclusively on the surfaces of gorgonians, despite scleractinian coral substrates being available (Yamashiro, Isomura and Sakai 2014). All colonies of *M. bouillonii* and *A. frontalis* encountered thus far within the Mariana Islands displayed variable, substrate-dependent growth forms that favor occupation of interstitial spaces on coral reefs. This suggests that *M. bouillonii* - *A. frontalis* colonies may not exhibit a similar propensity to overgrow habitats dominated by hard coral as to that reported for habitats dominated by octocoral (Yamashiro et al. 2014). While neither the current study nor the Yamashiro et al. (2014) work directly measured growth dynamics, the static growth morphologies documented between these studies contrast sharply.

Although unusual, evidence of gorgonian tissue damage, and apparent boring and anchoring of cyanobacterial filaments into gorgonian tissue at Aka Jima strongly suggest that the shrimp could sustain *M. bouillonii*-*A. frontalis* colonies long enough to overgrow *A. reticulata*. What are not clear are the factors that may have influenced relatively cryptic, interstitial colonies to transition into conspicuous colonies inhabiting sea fans in Aka Jima. One hypothetical explanation could be that colonies at Aka Jima were initiated by a mechanism alternative to *A. frontalis* weaving *M. bouillonii* directly on the sea fans, such as a typhoon or other dramatic hydrographic activity that could dislodge colonies from the benthos. Such colonies could than have become entangled in the finely branched arms of gorgonians, and subsequently sustained by a shrimp which survived the transplant. Consistent with this hypothesis, Aka Jima has been described in the literature and field station records as having a high frequency of typhoons and other strong oceanographic conditions (Iwao 2018). Dislodgement of benthic cyanobacteria from their growth substrate by wave action has been previously reported in the scientific literature (Becerro, Bonito and Paul 2006). Yamashiro et al. (2014) report that other algal species were also tangled amongst the gorgonians, and gorgonians are well known to position their growth axis perpendicular to water flow in order to facilitate filtration of food particles from the water column (Leversee 1976). Thus, it is conceivable that dislodged benthic cyanobacterial colonies could become entangled in the finely branched structure of a gorgonian coral.

Furthermore, the observed patterns in pigmentation change, considered alongside the nutrient analyses, help to better characterize the commensal interactions occurring between *M. bouillonii* and *A. frontalis,* and provide further evidence for the notions presented by Yamashiro et al. (2014) in that with *A. frontalis,* the chemically defended *M. bouillonii* could persist on sea fans in nutrient-depleted waters. *M. bouillonii* is potentially gaining, through its association with *A. frontalis,* the benefit of greater access to nutrients via shrimp excrement. By the end of the experiment, the pigmentation of the non-shrimp-associated cyanobacteria being grown with a shrimp matched that of the shrimp-associated cyanobacteria being grown with a shrimp, suggesting that non-shrimp-associated cyanobacteria with exposure to shrimp had access to more nutrients, and so were able to further develop their nitrogen-rich pigmentation (e.g. chlorophyll a and phycoerythrin). In contrast, shrimp-associated cyanobacteria that were denied access to shrimp were negatively impacted by their new nutrient-poor environment and required approximately two weeks to acclimate to the new conditions. This interpretation is further supported by the three replicates of shrimp-associated cyanobacteria growing with shrimp whose shrimp-driven nutrient influence was removed during the course of the experiment via death or escape; in all three of these cases, cyanobacterial pigmentation deteriorated as the nutrient-supplying shrimp was no longer actively excreting amongst the filaments. Further support can be found in the results of the nutrient analyses, which found statistically significantly higher levels of phosphate, nitrate, nitrite, and ammonia within the tubes of *M. bouillonii - A. frontalis* colonies, as compared to ambient seawater. This suggests that enclosed tubes and chambers of *M. bouillonii* limit water exchange, allowing for shrimp excrements to accumulate and create a beneficial nutrient-enriched environment within their confines.

A. frontalis has been described as an obligate tube dweller that benefits from its association with *M. bouillonii* by deriving chemically defended shelter (Fishelson 1966; Cruz-Rivera and Paul 2002, 2006; Banner and Banner 1982). This benefit is exemplified during the collection of *A. frontalis*, as even very short periods of time during which shrimp are not completely enshrouded in cyanobacterial filaments can result in swift predation by triggerfishes and likely other invertivorous fishes (Leber and Gerwick, unpublished observations). Another reported benefit is use of *M. bouillonii* by *A. frontalis* as a food source; Fishelson (1966) reported observing *A. frontalis* eating filaments of *M. bouillonii* and detecting cyanobacterial cells in shrimp excrement. The changes in wet weight recorded in the culturing experiment, namely the statistically significant effect of shrimp presence on changes in wet weight, provide further support that one benefit shrimp are gleaning from their cyanobacterial home is a source of food. Cultures that included shrimp saw notable decreases in wet weight, resulting from shrimp feeding upon the filaments. This brings into focus a seemingly paradoxical situation where *M. bouillonii* appears to benefit from the nutrient-enrichment provided by *A. frontalis* that allows for the development of its rich pigmentation, while also being limited to a degree in its ability to increase in biomass as *A. frontalis* benefits from *M. bouillonii* as a food source.

While there is more to be learned about the association between *M. bouillonii* and *A. frontalis*, this report expands on the limited knowledge of the habitat ecology of this unique symbiosis through the consideration of patterns in growth morphology in the Mariana Islands. Specifically, we demonstrate that *M. bouillonii* and *A. frontalis* consistently occupy the interstitial spaces of coral structures and other hard substrates, rather than overgrowing the external surfaces of living coral colonies. Furthermore, this report gives more dimension to the commensal partnership between *M. bouilloni* and *A. frontalis* by providing additional evidence to explain how both organisms benefit from their intimate cooperation with each other.

Chapter 1, in full, has been submitted for publication of the material as it may appear in Leber, Christopher A.; Reyes, Andres Joshua; Biggs, Jason S.; Gerwick, William H. "Cyanobacteria-Shrimp Colonies in the Mariana Islands". The dissertation author was the primary investigator and author of this paper. The dissertation author co-conceived of the work, designed and implemented the surveys and experiments, participated in collections, performed all data analyses, and was the primary author of the work.

1.4: Appendix: Supplemental Information

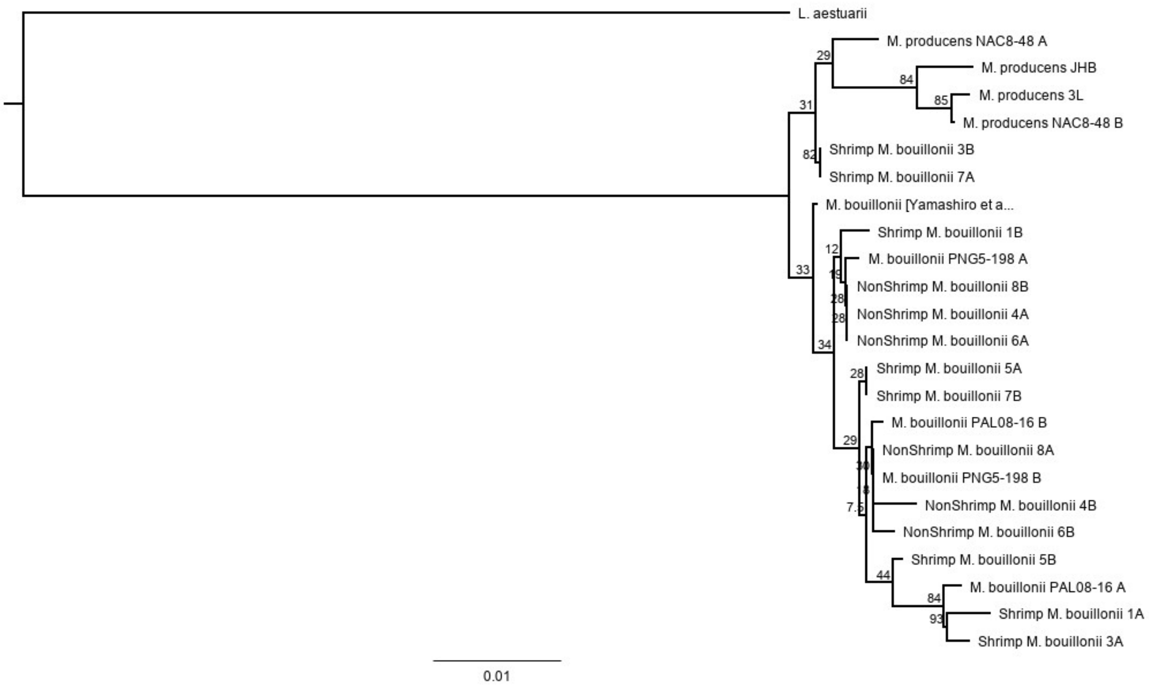


Figure 1.S1 - Phylogenetic tree of shrimp-associated and non-shrimp-associated *M. bouillonii* partial 16S rRNA gene sequences from Apra Harbor.

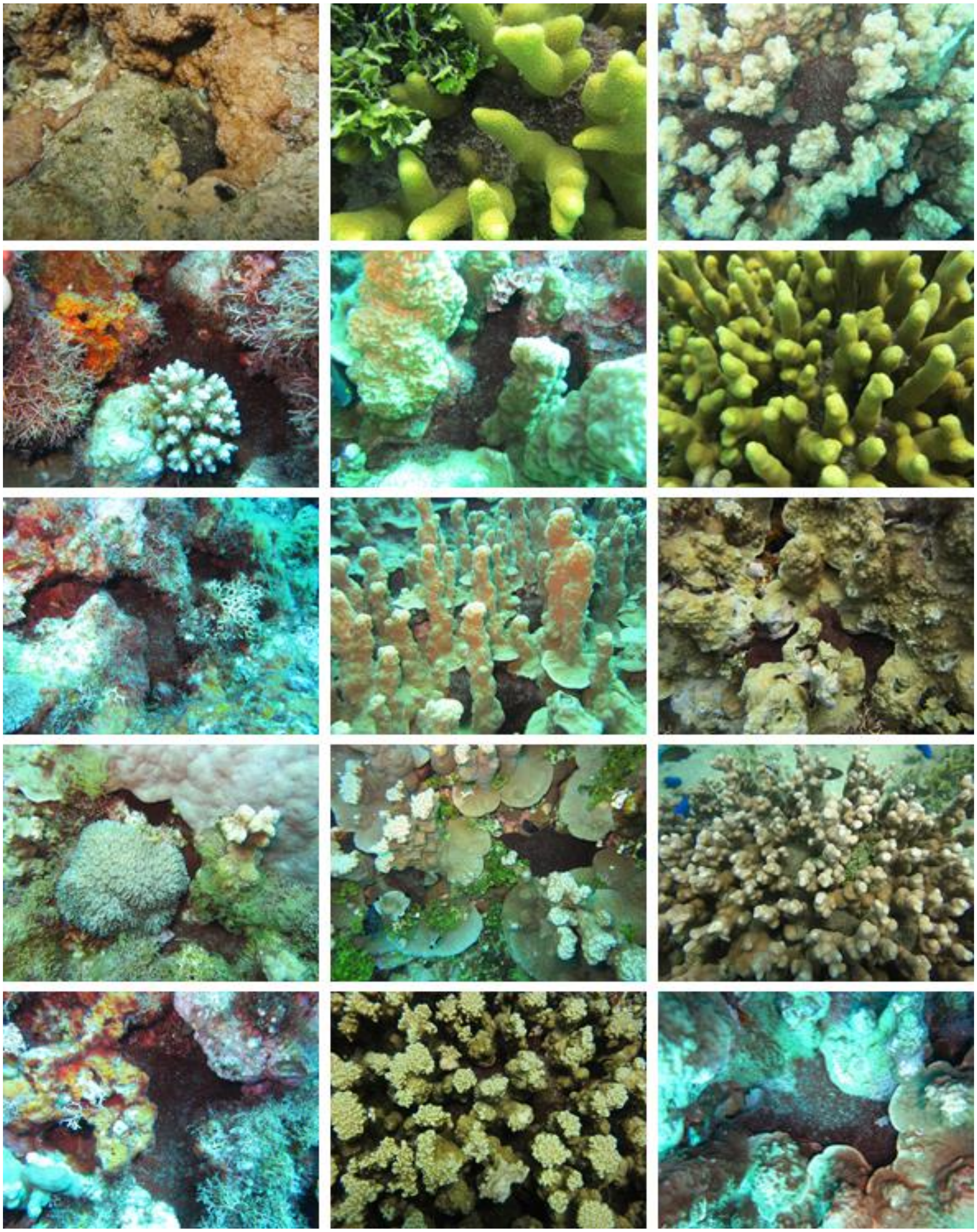


Figure 1.S2 - Representative photos of *M. bouillonii-A. frontalis* colonies on reefs in Saipan and Guam

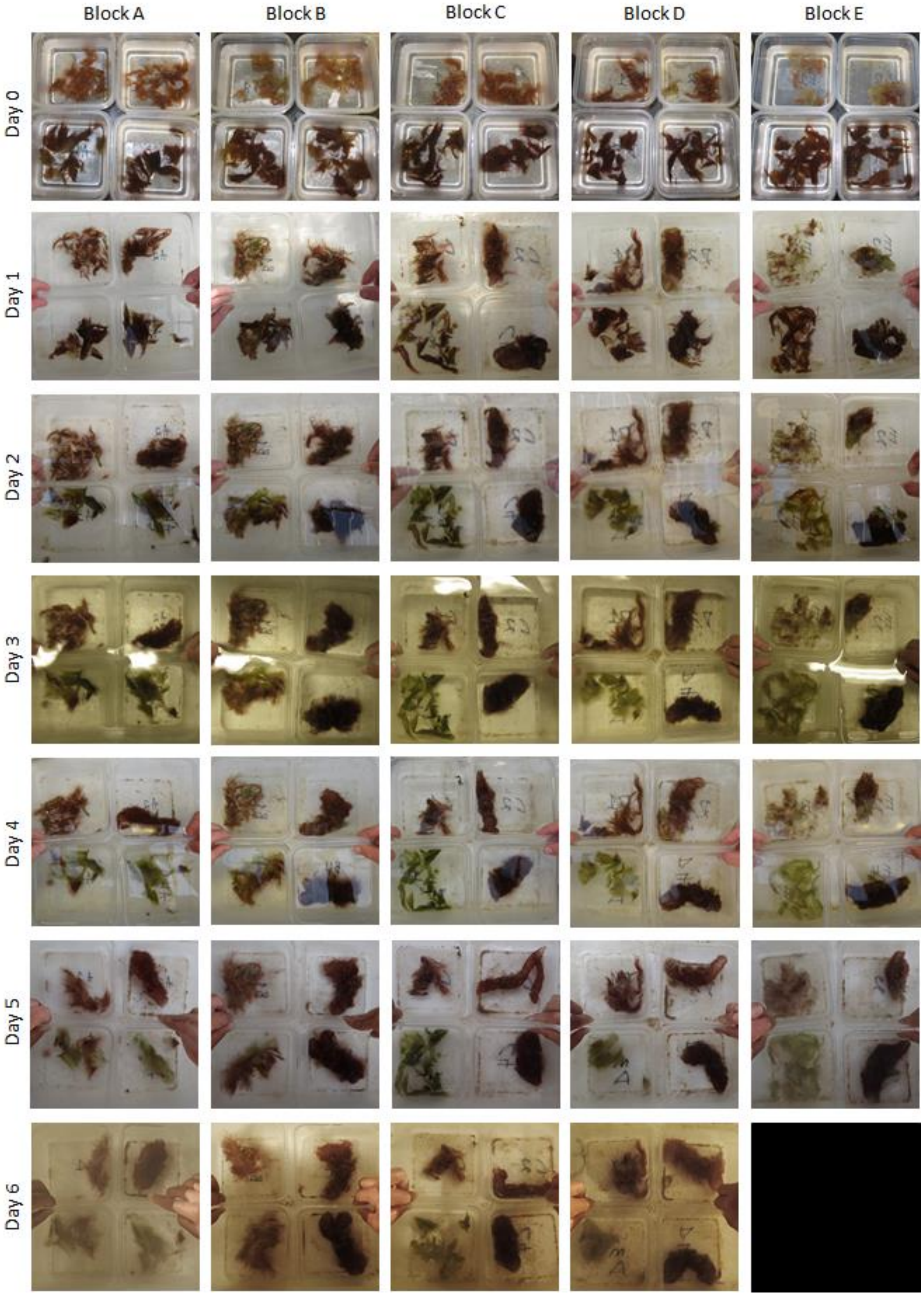


Figure 1.S3 - Photographic time series of *M. bouillonii* colony growth experiment. Clockwise from top left: non-shrimp-associated cyanobacterium without a shrimp, non-shrimp-associated cyanobacterium with a shrimp, shrimp-associated cyanobacterium with a shrimp, and shrimp-associated cyanobacterium without a shrimp.

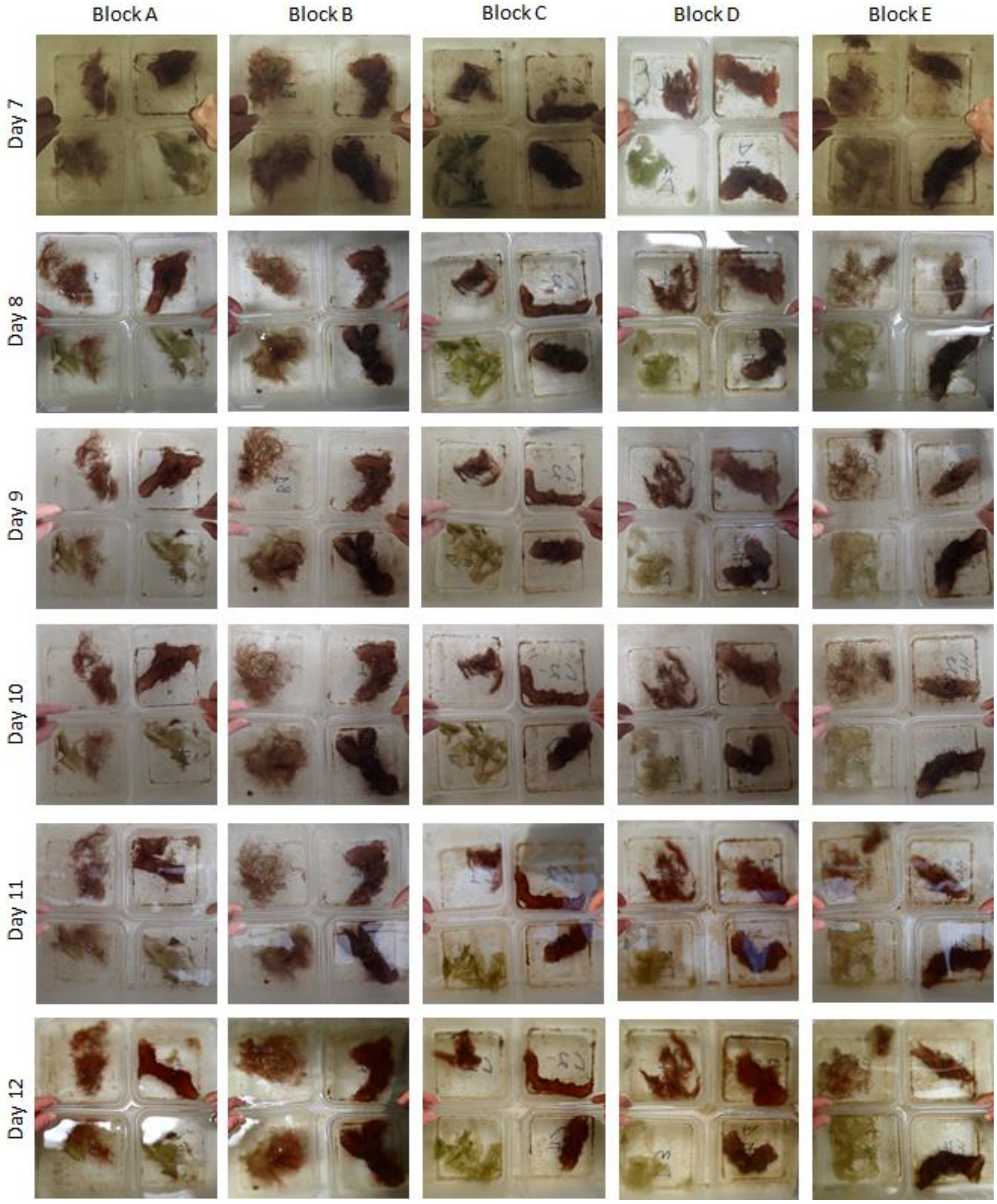


Figure 1.S3 (cont.)

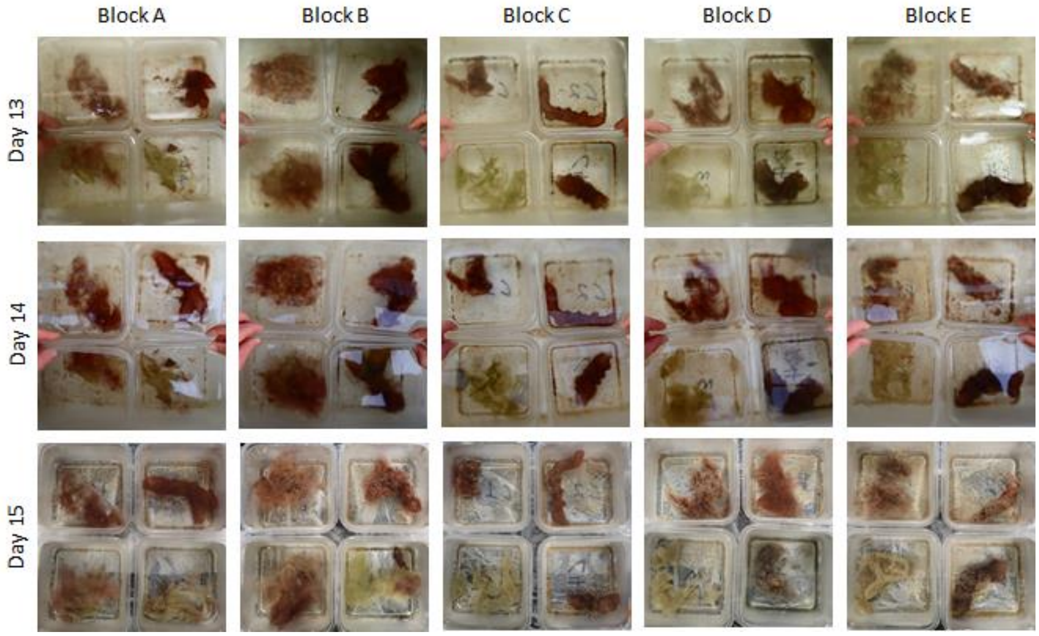


Figure 1.S3 (cont.)

Table 1.S1 - *M. bouillonii* – *A. frontalis* colony density and abundance

	Laulau Bay 1		Laulau Bay 2		Finger Reef 1		Finger Reef 2	
	Colonies	Colonies/m^2	Colonies	Colonies/m^2	Colonies	Colonies/m^2	Colonies	Colonies/m^2
0-10 m	84	4.20	67	3.35	30	1.50	75	3.75
10-20 m	90.5	4.53	70	3.50	10.5	0.53	32.5	1.63
20-30 m	116.5	5.83	57	2.85	2.5	0.13	20	1.00
30-40 m	77.5	3.88	45	2.25	2	0.10	4	0.20
40-50 m	104	5.20	62	3.10	1.5	0.08	1	0.05
Mean		4.725		3.01		0.465		1.325
s.d.		0.786		0.492		0.607		1.497

Table 1.S2 - *M. bouillonii – A. frontalis* colony substrate distribution

Substrate		Finger Reef Flat	Finger Reef Slope	Piti Bomb Holes 1	Piti Bomb Holes 2	Piti Bomb Holes 3	Merizo Fringing Reef 1	Merizo Fringing Reef 2
P. rus	Count[a]	73	32	13	0	0	6	1
	Proportion[b]	0.83	0.91	0.76	0.00	0.00	0.17	0.03
P. cylindrica	Count	2	0	2	37	99	28	30
	Proportion	0.02	0.00	0.12	0.97	0.93	0.78	0.97
P. cf. *deformis*	Count	0	0	0	0	0	2	0
	Proportion	0.00	0.00	0.00	0.00	0.00	0.06	0.00
Bare Reef	Count	13	3	2	1	7	0	0
	Proportion	0.15	0.09	0.12	0.03	0.07	0.00	0.00
Total		88	35	17	38	106	36	31

[a] Number of *M. bouillonii – A. frontalis* colonies counted along a transect occupying a particular substrate
[b] Proportion of all colonies counted along a transect occupying a particular substrate

Table 1.S3 - *M. bouillonii* colony growth experiment: changes in wet weight

Block	Factor: Cyano	Factor: Shrimp	Initial (g)	Final (g)	Difference
A	NS	NS	3.11	3.86	0.75
B	NS	NS	3.22	3.79	0.57
C	NS	NS	3.08	2.58	-0.5
D	NS	NS	3.31	3.52	0.21
E	NS	NS	3.35	4.15	0.8
A	NS	S	3.34	3.02	-0.32
B	NS	S	3.68	3.82	0.14
C	NS	S	3.11	1.97	-1.14
D	NS	S	3.11	2.02	-1.09
E	NS	S	3.7	3.07	-0.63
A	S	NS	3.38	3.35	-0.03
B	S	NS	3.38	4.02	0.64
C	S	NS	3.24	2.3	-0.94
D	S	NS	2.35	2.42	0.07
E	S	NS	3.57	3.13	-0.44

A	S	S	3.28	2.31	-0.97
B	S	S	3.14	3.02	-0.12
C	S	S	3.37	1.16	-2.21
D	S	S	2.64	2.02	-0.62
E	S	S	3.68	3.01	-0.67

Table 1.S4 - *M. bouillonii* colony growth experiment: Two-way ANOVA outputs

	sum of squares	df	F	P-value
factor - cyano	0.83	1.0	7.41	0.02
factor - shrimp	3.84	1.0	34.14	< 0.01
block	4.82	4.0	10.71	< 0.01
interaction - cyano:shrimp	0.05	1.0	0.43	0.53
residual	1.35	12.0	NaN	NaN

Table 1.S5 - *M. bouillonii-A. frontalis* water column and colony comparative water analyses nutrient data

	NO_3^-	PO_4^{3-}	Silicate	NO_2^-	NH_4^+
	µmol/L	µmol/L	µmol/L	µmol/L	µmol/L
water column 1	0.38	0.04	2.3	0	0.55
water column 2	0.31	0.03	2.4	0	0.43
water column 3	0.69	0.06	2.5	0.02	0.51
water column 4	0.34	0.03	2.2	0	0.4
water column 5	0.29	0.03	2.1	0	0.33
colony 1	1.59	0.19	2.3	0.07	0.95
colony 2	0.71	0.14	2.2	0.04	0.7
colony 3	1.46	0.21	2.2	0.06	0.69
colony 4	0.86	0.07	1.9	0.04	0.76
colony 5	0.54	0.13	1.9	0.04	0.29
p-value[a]	0.020000	0.005571	0.021743	0.001556	0.040447

[a] P-values from Student's T-test (paired, two tailed)

Table 1.S6 - *M. bouillonii* knowledge aggregation questionnaire questions and responses

Questions	Answers
During which years were you involved with *M. bouillonii* collection or field study?	2011 and 2013
	2000
	2006, 2009
	2000-2015
	1993-present
	2002-2004
	2008 to 2012
	2014, 2016
	March 2016 to July 2017
	2000 - 2016
	2005-2008
	2002 - 2009
	2012-2016
	2000-2004 (collection in 2002)
	2006
	2016 to present
	2007 & 2009
Where did you collect or study *M. bouillonii*?	Puerto Rico and Saipan
	Various reef systems at the northern coast of New Britain, Papua New Guinea.
	Papua New Guinea (near Alotau & Milne Bay); Palmyra Atoll
	Papua New Guinea, Costa Rica, Panama
	Guam
	Papua New Guinea (Bismark Sea) and Oahu Hawaii

	Palmyra Atoll
	American Samoa, Guam
	Apra Harbor, Guam; Mamaon Channel (Merizo), Guam; Piti, Guam
	Papua New Guinea, Palmyra Atoll, Lakadeep Islands, Guam, Saipan
	Panama (i think)
	Papua New Guinea - off the coast of Rabaul and New Ireland
	Panama, Saipan, American Samoa
	Papua New Guinea
	Papua New Guinea
	Guam, Saipan
	Papua New Guinea & Palmyra Atoll
How would you rank your ability to accurately identify *M. bouillonii* in the field? (1: no ability, 5: optimal ability)	3
	5
	5
	5
	5
	4
	2
	5
	5
	5
	3
	5
	5
	5
	5

	5
	5
In your own words, please describe *M. bouillonii* colony growth on coral reefs, including any patterns that you observed. (Responses modified to maintain anonymity)	A dense assemblage with lighter or bleached filaments and a minimal mucous layer near the exterior of the colony and dark brightly colored fresh filaments within the assemblage that were not directly exposed to the elements.
	This particular cyanobacterial species is usually found in shallow warm waters and is readily recognized in the field by its thallus morphology. The thalli form tenacious mats and are found attached to either dead corals of the Acropora species or other calcareous matter.
	Papua New Guinea - A reddish cobweb form in reef crevices (w/snapping shrimp); Palmyra - a reddish tube form in reef crevices. Both sites were ~20-50 ft depth
	Grows in small, flat mat across small openings in the reef, typically at depths between 10-25 meters. Often associated with snapping shrimp.
	The most obvious growth morph of M.b. is cocoon woven through and deep into living and/or dead corals (often Porites rus) by a shrimp. Free living morphs tend to be much more of a "puffball" with strings extending out. Both have a distinctive red to maroon pigmentation to them.
	It has been a long time, but I remember finding areas where it would form thick mats on dead coral.
	Upright tufts
	Usually arranged in shrimp nests. Occasionally free-growing or disorganized.
	First, the "free-growing tufts" (the unknitted algal strings) that we collected at Gab Gab Beach 1(Apra Harbor) had a greenish-red color, whereas the tufts (knitted by A. frontalis) within the coral had a dark red

	color. Second, the tufts that we collected at Middle shoals (Apra Harbor) fit the same profile. Most of the tufts we collected at this site were a dark red color and within coral, presumably knitted by the shrimp.
	grows as coverings over holes in reef, or as tubes around the base of corals and rocks
	as i recall, long tufts of dark green (in marine environment) filaments on sandy/loose substrate
	Attached to coral, grows in sheets with a cross-hatch pattern.
	The typical growth morphology found is the "woven" form, assembled by the pistol shrimp. This can appear slightly different in each case, but generally colonies will have a few tubes that run through the coral heads and connect larger chambers. These chambers can be somewhat buried into the coral, or in some cases exposed to the water column. Usually, the tubes and chambers are a deep red color and tend to be fairly clean from marine debris. The tunnels can be intricate, and depending on the coral they are growing on, they can be difficult to spot and collect. Additionally, some tufts of M.b. can be found unwoven, but it is much more rare. These typically look like a typical "hairball" of red filaments, and they tend to contain more discolored filaments.
	It always had a very cobweb-like appearance and was often draped over the opening of homes for snapping shrimp, about the size of a silver dollar. The shrimp would often make popping noises when you would steal their cover. When draped like that the surface was opaque and you could not see through it. The color of the algae was a deep wine-colored red. It was growing on the corals or had been placed there by the shrimp,

	but it was not extensively growing. Only showed up in pockets or small areas where we were collecting.
	[…]
	Grows with (in woven tubes and chambers, winding between coral heads and columns, through crevices in the reef, etc) and without shrimp (loose tufts, typically under overhanging coral structures)
	Moorea boillonii was often found as a patch of red filamentous growth among coral. It is a blood red color and fairly bright when a flash is used. The filaments are often in a somewhat round patch when observed in PNG and found weaving around coral in Palmyra. This pattern was discussed with Bill when it was observed in the field and we thought that it might have to do with the shrimp that is almost always found inside. The growth was never completely covering the coral it was found with. It was always a relatively confined growth to the interior of the coral or bottom. Not on the top.
On a scale from 1-5, how would you characterize the degree to which you observed *M. bouillonii* colonies overgrowing corals? (1: no overgrowth, 5: complete overgrowth)	2
	1
	1
	3
	1
	4
	3
	2
	1
	2

	2
	3
	1
	2
	2
	1
	3
On a scale from 1-5, based on what you observed in the field, how would you characterize the relationship between *M. bouillonii* colonies and coral reefs? (1: harmonious, 5: in conflict)	2
	2
	3
	1
	2
	4
	3
	2
	3
	2
	2
	2
	1
	2
	2
	1
	3
On a scale from 1-5, based on what you observed in the field,	2
	1

how would you characterize the impact of *M. bouillonii* colonies on coral reefs? (1: benign, 5: deleterious)	1
	3
	1
	4
	3
	2
	3
	1
	2
	2
	1
	2
	2
	1
	3
On a scale from 1-5, based on what you observed in the field, how would you characterize the degree to which *M. bouillonii* colony growth on coral reefs is recessed vs exposed? (1: cryptic & recessed, 5: overexposed)	1
	3
	2
	2
	1
	3
	2
	2
	3
	2
	3

	3
	3
	3
	5
	2
	1
Any additional observations on the growth of *M. bouillonii* that were not addressed in the above questions? (Responses modified to maintain anonymity)	In Saipan the largest assemblage was interwoven between a coral. There the assemblage didn't seem to be overgrowing the coral, I make that judgement because i didn't see it protruding out over the surface of the coral but woven under the exterior arms of the coral and because it didn't seem very tightly woven in with the coral, where it could be pulled out rather easily. That colony was also associated with snapping shrimp which snapped at us as we tore up the assemblage.
	Snapping shrimps are found to be associated with field collection of this cyanobacterial species.
	M. bouillonii is probably not obvious to the non-trained eye, but once you know what to look for you see it quite often. In Palmyra we only saw it on the open reef on SCUBA (one of the few cyanos out there), while other filamentous cyanobacteria were more common in the lagoon or lower flow/shallow areas (snorkeling or collecting by hand). I also don't think we saw it snorkeling in shallow areas in PNG, just on SCUBA.
	Growth of the cyanobacterium was opportunistic on the coral reef system, but not harmful and provides habitat for snapping shrimp.

	I have known, worked on, and collected this cyano for decades and have a intimate knowledge of its presence on Guam. Within Apra Harbor, you often see "tubes" of Mb woven through the architectural framework of columnar Porites rus collonies. The shrimp weaves this cocoon into a multi-branched network, and purposefully attaches this to points along the way. Much of the cocoon does not come into direct contact with live coral tissue, and when collected, the once shaded areas of the colony do not appear discolored. In fact, although the places in which these cocoons enter into the interior crevices are often "dead" (i.e., with a distinct colony edge) these entrance areas look the same in other colonies that are devoid of shrimp. This is the same for P.rus plates, as they are often "tucked" far enough back as to be growing where direct sunlight does not reach. Outside of Apra, Mb cocoons are also prevalent on both fore- and back-reef habitats. On fore-reefs, Mb can be most readily found in pockets and holes of the pavement and thus, it also exists quite readily where corals are not growing.
	It often growing as mats on coral sand as well as on dead coral. In an area north of Papua New Guinea I remember finding a lot of M. boullonii in areas that recently had undergone significant coral bleaching.
	none at this time

	Besides the places I collected it, I also saw M.b. in the Philippines (in Mabini, Batangas). I would guess it is pan-tropical, now that I know what to look for I can find it on almost any reef dive I take.
	[…]
	[…]
	I took many pictures of the colonies and patches that were observed in Palmyra. There were less pictures from PNG. I also happened to observe what I believe was M. bouillonii in Thailand during a vacation in 2013. I was near Ko Pi Pi on a relatively shallow dive near an overhang. The cobweb morphology and the color was obvious! I didn't collect the sample and I did not probe to see if a shrimp was observed.

Supplemental Methods – 16S rRNA gene sequencing

Sections of *M. bouillonii – A. frontalis* colonies and non-shrimp-associated *M. bouillonii* were collected in bulk from Apra Harbor, Guam on 10 June 2016 and 15 June 2017. RNAlater solution was used to store preserved biomass samples at -20°C until processing. One to two aliquots of biomass from each of four samples (two shrimp-associated and two non-shrimp-associated) RNAlater sample were blotted dry with paper towel and macerated with mortar and pestle in liquid nitrogen. Frozen homogenized aliquots were processed via the G20 Genomic Tip - Qiagen bacterial DNA isolation protocol, followed by cleanup via G20 genomic tip, by manufacturer's instructions (Qiagen). Universal specific cyanobacterial primers were used to amplify partial 16S sequences, using 106F (CGGACGGGTGAGTAACGCGTGA) and 781R(a) (GACTACTGGGGTATCTAATCCCATT) (Nübel, Garcia-Pichel and Muyzer 1997). Taq polymerase (Promega) was used, with an extension time of 1:00. Products were inserted into a cloning vector by TOPO-TA cloning kit (Life Technologies) into *E. coli* DH5α as per manufacturer's instructions. Colonies were generated overnight, transferred to liquid media and then grown overnight again. The QIAprep Miniprep kit and protocol (Qiagen) were used to harvest plasmids, followed by Sanger sequencing with M13 forward and reverse primers of two colonies per original aliquot. Sequences were compared using, and a phylogenetic tree was generated using Geneious (Geneious v2019.2). All sequences, including those featured for reference in the phylogenetic tree, are available via NCBI GenBank at the following accession numbers: *L. aestuarii* = NR_114680.1; *M. producens* NAC8-48 A = GU724199.1; *M. producens* NAC8-48 B = GU724200.1; *M. producens* JHB = FJ151521.1; *M. producens* 3L = FJ151527.1; *M. bouillonii* [Yamashiro et al.] = AB922817.1; *M. bouillonii* PNG5-198 A = FJ041298.1; *M. bouillonii* PNG5-198 B = FJ041299.1; *M. bouillonii* PAL08-16 A =

GU111927.1; *M. bouillonii* PAL08-16 B = GU182894.1; Shrimp *M. bouillonii* 1A = MT826199.1; Shrimp *M. bouillonii* 1B = MT826200.1; Shrimp *M. bouillonii* 3A = MK299234.1; Shrimp *M. bouillonii* 3B = MK299235.1; Non-shrimp *M. bouillonii* 4A = MT826201.1; Non-shrimp *M. bouillonii* 4B = MT826202.1; Shrimp *M. bouillonii* 5A = MK299236.1; Shrimp M. bouillonii 5B = MK299237.1; Non-shrimp *M. bouillonii* 6A = MT826203.1; Non-shrimp *M. bouillonii* 6B = MT826204.1; Shrimp *M. bouillonii* 7A = MT826205.1; Shrimp *M. bouillonii* 7B = MT826206.1; Non-shrimp *M. bouillonii* 8A = MT826207.1; Non-shrimp *M. bouillonii* 8B = MT826208.1.

Supplemental Methods – Knowledge Aggregation Questionnaire

In March of 2017, twenty natural products researchers with previous field experience with *M. bouillonii* were requested to complete a knowledge aggregation questionnaire designed to capture their observations on the distribution and patterns of *M. bouillonii* (See Supplementary Table 1.S6 for questions and tabulation of responses). The majority of scientific literature regarding *M. bouillonii* and *A. frontalis* is focused on the natural products chemistry of *M. bouillonii*, and collecting *M. bouillonii* from the field is a necessity for studying its natural products chemistry, suggesting that the population of natural products researchers responsible for these scholarly efforts could hold previously undocumented insights about *M. bouillonii* and *A. frontalis* colonies in an ecological context. The aggregation of recalled observations was administered via Google Forms. Seventeen responses were received, including those of the authors. Precedent for the use of this researcher questionnaire is provided by prior studies in which fishermen anecdotes were used to gain knowledge of historical fish abundances (Paterson 2010) and the reliance on surveys completed by the Florida dive community to assess the distribution of a blooming cyanobacterium (Paul et al. 2005). In the absence of substantial ecological study of *M. bouillonii* and *A. frontalis*, gaining insights from researchers with experience in collecting *M. bouillonii* and observing it in the field proved to be a useful approach in aggregating collective knowledge of the organisms' growth patterns and habits. Confidence intervals (95%) were calculated for answers to each question in the questionnaire, based on sample standard deviation using Microsoft Excel. There are limitations to this approach; the researchers who participated in this questionnaire have all previously worked in or with the labs of the study authors, potentially limiting the diversity and independence of perspectives captured by this questionnaire. Additionally, in some cases, participants in the questionnaire had not studied *M. bouillonii* for many years, leading to mistaken recollections about the organism (e.g. several respondents listed Panama, Puerto Rico, or Costa Rica as locations that they studied *M. bouillonii*; these locations are beyond the organism's native range).

Supplemental Results – Knowledge Aggregation Questionnaire

To gain a broader spatial and temporal perspective on the interactions of *M. bouillonii* and *A. frontalis* with corals, seventeen researchers with experience studying and collecting *M. bouillonii*, including the authors of this study, submitted answers to a knowledge aggregation questionnaire (Supplementary Table 1.S6). Their responses spanned twenty-four years of direct study of *M. bouillonii* (1993-2017, collectively over 100 years of study) and encompassed observations from the Philippines, Guam, Saipan, Papua New Guinea, Palmyra Atoll, American Samoa, and India's Lakshadweep Islands. Researchers were asked to rank their ability to identify *M. bouillonii* in the field, on a scale from 1 (no ability) to 5 (optimal ability). Responses ranged from 2 to 5, with a mean of 4.4±1.04 (s.d.) [4.5±0.94 (s.d.), including authors], suggesting a high level of familiarity and perceived competence in the identification of *M. bouillonii* amongst respondents.

Respondents to the questionnaire rated *M. bouillonii* growth on coral reefs based on four different characteristics, all on a scale of 1 to 5. These included the degree to which overgrowth was observed, perceived negative impact, perceived conflict with other organisms, and level of exposure. Together, these researchers reported that *M. bouillonii* has little tendency to overgrow corals (2.23±0.44, 95% C.I.) [2.00±0.44, 95% C.I., including authors], with four respondents emphasizing the growth of *M. bouillonii* as occurring on old coral structures and other hard substrates, five individuals describing *M. bouillonii* as covering or filling holes and crevices in the reef, and five researchers expressing that a good portion of *M. bouillonii* growth occurs through the interior of coral reef structures as opposed to overgrowing structures. Producing a similar distribution of answers (2.15±0.43, 95% C.I.) [2.00±0.44, 95% C.I., including authors], respondents characterized the impact of *M. bouillonii* on corals as not quite benign, but far from deleterious. When asked to characterize the *M. bouillonii*-coral relationship as harmonious (1) or in conflict (5), respondents judged the relationship to be more harmonious than conflicted (2.23±0.40, 95% C.I.) [2.18±0.38, 95% C.I., including authors]. Respondents described M. bouillonii as tending to mostly “not come into direct contact with live coral tissue,” to be “opportunistic on the coral reef system, but not harmful,” and as existing “quite readily where corals are not growing.”

There was slightly less agreement by respondents when asked to assess *M. bouillonii* growth as cryptic and recessed (1) versus exposed (5). Answers of 1, 2, 3, and 5 were all given, with a mean of 2.54±0.50 (95% C.I.) [2.41±0.48, 95% C.I.]. This variability is indicative of the array of different growth patterns described by respondents and observed in different locales. When associated with shrimp, *M. bouillonii* is described as growing with a woven, knitted, or cobweb-like appearance that forms mats, tubes, and complex chambers. Without the shrimp, *M. bouillonii* is morphologically characterized as a 'puff ball', a 'hair ball', or simply as a loose tuft of filaments. One respondent commented on the growth location of filaments not associated with shrimp, describing their tendency to be found dangling beneath overhanging reef structures. Descriptions of *M. bouillonii* growth when associated with shrimp were considerably more diverse; five respondents reported cyanobacterial colonies as growing recessed in or through coral reef structures, and four expressed a tendency of *M. bouillonii* tubes to be found winding around or encircling the base of calcareous structures, while one individual asserted that colonies are occasionally located in exposed areas.

Supplemental References

Geneious version 2019.2 created by Biomatters. Available from https://www.geneious.com

Nübel U, Garcia-Pichel F, Muyzer G (1997) PCR primers to amplify 16S rRNA genes from cyanobacteria. Appl Environ Microbiol 63:3327-3332

Paterson B (2010) Integrating fisher knowledge and scientific assessments. Anim Conserv 13:536-537. https://doi.org/10.1111/j.1469-1795.2010.00419.x

Paul VJ, Thacker RW, Banks K, Golubic S (2005) Benthic cyanobacterial bloom impacts the reefs of South Florida (Broward County, USA). Coral Reefs 24:693-697. https://doi.org/10.1007/s00338-005-0061-x

1.5: References

Banner DM, Banner AH (1982) The alpheid shrimp of Australia. Rec Aust Mus Suppl 34:359–362

Becerro MA, Bonito V, Paul VJ (2006) Effects of monsoon-driven wave action on coral reefs of Guam and implications for coral recruitment. Coral Reefs 25:193-199. https://doi.org/10.1007/s00338-005-0080-7

Carpenter EJ, Subramaniam A, Capone DG (2004) Biomass and primary productivity of the cyanobacterium *Trichodesmium* spp. in the tropical N Atlantic ocean. Deep Sea Res Part I Oceanogr Res Pap 52:173-203. https://doi.org/10.1016/j.dsr.2003.10.006

Cowles RP (1913) The habits of some tropical Crustacea. Philipp J Sci 8:119-125

Cruz-Rivera E, Paul VJ (2002) Coral reef benthic cyanobacteria as food and refuge: diversity, chemistry and complex interactions. In: Proceedings of 9th international coral reef symposium, vol 1, pp 515–520

Cruz-Rivera E, Paul VJ (2006) Feeding by coral reef mesograzers: algae or cyanobacteria?. Coral Reefs 25: 617-627. https://doi.org/10.1007/s00338-006-0134-5

Davie PJF, Short JW (2001) Decapod Crustacea of North East Cay, Herald Cays, Coral Sea. In: Herald Cays Scientific Study Report. The Royal Geographical Society of Queensland, Inc., Brisbane. Geography Monograph Series 6:79-84

Engene N, Coates RC, Gerwick WH (2010) 16S rRNA heterogeneity in the filamentous marine cyanobacterial genus *Lyngbya*. J Phycol 46:591-601. https://doi.org/10.1111/j.1529-8817.2010.00840.x

Engene N, Rottacker EC, Kaštovský J, Byrum T, Choi H, Ellisman MH, Komárek J, Gerwick WH (2012) *Moorea producens* gen. nov., sp. nov. and *Moorea bouillonii* comb. nov., tropical marine cyanobacteria rich in bioactive secondary metabolites. Int J Syst Evol Microbiol 62:1171–1178. https://doi.org/10.1099/ijs.0.033761-0

Fishelson L (1966) Observations on the littoral fauna of Israel, V. on the habitat and behaviour of *Alpheus frontalis* H. Milne-Edwards (Decapoda, Alpheidae). Crusteceana 11:98-104. https://doi.org/10.1163/156854066X00496

Frias-Lopez J, Bonheyo GT, Jin Q, Fouke BW (2003) Cyanobacteria Associated with Coral Black Band Disease in Caribbean and Indo-Pacific Reefs. Appl Environ. Microbiol 69:2409-2413. https://doi.org/10.1128/AEM.69.4.2409–2413.2003

Gochfeld DJ (2010) Territorial damselfishes facilitate survival of corals by providing an associational defense against predators. Mar Ecol Prog Ser 398**:**137-148. DOI: https://doi.org/10.3354/meps08302

Hoffmann L, Demoulin V (1991) Marine Cyanophyceae of Papua New Guinea. II. *Lyngbya bouillonii* Sp. Nov., A remarkable tropical reef-inhabiting blue-green alga. Belj J Bot 124:82-88

Iwao K (2018) Chapter 6b. – Status of coral reefs around the country: Kerama Islands. In: Iguchi A and Hongo C (eds.) *Coral Reefs of Japan*. Springer, pp 185-189

Johnson MW, Everest FA, Young RW (1947) The role of snapping shrimp (*Crangon* and *Synalpheus*) in the production of underwater noise in the sea. Biol Bull 93:122-138. https://doi.org/10.2307/1538284

Leão T, Castelão G, Korobeynikov A, Monroe EA, Podell S, Glukhov E, Allen EE, Gerwick WH, Gerwick L (2017) Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus *Moorea*. Proc Natl Acad Sci USA 114:3198-3203. https://doi.org/10.1073/pnas.1618556114

Leversee GJ (1976) Flow and feeding in fan-shaped colonies of the gorgonian coral, *Leptogorgia*. Biol Bull 151:344-356. https://doi.org/10.2307/1540667

Luesch H, Yoshida WY, Moore RE, Paul VJ, Corbett TH (2001) Total structure determination of apratoxin A, a potent novel cytotoxin from the marine cyanobacterium *Lyngbya majuscula*. J Am Chem Soc 123:5418-5423. https://doi.org/10.1021/ja010453j

Matthew S, Schupp PJ, Luesch H (2008) Apratoxin E, a cytotoxic peptolide from a Guamanian collection of the marine cyanobacterium *Lyngbya bouillonii*. J Nat Prod 71:1113-1116. https://doi.org/10.1021/np700717s

Milne-Edwards H (1837) Histoire Naturelle des Crustacés, Comprenant l´Anatomie, la Physiologie et la Classification de ces Animaux. In: Librairie Encyclopédique de Roret (1834-1840). Paris, vol 3, p 638, plates 1–42

Nagle DG, Paul VJ (1999) Production of secondary metabolites by filamentous tropical marine cyanobacteria: ecological functions of the compounds. J Phycol 35:1412-1421. https://doi.org/10.1046/j.1529-8817.1999.3561412.x

NOAA's environmental sensitivity index: Guam and the Northern Mariana Islands shapefile. (2005) National Oceanic and Atmospheric Administration (NOAA), National Ocean Service, Office of Response and Restoration, Emergency Response Division, Seattle, Washington. https://response.restoration.noaa.gov/maps-and-spatial-data/download-esi-maps-and-gis-data.html

Nomura K, Nagai S, Asakura A, Komai T (1996) A preliminary list of shallow water decapod Crustacea in the Kerama Group, the Ryukyu Archipelago. Bull Biogeogr Soc Japan 51:7-21

Paerl HW, Paul VJ (2012) Climate change: Links to global expansion of harmful cyanobacteria. Water Res 46:1349-1363. https://doi.org/10.1016/j.watres.2011.08.002

Poupin J (1998) Crustacea Decapoda and Stromatopoda of French Polynesia. Atoll Res Bull 451:1-62. https://doi.org/10.5479/si.00775630.451.1

Richters F (1880) Decapoda. In: Möbius K, Richters F and von Martens E (eds) Beiträge zur Meeresfauna der Insel Mauritius und der Seychellen. Gutmann'schen Buchhandlung, Berlin, pp 139-178, plates 15-18

Schorn MA, Jordan PA, Podell S, Blanton JM, Agarwal V, Biggs JS, Allen EE, Moore BS (2019) Comparative Genomics of Cyanobacterial Symbionts Reveals Distinct, Specialized Metabolism in Tropical *Dysideidae* Sponges. mBio 10: e00821-19. https://doi.org/10.1128/mBio.00821-19

Simões N, Apel M, Jones DA (2001) Intertidal habitats and decapod faunal assemblages (Crustacea: Decapoda) of Socotra Island, Republic of Yemen. Hydrobiologia 449**:**81-97. https://doi.org/10.1023/A:1017541019388

Sweet M, Burian A, Fifer J, Bulling M, Elliott D, Raymundo L (2019) Compositional homogeneity in the pathobiome of a new, slow-spreading coral disease. Microbiome 7:139. https://doi.org/10.1186/s40168-019-0759-6

Tan LT, Márquez BL, Gerwick WH (2002) Lyngbouilloside, a novel glycosidic macrolide from the marine cyanobacterium *Lyngbya bouillonii*. J Nat Prod 65:925-928. https://doi.org/10.1021/np010526c

Tidgewell K, Engene N, Byrum T, Media J, Doi T, Valeriote FA, Gerwick WH (2010) Evolved Diversification of a Modular Natural Product Pathway: Apratoxins F and G, Two Cytotoxic Cyclic Depsipeptides from a Palmyra Collection of *Lyngbya bouillonii*. Chembiochem. 11:1458–1466. https://doi.org/10.1002/cbic.201000070

Titlyanov EA, Yakovleva IM, Titlyanov TV (2007) Interaction between benthic algae (*Lyngbya bouillonii*, *Dictyota dichotoma*) and scleractinian coral *Porites lutea* in direct contact. J Exp Mar Bio Ecol 342:282-291. https://doi.org/10.1016/j.jembe.2006.11.007

Tronholm A, Engene N (2019) *Moorena* gen. nov., a valid name for "*Moorea* Engene & al." nom. inval. (Oscillatoriaceae, Cyanobacteria). Notulae Algarum 122:1–2

Yamashiro H, Isomura N, Sakai K (2014) Bloom of cyanobacterium *Moorea bouillonii* on the gorgonian coral *Annella reticulata* in Japan. Sci Rep 4:6032. https://doi.org/10.1038/srep06032

Chapter 2: Applying a Chemogeographic Strategy for Natural Product Discovery from the Marine Cyanobacterium *Moorena bouillonii*

2.0: Abstract

The tropical marine cyanobacterium *Moorena bouillonii* occupies a large geographic range across the Indian and western tropical Pacific Oceans and is a prolific producer of structurally unique and biologically active natural products. An ensemble of computational approaches, including the creation of the ORCA (Objective Relational Comparative Analysis) pipeline for flexible MS^1 feature detection and multivariate analyses, were used to analyze various *M. bouillonii* samples. The observed chemogeographic patterns suggested the production of regionally specific natural products by *M. bouillonii*. Analyzing the drivers of these chemogeographic patterns allowed for the identification, targeted isolation, and structure elucidation of a regionally specific natural product, doscadenamide A (**1**). Analyses of MS^2 fragmentation patterns further revealed this natural product to be part of an extensive family of herein annotated, proposed natural structural analogs (doscadenamides B–J, **2**–**10**); the ensemble of structures reflect a combinatorial biosynthesis using nonribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) components. Compound **1** displayed synergistic in vitro cancer cell cytotoxicity when administered with lipopolysaccharide (LPS). These discoveries illustrate the utility in leveraging chemogeographic patterns for prioritizing natural product discovery efforts.

2.1: Introduction

Natural products discovery programs operate with the general goal of detecting and characterizing chemically unique or biologically active substances. A common obstacle in discovery efforts is the rediscovery of known compounds, suggesting a need for tools and techniques that allow researchers to give priority to samples that possess new or otherwise interesting chemical substances. Various strategies have been employed for the dereplication of known chemicals within samples, and for the prioritization of samples based on chemical composition. In this regard, mass spectrometric analyses, usually in combination with liquid chromatography (e.g., LC-MS), have found great utility in natural products research due to the rapidity, small sample size requirements, and high amount of data generated. As a result, a number of approaches and algorithms have been developed to sift through LC-MS data so as to rapidly detect molecules of greater structural novelty and interest.

PoPCAR (Planes of Principal Component Analysis in R) applies principal component analysis (PCA) to a processed bucket table of sample features, selects outlying samples across different PCA planes, and then leverages the PCA feature loadings to identify the features that make the outlying samples unique (Chanana et al 2017). IDBac integrates proteomics and metabolomics data captured via MALDI-TOF MS applied to bacterial colonies on agar plates to classify bacterial strains and distinguish between closely related strains (Clark et al 2018). Global Natural Products Social Molecular Networking (GNPS) is a platform that facilitates the sharing of mass spectral data and provides tools for performing MS^2-based networking analyses (Wang et al 2016). GNPS continues to expand the repertoire of innovative approaches and techniques that it offers, with recent additions including a pipeline for Feature-Based Molecular Networking (FBMN) (Nothias et al 2020). FBMN utilizes a processed bucket table of sample

MS^1 features in conjunction with MS^2 fragmentation data to produce highly sensitive molecular networks well suited for quantitation and differentiation of isomeric compounds. In addition to these more specific tools, multiple tools are available for the processing and/or statistical analyses of MS-based chemical profile data, including XCMS (Gowda et al 2014), MZmine (Pluskal et al 2010), and Metaboanalyst (Chong et al 2018). The GNPS classical molecular networking approach (Wang et al 2016) is of particular note. While many approaches are sensitive to sample set heterogeneity and rely on specific or consistent sample preparations and data acquisitions in order to provide appropriate results, the classical molecular networking approach is much more flexible, and its outcomes are insulated from imperfect data. This allows classical molecular networking to be used in analyzing datasets that vary across numerous dimensions (instrument type, chromatographic method, sample preparation, etc.), providing many more opportunities for connecting disparate data sources.

The cyanobacterial genus *Moorena* (previously *Lyngbya*, then *Moorea*) is a prolific source of biologically active natural products, with biosynthetic gene clusters accounting for 18% of *Moorena* spp. genomes, on average (Engene et al 2012; Leão et al 2017; Tronholm and Engene 2019). Consistent with this finding, some 70 different isolated and structurally defined compounds have been reported from *M. bouillonii* (Table 2.S1) (Klein et al 1996, 1997, 1999a, 1999b; Luesch et al 1999, 2000a, 2000b, 2000c, 2001, 2002a, 2002b, 2002c, 2002d; Tan, Márquez, and Gerwick 2002; Williams et al 2003; Gutiérrez et al 2008; Matthew, Schupp, and Luesch 2008; Soria-Mercado et al 2009; Matthew et al 2010; Pereira, McCue, and Gerwick 2010; Rubio et al 2010; Tidgewell et al 2010; Choi et al 2012; Tan, Okino, and Gerwick 2013; Thornburg et al 2013; Mevers et al 2014; Kleigrewe et al 2015; Sumimoto et al 2016; Lopez et al 2017; Cai et al 2018; Nakamura et al 2018; Liang et al 2019; Mehjabin et al 2020; Sweeney-

Jones et al 2020). These display a broad structural diversity, and include peptides (Klein et al 1999a), cyclodepsipeptides (Luesch et al 2001), macrolides (Matthew et al 2010) and glycosidic macrolides (Tan, Márquez, and Gerwick 2002), and lipids (Mevers et al 2014). These compounds are also notable for their biological activities, including cytotoxins such as bouillonamide (Tan, Okino, and Gerwick 2013), lyngbouilloside (Tan, Márquez, and Gerwick 2002), multiple lyngbyabellins (Matthew et al 2010; Choi et al 2012), and the exquisitely potent apratoxin A (Luesch et al 2001). Other *M. bouillonii* compounds have been reported with cannabimimetic properties, such as columbamides A–C (Kleigrewe et al 2015) and mooreamide A (Mevers et al 2014), or as modulators of intracellular calcium mobilization such as alotamide A (Soria-Mercado et al 2009). *M. bouillonii* has a wide distribution across the tropical Western Pacific and Indian Oceans. However, *M. bouillonii* metabolites have only been described from collections made from a limited number of discrete locations, including Papua New Guinea (Klein et al 1996, 1997, 1999a, 1999b; Tan, Márquez, and Gerwick 2002; Gutiérrez et al 2008; Soria-Mercado et al 2009; Pereira, McCue, and Gerwick 2010; Tan, Okino, and Gerwick 2013; Mevers et al 2014; Kleigrewe et al 2015), Guam (Luesch et al 1999, 2000a, 2000b, 2000c, 2001, 2002c; Williams et al 2003; Matthew, Schupp, Luesch 2008; Matthew et al 2010; Rubio et al 2010; Cai et al 2018; Liang et al 2019), Palau (Luesch et al 2002a, 2002b, 2002c, 2002d; Williams et al 2003), Malaysia (Lopez et al 2017; Mehjabin et al 2020), Palmyra Atoll (Tidgewell et al 2010; Choi et al 2012), Fiji (Sweeney-Jones et al 2020) (The organism in this manuscript is reported as *M. producens*, however the manuscript includes a photo of the organism, which displays a morphology characteristic of shrimp-woven *M. bouillonii*. The 16S rRNA gene-based classification was inconclusive and known compounds previously isolated from *M. bouillonii* were reported.), the Red Sea (Thornburg et al 2013) (The organism in this manuscript is reported as *M. producens*, however the 16S rRNA gene-based classification is inconclusive and known chemistry associated with *M. bouillonii* was reported.), and the islands of southern Japan (Sumimoto et al 2016; Nakamura et al 2018). Collections from these diverse geographical regions differ substantially in their composition of metabolites, suggesting that even though many compounds are already known from *M. bouillonii*, comparing samples of different geographical origin could reveal distributional patterns in chemodiversity that would facilitate the identification of new natural products.

Much of the previous work connecting natural products chemistry and geography has focused on the latitudinal herbivory-defense hypothesis (LHDH). The LHDH suggests that tropical species display more developed defense phenotypes (including chemical defenses) than temperate species, due to higher levels of biotic stressors (Levin 1976; Coley and Aide 1991; Coley and Barone 1996). Studies in both terrestrial organisms (Levin 1976; Coley and Aide 1991; Coley and Barone 1996; Rasmann and Agrawal 2011) and marine organisms (Bakus and Green 1974; Hay and Fenical 1988; Bolser and Hay 1996) lend support to this hypothesis, but many examples counter to LHDH have also been reported, layering the theory with some degree of controversy while also revealing the complexity of drivers that influence chemical defense (Anstett et al 2016; Kooyers, Blackman, and Holeski 2017). Orthogonally, it has become a common strategy to look in underexplored geographical locations in order to find new and unique natural products. This has led natural products discovery efforts to interesting and exotic habitats, including tropical coral reefs (Klein et al 1996, 1997, 1999a, 1999b; Luesch et al 1999, 2000a, 2000b, 2000c, 2001, 2002a, 2002b, 2002c, 2002d; Tan, Márquez, and Gerwick 2002; Williams et al 2003; Gutiérrez et al 2008; Matthew, Schupp, and Luesch 2008; Soria-Mercado et al 2009; Matthew et al 2010; Pereira, McCue, and Gerwick 2010; Rubio et al

2010; Tidgewell et al 2010; Choi et al 2012; Tan, Okino, and Gerwick 2013; Thornburg et al 2013; Mevers et al 2014; Kleigrewe et al 2015; Sumimoto et al 2016; Lopez et al 2017; Cai et al 2018; Nakamura et al 2018; Liang et al 2019; Mehjabin et al 2020; Sweeney-Jones et al 2020), hypersaline lakes (Shang et al 2019), the Arctic (Marcolefas et al 2019) and Antarctic (Bory et al 2020), hydrothermal vents (Zhou et al 2019), and the deep sea (Zhang et al 2020). In spite of the acknowledgement that sampling in new geographical locations can allow access to new natural products, there are few examples of systematically applying geographical knowledge in order to inform natural product discovery. However, in one study the crude extracts and fractions from 300 geographically and taxonomically diverse cyanobacterial and algal collections were profiled by LC-MS/MS (Luzzatto-Knaan et al 2017). Analyses by GNPS classical molecular networking revealed geographic hotspots for chemodiversity, thus allowing for a molecular feature to be prioritized based on its chemogeographical distribution. In this case, it led to the characterization of a new metabolite given the common name yuvalamide A. Another example study focused on cyanobacteria from one specific genus, analyzing 10 samples of *Symploca* spp. collected at different times and in different places. This led to the efficient and targeted discovery of a new sample-specific bioactive natural product, samoamide A (Naman et al 2017).

In the present study, we illustrate the value of leveraging geographical patterns in chemodiversity to find previously uncharacterized natural products and apply this strategy to the marine filamentous cyanobacterial species *M. bouillonii*. This is a particularly interesting organism because of its wide geographical range and richness in natural products. To enable analyses and inform current discovery efforts based on legacy data, we were inspired to develop a flexible data pipeline described as the Objective Relational Comparative Analysis (ORCA) of chemical profiles from LC-MS data. Analyses of the LC-MS profiles from geographically disparate chemical extracts of *M. bouillonii*, used in conjunction with GNPS classical molecular networking, allowed for the prioritization of a molecular feature that led to the isolation and characterization of a new compound we called laulauamide (**1**) (The discovery, isolation, and structure elucidation of **1** were presented at the 2017 Annual Meeting of the American Society of Pharmacognosy. The name laulauamide was used for a poster presentation, and the associated abstract can be found under abstract P-219 at the following link [http://asp2017.org/wp-content/uploads/2016/12/ASP20201720Annual20Meeting_web.pdf]). Molecular networks along with detailed MS^2 fragmentation analyses revealed the presence of an extensive collection of proposed natural analogs. These display diversification through varied combinations of fatty acid side chains at two locations. Assays for biological activity yielded synergistic cytotoxic activity between **1** and lipopolysaccharide (LPS). Late in the performance of this work, a manuscript appeared from another laboratory that reported the isolation and structure elucidation of the main component of this new natural product family, and assigned it the common name "doscadenamide A" (Liang et al 2019), a name we retain so as to not create confusion in the literature record.

2.2: Results & Discussion

2.2.1: ORCA Pipeline

To allow for the comparison of LC-MS traces of extracts from different collections of *M. bouillonii*, a new pipeline was created called the Objective Relational Comparative Analysis (ORCA) pipeline (https://github.com/c-leber/ORCA) (Figure 2.1). ORCA is a flexible, modular pipeline that includes capabilities for simple and customizable MS^1 feature processing. ORCA can also accept any bucket table of samples vs. features as input, allowing for the comparison of data from any source that can be tabulated in such a manner. To accommodate heterogeneous data and to allow for the comparison of diverse datasets, ORCA MS^1 feature processing starts with an input directory of mzXML files, from which the MS^1 features are picked and integrated based on the mass-to-charge ratio (*m/z*) and a user-selected variant of retention time (rt). Feature picking is parameterized with the user-defined *m/z* and rt tolerances, and the peak size and shape parameters. Subsequently, MS^1 features picked from each sample file are consolidated based on the user-defined *m/z* and rt tolerance parameters and are organized into a samples vs. features bucket table containing feature integration values, with options to apply transformations based on the goals of the downstream analyses.

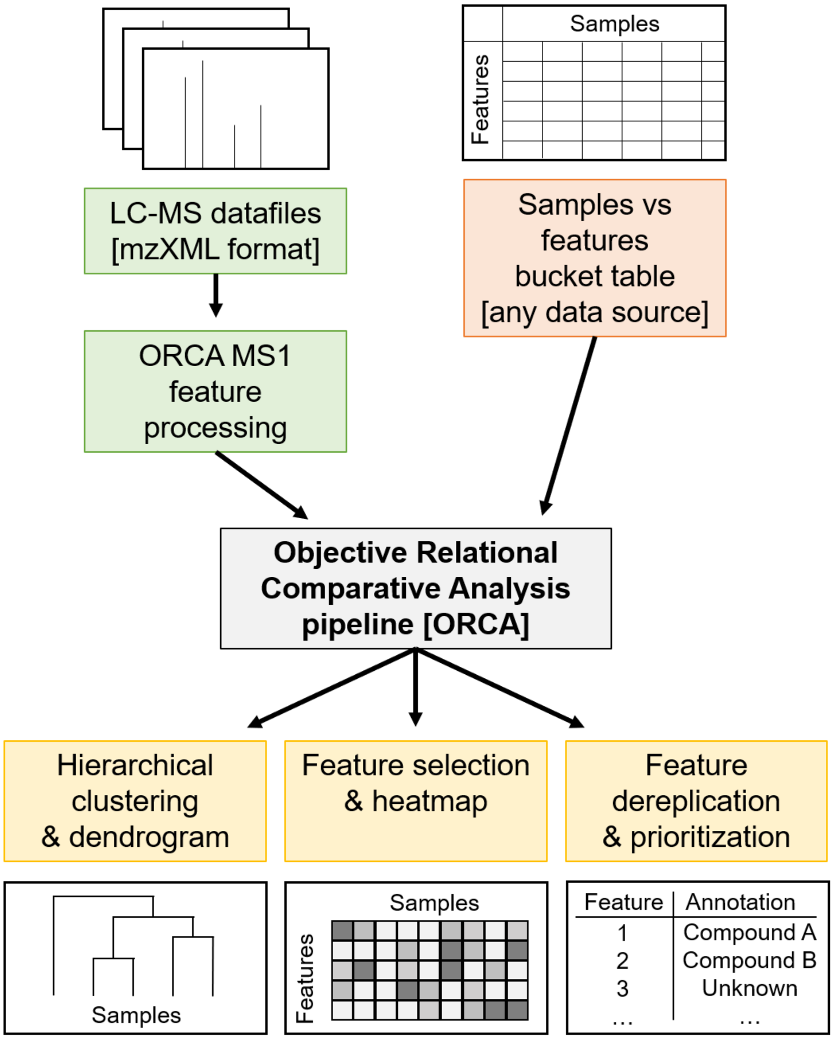

Samples
Features
LC-MS datafiles [mzXML format]
Samples vs features bucket table [any data source]
ORCA MS1 feature processing
Objective Relational Comparative Analysis pipeline [ORCA]
Hierarchical clustering & dendrogram
Feature selection & heatmap
Feature dereplication & prioritization
Samples
Samples
Features
Feature
Annotation
1
Compound A
2
Compound B
3
Unknown
…
…

Figure 2.1 - Illustration of the Objective Relational Comparative Analysis (ORCA) pipeline. The pipeline accepts inputs of either LC-MS datafiles in mzXML format, which can then undergo MS^1 feature processing, or an externally created samples vs. features bucket table coming from any data source. Analyses currently offered as a part of the ORCA pipeline include hierarchical clustering, feature selection, and feature dereplication based on user-provided reference data.

After processing of the sample MS^1 features, or the input of an externally generated sample vs. feature bucket table, the vectors of the feature values can then be utilized to initiate a diverse array of analyses, including hierarchical clustering of the samples to gain insights into the relationships between the samples, and univariate feature selection to learn about what specific MS^1 features are driving the differences between groups of samples. These analyses can then be visualized as dendrograms or heat maps, respectively. ORCA can also be used to generate a list of the most prominent MS^1 features across samples and to assign putative identifications from a user-supplied spreadsheet, allowing one to efficiently detect expected peaks across many samples, and to quickly determine the mass spectral signature of new potential isolation targets. ORCA was designed for assessing the relationships between heterogeneous samples and generating hypotheses regarding which features are driving these relationships; this makes ORCA a useful framework for not only learning about chemogeographical patterns, but also for comparing chemical profiles across different growth conditions (Crnkovic, May, and Orjala 2018), detecting contamination of botanical extracts (Pallarés et al 2019), identifying chemotaxonomic patterns (Engene, Tronholm, and Paul 2018), and many other potential uses.

2.2.2: Chemogeographic Analyses with ORCA and GNPS

Crude extracts of field-collected samples of *M. bouillonii* from American Samoa, Guam, Kavaratti (Lakshadweep Islands, India), Saipan, and the Paracel Islands (Xisha) in the South China Sea, as well as an in-house culture from Papua New Guinea, were profiled via LC-MS/MS; the resultant chromatograms were used as inputs for MS^1 feature processing in ORCA. Hierarchical clustering was performed on the MS^1 features and a dendrogram was produced with a cophenetic correlation coefficient of 0.905, indicating that the displayed structure in the dendrogram is highly correlated to the cosine distances between samples, and thus is representative of the data (Figure 2.2). The structure in the dendrogram suggests clustering of samples according to geographical region, a phenomenon that has previously been observed across other cyanobacterial samples (Luzzatto-Knaan et al 2017) but has not been specifically reported as a pattern for *M. bouillonii*. It is worth noting that, while samples with shared geographical origin are indeed arranged in clusters together in the dendrogram, the branch points for each geographical cluster are quite large, ranging from 0.4763 cosine distance for the two samples from Guam to 0.7248 cosine distance for where the three samples from Saipan converge. This is likely the result of a combination of the high variability and complexity in the composition of the studied samples, as well as the "curse of dimensionality" that artificially enlarges distance values when large numbers of features are being considered (Jayaram and Klawonn 2012). Classical molecular networking analysis using GNPS provided an orthogonal view, supporting the idea of chemogeographical specificity in these *M. bouillonii* samples, as numerous clusters of location-specific nodes are visible in the resultant network (Figures 2.3 and 2.S1). Furthermore, hierarchical clustering performed on presence-absence data of the MS^2 nodes from GNPS, as visualized with a dendrogram (Figure 2.4), revealed a chemogeographical clustering similar to that produced from the ORCA MS^1 features (Figure 2.2). The geographically associated structure in the data, as observable via both ORCA dendrograms, along with the presence of numerous location-specific clusters in the molecular network, led us to generate the hypothesis that the geographically specific distributions of natural products in our samples could be leveraged to identify previously unreported metabolites. The clustering of samples by specific geographical location stimulated further analyses to determine which molecular features were driving the observed geographic clusters, and which peaks were regionally specific. Particular attention was paid to Saipan, as it represented one region from which no new natural products from *M. bouillonii* had been reported in the scientific literature.

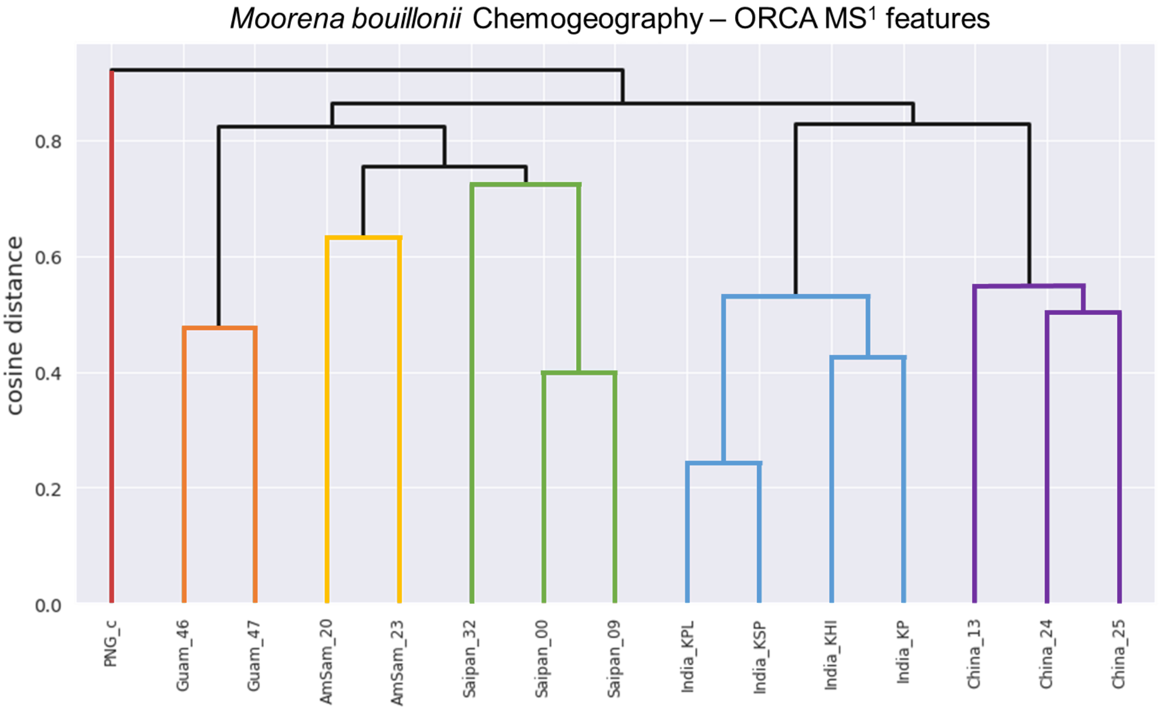


Figure 2.2 - ORCA-generated dendrogram (cophenetic correlation coefficient = 0.905) displaying the results of hierarchical clustering of the MS[1] features from *M. bouillonii* crude extracts. Samples are labeled with aliases comprising a general collection location concatenated to an abbreviated sample code. The structure in the dendrogram suggests that samples collected from the same geographical area are chemically more similar. Colorized for emphasis. Red: Papua New Guinea; Orange: Guam; Gold: American Samoa; Green: Saipan; Blue: Kavaratti (Lakshadweep Islands, India); Purple: Paracel Islands (Xisha) in the South China Sea.

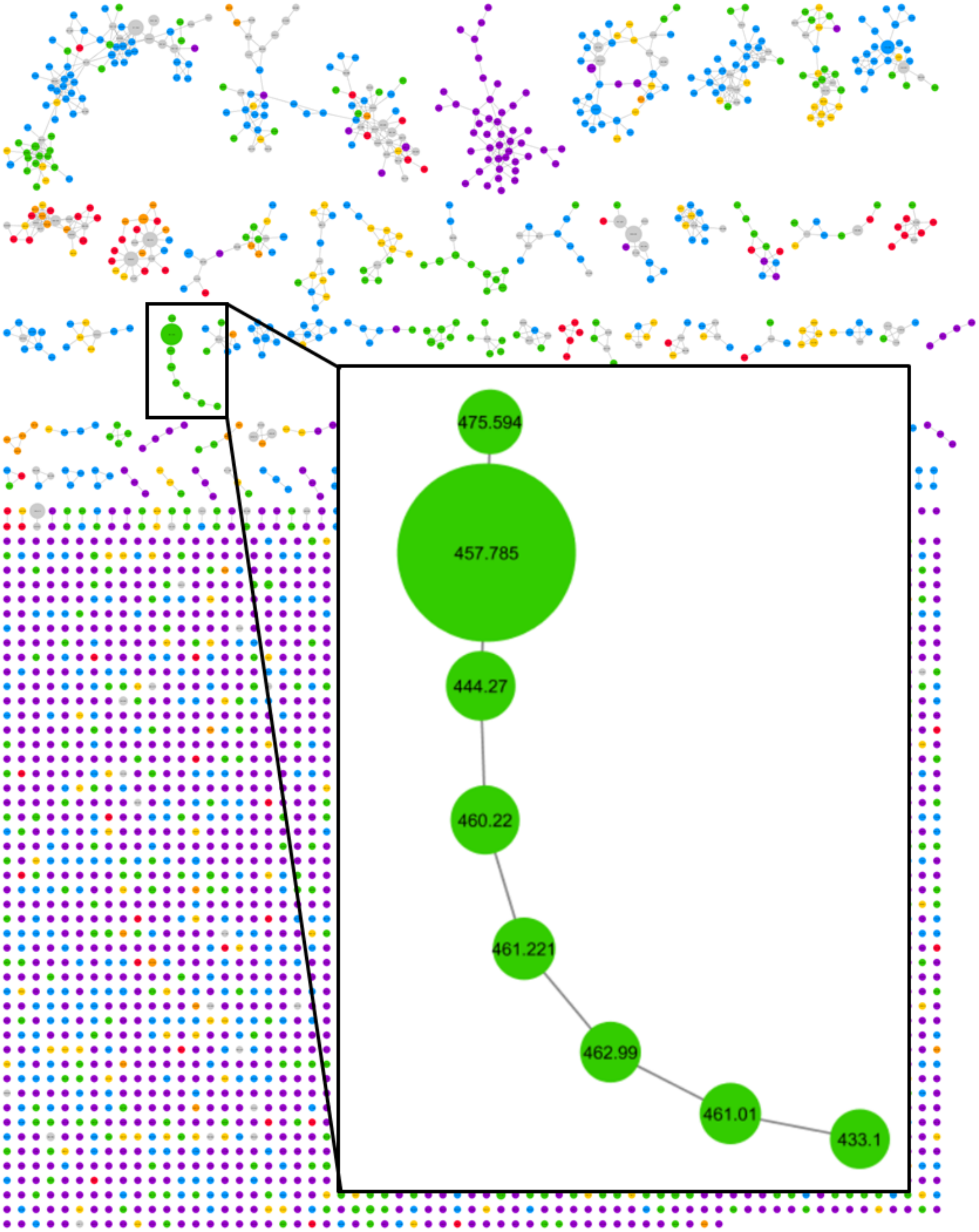


Figure 2.3 - Global Natural Products Social Molecular Networking (GNPS) classical molecular network of fifteen *M. bouillonii* crude extracts with the enlarged inset showing a cluster containing **1** (denoted with precursor mass *m/z* 457.785) and seven other nodes representing potential doscadenamide analogs (based on LR-MS/MS data). The green coloring of the nodes indicates that they represent features only detected in samples from Saipan. Nodes are scaled to summed precursor intensity. Grey nodes represent the MS^2 features that are present in samples from more than one geographical region. Geographical location of samples is colorized as follows: Red: Papua New Guinea; Orange: Guam; Gold: American Samoa; Green: Saipan; Blue: Kavaratti (Lakshadweep Islands, India); Purple: Paracel Islands (Xisha) in the South China Sea.

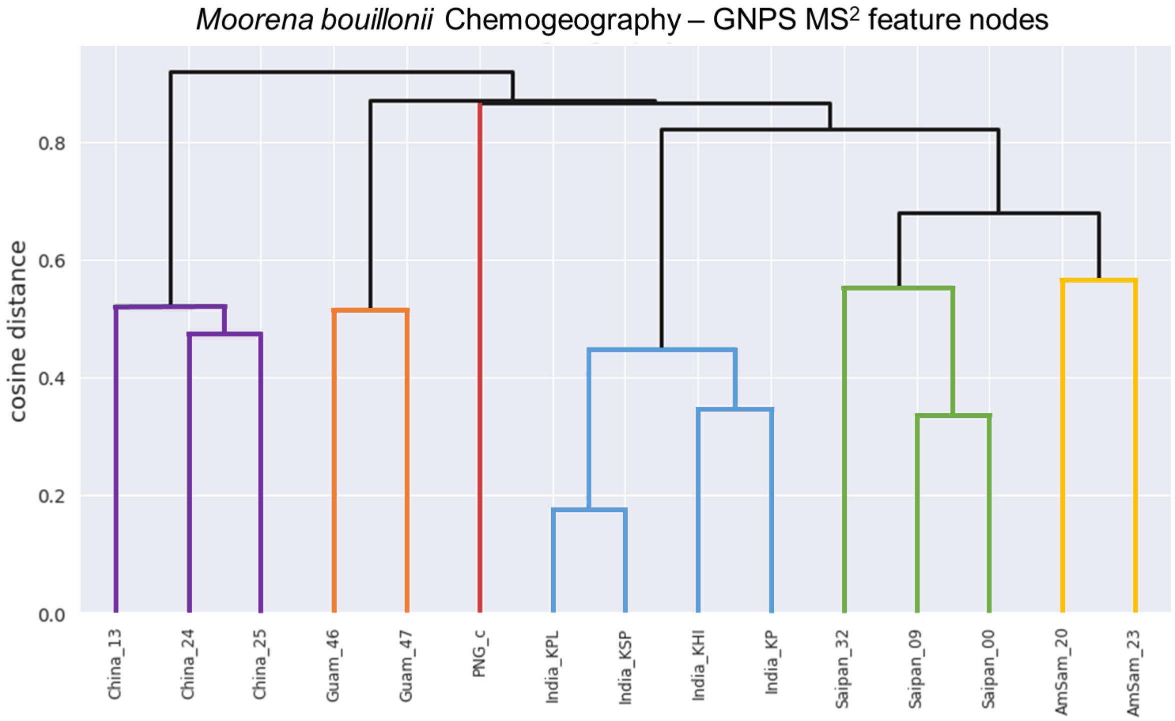


Figure 2.4 - ORCA-generated dendrogram (cophenetic correlation coefficient = 0.960) displaying the results of hierarchical clustering of the *M. bouillonii* crude extracts presence–absence data regarding GNPS nodes. Samples are labeled with aliases comprising a general collection location concatenated to an abbreviated sample code. Similar to Figure 2, the structure in the dendrogram suggests that samples collected from the same geographical area are chemically more similar. Colorized for emphasis. Red: Papua New Guinea; Orange: Guam; Gold: American Samoa; Green: Saipan; Blue: Kavaratti (Lakshadweep Islands, India); Purple: Paracel Islands (Xisha) in the South China Sea.

One cluster in the GNPS molecular network comprising only nodes originating from the Saipan-collected samples contained a particularly intense node for a feature with *m/z* 457.785 (Figure 2.3). Further investigations in ORCA revealed that this feature was present in high abundance in all samples from Saipan but was undetected or detectable at very low levels in the MS^1 spectra of samples from all the other studied locales (Table 2.S2). Additionally, this feature was not dereplicated when queried against all published compounds from *M. bouillonii* at the time (Table 2.S3) and when searched against the MarinLit database (http://pubs.rsc.org/marinlit/). This intriguing chemogeographic pattern prompted prioritization of this feature for isolation and structure elucidation, ultimately resulting in the characterization of a region-specific metabolite. Based on the specific collection site from which the Saipan samples originated (Laulau Bay, Saipan), we originally termed this metabolite "laulauamide". A molecular feature with *m/z* 721.10 was found to have a very similar geographic distribution. It was detected with high intensity in samples from Saipan, while being undetected or detectable at very low levels in other samples (Table 2.S2), and thus was another strong driver of the clustering of the Saipan samples. Isolation and analytical characterization revealed that this MS feature was the sodiated adduct of the known compound lyngbyapeptin A (Klein et al 1999a) (Table 2.S3).

2.2.3: Isolation and Structure Elucidation of Compound **1**

M. bouillonii biomass (1 L sample, 132 g dry biomass yielding 10 g of crude extract) from Saipan's Laulau Bay (denoted as Saipan_32 in Figures 2.2 and 2.4) was thoroughly extracted with 2:1 dichloromethane and methanol, and the resulting crude extract was fractionated over silica using vacuum liquid chromatography. LC-MS/MS analysis of the fractions revealed the MS^1 feature of interest to be in highest abundance in two relatively polar fractions. Reverse phase HPLC was used to initially isolate 1.5 mg of this compound from the two fractions. 1D and 2D NMR experiments were utilized to establish the planar structure of **1**, with major contributions from the ^1H-^1H Correlated Spectroscopy (COSY), ^1H-^{13}C Heteronuclear Single Quantum Coherence (HSQC), ^1H-^{13}C Heteronuclear Multiple Bond Coherence (HMBC), HSQC-Total Correlation Spectroscopy (TOCSY), and long-range ^1H-^{13}C Heteronuclear Single Quantum Multiple Bond Coherence (HSQMBC) data.

Compound **1** analyzed by high resolution electrospray ionization mass spectrometry (HRESIMS) suggested a molecular formula of $C_{27}H_{40}N_2O_4$ via observation of the sodiated molecular ion (observed *m/z* at 479.2877, calculated 479.2880), indicating 9 degrees of unsaturation. IR absorptions at 1724.31 and 3310.62 cm^{-1} were indicative of carbonyl and alkyne functionalities, respectively. An ultraviolet absorption at 217 nm was suggestive of an α,β-unsaturated carbonyl functionality. By ^{13}C NMR analysis, there were three carbonyls at shifts consistent with amides or esters, a highly polarized double bond consistent with one β-oxygenated enone (δ 94.2 and 179.2), and four acetylenic carbons (δ 84.7, 84.6, 68.52, and 68.50), accounting for 8 of the degrees of unsaturation, and thus indicating a monocyclic species.

The ^{1}H and ^{13}C NMR chemical shifts for the two acetylene groups were highly similar, and by HMBC correlations both had an adjacent methylene group at the same shift (δ 2.18, H_2-15 and H_2-24). In one of these two cases, sequential correlations deduced from the ^{1}H-^{1}H COSY data, and supported by the results of a ^{1}H-^{13}C HSQC-TOCSY experiment, provided a spin system involving three additional shielded methylene groups at δ 1.44, 1.39, and 1.75 and 1.42 (H_2-23, H_2-22, and H_2-21). The final of these methylene groups was positioned adjacent to a deshielded methine group at δ 3.75 (H_2-20). By COSY, the methine was determined to be adjacent to a shielded methyl group at δ 1.12 (H_3-27), and its chemical shift was explained by an HMBC correlation placing it adjacent to an ester or amide carbonyl (δ 176.4, C-19). The spin system of the second acetylene-terminating partial structure was highly similar and partially overlapped but terminated with a more shielded methine proton at δ 2.13 (H-11) with an adjacent methyl group (δ 1.11, H_3-18) and amide or ester carbonyl (δ 177.0, C-10). Summarizing, two essentially identical 2-methyl-7-octynoic acid structural units were thus defined from highly similar but non-identical data subsets.

The remainder of the molecule was thus composed of $C_9H_{14}N_2O_2$ with 3 degrees of unsaturation resulting from an enone and one ring structure. Two ^{1}H NMR singlets (δ 5.04, H-2 and 3.84, H_3-9) along with a 9-proton connected spin system remained unassigned. The singlet at 3.84 ppm was assignable to a methoxy group at the β-position of the enone by virtue of its relatively deshielded chemical shift and HMBC correlations to the highly deshielded olefinic carbon at δ 179.2 (C-3). The other singlet was thus assigned to the α-position of this enone as it was attached to a shielded olefinic carbon at δ 94.2 (C-2) and showed an HMBC correlation to the carbonyl carbon at δ 170.0 (C-1). As this partial structure accounted for all oxygen atoms in compound **1**, the shielded nature of this carbonyl necessarily required it to be attached to a nitrogen atom, forming an amide. Based on ^{1}H and ^{13}C NMR chemical shift data (δ 4.64, H-4; δ 59.2, C-4), one terminus of the remaining spin system was assigned to a methine with an attached nitrogen atom. ^{1}H-^{1}H COSY data, in conjunction with the ^{1}H-^{13}C HSQC-TOCSY, allowed formulation of four sequential methylene groups. The final methylene was also relatively deshielded (δ 3.22 and 3.13, H_2-8; δ 39.3, C-8), consistent with its attachment to a nitrogen atom. At this point, all atoms in the molecular formula of compound **1** were accounted for, except for one proton that was attached to a heteroatom by evaluation of the HSQC data (e.g., only 39 protons were found attached to carbon atoms); this was deduced to be an NH as three of the four oxygen atoms were assigned as carbonyls and one as a methylated enol.

HMBC correlations from the two diastereotopic protons at δ 3.13/3.22 ppm (H_2-8) to the carbonyl at δ 177.0 (C-10) connected these two partial structures. The other 2-methyl-7-octynoic acid was therefore connected to the only remaining heteroatom, the N-atom connected to the δ 170.0 (C-1) carbonyl of the enone functionality. Remaining structural features at this point included the formation of one ring, and placement of a proton on one of the two nitrogen atoms; two possibilities emerged (**1a** and **1b**) (Figure 2.5).

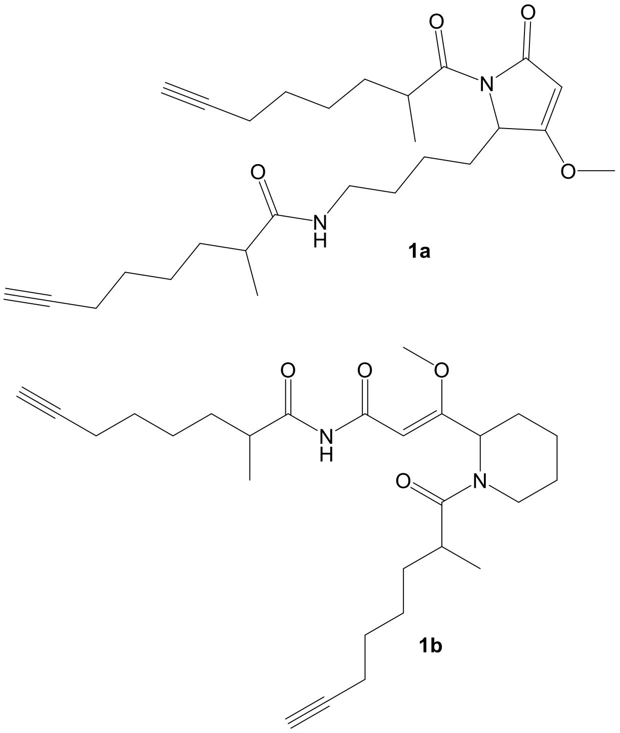


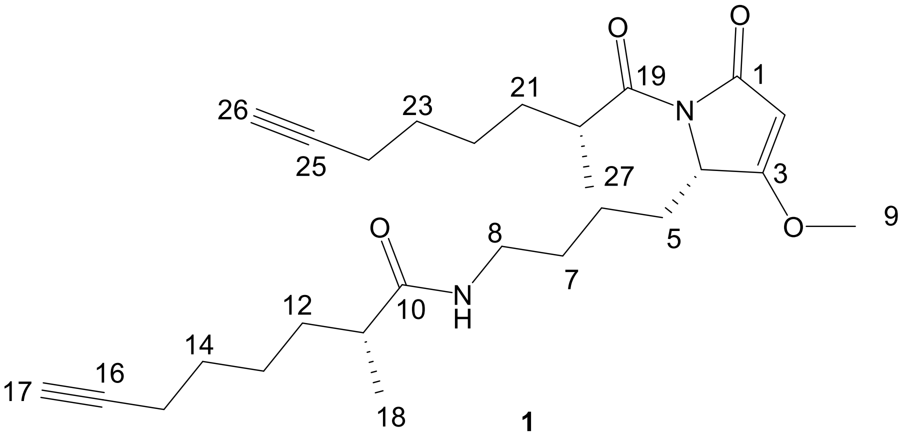
Figure 2.5 - Competing structural hypotheses for the two-dimensional structure of compound **1.**

Both structural possibilities had features that were attractive and unattractive from a predicted biosynthetic perspective. In **1a**, the fundamental assembly of the PKS derived octynoic acid; its passage to an NRPS to incorporate a lysine residue, followed by a ketide extension, *O*-methylation of the β-enol, and cyclization to a pyrrolidone ring, is well precedented within cyanobacterial natural products (Milligan et al 2000; Edwards et al 2004; Linington et al 2009). However, the acylation of a second octynoic acid residue to the lysine side chain nitrogen is an unprecedented event. Alternative structure **1b** has the attractiveness of a regular, predicted PKS(4)-NRPS(glycine)-PKS(3)-NRPS(glycine)-PKS architecture; however, it is quite awkward in requiring several unusual adjustments to the oxidation state of the carbon atoms, and creation of the second 2-methyl-7-octynoic acid residue via a completely different set of biosynthetic steps from the first one.

Modeling of these two alternative cyclization products for ^{13}C NMR shifts (see Figures 2.S2 and 2.S3 for the predicted ^{13}C NMR shifts for **1a** and **1b**, respectively) and comparison with those experimentally measured for **1** revealed that both possibilities were reasonably good fits, but the predicted values for 1a tended to be closer to the shifts experimentally derived for compound **1**. For both C-2, the methyl enol carbon (**1a** δ 95.5, **1b** δ 101.4, **1** δ 94.2), and C-3, the deshielded olefinic carbon (**1a** δ 180.7, **1b** δ 171.5, **1** δ 179.2), the fit for alternative **1a** was considerably better. Only at C-6 was the cyclization product proposed in **1b** favored (**1**a δ 24.3, **1b** δ 19.6, **1** δ 20.4). A deeper look into the long-range ^{1}H-^{13}C HSQMBC data was undertaken. The key proton distinguishing these two possible structures, H-4 at δ 4.64, showed correlations to several resonances, including two of the three carbonyl resonances (δ 170.0, C-1; δ 176.4, C-19) and the β-oxygenated enone (δ 179.2, C-3); these correlations were compatible with structure **1a** (one 2-bond and two 3-bond correlations), while in structure **1b** these correlations would result from one 2-bond, one 4-bond, and one 6-bond ^{1}H-^{13}C coupling. Furthermore, analysis of the HMBC correlations observed for H-4 and H_2-8, both from the lysine-derived residue, showed mutual signals with only C-5 at δ 29.0 and C-6 at δ 20.4 that would be consistent with either proposed structure. There were no shared correlations observed between these protons and the equally 3-bond proximal carbonyl in **1b**, nor 3-bond correlations from H-4 to C-8 and H-8 to C-4 that would be reasonably expected to be observed from **1b**, lending further support for **1a** as being the correct structure of **1**.

Compound **1** contains three stereocenters—two associated with the two 2-methyl-7-octynoic acid side chains, and one contained in the central heterocycle. A racemic standard of

2-methyloctanoic acid was derivatized with (*S*)-(+)-2-phenylglycine methyl ester. A chiral standard of (*S*)-2-methyloctanoic was generated via the zirconium-catalyzed asymmetric carbo-alumination (ZACA) reaction (Negishi et al 2004) of 1-octene to stereoselectively install a methyl group at the C-2 position, followed by an oxidation to 2-methyloctanoic acid and derivatization with (*S*)-(+)-2-phenylglycine methyl ester. Configuration of both 2-methyloctynoic acid moieties of compound **1** was established to be *R* through catalytic hydrogenation, acid hydrolysis, derivatization with (*S*)-(+)-2-phenylglycine methyl ester, and comparison via LC-MS to the generated standards of 2-methyloctanoic acid coupled with the same chiral auxiliary group (Figure 2.S4). Ozonolysis with an oxidative work-up (Pereira et al 2011), followed by acid hydrolysis, was used to open the heterocyclic ring structure and liberate lysine from compound **1**. The lysine was then derivatized with Marfey's reagent (L-FDAA) and compared to racemic and L-lysine standards derivatized with the same Marfey's reagent, indicating an *S* configuration of this residue (Figure 2.S5). The fully elucidated structure of compound **1** was thus determined as in Figure 2.6.



1

Figure 2.6 - Complete structure of compound **1**.

2.2.5: Annotation of Predicted Analogs **2-10**

Low-resolution LC-MS/MS fragmentation data for **1** consistently showed three peaks at *m/z* 321, 303, and 168 (Figure 2.S6), which we predicted to represent a side-chain loss, a side-chain loss plus the loss of an amine, and the loss of both side chains plus an amine, respectively. To better understand the fragmentations of compound **1** and use this information for identifying analogs based on repeating the MS^2 fragmentation patterns, high-resolution MS^2 fragmentation data were acquired for compound **1**. Numerous fragment peaks were recorded, including peaks observed at *m/z* 321.2171, 303.1901, and 168.1016. These values match very well to the calculated monoisotopic masses of the predicted fragment structures shown in Figure 2.7 (*m/z* 321.217, 303.183, 168.102; allowing for hydrogen rearrangements), lending support to our fragmentation hypothesis, and providing a starting point for understanding and proposing the structures of analogs via their fragmentation patterns.

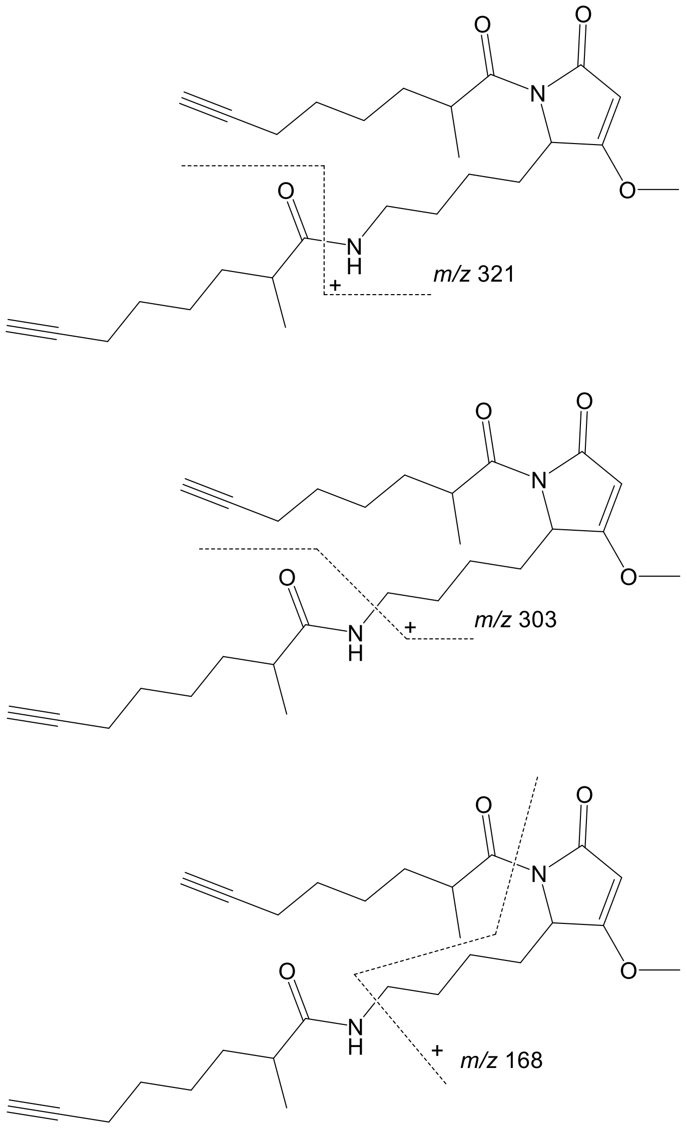
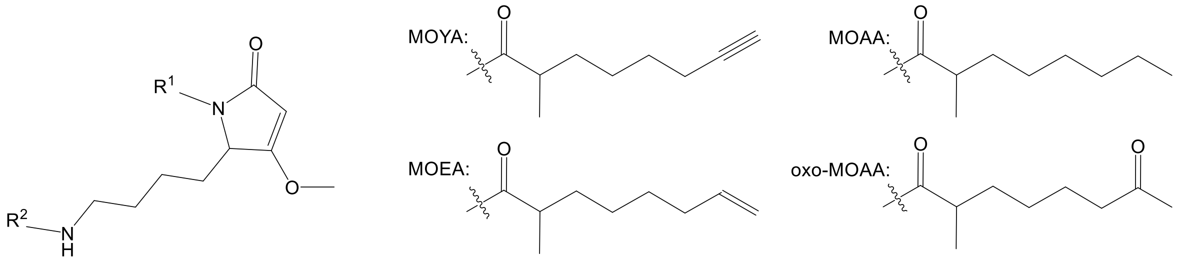


Figure 2.7 - Hypothesized fragment structures of compound **1**.

GNPS classical molecular networking placed compound **1** as a node in a cluster with seven other nodes originating from the Saipan *M. bouillonii* samples (Figure 2.3), suggesting several naturally occurring analogs were present. ORCA revealed that compound **1** is also present in samples from Guam, though detected with a much lower MS^1 intensity than in samples from Saipan. This inspired the generation of a more detailed molecular network composed of both crude extracts and fractions from a Saipan sample and a Guam sample (denoted as Saipan_32 and Guam_46 in the above dendrograms), revealing an even larger cluster of potential analogs that contained 33 nodes, including compound **1** (Figure 2.S7). Some nodes in the cluster had very similar masses, which could be the result of an artifact from the particular parameter set selected for the analysis, an artifact of the low resolution MS data analyzed, or be an indicator of isomeric analogs; therefore, further analysis was needed.

Analysis using the GNPS in browser network visualizer suggested that there was a common connection between many of the potential analogs (23 out of 33, including **1**), namely the presence of an MS^2 fragment peak at *m/z* 168 (Figure 2.S7). To facilitate further analysis of MS^2 spectra and the presence of potential analogs, the ORCA MS^2 auxiliary pipeline was developed. MS^2 scans from the Saipan and Guam crude extract and fractions were binned based on precursor mass, and then filtered to only precursor masses with scans that included a *m/z* 168 fragment peak. Clustering scans from each relevant precursor mass by cosine distance, paired with manual analysis, allowed the structures of 9 analogs (**2**–**10**) to be proposed (Figure 2.8; see Figures 2.S8–2.S26 for the proposed structures, consensus spectra, and predicted fragment structures). It must be noted that alternative structural proposals are conceivable for these analogs; however, given the literature precedent for cyanobacteria to produce families of natural products with the same array of variations in desaturation and oxidation as proposed here, e.g., (Hooper et al 1998; Edwards et al 2004; Boudreau et al 2012), and the predictable MS^2 fragmentation spectra observed, these proposals represent the most parsimonious and best supported structural hypotheses. Ambiguities in the remaining related MS^2 spectra prevent the definitive assignment of carbon chain isomers and positional isomers, and the proposal of additional analogs, but suggest a process of combinatorial biosynthesis in generating this expansive natural product family. While quantities of these minor metabolites in our samples were not sufficient for isolation and further characterization, the total synthesis published alongside the characterization of **1** (Liang et al 2019) is very amenable to incorporating alternative side chains, and this could be used for generating these proposed analogs for further study.



1 doscadenamide A: R^1 = MOYA, R^2 = MOYA; **2** doscadenamide B: R^1 = MOEA, R^2 = MOEA;
3 doscadenamide C: R^1 = MOYA, R^2 = MOEA; **4** doscadenamide D: R^1 = MOEA, R^2 = MOYA;
5 doscadenamide E: R^1 = MOYA, R^2 = MOAA; **6** doscadenamide F: R^1 = MOAA, R^2 = MOYA;
7 doscadenamide G: R^1 = MOEA, R^2 = MOAA; **8** doscadenamide H: R^1 = MOAA, R^2 = MOEA;
9 doscadenamide I: R^1 = MOYA, R^2 = oxo-MOAA; **10** doscadenamide J: R^1 = oxo-MOAA, R^2 = MOYA

Figure 2.8 - The doscadenamides: compound **1**, along with its analogs whose proposed structures were annotated via informative patterns in the MS^2 fragmentation data (see Figures 2.S9–2.S26). Each analog consists of a heterocyclic core with two fatty acid side chains with the following possibilities: MOYA = 2-methyl octynoic acid; MOEA = 2-methyl octenoic acid; MOAA = 2-methyl octanoic acid; oxo-MOAA = 2-methyl 7-oxo octanoic acid.

2.2.6: Structural Features and Biological Activity of **1**

Compound **1** contains unusual structural features that, while having precedent in other cyanobacterial natural products, have not previously been seen together. Terminal alkynes can be found in several other natural products from *Moorena* spp., including jamaicamide B (Edwards et al 2004), carmabin A (Hooper et all 1998), and vatiamides A, C, and E (Moss et al 2019), but having two is notable. While ribosomally synthesized and post-translationally modified peptides (RiPPs) and NRPS-derived natural products with amino acid subunits are common in cyanobacteria, lysine is not often seen, especially in the natural products of marine cyanobacteria (Gerwick, Tan, and Sitachitta 2001; Tidgewell, Clark, and Gerwick 2010). The heterocycle in **1**, composed of an acetate extended amino acid, has been observed in the malyngamides (Milligan et al 2000), jamaicamides (Edwards et al 2004), gallinamides (Linington et al 2009), and other cyanobacterial natural products, but again, never has it been reported involving a lysine residue. Two curiosities of the biosynthesis of compound **1**, namely the origin of the two 2-methyl octynoic acid residues and the formation of the heterocycle, can be explained by analogy to what is known about the biosynthesis of the jamaicamides (Edwards et al 2004). To generate 2-methyl octynoic acid, a fatty-acid desaturase analogous to JamB could act upon an octanoic acid precursor, or a smaller precursor that has been PKS-extended to the appropriate size. The placement of the methyl group in the 2 position suggests incorporation via S-adenosyl methionine (SAM). Formation of the heterocycle likely occurs as the result of an acetate extension of the carboxyl group of lysine, followed by a Claisen-like condensation and cyclization directed by a cyclase analogous to JamQ. As noted above in the discussion of structural possibilities **1**a and **1**b, what is less clear is how 2-methyl octynoic acid is appended to the terminus of the lysine side chain; the peptide bond formed is far from unusual, but its placement suggests enzymatic activity occurring beyond the otherwise linear PKS-NRPS assembly of the molecule.

To further evaluate the relationships of the structural features found together in compound **1** to the known natural product chemical space, we applied a Small Molecule Accurate Recognition Technology (SMART) (Zhang et al 2017) analysis to search for structurally similar molecules based on HSQC spectra. SMART did not yield any similar compounds with a cosine value higher than 0.84, further revealing the structural uniqueness of compound **1**. We also utilized the structure similarity search function in SciFinder (https://scifinder.cas.org/), which yielded only the sintokamides (Figure 2.S27). The sintokamides share a similar heterocycle and are halogenated natural products from sponges

(Sadar et al 2008). While not suggested by either structure similarity query, tetramic acids (Lowery et al 2009) and prostaglandins (PGE_2, for example) (Ricciotti and FitzGerald 2011) (Figure 2.S27) are two chemical classes that possess some distant level of structural similarity to compound **1**, and this inspired additional bioactivity testing efforts, as described below.

Structural similarity to tetramic acids inspired in silico antibiotic screening (http://chemprop.csail.mit.edu/) (Stokes et al 2020). The known antibacterials C_{12}-tetramic acid and C_{14}-tetramic acid scored over five times greater than the highest scoring doscadenamide (Table 2.S4), providing little incentive to further evaluate the doscadenamides for antibiotic activity.

Compound **1** was assayed for cytotoxicity against human NCI-H460 cells and yielded an $IC_{50} > 22$ μM, suggesting negligible cytotoxicity. This lack of cytotoxicity, plus some distant structural similarity to prostaglandins, inspired the screening of compound **1** in a Griess assay for anti-inflammation (as well as cytotoxicity) toward murine macrophages RAW264.7 cells at a range of 7–55 μM. Curiously, rather than producing inflammatory or anti-inflammatory effects, compound **1** yielded dose-dependent synergistic cytotoxicity with lipopolysaccharide (LPS). This anomalous result was confirmed through multiple replicates of the assay (Figures 2.S28–2.S30).

2.2.7: Conclusions

The doscadenamides were discovered based on global scale patterns in *M. bouillonii* chemical diversity. This illustrates that cyanobacteria harbor intraspecific chemogeographic patterns, and that these patterns can be utilized to direct discovery efforts towards new, regionally specific natural product families. There are many tools available for pursuing chemogeographic and other metabolite patterns in sample sets that can inform discovery efforts, each with their own strengths and limitations. While tools like the ORCA pipeline and GNPS classical molecular networking may be of limited utility in terms of quantitative analyses and effective separation of isomeric features, their flexibility in handling heterogeneous sample sets allows for comparative analyses between samples that could otherwise not be conducted. Furthermore, the intrinsic imperfection of real-world data and the deficiencies inherent to various tools and approaches encourages that an ensemble of tools and approaches be applied. By using ORCA in conjunction with GNPS, we were able to generate convergent results that increased confidence in our conclusions. Converging results from ORCA and GNPS were also helpful in giving confidence to the parameters selected for our analyses; parameter selection is often a challenge when applying computational techniques and requires deep knowledge of the dataset as well as manual validation. The chemogeographic patterns in *M. bouillonii* natural products that are qualitatively presented in this manuscript highlight the opportunity to further explore *M. bouillonii* natural products chemistry and how compounds and compound families are distributed ubiquitously vs. regionally, at different geographical scales. Studying *M. bouillonii* metabolomics in a more controlled, semi-quantitative fashion would allow these patterns to be evaluated more deeply and will be the focus of a future manuscript.

Doscadenamide A (**1**), when considered in isolation, is a structurally intriguing compound. Being composed of a heterocyclized, acetate-extended amino acid core appended with terminal alkyne containing side chains, it blends structural features common among cyanobacterial natural products with a flair of the unusual: the inclusion of lysine, the dual terminal alkynes, and the acylation of the lysine side chain with one of those terminal alkyne containing side chains. In considering the doscadenamides as a family of cyanobacterial natural products, it is likely they are produced via a seemingly combinatorial addition of different acyl groups to a consistent core structure. From a biosynthetic perspective, this suggests a low level of fidelity in the assembly process. Connecting this family of compounds to the biosynthetic gene cluster responsible for their production would elevate our understanding of how cyanobacteria diversify their natural product arsenals. Since the aforementioned procedure for the total synthesis of compound **1** (Liang et al 2019) is amenable to incorporating alternative sidechains, this could be used for generating the nine proposed natural structural analogs reported here (**2**–**10**), as well as for evaluating their activities as quorum sensing modulators (Liang et al 2019) and their cytotoxic synergism with LPS.

2.3: Methods

2.3.1: General Experimental Procedures

Optical rotation was measured using a JASCO P-2000 polarimeter (Easton, MD, USA), UV/Vis data were obtained using a Beckman DU800 spectrophotometer (Brea, CA, USA), and IR spectra were recorded on a ThermoScientific Nicolet 6700 FT-IR spectrometer (Waltham, MA, USA). NMR experiments were conducted using a JEOL ECZ 500 NMR spectrometer (Akishima, Tokyo, Japan) equipped with a 3 mm inverse probe (H3X), a Bruker AVANCE III 600 MHz NMR with a 1.7 mm dual tune TCI cryoprobe (Billerica, MA, USA), and a Varian VX500 (Palo Alto, CA, USA). NMR data were processed using Mestrenova (Mestrelab, Santiago de Compostela, Spain) and TopSpin (Bruker, Billerica, MA, USA). NMR data were recorded in $CDCl_3$ and referenced to the solvent peak (7.260, 77.160). For the low-resolution LC-MS/MS analysis, a ThermoFinnigan Surveyor HPLC System (San Jose, CA, USA) with a Phenomenex Kinetex 5 μm C18 100 × 4.6 mm column (Torrance, CA, USA) coupled to a

ThermoFinnigan LCQ Advantage Max Mass Spectrometer (San Jose, CA, USA) in positive ion mode was used. Samples were analyzed using one of two linear gradients from 30% CH_3CN + 0.1% formic acid to 99% CH_3CN + 0.1% formic acid in H_2O + 0.1% formic acid at a flow rate of either 0.6 mL/min or 0.7 mL/min over 32 min or 30 min, respectively. Samples were run at a concentration of 1 mg/mL, with concentrations increased up to 4 mg/mL in situations where the peak intensities were insufficient. For the HiRes-ESI-MS analysis, an Agilent 6230 time-of-flight mass spectrometer (TOFMS) (Santa Clara, CA, USA) with Jet Stream ESI source was used. For HiResMS^2 fragmentation data, a ThermoScientific Orbitrap XL mass spectrometer (Waltham, MA, USA) with direct infusion of the sample into the Thermo IonMax electrospray interface was used.

Compound isolation was performed using two semi-preparative HPLCs: a Thermo Scientific Dionex UltiMate 3000 HPLC (Waltham, MA, USA) system with automated fraction collector, a Waters HPLC system with 1500 series pumps (Milford, MA, USA), and a 996 photodiode array detector with manual fraction collection. HPLC separation was performed using a Phenomenex Kinetex 5 µm C18 10 × 150 mm column (Torrance, CA, USA) and reverse phase gradients of acetonitrile in H_2O, with both solvents containing 0.1% (*v/v*) formic acid. HPLC grade organic solvents and Millipore Milli-Q system (Burlington, MA, USA) purified water were used.

All reagents, catalysts, and solvents used for the synthetic experiments were purchased in their purest and driest form. All experiments were carried out under an inert atmosphere (Ar) unless otherwise specified.

National Cancer Institute (NCI) H460 hypotriploid human cells [American Type Culture Collection (ATCC) HTB-177] and RAW 264.7 murine macrophages (ATCC TIB-71) were purchased from the ATCC (Manassas, VA, USA).

2.3.2: Sample Collection

Fifteen benthic filamentous tropical marine cyanobacterial samples were hand-collected via self-contained underwater breathing apparatus (SCUBA) or snorkeling in American Samoa, Guam, Kavaratti (Lakshadweep Islands, India), Papua New Guinea, Saipan, and the Paracel Islands (Xisha) in the South China Sea between the years 2005 and 2018. Samples from all locations besides Papua New Guinea were preserved in 1:1 seawater and either ethyl or isopropyl alcohol, transported back to laboratories, and stored frozen until extraction. The sample from Papua New Guinea was transported back to the laboratory in a culture flask and propagated in seawater (SW) BG-11 media (Moss et al 2018). For additional metadata about these samples, see Table 2.S5.

2.3.3. Sample Preparation

Cyanobacterial biomass was exhaustively extracted with 2:1 dichloromethane and methanol, concentrated under vacuum, and resuspended in methanol or acetonitrile at a concentration of 1 mg/mL. Samples were prepared for LC-MS/MS analysis via elution through C18 solid phase extraction (SPE) cartridges.

2.3.4. ORCA Pipeline

Code, data files, and supporting documentation on use and workings of the ORCA pipeline are available at https://github.com/c-leber/ORCA, while the parameter sets used for the various analyses reported in this study are available in Tables 2.S6–2.S8. ORCA was written in Python (Van Rossum and Drake 1995) and is built off the following Python packages: pandas (0.25.2) (McKinney 2010; Pandas-Dev/Pandas 2018), numpy (1.16.5) (Oliphant 2006; van der Walt, Colbert, and Varoquaux 2011), pyteomics (4.1.2) (Goloborodko et al 2013; Levitsky et al 2019), scipy (1.3.1) (Virtanen et al 2020), networkx (2.4) (Hagberg, Schult, and Swart 2008), matplotlib (3.0.3) (Hunter 2007), sklearn (0.21.3) (Pedregosa et al 2011), and seaborn (0.9.0) (Mwaskom/Seaborn 2018). ORCA is available in the form of a Jupyter Notebook (Pérez and Granger 2007; Kluyver et al 2016), to facilitate customization and interactive experimentation. Prior to analyses in ORCA, proprietary LC-MS datafiles were converted to mzXML using MSCONVERT (Holman, Tabb, and Mallick 2014), which is a part of the ProteoWizard Library (Chambers et al 2012). MSCONVERT was also used to convert proprietary LC-MS/MS datafiles to mzML for the ORCA MS^2 Auxiliary pipeline, and to mzXML or mzML for GNPS.

2.3.5: GNPS Classical Molecular Networking

Molecular networks were created using the online workflow (https://ccms-ucsd.github.io/GNPSDocumentation/) on the GNPS website (http://gnps.ucsd.edu) and were visualized using Cytoscape (3.7.2) (https://cytoscape.org/) (Shannon et al 2003) and the GNPS in-browser network visualizer. For full accounting of the networking parameter sets, see Tables 2.S9–2.S11.

2.3.6: Compound Isolation

M. bouillonii biomass from Laulau Bay, Saipan, was thoroughly extracted with 2:1 dichloromethane and methanol, yielding 10 g crude extract from 132 g (1 L) biomass. A portion of the crude extract was fractionated over silica with vacuum liquid chromatography and a standardized solvent system protocol (Table 2.S12). Two relatively polar fractions (fractions F and G) were found to contain the bulk of compound **1**. Reverse phase HPLC was used to isolate 2.6 mg of this compound from these two fractions. A gradient method from 37% to 50% CH_3CN + 0.1% formic acid in H_2O + 0.1% formic acid over 60 min at a flow rate of 4 mL/min resulted in the elution of compound **1** starting at a retention time of approximately 38 min.

2.3.7: Planar Structure Characterization

Compound **1**: white solid, $[\alpha]^{26}_{D}$ +17.7 (*c* 0.1, MeOH); UV/Vis (Figure 2.S31); IR (Figure 2.S32);. NMR data (Table 2.S13); ^{1}H, ^{13}C, COSY, HSQC, HMBC, HSQC-TOCSY, and long-range HSQMBC spectra (Figures 2.S33–2.S39); HR ESIMS (obs. *m/z* $[M + Na]^+$ at 479.2877, $C_{27}H_{40}N_2O_4$, calc. 479.2880).

2.3.8. Structure Elucidation—Standard Preparation and Derivatization for Configurational Characterization

Methods for ZACA methylalumination-oxidation (Negishi et al 2004), catalytic hydrogenation (Iwasaki et al 2015), ozonolysis (Pereira et al 2011; Linington et al 2007), acid hydrolysis (Linington et al 2007; Pereira et al 2011; Iwasaki et al 2015), peptide coupling (Yoshimura et al 2014; Iwasaki et al 2015), and derivatization with Marfey's reagent (Linington et al 2007; Iwasaki et al 2015) were adapted from the literature.

2.3.8.1: Synthesis of (S)-2-methyloctanoic Acid

To a solution of trimethylaluminum (891 μL, 1.782 mmol) and (+)-$(NMI)_2ZrCl_2$ (23.84 mg, 0.036 mmol) in 1.5 mL of CH_2Cl_2 was added a solution of oct-1-ene (100 mg, 0.891 mmol) in 1.5 mL of CH_2Cl_2. After stirring overnight at 23 °C, the mixture was treated with a vigorous stream of O_2 for 1 h at 0 °C and then stirred for 5 h under an atmosphere of O_2 at room temperature. The reaction mixture was quenched with 1 M HCl, extracted with CH_2Cl_2, washed with brine, dried over $MgSO_4$ and concentrated. The residue was purified via silica flash column chromatography (20% ethyl acetate/hexanes) to yield (*S*)-2-methyloctan-1-ol (50 mg, 0.347 mmol, 39% yield) as a clear oil. The crude product was used in the next step without further purification.

To a solution of (*S*)-2-methyloctan-1-ol (50 mg, 0.347 mmol) in acetonitrile (1.4 mL) was added *N*-methyl morpholine *N*-oxide (NMO) solution in H_2O (468 mg, 3.47 mmol) and tetrapropylammonium perruthenate (TPAP) (12.18 mg, 0.035 mmol) sequentially at room temperature and the mixture was stirred for 2 h. The mixture was then concentrated, and the residue passed through a pad of silica gel using hexanes:diethyl ether (3:1) containing 0.1% acetic acid. The eluted solvent was concentrated to yield (*S*)-2-methyloctanoic acid (48 mg, 0.303 mmol, 88% yield). $[\alpha]_{D}^{26}$ +10.0 (c 1.05, MeOH); ^{1}H NMR (500 MHz, $CDCl_3$) δ 2.46 (ddq, J = 9.7, 6.8, 3.1 Hz, 1H), 1.69 (m, 2H), 1.44 (m, 2H), 1.37–1.24 (m, 6H), 1.19 (m, 3H), 0.89 (m, 3H); ^{13}C NMR (126 MHz, $CDCl_3$) δ 183.2, 39.6, 33.7, 31.9, 29.4, 27.3, 22.8, 17.1, 14.3. HRESIMS *m/z* $[M + H]^+$ 159.1394 (calc. for $C_9H_{19}O_2$, 159.1385).

2.3.8.2: Derivatization of 2-methyloctanoic Acid with 2-phenylglycine Methyl Ester

To generate a 1:1 standard mixture of both possible diastereomers of 2-methyloctanoic acid, 6.0 mg (37.9 µmol) of racemic 2-methyloctanoic acid was combined with 1-[bis(dimethylamino)methylene]-1H-1,2,3-triazolo[4,5-b]pyridinium 3-oxid hexafluorophosphate (HATU) (14.4 mg, 37.9 µmol), (*S*)-(+)-2-phenylglycine methyl ester hydrochloride (7.6 mg, 37.9 µmol) and *N*,*N*-diisopropylethylamine (DIPEA) (30 µL) in dimethylformamide (DMF) (300 µL). This was stirred overnight at room temperature and ambient atmosphere. The reaction mixture was then diluted with 1.0 mL of EtOAc, washed with saturated aqueous NH_4Cl (3 × 1.0 mL), concentrated under vacuum, and prepared for LC-MS analysis. To generate a chiral standard, 1.8 mg (11.4 µmol) of (*S*)-2-methyloctanoic acid was combined with HATU (4.3 mg, 11.4 µmol), (*S*)-(+)-2-phenylglycine methyl ester hydrochloride (2.3 mg, 11.4 µmol) and DIPEA (30 µL) in DMF (300 µL), and stirred overnight at room temperature and ambient atmosphere. The reaction mixture was then diluted with 1.0 mL of EtOAc, washed with saturated aqueous NH_4Cl (3 × 1.0 mL), concentrated under vacuum, and prepared for LC-MS analysis. The diastereomeric ratio of the chiral standard was 3:1 by area-under-curve analysis.

2.3.8.3: Derivatization of Lysine with Marfey's Reagent (FDAA)

To generate a racemic standard, 0.8 mg (4 µmol) of racemic lysine hydrochloride and 0.1 M $NaHCO_3$ (200 µL) were added to a solution of L-FDAA (4.4 mg, 16 µmol) in acetone (600 µL). The reaction mixture was sealed in a vial with ambient atmosphere, stirred at 90 °C for 5 min, neutralized with 6 M HCl, concentrated under vacuum, and prepared for LC-MS analysis. To generate a chiral standard, 1.0 mg (6 µmol) of L-lysine monohydrate and 0.1 M $NaHCO_3$ (200 µL) were added to a solution of L-FDAA (6.5 mg, 24 µmol) in acetone (600 µL). The reaction mixture was sealed in a vial with ambient atmosphere, stirred at 90 °C for 5 min, neutralized with 6 M HCl, concentrated under vacuum, and prepared for LC-MS analysis.

2.3.8.4: Derivatization of Compound **1**

Compound **1** (0.5 mg) was combined with 1.0 mg of Pd/C in 1 mL EtOH and stirred under an atmosphere of H_2 for 8 h. The mixture was filtered through glass wool, rinsed with EtOH (3 × 1.0mL), and concentrated in vacuo.

Hydrogenated compound **1** (0.25 mg) was dissolved in 1 mL CH_2Cl_2, into which a stream of ozone gas was bubbled at −78 °C for 25 min. The reaction was concentrated under vacuum, and the residue was treated with 1 mL of 1:2 35% H_2O_2:HCOOH at 70 °C and ambient atmosphere for 20 min. The reaction was again concentrated in vacuo, followed by the addition of 1 mL 6 M HCl. The reaction mixture was stirred in a sealed vial with ambient atmosphere overnight at 110 °C, and then concentrated under vacuum. To the residue, a solution of L-FDAA (0.6 mg, 2 µmol) in acetone (200 µL) and 0.1 M $NaHCO_3$ (200 µL) were added. The reaction mixture was sealed in a vial with ambient atmosphere, stirred at 90 °C for 5 min, neutralized with 6 M HCl, concentrated in vacuo, and prepared for LC-MS analysis.

Hydrogenated compound **1** (0.25 mg) was dissolved in 1 mL 6 M HCl. This reaction mixture was stirred in a sealed vial with ambient atmosphere overnight at 110 °C, and then concentrated under vacuum. The residue was combined with HATU (0.4 mg, 1 µmol), (*S*)-(+)-2-phenylglycine methyl ester hydrochloride (0.2 mg, 1 µmol), and DIPEA (20 µL) in DMF (200 µL) and stirred for 6 h at room temperature and at ambient atmosphere. The reaction mixture was then diluted with 1.0 mL of EtOAc, washed with saturated aqueous NH_4Cl (3 × 1.0 mL), concentrated under vacuum, and prepared for LC-MS analysis.

2.3.9: ORCA MS^2 Auxiliary Pipeline

Code, data files, and supporting documentation on the use and workings of the ORCA MS^2 Auxiliary pipeline are available at https://github.com/c-leber/ORCA. MS^2 spectra were binned with the bin_OOM parameter set to 0, and the cutoff parameter for the hierarchical clustering of the MS^2 scans for a particular precursor mass was set to 0.15. The MS^2 scans were filtered to only include scans for precursor masses, which contained fragment peaks with *m/z* 168, resulting in 78 precursor masses. These precursor masses were individually analyzed via the hierarchical clustering of scan fragmentation patterns followed by the generation of consensus spectra for each cluster. Consensus spectra were manually inspected to detect interpretable fragmentation patterns similar to those of compound **1**.

2.3.10: Bioassays

Methods for the NCI-H460 cytotoxicity assay (Tao et al 2108) and the Griess assay (Green et al 1982; Choi et al 2012) were adapted from the literature.

2.3.10.1: Cytotoxicity Assay of Compound **1** with NCI-H460 Cell Line

NCI-H460 hypotriploid human cells (ATCC HTB-177) were grown in monolayers to near confluence in flasks and then seeded into wells at 3.33 × 10^4 cells/mL of Roswell Park Memorial Institute (RPMI) medium with standard fetal bovine serum (FBS), 180 μL/well, and incubated for 24 h at 37 °C in 96-well plates. Cells were exposed to compound **1** at ten half log concentrations, the highest being 21.9 μM with 1% dimethyl sulfoxide (DMSO) present, while the lowest was 0.7 nM. Plates were incubated for an additional 48 h and then stained with 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT), for 25 min, after which the optical densities were recorded at 630 and 570 nm for each well on a SpectraMax M2 microplate reader with SoftMax® Pro Microplate Data Acquisition and Analysis Software (Molecular Devices, LLC, Version No. M2, Sunnyvale, CA, USA). The test samples were compared with a negative control of 1% DMSO and a positive control of doxorubicin (0.1 μg/mL and 1 μg/mL), both in RPMI medium. Due to the limited availability of compound **1**, fully toxic concentrations were not reached; hence, the resultant dose–response curve was incomplete. Nevertheless, the IC_{50} value for compound **1** is greater than 21.9 μM.

2.3.10.2: Griess Assay and Cytotoxicity of Compound **1** in RAW 264.7 Cells

RAW 264.7 murine macrophages (ATCC TIB-71) were seeded at 5×10^4 cells in 96-well plates in Dulbecco's Modified Eagle Medium (DMEM; Gibco, Carlsbad, CA, USA) supplemented with 10% endotoxin-low FBS (HyClone, characterized, Endotoxin: ≤ 25 EU/mL), 190 μL/well, and incubated for 24 h at 37 °C. Compound **1** at concentrations of 55, 28, 14, or 7 μM was applied in triplicate, and after 1 h lipopolysaccharide (LPS from Escherichia coli 026:B6, =10,000 EU/mg, Sigma-Aldrich, Oakville, ON, Canada) was added (0.5 or 1.5 μg/mL) to all wells except those for the LPS-free controls and those for evaluating the pro-inflammatory effects of compound **1**. LPS alone was used as a negative control, whereas the same LPS concentration with 1% DMSO served as the positive control in the Griess assay. After 24 h, Griess reactions (section 2.3.10.3) were used to assess NO generation as a proxy for inflammation, and MTT staining (section 2.3.10.1) was used to assess cell viability. Doxorubicin at 3.3 μg/mL was used as a positive control for assessing cell viability. Cell survival was calculated as a percentage compared to wells with 1% or 1.5% EtOH and no LPS. A NO concentration standard curve was prepared in Microsoft Excel based on eight serial dilutions of a nitrite standard (0–100 μM) with DMEM. One-way ANOVA and Tukey's method were used to test for significance in the cell survival results from the assay; high mortality in certain conditions made statistical analyses of the inflammation data inappropriate. Statistical analyses were applied using GraphPad Prism version 8.0.0 for Windows. Batch variability in LPS potency and RAW 264.7 murine macrophage sensitivity, as well as limited availability of compound **1** necessitated using differing reagent concentrations across the biological replicates.

2.3.10.3: Griess Reaction

Supernatant from each sample well (50 μL) was added to the experimental wells in triplicate. A 1:1 mixture of 1% sulfanilamide solution in 5% phosphoric acid and 0.1% *N*-1-napthylethylenediamine dihydrochloride (100 μL) was added to each well and the plate was incubated in the dark for 20 min. Optical density was measured at 570 nm on a SpectraMax M2 microplate reader. The raw data were exported to a Microsoft Excel work sheet and the concentration of nitrite in the samples was determined by comparison to the standard curve using regression analysis.

2.3.10.4: In silico Antibiotic Screening

The simplified molecular-input line-entry system (SMILES) structures of compound **1**, proposed analogs **2**–**10**, C_{12}-tertramic acid, and C_{14}-tetramic acid were submitted to Chemprop Predict (http://chemprop.csail.mit.edu/predict) (Stokes et al 2020), using the Antibiotics model checkpoint.

Chapter 2, in full, is a reprint of the material as it appears in Leber, Christopher A.; Naman, C. Benjamin; Keller, Lena; Almaliti, Jehad ; Caro-Diaz, Eduardo J. E.; Glukhov, Evgenia; Joseph, Valsamma; Sajeevan, T. P.; Reyes, Andres Joshua; Biggs, Jason S.; Li, Te; Yuan, Ye; He, Shan; Yan, Xiaojun; Gerwick, William H. “Applying a Chemogeographic Strategy for Natural Product Discovery from the Marine Cyanobacterium *Moorena bouillonii*”, Marine Drugs, vol. 18, 2020. The dissertation author was the primary investigator and author of this paper. The dissertation author co-conceived of the work, acquired and prepared necessary samples, collected mass spectral and NMR data, wrote code and performed all analyses, led structure elucidation efforts, conducted or otherwise participated in synthetic reactions for determination of configuration, implemented *in silico* antibacterial screening, and was the primary author of this work.

2.4: Appendix: Supplemental Information

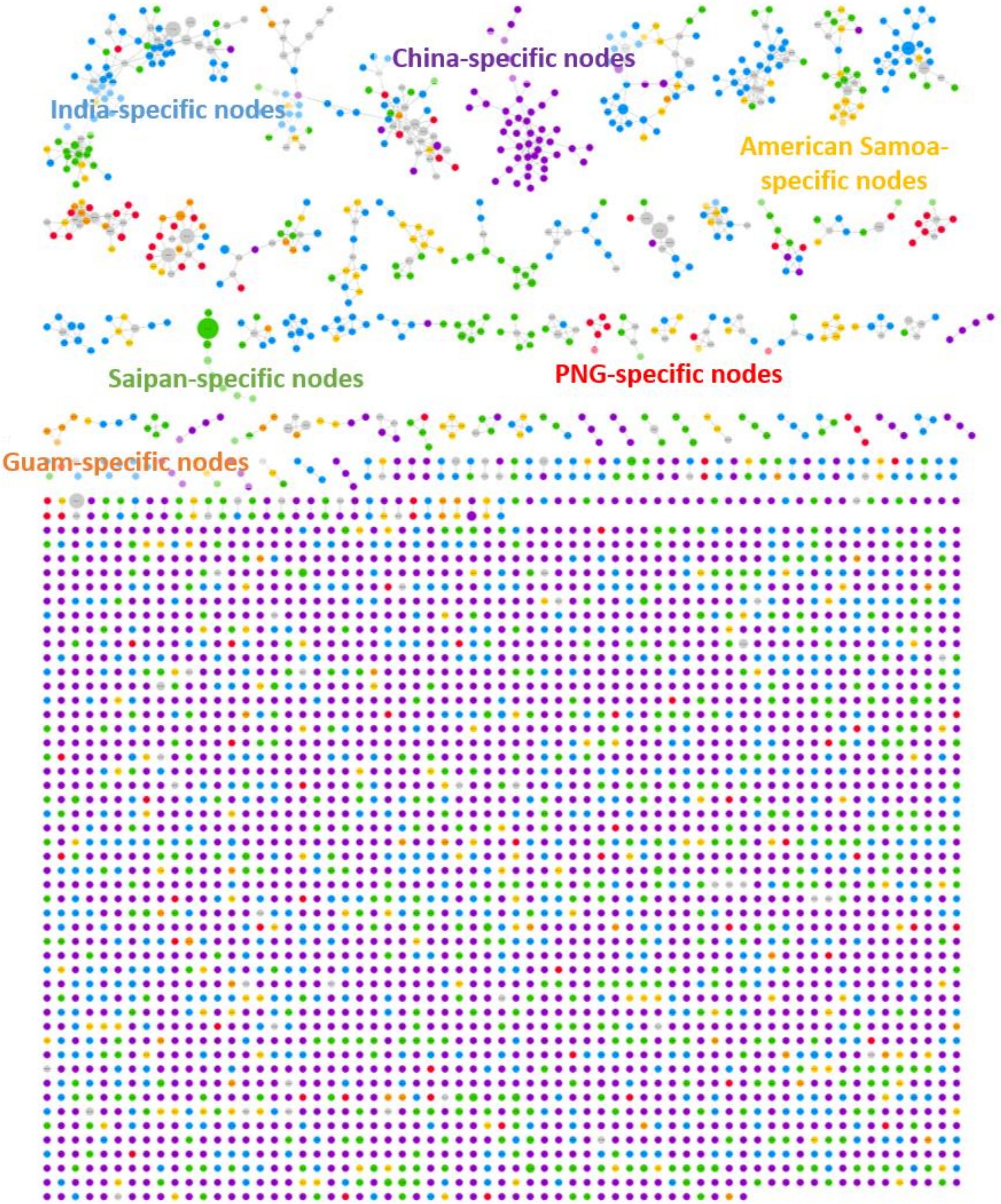


Figure 2.S1 - Molecular network of *M. bouillonii* crude extracts. GNPS classical molecular network of *M. bouillonii* crude extracts showing clusters of regionally specific nodes. Grey nodes represent MS^2 features that are present in samples from more than one geographically region. Nodes are scaled to summed precursor intensity. Red: Papua New Guinea, Orange: Guam, Gold: American Samoa, Green: Saipan, Blue: Kavaratti (Lakshadweep Islands, India), Purple: Xisha (Paracel) Islands in the South China Sea. See Table 2.S9 for network parameters.

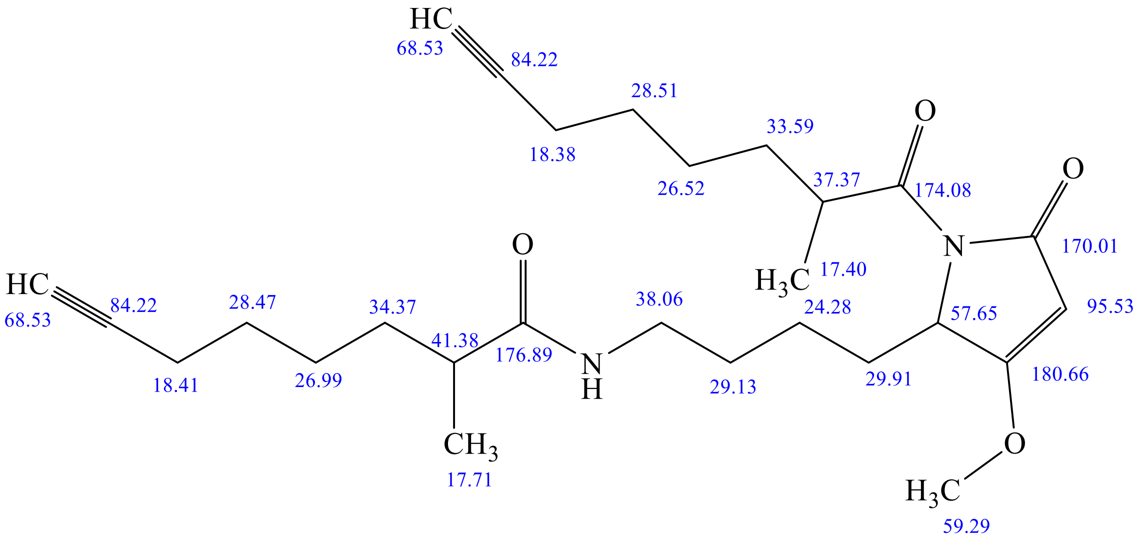


Figure 2.S2 - Predicted ^{13}C shifts for candidate structure **1a**. ^{13}C NMR shifts were calculated using ACD/Labs 2019.2.1 (ACD/C+H Predictors and DB 2019.2.1) (https://www.acdlabs.com/index.php).

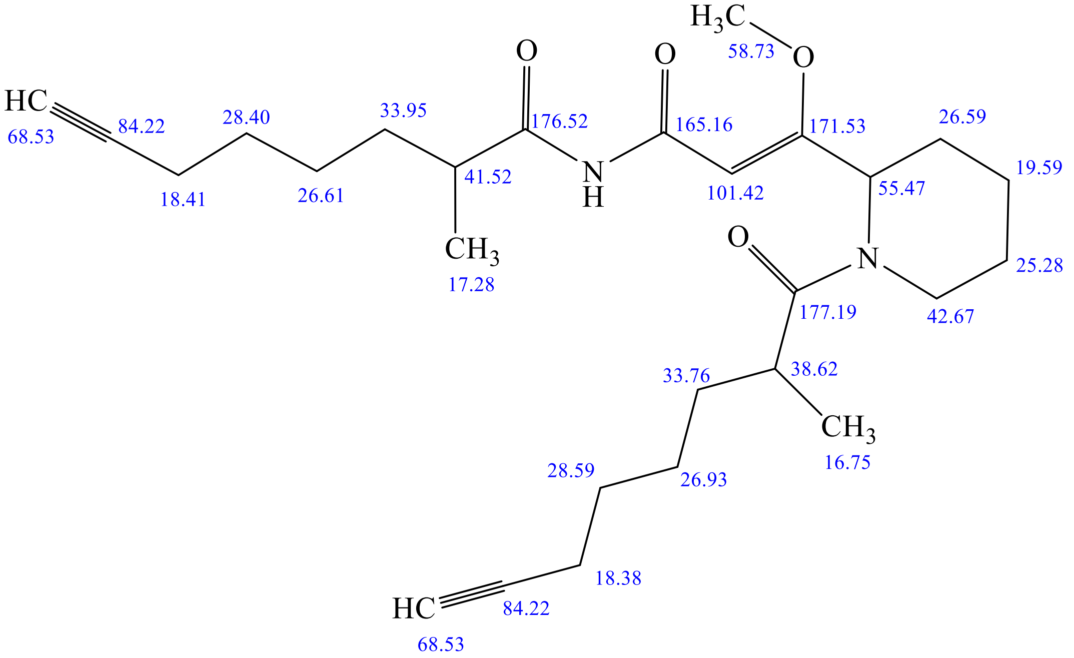


Figure 2.S3 - Predicted ^{13}C shifts for candidate structure **1b**. ^{13}C NMR shifts were calculated using ACD/Labs 2019.2.1 (ACD/C+H Predictors and DB 2019.2.1) (https://www.acdlabs.com/index.php).

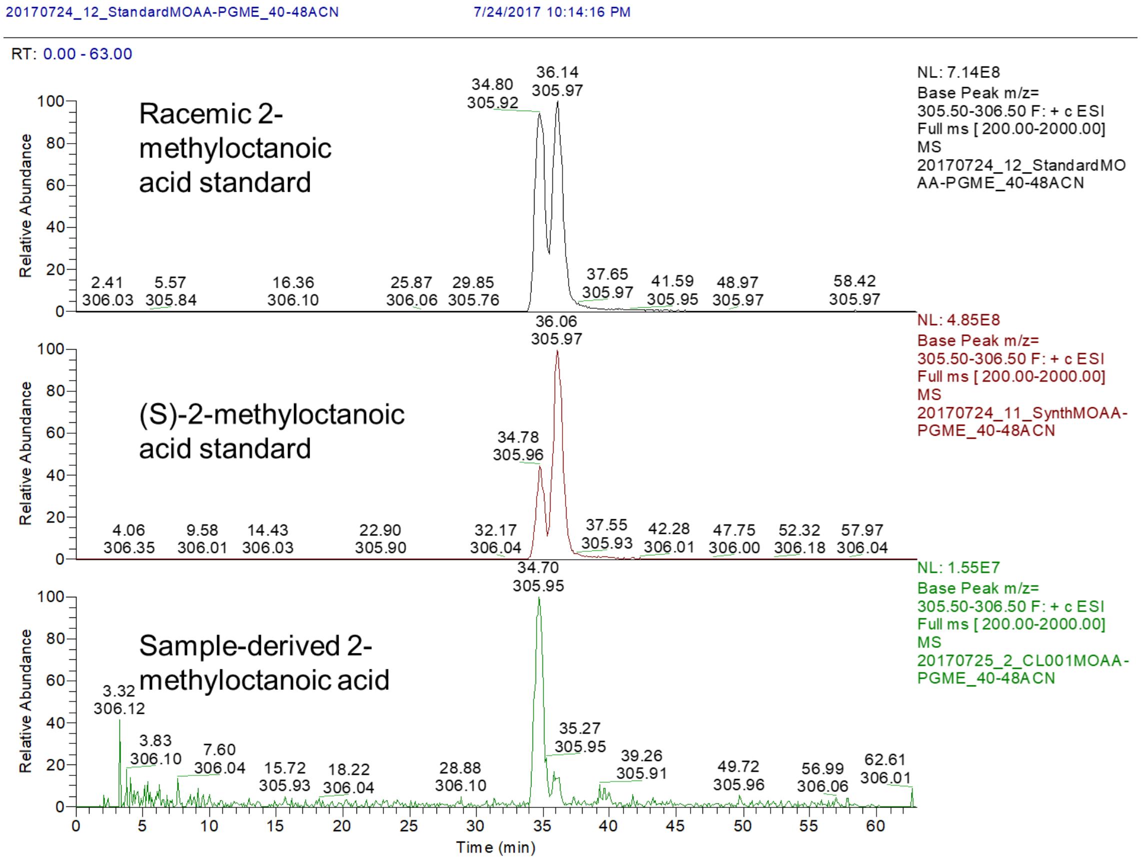


Figure 2.S4 - Compound **1** derived 2-methyoctanoic acid compared to standards. LC-MS TIC traces comparing (*S*)-(+)-2-phenylglycine methyl ester derivatized racemic 2-methyloctanoic acid and (*S*)-2-methyloctanoic acid standards to sample-derived 2-methyl octanoic acid, indicating the sample-derived 2-methyl octanoic acid to be of the *R* configuration.

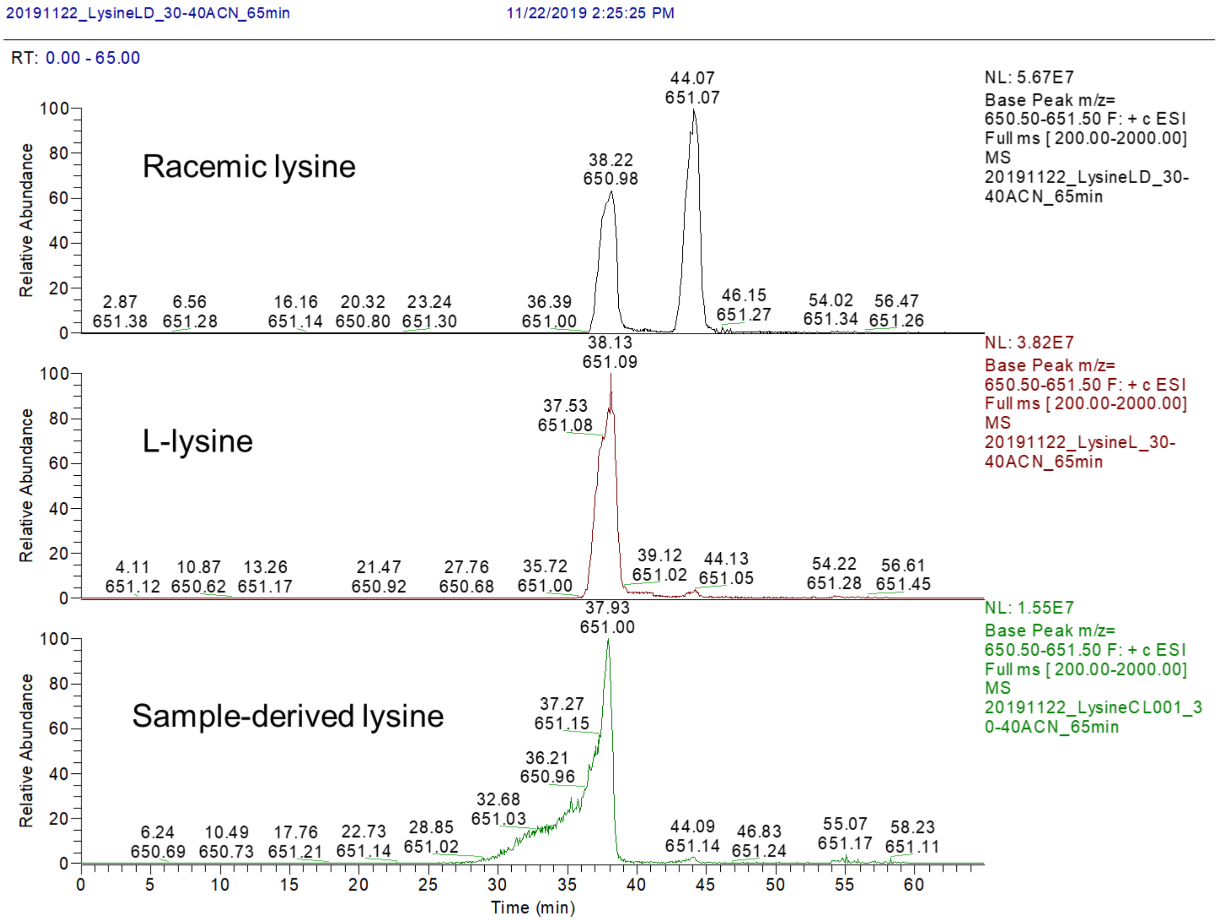


Figure 2.S5 - Compound **1** derived lysine compared to standards. LC-MS TIC traces comparing L-FDAA derivatized racemic lysine and L-lysine standards to sample-derived lysine, indicating the sample-derived lysine to be of the *S* configuration (L-lysine).

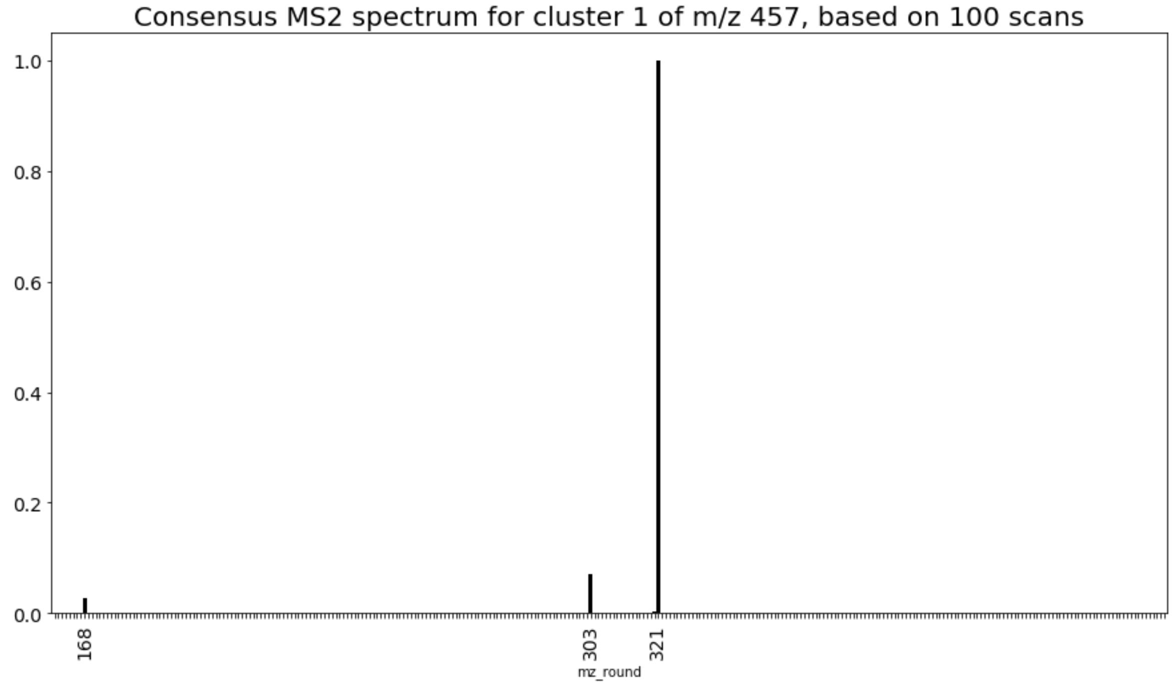
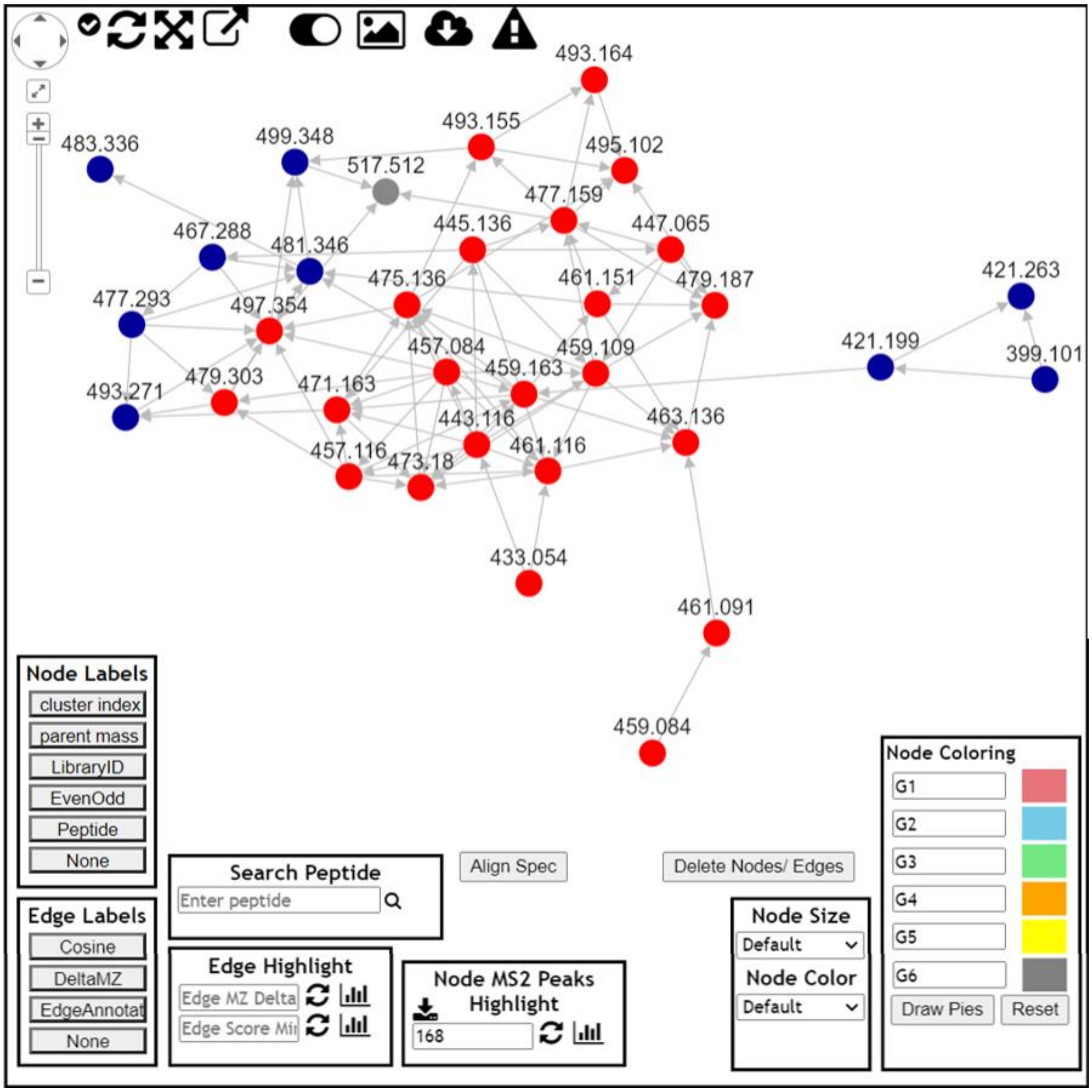


Figure 2.S6 - Doscadenamide A (**1**) consensus MS^2 spectrum. Consensus MS^2 spectrum representing a cluster of 100 scans for precursor mass *m/z* 457 and displaying the fragmentation spectrum for doscadenamide A (**1**).

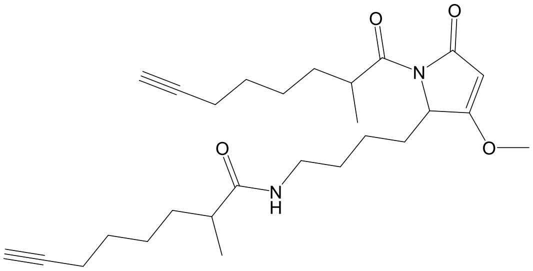


Node MS2 Peaks Highlight: 168

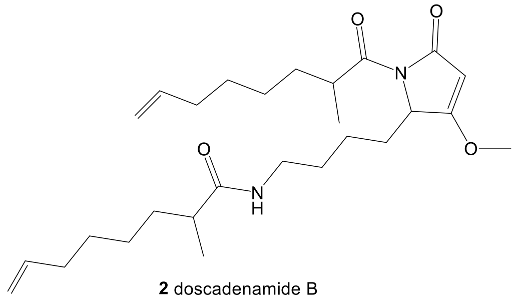
Node Size: Default

Node Color: Default

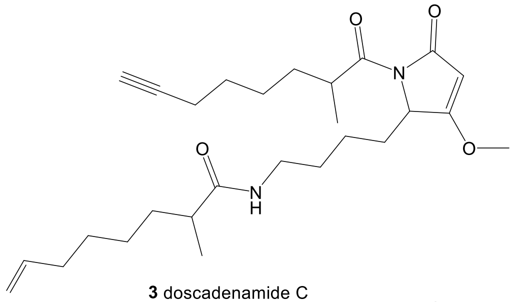
Figure 2.S7 - Molecular network cluster of compound **1** and analogs, highlighting *m/z* 168 frag. peak. A cluster of 33 MS^2 spectral nodes, including compound **1** (*m/z* 457.084), as visualized in the GNPS in browser network visualizer. This cluster is a part of a GNPS classical molecular network generated with crude extracts and fractions from two *M. bouillonii* samples: one from Saipan and one from Guam. All nodes colored red (23 out of 33) possess a fragment peak at *m/z* 168, suggesting that the structures they represent include a heterocyclic core identical to compound **1**.



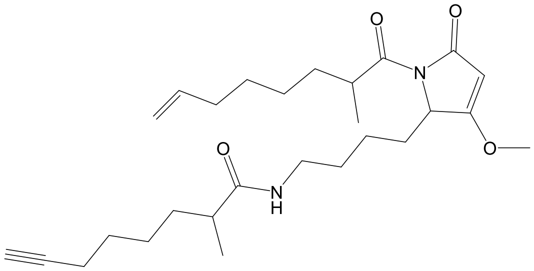
1 doscadenamide A



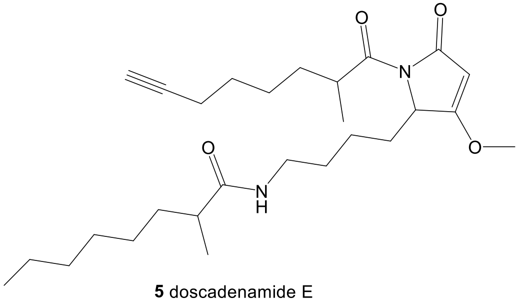
2 doscadenamide B



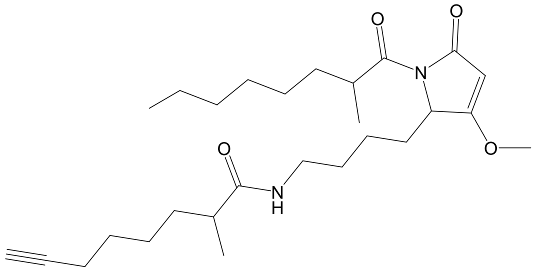
3 doscadenamide C



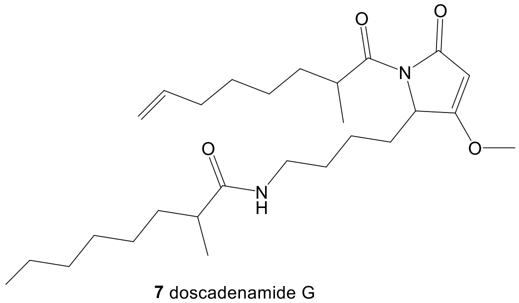
4 doscadenamide D



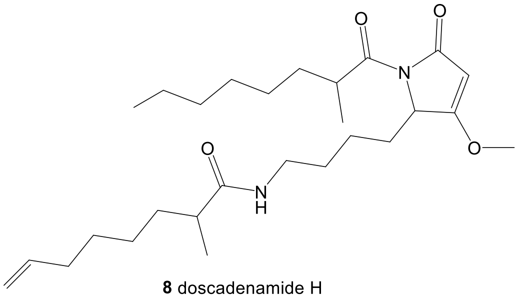
5 doscadenamide E



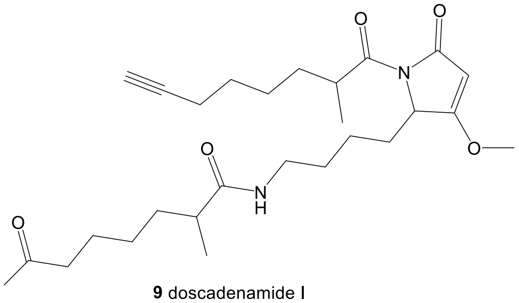
6 doscadenamide F



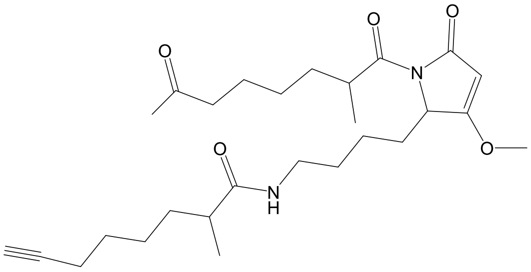
7 doscadenamide G



8 doscadenamide H



9 doscadenamide I



10 doscadenamide J

Figure 2.S8 - Structure of compound **1** with structure proposals for analogs (**2-10**). The doscadenamides: compound **1**, along with analogs whose proposed structures were annotated via informative patterns in the MS^2 fragmentation data (See Figures 2.S9-2.S26).

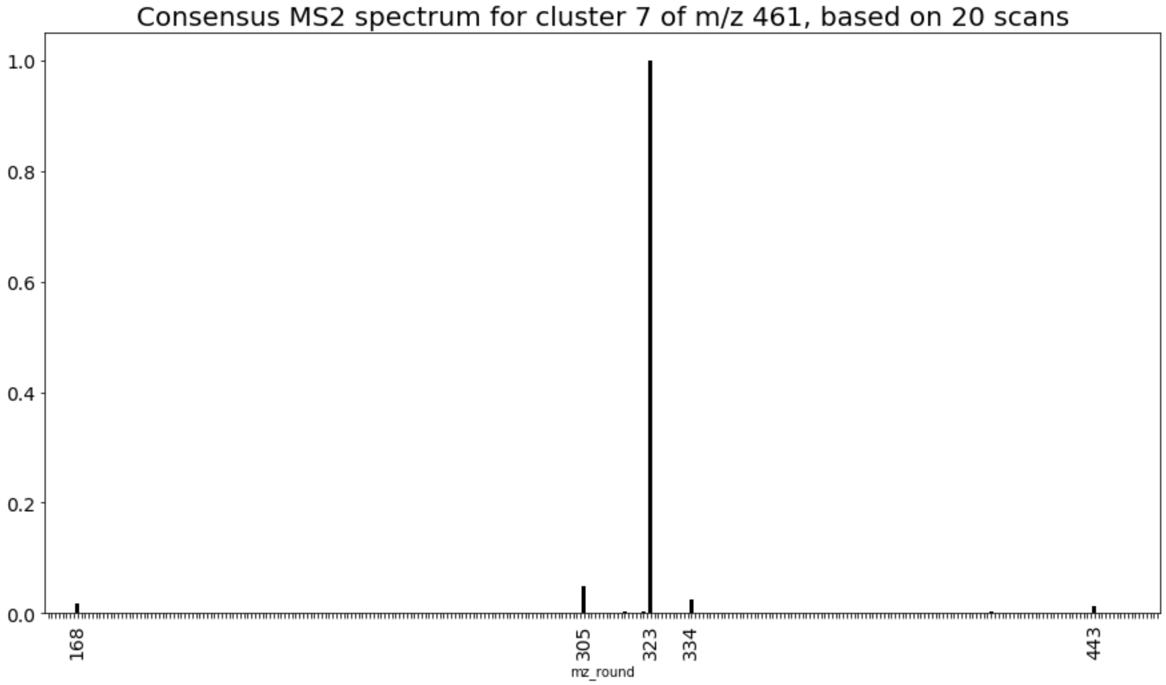


Figure 2.S9 - Doscadenamide B (**2**) consensus MS^2 spectrum. Consensus MS^2 spectrum representing a cluster of 20 scans for precursor mass *m/z* 461, representing the fragmentation spectrum of the proposed analog doscadenamide B (**2**).

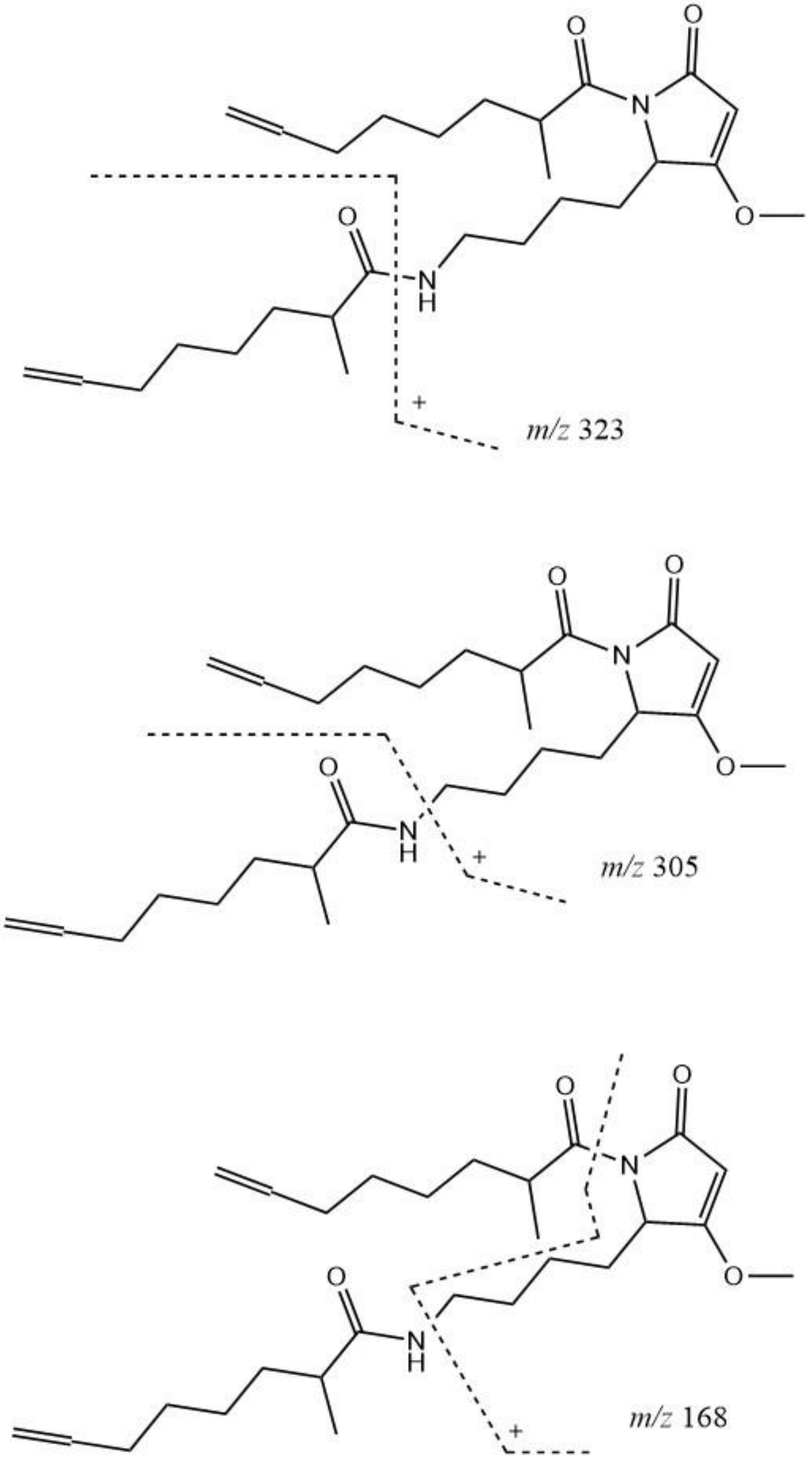


Figure 2.S10 - Doscadenamide B (**2**) proposed fragmentation

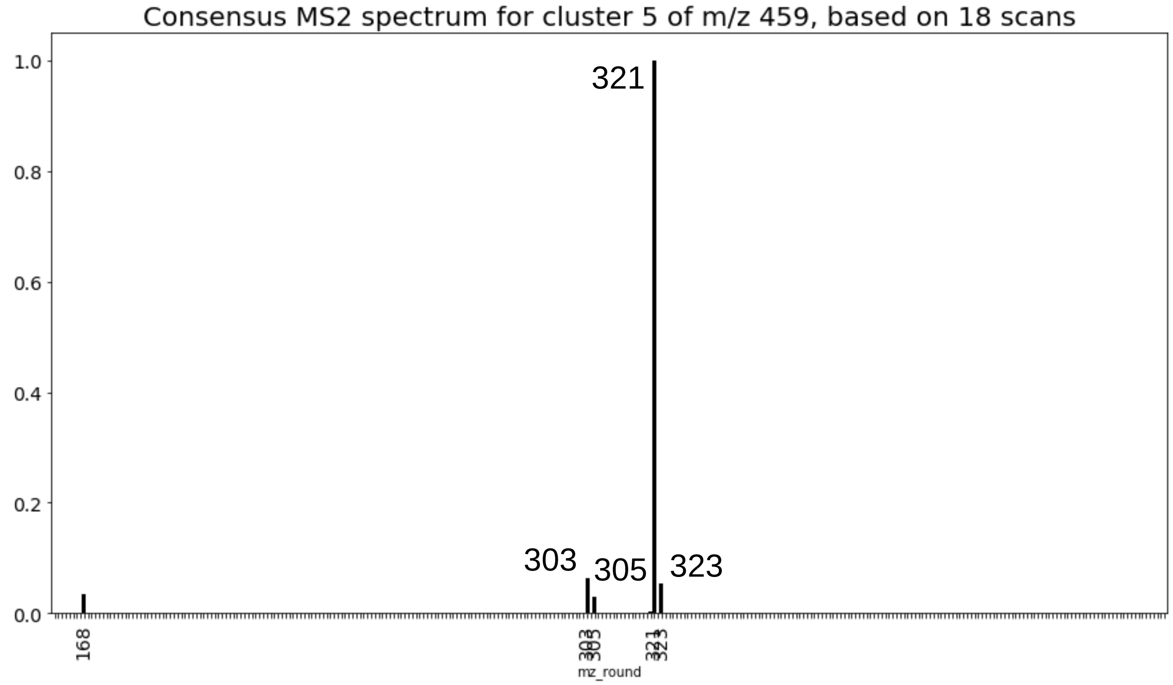


Figure 2.S11 - Doscadenamide C (**3**) consensus MS^2 spectrum. Consensus MS^2 spectrum representing a cluster of 18 scans for precursor mass *m/z* 459, representing the fragmentation spectrum of the proposed analog doscadenamide C (**3**).

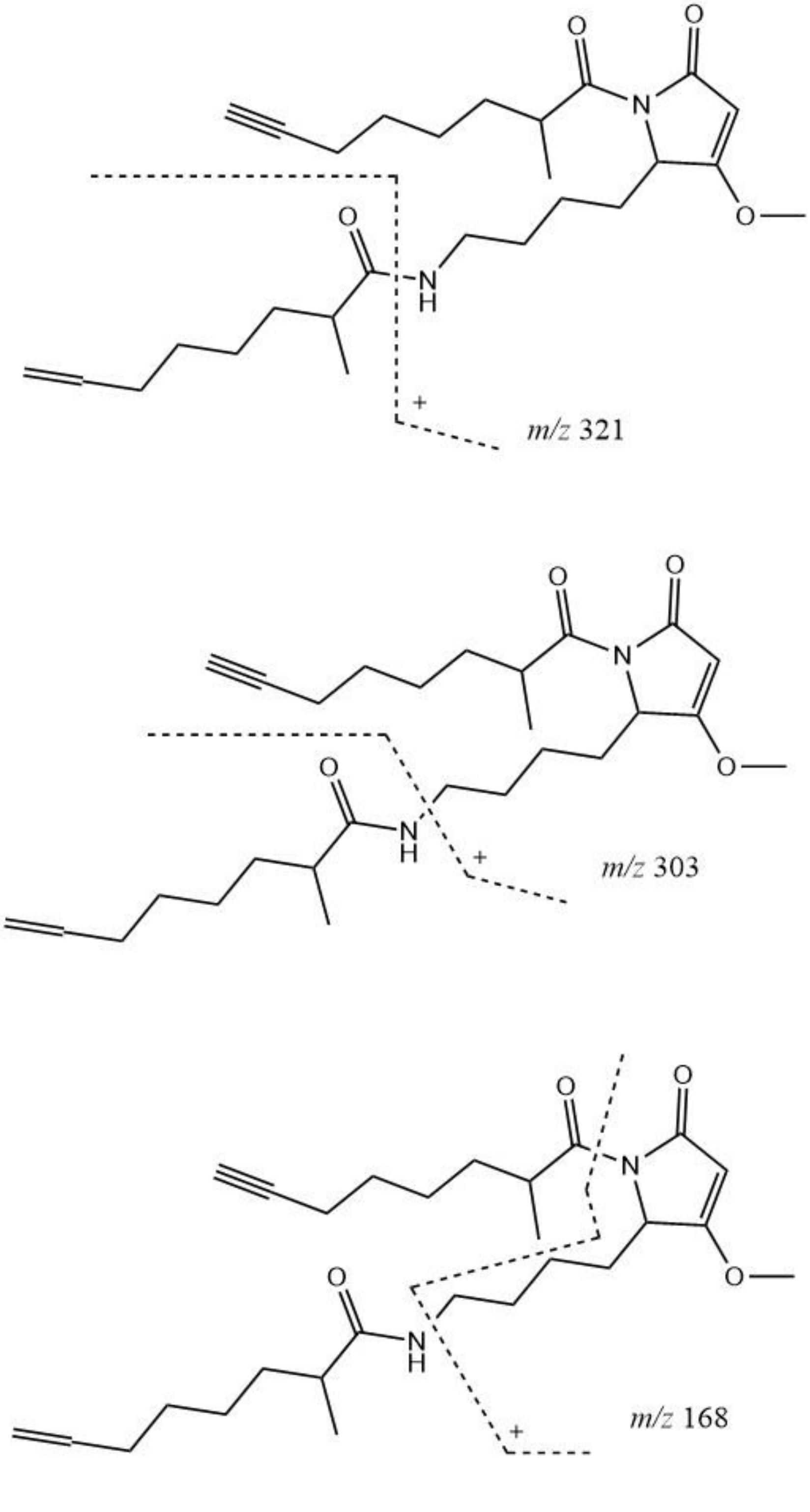


Figure 2.S12 - Doscadenamide C (**3**) proposed fragmentation

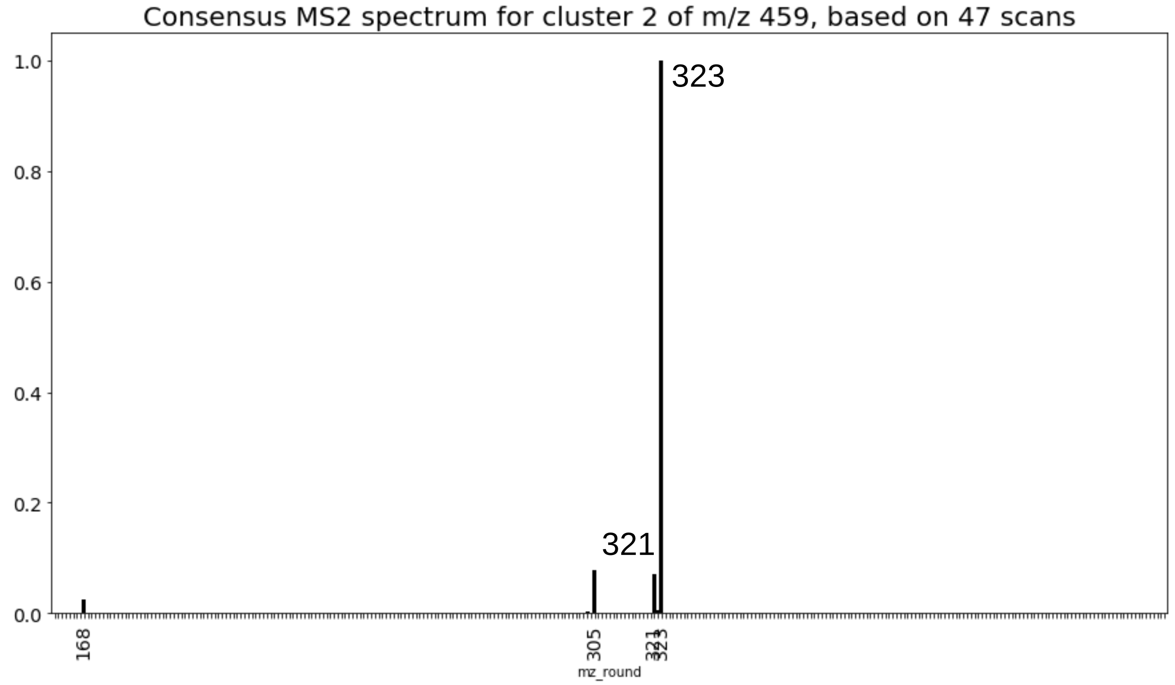


Figure 2.S13 - Doscadenamide D (**4**) consensus MS^2 spectrum. Consensus MS^2 spectrum representing a cluster of 47 scans for precursor mass *m/z* 459, representing the fragmentation spectrum of the proposed analog doscadenamide D (**4**).

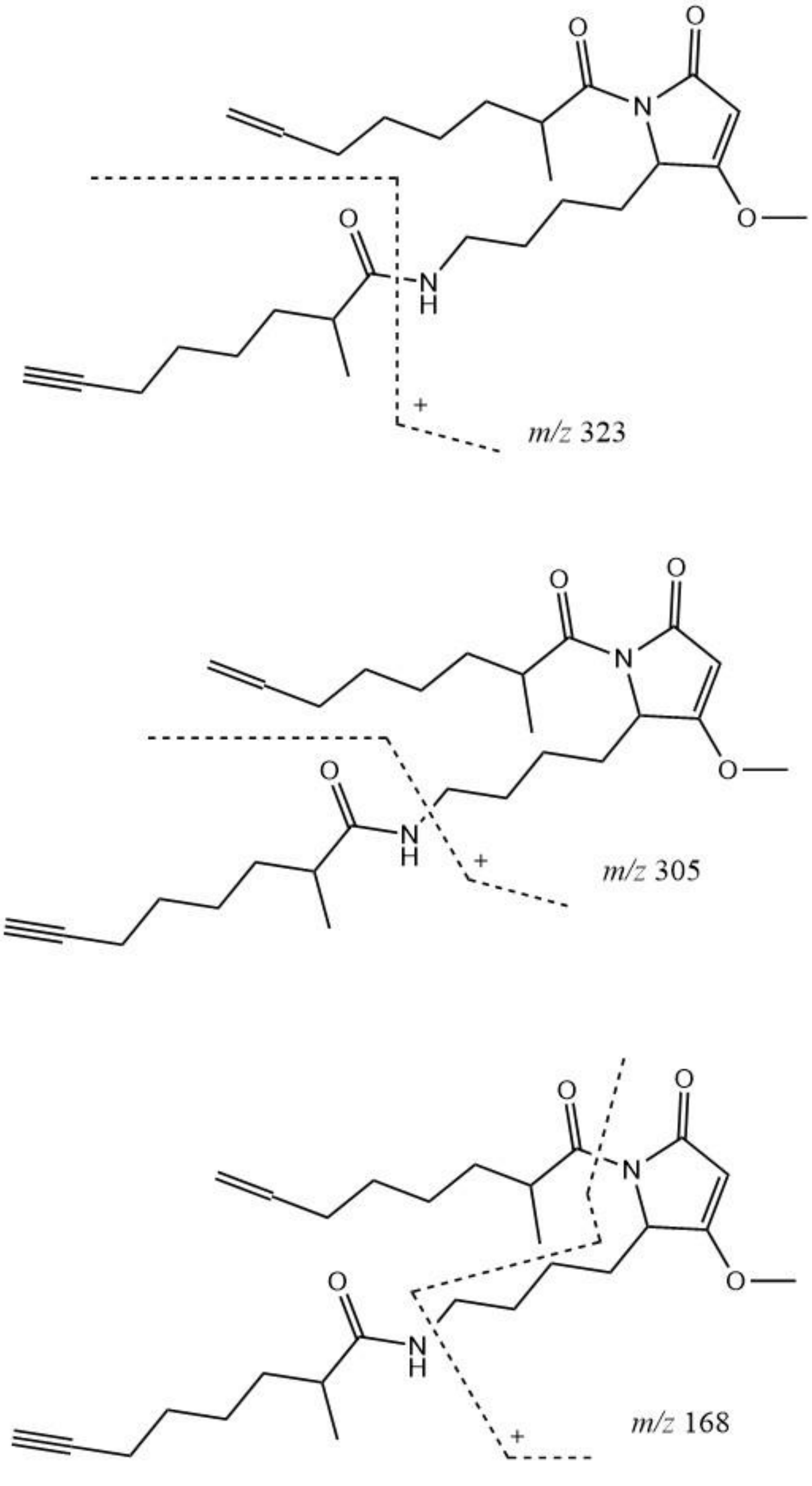


Figure 2.S14 - Doscadenamide D (**4**) proposed fragmentation

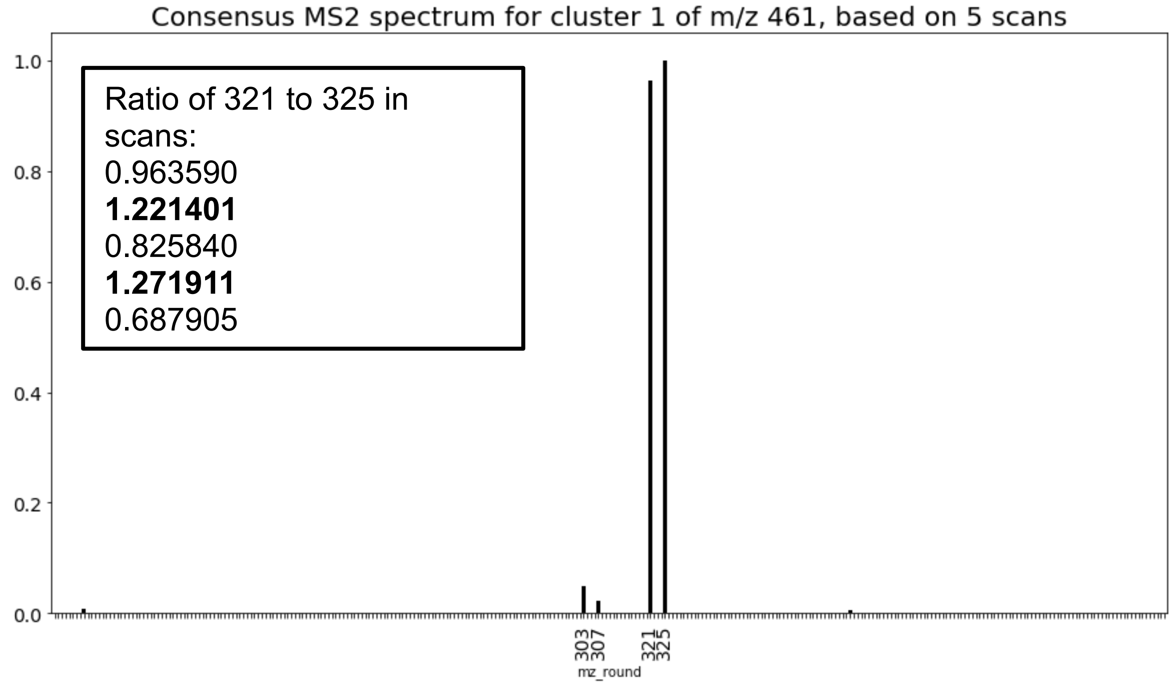


Figure 2.S15 - Doscadenamide E (**5**) consensus MS^2 spectrum. Consensus MS^2 spectrum representing a cluster of 5 scans for precursor mass *m/z* 461, representing the fragmentation spectrum of the proposed analog doscadenamide E (**5**). The expected fragmentation spectrum would have a fragment peak at *m/z* 321 that is more intense than the fragment peak at *m/z* 325, indicating the apparent propensity for side chains acylated to the terminus of the lysine side chain to fragment. The inlay in the top right hand corner reports the ratio of the *m/z* 321 peak relative intensity to *m/z* 325 peak relative intensity for the 5 scans represented in the consensus. These ratios reveal that in 2 scans, the fragment peak at *m/z* 321 is indeed more intense than the fragment peak at *m/z* 325.

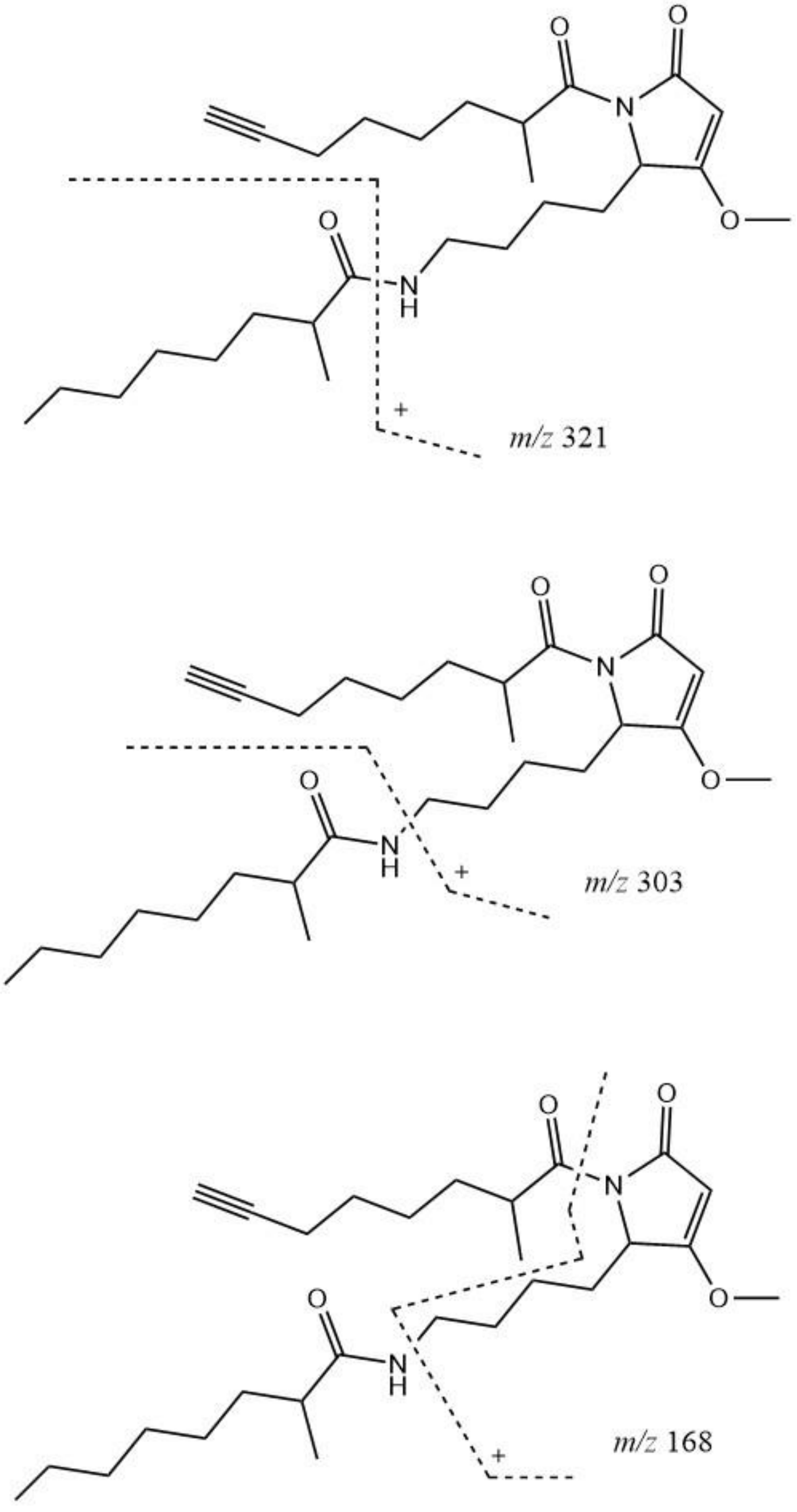


Figure 2.S16 - Doscadenamide E (**5**) proposed fragmentation

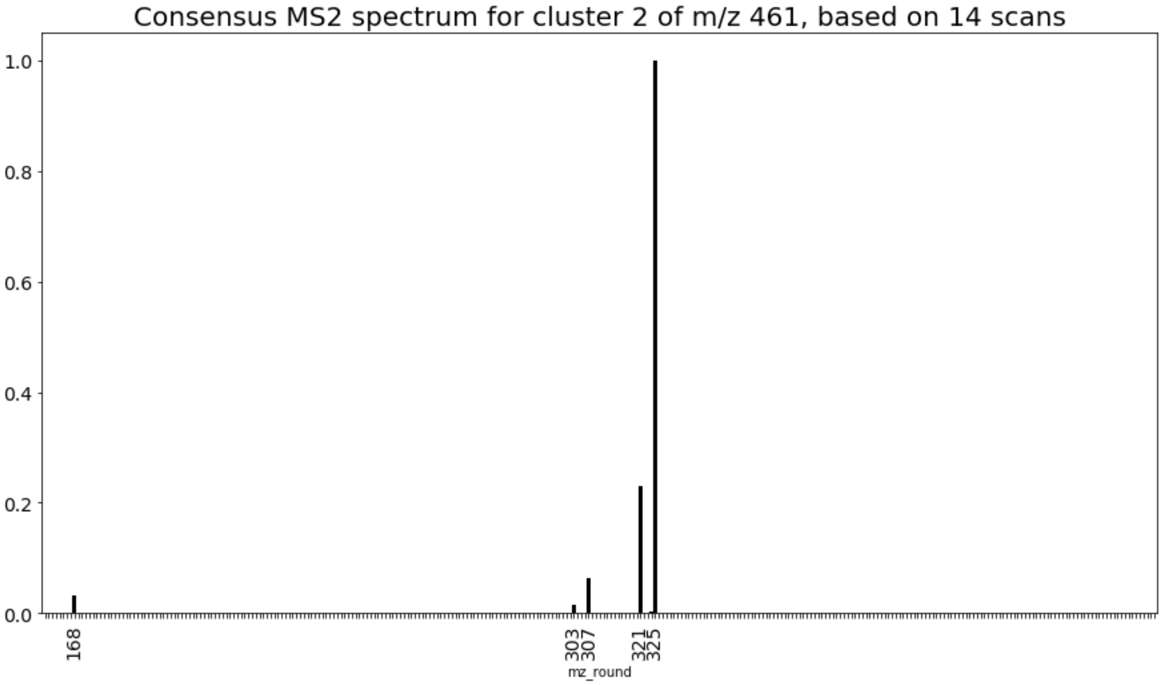


Figure 2.S17 - Doscadenamide F (**6**) consensus MS_2 spectrum. Consensus MS^2 spectrum representing a cluster of 14 scans for precursor mass *m/z* 461, representing the fragmentation spectrum of the proposed analog doscadenamide F (**6**).

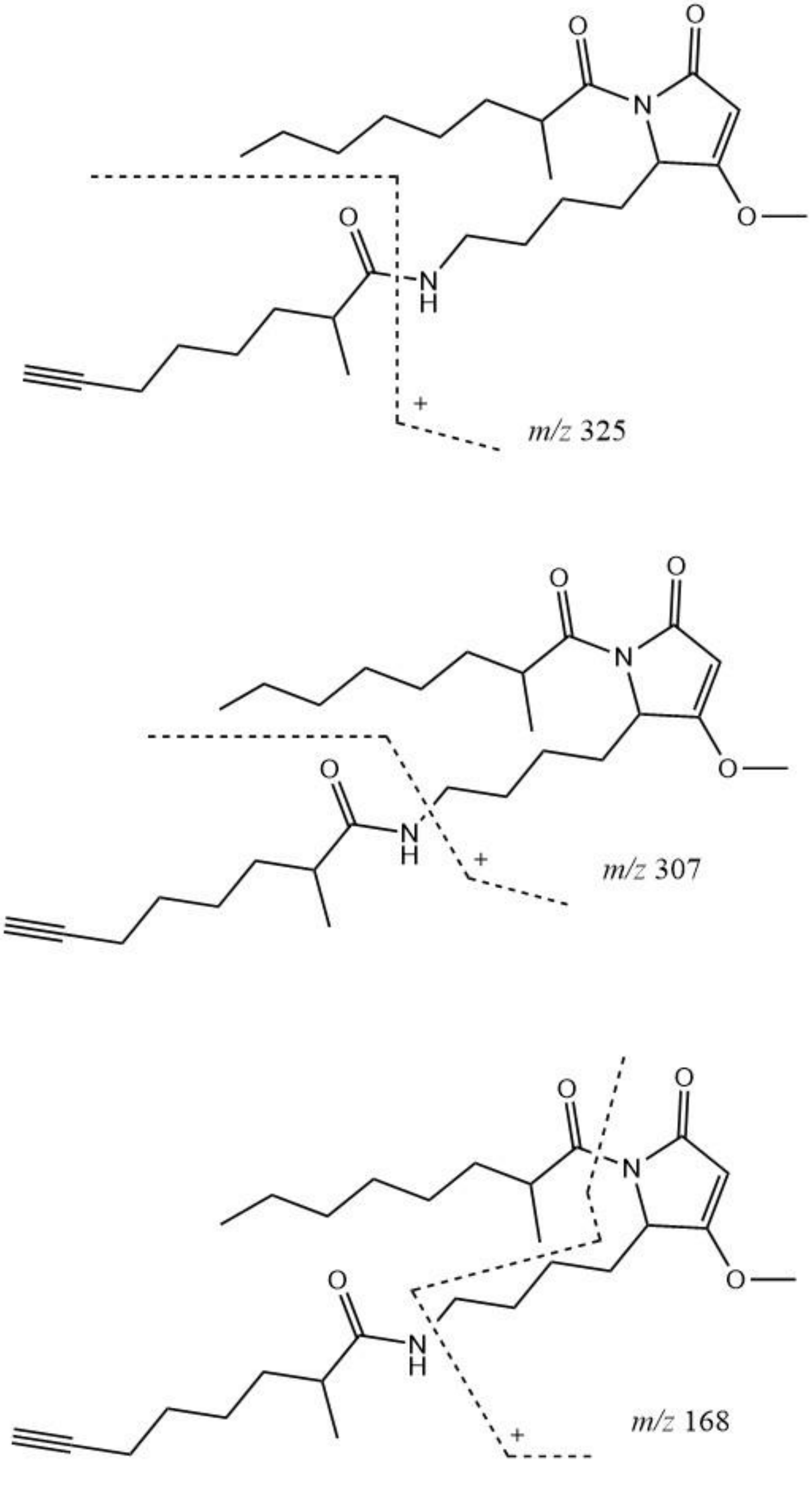


Figure 2.S18 - Doscadenamide F (**6**) proposed fragmentation

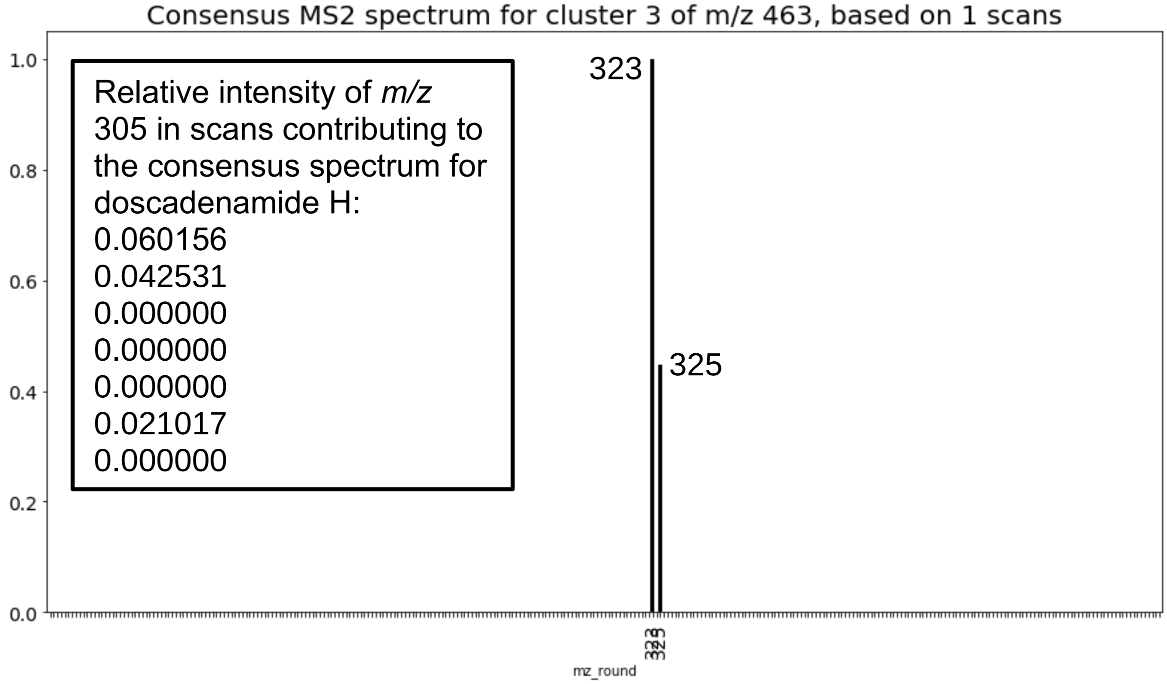


Figure 2.S19 - Doscadenamide G (**7**) consensus MS^2 spectrum. MS^2 spectrum captured in one scan for precursor mass *m/z* 463, representing the partial fragmentation spectrum of the proposed analog doscadenamide G (**7**). The expected *m/z* 305 and *m/z* 168 peaks were detected at too low of intensity to appear in this output. However, the *m/z* 305 peak is detected in several of the scans that make up the consensus spectrum representing doscadenamide H (Figure 2.S20) – the relative intensity of these *m/z* 305 fragment peaks are displayed in the above figure inlay. Detection of this m/z 305 peak is important because doscadenamides G and H coelute and such a fragment peak would not be produced by doscadenamide H. Therefore, this lends further support to the structural proposal for doscadenamide G.

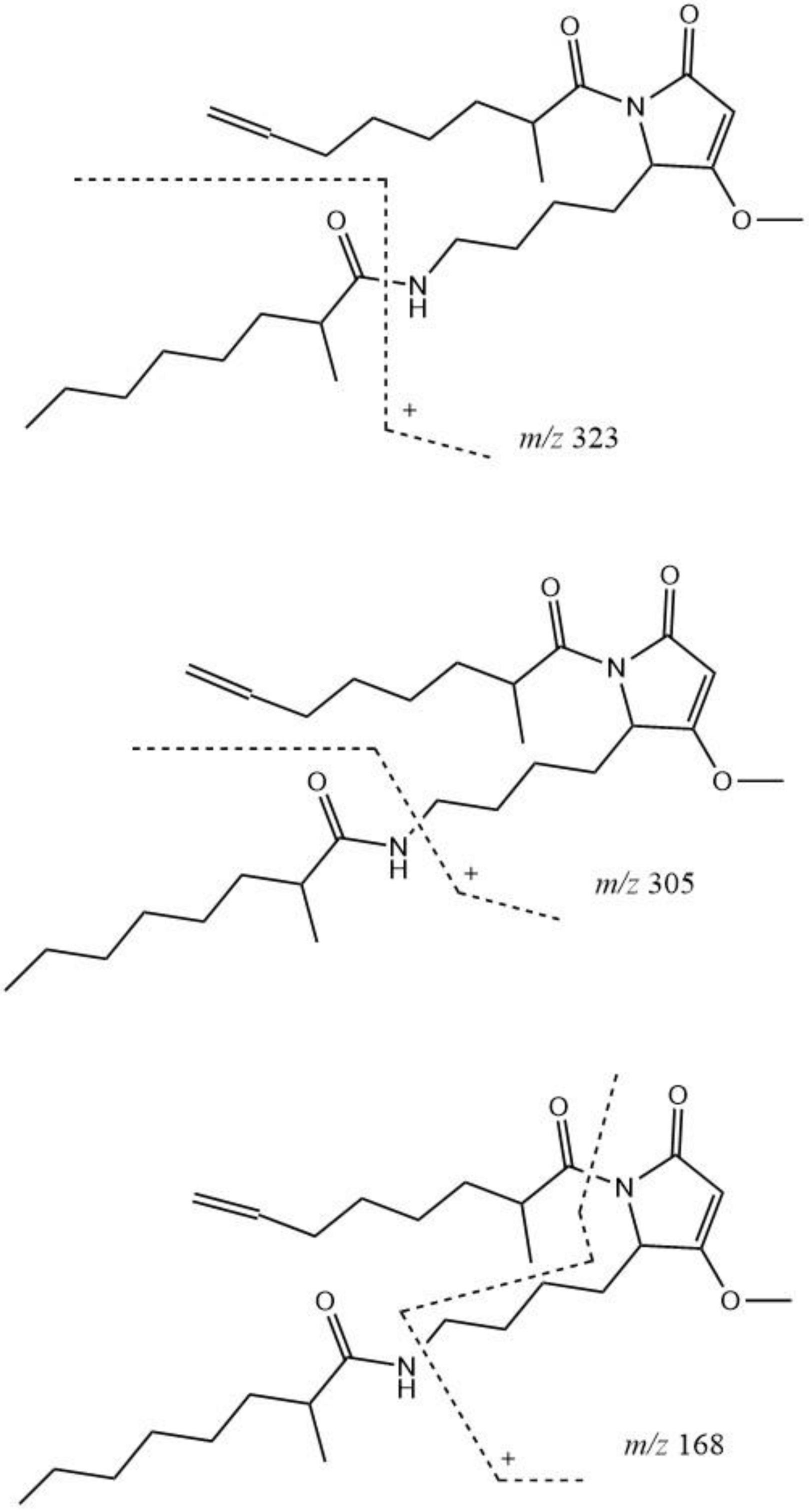


Figure 2.S20 - Doscadenamide G (**7**) proposed fragmentation

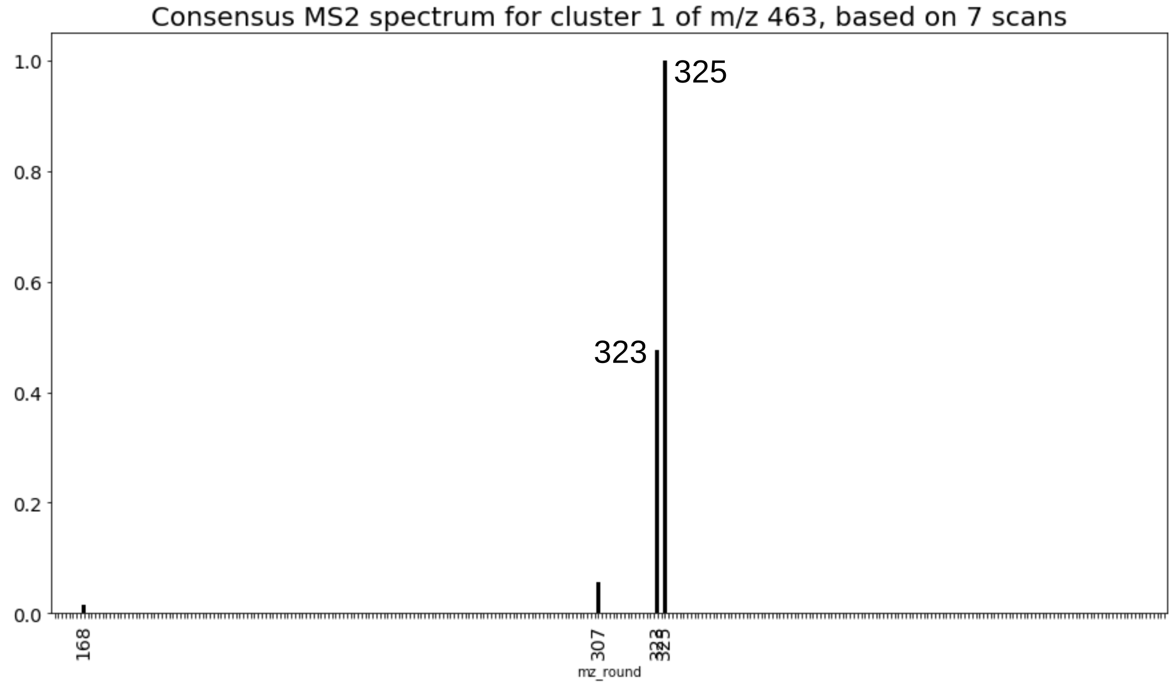


Figure 2.S21 - Doscadenamide H (**8**) consensus MS^2 spectrum. Consensus MS^2 spectrum representing a cluster of 7 scans for precursor mass *m/z* 463, representing the fragmentation spectrum of the proposed analog doscadenamide H (**8**).

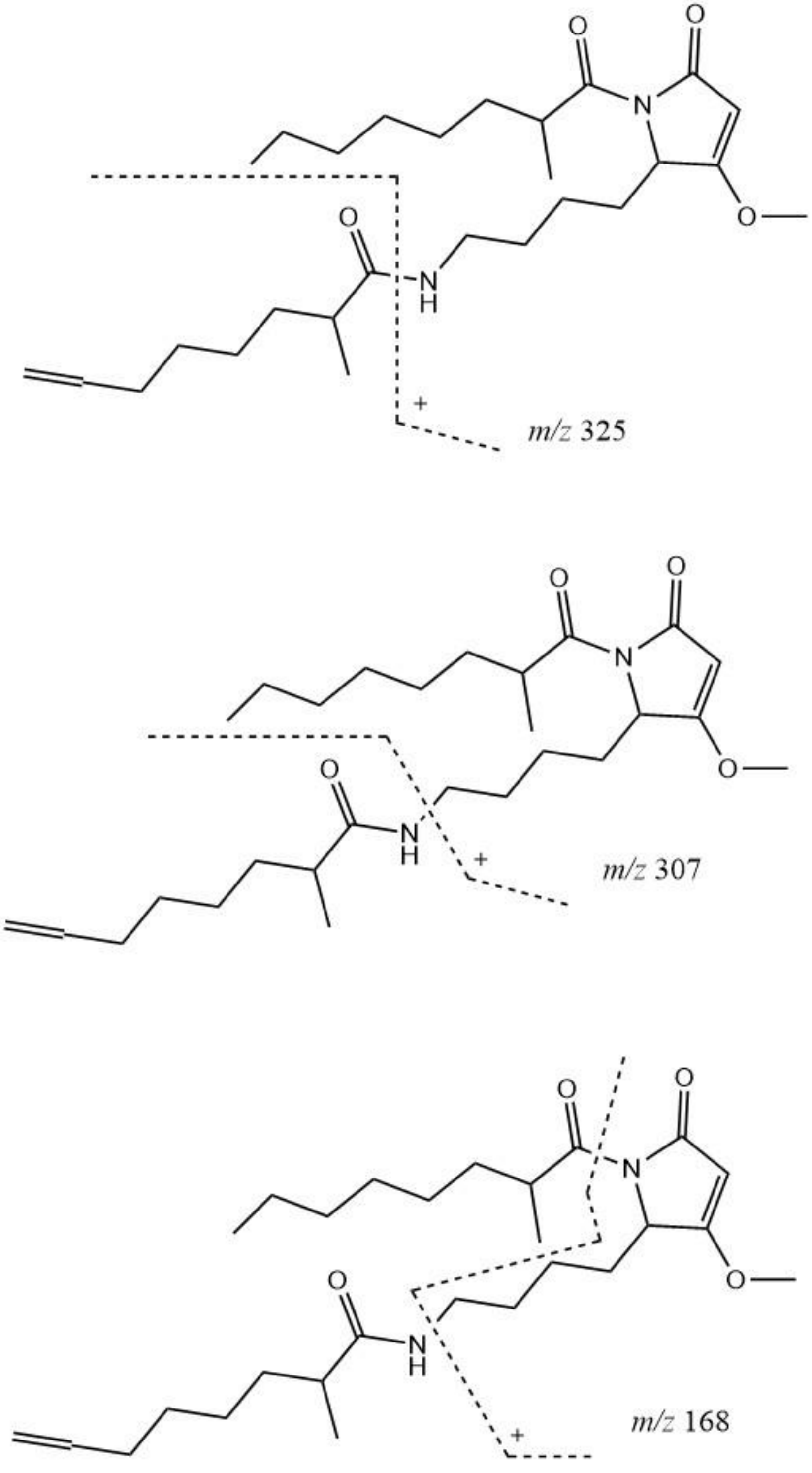


Figure 2.S22 - Doscadenamide H (**8**) proposed fragmentation

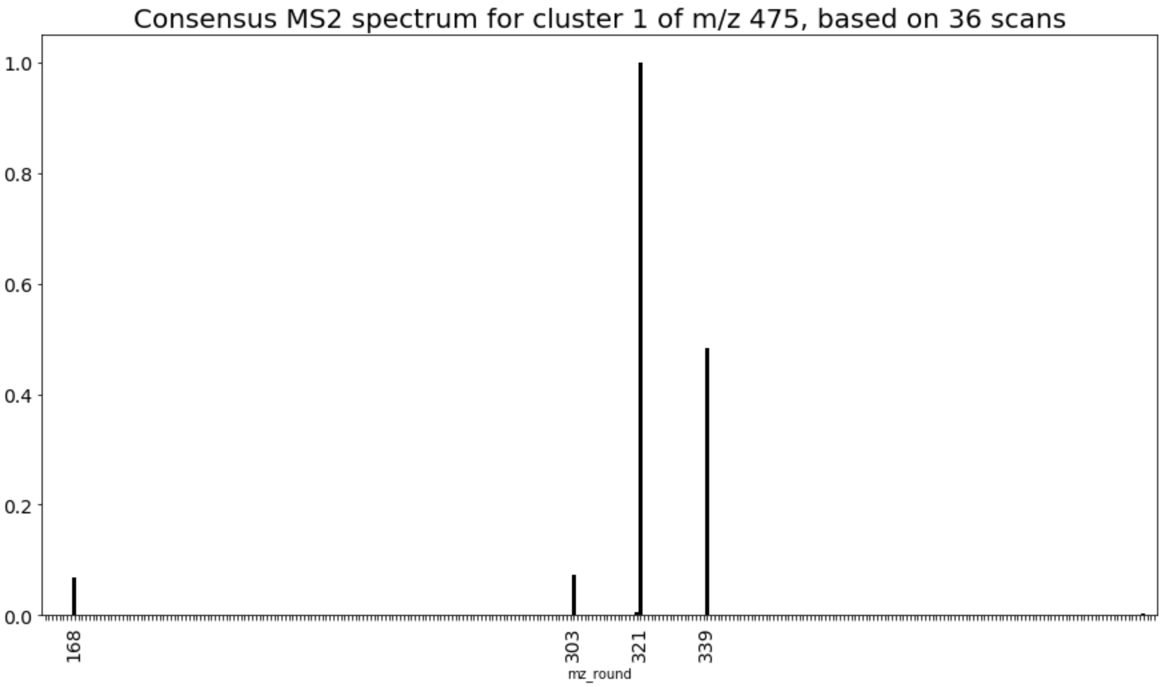


Figure 2.S23 - Doscadenamide I (**9**) consensus MS^2 spectrum. Consensus MS^2 spectrum representing a cluster of 36 scans for precursor mass *m/z* 475, representing the fragmentation spectrum of the proposed analog doscadenamide I (**9**).

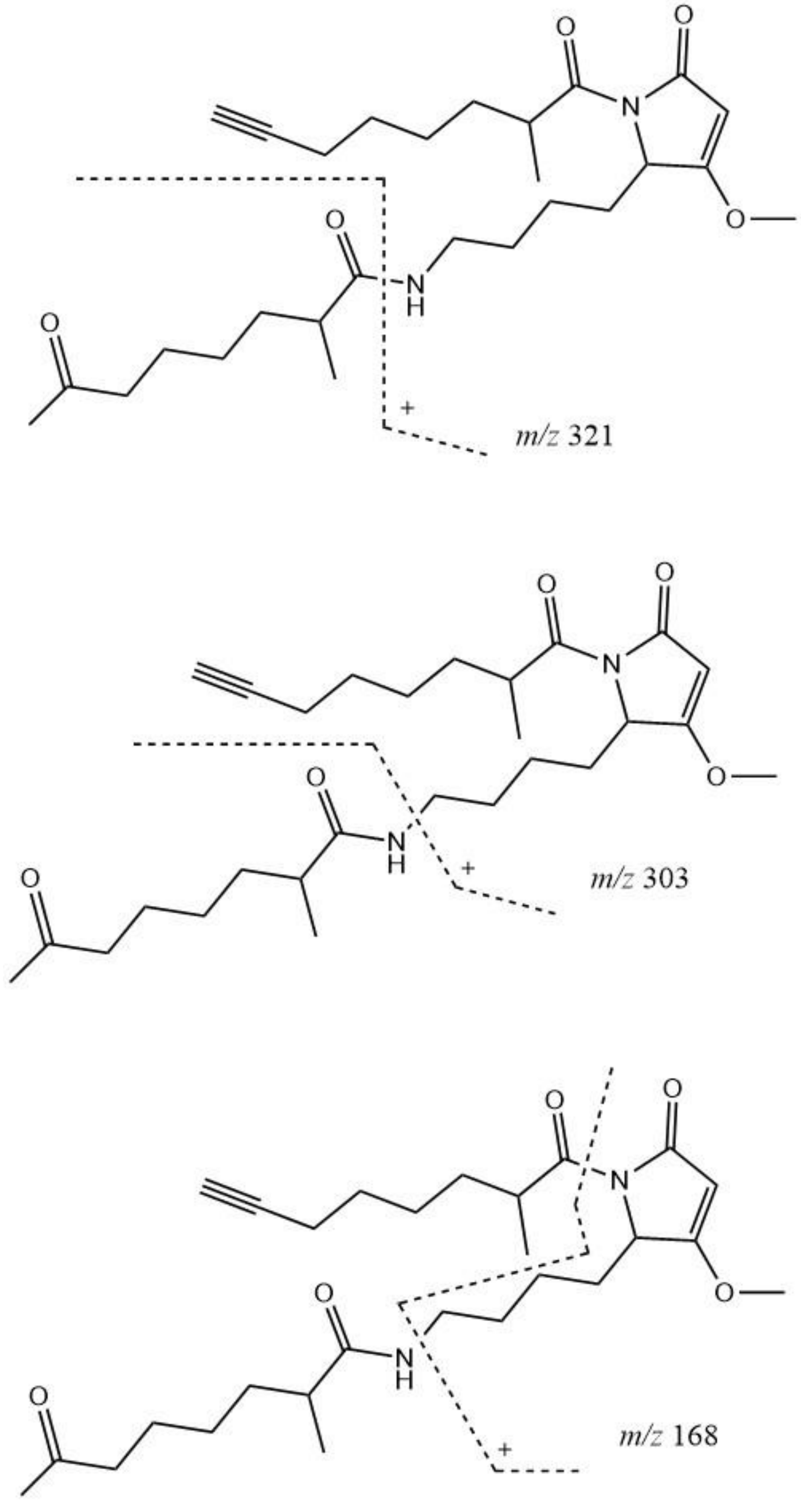


Figure 2.S24 - Doscadenamide I (**9**) proposed fragmentation

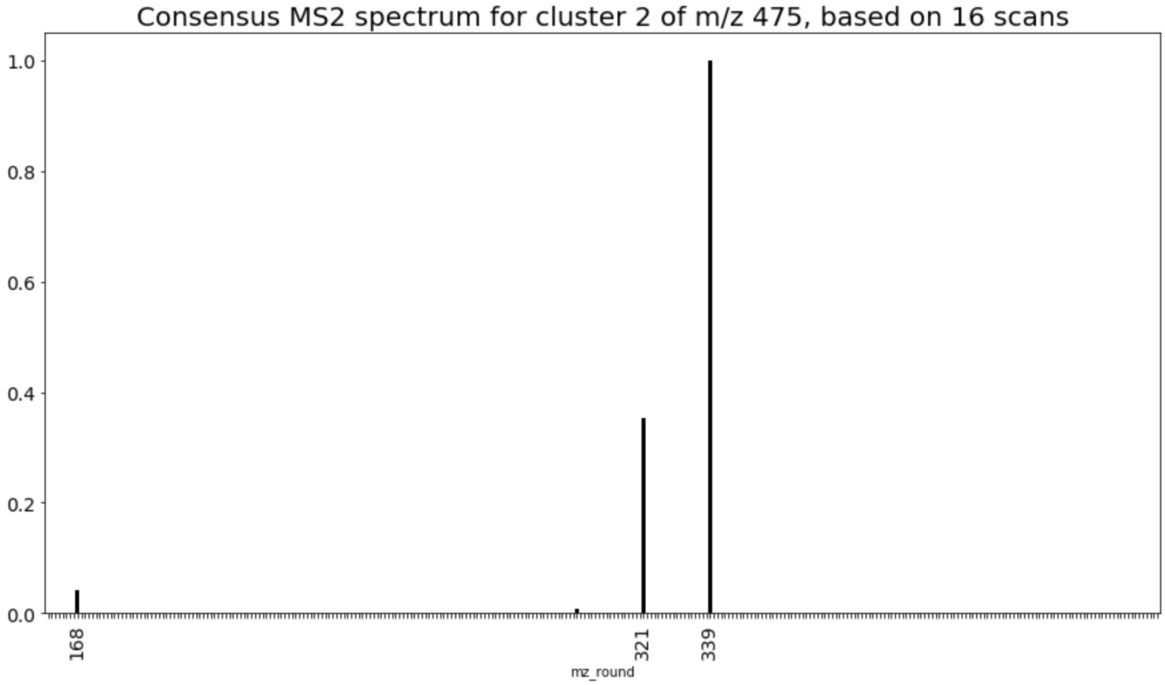


Figure 2.S25 - Doscadenamide J (**10**) consensus MS^2 spectrum. Consensus MS^2 spectrum representing a cluster of 16 scans for precursor mass *m/z* 475, representing the fragmentation spectrum of the proposed analog doscadenamide J (**10**).

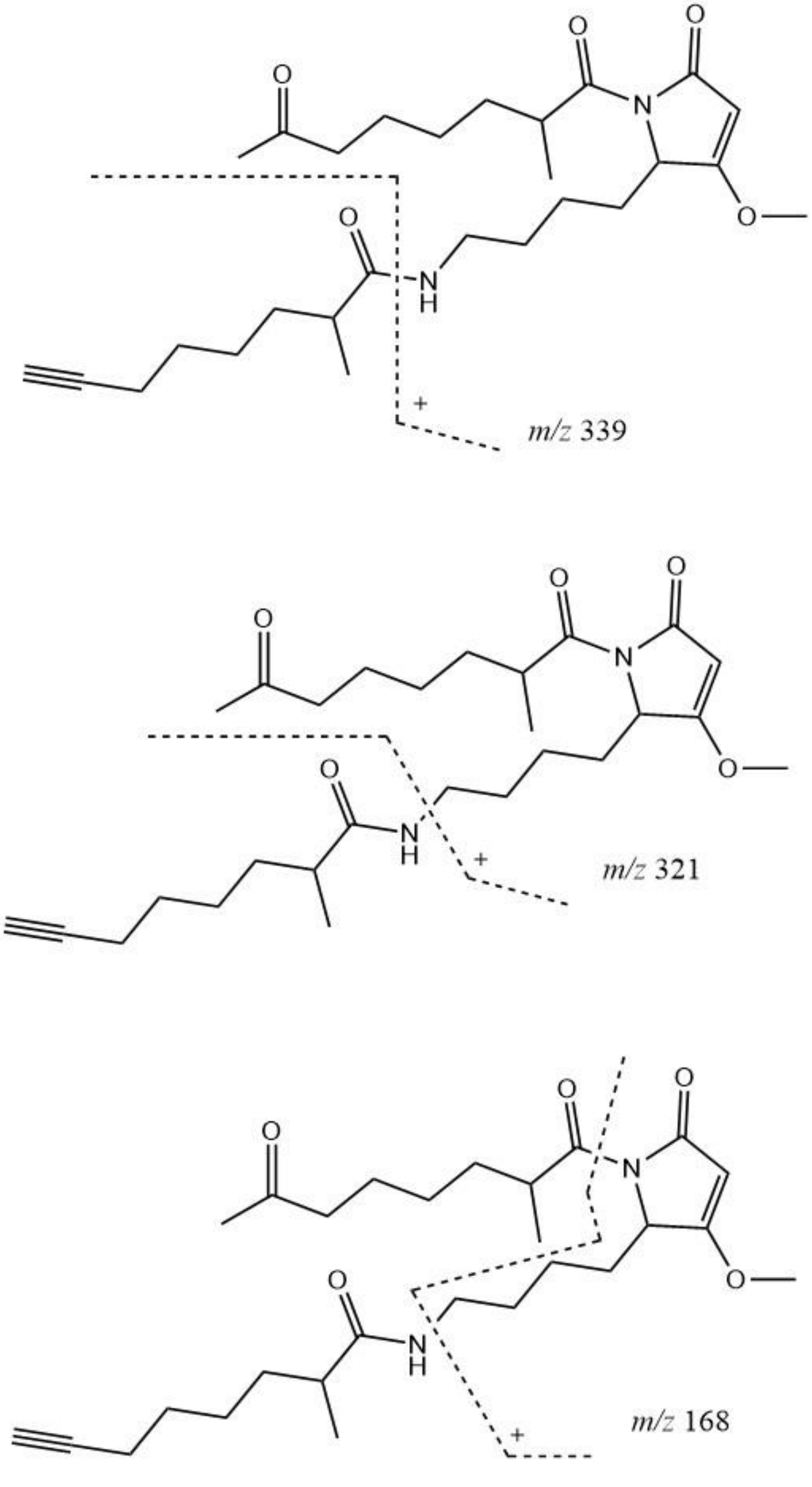


Figure 2.S26 - Doscadenamide J (**10**) proposed fragmentation

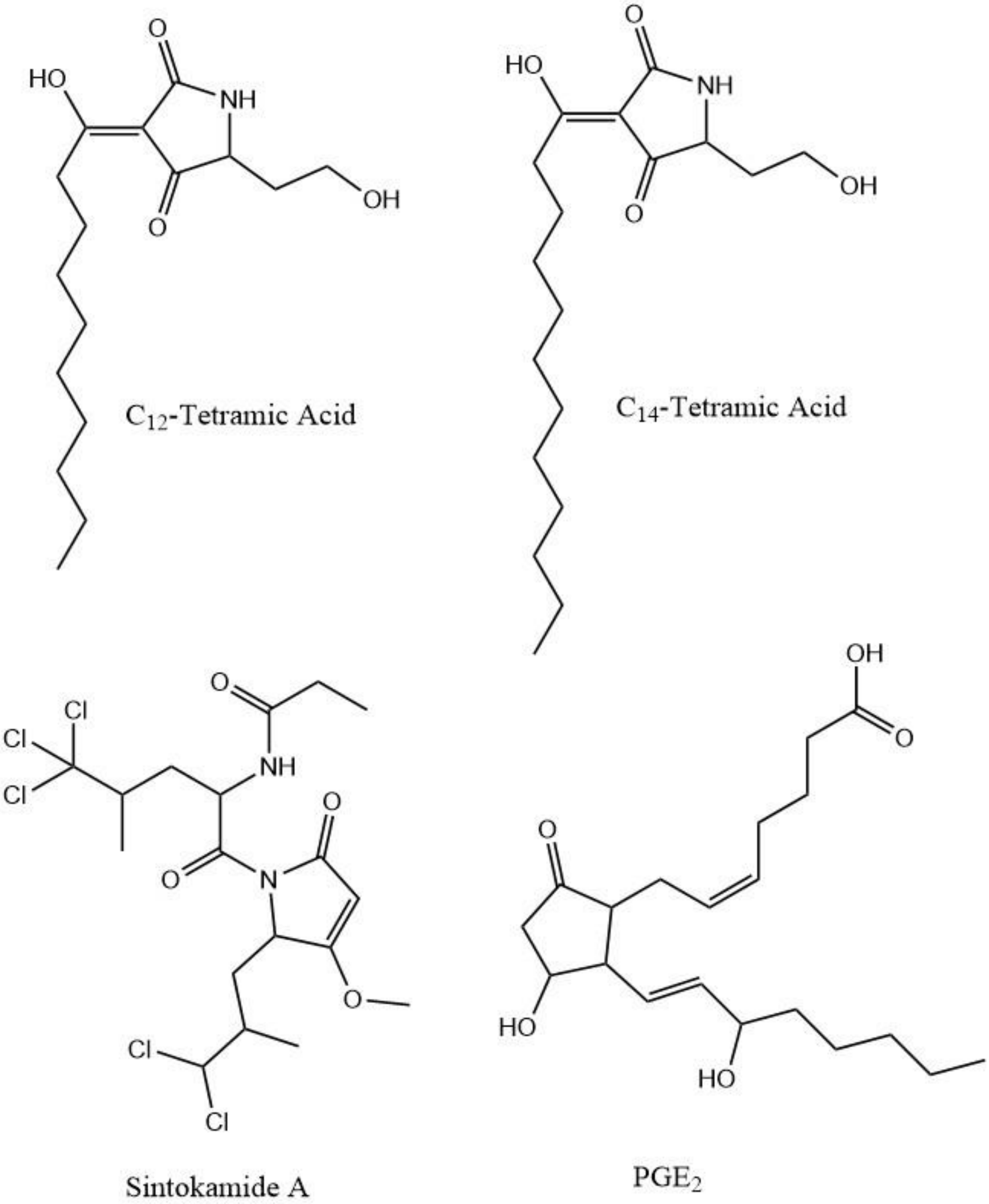


Figure 2.S27 - Representative structures from compound families similar to the doscadenamides

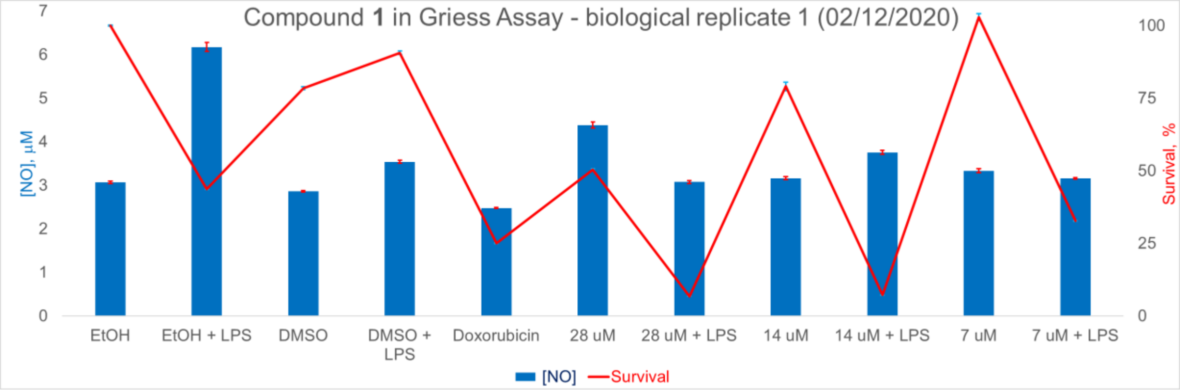


Figure 2.S28 - Results of compound **1** in Griess assay – biological replicate 1. Reagents were applied at the following concentrations: EtOH (1.5%), LPS (0.5 µg/mL), DMSO (1.0%), and doxorubicin (3.3 µg/mL). One-way ANOVA applied to the survival data indicated statistically significant differences between conditions (p-value < 0.01). Tukey's method was used to determine significance groups: EtOH (a), EtOH + LPS (b), DMSO (c), DMSO + LPS (ac), doxorubicin (d), 28 µM (b), 28 µM + LPS (e), 14 µM (c), 14 µM + LPS (e), 7 µM (a), 7 µM + LPS (d). This result indicates that when compound **1** is applied with LPS, it has a statistically significant negative impact on cell survival, as compared to compound **1** or LPS applied individually, at all three concentrations tested and in a dose-dependent fashion (e.g. when compound **1** was applied at 28 µM and 14 µM, with LPS, it had a statistically significant more negative impact on cell survival than when it was applied at 7 µM with LPS).

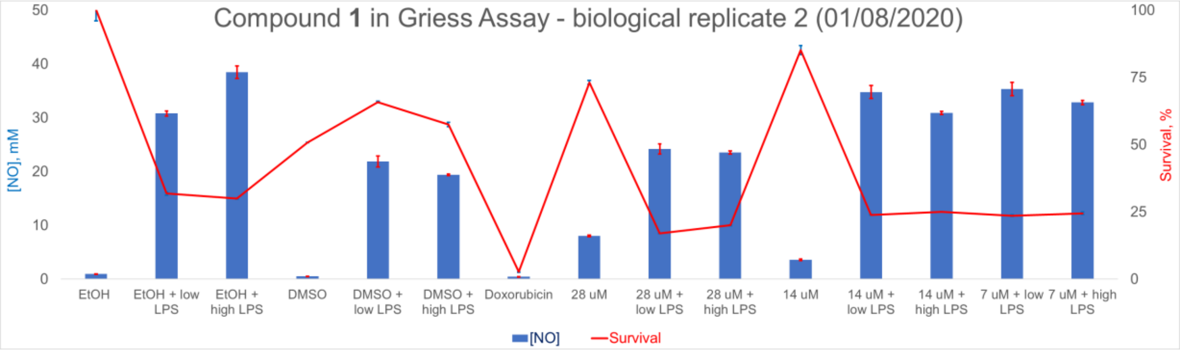


Figure 2.S29 - Results of compound **1** in Griess assay – biological replicate 2. Reagents were applied at the following concentrations: EtOH (1.0%), low LPS (0.5 µg/mL), high LPS (1.5 µg/mL), DMSO (1.0%), and doxorubicin (3.3 µg/mL). One-way ANOVA applied to the survival data indicated statistically significant differences between conditions (p-value < 0.01). Tukey's method was used to determine significance groups: EtOH (a), EtOH + low LPS (b), EtOH + high LPS (bc), DMSO (d), DMSO + low LPS (e), DMSO + high LPS (de), doxorubicin (f), 28 µM (eg), 28 µM + low LPS (c), 28 µM + high LPS (bc), 14 µM (g), 14 µM + low LPS (bc), 14 µM + high LPS (bc), 7 µM + low LPS (bc), 7 µM + high LPS (bc). No statistically significant difference was found between LPS conditions and compound **1** plus LPS conditions. However, the results still reveal a trend towards compound **1** synergistic cytotoxicity when applied with 0.5 µg/mL LPS, producing an average survival percentage of 17.1, 23.9 and 23.7% at 28, 14 and 7 µM, respectively. In comparison, 0.5 µg/mL LPS applied with EtOH resulted in an average survival of 31.8%.

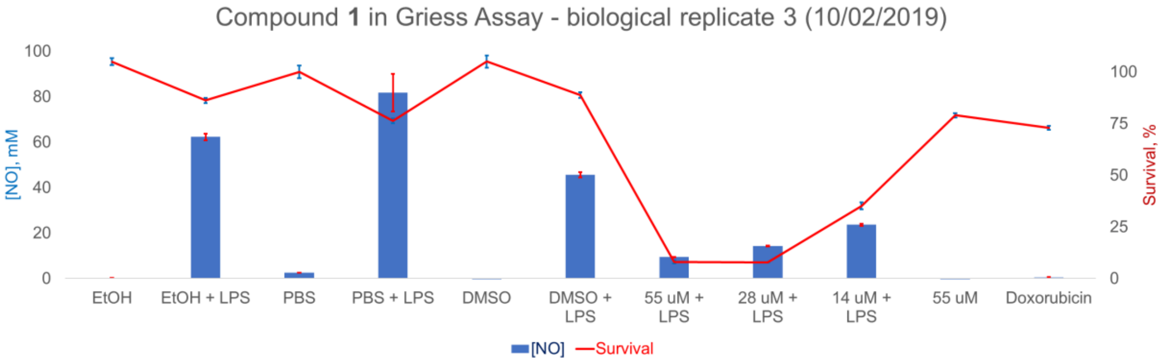


Figure 2.S30 - Results of compound **1** in Griess assay – biological replicate 3. Reagents were applied at the following concentrations: EtOH (1.0%), LPS (1.5 µg/mL), DMSO (1.0%), and doxorubicin (3.3 µg/mL). Additional control conditions were included in this assay run; phosphate-buffered saline (PBS) with and without LPS was tested. One-way ANOVA applied to the survival data indicated statistically significant differences between conditions (p-value < 0.01). Tukey's method was used to determine significance groups: EtOH (a), EtOH + LPS (ab), PBS (ab), PBS + LPS (ab), DMSO (a), DMSO + LPS (ab), 55 µM + LPS (c), 28 µM + LPS (c), 14 µM + LPS (b), 55 µM (ab), doxorubicin (b). This result illustrates the dose-dependent cytotoxicity of compound **1** when applied with LPS. Statistically significant negative impacts on cell survival were observed when compound **1** was applied with LPS at concentrations of 55 µM and 28 µM, as compared to LPS applied with negative control (EtOH or PBS) and compound **1** applied at 55 µM without LPS.

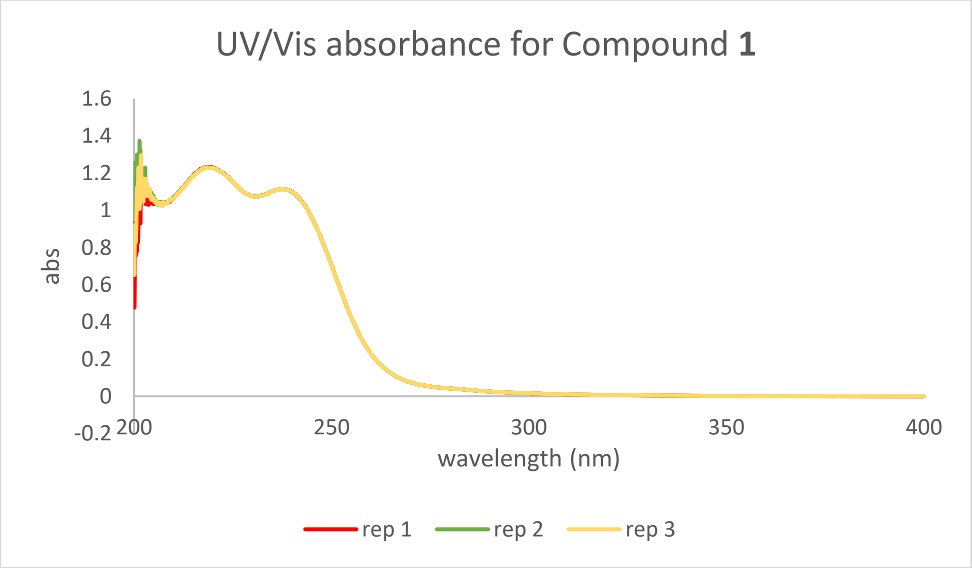


Figure 2.S31 - UV/Vis absorbance spectrum (200-400 nm) for compound **1**

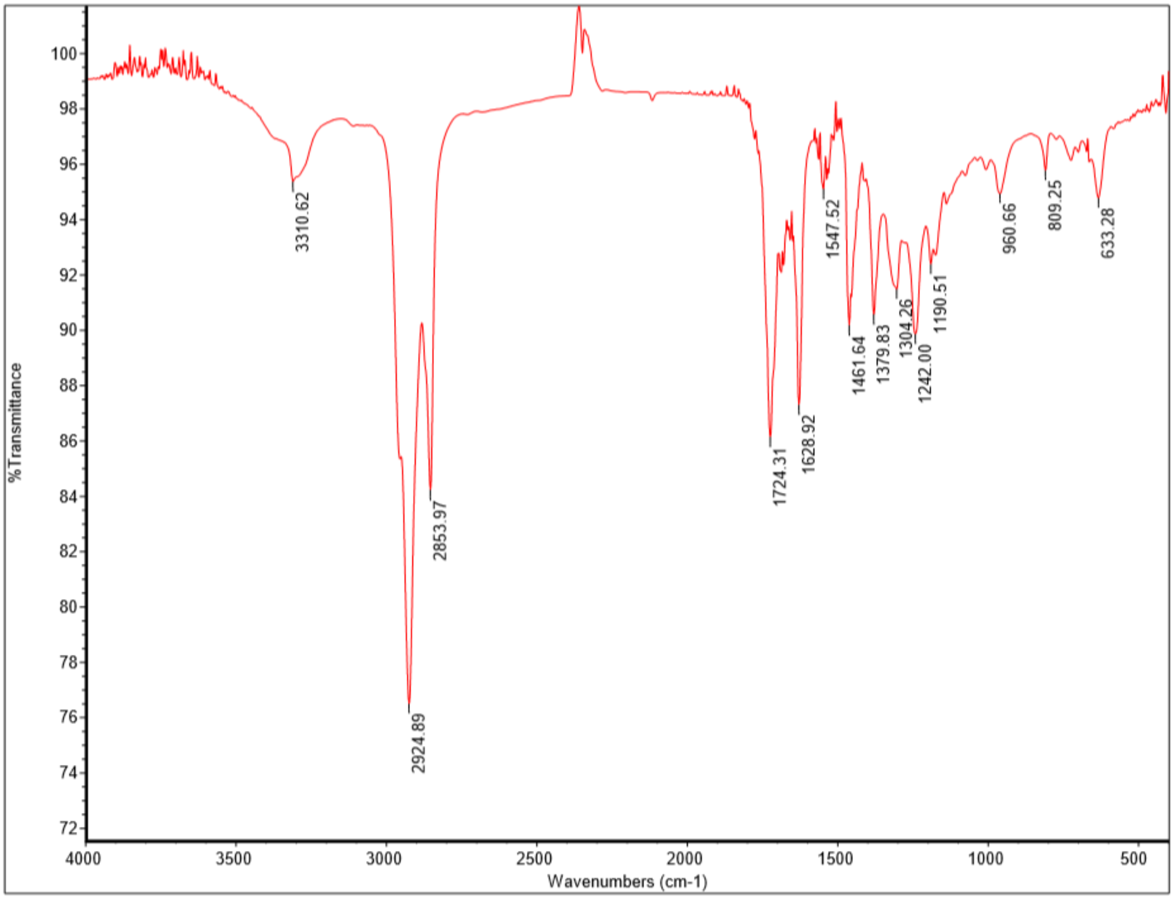


Figure 2.S32 - IR spectrum for compound **1**

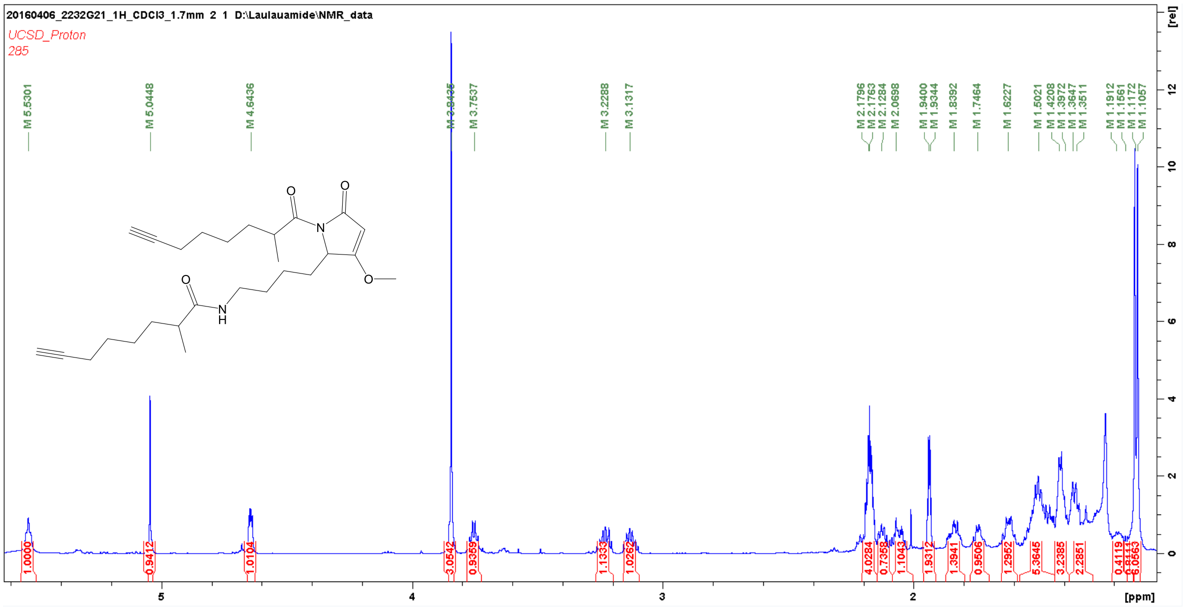


Figure 2.S33 - ^{1}H NMR spectrum for compound **1**

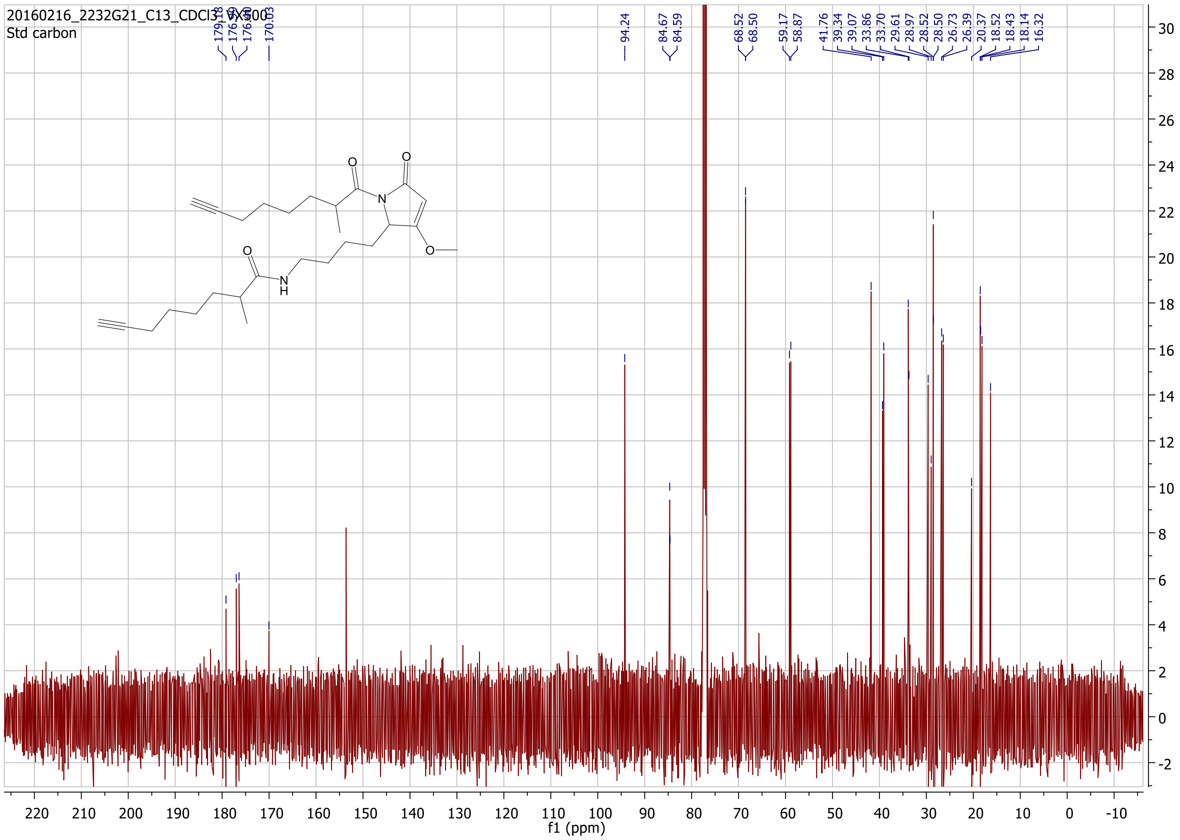


Figure 2.S34 - ^{13}C NMR spectrum for compound **1**

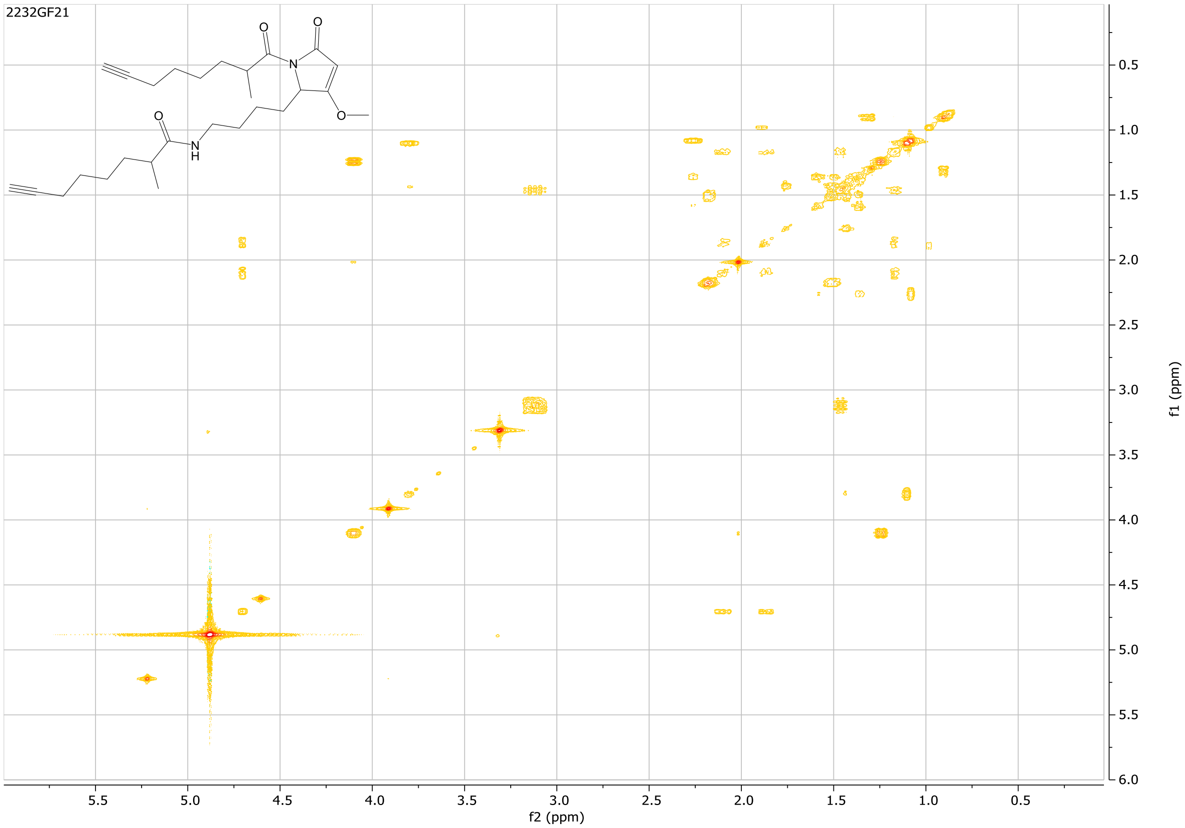


Figure 2.S35 - ^{1}H-^{1}H COSY spectrum for compound **1**

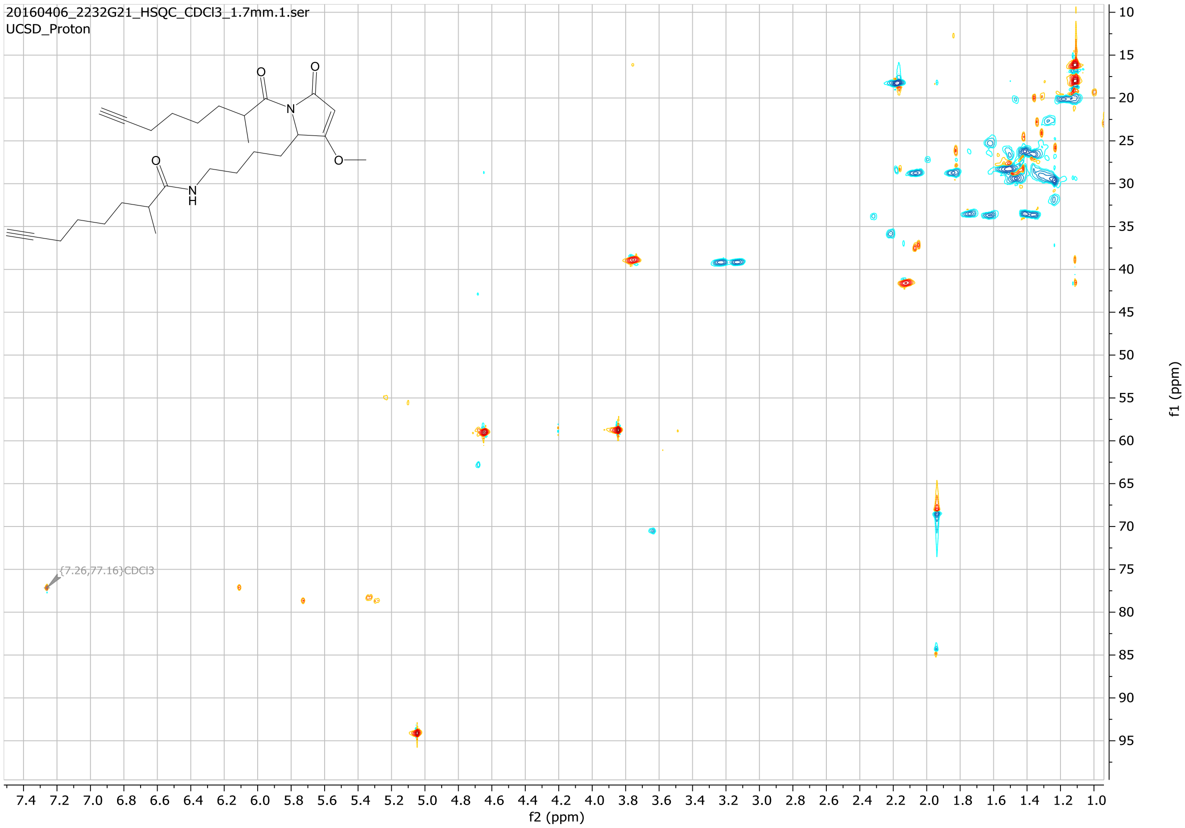


Figure 2.S36 - ^{1}H-^{13}C HSQC spectrum for compound **1**

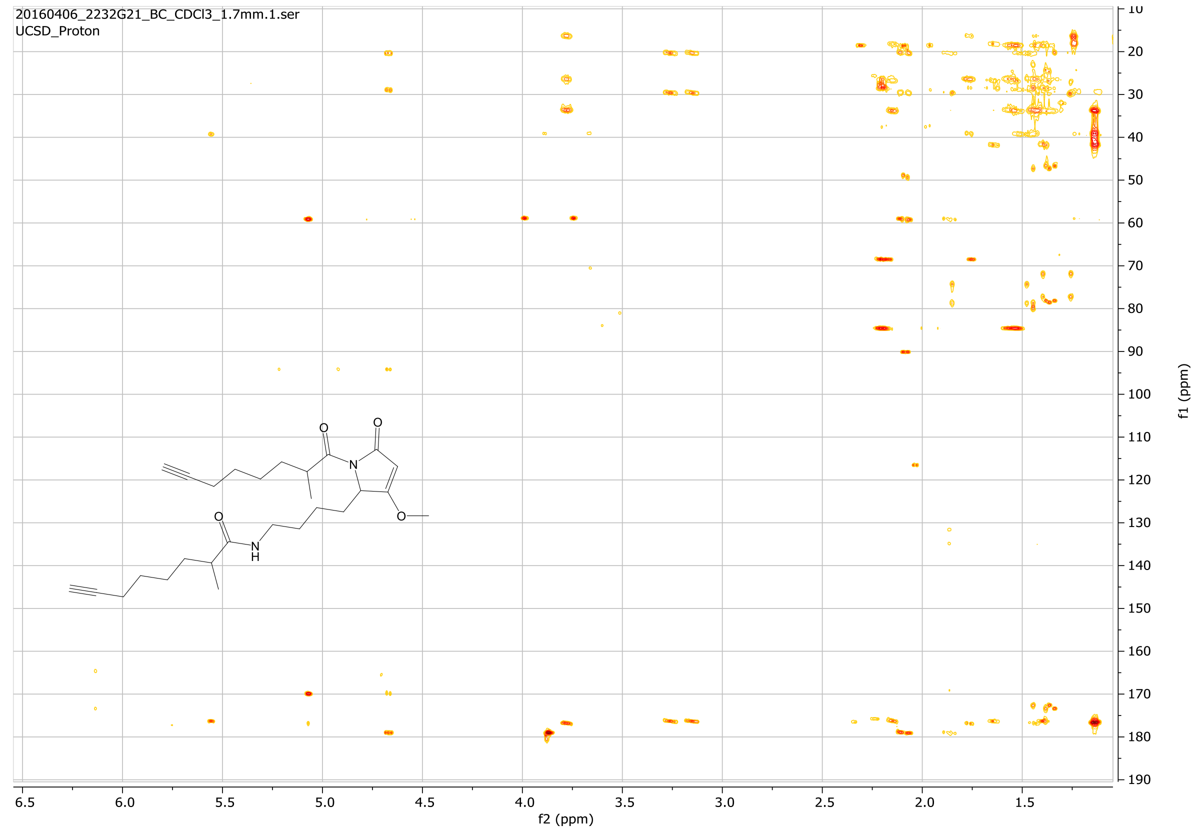


Figure 2.S37 - ^{1}H-^{13}C HMBC spectrum for compound **1**

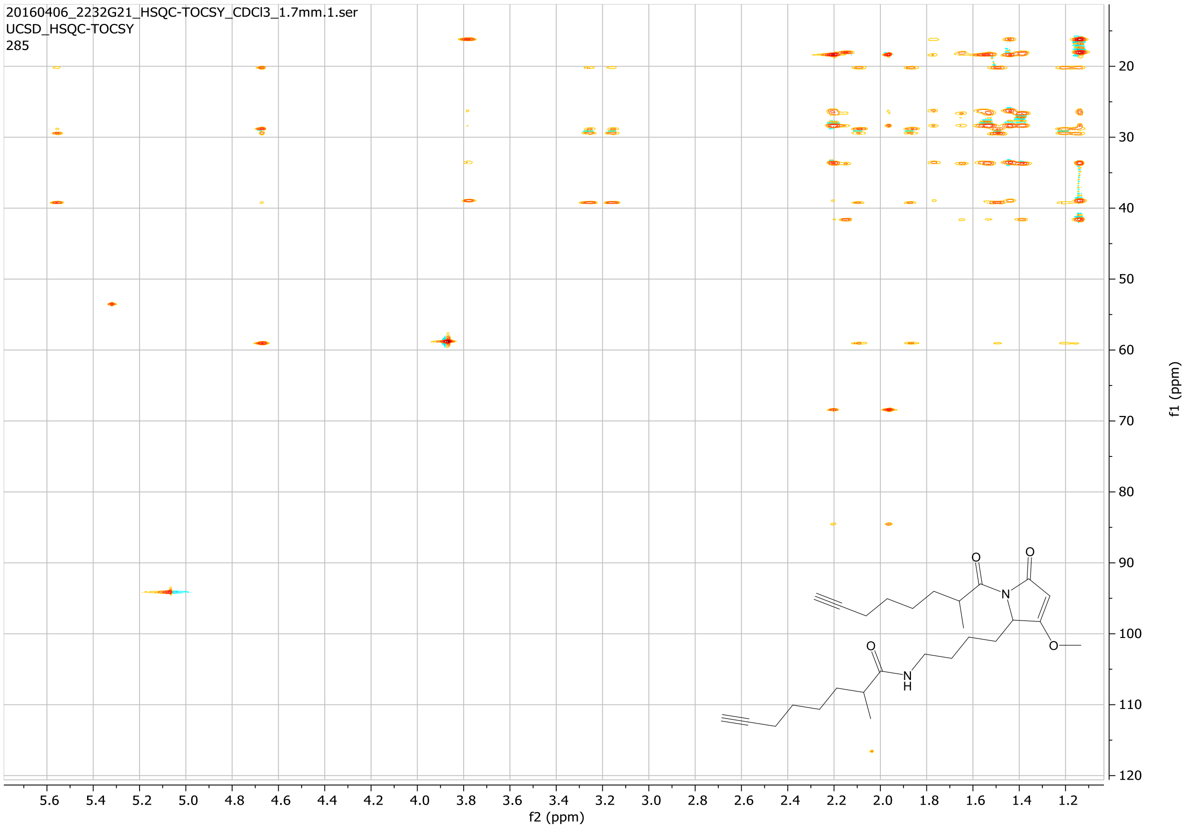


Figure 2.S38 - ^{1}H-^{13}C HSQC-TOCSY spectrum for compound **1**

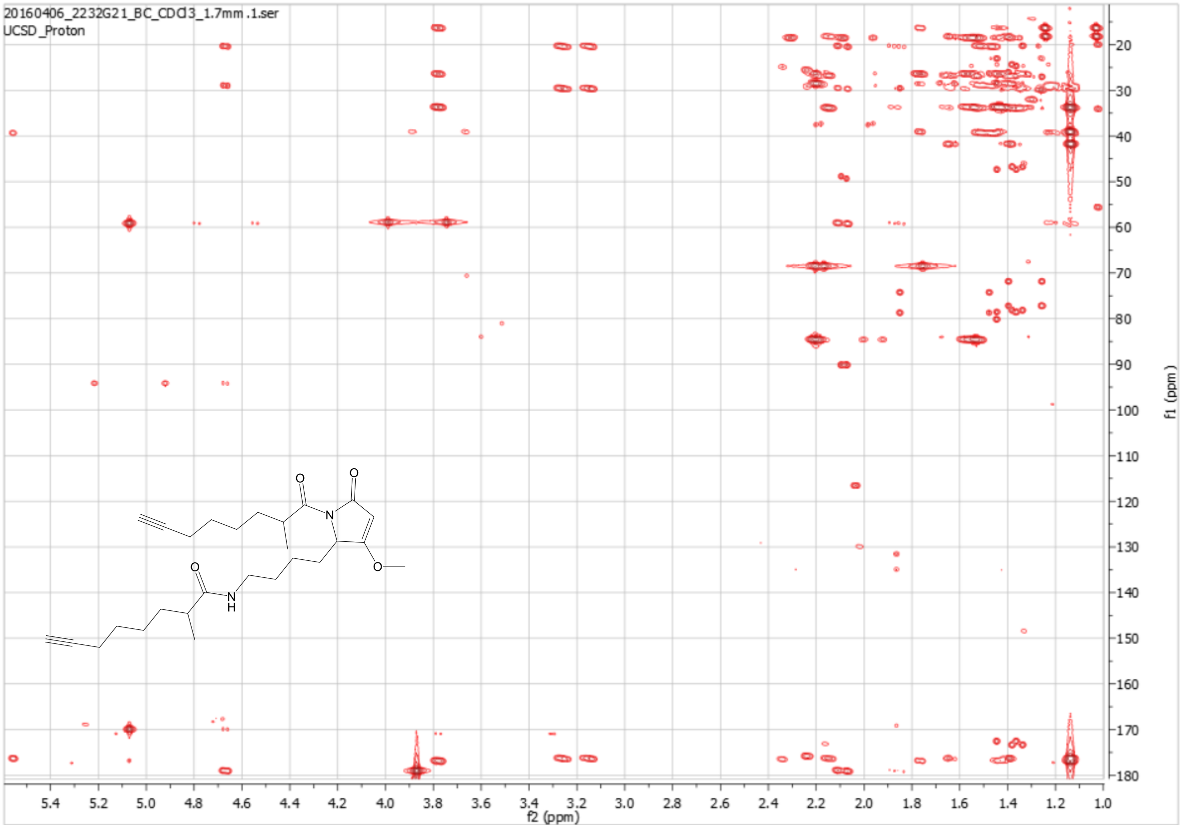


Figure 2.S39 - ^{1}H-^{13}C Long-range HSQMBC spectrum for compound **1**

Table 2.S1 Known compounds isolated from *M. bouillonii*

Name	Monoisotopic Mass	Protenated Peak	Sodiated Peak	Location of initial isolation	Notes
15-norlyngbyapeptin A	683.3716	684.3786	706.3608	Palau, Guam	1
18E-lyngbyaloside C	648.2509	649.2579	671.2401	Guam	
18Z-lyngbyaloside C	648.2509	649.2579	671.2401	Guam	
27-deoxylyngbyabellin A	674.1766	675.1836	697.1658	Guam	
2-epi-lyngbyaloside	660.2509	661.2579	683.2401	Guam	
7-epilyngbyabellin L	544.1105	545.1175	567.0997	Palmyra Atoll	
alotamide A	587.3393	588.3463	610.3285	Papua New Guinea	
apramide A	976.5819	977.5889	999.5711	Guam	
apramide B	962.5663	963.5733	985.5555	Guam	
apramide C	978.5976	979.6046	1001.5868	Guam	
apramide D	1002.5976	1003.6046	1025.5868	Guam	
apramide E	988.5819	989.5889	1011.5711	Guam	
apramide F	1004.6132	1005.6202	1027.6024	Guam	
apramide G	827.5343	828.5413	850.5235	Guam	
apratoxin A	839.4866	840.4936	862.4758	Guam	2
apratoxin A sulfoxide	855.4816	856.4886	878.4708	Red Sea	3
apratoxin B	825.471	826.4780	848.4602	Guam	4
apratoxin C	825.471	826.4780	848.4602	Palau	4
apratoxin D	882.5415	883.5485	905.5307	Papua New Guinea	5
apratoxin E	795.4604	796.4674	818.4496	Guam	
apratoxin F	827.4866	828.4936	850.4758	Palmyra Atoll	
apratoxin G	813.471	814.4780	836.4602	Palmyra Atoll	
apratoxin H	853.5023	854.5093	876.4915	Red Sea	3
apratyramide	804.4309	805.4379	827.4201	Guam	
bouillonamide	817.49896	818.5060	840.4882	Papua New Guinea	
bouillomide A (lyngbyastatin 9)	960.4956	961.5026	983.4848	Guam	
bouillomide B (lyngbyastatin 10)	1038.4062	1039.4132	1061.3954	Guam	
columbamide A	465.2413	466.2483	488.2305	Papua New Guinea	
columbamide B	499.2023	500.2093	522.1915	Papua New Guinea	
columbamide C	423.2307	424.2377	446.2199	Papua New Guinea	
columbamide D	451.262	452.2690	474.2512	Malaysia	
columbamide E	485.223	486.2300	508.2122	Malaysia	
columbamide F	493.2726	494.2796	516.2618	Malaysia	
columbamide G	527.2336	528.2406	550.2228	Malaysia	
columbamide H	417.301	418.3080	440.2902	Malaysia	
cyanolide A	832.482	833.4890	855.4712	Papua New Guinea	
doscadenamide A	456.2988	457.3058	479.2880	Guam	
kakeromamide A	790.4088	791.4158	813.3980	Japan	
kakeromamide B	790.4088	791.4158	813.3980	Fiji	6
kanamienamide	492.3563	493.3633	515.3455	Japan	
laingolide	351.2773	352.2843	374.2665	Papua New Guinea	
laingolide A	337.2617	338.2687	360.2509	Papua New Guinea	
laingolide B	369.2071	370.2141	392.1963	Guam	
lyngbouilloside	584.356	585.3630	607.3452	Papua New Guinea	
lyngbyabellin A	690.1715	691.1785	713.1607	Guam	2
lyngbyabellin B	678.1715	679.1785	701.1607	Guam	2
lyngbyabellin C	608.082	609.0890	631.0712	Palau	1
lyngbyabellin D	895.2553	896.2623	918.2445	Palau, Guam	1
lyngbyabellin J	863.2291	864.2361	886.2183	Guam	
lyngbyabellin K	578.0715	579.0785	601.0607	Palmyra Atoll	
lyngbyabellin L	544.1105	545.1175	567.0997	Palmyra Atoll	
lyngbyabellin M	624.1133	625.1203	647.1025	Palmyra Atoll	
lyngbyabellin N	904.292	905.2990	927.2812	Palmyra Atoll	
lyngbyaloside	660.2509	661.2579	683.2401	Papua New Guinea	
lyngbyaloside B	648.2509	649.2579	671.2401	Palau	1
lyngbyapeptin A	697.3873	698.3943	720.3765	Papua New Guinea	
lyngbyapeptin B	721.3509	722.3579	744.3401	Palau	1
lyngbyapeptin C	735.3665	736.3735	758.3557	Palau	1
lyngbyapeptin D	683.3716	684.3786	706.3608	Guam	
lyngbyastatin 2	1058.6878	1059.6948	1081.6770	Guam	2
mandangolide	377.561	378.5680	400.5502	Papua New Guinea	

mooreamide A	389.293	390.3000	412.2822	Papua New Guinea	
norlyngbyastatin 2	1044.6722	1045.6792	1067.6614	Guam	2
palau'imide	428.2675	429.2745	451.2567	Palau	1
ulongamide A	627.309	628.3160	650.2982	Palau	1
ulongamide B	643.3039	644.3109	666.2931	Palau	1
ulongamide C	691.3039	692.3109	714.2931	Palau	1
ulongamide D	671.3352	672.3422	694.3244	Palau	1
ulongamide E	685.3509	686.3579	708.3401	Palau	1
ulongamide F	607.3403	608.3473	630.3295	Palau	1

[1] Reported as *Lyngbya sp.*; cited in subsequent publications as *M. bouillonii*

[2] Reported as *Lyngbya majuscula*; cited in subsequent publications as *M. bouillonii*

[3] Reported as *Moorea producens*; 16S classification inconclusive; chemistry associated with *M. bouillonii*

[4] Reported as *Lyngbya sp.*; but cited in subsequent publications as *M. bouillonii* and reported to grow with *Alpheus frontalis*

[5] Reported as *Lyngbya majuscula* and *Lyngbya sordid*; 16S classification inconclusive; chemistry associated with *M. bouillonii*

[6] Reported as *Moorea producens*; manuscript includes a photo of woven *M. bouillonii*; 16S classification inconclusive; compound isolated along with known compounds previously isolated from *M. bouillonii*

Table 2.S2 - Average relative abundances and feature selection scores for top 10 Saipan MS[1] features

m/z	**relative rt**	**F-value[1]**	**p-value[1]**	**Am. Samoa[2]**	**China[2]**	**Guam[2]**	**India[2]**	**PNG[2]**	**Saipan[2]**
721.10	0.6065	0.869	0.5371	0.045	0.025	0	0	0.007	0.210
457.05[3]	**0.4852**	**17.402**	**0.0002**	**0.012**	**0.023**	**0.006**	**0**	**0.005**	**0.176**
1368.03	0.5654	5.049	0.0176	0.011	0.035	0.014	0.008	0.005	0.175
609.10	0.6263	1.229	0.3704	0.049	0.028	0.050	0.126	0.010	0.172
535.10	0.6177	2.073	0.1613	0.065	0.034	0.214	0.005	0	0.158
1367.06	0.5626	2.314	0.1296	0.008	0.027	0.030	0.069	0	0.136
459.06	0.5141	31.678	< 0.0001	0	0.024	0.006	0.005	0	0.112
536.10	0.6270	3.205	0.0617	0.094	0.014	0.034	0.006	0.006	0.102
1687.01	0.4942	1.086	0.4296	0.005	0.008	0	0.019	0.005	0.100
722.04	0.5938	0.703	0.6356	0.035	0.010	0.015	0.008	0.004	0.096

[1] F-values and p-values are generated in ORCA using the scikit learn (https://scikit-learn.org/stable/) implementation of univariate feature selection. These scores should be interpreted cautiously, as the dataset does not meet the assumptions necessary for univariate features selection but can still help in generating hypotheses about which features are driving differences between samples collected from different geographical regions.

[2] Average of the unit vector normalized integrated feature values for all samples from the geographical region.

[3] Compound **1** was detected in high abundance in samples from Saipan, ranking as the second most abundant MS[1] feature, while not being detected or being detected at very low levels in other samples.

Table 2.S3 - Putative identifications for the top 30 MS^1 features in the *M. bouillonii* crude extract dataset

	m/z	relative rt	max transformed integral	putative ids	differenc
1	815.10	0.6383	0.857928	['apratoxin G [M+H]+']	[0.62]
2	814.06	0.5968	0.765564	['apratoxin G [M+H]+', 'kakeromamide A [M+Na]+', 'kakeromamide B [M+Na]+']	[0.42, 0.66, 0.66]
3	721.10	0.6065	0.572637	['lyngbyapeptin A [M+Na]+']	[0.73]
4	840.07	0.6350	0.510516	['apratoxin A [M+H]+', 'bouilllonamide [M+Na]+']	[0.43, 0.42]
5	862.08	0.5894	0.508103	['apratoxin A [M+Na]+']	[0.39]
6	623.10	0.6501	0.440651	['None']	[0]
7	611.09	0.5845	0.427792	['alotamide A [M+Na]+']	[0.77]
8	678.10	0.5988	0.386879	['None']	[0]
9	535.10	0.6177	0.382478	['None']	[0]
10	609.10	0.6263	0.369167	['lyngbyabellin C [M+H]+', 'ulongamide F [M+H]+']	[0.01, 0.75]
11	378.06	0.5832	0.35347	['mandangolide [M+H]+']	[0.5]
12	625.11	0.6320	0.334295	['lyngbyabellin M [M+H]+']	[0.01]
13	793.11	0.6292	0.326725	['None']	[0]
14	827.10	0.6185	0.312621	['apratoxin B [M+H]+', 'apratoxin C [M+H]+', 'apratyramide [M+Na]+']	[0.62, 0.62, 0.32]
15	744.09	0.6160	0.31144	['lyngbyapeptin B [M+Na]+']	[0.25]
16	836.08	0.6534	0.286632	['apratoxin G [M+Na]+']	[0.38]
17	1368.02	0.5654	0.280465	['None']	[0]
18	639.08	0.6220	0.279184	['None']	[0]
19	581.10	0.6134	0.277933	['None']	[0]
20	632.10	0.6188	0.274745	['None']	[0]
21	722.04	0.5938	0.266215	['lyngbyapeptin B [M+H]+']	[0.32]
22	886.05	0.6135	0.258829	['lyngbyabellin J [M+Na]+']	[0.17]
23	841.13	0.6118	0.24873	['apratoxin A [M+H]+', 'bouilllonamide [M+Na]+']	[0.63, 0.64]
24	651.11	0.6804	0.246204	['ulongamide A [M+Na]+']	[0.81]
25	1687.01	0.4942	0.245016	['None']	[0]
26	**457.05**	**0.4852**	**0.242068**	**['None']**[1]	**[0]**
27	1367.06	0.5627	0.22625	['None']	[0]
28	816.09	0.6681	0.224017	['None']	[0]
29	494.07	0.6387	0.21453	['columbamide F [M+H]+', 'kanamienamide [M+H]+']	[0.21, 0.71]
30	863.09	0.6430	0.209889	['apratoxin A [M+Na]+']	[0.61]

[1] No putative identifications were assigned to compound **1**, suggesting it to be a new natural product.

Table 2.S4 - *In silico* antibiotic screening results for the doscadenamides and tetramic acids

compound name	SMILES	score[1]
doscadenamides A	COC1=CC(N(C(C(C)CCCCC#C)=O)C1CCCCNC(C(C)CCCCC#C)=O)=O	0.031352468
doscadenamides B	COC1=CC(N(C(C(C)CCCCC=C)=O)C1CCCCNC(C(C)CCCCC=C)=O)=O	0.056482284
doscadenamides C	COC1=CC(N(C(C(C)CCCCC#C)=O)C1CCCCNC(C(C)CCCCC=C)=O)=O	0.038773874
doscadenamides D	COC1=CC(N(C(C(C)CCCCC=C)=O)C1CCCCNC(C(C)CCCCC#C)=O)=O	0.038728081
doscadenamides E	COC1=CC(N(C(C(C)CCCCC#C)=O)C1CCCCNC(C(C)CCCCCC)=O)=O	0.062902918
doscadenamides F	COC1=CC(N(C(C(C)CCCCCC)=O)C1CCCCNC(C(C)CCCCC#C)=O)=O	0.0636456
doscadenamides G	COC1=CC(N(C(C(C)CCCCC=C)=O)C1CCCCNC(C(C)CCCCCC)=O)=O	0.091586996
doscadenamides H	COC1=CC(N(C(C(C)CCCCCC)=O)C1CCCCNC(C(C)CCCCC=C)=O)=O	0.092198204
doscadenamides I	COC1=CC(N(C(C(C)CCCCC#C)=O)C1CCCCNC(C(C)CCCCC(C)=O)=O)=O	0.049369105
doscadenamides J	COC1=CC(N(C(C(C)CCCCC(C)=O)=O)C1CCCCNC(C(C)CCCCC#C)=O)=O	0.04936365
C_{12}-tetramic acid	O=C(/C1=C(O)/CCCCCCCCC)NC(CCO)C1=O	0.545725947
C_{14}-tetramic acid	O=C(/C1=C(O)/CCCCCCCCCCC)NC(CCO)C1=O	0.590775286

[1] Scores represent the probability that the screened compound would inhibit *E. coli* growth at 50 μM (Stokes et al 2020).

Table 2.S5 - *M. bouillonii* crude extract sample metadata

filename	alias	extract number	collection code	collection country/territory	collection region	collection site	collection date
20170915_CBN_XSSCB2017_13.mzXML	China_13	-	XSSCB2017_13	China/Xisha	Sanshax	16° 51' 05.52", 112° 20' 56.13"	5/16/2017
20170915_CBN_XSSCB2017_24.mzXML	China_24	-	XSSCB2017_24	China/Xisha	Sanshax	16° 51' 05.52", 112° 20' 56.13"	5/19/2017
20170915_CBN_XSSCB2017_25.mzXML	China_25	-	XSSCB2017_25	China/Xisha	Sanshax	16° 51' 05.52", 112° 20' 56.13"	5/19/2017
2019-08-23_CBN_KHI-18-1.mzXML	India_KHI	-	KHT08APR18-3	India/Lakshadweep	Kavaratti	Heaven's Treat lagoon	4/8/2018
2019-08-23_CBN_KP-16-1.mzXML	India_KP	-	KP-16	India/Lakshadweep	Kavaratti	Paradise Hut lagoon	2/6/2016
2019-08-23_CBN_KPL-18-1.mzXML	India_KPL	-	KPL08APR18-1	India/Lakshadweep	Kavaratti	Paradise Hut lagoon	4/8/2018
2019-08-23_CBN_KSP-18-1.mzXML	India_KSP	-	KSP07APR18-1	India/Lakshadweep	Kavaratti	south of Paradise Hut pier	4/7/2018
2200.mzXML	Saipan_00	2200	SPB31JAN13-1	Saipan	-	Laulau Bay	1/31/2013
2209.mzXML	Saipan_09	2209	SPD29JAN13-6	Saipan	-	Laulau Bay	1/29/2013
2220.mzXML	AmSam_20	2220	ASA12JUL14-1	American Samoa	-	Afao	7/12/2014
2223.mzXML	AmSam_23	2223	ASG15JUL14-1	American Samoa	-	Fagasa Bay	7/15/2014
2232.mzXML	Saipan_32	2232	SPB01FEB13-1	Saipan	-	Laulau Bay	2/1/2013
2246.mzXML	Guam_46	2246	GBB21MAR16-1	Guam	-	Apra Harbor	3/21/2016
2247.mzXML	Guam_47	2247	GGG21MAR16-1	Guam	-	Apra Harbor	3/21/2016
Mb.mzXML	PNG_c	-	PNG19MAY05-8	Papua New Guinea	New Ireland	Pigeon Island	5/19/2005

Table 2.S6 - ORCA parameter set for MS^1 feature dendrogram

parameter	value
bin_width	0.5
bin_offset	0
bins_start	200
bins_end	2000
peak_consecutivity	0
peak_cluster_size_cutoff	3
min_integral	100000
rt_setting	‘relative’
rrt_tolerance	0.05
transforms	None
metric	‘cosine’
method	‘average’
color_cutoff	N/A (custom colorization)

Table 2.S7 - ORCA parameter set for GNPS MS^2 feature presence/absence dendrogram

parameter	value
drop_columns	[‘#OTU ID’]
drop_rows	[]
transpose_buckettable	False
transforms	presence_absence = True
metric	‘cosine’
method	‘average’
color_cutoff	N/A (custom colorization)

Table 2.S8 - ORCA parameter set for MS^1 feature selection

parameter	value
bin_width	1
bin_offset	0
bins_start	200
bins_end	2000
peak_consecutivity	0
peak_cluster_size_cutoff	3
min_integral	100000
rt_setting	‘relative’
rrt_tolerance	0.05
transforms	None
metric	‘cosine’
method	‘average’
color_cutoff	N/A (custom colorization)

Table 2.S9 - GNPS parameter set for *M. bouillonii* crude extract molecular network

parameter	value
workflow version	1.2.5
PAIRS_MIN_COSINE	0.7
ANALOG_SEARCH	1
tolerance.PM_tolerance	2.0
tolerance.Ion_tolerance	0.5
MIN_MATCHED_PEAKS	2
TOPK	10
CLUSTER_MIN_SIZE	1
MAXIMUM_COMPONENT_SIZE	100
MIN_PEAK_INT	50
FILTER_STDDEV_PEAK_INT	0.0
RUN_MSCLUSTER	on
FILTER_PRECURSOR_WINDOW	1
FILTER_LIBRARY	1
WINDOW_FILTER	1
SCORE_THRESHOLD	0.7
MIN_MATCHED_PEAKS_SEARCH	2
MAX_SHIFT_MASS	100.0

Table 2.S10 - GNPS parameter set for *M. bouillonii* crude extract MS^2 feature bucket table

parameter	**value**
workflow version	release_22
PAIRS_MIN_COSINE	0.7
ANALOG_SEARCH	1
tolerance.PM_tolerance	2.0
tolerance.Ion_tolerance	0.5
MIN_MATCHED_PEAKS	4
TOPK	10
CLUSTER_MIN_SIZE	1
MAXIMUM_COMPONENT_SIZE	100
MIN_PEAK_INT	0.0
FILTER_STDDEV_PEAK_INT	0.0
RUN_MSCLUSTER	on
FILTER_PRECURSOR_WINDOW	1
FILTER_LIBRARY	1
WINDOW_FILTER	1
SCORE_THRESHOLD	0.7
MIN_MATCHED_PEAKS_SEARCH	4
MAX_SHIFT_MASS	100.0

Table 2.S11 - GNPS parameter set for Saipan and Guam sample crudes and fractions molecular network

parameter	**value**
workflow version	release_17
PAIRS_MIN_COSINE	0.7
ANALOG_SEARCH	1
tolerance.PM_tolerance	1.0
tolerance.Ion_tolerance	0.5
MIN_MATCHED_PEAKS	4
TOPK	10
CLUSTER_MIN_SIZE	2
MAXIMUM_COMPONENT_SIZE	100
MIN_PEAK_INT	0.0
FILTER_STDDEV_PEAK_INT	0.0
RUN_MSCLUSTER	on
FILTER_PRECURSOR_WINDOW	1
FILTER_LIBRARY	1
WINDOW_FILTER	1
SCORE_THRESHOLD	0.7
MIN_MATCHED_PEAKS_SEARCH	4
MAX_SHIFT_MASS	100.0

Table 2.S12 - VLC fractionation solvent systems

fraction	**composition**
A	100% hexane
B	90% hexane : 10% ethyl acetate
C	80% hexane : 20% ethyl acetate
D	60% hexane : 40% ethyl acetate
E	40% hexane : 60% ethyl acetate
F	20% hexane : 80% ethyl acetate
G	100% ethyl acetate
H	75% ethyl acetate : 25% methanol
I	100% methanol

Table 2.S13 - ^{1}H and ^{13}C NMR Data for doscadenamide A (**1**) in CDCl3[a]

Residue	**Position**	**δ_C, type**	**δ_H, mult. (*J*, Hz)**
pyLys-OMe	1	170.0, C	
	2	94.2, CH	5.04, s
	3	179.2, C	
	4	59.2, CH	4.64, dd (*J* = 5.7, 2.9 Hz)
	5	29.0, CH$_2$	2.07, m 1.84, m
	6	20.4, CH$_2$	1.19, m 1.16, m
	7	29.6, CH$_2$	1.44, m
	8	39.3, CH$_2$	3.22, dp (*J* = 19.2, 6.4 Hz) 3.13, dp (*J* = 19.2, 6.4 Hz)
	9	58.9, CH$_3$	3.84, s
	NH		5.53, brs
Moya-1	10	177.0, C	
	11	41.8, CH	2.13, m
	12	33.9, CH$_2$	1.62, m 1.35, m
	13	28.5, CH$_2$	1.36, m
	14	28.5, CH$_2$	1.50, m
	15	18.5[b], CH$_2$	2.18, m
	16	84.7, C	
	17	68.52[c], CH	1.934[b], t (*J* = 2.7 Hz)
	18	18.1[d], CH$_3$	1.11, d (*J* = 1.1 Hz)
Moya-2	19	176.4, C	
	20	39.1, CH	3.75, m
	21	33.7, CH$_2$	1.75, m 1.42, m
	22	26.7, CH$_2$	1.39, m
	23	26.4, CH$_2$	1.44, m
	24	18.4[b], CH$_2$	2.18, m
	25	84.6, C	
	26	68.50[c], CH	1.940[b], t (*J* = 2.7 Hz)
	27	16.3[d], CH$_3$	1.12, d (*J* = 1.1 Hz)

[a] Data recorded at 600 MHz (^{1}H NMR) and 125 MHz (^{13}C NMR). [b,c,d] Assignments with the same superscripted letter could be reversed.

2.5: References

Anstett DN, Nunes KA, Baskett C, Kotanen PM (2016) Sources of Controversy Surrounding Latitudinal Patterns in Herbivory and Defense. Trends Ecol. & Evol 31:789–802. https://doi.org/10.1016/j.tree.2016.07.011

Bakus GJ, Green G (1974) Toxicity in sponges and holothurians: a geographic pattern. Science 185:951–953. https://doi.org/10.1126/science.185.4155.951

Bolser RC, Hay ME (1996) Are tropical plants better defended? Palatability and defenses of temperate vs tropical seaweeds. Ecology 77:2269–2286. https://doi.org/10.2307/2265730

Bory A, Shilling AJ, Allen J, Azhari A, Roth A, Shaw LN, Kyle DE, Adams JH, Amsler CD, McClintock JB, Baker BJ (2020) Bioactivity of Spongian Diterpenoid Scaffolds from the Antarctic Sponge Dendrilla antarctica. Mar Drugs 18:327. https://doi.org/10.3390/md18060327

Boudreau PD, Byrum T, Liu WT, Dorrestein PC, Gerwick WH (2012) Viequeamide A, a Cytotoxic Member of the Kulolide Superfamily of Cyclic Depsipeptides from a Marine Button Cyanobacterium. J Nat Prod 75:1560–1570. https://doi.org/10.1021/np300321b

Cai W, Salvador-Reyes LA, Zhang W, Chen QY, Matthew S, Ratnayake R, Seo SJ, Dolles S, Gibson DJ, Paul VJ, Luesch H (2018) Apratyramide, a Marine-Derived Peptidic Stimulator of VEGF-A and Other Growth Factors with Potential Application in Wound Healing. ACS Chem Biol 13: 91–99. https://doi.org/10.1021/acschembio.7b00827

Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak MY, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P (2012) A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 30:918–920. https://doi.org/10.1038/nbt.2377

Chanana S, Thomas CS, Braun DR, Hou Y, Wyche TP, Bugni TS (2017) Natural Product Discovery Using Planes of Principal Component Analysis in R (PoPCAR). Metabolites 7:34. https://doi.org/10.3390/metabo7030034

Choi H, Mascuch SJ, Villa FA, Byrum T, Teasdale ME, Smith JE, Preskitt LB, Rowley DC, Gerwick L, Gerwick WH (2012) Honaucins A-C, potent inhibitors of inflammation and bacterial quorum sensing: synthetic derivatives and structure-activity relationships. Chem Biol 19:589–98. https://doi.org/10.1016/j.chembiol.2012.03.014

Choi H, Mevers E, Byrum T, Valeriote FA, Gerwick WH (2012) Lyngbyabellins K-N from two Palmyra atoll collections of the marine cyanobacterium *Moorea bouillonii*. European J Org Chem 2012:5141–5150. https://doi.org/10.1002/ejoc.201200691

Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, Wishart DS, Xia J (2018) MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Res 46:W486–W494. https://doi.org/10.1093/nar/gky310

Clark CM, Costa MS, Sanchez LM, Murphy BT (2018) Coupling MALDI-TOF mass spectrometry protein and specialized metabolite analyses to rapidly discriminate bacterial function. Proc Natl Acad Sci USA 115:4981–4986. https://doi.org/10.1073/pnas.1801247115

Coley PD, Aide TM (1991) Comparison of herbivory and plant defenses in temperate and tropical broad-leaved forests. In: Price PW, Lewinsohn TM, Fernandes GW, Benson WW (eds.) Plant–Animal Interaction: Evolutionary Ecology in Tropical and Temperate Regions. Wiley-Interscience, New York, USA, pp. 25–49

Coley PD, Barone JA (1996) Herbivory and plant defenses in tropical forests. Annu Rev Ecol Syst 27:305–335. https://doi.org/10.1146/annurev.ecolsys.27.1.305

Crnkovic CM, May DS, Orjala J (2018) The impact of culture conditions on growth and metabolomic profiles of freshwater cyanobacteria. J Appl Phycol 30:375–384. https://doi.org/10.1007/s10811-017-1275-3

Edwards DJ, Marquez BL, Nogle LM, McPhail K, Goeger DE, Roberts MA, Gerwick WH (2004) Structure and Biosynthesis of the Jamaicamides, New Mixed Polyketide-Peptide Neurotoxins from the Marine Cyanobacterium *Lyngbya majuscula*. Chem Biol 11:817–833. https://doi.org/10.1016/j.chembiol.2004.03.030

Engene N, Rottacker EC, Kaštovský J, Byrum T, Choi H, Ellisman MH, Komárek J, Gerwick WH (2012) *Moorea producens* gen. nov., sp. nov. and *Moorea bouillonii* comb. nov., tropical marine cyanobacteria rich in bioactive secondary metabolites. Int J Syst Evol Microbiol 62:1171–1178. https://doi.org/10.1099/ijs.0.033761-0

Engene N, Tronholm A, Paul VJ (2018) Uncovering cryptic diversity of *Lyngbya* : the new tropical marine cyanobacterial genus *Dapis* (Oscillatoriales). J Phycol 54:435–446. https://doi.org/10.1111/jpy.12752

Gerwick WH, Tan LT, Sitachitta N (2001) Nitrogen-containing metabolites from marine cyanobacteria. In: Cordell GA (ed.) The Alkaloids: Chemistry and Biology, 1st ed. Academic Press, Cambridge, MA, USA. vol 57, pp. 75-184. https://doi.org/10.1016/S0099-9598(01)57003-0

Goloborodko AA, Levitsky LI, Ivanov MV, Gorshkov MV (2013) Pyteomics - a Python Framework for Exploratory Data Analysis and Rapid Software Prototyping in Proteomics. J Am Soc Mass Spectr 24:301–304. https://doi.org/10.1007/s13361-012-0516-6

Gowda H, Ivanisevic J, Johnson CH, Kurczy ME, Benton HP, Rinehart D, Nguyen T, Ray J, Kuehl J, Arevalo B, Westenskow PD, Wang J, Arkin AP, Deutschbauer AM, Patti GJ, Siuzdak G (2014) Interactive XCMS Online: Simplifying Advanced Metabolomic Data Processing and Subsequent Statistical Analyses. Anal Chem 86:6931–6939. https://doi.org/10.1021/ac500734c

Green LC, Wagner DA, Glogowski J, Skipper PL, Wishnok JS, Tannenbaum SR (1982) Analysis of nitrate, nitrite, and [N-15]-labeled nitrate in biological-fluids. Anal Biochem 126:131–138. https://doi.org/10.1016/0003-2697(82)90118-x

Gutiérrez M, Suyama TL, Engene N, Wingerd JS, Matainaho T, Gerwick WH (2008) Apratoxin D, a Potent Cytotoxic Cyclodepsipeptide from Papua New Guinea Collections of the Marine Cyanobacteria *Lyngbya majuscula* and *Lyngbya sordida*. J Nat Prod 71:1099–1103. https://doi.org/10.1021/np800121a

Hagberg AA, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux G, Vaught T, Millman J (eds.) Proceedings of the 7th Python in Science Conference (SciPy2008), Pasadena, CA, August 19-24, 2008. pp. 11–15

Hay ME, Fenical W (1988) Marine plant-herbivore interactions: the ecology of chemical defense. Annu Rev Ecol Syst 19:111–145.

Holman JD, Tabb DL, Mallick P (2014) Employing ProteoWizard to Convert Raw Mass Spectrometry Data. Curr Protoc Bioinformatics 46:1–9. https://doi.org/10.1002/0471250953.bi1324s46

Hooper GJ, Orjala J, Schatzman RC, Gerwick WH (1998) Carmabins A and B, New Lipopeptides from the Caribbean Cyanobacterium *Lyngbya majuscula*. J Nat Prod 61:529–533. https://doi.org/10.1021/np970443p

Hunter JD (2007) Matplotlib: A 2D graphics environment. Comput Sci Eng 9:90–95. https://doi.org/ 10.1109/MCSE.2007.55

Iwasaki A, Ohno O, Sumimoto S, Ogawa H, Nguyen KA, Suenaga K (2015) Jahanyne, an Apoptosis-Inducing Lipopeptide from the Marine Cyanobacterium *Lyngbya sp.* Org Lett 17:652–655. https://doi.org/10.1021/ol5036722

Jayaram B, Klawonn F (2012) Can unbounded distance measures mitigate the curse of dimensionality? International Journal of Data Mining, Modelling and Management 4:361–383. https://doi.org/10.1504/IJDMMM.2012.049883

Kleigrewe K, Almaliti J, Tian IY, Kinnel RB, Korobeynikov A, Monroe EA, Duggan BM, Di Marzo V, Sherman DH, Dorrestein PC, Gerwick L, Gerwick WH (2015) Combining Mass Spectrometric Metabolic Profiling with Genomic Analysis: A Powerful Approach for Discovering Natural Products from Cyanobacteria. J Nat Prod 78:1671–1682. https://doi.org/10.1021/acs.jnatprod.5b00301

Klein D, Braekman JC, Daloze D, Hoffmann L, Castillo G, Demoulin V (1999a) Lyngbyapeptin A, a modified tetrapeptide from *Lyngbya bouillonii* (Cyanophyceae). Tetrahedron Lett 40:695–696. https://doi.org/10.1016/S0040-4039(98)02451-4

Klein D, Braekman JC, Daloze D, Hoffmann L, Castillo G, Demoulin V (1999b) Madangolide and Laingolide A, Two Novel Macrolides from *Lyngbya bouillonii* (Cyanobacteria). J Nat Prod 62:934–936. https://doi.org/10.1021/np9900324

Klein D, Braekman JC, Daloze D, Hoffmann L, Demoulin V (1996) Laingolide, a novel 15-membered macrolide from *Lyngbya bouillonii* (cyanophyceae). Tetrahedron Lett 37:7519–7520. https://doi.org/10.1016/0040-4039(96)01728-5.

Klein D, Braekman JC, Daloze D, Hoffmann L, Demoulin V (1997) Lyngbyaloside, a Novel 2,3,4-Tri-O-methyl-6-deoxy-α-mannopyranoside Macrolide from *Lyngbya bouillonii* (Cyanobacteria). J Nat Prod 60:1057–1059. https://doi.org/10.1021/np9702751

Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick J, Grout J, Corlay S, Ivanov P, Avila D, Abdalla S, Willing C, Jupyter Development Team (2016) Jupyter Notebooks—a publishing format for reproducible computational workflows, In: Loizides F, Scmidt B (eds.) Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press, Amsterdam, The Netherlands. pp. 87–90. https://doi.org/10.3233/978-1-61499-649-1-87

Kooyers NJ, Blackman BK, Holeski LM (2017) Optimal defense theory explains deviations from latitudinal herbivory defense hypothesis. Ecology 98:1036–1048. https://doi.org/10.1002/ecy.1731

Leão T, Castelão G, Korobeynikov A, Monroe EA, Podell S, Glukhov E, Allen EE, Gerwick WH, Gerwick L (2017) Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus *Moorea*. Proc Natl Acad Sci USA 114:3198-3203. https://doi.org/10.1073/pnas.1618556114

Levin DA (1976) Alkaloid-bearing plants: an ecogeographic perspective. Am Nat 110:261–284

Levitsky LI, Klein J, Ivanov MV, Gorshkov MV (2019) Pyteomics 4.0: five years of development of a Python proteomics framework. J Proteome Res 18:709–714. https://doi.org/10.1021/acs.jproteome.8b00717

Liang X, Matthew S, Chen QY, Kwan JC, Paul VJ, Luesch H (2019) Discovery and Total Synthesis of Doscadenamide A: A Quorum Sensing Signaling Molecule from a Marine Cyanobacterium. Org Lett 21:7274–7278. https://doi.org/10.1021/acs.orglett.9b02525

Linington RG, Clark BR, Trimble EE, Almanza A, Ureña LD, Kyle DE, Gerwick WH (2009) Antimalarial Peptides from Marine Cyanobacteria: Isolation and Structural Elucidation of Gallinamide A. J Nat Prod 72:14–17. https://doi.org/10.1021/np8003529

Linington RG, González J, Ureña LD, Romero LI, Ortega-Barría E, Gerwick WH (2007) Venturamides A and B: Antimalarial Constituents of the Panamanian Marine Cyanobacterium *Oscillatoria sp*. J Nat Prod 70:397–401. https://doi.org/10.1021/np0605790

Lopez JAV, Petitbois JG, Vairappan CS, Umezawa T, Matsuda F, Okino T (2017) Columbamides D and E: Chlorinated Fatty Acid Amides from the Marine Cyanobacterium *Moorea bouillonii* Collected in Malaysia. Org Lett 19:4231–4234. https://doi.org/10.1021/acs.orglett.7b01869

Lowery CA, Park J, Gloeckner C, Meijler MM, Mueller RS, Boshoff HI, Ulrich RL, Barry III CE, Bartlett DH, Kravchenko VV, Kaufmann GF, Janda KD (2009) Defining the Mode of Action of Tetramic Acid Antibacterials Derived from Pseudomonas aeruginosa Quorum Sensing Signals. J Am Chem Soc 131:14473–14479. https://doi.org/10.1021/ja9056079

Luesch H, Yoshida WY, Moore RE, Paul VJ (1999) Lyngbyastatin 2 and Norlyngbyastatin 2, Analogues of Dolastatin G and Nordolastatin G from the Marine Cyanobacterium *Lyngbya majuscula*. J Nat Prod 62:1702–1706. https://doi.org/10.1021/np990310z

Luesch H, Yoshida WY, Moore RE, Paul VJ (2000a) Apramides A−G, Novel Lipopeptides from the Marine Cyanobacterium *Lyngbya majuscula*. J Nat Prod 63:1106–1112. https://doi.org/10.1021/np000078t

Luesch H, Yoshida WY, Moore RE, Paul VJ (2000b) Isolation and Structure of the Cytotoxin Lyngbyabellin B and Absolute Configuration of Lyngbyapeptin A from the Marine Cyanobacterium *Lyngbya majuscula*. J Nat Prod 63:1437–1439. https://doi.org/10.1021/np000104n

Luesch H, Yoshida WY, Moore RE, Paul VJ, Mooberry SL (2000c) Isolation, Structure Determination, and Biological Activity of Lyngbyabellin A from the Marine Cyanobacterium *Lyngbya majuscula*. J Nat Prod 63:611–615. https://doi.org/10.1021/np990543q

Luesch H, Yoshida WY, Moore RE, Paul VJ, Corbett TH (2001) Total Structure Determination of Apratoxin A, a Potent Novel Cytotoxin from the Marine Cyanobacterium *Lyngbya majuscula*. J Am Chem Soc 123:5418–5423. https://doi.org/10.1021/ja010453j

Luesch H, Yoshida WY, Harrigan GG, Doom JP, Moore RE, Paul VJ (2002a) Lyngbyaloside B, a New Glycoside Macrolide from a Palauan Marine Cyanobacterium, *Lyngbya* sp. J Nat Prod 65:1945–1948. https://doi.org/10.1021/np0202879

Luesch H, Williams PG, Yoshida WY, Moore RE, Paul VJ (2002b) Ulongamides A−F, New β-Amino Acid-Containing Cyclodepsipeptides from Palauan Collections of the Marine Cyanobacterium *Lyngbya* sp. J Nat Prod 65:996–1000. https://doi.org/10.1021/np0200461

Luesch H, Yoshida WY, Moore RE, Paul VJ (2002c) New apratoxins of marine cyanobacterial origin from guam and palau. Bioorg Med Chem 10:1973–1978. https://doi.org/10.1016/S0968-0896(02)00014-7

Luesch H, Yoshida WY, Moore RE, Paul VJ (2002d) Structurally diverse new alkaloids from Palauan collections of the apratoxin-producing marine cyanobacterium *Lyngbya sp.* Tetrahedron 58:7959–7966. https://doi.org/10.1016/S0040-4020(02)00895-5

Luzzatto-Knaan T, Garg N, Wang M, Glukhov E, Peng Y, Ackermann G, Amir A, Duggan BM, Ryazanov S, Gerwick L, Knight R, Alexandrov T, Bandeira N, Gerwick WH, Dorrestein PC (2017) Digitizing mass spectrometry data to explore the chemical diversity and distribution of marine cyanobacteria and algae. eLife 6:e24214. https://doi.org/10.7554/eLife.24214.001

Marcolefas E, Leung T, Okshevsky M, McKay G, Hignett E, Hamel J, Aguirre G, Blenner-Hassett O, Boyle B, Lévesque RC, Nguyen D, Gruenheid S, Whyte1 L (2019) Culture-Dependent Bioprospecting of Bacterial Isolates From the Canadian High Arctic Displaying Antibacterial Activity. Front Microbiol 10:1836. https://doi.org/10.3389/fmicb.2019.01836

Matthew S, Salvador LA, Schupp PJ, Paul VJ, Luesch H (2010) Cytotoxic Halogenated Macrolides and Modified Peptides from the Apratoxin-Producing Marine Cyanobacterium *Lyngbya bouillonii* from Guam. J Nat Prod 73:1544–1552. https://doi.org/10.1021/np1004032

Matthew S, Schupp PJ, Luesch H (2008) Apratoxin E, a Cytotoxic Peptolide from a Guamanian Collection of the Marine Cyanobacterium *Lyngbya bouillonii*. J Nat Prod 71:1113–1116. https://doi.org/10.1021/np700717s

McKinney W (2010) Data Structures for Statistical Computing in Python. In: van der Walt S, Millman J (eds.) Proceedings of the 9th Python in Science Conference (SciPy2010), Austin, TX, USA, June 28 - July 3, 2010. pp. 51–56. https://doi.org/10.25080/Majora-92bf1922-00a

Mehjabin JJ, Wei L, Petitbois JG, Umezawa T, Matsuda F, Vairappan CS, Morikawa M, Okino T (2020) Biosurfactants from Marine Cyanobacteria Collected in Sabah, Malaysia. J Nat Prod 83:1925–1930. https://doi.org/10.1021/acs.jnatprod.0c00164

Mevers E, Matainaho T, Allara M, Di Marzo V, Gerwick WH (2014) Mooreamide A: A cannabinomimetic lipid from the marine cyanobacterium *Moorea bouillonii*. Lipids 49:1127–1132. https://doi.org/10.1007/s11745-014-3949-9

Milligan KE, Márquez B, Williamson RT, Davies-Coleman M, Gerwick WH (2000) Two New Malyngamides from a Madagascan *Lyngbya majuscula*. J Nat Prod 63:965–968. https://doi.org/10.1021/np000038p

Moss NA, Leão T, Glukhov E, Gerwick L, Gerwick WH (2018) Collection, Culturing, and Genome Analyses of Tropical Marine Filamentous Benthic Cyanobacteria. In: Tawfik DS (ed.) Methods in Enzymology, 1st ed. Academic Press, Cambridge, MA, USA. vol 604, pp. 3–43. https://doi.org/10.1016/bs.mie.2018.02.014

Moss NA, Seiler G, Leão TF, Castro-Falcón G, Gerwick L, Hughes CC, Gerwick WH (2019) Nature's Combinatorial Biosynthesis Produces Vatiamides A–F. Angew Chem Int Ed 58:9027–9031. https://doi.org/10.1002/anie.201902571

Mwaskom/seaborn, v0.9.0 (2018) https://doi.org/10.5281/zenodo.1313201

Nakamura F, Maejima H, Kawamura M, Arai D, Okino T, Zhao M, Ye T, Lee J, Chang YT, Fusetani N, Nakao Y (2018) Kakeromamide A, a new cyclic pentapeptide inducing astrocyte differentiation isolated from the marine cyanobacterium *Moorea bouillonii*. Bioorg Med Chem Lett 28:2206–2209. https://doi.org/10.1016/j.bmcl.2018.04.067

Naman CB, Rattan R, Nikoulina SE, Lee J, Miller BW, Moss NA, Armstrong L, Boudreau PD, Debonsi HM, Valeriote FA, Dorrestein PC, Gerwick WH (2017) Integrating molecular networking and biological assays to target the isolation of a cytotoxic cyclic octapeptide, samoamide A, from an American Samoan marine cyanobacterium. J Nat Prod 80:625–633. https://doi.org/10.1021/acs.jnatprod.6b00907

Negishi E, Tan Z, Liang B, Novak T (2004) An efficient and general route to reduced polypropionates via Zr-catalyzed asymmetric C-C bond formation. Proc Natl Acad Sci USA 101:5782–5787. https://doi.org/10.1073/pnas.0307514101

Nothias LF, Petras D, Schmid R, Dührkop K, Rainer J, Sarvepalli A, Protsyuk I, Ernst M, Tsugawa H, Fleischauer M, Aicheler F, Aksenov AA, Alka O, Allard PM, Barsch A, Cachet X, Caraballo-Rodriguez AM, Da Silva RR, Dang T, Garg N, Gauglitz JM, Gurevich A, Isaac G, Jarmusch AK, Kameník Z, Kang KB, Kessler N, Koester I, Korf

A, Le Gouellec A, Ludwig M, Martin H C, McCall LI, McSayles J, Meyer SW, Mohimani H, Morsy M, Moyne O, Neumann S, Neuweger H, Nguyen NH, Nothias-Esposito M, Paolini J, Phelan VV, Pluskal T, Quinn RA, Rogers S, Shrestha B, Tripathi A, van der Hooft JJJ, Vargas F, Weldon KC, Witting M, Yang H, Zhang Z, Zubeil F, Kohlbacher O, Böcker S, Alexandrov T, Bandeira N, Wang M, Dorrestein PC (2020) Feature-based Molecular Networking in the GNPS Analysis Environment. Nat Methods 17:905–908. https://doi.org/10.1038/s41592-020-0933-6

Oliphant TE (2006) A guide to NumPy, 2nd ed. Trelgol Publishing, USA.

Pallarés N, Tolosa J, Mañes J, Ferrer E (2019) Occurrence of Mycotoxins in Botanical Dietary Supplement Infusion Beverages. J Nat Prod 82:403–406. https://doi.org/10.1021/acs.jnatprod.8b00283

Pandas-dev/pandas, Version v0.25.2 (2018) Zenodo. https://doi.org/10.5281/zenodo.3509135

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine Learning in Python. J of Mach Learn Res 12:2825–2830.

Pereira A, Etzbach L, Engene N, Müller R, Gerwick WH (2011) Molluscicidal Metabolites from an Assemblage of Palmyra Atoll Cyanobacteria. J Nat Prod 74:1175–1181. https://doi.org/10.1021/np200106b

Pereira AR, McCue CF, Gerwick WH (2010) Cyanolide A, a Glycosidic Macrolide with Potent Molluscicidal Activity from the Papua New Guinea Cyanobacterium *Lyngbya bouillonii*. J Nat Prod 73:217–220. https://doi.org/10.1021/np9008128

Pérez F, Granger BE (2007) IPython: A System for Interactive Scientific Computing, Comput Sci Eng 9:21–29. https://doi.org/10.1109/MCSE.2007.53

Pluskal T, Castillo S, Villar-Briones A, Orešič M (2010) MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11:395. https://doi.org/10.1186/1471-2105-11-395

Rasmann S, Agrawal AA (2011) Latitudinal patterns in plant defense: evolution of cardenolides, their toxicity and induction following herbivory. Ecol Lett 14:476–483. https://doi.org/10.1111/j.1461-0248.2011.01609.x

Ricciotti E, FitzGerald GA (2011) Prostaglandins and Inflammation. Arterioscler Thromb Vasc Biol 31:986–1000. https://doi.org/10.1161/ATVBAHA.110.207449

Rubio BK, Parrish SM, Yoshida W, Schupp PJ, Schils T, Williams PG (2010) Depsipeptides from a Guamanian marine cyanobacterium, *Lyngbya bouillonii*, with selective inhibition of serine proteases. Tetrahedron Lett 51:6718–6721. https://doi.org/10.1016/j.tetlet.2010.10.062

Sadar MD, Williams DE, Mawji NR, Patrick BO, Wikanta T, Chasanah E, Irianto HE, Van Soest R, Andersen RJ (2008) Sintokamides A to E, Chlorinated Peptides from the Sponge Dysidea sp. that Inhibit Transactivation of the N-Terminus of the Androgen Receptor in Prostate Cancer Cells. Org Lett 10:4947–4950. https://doi.org/10.1021/ol802021w

Shang Z, Winter J.M, Kauffman CA, Yang I, Fenical W (2019) Salinipeptins: Integrated Genomic and Chemical Approaches Reveal D-Amino Acid-Containing Ribosomally Synthesized and Post-Translationally Modified Peptides from a Great Salt Lake *Streptomyces sp.* ACS Chem Biol 14:415–425. https://doi.org/10.1021/acschembio.8b01058

Shannon P, Markiel1 A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski1 B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–504. https://doi.org/10.1101/gr.1239303

Soria-Mercado IE, Pereira A, Cao Z, Murray TF, Gerwick WH (2009) Alotamide A, a novel neuropharmacological agent from the marine cyanobacterium *Lyngbya bouillonii*. Org Lett 11:4704–4707. https://doi.org/10.1021/ol901438b

Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, MacNair CR, French S, Carfrae LA, Bloom-Ackermann Z, Tran VM, Chiappino-Pepe A, Badran AH, Andrews IW, Chory EJ, Church GM, Brown ED, Jaakkola TS, Barzilay R, Collins JJ (2020) A Deep Learning Approach to Antibiotic Discovery. Cell 181:475–483. https://doi.org/10.1016/j.cell.2020.01.021

Sumimoto S, Iwasaki A, Ohno O, Sueyoshi K, Teruya T, Suenaga K (2016) Kanamienamide, an Enamide with an Enol Ether from the Marine Cyanobacterium *Moorea bouillonii*. Org Lett 18:4884–4887. https://doi.org/10.1021/acs.orglett.6b02364

Sweeney-Jones AM, Gagaring K, Antonova-Koch J, Zhou H, Mojib N, Soapi K, Skolnick J, McNamara CW, Kubanek J (2020) Antimalarial Peptide and Polyketide Natural Products from the Fijian Marine Cyanobacterium *Moorea producens*. Mar Drugs 18:167. https://doi.org/10.3390/md18030167

Tan LT, Márquez BL, Gerwick WH (2002) Lyngbouilloside, a Novel Glycosidic Macrolide from the Marine Cyanobacterium *Lyngbya bouillonii*. J Nat Prod 65:925–928. https://doi.org/10.1021/np010526c

Tan LT, Okino T, Gerwick WH (2013) Bouillonamide: A Mixed Polyketide-Peptide Cytotoxin from the Marine Cyanobacterium *Moorea bouillonii*. Mar Drugs 11:3015–3024. https://doi.org/10.3390/md11083015

Tao Y, Li P, Zhang D, Glukhov E, Gerwick L, Zhang C, Murray TF, Gerwick WH (2018) Samholides, Swinholide-Related Metabolites from a Marine Cyanobacterium cf. *Phormidium* sp. J Org Chem 83:3034–3046. https://doi.org/10.1021/acs.joc.8b00028

Thornburg CC, Cowley ES, Sikorska J, Shaala LA, Ishmael JE, Youssef DTA, McPhail KL (2013) Apratoxin H and Apratoxin A Sulfoxide from the Red Sea Cyanobacterium *Moorea producens*. J Nat Prod 76:1781–1788. https://doi.org/10.1021/np4004992

Tidgewell K, Clark BR, Gerwick WH (2010) The Natural Products Chemistry of Cyanobacteria. In: Mander LN, Liu HW (eds.) Comprehensive Natural Products II: Chemistry and Biology, 1st ed. Elsevier: Amsterdam, The Netherlands. vol 2, pp. 144–188. https://doi.org/10.1016/B978-008045382-8.00041-1

Tidgewell K, Engene N, Byrum T, Media J, Doi T, Valeriote FA, Gerwick WH (2010) Evolved Diversification of a Modular Natural Product Pathway: Apratoxins F and G, Two Cytotoxic Cyclic Depsipeptides from a Palmyra Collection of *Lyngbya bouillonii*. Chembiochem. 11:1458–1466. https://doi.org/10.1002/cbic.201000070

Tronholm A, Engene N (2019) *Moorena gen. nov.*, a valid name for "*Moorea* Engene & *al*." *nom. inval.* (*Oscillatoriaceae, Cyanobacteria*). Notulae algarum 122:1–2

van der Walt S, Colbert SC, Varoquaux G (2011) The NumPy Array: A Structure for Efficient Numerical Computation. Comput Sci Eng 13:22–30, https://doi.org/10.1109/MCSE.2011.37

Van Rossum G, Drake Jr. FL (1995) Python reference manual, Release 2.0.1. PythonLabs, Amsterdam, The Netherlands. https://docs.python.org/2.0/ref/ref.html

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat I, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, SciPy 1.0 Contributors (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2

Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu WT, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu CC, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw CC, Yang YL, Humpf HU, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, Boya P CA, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodríguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard PM, Phapale P, Nothias LF, Alexandrov T, Litaudon M, Wolfender JL, Kyle JE, Metz TO, Peryea T, Nguyen DT, VanLeer D, Shinn P, Jadhav A, Müller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BØ, Pogliano K, Linington RG, Gutiérrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34:828–837. https://doi.org/10.1038/nbt.3597

Williams PG, Luesch H, Yoshida WY, Moore RE, Paul VJ (2003) Continuing Studies on the Cyanobacterium *Lyngbya sp.*: Isolation and Structure Determination of 15-Norlyngbyapeptin A and Lyngbyabellin D. J Nat Prod 66:595–598. https://doi.org/10.1021/np030011g

Yoshimura A, Kishimoto S, Nishimura S, Otsuka S, Sakai Y, Hattori A, Kakeya H (2014) Prediction and Determination of the Stereochemistry of the 1,3,5-Trimethyl-Substituted Alkyl Chain in Verucopeptin, a Microbial Metabolite. J Org Chem 79:6858–6867. https://doi.org/10.1021/jo500906v

Zhang C, Idelbayev Y, Roberts N, Tao Y, Nannapaneni Y, Duggan BM, Min J, Lin EC, Gerwick EC, Cottrell GW, Gerwick WH (2017) Small Molecule Accurate Recognition Technology (SMART) to Enhance Natural Products Research. Sci Rep 7:14243. https://doi.org/10.1038/s41598-017-13923-x

Zhang S, Gui C, Shao M, Kumar PS, Huang H, Ju J (2020) Antimicrobial tunicamycin derivatives from the deep sea-derived Streptomyces xinghaiensis SCSIO S15077, Nat Prod Res 34:1499–1504. https://doi.org/10.1080/14786419.2018.1493736

Zhou H, He Y, Tian Y, Cong B, Yang, H (2019) Bacilohydrin A, a New Cytotoxic Cyclic Lipopeptide of Surfactins Class Produced by *Bacillus sp.* SY27F from the Indian Ocean Hydrothermal Vent. Nat Prod Commun 14:141–146. https://doi.org/10.1177/1934578X1901400137

Chapter 3: Induction versus Selection: Examining Alternative Mechanisms Driving *Alpheus frontalis* Association and Chemical Composition in *Moorena bouillonii*

3.0: Abstract

Alpheus frontalis shrimp weave and inhabit specific chemotypes of the cyanobacterium *Moorena bouillonii*, while not associating with other similar chemotypes. This pattern inspired competing hypotheses: shrimp induce changes in chemical composition by interacting with the cyanobacterium, or shrimp select a preferred chemotype to associate with. Investigations into these hypotheses yielded no evidence of shrimp inducing changes in cyanobacterial chemical composition and revealed that shrimp select their preferred chemotype of *M. bouillonii* via a waterborne cue. Further behavioral assays suggested that the waterborne cue was lipophilic and could be captured from seawater with C18 solid phase extraction and hexane elution. Comparative gas chromatography-mass spectrometry (GC-MS) allowed for the identification of a lead compound for the shrimp attractant waterborne chemical cue.

3.1: Introduction

Chemical substances are known in a number of cases to mediate interactions between or within marine species. Seabirds track dimethyl sulfide (DMS) odor plumes resulting from phytoplankton release of dimethylsulfoniopropionate (DMSP) when they are fed upon by zooplankton; this facilitates the ability of seabirds to locate productive ocean waters (Nevitt 2008). Larvae of the tube worm *Phragmatopoma californica* are induced to settle when exposed to unsaturated fatty acids found in the sandy tubes of conspecific adults, resulting in large localized aggregations of tube worms along California's rocky shore (Pawlik 1986). Similarly, the barnacle *Balanus glandula* produces a glycoprotein in its cuticle, which triggers conspecific larval settlement upon contact (Zimmer et al 2016). Aptly named 'MULTIFUNCin,' this glycoprotein is also relied upon by the predatory whelk *Acanthinucella spirata* for identifying live barnacles to feast upon (Zimmer et al 2016). Mark Hay masterfully articulated in the title of his 2009 review that "chemical signals and cues structure marine populations, communities, and ecosystems" (Hay 2009).

Chemical substances can play roles at multiple levels in biological interactions. The coral *Acropora nasuta* releases an attractive cue to direct symbiotic goby fishes to remove the toxic alga *Chlorodesmis fastigiata* when it encroaches on its space (Dixson and Hay 2012). The coral was triggered to release its goby attractant via contact with *C. fastigiata* as well as via contact with an algal mimic that had been treated with *C. fastigiata* extract, suggesting a chemical contact cue from the algae triggers the coral's chemical response. The gobies only respond when *A. nasuta*, their host *Acropora* sp., is contacted by the algae, and not when the congener *A. millepora* is contacted, showing specificity in the cue's production. Furthermore, one of the species of goby, *Gobiodon histrio*, eats the removed *C. fastigiata*; doing so may allow *G. histrio* to produce skin secretions that are more effectively noxious as a chemical defense. By working to understand the vast complexity of invisible chemical signals and modulators that are swirling around the marine environment, the mechanisms underlying biological interactions can be elucidated (Hay 2009). Moreover, understanding of these interactions, and how they are mediated by chemical substances, has the potential to inform the development of new targeted, ecologically relevant marine ecosystem management practices (Saha et al 2019).

One category of interactions in marine ecosystems known to have a chemical component is the location and identification of a preferred host or host environment. A classic example is the settlement of coral larvae driven by tetrabromopyrrole produced by bacterial biofilms (Sneed et al 2014); an alternative hypothesis is that compounds termed ‘morphogens’ produced by crustose coralline algae (with some less significant contributions by biofilm microbes) are the responsible chemical party (Gomez-Lemos et al 2018). Regardless, the scientific literature indicates coral larval settlement to be non-random and driven by chemical cues in the environment. For non-sessile organisms, such as sea hares of the *Aplysia* genus, host chemical cues may not be as important for finding a specific preferred site for settlement and metamorphosis, but may instead be used post-metamorphosis to locate preferred algal species for feeding and association (Pawlik 1989; Nocchi et al 2017). Both anemone fish (Elliot et al 1995; Scott and Dixson 2016) and anemone shrimp (Guo et al 1996) have been documented to make use of host chemical cues to locate potential host anemones and to select preferred species of host anemone over other species. Marine organisms also rely on chemicals in determining which potential hosts should be avoided. The symbiotic pea crab *Dissodactylus crinitichelis* is not significantly attracted to chemical cues released by its sand dollar host *Encope emarginata*, but rather, is only significantly attracted when exposed to host and conspecific cues in combination (Souza et al 2019). Notably, crabs were deterred by sand dollar conditioned water, and so avoided potential hosts, when the sand dollars were exposed to stressful conditions. Similarly, anemone fish have been shown in a flume environment to prefer the chemical cues of host anemones with intact *Symbiodinium* spp. symbionts to those of bleached potential hosts (Scott and Dixson 2016).

Another chemobiological interaction of note is the induction of chemotypic changes in response to a biological or chemical exposure. The previously described release of goby-attractant chemistry by *A. nasuta* in response to contact from *C. fastigiata* (Dixson and Hay 2012) is an example of a chemical response induced by a competitor, but chemical defense induction is very often associated with herbivory. Feeding by the amphipod *Ampithoe longimana* on the brown alga *Dictyota menstrualis* resulted in increased concentrations of the known feeding deterrent natural products dictyol E and pachydictyol A, and their analog dictyodial (Cronin and Hay 1996). In the case of the red alga *Laurencia dendroidea,* decreases in palatability and changes in sesquiterpene production were induced not only by grazing by the sea hare *Aplysia brasiliana* as well as simulated grazing, but also via exposure to chemical cues released into the water column by other grazed *L. dendroidea* thalli (Pereira et al 2020). For the dinoflagellate *Alexandrium minutum* and the diatom *Pseudo-nitzschia seriata*, grazing is not even necessary for the inducement of increased concentrations of paralytic shellfish toxins and domoic acid, respectively (Bergkvist et al 2008; Selander et al 2019). Toxin production in these phytoplankton is instead induced via exposure to the copapodamides, a family of metabolites exuded by predatory copepods (Selander et al 2015); various species of copepods release differing ratios of copapodamide species, resulting in differential toxin production in *A. minutum* when exposed to different predators (Bergkvist et al 2008).

Crustaceans are well-known for engaging in biological interactions that involve coopting the chemical defenses of other marine organisms, particularly eukaryotic and prokaryotic algae. Juvenile decorator crabs of the species *Libinia dubia* selectively decorate their carapaces with *D. menstrualis* based on the presence of the secondary metabolite dictyol E, in order to deter predation (Stachowicz and Hay 1999). Similarly, the amphipod *Pseudamphithoides incurvaria* collects the brown alga *Dictyota bartayresii* and uses it to build a shelter; *P. incurvaria* can be compelled to build its shelter out of other algal biomass when that biomass is coated in the *Dictyota* secondary metabolite pachydictyol A, with more pachydictyol A proportionally increasing shelter building (Hay, Duffy, and Fenical 1990). Isopods of the *Santia* genus grow biofilms of unicellular cyanobacteria and other microbes on their carapaces, providing the dual benefits of being an always accessible source of food and a mobile source of predator-deterring chemical protection (Lindquist et al 2005).

The association between *Alpheus frontalis* H. Milne Edwards 1837 (Alpheidae) and *Moorena bouillonii* (L.Hoffmann & Demoulin) Engene & Tronholm 2019 (Oscillatoriaceae) appears to be yet another example of a crustacean leveraging the natural products of an alga for its own protective benefit, an arrangement known as an associational defense (Bergvall et al 2006). *A. frontalis* is considered to be an obligate tube builder, using its woven tubes of *M. bouillonii* as a source of both food and shelter (Banner and Banner 1982; Fishelson 1966). The association is highly specific; in field surveys, *A. frontalis* is found only associated with *M. bouillonii,* and habitation of *M. bouillonii* is dominated by, if not solely restricted to, *A. frontalis* (Cruz-Rivera and Paul 2002, 2006). When *M. bouillonii* is not available under laboratory conditions, *A. frontalis* will weave its shelter out of other available algal filaments but will immediately incorporate *M. bouillonii* into this structure when access to it is returned (Fishelson 1966).

At Finger Reef, in Apra Harbor on the island of Guam, the curious observation was made that *M. bouillonii* can be found in close proximity existing in two different forms: one which is associated with *A. frontalis* and possessing a particular suite of secondary metabolites, and another that grows without the shrimp and possesses a closely related yet distinct chemical composition (Matthew, Schupp, and Luesch 2008). This pattern of association inspired the generation of two competing hypotheses: 1) *A. frontalis* interaction with *M. bouillonii* induces a change in the secondary metabolome, or 2) *A. frontalis* selects *M. bouillonii* filaments that produce a preferred composition of natural products. In the thesis chapter that follows, these two hypotheses were tested with the goals of explaining the observed pattern in *A. frontalis* – *M. bouillonii* association.

3.2: Results & Discussion

3.2.1: Pattern of shrimp association and chemical composition

Two initial collections were made in close proximity to one another in Apra Harbor, Guam. One sample of *M. bouillonii* was found to be associated with *A. frontalis* whereas the second sample was not. Both samples were extracted targeting nonpolar to midpolar chemical constituents, and their crude extracts were profiled via LC-MS/MS. The resultant chromatograms both contained peaks consistent with the known metabolites apratoxin A, lyngbyastatin 2, and apratoxin E, but with varied relative abundances. Apratoxin A was more prevalent in the sample of shrimp-associated *M. bouillonii* (Figure 3.S1), while the peaks putatively representing lyngbyastatin 2 and apratoxin E were in higher relative abundance in the sample found not in association with *A. frontalis* (Figures 3.S2, 3.S3). This is consistent with previous reports (Matthew, Schupp, and Luesch 2008).

The apparent differential production of apratoxin A, lyngbyastatin 2, and apratoxin E exemplify how the shrimp-associated and the non-shrimp-associated *M. bouillonii* samples differ in their chemical compositions. Specifically, both samples were found to possess many of the same molecular features, but at differing relative abundances. Of the 799 features detected for the shrimp-associated cyanobacterium and the 716 features detected for the non-shrimp-associated cyanobacterium, 271 of the features were shared by both. This significant overlap in chemical compositions, albeit with varied relative abundances, results in these two samples being more chemically similar to one another than to *M. bouillonii* samples collected from other geographic locations, irrespective of their shrimp associations (Leber et al 2020). This observation further adds to the curiosity surrounding the distributional patterns of *A. frontalis*, and how it becomes associated with one cyanobacterial chemotype but not with a very similar cyanobacterial chemotype.

It must be considered whether these two chemotypes of *M. bouillonii* are truly two types of the same species with different chemical compositions and ecological functionalities, or if they may represent two different operational taxonomic units. The non-shrimp-associated cyanobacterium was originally reported as *Lyngbya bouillonii*, based on its microscopic morphological characteristics (Matthew, Schupp, and Luesch 2008). This taxonomic assignment is further supported by partial 16S rRNA gene sequence data, which fails to clade shrimp-associated *M. bouillonii* separate from non-shrimp-associated *M. bouillonii* (See Chapter 1: Figure 1.S1). However, 16S rRNA sequencing is known to lack the phylogenetic resolution necessary to differentiate *M. bouillonii* from its congener *Moorena producens* (Engene et al 2012), and so is insufficient in definitively determining shrimp-associated and non-shrimp-associated *M. bouillonii* to be the same or different taxonomic units. Due to their high similarity in terms of filament and cell size and shape, 16S rRNA gene sequences, and distinctive chemical composition, they will be referred to herein as chemotypes of the species *M. bouillonii*, with the acknowledgement that additional scientific inquiry is required for improved understanding of how these cyanobacterial types are related.

3.2.2: Testing the induction hypothesis

We first sought to interrogate the hypothesis that natural product induction was the mechanism by which the observed pattern in *A. frontalis* association and *M. bouillonii* chemical composition is established. *A. frontalis* and *M. bouillonii* growing with (shrimp-associated cyanobacterium, or shrimp cyano) and without *A. frontalis* (non-shrimp-associated cyanobacterium, or non-shrimp cyano) were collected and used to assemble four different growth conditions: shrimp-associated cyanobacterium with a shrimp, shrimp-associated cyanobacterium without a shrimp, non-shrimp-associated cyanobacterium with a shrimp, and non-shrimp-associated cyanobacterium without a shrimp (Figure 3.1). Five replicates of each of these four growth conditions were placed in containers, covered with a mesh screen, and submerged in a blocked arrangement in a flow-through, raw seawater table. Submersion allowed for water exchange and the deposition of materials for shrimp to scavenge in each container, while the screen prevented shrimp escape. The experiment lasted 15 days, after which cyanobacterial biomass was preserved for extraction and analysis via LC-MS/MS. Hierarchical clustering of the resultant metabolomics data via the ORCA pipeline (Leber et al 2020) revealed that regardless of whether these two chemotypes of cyanobacteria were grown with or without shrimp, their chemical compositions reflected their original field collection profiles (Figure 3.2). Whereas there is evident structure in the dendrogram resulting from hierarchical clustering that suggests samples from the same cyanobacterial strain are more closely related in their chemical composition, samples cultured with or without a shrimp are intermixed for a given chemical strain, and do not alter between the two clusters. Therefore, there was no evidence that interactions with shrimp triggered an induction or any other modulation of natural products composition.

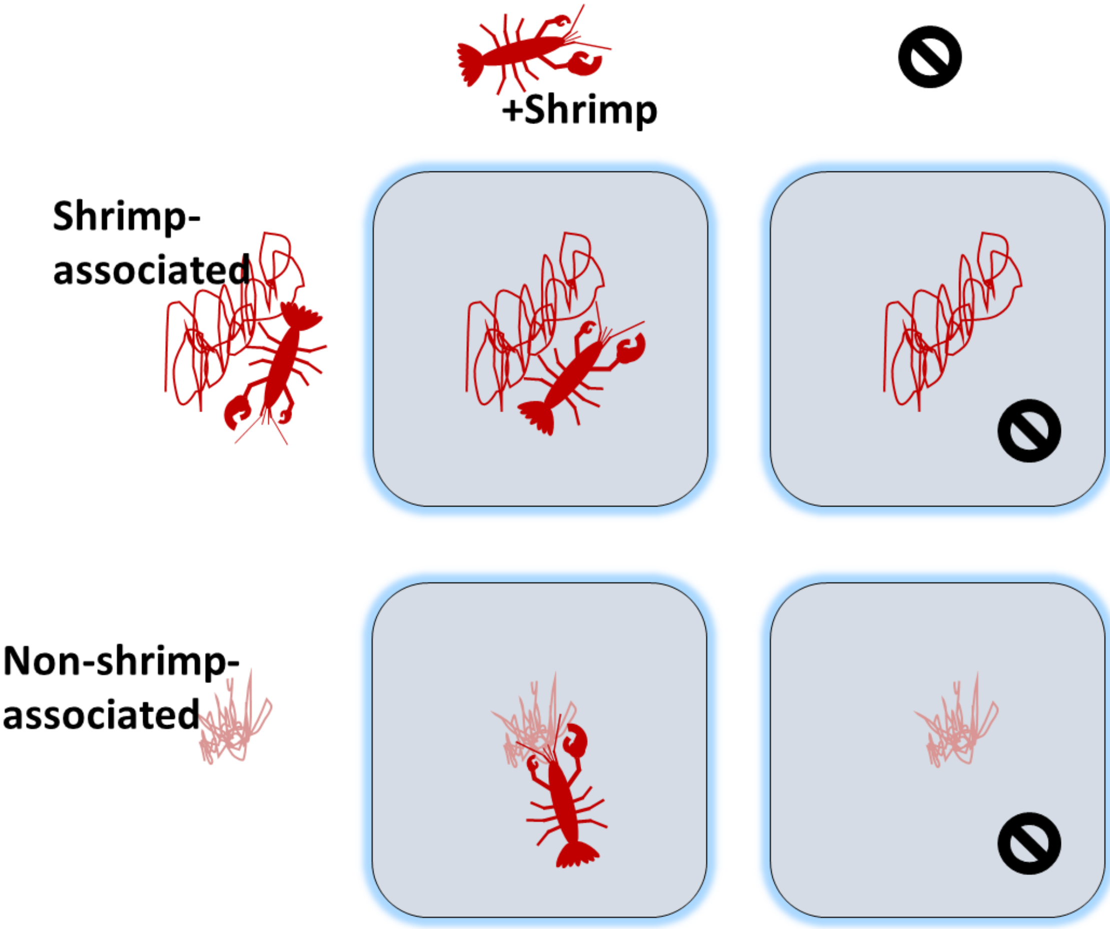


Figure 3.1 – Illustration of the two-factorial induction experiment's four experimental conditions. Shrimp-associated or non-shrimp-associated *M. bouillonii* was grown with or without a shrimp for 15 days. There were five replicates of each condition.

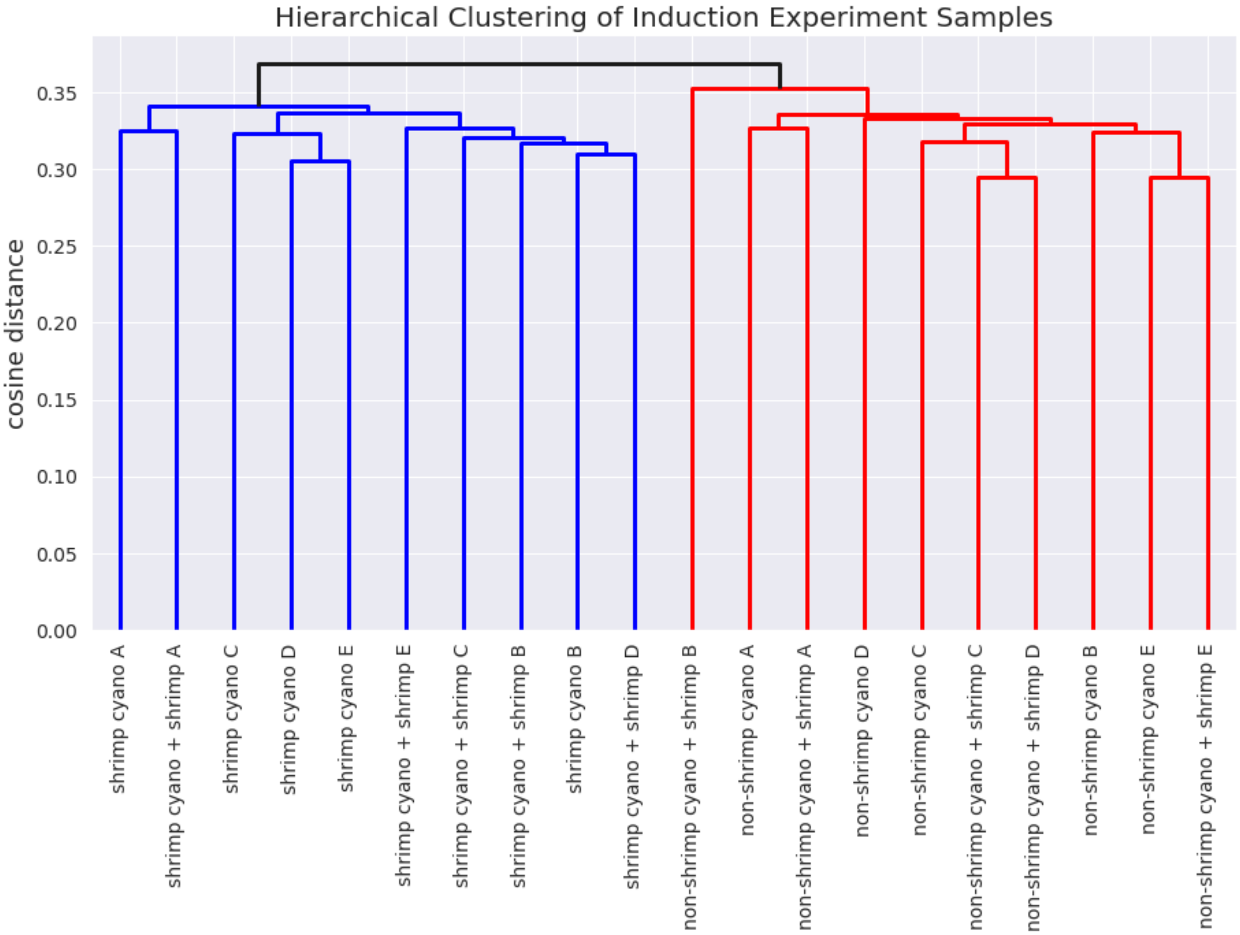


Figure 3.2 - ORCA dendrogram illustrating that regardless of if samples were grown with shrimp or without shrimp, they maintained the chemical composition of the collected organism. The blue cluster contains all samples of the shrimp-associated chemotype, regardless of whether they were grown with shrimp during the experiment. The red cluster contains all samples of the non-shrimp-associated chemotype, regardless of whether they were grown with shrimp during the experiment. The letters appended to the end of each sample name indicate which replicate block they were a part of (A-E).

3.2.3: Testing the preference hypothesis

To investigate the hypothesis that shrimp select a preferred *M. bouillonii* chemotype with which to engage, a behavioral assay for measuring shrimp preference was deployed. It was necessary to conduct behavioral assays at night, as the shrimp are typically quiescent (or very excitable and erratic when exposed) during the day but will move around freely at night. In the preference assay, small portions of shrimp-associated cyanobacterium and non-shrimp-associated cyanobacterium were attached to a cement tile and placed at one end of a small tank. A shrimp was then introduced to the other end of the tank, and the tank was left alone for 30 minutes. After the allotted time, the shrimp's end location was recorded. In 26 out of 30 individual shrimp trials, shrimp were found to have woven themselves into the shrimp-associated cyanobacterium. In the remaining four trials, shrimp were found to have woven mainly the shrimp-associated cyanobacterium and used small pieces of the non-shrimp-associated cyanobacterium to fill any gaps. These results strongly suggested a preference for the shrimp-associated cyanobacterium by *A. frontalis* but did not constrain what that preference was based on. The experiment was structured in such a way that shrimp could feasibly be making their cyanobacterial selection based on visual, chemical, or tactile cues, or a combination of these. In order to more specifically identify the means by which shrimp select their preferred cyanobacteria, more experiments were undertaken.

3.2.4: Testing for preference via waterborne chemical cue

Two different Y-tube assays were conducted to assess whether or not a waterborne chemical cue was responsible for *A. frontalis* selection of its preferred cyanobacterial chemotype. The basic configuration for these assays involved a Y-tube with chambers at the end of each arm and the tail, with seawater flowing through it from the arms to the tail (Figure 3.3). Chemical stimuli were placed in the arm chambers, a shrimp was placed in the tail chamber. The shrimp was continuously observed for 20 minutes, or until it fully entered and remained in one of the terminal arm chambers. In the first round of Y-tube assays, one arm chamber contained non-shrimp-associated cyanobacteria while the other was empty. Of the 20 shrimp assayed, eight travelled to the empty chamber, six to the non-shrimp-associated cyanobacterium chamber, and six shrimp failed to reach an arm chamber within 20 minutes (Figure 3.4A).[3] Chi-square analysis of these results yield a p-value of 0.82, as the behavior exhibited by the shrimp was essentially random. In the second set of Y-tube assays, one chamber contained shrimp-associated cyanobacteria and the other contained non-shrimp-associated cyanobacteria. Out of the 20 shrimp assayed, all 20 shrimp sought out the chamber containing the shrimp-associated cyanobacteria (Figure 3.4B). Chi-square analysis of these results yielded a p-value of 2.1 x 10^{-9}, clearly indicating statistical significance. This result provided strong support for the hypothesis that shrimp are selecting a preferred *M. bouillonii* chemotype and suggested that there is a water-borne chemical cue that is facilitating shrimp efforts to locate the preferred *M. bouillonii* chemotype.

[3] The rationale for setting the null hypothesis of the chi-square analysis to give equal likelihood to shrimp terminating the experiment in either terminal chamber or no chamber at all was that shrimp not reaching a terminal chamber was a very common occurrence, specifically when they did not have access to the preferred chemotype of *M. bouillonii*, and so it was unreasonable to not include this common outcome as a valid experimental endpoint. A question could be raised that if all shrimp reached terminal chambers, with an even split between the two chambers and a statistically significant result, how would this be interpreted? All shrimp reaching terminal chambers, with an equal division across both chambers, would suggest that shrimp are stimulated to travel towards seawater flow, either due to chemical stimuli intrinsic to the raw seawater system or as a behavioral response to hydrodynamic flow.

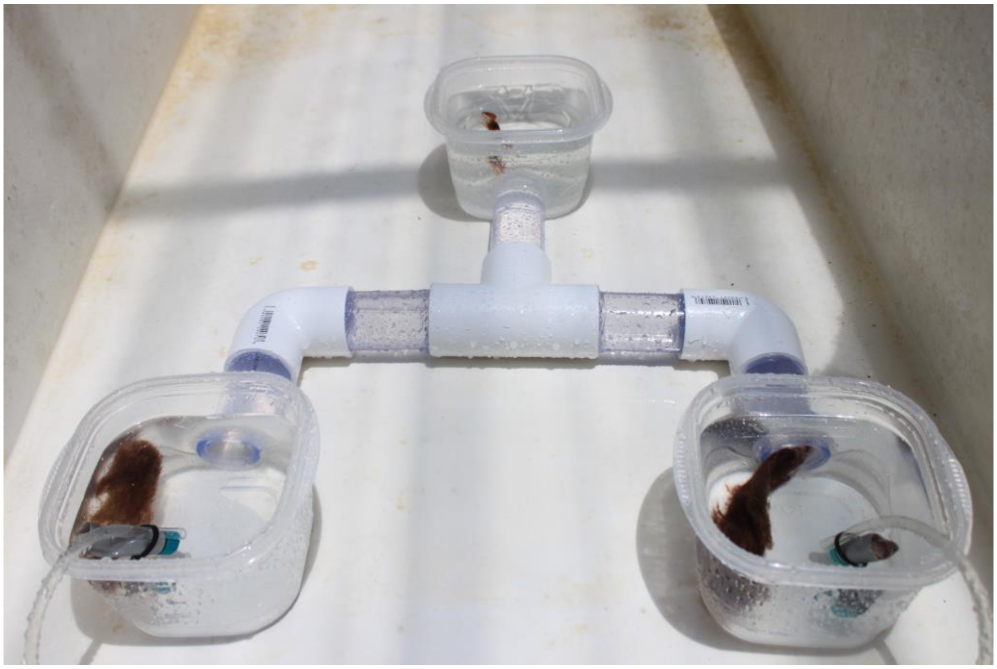


Figure 3.3 – Photograph of the Y-tube experimental setup. This view looks from the arms of the Y, where seawater enters chambers containing chemical stimuli, and towards the tail of the Y, where shrimp are placed into a chamber at the beginning of each trial and where seawater flows from the apparatus.

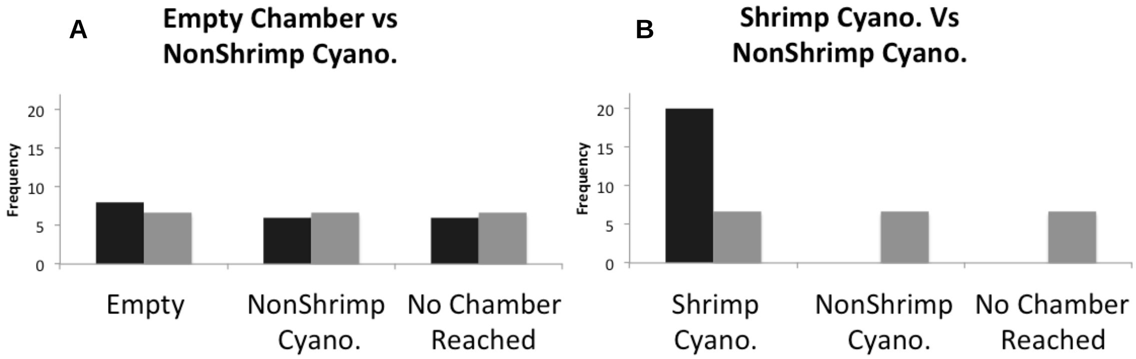


Figure 3.4 – Comparing shrimp Y-tube responses when given A) the choice between non-shrimp-associated *M. bouillonii* versus seawater and B) non-shrimp-associated *M. bouillonii* versus shrimp-associated *M. bouillonii*.

Not only did shrimp show a unanimous preference for the shrimp-associated chemotype, they also tended to complete Y-tube trials much more rapidly when this chemotype was available to them (Table 3.S1). 65% of shrimp completed the Y-tube in less than 2 minutes, and 95% completed in less than 10 minutes when the shrimp-associated chemotype of *M.*

bouillonii was presented as a chemical stimulus in one of the Y-tube arm chambers. In comparison, only 25% of shrimp completed the Y-tube within 2 minutes while 55% completed within 10 minutes when shrimp were only presented with the non-shrimp-associated chemotype of *M. bouillonii* versus seawater. This lends further support that shrimp are actively selecting the shrimp-associated chemotype and are stimulated into action when they detect its waterborne cue.

3.2.5: Defining the waterborne cue

Characterizing the waterborne cue responsible for the ability of *A. frontalis* to locate its preferred *M. bouillonii* chemotype has proven to be exceedingly difficult, with numerous environmental, chemical, and technical challenges. Beyond the challenge of daylight causing shrimp to immediately (and frantically) seek shelter, much more nuanced and (sometimes) less dramatic environmental challenges also arose. Disruptions in shrimp behavior during Y-tube experiments were observed resulting from the light of a full moon emerging from cloud cover, the shining of headlights from vehicles in the Guam Marine Laboratory parking lot, and the vibrations of music playing loudly nearby. Light and sound caused shrimp to take shelter in the elbow joints and 'T' joint of the Y-tube; these joints were made of opaque PVC as compared to the transparent vinyl tubes and plastic containers that comprised the rest of the Y-tube's construction. Oceanographic events also occasionally interfered with obtaining assay results, as upwelling events and storm surges caused the raw flow-through seawater system that supplied the Y-tube to be dense with plankton and particulates that the shrimp found particularly appetizing and much more interesting that any waterborne cyanobacterial cue that they may have been detecting.

Irrespective of the environmental challenges, which posed more nuisance than detriment, the task of isolating the waterborne cue from *M. bouillonii* or *M. bouillonii* seawater proved to be a major obstacle. A combination of XAD resins and multiple eluting solvents, as well as biomass crude extracts, dissolved into seawater and tested at various concentrations in the Y-tube failed to elicit consistent and robust shrimp responses. Further complicating the matter, while efforts were made to apply ecologically relevant concentrations of resin eluent, because no consistent behavior was observed in response to this eluent, it was near impossible to determine if methods needed to change in terms of capturing the cue, applying the cue in the assay, or both. Nevertheless, it was indeed clear that methods needed to change.

After encountering the various challenges and reflecting on how they limited progress in better understanding the shrimp attractant waterborne cue, the focus of the work shifted from identifying the cue outright to systematically constraining characteristics of a possible cue, and thus limiting possibilities of what the cue might be. A progression of Y-tube experiments was conducted (Table 3.1), starting where the original Y-tube experiment began with a comparison of shrimp-associated *M. bouillonii* in one Y-tube arm versus seawater. In agreement with prior results, 10 out of the 10 shrimp assayed selected the Y-tube chamber containing *M. bouillonii* over the control chamber. The same experiment was then conducted with *M. bouillonii* in a secondary container, from which infused seawater flowed into the Y-tube. Again, 10 out of 10 shrimp selected the Y-tube chamber receiving infused seawater. In the next iteration, seawater flowing out of a container with *M. bouillonii* was first passed through cheesecloth before reaching the Y-tube. The cheesecloth limited access to the Y-tube for macroscopic particles, trapping broken cyanobacterial filaments and other particulate matter. In this case, 9 out of 10 shrimp selected for the Y-tube chamber receiving *M. bouillonii* infused and coarsely filtered seawater, illustrating a continued statistically significant preference of shrimp for *M. bouillonii* exposed seawater (p-value = 6.71 x 10^{-4}).

Table 3.1 – Results of Y-tube experiments aimed at constraining the waterborne shrimp attractant cue.

	cyano cue	**seawater control**	**no chamber reached**	**p-value**
expected	3.33	3.33	3.33	
M. bouillonii	10	0	0	4.49 x 10^{-5}
M. bouillonii flowthrough	10	0	0	4.49 x 10^{-5}
M. bouillonii flowthrough, cheesecloth	9	1	0	6.71 x 10^{-4}
M. bouillonii in dialysis tubing	6	2	2	0.202
M. bouillonii in benzolated dialysis tubing	6	2	2	0.202

Following filtration via cheesecloth, the Y-tube experiments progressed to finer scale filtration using dialysis tubing. When *M. bouillonii* sealed in dialysis tubing with seawater was placed in a secondary container, and seawater was directed through this container and into the Y-tube, only 6 out of 10 shrimp favored the associated Y-tube arm, while 2 selected the control arm and 2 failed to reach a Y-tube arm chamber after the 10 minutes of allotted experimental time. This outcome was found to not be statistically significant from random expectation (p-value = 0.202). The final iteration of Y-tube experiments involved an identical arrangement, aside from *M. bouillonii* being sealed instead in benzoylated dialysis tubing featuring smaller pores. This round of Y-tube trials yielded an identical result: out of 10 shrimp, 6 selected the Y-tube chamber with *M. bouillonii* dialysis tubing flow through, while 2 shrimp selected the control chamber and 2 shrimp failed to finish.

The results of the dialysis tubing Y-tube trials required careful consideration, as they presented some ambiguity in how to be interpreted. From the standpoint of statistical significance, both types of dialysis tubing failed to provide statistically significant evidence of the waterborne cue being transmitted through the membrane pores in sufficient concentration to generate a robust shrimp response. The regular dialysis tubing was reported to contain pores with a size cutoff between 12,000 and 14,000 Da, while a size cutoff between 1,200 and 2,000 Da was attributed to the benzoylated dialysis tubing. One would then expect that after seeing a suggestive yet non-consistent, non-significant response from the regular dialysis tubing, no shrimp response should be observable when *M. bouillonii* is sealed into benzoylated dialysis tubing. One might conclude from such an outcome that the waterborne cue is a macromolecule that is too big to diffuse through dialysis tubing. However, the results of the benzoylated dialysis tubing trials were identical to those of the regular dialysis tubing, and the observations of shrimp behavior during these trials suggested an alternative scenario. When shrimp did select the Y-tube chamber that was receiving seawater exposed to dialysis tubing encased *M. bouillonii*, they tended to respond very rapidly and displayed behaviors associated with cue tracking (e.g. direct travel interrupted by sharp pauses during which antennules are animatedly articulated). If one accepts that both sets of dialysis tubing trials (regular and benzoylated) displayed non-statically significant trends towards the cyanobacteria-exposed Y-tube chambers, and also considers the shrimp behaviors during these experiments, this could suggest that the chemical cue was diffusing through both types of dialysis tubing, but not doing so unencumbered. This potential limited propensity for diffusing through both types of dialysis tubing would suggest that a factor other than size, such as solubility, may have been contributing to the diminished capacity of the cue for diffusion across these semipermeable membranes.

3.2.6: A breakthrough

A waterborne cue with poor solubility in water may sound paradoxical, but the curious results from the dialysis tubing Y-tube experiments were not alone in suggesting a lipophilic and/or volatile cue. It was observed that water infused with *M. bouillonii* seemed to lose shrimp attractant activity over time, suggesting either evaporation or degradation. It was also observed that filtering infused seawater with a vacuum pump, which thoroughly aerated the seawater during the process, resulted in the loss of shrimp attractant activity. The aeration could facilitate the off-gasing of a volatile cue, or a lipophilic cue may be lost through adsorption to particulates captured by the filter. These observations, bolstered by the results of the dialysis tubing Y-tube experiments and considered alongside the numerous failures to elute the cue from resins with more polar solvents, stimulated a new approach with a focus on a potentially lipophilic and/or volatile cue.

A smaller assay format, involving cotton balls impregnated with hexane eluent from C18 solid phase extraction (SPE) of *M. bouillonii* infused seawater versus hexane control placed on opposing sides of a glass dish, was adopted to avoid challenges with adequate concentrations in the flowing seawater Y-tube system. Quantification of shrimp contact with treatment versus control cotton balls over seven trials resulted in an average of 164.9 seconds in contact with treated cotton balls and 102.1 seconds in contact with control. Using a paired-sample, single-tailed t-test yielded a p-value of 0.060, suggesting shrimp preference for association with the treated cotton ball but falling just short of statistical significance, given a p-value cutoff of 0.05 (Table 3.S2). More compelling than what was captured by quantifying the time spent by the shrimp in contact with treated versus control cotton balls were the dramatic behaviors displayed by some shrimp in response to contacting the treated cotton ball. Shrimp could be seen attempting to burrow into the treated cotton balls, and would curl around them, just as they had been observed behaving when offered a ball of shrimp-associated *M. bouillonii* (Figure 3.5). This striking behavior, of interacting with the treated cotton balls as they would with a ball of cyanobacteria, represents the clearest indicator to date of successful capture and subsequent deployment of the shrimp attractant chemical cue.

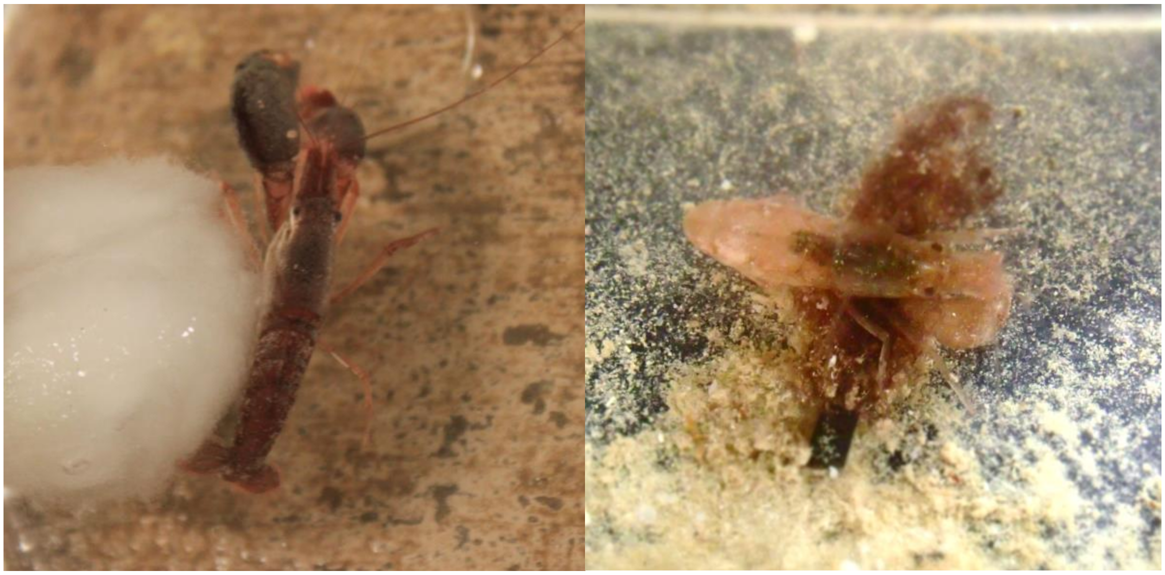


Figure 3.5 – A) Shrimp curling around a treated cotton ball. This behavior was highly similar to what was observed in B) shrimp approaching shrimp-associated *M. bouillonii*.

3.2.6: Comparative gas chromatography for cue identification

Cultures of shrimp-associated *M. bouillonii*, collected from Guam (by the author) and Papua New Guinea, and a culture of non-shrimp-associated *M. bouillonii*, collected from Guam (by the author), provided a continuous supply of spent seawater media that had been infused with the cue. Detection and characterization of the cue via gas chromatography mass spectrometry (GC-MS) presented method development challenges. Low strength of signals detected from the headspace of live cultures, as well as those detected from liquid-liquid extraction of spent media suggested an apparent low concentration of the cue. The methods used for preparing the cue for the above described cotton ball assays, specifically the concentration of the cue on C18 SPE cartridges followed by elution with hexane, were then employed with much greater success. Partial least squares discriminant analysis (PLS-DA) (Figure 3.S4) and a random forest model (Figure 3.S5) were used to independently rank molecular features of importance in differentiating between samples from the non-shrimp-associated Guam culture and the two shrimp-associated cultures. Both methods ranked feature 2939 as the top hit with a distribution of high abundance in the shrimp-associated cultures and low abundance in the non-shrimp-associated culture. Improved signal strength resulting from the methodological shift to SPE concentration preparation allowed for multiple strong annotations of this feature as isopropyl myristate (cosine scores 0.92, 0.88, 0.84, 0.83) (Figure 3.6).

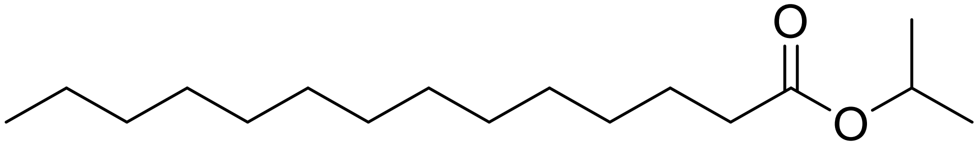


Figure 3.6 – Isopropyl myristate, potential waterborne chemical cue that attracts *A. frontalis* to the shrimp-associated chemotype of *M. bouillonii.*

In addition to its compelling distribution measured between *M. bouillonii* cultures, isopropyl myristate possesses a number of characteristics that support its potential as our sought after waterborne chemical cue. While there is no literature precedent for its involvement in chemical signaling, isopropyl myristate has previously been detected in algal extracts (Yasmeen et al 2018) and cyanobacterial cultures (Huang, Cheng, and Cheng 2008; Huang, Lai, and Cheng 2007). The lipophilicity of this compound aligns well with the peculiar cue behaviors observed that suggested a lipophilic cue, but with a boiling point of 315 °C at 760 mm Hg (PubChem [https://pubchem.ncbi.nlm.nih.gov/]), it would not be expected to be particularly volatile under ambient conditions. The ester functionality fits particularly well with our observed lessening of cue efficacy over time, as the ester would be quite susceptible to hydrolysis in seawater.

Isopropyl myristate is a promising lead for the waterborne chemical cue, but it is not the only compound worth pursuing. Isopropyl myristate was the top scoring annotation for feature 2939, but a total of 32 annotations were assigned with cosine scores of 0.66 or greater. Of those 32 annotations, 3 of the top 6 were isopropyl myristate, and many of the others could be deemed implausible after manual inspection of the query spectrum as compared to the library annotation spectrum. However, alternative annotations such as butyl myristate, myristic anhydride, and myristic acid isobutyl ester, as well as various tetradecanoic acid esters, displayed library spectra that matched well with the query spectrum of 2939, and so warrant further consideration. Additionally, while 2939 was ranked as the highest importance feature with a distribution of high abundance in the shrimp-associated cultures and low abundance in the non-shrimp-associated culture according to both PLS-DA VIP scoring and a random forest model, three other features in the top 15 ranked features of both models had the same distribution. These features were 1460, 1159, and 1807. Based on annotation results, features 1460 and 1159 are likely dioxolane compounds, while feature 1807 is suggested to be in the benzoquinone class of compounds.

At present, collaborative efforts between the Gerwick, Dorrestein, and Biggs laboratories continue work on verifying the identity of the waterborne shrimp attractant produced by shrimp-associated *M. bouillonii*, and biologically validating it. While the annotations of isopropyl myristate were strongly supported, its identity must be confirmed with more replicates, comparison to a pure standard, and co-injection. Candidate compounds for features 1460, 1159, and 1807 should also have their identities confirmed. These potential cues could then be biologically validated by applying them in dilution to cotton balls and offering them to *A. frontalis* either in the format of the cotton ball assay or in the Y-tube.

3.3: Methods

3.3.1: Preliminary collections, sample processing, and LC-MS/MS profiling

Initial shrimp-associated and non-shrimp-associated *M. bouillonii* samples were collected from by the blue barge dry dock (13°26'57"N:144°39'27"E) and Gab Gab Beach (13°26'38"N:144°38'36"E), respectively, in Apra Harbor, Guam on March 21, 2016. Samples were preserved in 1:1 isopropyl alcohol and seawater and stored at -20 °C when not in transit. Both samples were exhaustively extracted with 2:1 dichloromethane and methanol to produce crude extracts. Extracts were concentrated in vacuum, resuspended in 1:1 acetonitrile and methanol, processed through C18 SPE prior to LC-MS/MS analysis, and prepared for analysis in 1:1 acetonitrile and methanol at a concentration of 4 mg/mL. LC-MS/MS analysis was conducted using a ThermoFinnigan Surveyor HPLC System (San Jose, CA, USA) with a Phenomenex Kinetex 5 µm C18 100 × 4.6 mm column (Torrance, CA, USA) coupled to a ThermoFinnigan LCQ Advantage Max Mass Spectrometer (San Jose, CA, USA) in positive ion mode. Samples were analyzed using a linear gradient from 30% CH_3CN + 0.1% formic acid to 99% CH_3CN + 0.1% formic acid in H_2O + 0.1% formic acid at a flow rate of either 0.6 mL/min over 32 min. MSCONVERT (Holman, Tabb, and Mallick 2014), a part of the ProteoWizard Library (Chambers et al 2012), was used to convert proprietary LC-MS/MS files to mzXML, and MZmine (Pluskal et al 2010) was used to visualize resultant chromatograms.

3.3.2: Induction experiment

A. frontalis and *M. bouillonii* growing with (shrimp-associated cyanobacteria) and without *A. frontalis* (non-shrimp-associated cyanobacteria) were collected from Apra Harbor, Guam (with additional shrimp collected near Merizo, Guam) and used to assemble four different growth conditions in reusable plastic food storage containers: shrimp-associated cyanobacteria with a shrimp, shrimp-associated cyanobacterium without a shrimp, non-shrimp-associated cyanobacterium with a shrimp, and non-shrimp-associated cyanobacterium without a shrimp, each with five replicates. Containers were covered with window screening, and placed in a blocked arrangement, weighted and submerged in a flow-through raw seawater table. Submersion allowed for water exchange and the deposition of materials for shrimp to scavenge in each container, while the mesh was intended to prevent shrimp escape. The water table was covered with shade cloth to prevent light stress. The experiment was conducted over 15 days (June 11, 2016 to June 26, 2016). At the completion of the experiment, cyanobacterial biomass was preserved for chemical analysis in 1:1 isopropyl alcohol and seawater and stored at -20 °C when not in transit. Samples were prepared and analyzed via LC-MS/MS as described above. Comparison of resultant chromatograms was undertaken using the ORCA pipeline (Leber et al 2020), which is publicly available at https://github.com/c-leber/ORCA. The MS^1 feature processing parameter set used was as follows: bin_width = 1, bin_offset = 0, bins_start = 200, bins_end = 2000, peak_consecutivity = 0, peak_cluster_size_cutoff = 3, min_integral = 100000, rt_setting = 'raw', rrt_tolerance = 0.5. The resultant samples vs MS^1 features bucket table was logarithmically transformed.

3.3.3: Selection assay

The selection assay was conducted on 30 individual *A. frontalis* shrimp, both males and females, ranging in size from 2 cm to 5.5 cm, and collected from Finger Reef, Apra Harbor, Guam. The assay was conducted during June 2017. Each trial was conducted in a 12 L rectangular tank that was refilled with fresh seawater before each trial. *M. bouillonii* of both chemotypes (shrimp-associated and non-shrimp-associated) was loosely attached side-by-side to a square ceramic tile with zip ties. This tile was placed at one end of the tank, while a shrimp was placed at the other. The tank was left in the dark for 30 minutes, after which it was recorded which cyanobacterium the shrimp had begun to weave.

3.3.4: Selection Y-tube experiments

The Y-tube was constructed from clear vinyl tubing (1 inch inner diameter x 1-1/4 inch outer diameter), PVC pipe fittings (1-1/4 inch size), and medium square plastic food storage containers (Up & Up brand), and joints were sealed with aquarium safe, all-purpose silicone adhesive sealant (DAP brand). The total volume of the Y-tube was 2,550 mL, and it was operated at a flowrate of approximately 729 mL/min. Twenty individual shrimp, collected from Finger Reef, Apra Harbor, Guam, were assayed in Y-tube trials in which one Y-tube arm chamber contained the non-shrimp-associated chemotype of *M. bouillonii* and one chamber contained nothing. Prior to each trial, the Y-tube was rinsed with fresh water, thoroughly flushed with seawater, and then filled. Chemical stimuli were added to the arm chambers, with random assignment based on a coin flip, and then allowed to establish plumes in the running seawater through the Y-tube. After 3:30 minutes, a shrimp was gently added to the tail chamber to begin navigating the Y-tube. Each trial lasted 20 minutes, or until the shrimp fully entered one of the arm chambers. Trials were conducted at night in the dark, as light disrupts *A. frontalis* behavior. Twenty individual shrimp were similarly assayed in Y-tube trials in which one Y-tube arm chamber contained the non-shrimp-associated chemotype of *M. bouillonii* and one chamber contained the shrimp-associated chemotype of *M. bouillonii*. Both sets of trials were conducted during June 2017.

3.3.5: Cue defining Y-tube experiments

Using the same Y-tube described above, a series of Y-tube experiments were conducted during May and June of 2019 to define the waterborne shrimp attractant chemical cue associated with shrimp-associated *M. bouillonii*. Each of these experiments involved 10 individual shrimp collected from Piti Bomb Holes, Guam. The Y-tube was operated at a flowrate between 375 mL/min and 417 mL/min. Chemical stimuli plumes were allowed to develop in the Y-tube for 3:00 minutes prior to the initiation of each trial with the addition of a shrimp. Trials lasted for 10 minutes, or until shrimp fully entered one of the Y-tube arm chambers. Chemical stimuli were randomly assigned to the left or right arms of the Y-tube based on a coin flip. These experiments were conducted at night in the dark. In the first of this series of experiments, one Y-tube arm chamber contained the shrimp-associated chemotype of *M. bouillonii*, while the other contained nothing. In the next iteration, lids with holes were placed on top of each Y-tube arm chamber, and a 1L glass beaker was placed on each lid. One beaker contained shrimp-associated cyanobacterial biomass while the other was empty. Seawater was directed into these beakers, with the overflow flowing through the holes in the lids of the Y-tube arm chambers, and into the Y-tube arm chambers. This allowed cyanobacteria-infused seawater to be directed through the Y-tube, without cyanobacterial biomass being present in the Y-tube. For the next iteration, the prior setup was further amended with cheese cloth fastened over the lids of the Y-tube arm chambers; this provided filtration of macroscopic particles prior to seawater advancing the Y-tube. In the remaining two experimental iterations, the cheese cloth was removed, and instead of cyanobacterial biomass or nothing being placed in the 1L beakers above the Y-tube, cyanobacterial biomass of the shrimp-associated chemotype of *M. bouillonii* and seawater and just seawater were sealed into dialysis tubing (Carolina Biological Supply; Burlington, NC), and then benzoylated dialysis tubing (Sigma-Aldrich; St. Louis, MO). For both of these iterations, cyanobacteria and seawater or just seawater in dialysis tubing and in benzoylated dialysis tubing were allowed to soak in the 1L beakers of seawater for 1:15 hours prior to beginning trials.

3.3.6: Cotton ball assay

Seven trials were conducted of this assay. Per trial, approximately 200 mL of the shrimp-associated chemotype of *M. bouillonii* was agitated in 500 mL of seawater and allowed to soak for 30 minutes. The seawater was filtered through cheesecloth to remove cyanobacterial biomass, and the biomass was compressed to release additional liquid. An SPE cartridge (SEClute SPE C18-max 2000mg/20mL) augmented with 2 cotton balls to trap sediment was washed with ~15 mL each of filtered water, methanol, hexane, and filtered water. The cyanobacteria infused seawater was vacuumed through the SPE cartridge, and the cotton balls were removed from the cartridge and discarded. 5 mL of hexanes was used elute from the SPE cartridge. 3.5 mL of the hexane eluent was added to a treatment cotton ball, while 3.5 mL of hexane was added to a control cotton ball. In a circular glass dish, filled with seawater and containing one shrimp, each cotton ball tethered with a safety pin attached to string was placed opposite the other. Placement of cotton balls (treatment on left or right side of dish) was determined via coin flip. Trials were filmed for 10 minutes, and time spent in contact with control and treatment cotton balls was measured. Trials were filmed with a Canon EOS Rebel T1i and a Canon PowerShot® ELPH 100 HS.

3.3.7: GC-MS profiling of *M. bouillonii* culture spent media

Three *M. bouillonii* cultures were analyzed: shrimp-associated chemotype originally collected from Pigeon Island, New Ireland, Papua New Guinea (4°16'04"S:152°20'16"E) on May 19, 2005; shrimp-associated chemotype collected from Finger Reef, Apra Harbor, Guam (13°26'40"N:144°38'11"E) on June 15, 2017; and non-shrimp-associated chemotype collected from Finger Reef, Apra Harbor, Guam (13°26'40"N:144°38'11"E) on May, 23, 2019. Samples were prepared by passing cyanobacterial biomass into fresh media and passing the spent media through a C18 SPE cartridge under vacuum (pre-washed with 15 mL hexane). The SPE cartridge was eluted with 5 mL of hexane, aliquots of which were transferred into 200 µL inserts in 2 mL vials and sealed with silicone septum containing caps. The GC-MS analysis was carried out using an Agilent 7200 GC-QTOF equipped with an autosampling system and an HP-5MS column (30 m x 0.25 mm x 0.25 µm). 1 µL of sample was injected at a 100:1 split. The injector temperature was set to 250 °C. The analysis protocol was as follows: starting temperature 40 °C for 1 min; 20 °C/min oven ramp to 110 °C; 10 °C/min oven ramp to 300 °C; 50 °C/min oven ramp to 320 °C to purge the column. The helium carrier gas was set to constant flowrate of 1.2 mL/min. The scanned *m/z* range was 35-400, with an acquisition rate of 20 spectra/s. Solvent blanks and vial blanks were interspersed with the samples to assess background signal.

Acquired data were converted from .d format to .mzml and analyzed using GNPS GC-MS workflows for deconvolution and library search (Aksenov et al 2020). The feature table was downloaded from the GNPS deconvolution workflow and filtered to remove features with balance scores below 65%. Feature importance analyses were performed using Metabolanlyst (Chong et al 2018). The data were normalized using quantile normalization and scaled with autoscale and log-transformation. Relevant GNPS jobs are available at the following links:
Deconvolution:
https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=8004c66ae8c94d37a615e4a9b26902e2
Library search with Wiley:
https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=3e0c81872a604e5881a51560c09d8bf2
Library search without Wiley:
https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=b37b05bd4f5f40868bce81f7ca49f072
Data will be made publicly available on MassIVE (Aron et al 2020) when the manuscript is submitted for publication.

Chapter 3 is coauthored with Naman, C. Benjamin; Aksenov, Alexander A.; Reyes, Andres Joshua; Glukhov, Evgenia; Dorrestein, Pieter C.; Biggs, Jason S.; Gerwick, William H. The dissertation author was the primary author of this chapter. The dissertation author co-conceived of the work, participated in cyanobacteria and shrimp collections, designed and implemented all experiments, collected LC-MS data and prepared samples for comparative GC-MS, conducted data analyses, and was the primary author of this work.

3.4: Appendix: Supplemental Information

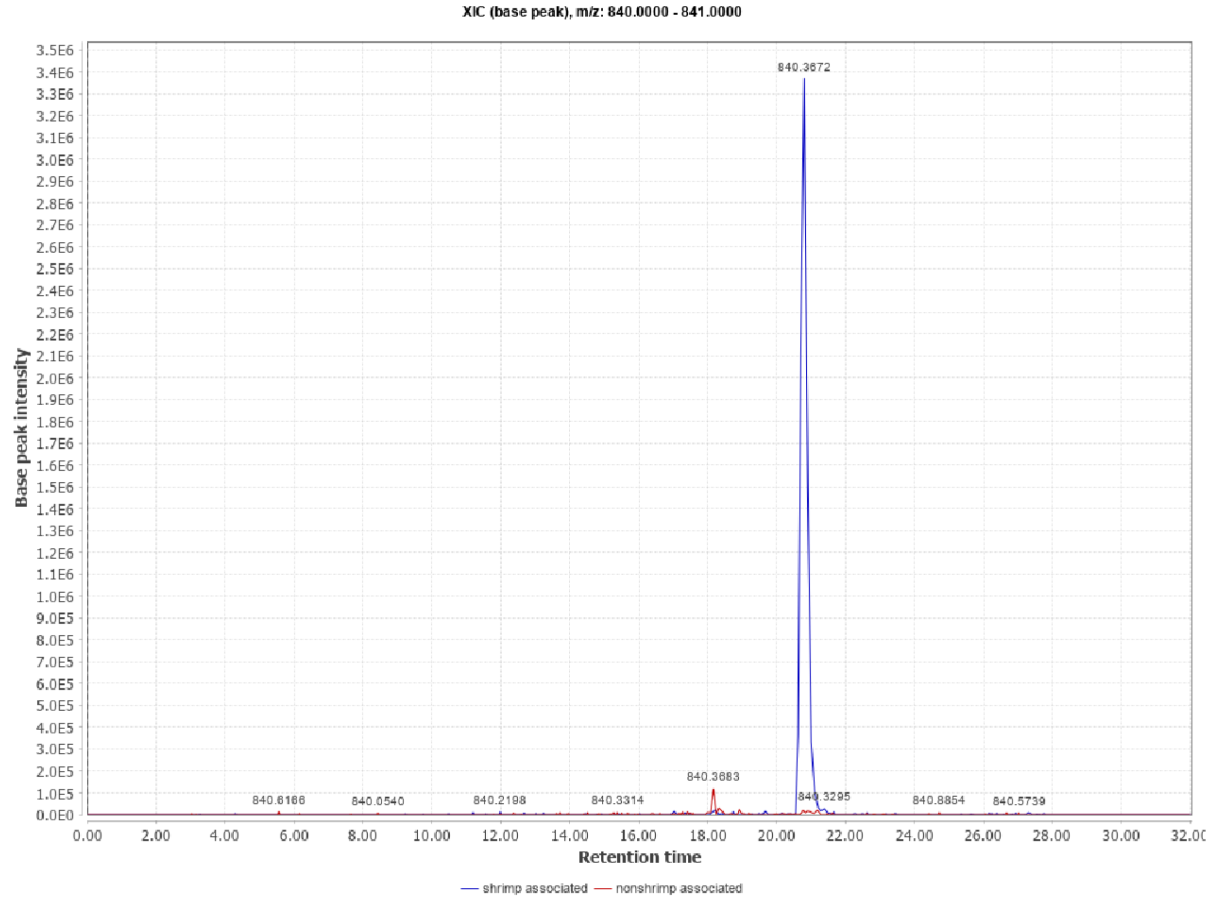


Figure 3.S1 – Extracted ion chromatogram between *m/z* 840 and 841, illustrating differential production of apratoxin A in a shrimp-associated *M.* bouillonii sample [blue] and a non-shrimp-associated *M. bouillonii* sample [red].

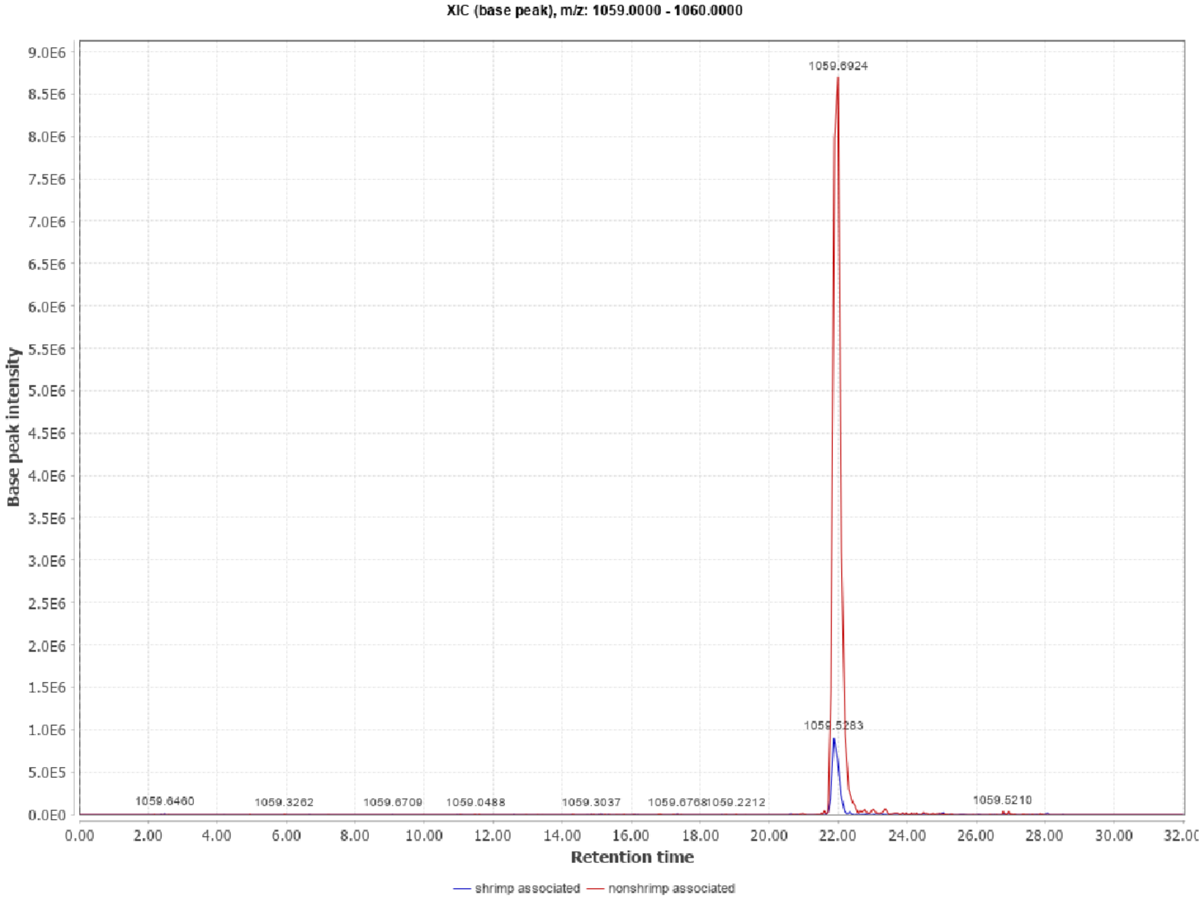


Figure 3.S2 - Extracted ion chromatogram between *m/z* 1059 and 1060, illustrating differential production of lyngbyastatin 2 in a shrimp-associated *M.* bouillonii sample [blue] and a non-shrimp-associated *M. bouillonii* sample [red].

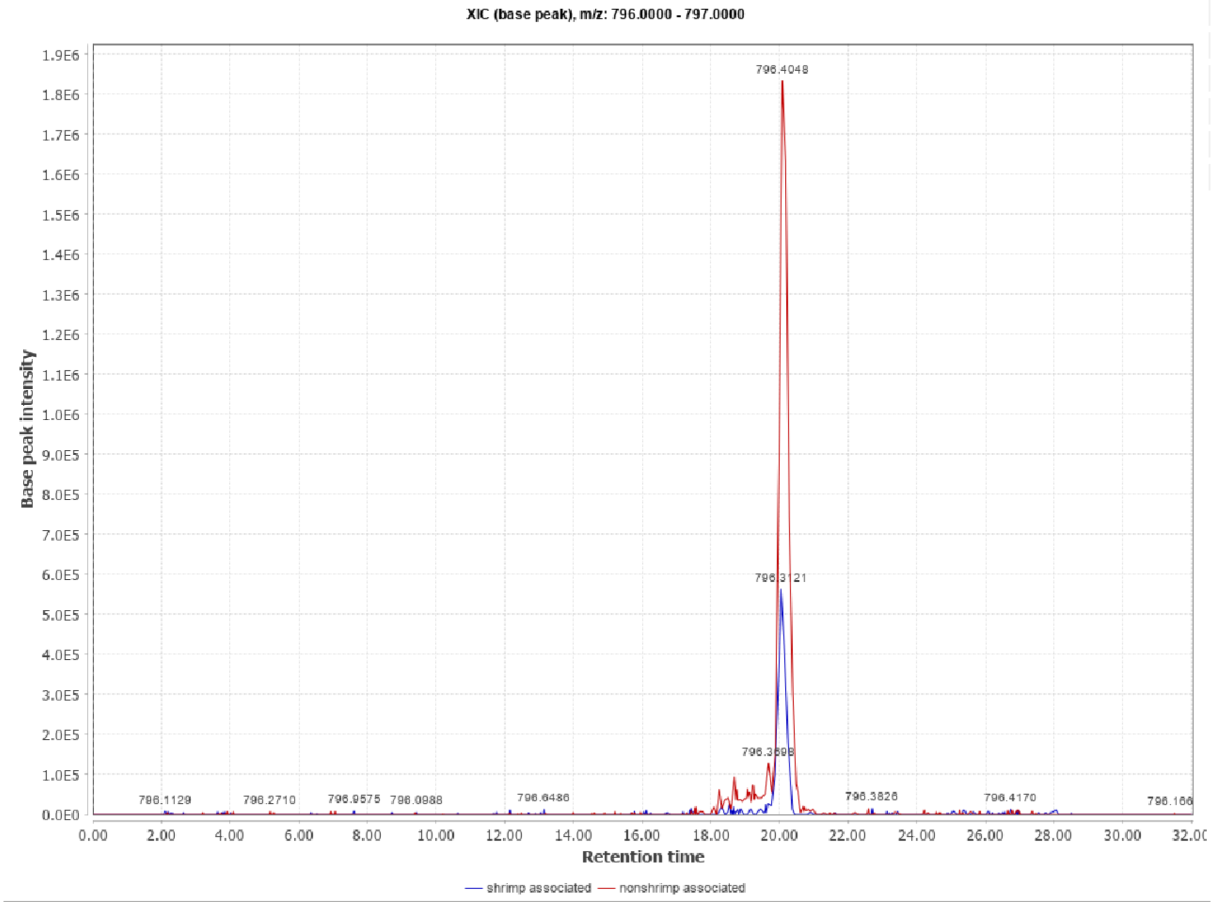


Figure 3.S3 - Extracted ion chromatogram between *m/z* 796 and 797, illustrating differential production of apratoxin E in a shrimp-associated *M*. bouillonii sample [blue] and a non-shrimp-associated *M. bouillonii* sample [red].

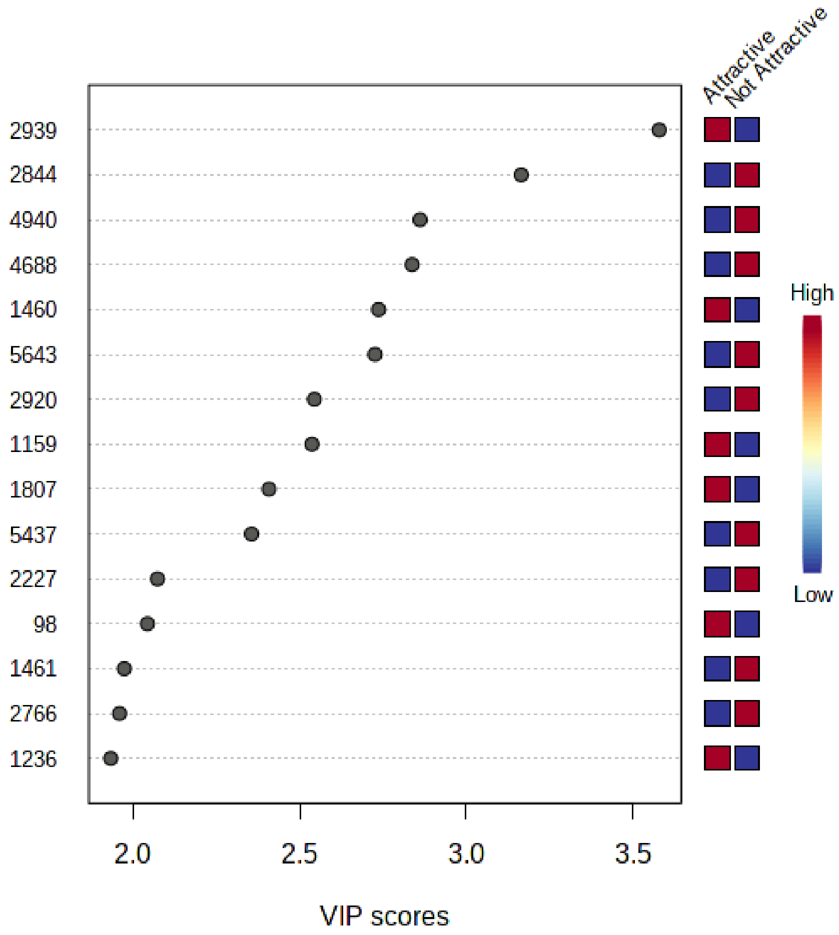


Figure 3.S4 – VIP scores plot resulting from PLS-DA comparing the GC-MS molecular features of the shrimp-associated chemotype of *M. bouillonii* cultures (denoted above as ‘Attractive’) versus the non-shrimp-associated chemotype culture (denoted as Not Attractive). Feature 2939, which corresponds to isopropyl myristate, had the highest VIP score and a distribution of high abundance in the shrimp-associated chemotype and low abundance in the non-shrimp-associated chemotype.

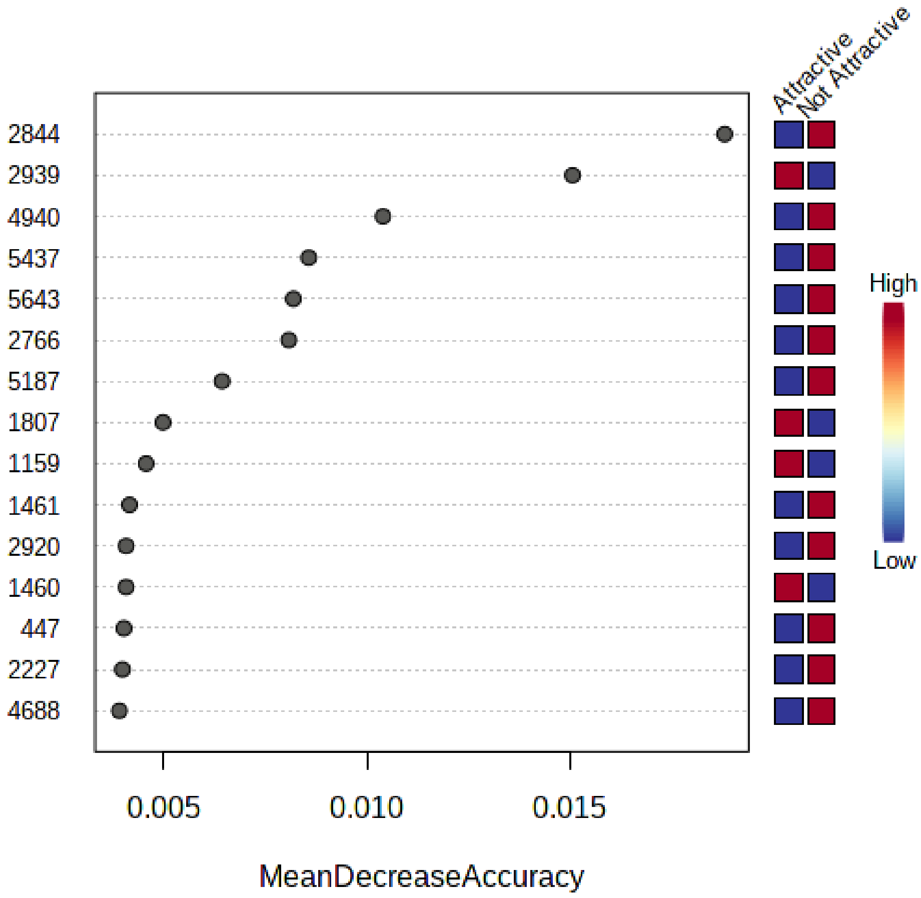


Figure 3.S5 – Random forest feature importance plot comparing the GC-MS molecular features of the shrimp-associated chemotype of *M. bouillonii* cultures (denoted above as 'Attractive') versus the non-shrimp-associated chemotype culture (denoted as Not Attractive). Feature 2939, which corresponds to isopropyl myristate, generated the largest mean decrease in accuracy when removed from the model of any feature with a distribution of high abundance in the shrimp-associated chemotype and low abundance in the non-shrimp-associated chemotype.

Table 3.S1 – Rapidity of response in Y-tube experiments. Shrimp responses were much more rapid when presented with the shrimp-associated chemotype of *M. bouillonii* as one of the chemical stimuli in the Y-tube. This lends further support to the conclusion that shrimp are actively responding to a waterborne cue released by that chemotype.

	empty vs non-shrimp cyano	**shrimp cyano vs non-shrimp cyano**
response time	# of shrimp (proportion)	# of shrimp (proportion)
< 10 mins	11 (55%)	19 (95%)
< 2 mins	5 (25%)	13 (65%)

Table 3.S2 – Time of shrimp association for cotton ball assay. Treatment indicates cotton balls that were impregnated with hexane SPE eluent, while control indicates cotton balls to which hexane was added. A paired-sample, single-tailed t-test was conducted to test for a statistically significant difference between the time shrimp spent in contact with the treated versus control cotton balls, yielded a p-value of 0.06.

	treatment	**control**	**difference**
trial 1	234	0	234
trial 2	100	99	1
trial 3	211	76	135
trial 4	55	79	-24
trial 5	221	167	54
trial 6	155	114	41
trial 7	178	180	-2
avg.	164.86	102.14	62.71
s.e.	25.17	22.92	34.72
p-value	0.060		

3.5: References

Aksenov AA, Laponogov I, Zhang Z, Doran SLF, Belluomo I, Veselkov D, Bittremieux W, Nothias LF, Nothias-Esposito M, Maloney KN, Misra BB, Melnik AV, Smirnov A, Du X, Jones II KL, Dorrestein K, Panitchpakdi M, Ernst M, van der Hooft JJJ, Gonzalez M, Carazzone C, Amézquita A, Callewaert C, Morton JT, Quinn RA, Bouslimani A, Orio AA, Petras D, Smania AM, Couvillion SP, Burnet MC, Nicora CD, Zink E, Metz TO, Artaev V, Humston-Fulmer E, Gregor R, Meijler MM, Mizrahi I, Eyal S, Anderson B, Dutton R, Lugan R, Le Boulch P, Guitton Y, Prevost S, Poirier A, Dervilly G, Le Bizec B, Fait A, Persi NS, Song C, Gashu K, Coras R, Guma M, Manasson J, Scher JU, Barupal DK, Alseekh S, Fernie AR, Mirnezami R, Vasiliou V, Schmid R, Borisov RS, Kulikova LN, Knight R, Wang M, Hanna GB, Dorrestein PC, Veselkov K (2020) Auto-deconvolution and molecular networking of gas chromatography–mass spectrometry data. Nat Biotechnol. https://doi.org/10.1038/s41587-020-0700-3

Aron AT, Gentry EC, McPhail KL, Nothias LF, Nothias-Esposito M, Bouslimani A, Petras D, Gauglitz JM, Sikora N, Vargas F, van der Hooft JJJ, Ernst M, Kang KB, Aceves CM, Caraballo-Rodríguez AM, Koester I, Weldon KC, Bertrand S, Roullier C, Sun K, Tehan RM, Boya P CA, Christian MH, Gutiérrez M, Ulloa AM, Mora JAT, Mojica-Flores R, Lakey-Beitia J, Vásquez-Chaves V, Zhang Y, Calderón AI, Tayler N, Keyzers RA, Tugizimana F, Ndlovu N, Aksenov AA, Jarmusch AK, Schmid R, Truman AW, Bandeira N, Wang M, Dorrestein PC (2020) Reproducible molecular networking of untargeted mass spectrometry data using GNPS. Nat Protoc 15:1954–1991. https://doi.org/10.1038/s41596-020-0317-5

Banner DM, Banner AH (1982) The alpheid shrimp of Australia. Rec Aust Mus Suppl 34:359–362

Bergkvist J, Selander E, Pavia H (2008) Induction of toxin production in dinoflagellates: the grazer makes a difference. Oecologia 156:147–154. https://doi.org/10.1007/s00442-008-0981-6

Bergvall UA, Rautio P, Kesti K, Tuomi J, Leimar O (2006) Associational effects of plant defences in relation to within- and between-patch food choice by a mammalian herbivore: Neighbour contrast susceptibility and defence. Oecologia 147:253-60. https://doi.org/10.1007/s00442-005-0260-8

Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak MY, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P (2012) A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 30:918–920. https://doi.org/10.1038/nbt.2377

Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, Wishart DS, Xia J (2018) MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Res 46:W486–W494. https://doi.org/10.1093/nar/gky310

Cronin G, Hay ME (1996) Induction of Seaweed Chemical Defenses by Amphipod Grazing. Ecology 77: 2287-2301. https://doi.org/10.2307/2265731

Cruz-Rivera E, Paul VJ (2002) Coral reef benthic cyanobacteria as food and refuge: diversity, chemistry and complex interactions. In: Proceedings of 9th international coral reef symposium, vol 1, pp 515–520

Cruz-Rivera E, Paul VJ (2006) Feeding by coral reef mesograzers: algae or cyanobacteria?. Coral Reefs 25: 617-627. https://doi.org/10.1007/s00338-006-0134-5

Dixson DL, Hay ME (2012) Corals Chemically Cue Mutualistic Fishes to Remove Competing Seaweeds. Science 338:804-807. https://doi.org/10.1126/science.1225748

Elliott JK, Elliott JM, Mariscal RN (1995) Host selection, location, and association behaviors of anemonefishes in field settlement experiments. Mar Biol 122:377-389. https://doi.org/10.1007/BF00350870

Engene N, Rottacker EC, Kaštovský J, Byrum T, Choi H, Ellisman MH, Komárek J, Gerwick WH (2012) *Moorea producens* gen. nov., sp. nov. and *Moorea bouillonii* comb. nov., tropical marine cyanobacteria rich in bioactive secondary metabolites. Int J Syst Evol Microbiol 62:1171–1178. https://doi.org/10.1099/ijs.0.033761-0

Fishelson L (1966) Observations on the littoral fauna of Israel, V. on the habitat and behaviour of *Alpheus frontalis* H. Milne-Edwards (Decapoda, Alpheidae). Crusteceana 11:98-104. https://doi.org/10.1163/156854066X00496

Gómez-Lemos LA, Doropoulos C, Bayraktarov E, Diaz-Pulido G (2018) Coralline algal metabolites induce settlement and mediate the inductive effect of epiphytic microbes on coral larvae. Sci Rep 8:17557. https://doi.org/10.1038/s41598-018-35206-9

Guo CC, Hwang JS, Fautin DG (1996) Host selection by shrimps symbiotic with sea anemones: A field survey and experimental laboratory analysis. J Exp Mar Biol Ecol 202: 165-176. https://doi.org/10.1016/0022-0981(96)00020-2

Hay ME, Duffy JE, and Fenical W (1990) Host-Plant Specialization Decreases Predation on a Marine Amphipod: An Herbivore in Plant's Clothing. Ecology 71:733-743. https://doi.org/10.2307/1940326

Hay ME (2009) Marine Chemical Ecology: Chemical Signals and Cues Structure Marine Populations, Communities, and Ecosystems. Ann Rev Mar Sci 1: 193–212. https://doi.org/10.1146/annurev.marine.010908.163708

Holman JD, Tabb DL, Mallick P (2014) Employing ProteoWizard to Convert Raw Mass Spectrometry Data. Curr Protoc Bioinformatics 46:1–9. https://doi.org/10.1002/0471250953.bi1324s46

Huang WJ, Cheng YL, Cheng BL (2008) Ozonation By-products and Determination of Extracellular Release in Freshwater Algae and Cyanobacteria. Environ Eng Sci 25:139-151. https://doi.org/10.1089/ees.2006.0113

Huang WJ, Lai CH, Cheng YL (2007) Evaluation of extracellular products and mutagenicity in cyanobacteria cultures separated from a eutrophic reservoir. Sci Total Environ 377: 214-223. https://doi.org/10.1016/j.scitotenv.2007.01.075

Leber CA, Naman CB, Keller L, Almaliti J, Caro-Diaz EJE, Glukhov E, Joseph V, Sajeevan TP, Reyes AJ, Biggs JS, Li T, Yuan Y, He S, Yan X, Gerwick WH (2020) Applying a Chemogeographic Strategy for Natural Product Discovery from the Marine Cyanobacterium *Moorena bouillonii*. Mar Drugs 18:515. https://doi.org/10.3390/md18100515

Lindquist N, Barber PH, Weiszl JB (2005) Episymbiotic microbes as food and defence for marine isopods: unique symbioses in a hostile environment. Proc Biol Sci 272:1209–1216. https://doi.org/10.1098/rspb.2005.3082

Matthew S, Schupp PJ, Luesch H (2008) Apratoxin E, a Cytotoxic Peptolide from a Guamanian Collection of the Marine Cyanobacterium *Lyngbya bouillonii*. J Nat Prod 71:1113–1116. https://doi.org/10.1021/np700717s

Nevitt GA (2008) Sensory ecology on the high seas: the odor world of the procellariiform seabirds. J Exp Biol 211:1706-1713. https://doi.org/10.1242/jeb.015412

Nocchi N, Soares AR, Souto ML, Fernández JJ, Martin MN, Pereira RC (2017) Detection of a chemical cue from the host seaweed *Laurencia dendroidea* by the associated mollusc *Aplysia brasiliana*. PLOS One 12:e0187126. https://doi.org/10.1371/journal.pone.0187126

Pawlik JR (1986) Chemical induction of larval settlement and metamorphosis in the reef-building tube worm *Phragmatopoma californica* (Sabellariidae: Polychaeta). Mar Biol 91:59–68. https://doi.org/10.1007/BF00397571

Pawlik JR (1989) Larvae of the sea hare *Aplysia californica* settle and metamorphose on an assortment of macroalgal species. Mar Ecol Prog Ser 51:195-199. https://doi.org/10.3354/meps051195

Pereira RC, Nocchi N, Souto ML, Fernández JJ, Norte M, Duarte HM, Soares AR (2020) The sea-hare *Aplysia brasiliana* promotes induction in chemical defense in the seaweed *Laurencia dendroidea* and in their congeneric neighbors. Plant Physiol Biochem154:295-303. https://doi.org/10.1016/j.plaphy.2020.05.020

Pluskal T, Castillo S, Villar-Briones A, Orešič M (2010) MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11:395. https://doi.org/10.1186/1471-2105-11-395

Saha M, Berdalet E, Carotenuto Y, Fink P, Harder T, John U, Not F, Pohnert G, Potin P, Selander E, Vyverman W, Wichard T, Zupo V, Steinke M (2019) Using chemical language to shape future marine health. Front Ecol Environ 17:530-537. https://doi.org/10.1002/fee.2113

Scott A, Dixson DL (2016) Reef fishes can recognize bleached habitat during settlement: sea anemone bleaching alters anemonefish host selection. Proc Royal Soc B 283:20152694. https://doi.org/10.1098/rspb.2015.2694

Selander E, Berglund EC, Engström P, Berggren F, Eklund J, Harðardóttir S, Lundholm N, Grebner W, Andersson MX (2019) Copepods drive large-scale trait-mediated effects in marine plankton. Sci Adv 5:eaat5096. https://doi.org/10.1126/sciadv.aat5096

Selander E, Kubanek J, Hamberg M, Andersson MX, Cervin G, Pavia H (2015) Predator lipids induce paralytic shellfish toxins in bloom-forming algae. Proc Natl Acad Sci USA 112:6395-6400. https://doi.org/10.1073/pnas.1420154112

Sneed JM, Sharp KH, Ritchie KB, Paul VJ (2014) The chemical cue tetrabromopyrrole from a biofilm bacterium induces settlement of multiple Caribbean corals. Proc Royal Soc B 281: 20133086. https://doi.org/10.1098/rspb.2013.3086

Souza J, Barroso D, Hirose GL (2019) Chemical recognition in the symbiotic pea crab *Dissodactylus crinitichelis* (Crustacea: Decapoda: Pinnotheridae): host and conspecific cues. J Exp Mar Biol Ecol 511:108-112. https://doi.org/10.1016/j.jembe.2018.12.002

Stachowicz JJ, Hay ME (1999) Reducing Predation Through Chemically Mediated Camouflage: Indirect Effects of Plant Defenses on Herbivores. Ecology 80:495-509. https://doi.org/10.1890/0012-9658(1999)080[0495:RPTCMC]2.0.CO;2

Yasmeen A, Qasim M, Ahmed A, Uddin N, Ahmed Z, Ali MS, Rasheed M (2018) GC-MS and Antioxidant Studies on Botanicals from *Sargassum wightii*: Natural Product Study Revealing Environmental Contaminants. J Chem Soc Pakistan 40:201-212

Zimmer RK, Ferrier GA, Kim SJ, Kaddis CS, Zimmer CA, Loo JA (2016) A multifunctional chemical cue drives opposing demographic processes and structures ecological communities. Ecology 97:2232–2239. https://doi.org/10.1002/ecy.1455

Chapter 4: *Moorena bouillonii* Chemogeography: Distributional Patterns of Compounds and Compound Families at Multiple Spatial Scales

4.0: Abstract

Geographic location of collection is a factor known to be associated with chemodiversity in the cyanobacterium *Moorena bouillonii*. However, *M. bouillonii* chemogeography, and chemogeography in general, is understudied. Forty-nine collaboratively-sourced potential *M. bouillonii* crude extracts were profiled via liquid chromatography – tandem mass spectrometry (LC-MS/MS). The resultant dataset was assessed for patterns in molecular features in aggregate, as well as distributional patterns of known molecular families. These analyses revealed that both shrimp-association and location of origin influenced chemical composition, but chemogeographic relationships were no longer evident when examined at intra-island scale. Compound families exhibited unique distributions, with some tending towards more regional specificity. Mining of the rich metabolomics dataset yielded a list of priority molecular features for future natural products discovery efforts in *M. bouillonii*.

4.1: Introduction

Chemical diversity is of great importance to the fields of natural product drug discovery and chemical ecology. Natural product drug discovery operates on the general strategy of accessing new chemical entities by examining previously unstudied or understudied biological entities (Singh and Pelaez 2008; Naman et al 2017); when previously unstudied taxa are investigated, there is potential to unearth new compounds with high structural novelty (Pye et al 2016). When studying biological interactions through a chemical ecological lens, knowledge of the chemical diversity present can inform the work in tangible and intangible ways. Having a good working knowledge of the diversity of chemical substances in a particular system can facilitate the identification of multiple active constituents that are working additively or synergistically to produce specific biological outcomes (e.g. Selander et al 2015). This knowledge can also simply reveal, and so allow an appreciation for, the complexity of the chemical environment in which the biological interaction is taking place.

Small molecule chemical diversity can be assessed in multiple different ways, with each strategy offering benefits and shortcomings. Genomic approaches allow for the study of the biosynthetic diversity or the potential chemical diversity in an organism by providing information about what biosynthetic enzymes, and by extension, depending on the clarity of the particular biosynthetic logic, what small molecule products, are encoded in the organism's DNA (Navarro-Muñoz et al 2019). Silent biosynthetic gene clusters (BGCs) are BGCs that are not expressed under standard conditions, and so represent a gap between the potential chemical diversity described in an organism's genome, and the practical chemical diversity being produced (Rutledge and Challis 2015). Another identifiable deficit, leading in this case to an underestimate of practical chemical diversity, is that genomic information may not reveal the multitude of analog structures that can be synthesized from a single BGC, resulting from biosynthetic enzyme substrate promiscuity (Pandya et al 2014; Chevrette et al 2020) or noncolinear pathways (Moss et al 2019) . Transcriptomic and proteomic approaches can also provide insights into the potential small molecular diversity of an organism, improving upon genomics as a proxy for practical chemical diversity in that these methods capture what genomic information is being transcribed and translated towards producing that actual chemical diversity. However, these methods are still indirect measurements of chemical diversity, and mechanisms such as regulation via post-translational modification and insufficient substrate pools can interrupt small molecule production (Fernie and Stitt 2012).

Metabolomics, particularly via mass spectrometry methods, allows for the observation and measurement of the practical chemical diversity in an organism (Luzzatto-Knaan et al 2017; Cantrell et al 2019). In more directly assessing actual (detectable) chemical diversity, metabolomics necessarily fails to define the full potential of an organism’s chemical diversity; metabolomics data presents only a snapshot of chemical entities detectable at a particular place, time, and biological and environmental condition in an organism. To make up for some of this deficit, researchers have conducted comparative metabolomics studies to include organisms collected across various geographical regions (Luzzatto-Knaan et al 2017; Leber et al 2020; Maloney et al 2020) as well as studies of organisms cultured under wide arrays of conditions (Bode et al 2002; Adpressa and Loesgen 2016; Crnkovic et al 2018), co-cultured with other organisms (Serrano et al 2017) or exposed to chemical elicitors (Okada et al 2016).

Another challenge with mass spectrometry-based metabolomics is that even if all of the chemical diversity in an organism can be measured, the information collected about each chemical entity is limited and coded. A standard liquid chromatography mass spectrometry

(LCMS) experiment provides a mass-to-charge ratio (*m/z*) and a chromatographic method dependent retention time for each feature captured; isotopic patterns can provide additional insights about elemental composition, but nothing comprehensive. More detailed structural information can be accessed by running tandem mass spectrometry experiments that produce MS^2 data. Tools such as the Global Natural Products Social molecular networking (GNPS) platform can leverage this data to cluster molecular features based on the structural similarity represented in their MS^2 fragmentation spectra and to annotate features with sufficient MS^2 spectral similarity to database entries (Wang et al 2016). Recent advances allow for an even greater depth and breadth of chemical diversity characterization from MS^2 data via extended annotations; DEREPLICATOR+ putatively identifies molecular features based on similarity to predicted fragmentation data from known compounds (Mohimani et al 2018), Network Annotation Propagation constrains *in silico* annotation predictions to improve utility and accuracy (da Silva et al 2018), and MS2LDA allows the assignment of substructures to molecular features (van der Hooft et al 2016). Some or all of these approaches can additionally be combined using MolNetEnhancer to overlay the various annotations and to classify molecular families (Ernst et al 2019).

It has been well established that accessing taxonomically novel organisms is an effective strategy for accessing new and novel chemical diversity (Pye et al 2016; Hoffmann et al 2018). However, there is still much to be learned, particularly in marine systems (Noyer, Thomas, and Becerro 2011), about the expansive chemical diversity that can exist within a single species. In Leber et al (2020), geographically distinct collections of the natural product rich marine cyanobacterium *Moorena bouillonii* (L.Hoffmann & Demoulin) Engene & Tronholm 2019 (Oscillatoriaceae) were subjected to chemogeographic analyses that directed the annotation of a family of regionally specific small molecules, the doscadenamides. The geographic patterns made visible by this previous work opened up a Pandora's box of questions about chemical diversity in *M. bouillonii*. How is small molecule production distributed across the species? Are there constitutive, regionally specific, and sample specific metabolites? How do patterns in chemogeography look at different spatial scales? How are known compound families distributed across geographical space, and do individual analogs tend to be produced in the same or different locations? What other patterns in the distribution of natural product families produced by *M. bouillonii* are there to be found? These questions could not be comprehensively answered in the previous work due to the relatively small sample set and the use of legacy data that precluded the use of more sophisticated analysis and annotation tools. In the manuscript chapter that follows, an extended, uniformly acquired, high quality LC-MS/MS dataset of broadly collected *M. bouillonii* crude extracts was explored in depth to examine the questions above and learn more about intraspecific chemical diversity in *M. bouillonii*. We report investigations into the distributions of molecular features and the revelations they provide about shared metabolites in aggregate, as well as our inspections of known compound families, which illustrate how analogs cooccur or define specific regions. We also translate these chemogeographic analyses into actionable discovery leads, by generating a priority list of major molecular features representing potentially new natural products.

4.2: Results & Discussion

4.2.1: MS^1 sample set overview

Across the 49 cyanobacterial samples analyzed in this study, 14,067 MS^1 molecular features were detected. Of these, 5,369 features (38.2%) were detected only in single samples, making them sample-specific; no features were found to be ubiquitous across the entire sample set (Figure 4.S1). Features detected per sample ranged from 176 to 3,320 (Figure 4.S2), and sample-specific features per sample ranged from 1 to 482 (Figure 4.S3). Sample-specific features must be considered with caution, as their origins are not known. A sample-specific feature, in the most interesting scenario, may be indicative of a unique compound that is only produced by the collected organism due to either specific biosynthetic capacity or inimitable environmental conditions present at collection. A sample-specific feature could also represent a peak not optimally aligned during preprocessing, a noise peak not properly filtered, a researcher-introduced contaminant acquired between sample collection and analysis, or an environmentally introduced contaminant courtesy of biotic interactions or abiotic exposures prior to collection. In the current study, care was taken during data preprocessing to limit instances of the first two possibilities, but it must be acknowledged that knowledge of the origins of these sample-specific features is limited. In the case of features occurring across samples, which will be the major focus of the analyses to follow, multiple detections within the dataset lend support that a feature is indeed associated with a subset of *M. bouillonii* samples, rather than being process derived. This, however, only suggests attribution of features to particular *M. bouillonii* holobionts and does not allow for direct attribution as being a product of *M. bouillonii*.

Viewing each feature count, sample-specific feature count, and maximum peak area plotted against each other (Figures 4.S4, 4.S5, 4.S6) allowed for an overview of these samples along these dimensions, and provided an opportunity to uncover problems related to data acquisition and processing. Inspecting number of features versus maximum peak area (Figure 4.S4) suggested that five samples with maximum peak areas below 10^7 and feature counts below 700 *may* have benefited from analysis at higher concentration, but the overall trend displayed is a very weak positive association between maximum peak area and feature count. Similarly, weak positive associations can be seen for sample-specific features versus feature counts (Figure 4.S5) and sample-specific features versus maximum peak area (Figure 4.S6). In both plots, it is illustrated that most samples contained 100 or less sample-specific features, but some samples with feature counts greater than 1,000 or maximum peak areas greater than 10^7 were found to have considerably more sample-specific features. While no detrimental systemic issues were detected, these plots suggest that some degree of under detection of features may be occurring, specifically in samples with lower maximum peak areas. Efforts were made during data preprocessing to combat this situation via feature gap-filling, but the issue in part remains. Conduction of the analyses that follow was informed by the potential for feature under detection.

4.2.2: Predominant patterns in similarity

In Leber et al (2020), hierarchical clustering based on cosine distance of MS^1 feature vectors of a small sample set of *M. bouillonii* crude extracts yielded clusters based on geographical origin. Conducting the same analysis on the 49 samples featured in this study resulted in a similar outcome, but with additional layers of complexity (Figure 4.1). All samples from the Lakshadweep Islands and the Paracel Islands clustered together, along with some samples from Palmyra Atoll and American Samoa. Samples from Guam, which make up the majority of samples in this study, were split amongst three different clusters, the third of which (denoted as C in Figure 4.1) also contained samples from the nearby island of Saipan. The cophenetic correlation was 0.866, indicating that the dendrogram is a fair representation of the underlying cosine distances calculated pairwise across the samples. In spite of the splitting of Guam samples across three clusters, and the lack of clustering of samples from certain regions, particularly Papua New Guinea, samples from the same collection region were found to be statistically significantly more similar to each other in terms of cosine distance (t statistic=-11.27, p-value < 0.001), number of shared features (t statistic=9.86, p-value < 0.001), and proportion of shared features (t statistic=13.04, p-value < 0.001), as compared to samples not from the same region.

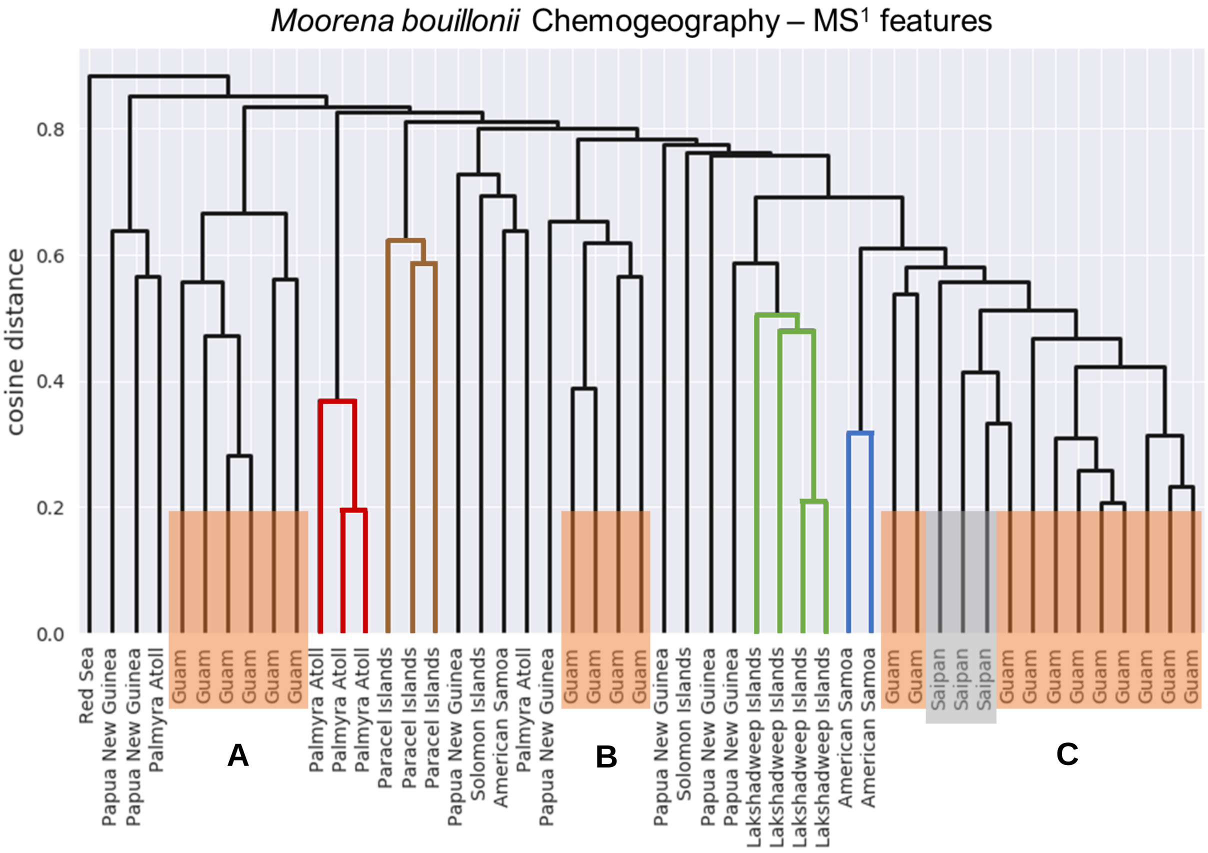


Figure 4.1 – ORCA dendrogram depicting the results of hierarchical clustering of *M. bouillonii* crude extracts. Clustering based on geography is evident, especially for samples from the Lakshadweep Islands (green), the Paracel Islands (brown), and, to a lesser extent, Palmyra Atoll (red) and American Samoa (blue). This suggests that samples are more chemically similar when they come from the same region. Samples from Guam are split across three clusters (A-C, denoted in orange); these three clusters suggest that additional factors are influencing chemical composition similarity. Cluster A is dominated by non-shrimp-associated samples of *M. bouillonii.* cluster B contains both shrimp-associated and non-shrimp-associated samples, as well as both cultured and environmental samples. Cluster C is all shrimp-associated environmental samples, and is interspersed with samples from Saipan (denoted in grey). The cophenetic correlation of the dendrogram is 0.866.

A closer inspection of the three clusters of samples from Guam helped reveal what other factors may be driving this disruption of a uniform Guamanian cluster. Of the six samples in Guam cluster A and the four samples in cluster B, five and two of the samples were from non-shrimp-associated collections of *M. bouillonii*, respectively. In contrast, all of the samples in cluster C, including those from Saipan, were from collections of *Alpheus frontalis* H. Milne Edwards 1837 (Alpheidae) associated *M. bouillonii*. The separation of cluster A from cluster B appears to be driven by the influence of culture vs collection, as cluster B contains both Gerwick Laboratory cultures of *M. bouillonii* from Guam, and the cluster is most closely associated with a sample from Papua New Guinea (featured in Figure 4.1 directly to the left of the cluster) that is also a Gerwick Laboratory culture.

The strength of shrimp-association in driving clustering may be partially obscured when visualized by dendrogram, but it becomes much more apparent when observed from an alternative viewpoint, namely Principal Component Analysis (PCA).[4] As can be seen in Figure 4.2A, geographical clustering is generally apparent when visualized via PCA, with three clusters of Guam samples distributed across the first principal component. Two clusters are in close proximity on the right side of the figure, while one cluster is separate from all other points on the PCA plot, in the upper left corner. By switching lenses to view the dataset from the point of view of shrimp association (Figure 4.2B), it is revealed that this distant cluster is composed on Guamanian samples of *M. bouillonii* that were not found to be associated with the shrimp symbiont *A. frontalis*. It should be noted that one non-shrimp-associated sample can be seen in the lower left of the plot, separated from the other non-shrimp samples. This is a non-shrimp-associated *M. bouillonii* sample from the Solomon Islands, suggesting that geographical differences in chemical composition are not limited to shrimp-associated samples, but also extend to non-shrimp-associated samples.

[4] PCA assumes normally distributed variables; this assumption is consistently violated in sparse metabolomics datasets. In spite of violated assumptions, PCA has been shown to have utility in metabolomics applications (e.g. Chanana et al 2017), particularly as a hypothesis generating tool. Usage of PCA with metabolomics data should be approached cautiously, with thoughtful and thorough consideration of how data should be preprocessed in order to visualize relationships of interest, and an understanding of its limitations.

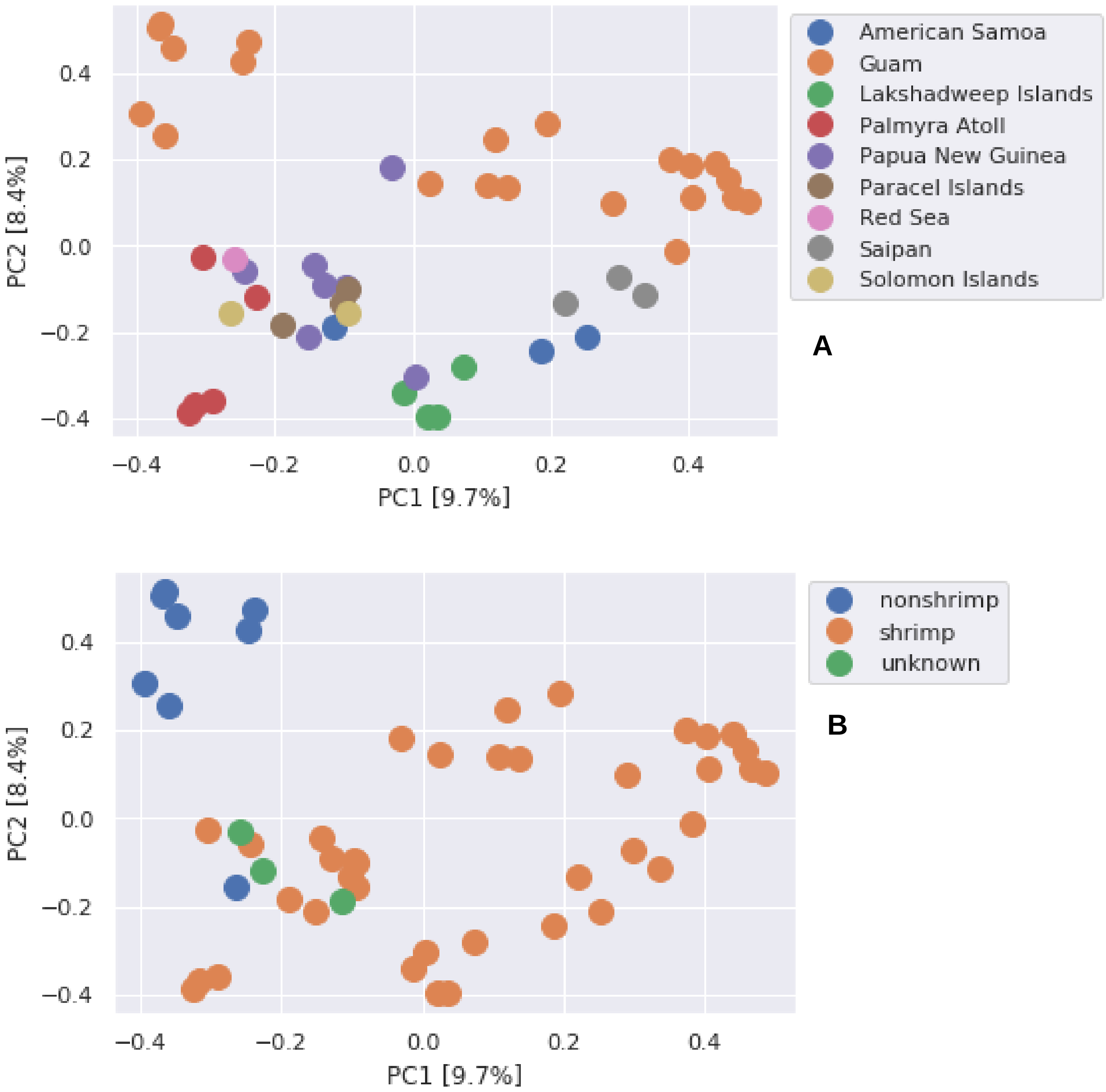


Figure 4.2 – Principal component analysis plots of *M. bouillonii* crude extracts, colorized based on A) collection region and B) shrimp-association.

<u>4.2.3: Features driving patterns in similarity</u>

In establishing both geographical location of collection and shrimp association as factors associated with similarity in sample chemical composition, relationships between samples can be more clearly delineated through the selection of relevant features. Samples were binned based on collection location, with samples from Guam additionally binned based on shrimp association. Features were then selected based on being shared by all samples within at least one assigned bin; features that are shared by the majority of samples within a bin, with 'majority' defined by a cutoff value between 0.50 and 1.00, will henceforth be referred to as the core features of that bin. Features that were not core to at least one bin (cutoff value of 1.00, in this case) were dropped. This reduced the set of considered features down to 1,175. Figure 4.3 shows the result of hierarchical clustering of the feature vectors distilled by feature selection. This same analysis was conducted with more flexibility to accommodate the potential under detection of features and produced a very similar result (Figure 4.S7). The two most notable changes to this dendrogram, as compared Figure 4.1, are as follows. First, Guam samples are now split across two clusters, with one containing all non-shrimp-associated samples and the other containing all shrimp-associated samples. This reaffirms shrimp association as a driving factor in *M. bouillonii* chemical composition. Secondly, samples from Saipan now appear further mixed with shrimp-associated Guam samples, with two samples from American Samoa also appearing to be closely chemically related. This suggests that distinct geographically driven clusters may represent too rigid of a framework for understanding how chemical diversity is distributed across samples of *M. bouillonii*. Instead, as geographical scale (or other impediments to the mixing of cyanobacterial populations) decreases, chemical compositional gradients may be truer to reality. In Leber et al (2020), hierarchical clustering of a small sample set of *M. bouillonii* crude extracts produced distinct clusters of samples from Guam, Saipan, and American Samoa. In this current study, when more samples are considered and so more variability is introduced, the lines between geographical clusters begin to blur. Chemical compositional similarity is still clearly associated with location of collection.

However, samples from neighboring regions, such as Guam and Saipan, can ‘look’ very similar, in a chemical compositional sense.

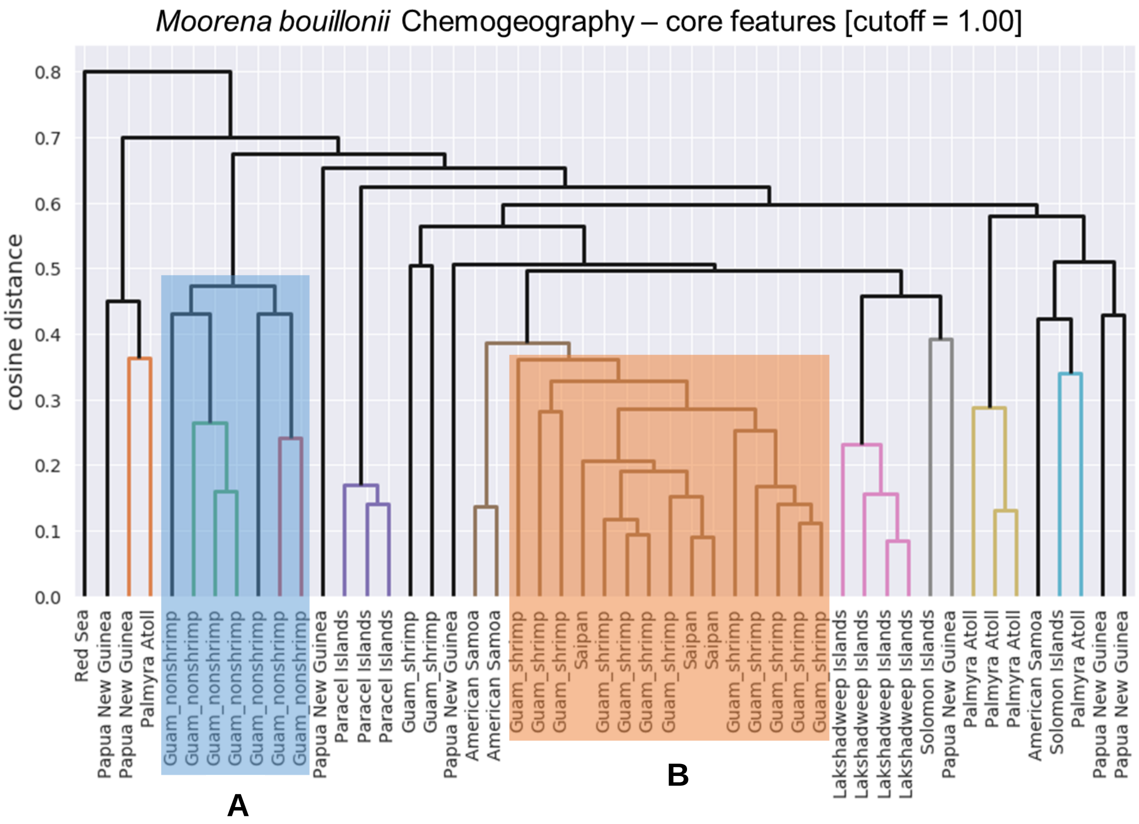


Figure 4.3 – ORCA dendrogram depicting the results of hierarchical clustering of *M. bouillonii* crude extracts, after feature selection for core molecular features per region and, for Guam samples, shrimp-association (cutoff = 1.00). Feature selection further clarifies patterns in the dataset, rearranging Guam samples into two clusters: A) non-shrimp associated, and B) shrimp-associated (including shrimp-associated samples from Saipan). The cophenetic correlation of the dendrogram is 0.877.

The cutoff value in determining core molecular features per bin was eased back to 0.75, indicating that at least 3 of 4 samples in a region would have to contain a feature for it to be counted as a core feature. This relaxed cutoff was used to generate more inclusive counts of core and sample-specific core features for each regional bin, and to compensate for potential issues of under detection. The results of these counts are summarized in Table 4.1. The counts were also conducted with a cutoff value of 1.00, resulting in smaller counts but supporting the same patterns (Table 4.S1). Samples from the Lakshadweep Islands, the Paracel Islands, Saipan, and shrimp-associated samples from Guam all contained large numbers of core features, indicating a higher degree of chemical compositional overlap and consistency within these individual groupings of samples. However, when one considers regionally-specific core features (core features not detected in any samples outside that region), the Paracel Islands and the Lakshadweep Islands outshine all other regional clusters (excluding Red Sea, as it is cluster of n = 1) by an order of magnitude, setting them a part as much more uniquely chemical diverse than the other regions examined. Whereas the Lakshadweep Islands and the Paracel Islands were recorded to possess 74 and 69 regionally specific core features, shrimp-associated samples from Guam were only recorded to possess 2 specific core features. Saipan had zero recorded specific core features, meaning that all of the 473 features found to be core in samples from Saipan were detected in other samples throughout the dataset. With American Samoa also only recording 2 regionally specific core features, this lends additional support to the conclusions arising from the hierarchical clustering that clusters of chemically similar samples organized by geographical origin begin to dissolve and homogenize as geographical distance and other barriers are reduced.

Table 4.1 – Counts of features, core features, and specific core features, per sample bin. Core feature counts were generated with a cutoff of 0.75.

	mean – feat./sample	**std – feat./sample**	**core feat.**	**specific core feat.**	**sample size**
American Samoa	878.67	181.68	197	2	3
Guam_non-shrimp	1564.29	1014.42	111	3	7
Guam_shrimp	1468.21	846.68	363	2	14
Lakshadweep Islands	1005.25	260.55	569	74	4
Palmyra Atoll	762.40	328.67	96	4	5
Papua New Guinea	897.71	525.25	33	0	7
Paracel Islands	1399.67	577.11	307	69	3
Red Sea	176.00	0.00	176	35	1
Saipan	1399.00	185.81	473	0	3
Solomon Islands	846.50	306.50	156	2	2

Another way to look at chemical compositional relatedness across geographical groups is to assess the degree of overlap between sets of core features (cutoff = 0.75, Table 4.2; cutoff = 1.00, Table 4.S2). Large portions of the core features of the Lakshadweep Islands register as core features in American Samoa (17%), shrimp-associated Guam samples (24%), the Paracel Islands (19%), and Saipan (31%). Similarly, many of core features of Saipan are shared with American Samoa (20%), shrimp-associated Guam samples (37%), the Lakshadweep Islands (37%), and the Paracel Islands (23%). In fact, all sample groupings can be seen to substantially overlap in their core features with other groups, with even the sample from the Red Sea, which is displayed as the most chemically distant sample in both dendrograms (Figures 4.1 and 4.3), has 20% of its core features co-occurring with the core features for Saipan. This heavy overlap in core features across sample groupings suggests that even while no features were found to be ubiquitous across the sample set, there does appear to be a large degree of shared chemical entities. This large degree of chemical overlap makes even more impressive the unique diversity of features that are attributable to the Lakshadweep Islands and the Paracel Islands.

Table 4.2 - Counts and proportions of overlapping core features between sample region and shrimp-association bins (cutoff = 0.75). Proportions, shown in parentheses, were calculated by taking the number of core features that overlap between each pair of groups, and dividing by the total core features of the group listed in the column headings.

	Amer. Samoa	**Guam _n.s.**	**Guam _shr.**	**Lakshad. Islands**	**Palmyra Atoll**	**Papua New Guinea**	**Paracel Islands**	**Red Sea**	**Saipan**	**Solom. Islands**
Total core features	**197**	**111**	**363**	**569**	**96**	**33**	**307**	**176**	**473**	**156**
American Samoa	**197**	24	78	99	48	15	38	23	95	38
	(1.00)	(0.22)	(0.21)	(0.17)	(0.50)	(0.45)	(0.12)	(0.13)	(0.20)	(0.24)
Guam _non-shrimp	24	**111**	32	27	14	11	23	14	28	17
	(0.12)	**(1.00)**	(0.09)	(0.05)	(0.15)	(0.33)	(0.07)	(0.08)	(0.06)	(0.11)
Guam _shrimp	78	32	**363**	139	30	28	70	26	173	39
	(0.40)	(0.29)	**(1.00)**	(0.24)	(0.31)	(0.85)	(0.23)	(0.15)	(0.37)	(0.25)
Lakshadweep Islands	99	27	139	**569**	51	24	103	28	177	68
	(0.50)	(0.24)	(0.38)	**(1.00)**	(0.53)	(0.73)	(0.34)	(0.16)	(0.37)	(0.44)
Palmyra Atoll	48	14	30	51	**96**	10	16	16	33	33
	(0.24)	(0.13)	(0.09)	(0.09)	**(1.00)**	(0.30)	(0.05)	(0.09)	(0.07)	(0.21)
Papua New Guinea	15	11	28	24	10	**33**	17	16	27	18
	(0.08)	(0.10)	(0.08)	(0.04)	(0.10)	**(1.00)**	(0.06)	(0.09)	(0.06)	(0.12)
Paracel Islands	38	23	70	103	16	17	**307**	16	107	25
	(0.19)	(0.21)	(0.19)	(0.18)	(0.17)	(0.52)	**(1.00)**	(0.09)	(0.23)	(0.16)
Red Sea	23	14	26	28	16	16	16	**176**	36	35
	(0.12)	(0.13)	(0.07)	(0.05)	(0.17)	(0.48)	(0.05)	**(1.00)**	(0.07)	(0.22)
Saipan	95	28	173	177	33	27	107	36	**473**	51
	(0.48)	(0.25)	(0.48)	(0.31)	(0.34)	(0.82)	(0.35)	(0.20)	**(1.00)**	(0.33)
Solomon Islands	38	17	39	68	33	18	25	35	51	**156**
	(0.19)	(0.15)	(0.11)	(0.12)	(0.34)	(0.55)	(0.08)	(0.20)	(0.11)	**(1.00)**

4.2.4: Patterns at intra-island scale

Of the 49 samples included in this study, 21 were collected from 10 sites in 6 sub-regions around the island of Guam. This allowed examination of how chemical composition varies at a smaller geographical intra-island scale. Hierarchical clustering of samples from Guam and labeling by sub-region produced three clusters of samples that did not suggest similarity as being associated with sub-region (Figure 4.4 and 4.5). Even after feature selection, selecting for features present in the core features of at least one sub-region (cutoff = 1.00), no structure in the hierarchical clustering was revealed that suggested increased similarity between samples from the same sub-region (Figure 4.6 and 4.7). However, feature selection did again result in reorganization of the dendrogram, resulting in two clusters organized by shrimp association. In agreement with these visualizations, greater similarity within groups, when organized by subregion, was not found to be statistically significant when assessed by cosine distance (t statistic = 1.82, p-value = 0.965), number of shared features (t statistic = -1.32, p-value = 0.906), or proportion of shared features (t statistic = -1.69, p-value = 0.954). Greater similarity within groups organized by shrimp-associated, however, was found to be statistically significant when assessed by cosine distance (t statistic = -15.95, p-value = 3.37 x 10^{-36}), number of shared features (t statistic = 7.99, p-value = 9.43 x 10^{-12}), or proportion of shared features (t statistic = 16.34, p-value = 1.89 x 10^{-39}).

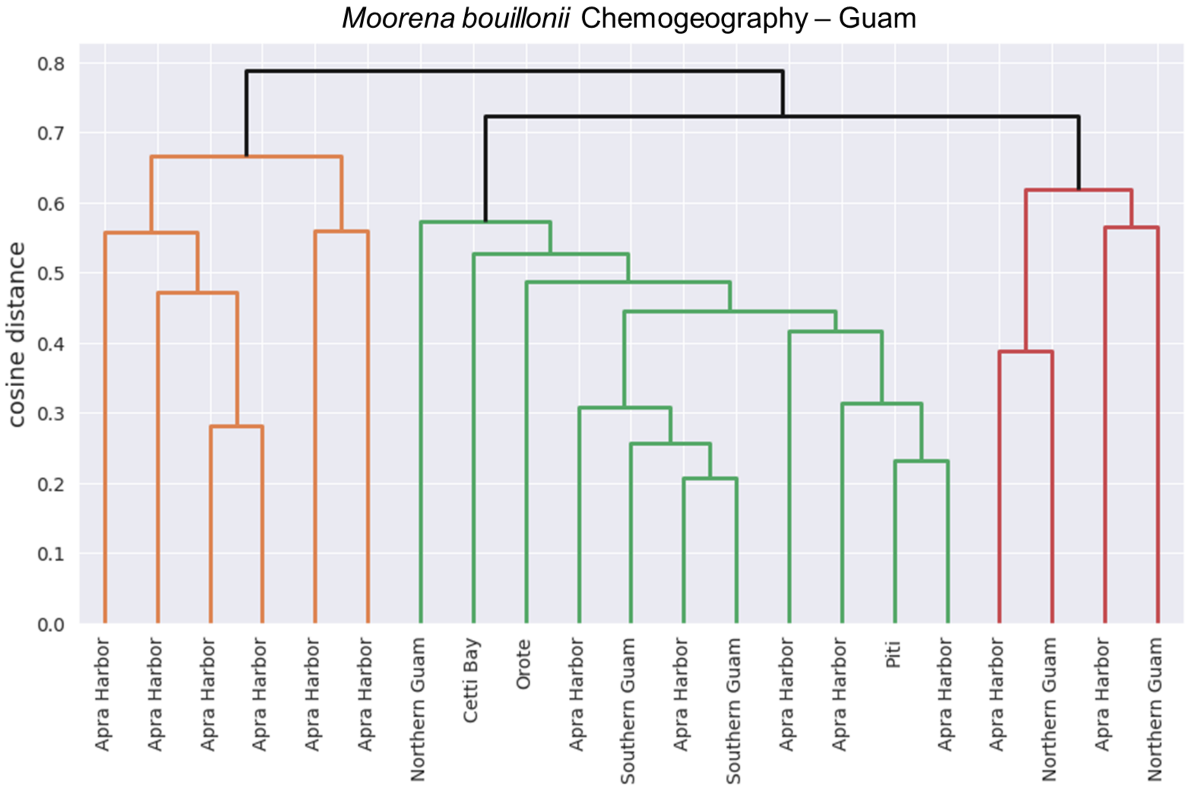


Figure 4.4 – ORCA dendrogram illustrating the results of hierarchical clustering of *M. bouillonii* crude extracts from Guam. Samples cluster into three groups, unrelated to the subregion that they were collected from. The green cluster contains only shrimp-associated environmental samples. The orange cluster contains mainly non-shrimp-associated samples and all environmental samples. The red cluster consists of half shrimp associated and half non-shrimp-associated as well as half culture and half environmental samples.

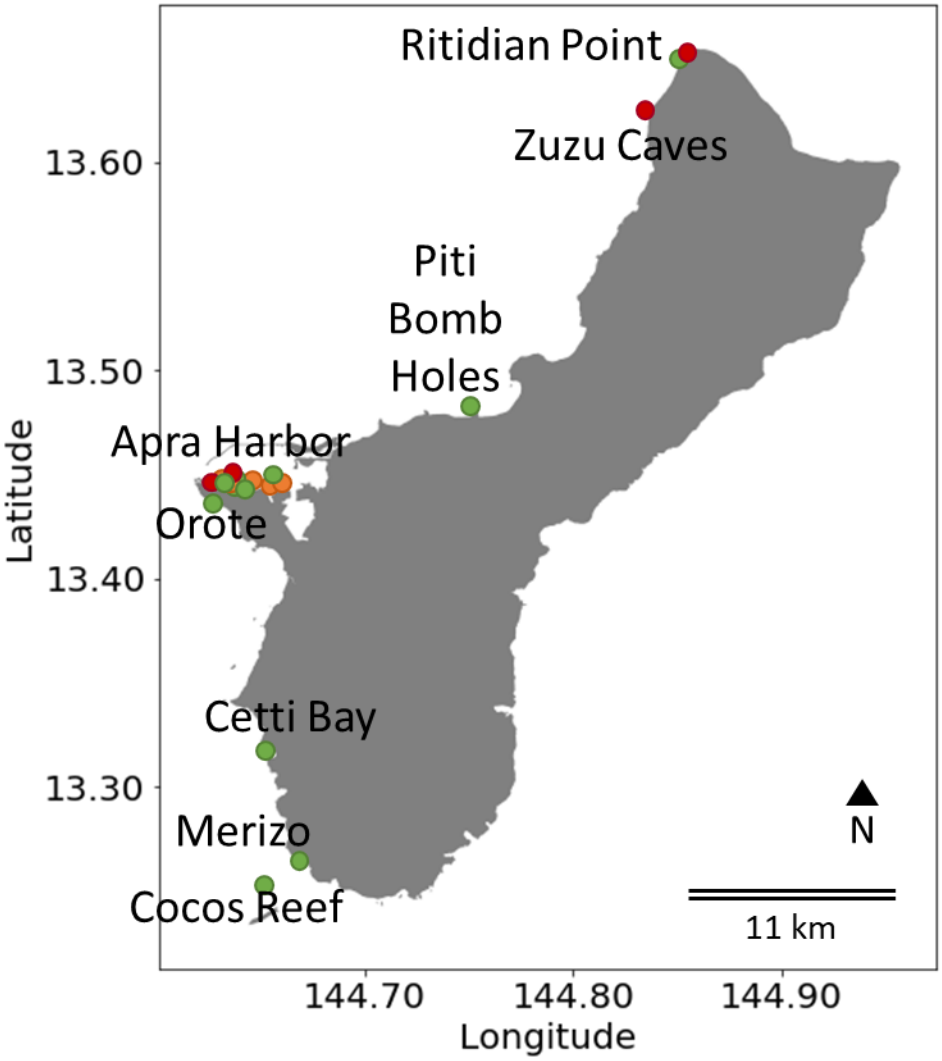


Figure 4.5 – Map of Guam, depicting locations of sample collection. Samples are colorized based on how they clustered in Figure 4.4. This view of the samples in geographic space suggests that location of samples collection does not explain chemical compositional similarity. The map was created using a NOAA-produced coastline shapefile (NOAA 2005).

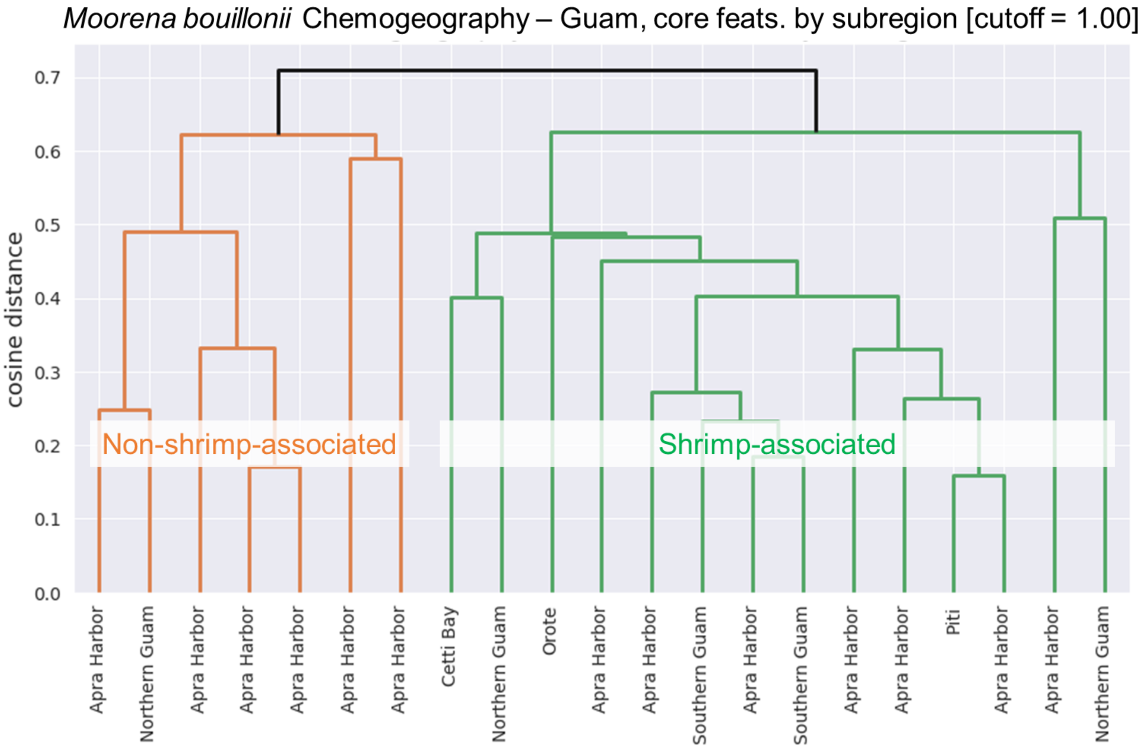


Figure 4.6 – ORCA dendrogram illustrating the results of hierarchical clustering of *M. bouillonii* crude extracts from Guam after feature selection by collection subregion. Feature selection results in a rearrangement of the samples into two clusters: one with all of the non-shrimp-associated samples, and one with all of the shrimp-associated samples. This suggests that shrimp-association is more correlated with similarity in chemical composition than location of collection at this intra-island scale.

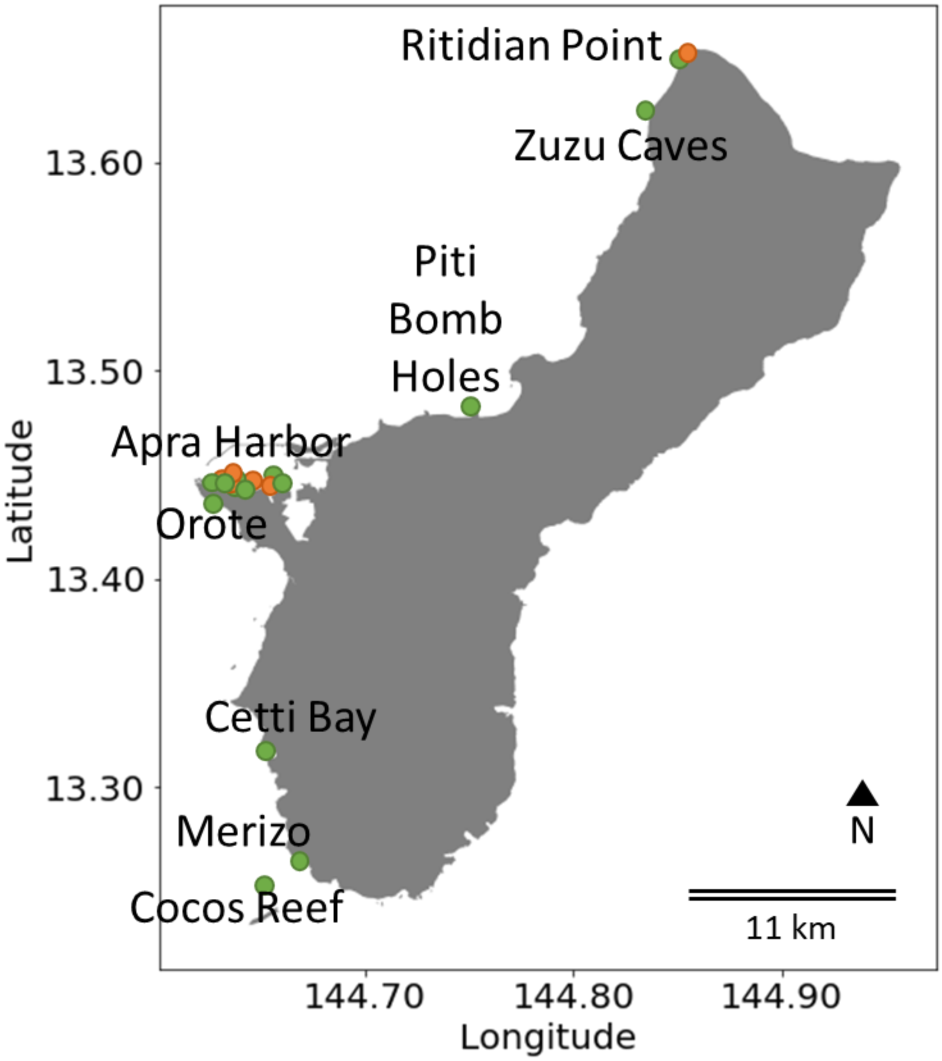


Figure 4.7 – Map of Guam, depicting locations of sample collection. Samples are colorized based on how they clustered in Figure 4.6, with orange samples corresponding to non-shrimp-associated *M. bouillonii* and green samples corresponding to shrimp-associated. This view of the samples in geographic space reiterates the conclusion that even after feature selection by subregion, hierarchical clustering did not reveal any structure in the data suggesting a relationship between chemical composition similarity and subregion of origin. The map was created using a NOAA-produced coastline shapefile (NOAA 2005).

4.2.5: Distributions of known compound families

Of the 14,067 MS^1 features previously discussed, 4,437 had matching MS^2 spectra, and so could be networked together in a feature-based molecular network (FBMN) (Nothias et al 2020) using GNPS (Wang et al 2016) (Figure 4.8). Combining FBMN with MS2LDA, DEREPLICATOR+, and NAP via MolNetEnhancer provided additional annotations for the dataset. In terms of general characterization of the chemical diversity present, these tools were limited in their efficacy, with MolNetEnhancer classifying only 20 out of the close to 300 molecular clusters. This is likely due to the limited degree with which annotations were made in the dataset, resulting from *M. bouillonii* natural products (and marine filamentous cyanobacterial natural products in general) being relatively poorly represented in the foundational databases on which these tools rely. In terms of identifying clusters corresponding to known compound families, a combination of FBMN spectral matching, DEREPLICATOR+ identifications, and manual exploration proved instrumental in identifying 5 known families: the apratoxins, the columbamides, the doscadenamides, the apramides, and the lyngbyapeptins. Each family displayed a unique distribution of how its member chemical species are differential produced across the geographical regions surveyed, as illustrated with relative quantitation by the pie charts overlaid by each node. The colors map to regional bins as follows: American Samoa = red; non-shrimp-associated Guam = orange; shrimp-associated Guam = yellow; Lakshadweep Islands = green; Palmyra Atoll = teal; Papua New Guinea = light blue; Paracel Islands = royal blue; Red Sea = purple; Saipan = pink; Solomon Islands = grey.

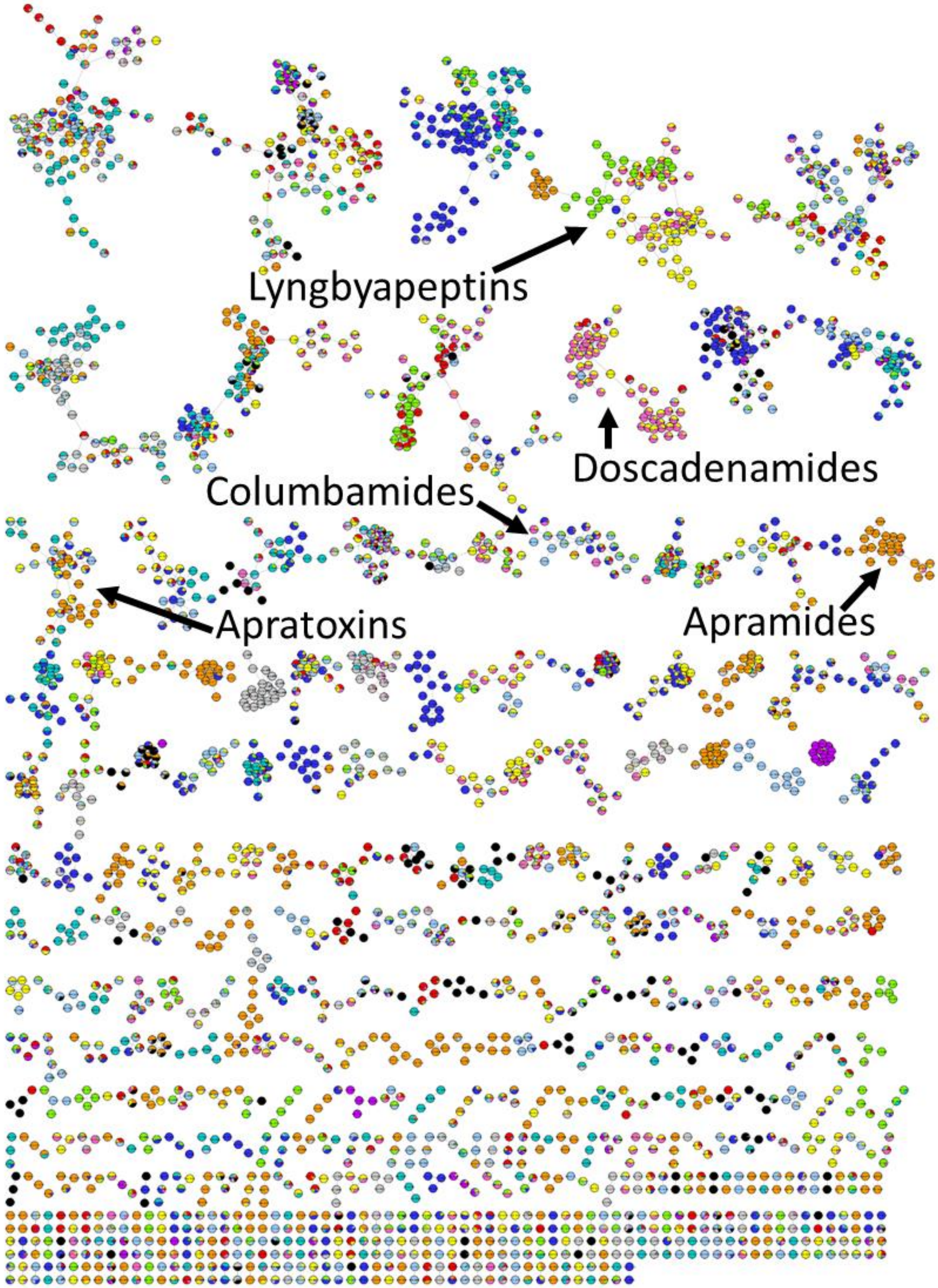


Figure 4.8 – Feature-based molecular network of *M. bouillonii* crude extracts, excluding singleton nodes. Relative quantification is illustrated by the pie chart overlay on each node. The colors map to bins as follows: American Samoa = red; non-shrimp-associated Guam = orange; shrimp-associated Guam = yellow; Lakshadweep Islands = green; Palmyra Atoll = teal; Papua New Guinea = light blue; Paracel Islands = royal blue; Red Sea = purple; Saipan = pink; Solomon Islands = grey; blanks = black.

Of the five molecular families identified in this dataset, the apratoxins (Figure 4.9) appear the most widely distributed, with the columbamides (Figure 4.10) in a close second. In fact, apratoxin A (denoted by *m/z* 840.4979) was detected from all regions, and is featured prominently in samples from American Samoa, Guam (shrimp-associated), Palmyra Atoll, Papua New Guinea, the Paracel Islands, and Saipan. Apratoxins B and C (denoted by the 2 nodes at *m/z* 826.4818 and 826.4826 also feature prominently across multiple regions. In contrast to the seeming ubiquity of some apratoxin analogs, two other subsets are apparent for their regional specificity. A relaxed gaze upon the cluster is characterized by opposing bursts of orange and, to a lesser extent, teal, highlighting suites of apratoxin analogs prominent in non-shrimp-associated Guam samples and Palmyra Atoll samples, respectively. Many of the predominantly orange nodes contain slivers of yellow, capturing the extensive overlap in natural products found between the shrimp and non-shrimp-associated samples from Guam. While they do have many chemical constituents in common, the differential production of those compounds, along with group-specific compounds, assists in driving two separate clusters of samples. The apratoxins are infamous for their cytotoxicity (Luesch et al 2001), leaving one to wonder whether the prominence of such a wide array of apratoxin analogs that are found at much lower relative levels in other samples, if detected at all, may have contributed to shrimp selection for other chemotypes.

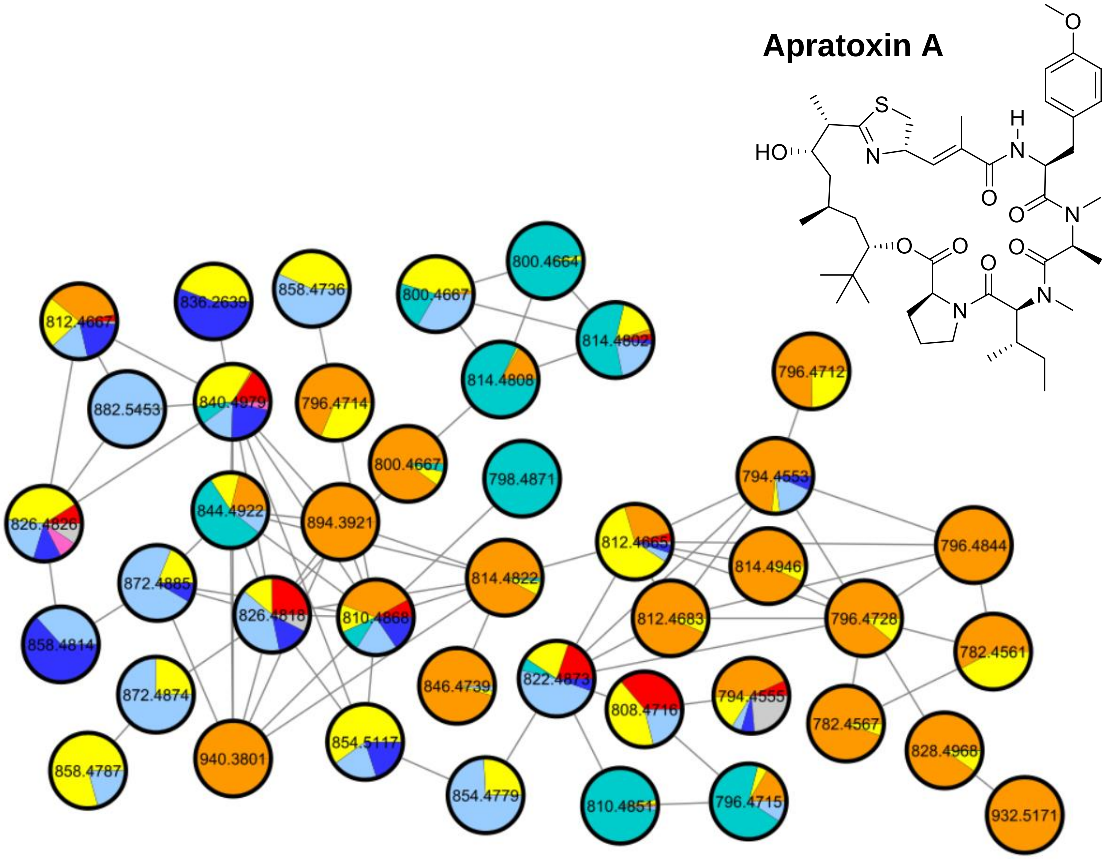


Figure 4.9 – Feature-based molecular network cluster of the apratoxins along with a representative structure from the compound family: apratoxin A. The pie charts, illustrating relative quantification, are colorized as follows: American Samoa = red; non-shrimp-associated Guam = orange; shrimp-associated Guam = yellow; Lakshadweep Islands = green; Palmyra Atoll = teal; Papua New Guinea = light blue; Paracel Islands = royal blue; Red Sea = purple; Saipan = pink; Solomon Islands = grey.

While the columbamides were not quite as ubiquitous in that they were not detected in the sample from the Red Sea, one could argue that they are generally more broadly distributed, with most analogs occurring across most regions sampled. Columbamides were found to be most prominent in samples from Papua New Guinea (light blue), followed by the Lakshadweep Islands (green) and the Paracel Islands (royal blue). Many columbamide analogs were also quite apparent in samples from Saipan (pink), with one analog being detected only in samples from Saipan.

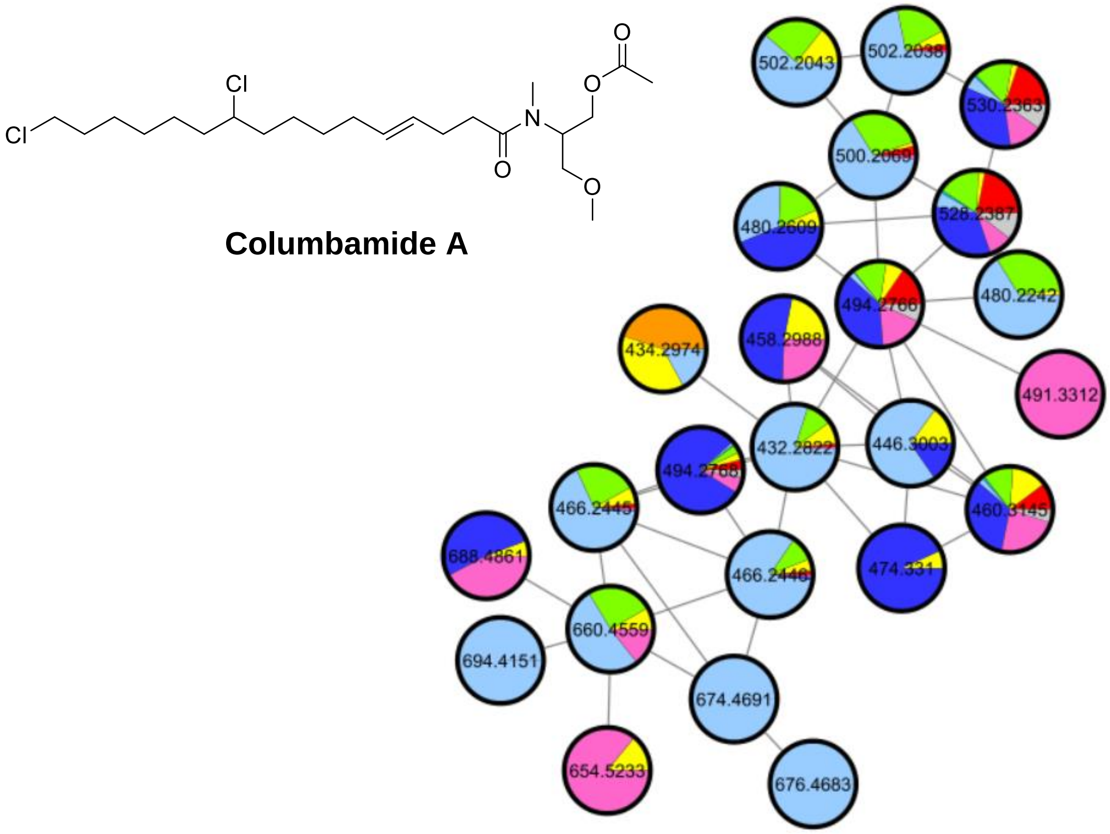


Figure 4.10 – Feature-based molecular network cluster of the columbamides along with a representative structure from the compound family: columbamide A. The pie charts, illustrating relative quantification, are colorized as follows: American Samoa = red; non-shrimp-associated Guam = orange; shrimp-associated Guam = yellow; Lakshadweep Islands = green; Palmyra Atoll = teal; Papua New Guinea = light blue; Paracel Islands = royal blue; Red Sea = purple; Saipan = pink; Solomon Islands = grey.

In contrast to the apratoxins and columbamides, the doscadenamides (Figure 4.11) and the apramides (Figure 4.12) tended to show more regional specificity. While several doscadenamide analogs appeared to be prominent in other regions, particularly American Samoa (red) and Papua New Guinea (light blue), most of the doscadenamides were only relatively abundant in the Marianas Islands, with prominence in samples from Guam (shrimp-associated) and Saipan. The apramides were also predominantly found in samples from Guam but were much more dominant in non-shrimp-associated samples. Just as for the apratoxins, the apramides present another dimension by which shrimp and non-shrimp-associated samples from Guam overlap in chemistry but are differentiable due to very different relative levels of production of those shared chemical components.

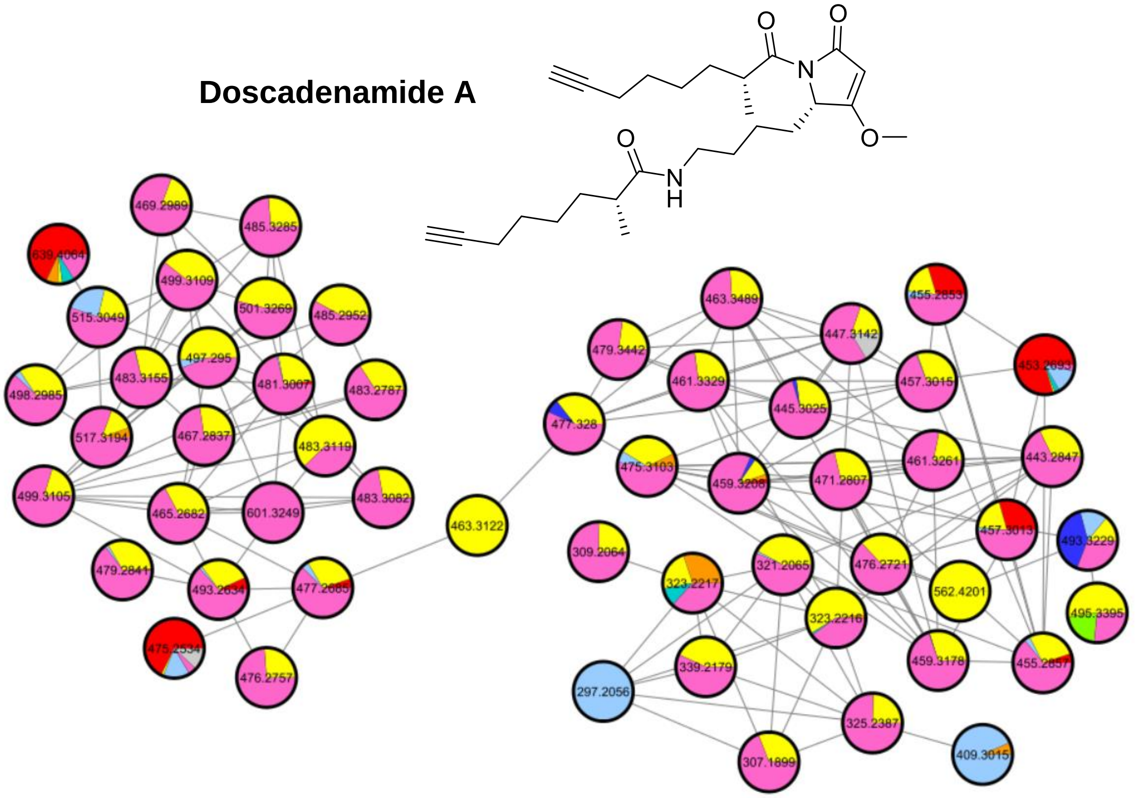


Figure 4.11 – Feature-based molecular network cluster of the doscadenamides along with a representative structure from the compound family: doscadenamide A. The pie charts, illustrating relative quantification, are colorized as follows: American Samoa = red; non-shrimp-associated Guam = orange; shrimp-associated Guam = yellow; Lakshadweep Islands = green; Palmyra Atoll = teal; Papua New Guinea = light blue; Paracel Islands = royal blue; Red Sea = purple; Saipan = pink; Solomon Islands = grey.

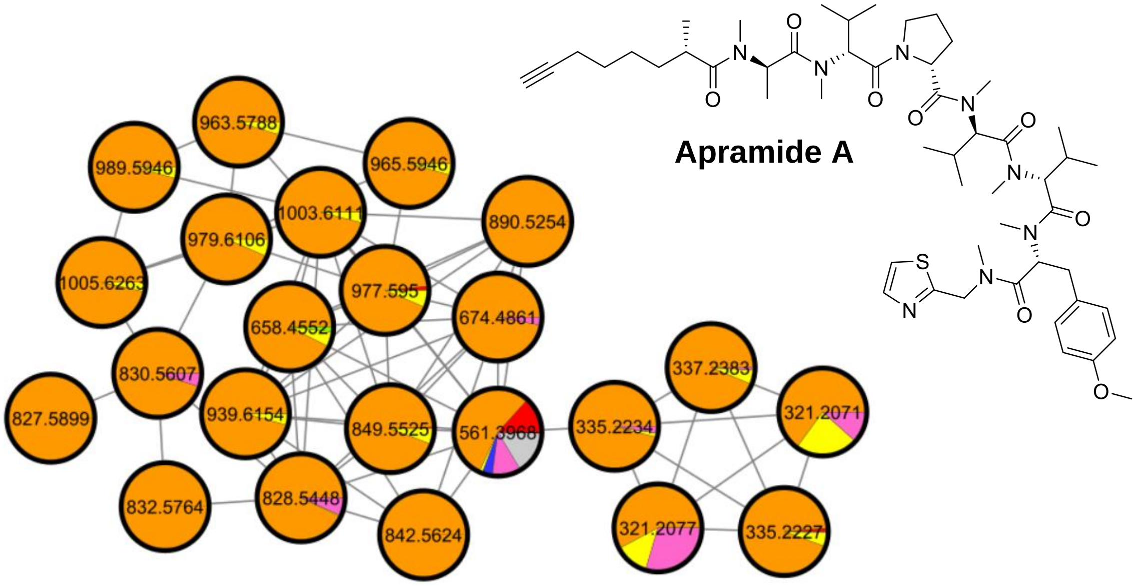


Figure 4.12 – Feature-based molecular network cluster of the apramides along with a representative structure from the compound family: apramide A. The pie charts, illustrating relative quantification, are colorized as follows: American Samoa = red; non-shrimp-associated Guam = orange; shrimp-associated Guam = yellow; Lakshadweep Islands = green; Palmyra Atoll = teal; Papua New Guinea = light blue; Paracel Islands = royal blue; Red Sea = purple; Saipan = pink; Solomon Islands = grey.

The lyngbyapeptin cluster (Figure 4.13) was composed of multiple subsets of analogs, and aside from a few nodes with wider distributions, each major subset of nodes was mainly recorded only from a limited geographical area. A large subset of nodes was predominantly detected in Guam (shrimp-associated) and Saipan samples, with an additional collection mainly found in shrimp-associated Guam samples only. Eight additional nodes were prolific in non-shrimp-associated Guam samples, while detected at much lower relative quantities in shrimp-associated samples. A final subset of lyngbyapeptin analogs was strongly regionally specific for the Lakshadweep Islands.

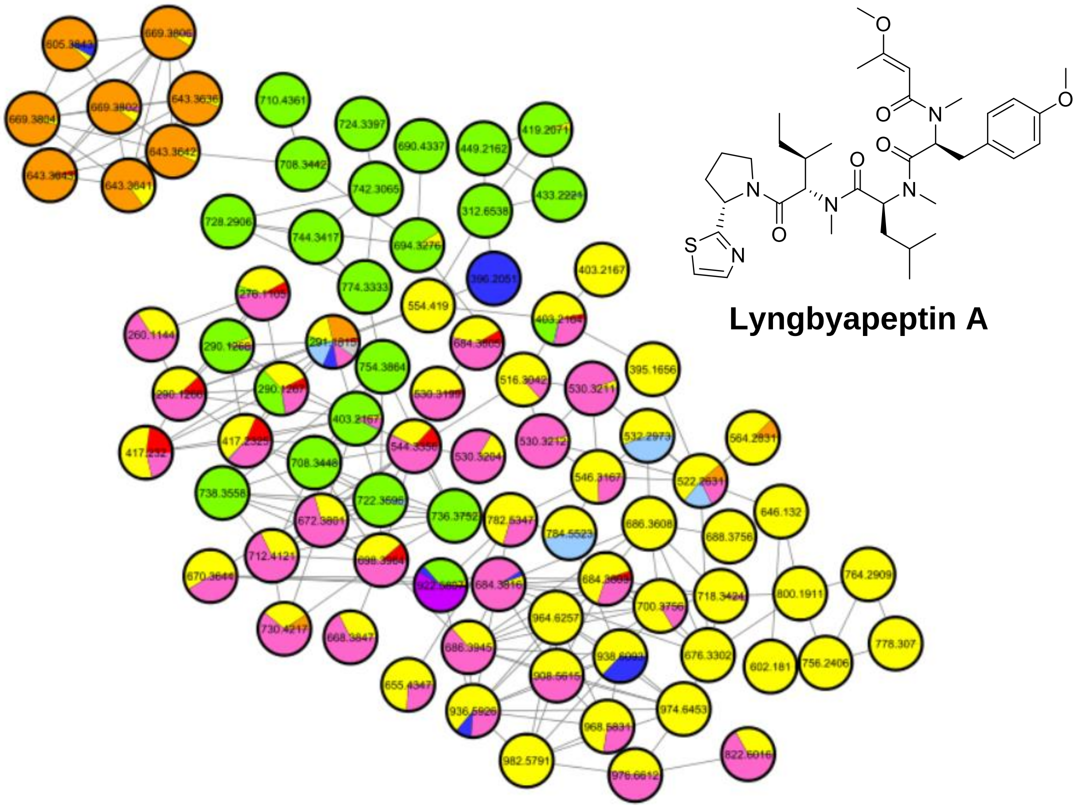


Figure 4.13 – Feature-based molecular network cluster of the lyngbyapeptins along with a representative structure from the compound family: lyngbyapeptin A. The pie charts, illustrating relative quantification, are colorized as follows: American Samoa = red; non-shrimp-associated Guam = orange; shrimp-associated Guam = yellow; Lakshadweep Islands = green; Palmyra Atoll = teal; Papua New Guinea = light blue; Paracel Islands = royal blue; Red Sea = purple; Saipan = pink; Solomon Islands = grey.

4.2.6: Major features prioritized for discovery

Chemogeographic analyses not only provided insights into how *M. bouillonii* molecular features and compound families are distributed across the sampled geographic space, but also allowed for the generation of a priority list of major molecular features that could immediately inform natural product drug discovery efforts. The top five most prominent features from each sample were compiled, resulting in a list of 115 major features (Table 4.3). The *m/z* of each feature was then queried against a dataset of all published compounds from *M. bouillonii* (Leber et al 2020). Out of the 115 features, 90 could not be putatively identified when queried against a database of published compounds from *M. bouillonii*. While this does not eliminate the possibility that any of these 90 molecular features represents a known compound, it does suggest that there is an abundance of opportunity for discovering new natural products from *M. bouillonii*.

Table 4.3 – Top five most prominent features from each *M. bouillonii* sample. 90 out of the 115 major features could not be dereplicated when queried against known *M. bouillonii* natural products.

m/z	retention time	source location	putative ids
222.1323	3.069602	['Guam']	['None']
247.152	6.989638	['Guam']	['None']
291.1815	1.645606	['Guam']	['None']
326.3316	5.588302	['Paracel Islands']	['None']
335.2227	3.055767	['Guam']	['None']
339.3355	5.735705	['Guam']	['None']
352.2761	3.772212	['American Samoa']	['laingolide [M+H]+']
354.3644	6.274413	['Paracel Islands']	['None']
355.2735	3.712698	['Papua New Guinea']	['None']
359.1289	3.134836	['Solomon Islands']	['None']
359.1963	0.426434	['Solomon Islands']	['None']
360.2254	2.458011	['American Samoa']	['laingolide A [M+Na]+']
364.3494	5.897046	['Guam', 'Red Sea', 'Papua New Guinea']	['None']
371.0934	5.672424	['Papua New Guinea']	['None']
371.0935	4.980381	['Papua New Guinea', 'Palmyra Atoll']	['None']
378.2929	3.989943	['American Samoa']	['None']
385.3395	7.05957	['Saipan']	['None']
399.3009	3.776899	['Papua New Guinea']	['None']
404.3457	4.708892	['Red Sea']	['None']
417.2325	2.407709	['Guam']	['None']
418.3023	4.419728	['Guam', 'Saipan', 'Paracel Islands']	['columbamide H [M+H]+']
426.3286	4.708892	['Red Sea']	['None']
432.2822	4.423393	['Papua New Guinea']	['None']
432.3779	5.578946	['Red Sea']	['None']
443.3296	3.849141	['Papua New Guinea']	['None']
454.2628	3.934073	['Paracel Islands']	['None']
457.3015	1.498591	['Guam', 'Saipan']	['doscadenamide A [M+H]+']
459.3178	2.512371	['Guam', 'Saipan']	['None']
460.3145	5.157204	['Guam', 'Paracel Islands']	['None']
462.3129	5.149448	['Guam']	['None']
466.1457	4.883285	['Guam']	['None']
466.2445	3.959899	['Papua New Guinea']	['columbamide A [M+H]+']
468.1428	4.76467	['Guam', 'Saipan']	['None']
468.242	3.974936	['Papua New Guinea']	['None']
474.2474	3.972247	['Paracel Islands']	['columbamide D [M+Na]+']

Table 4.3 (cont.) – Top five most prominent features from each *M. bouillonii* sample. 90 out of the 115 major features could not be dereplicated when queried against known *M. bouillonii* natural products.

m/z	retention time	source location	putative ids
479.2841	1.496712	['Saipan']	['doscadenamide A [M+Na]+']
482.2973	5.163594	['Guam']	['None']
488.2254	4.013767	['Papua New Guinea']	['columbamide A [M+Na]+']
493.3601	3.212118	['Guam']	['kanamienamide [M+H]+']
494.2766	4.704009	['Paracel Islands']	['columbamide F [M+H]+']
505.3221	2.280325	['Paracel Islands']	['None']
509.2223	2.630083	['Papua New Guinea']	['None']
514.3833	1.074207	['Paracel Islands']	['None']
528.3774	3.984582	['Guam']	['None']
528.3985	1.442634	['Paracel Islands']	['None']
535.268	4.961861	['Saipan', 'Palmyra Atoll', 'American Samoa']	['None']
536.2712	4.952996	['American Samoa']	['None']
537.4436	6.512034	['Lakshadweep Islands']	['None']
544.3731	4.186979	['Guam']	['None']
561.1286	4.982971	['Papua New Guinea']	['None']
564.3413	0.929901	['Guam']	['None']
578.358	0.999633	['Guam']	['None']
579.3617	1.085258	['Guam']	['None']
593.2757	4.736265	['Guam', 'Solomon Islands', 'Palmyra Atoll', 'American Samoa']	['None']
594.3533	1.069336	['Guam']	['None']
594.4106	3.620372	['Palmyra Atoll']	['None']
602.3771	2.98317	['Paracel Islands']	['None']
607.2913	5.003193	['Palmyra Atoll']	['lyngbouilloside [M+Na]+']
607.2914	5.608061	['Solomon Islands']	['lyngbouilloside [M+Na]+']
607.2917	6.342993	['American Samoa']	['lyngbouilloside [M+Na]+']
609.2709	4.390828	['Guam']	['None']
609.2714	4.826529	['Solomon Islands']	['None']
611.2864	2.585038	['Solomon Islands']	['None']
612.4202	4.624253	['Palmyra Atoll']	['None']
615.3853	2.75359	['Papua New Guinea']	['None']
616.3928	3.558299	['Palmyra Atoll']	['None']
621.3072	5.811726	['Lakshadweep Islands', 'Palmyra Atoll', 'Papua New Guinea']	['None']

Table 4.3 (cont.) – Top five most prominent features from each *M. bouillonii* sample. 90 out of the 115 major features could not be dereplicated when queried against known *M. bouillonii* natural products.

m/z	retention time	source location	putative ids
621.3074	5.49306	['Lakshadweep Islands', 'Papua New Guinea', 'Palmyra Atoll']	['None']
623.2867	4.776902	['Papua New Guinea']	['None']
623.2876	5.283011	['Solomon Islands']	['None']
625.3007	4.182387	['Solomon Islands']	['None']
635.3234	6.049831	['Guam', 'Lakshadweep Islands', 'American Samoa']	['None']
637.3671	2.831935	['Papua New Guinea']	['None']
638.3055	5.751461	['Lakshadweep Islands']	['None']
642.2108	6.556206	['Papua New Guinea', 'Palmyra Atoll']	['None']
648.4201	3.70225	['Lakshadweep Islands', 'Palmyra Atoll']	['None']
651.3189	5.758731	['Guam', 'Lakshadweep Islands', 'Papua New Guinea', 'American Samoa', 'Saipan']	['None']
651.3837	3.061406	['Papua New Guinea']	['None']
652.322	5.617398	['Guam', 'Lakshadweep Islands', 'American Samoa']	['None']
653.2983	5.46834	['Papua New Guinea']	['None']
655.512	7.198186	['Solomon Islands']	['None']
662.4358	4.097888	['Palmyra Atoll']	['None']
665.3344	6.157737	['Lakshadweep Islands']	['None']
667.3141	5.753695	['Lakshadweep Islands', 'Papua New Guinea']	['None']
673.4171	4.101458	['Guam', 'Lakshadweep Islands', 'Saipan']	['None']
681.3301	6.106998	['Papua New Guinea']	['None']
682.3331	6.08975	['Papua New Guinea']	['None']
685.2803	6.617158	['Saipan']	['None']
688.6183	7.671948	['Guam', 'Papua New Guinea']	['None']
699.4169	7.071523	['Paracel Islands']	['None']
708.3448	1.048381	['Lakshadweep Islands']	['ulongamide E [M+Na]+']
722.3596	1.357359	['Lakshadweep Islands']	['lyngbyapeptin B [M+H]+']
765.5147	6.069748	['Palmyra Atoll']	['None']
777.5781	4.001039	['American Samoa']	['None']
791.4199	2.481706	['Paracel Islands']	['kakeromamide A [M+H]+', 'kakeromamide B [M+H]+']
796.4728	2.623137	['Guam']	['apratoxin E [M+H]+']
814.4822	1.959695	['Guam']	['apratoxin G [M+H]+']
815.4848	2.01454	['Guam']	['None']
826.4826	3.286896	['Guam', 'Papua New Guinea']	['apratoxin B [M+H]+', 'apratoxin C [M+H]+']
827.4855	3.325879	['Guam']	['apratyramide [M+Na]+']

Table 4.3 (cont.) – Top five most prominent features from each *M. bouillonii* sample. 90 out of the 115 major features could not be dereplicated when queried against known *M. bouillonii* natural products.

m/z	retention time	source location	putative ids
828.4893	3.149444	['Palmyra Atoll']	['apramide G [M+H]+', 'apratoxin F [M+H]+']
829.6838	4.708892	['Red Sea']	['None']
833.4937	4.268341	['Papua New Guinea']	['cyanolide A [M+H]+']
841.5007	3.376924	['Guam', 'Papua New Guinea']	['None']
887.5724	7.961793	['Lakshadweep Islands']	['None']
887.5727	7.655898	['Paracel Islands']	['None']
904.5412	3.826858	['Guam']	['None']
905.5369	4.86486	['Papua New Guinea']	['apratoxin D [M+Na]+']
912.706	6.839769	['Lakshadweep Islands']	['None']
917.5838	6.98196	['Palmyra Atoll']	['None']
918.5881	6.598318	['Palmyra Atoll']	['None']
977.595	3.036765	['Guam']	['apramide A [M+H]+']
999.5778	3.030814	['Guam']	['apramide A [M+Na]+']
1035.616	2.921822	['Guam']	['None']
1367.815	4.102983	['Lakshadweep Islands', 'Saipan']	['None']

4.2.7: Conclusions

M. bouillonii is well known as a prolific producer of natural products; however, comparing samples from various geographical locations reveals chemical diversity and complexity that goes well beyond the narrow subset of natural products previously described from this organism. While one must consider that the quantities of MS^1 features discussed early in the chapter may in part be driven by numerous influences, ranging from sample contamination to imperfect processing, cautious appreciation of those numbers paired with examination of known molecular families provides robust support for the idea that the natural products arsenal of *M. bouillonii* is vast and diverse. While there are undoubtedly numerous factors driving differential production of secondary metabolites in *M. bouillonii*, this work has shown geographical region and shrimp association to be statistically significant organizational frameworks with which observed patterns in *M. bouillonii* chemical composition align. It is plausible that genomic differences underly the observed correlations between region of origin or shrimp-association and chemical composition, and so exploration of genomic and biosynthetic diversity along geographic and ecofunctional gradients in *M. bouillonii* could be a logical and exciting extension of this work. The broad overlap in structurally unique compounds and compound families in *M. bouillonii*, such as the apratoxins, suggests that such genomic differences may exist in modified biosynthetic pathways, and that these variations may best be described as representing different chemotypes or 'chemical races' of the same species.

While geographical location of collection has been clearly shown to be associated with specific patterns of chemical composition, scale is important. At regional scale, when comparing different island groups, distinct clusters of samples could be assembled. As scale decreased, namely in considering shrimp-associated samples from Guam and samples from

Saipan, a high degree of overlap in core features made drawing a distinction between regions more difficult. Though geographically quite distant from the Mariana Islands, samples from American Samoa were also quite similar to samples from Guam and Saipan, though not as much as their similarity to each other. This suggests that gradients of similarity present a more useful construct, rather than distinct clusters, in understanding relatedness between the chemical compositions of samples of *M. bouillonii*. When samples from various sub-regions around Guam were examined for geographical patterns at the intra-island scale, no geographically driven structure was revealed. In this case, samples were able to be organized by shrimp association. This illustrated not only that factors beyond geographical region influence patterns in *M. bouillonii* chemical composition, but also presented a lower limit to the geographic scale that is meaningful for distinguishing samples based on chemical composition.

Counts of regionally specific core MS^1 features highlighted the Paracel Islands and the Lakshadweep Islands as locations with substantial and unique chemical diversity; this conclusion is further supported by the greater MS^2 molecular network beyond the five molecular families examined, with its numerous green and royal blue clusters and cluster subsets. While one may be tempted to attribute this unique chemical diversity to some special characteristic of these regions, it is likely at least in part a function of sampling density. The vast majority of samples included in this study, and really most of the samples of *M. bouillonii* that have thus far been studied for their natural products chemistry, originated from the Tropical Western Pacific. One could hypothesize that sampling in the western Philippines, eastern Vietnam, Taiwan, or other sites with connectivity to the South China Sea would result in samples with chemical compositions much more similar to those from the Paracel Islands.

Similarly, one would expect samples from Sri Lanka and other sites on the Arabian Sea to share more chemical constituents with the samples from the Lakshadweep Islands.

Samples included in this study were collected haphazardly where available from accessible locations, and were aggregated from collaborators based on availability, rather than via an explicit sampling protocol. This undoubtedly means that the patterns herein fail to capture the true extent and complexity of *M. bouillonii* chemical diversity across geographic locations and between shrimp-associated and non-shrimp-associated chemotypes. As a foundational study of *M. bouillonii* chemogeography, it is the hope of the author that continued and geographically diversified sampling of *M. bouillonii* can build upon this work. On the micro scale, thorough random sampling of specific sites could provide additional geographic resolution that may reveal additional factors by which *M. bouillonii* chemical composition is influenced, such as depth. On the macro scale, additional samples from under-sampled areas, such as the Solomon Islands and the Red Sea, would provide a more balanced dataset for better understanding how chemodiversity varies by location. Sampling previously unsampled areas, such as Sri Lanka and the Philippines, would help to fill in geographical gaps, and so better illustrate how gradients of *M. bouillonii* natural products exist across geographical space. Though seemingly less prominent, and certainly more difficult to find, it would be fascinating to increase efforts in the sampling of non-shrimp-associated *M. bouillonii* from a breadth of geographic locations to assess how this chemotype may mirror the chemogeographic patterns evident in its shrimp-associated counterpart.

While there is much support for chemogeographical patterns in *M. bouillonii* presented herein, one region that was incredibly inconsistent in following such patterns was Papua New Guinea. While multiple factors likely contributed to this inconsistency, including age of samples, higher variability in field identification and extraction, and chemical compositional differences intrinsic to cultures versus collections, one must also consider the geographical scale over which these samples were collected. Papua New Guinea is a vast archipelago, and samples were collected from disparate sites in different seas surrounding the archipelago, from both the main island as well as the island of New Britain. To group all of these samples together under the name ‘Papua New Guinea,’ while technically correct by our human-centric organization of geographical locations, represents a collection of samples encompassing a much larger region than all of the other regional groupings.

Exploring how different analogs are geographically distributed within known molecular families provokes many questions related to how different suites of compounds are produced predominantly in one region, while entirely different suites of analogs from the same family are the major products somewhere else. Along with this, one must consider the symbiont of *M. bouillonii*, the shrimp *Alpheus frontalis*. The shrimp can be found weaving the cyanobacterium from the eastern coast of Africa all of the way to the central tropical Pacific. This study clearly shows that while the shrimp are conducting the same behaviors, namely weaving and constructing domiciles of living filaments as well as feeding on the cyanobacterial cells, they are experiencing very diverse and very different chemical environments while doing so. This suggests an extremely broad tolerance for biologically active compounds; the limits of this tolerance may very well be on display in the apratoxin and apramide family clusters. As noted above, there are numerous apratoxin and apramide analogs that are much more prominent in non-shrimp-associated Guam samples as compared to those that are shrimp-associated.

Beyond the insights that this study has provided on the diversity, patterns, and distribution of natural products in *M. bouillonii*, this work also represents a road map for future scientific collaborations between laboratories in San Diego, Mangilao, Kochi, Ningbo, Corvallis, and Salt Lake City. By sharing samples and analyzing these samples with each other as context, our groups can more efficiently target molecular features most unique to our sampling regions, while effectively avoiding competition. Furthermore, orchestrated targeting of analogous features within the same family but produced in different regions can accelerate the broad characterization related suites of metabolites. It is my hope that this work will provide a foundation for cooperative exploration and elucidation of the still greatly underexplored secondary metabolome of *M. bouillonii*.

4.3: Methods

4.3.1: *Moorena bouillonii* crude extract sample preparation

Forty-nine samples of cyanobacteria, both field-collected and laboratory-cultured, were studied in this project. Samples were selected based on collection location, putative field identification, woven ‘cobweb’ morphology, and shrimp association; some or all of these factors were used in each case for tentatively identifying each sample as *Moorena bouillonii*. In some cases, 16S rRNA gene sequence information was also available, however this was considered only in a secondary sense, as 16S rRNA sequencing lacks the phylogenetic resolution necessary to differentiate *M. bouillonii* from its congener *Moorena producens* (Engene et al 2012). The dataset was assembled via an inclusive approach, opting to retain edge cases rather than discard them. Samples originated from collection sites around Guam, Saipan (Commonwealth of the Northern Mariana Islands), Palmyra Atoll, Papua New Guinea, American Samoa, Kavaratti (Lakshadweep, India), the Paracel Islands (Xisha, China), the Solomon Islands, and the Red Sea (via Egypt) (See Table 4.S3 for selected metadata. Full metadata is available at MassIVE accession MSV000086109). Crude extracts were generated by exhaustively extracting cyanobacterial biomass with 2:1 dichloromethane and methanol. Extracts were concentrated and resuspended in LC-MS grade acetonitrile, and desalted via elution through C18 solid phase extraction (SPE) cartridges with acetonitrile. Volumes equivalent to 100 µg of crude extract were deposited in a 96-well plate and solvent evaporated to dryness.

Samples were resuspended in 50 µL of LC-MS grade methanol containing 2 µM sulfamethazine as an internal standard. LC-MS/MS was performed in an UltiMate 3000 UPLC system (Thermo Scientific; Waltham, MA) using a Kinetex 1.7 ųm C18 reversed phase UHPLC column (50 mm X 2.1 mm) and Maxis Impact Q-TOF mass spectrometer (Bruker Daltonics; Billerica, MA) equipped with an ESI source. Data was acquired in positive ion mode.

4.3.2: LC-MS/MS conditions

2 µL of each sample was injected, and samples were run in duplicate. The column was equilibrated with 50% solvent A (water, 0.1% formic acid) and 50% solvent B (LC-MS grade acetonitrile, 0.1% formic acid) for 1 min, followed by a linear gradient from 50% B to 100% B over 5 min, and a 2 min hold at 100% B. The method continued with 100% to 50% B over 1 min and a hold at 50% B for 1 min for a total run time of 10 min. The solvent flow rate was 0.5 mL/min throughout the run. MS spectra were acquired in positive ion mode, in the *m/z* range of 50-2000. After every 10 injections, a mixture of 10 mg/mL of each sulfamethazine, sulfamethizole, sulfachloropyridazine, sulfadimethoxine, amitriptyline, and coumarin was run for quality control. An external calibration with ESI-Low Concentration Tuning Mix (*m/z* 118.086255; 322.048121; 622.028960; 922.009798; 1221.990637; 1521.971475;

1821.952313) (Agilent Technologies; Santa Clara, CA) was performed prior to data collection. The internal calibrant Hexakis(1H,1H,2H-perfluoroethoxy)phosphazene (CAS 186817-57-2) was used throughout the runs. The following parameters were set for acquisition in positive ion mode: capillary voltage = 4200 V, nebulizer gas pressure (N_2) = 2 bar, ion source temperature = 200 ^{o}C, dry gas flow = 9 L/min, spectral rate = 3 Hz for MS^1, 10 Hz for MS^2. For acquiring MS/MS fragmentation spectra, the 5 most intense ions per MS^1 scan were selected, the MS/MS active exclusion parameter was enabled; it was set to 5 and set to release after 30 s. The precursor ion was reconsidered for MS/MS if current intensity/previous intensity ratio was >2. An advanced stepping function was used to fragment ions in positive acquisition according to Table 4.S4, and CID energies for MS/MS data acquisition were used according to Table 4.S5. The mass of the internal calibrant was excluded from the MS/MS list using a mass range of *m/z* 621.5–623.0. All data is available on MassIVE (Aron et al 2020), accession MSV000086109.

4.3.3: Data preprocessing

Mass spectral data was preprocessed using MZmine 2.53 (Pluskal et al 2010). Data was downloaded from MassIVE in mzXML format and loaded into MZmine. Masses were detected for MS1 (noise level = 1.0E3) and MS^2 (noise level = 1.0E2), with mass detector set to ‘Centroid’, and a common mass list name used for both. The mass list was then used to build chromatograms with ADAP Chromatogram Builder (MS level = 1; Min group size in # of scans = 7; Group intensity threshold = 1.0E3; Min highest intensity = 5.0E3; m/z tolerance = 0.005 m/z) (Myers et al 2017), followed by chromatogram deconvolution (Algorithm = baseline cutoff; min. peak height = 5.0E3; peak duration range = 0-10 mins; baseline level = 1.0E3; m/z center calculation = median; m/z range for MS2 scan pairing = 0.01 Da; RT range for MS2 scan pairing = 0.1 min), and isotopic peaks grouping (m/z tolerance = 0.005 m/z; retention time tolerance = 0.1 min; maximum charge = 5; representative isotope = most intense). Join aligner was then used to align features (m/z tolerance = 0.01 m/z; weight for m/z = 0.75; retention time tolerance = 0.5 min; weight for RT = 0.25), and then peak finder was applied (Intensity tolerance = 50%; m/z tolerance = 0.01 m/z; retention time tolerance = 0.3 min). Peaks were filtered to between 5.0E3 and 1.0E7 peak heights, deduplicated (Filter mode = new average; m/z tolerance = 0.1 m/z; RT tolerance = 0.3 min), and then further filtered so that only peaks with a minimum of 2 peaks per row were retained. This set of peaks was used for MS^1 in ORCA (Leber et al 2020). An additional filtering step of retaining only peaks with paired MS^2 scans was undergone to prepare a dataset for FBMN (Nothias et al 2020).

4.3.4: MS^1 analyses

MS^1 analyses were conducted using the Objective Relational Comparative Analyses (ORCA) pipeline (Leber et al 2020). The ORCA pipeline is available online at https://github.com/c-leber/ORCA. Additional preprocessing steps in ORCA included applying a logarithmic (base 10) transformation[5], removing features that had a median area greater than 0 for at least one of four different sets of blanks (process, solvent, qc_mix, internal_standard), removing blanks and extraneous samples from the dataset, removing peaks that did not occur in both duplicates of each sample, removing duplicates so that only one of each sample was present, and dropping features with areas of 0 for all remaining samples. Additionally, prior to PCA, unit vector normalization was applied. Hierarchical clustering was conducted with cosine

[5] Logarithmic transformations are applied to sparse feature vectors by first setting values less than the minimum peak area (e.g. values of 0) to 1, so that following transformation, these values are again 0.

distance as the metric, and ‘average’ as the method of clustering. To test the statistical significance of similarity within groups versus not within groups, pairwise calculations of cosine distance, number of overlapping features, and proportion of overlapping features were conducted, and these were labelled as True (indicating sample pairs both from the same group) or False (sample pairs not from the same group). A t-test for each of the three metrics (cosine distance, shared feature count, and proportion of shared features) was then conducted, comparing within group pairs to not within group pairs. The priority feature list was queried for dereplication using a m/z cutoff of 0.1 m/z.

4.3.5: MS^2 analyses and molecular networking

Data was submitted to GNPS (Wang et al 2016) and analyzed via multiple workflows, including classical molecular networking (https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=3aced16d710742a5a2df502a9ba151a9) and FBMN (https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=563bcf3c54304f3891d372683e07a45b) (Nothias et al 2020). Prior to submission to FBMN, the peak table generated via MZmine was loaded into ORCA, and peaks were normalized to the internal standard, for improved relative quantitation. The output from FBMN was further analyzed via DEREPLICATOR+ (https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=43557fc5a1a64eeab6aa13bfb9c044cb) (Mohimani et al 2018), MS2LDA (https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=55adeb6b0011485eb79e9c0aeff82af7) (van der Hooft et al 2016), and NAP (https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=9cdf011351dc439c9b476abf10117

985) (da Silva et al 2018). Results from DEREPLICATOR+ and MS2LDA were combined via MolNetEnhancer (https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=0653e8521d83464684fd3e9db2c56a70) (Ernst et al 2019). Resultant molecular networks were visualized using Cytoscape (3.7.2) (https://cytoscape.org/) (Shannon et al 2003).

Chapter 4 is coauthored with Caraballo-Rodríguez, Andrés Mauricio; Naman, C. Benjamin; Glukhov, Evgenia; Reyes, Andres Joshua; Joseph, Valsamma; Sajeevan, T. P.; Li, Te; Yuan, Ye; He, Shan; Yan, Xiaojun; Miller, Bailey W.; Thornburg, Christopher C.; McPhail, Kerry L.; Schmidt, Eric W.; Haygood, Margo G.; Dorrestein, Pieter C.; Biggs, Jason S.; Gerwick, William H. The dissertation author was the primary author of this chapter. The dissertation author co-conceived of this work, participated in cyanobacterial collections on Guam and Lakshadweep, acquired and prepared all samples for LC-MS/MS data collection, wrote code and conducted all analyses, and was the primary author of this work.

4.4 Appendix: Supplemental Information

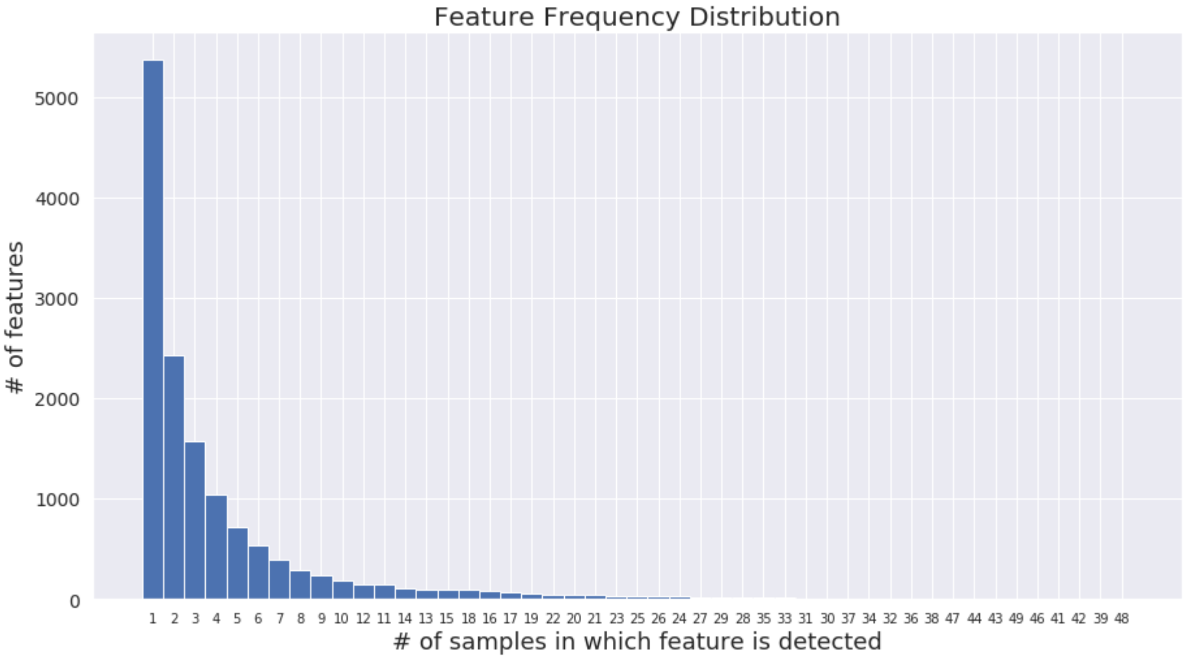


Figure 4.S1 – Frequency distribution of features occurring in different sized subsets of samples. 5,369 detected features were sample-specific, while no features were detected across all samples after background removal.

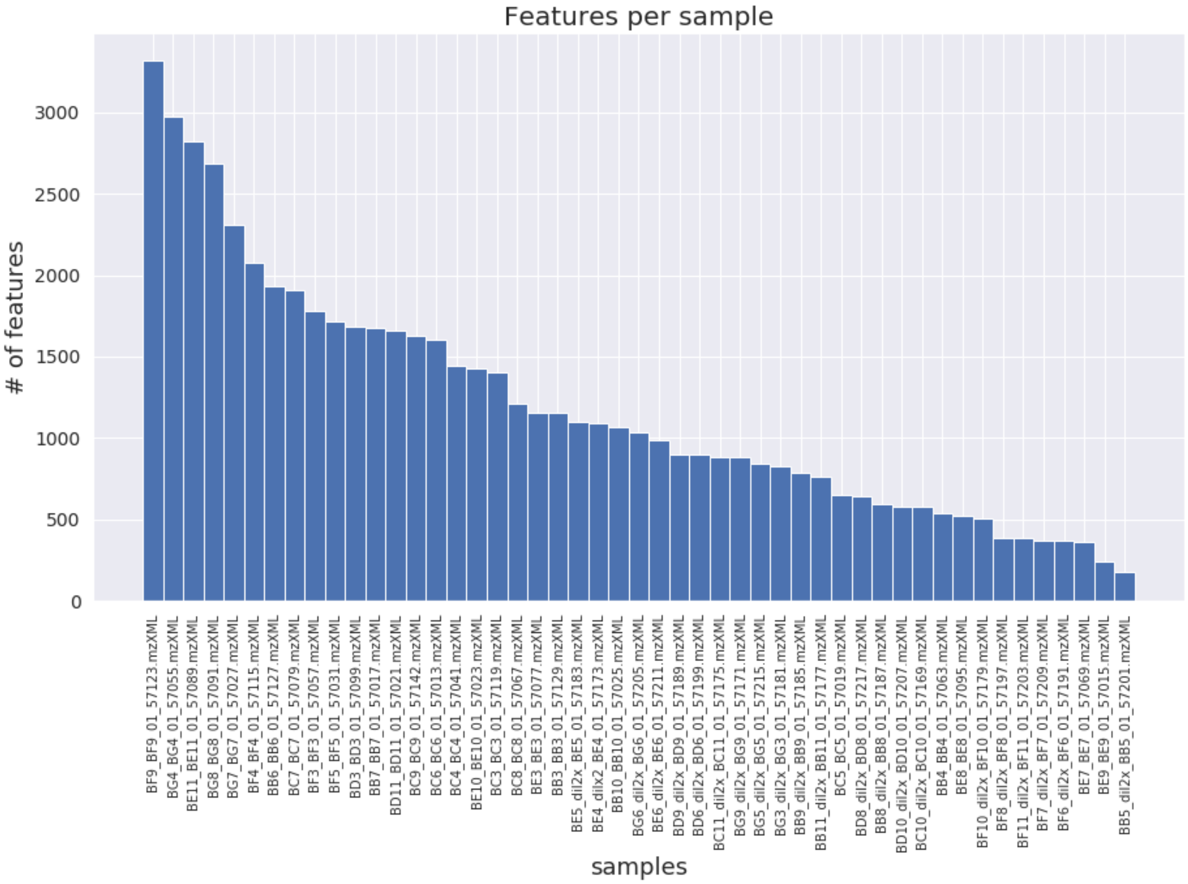


Figure 4.S2 – Number of features per sample, with samples ordered from left to right based on feature count. The maximum number of features detected in a sample was 3,320, while the minimum was 176 features.

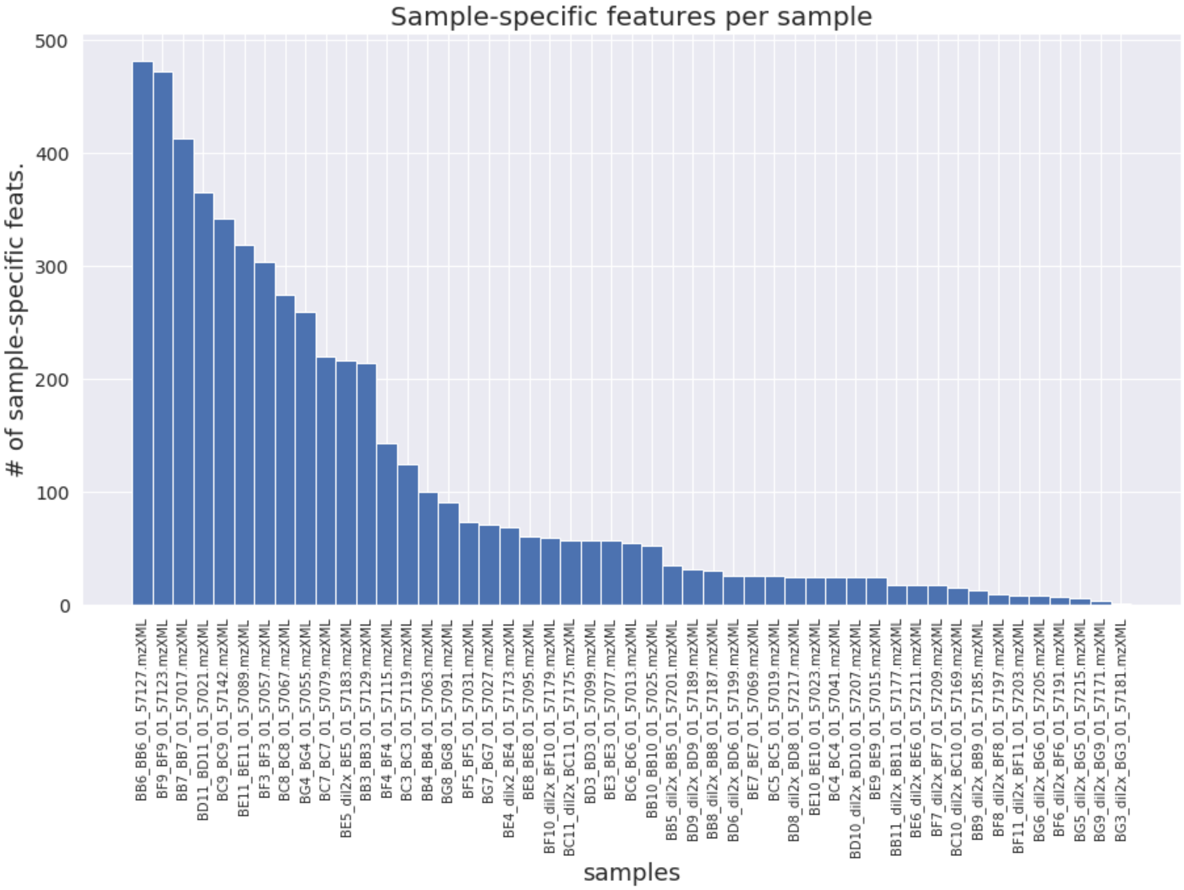


Figure 4.S3 – Number of singletons (features detected in only single samples) per sample, with samples ordered from left to right based on singleton count. The maximum number of singletons detected in a sample was 482, while the minimum was 1 singleton.

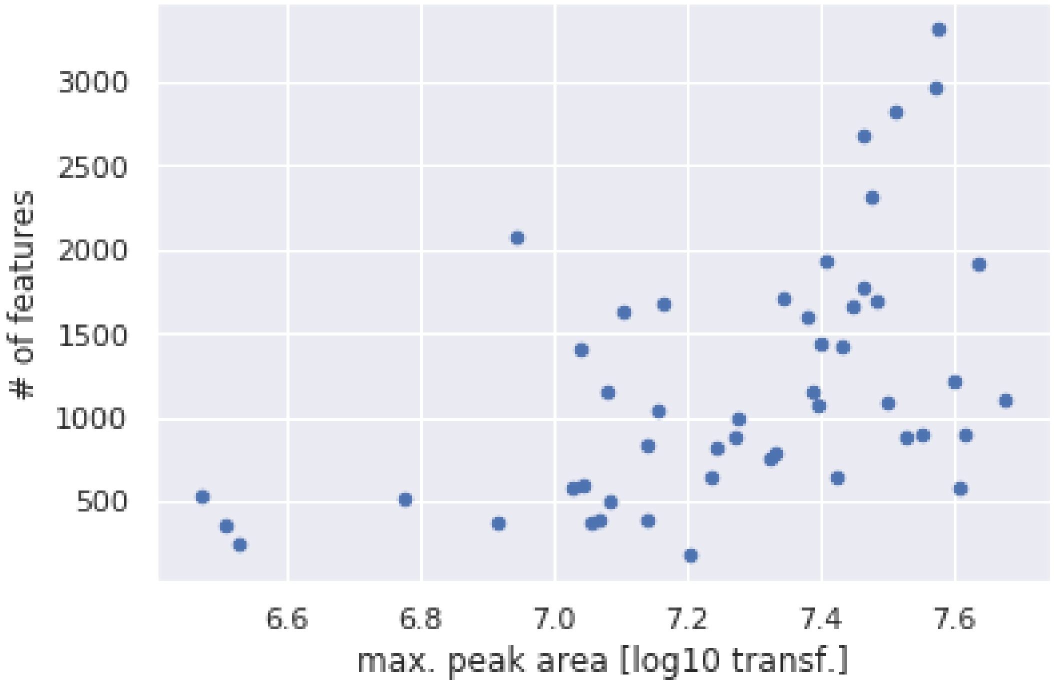


Figure 4.S4 – Plot of the number of features in a sample versus the area of the maximum peak in that sample.

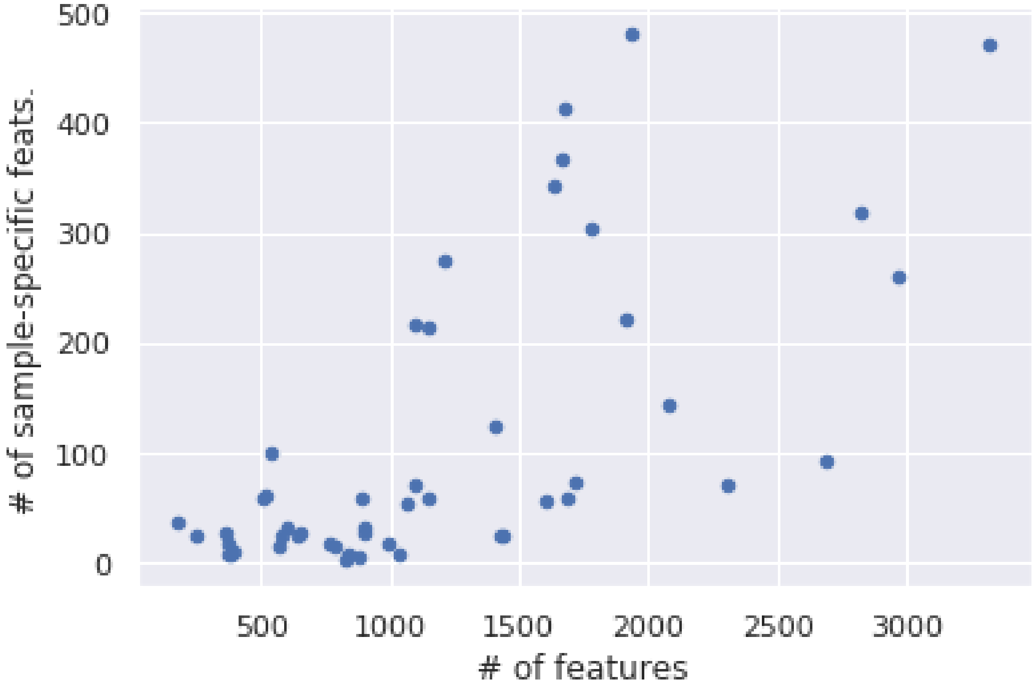


Figure 4.S5 – Plot of number of sample specific features versus the total number of features in a sample.

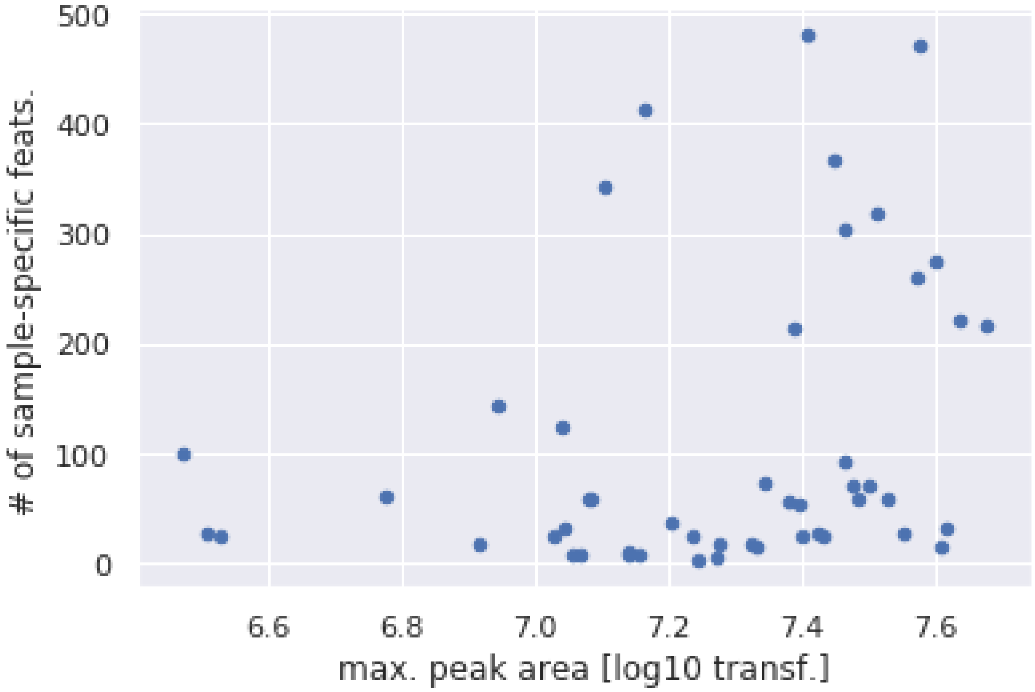


Figure 4.S6 – Plot of number of sample-specific features versus the area of the maximum peak in each sample.

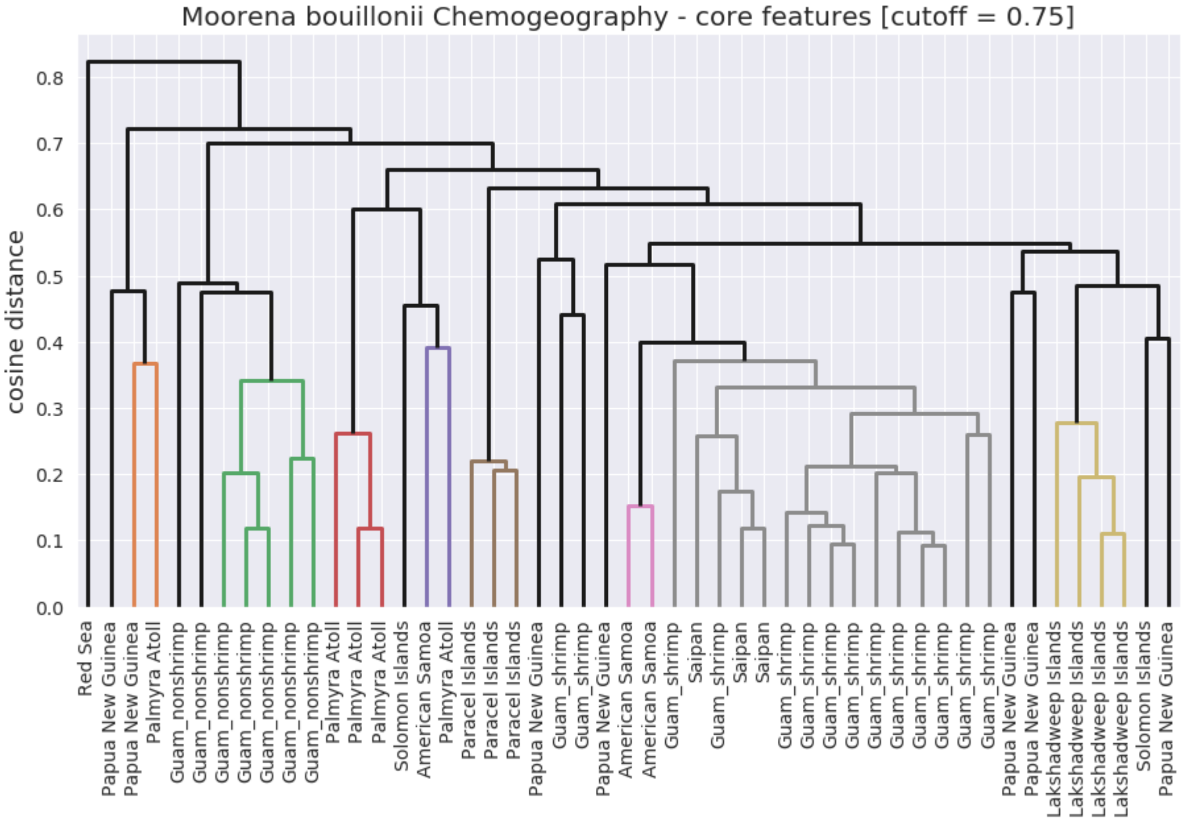


Figure 4.S7 – ORCA dendrogram illustrating the results of hierarchical clustering after selecting for core features per region, with a cutoff value of 0.75.

Table 4.S1 – Counts of features, core features, and specific core features for each regional bin. A cutoff of 1.00 was applied for determining core features.

	mean - feats/sample	**std - feats/sample**	**core feats**	**specific core feats**	**sample size**
American Samoa	878.67	181.68	197	2	3
Guam_non-shrimp	1564.29	1014.42	38	1	7
Guam_shrimp	1468.21	846.68	29	0	14
Lakshadweep Islands	1005.25	260.55	320	48	4
Palmyra Atoll	762.40	328.67	10	0	5
Papua New Guinea	897.71	525.25	12	0	7
Paracel Islands	1399.67	577.11	307	69	3
Red Sea	176.00	0.00	176	35	1
Saipan	1399.00	185.81	473	0	3
Solomon Islands	846.50	306.50	156	2	2

Table 4.S2 – Pairwise counts of shared core features, with core features determined based on of a cutoff value of 1.00.

	Amer. Samoa	Guam _n.s.	Guam _shr.	Lakshad . Islands	Palmyra Atoll	Papua New Guinea	Paracel Islands	Red Sea	Saipan	Solom. Islands
American Samoa	**197** **(1.00)**	9 (0.23)	18 (0.62)	66 (0.21)	9 (0.90)	10 (0.83)	38 (0.12)	23 (0.13)	95 (0.20)	38 (0.24)
Guam _non-shrimp	9 (0.05)	**38** **(1.00)**	6 (0.21)	8 (0.03)	4 (0.40)	7 (0.58)	12 (0.04)	5 (0.03)	11 (0.02)	6 (0.04)
Guam _shrimp	18 (0.09)	6 (0.16)	**29** **(1.00)**	20 (0.06)	4 (0.40)	7 (0.58)	18 (0.06)	9 (0.05)	25 (0.05)	10 (0.06)
Lakshadweep Islands	66 (0.34)	8 (0.21)	20 (0.69)	**320** **(1.00)**	6 (0.60)	9 (0.75)	76 (0.25)	18 (0.10)	127 (0.27)	31 (0.20)
Palmyra Atoll	9 (0.05)	4 (0.11)	4 (0.14)	6 (0.02)	**10** **(1.00)**	6 (0.50)	8 (0.03)	4 (0.02)	10 (0.02)	7 (0.04)
Papua New Guinea	10 (0.05)	7 (0.18)	7 (0.24)	9 (0.03)	6 (0.60)	**12** **(1.00)**	11 (0.04)	9 (0.05)	12 (0.03)	11 (0.07)
Paracel Islands	38 (0.19)	12 (0.32)	18 (0.62)	76 (0.24)	8 (0.08)	11 (0.92)	**307** **(1.00)**	16 (0.09)	107 (0.23)	25 (0.16)
Red Sea	23 (0.12)	5 (0.13)	9 (0.31)	18 (0.06)	4 (0.04)	9 (0.75)	16 (0.05)	**176** **(1.00)**	36 (0.08)	35 (0.22)
Saipan	95 (0.48)	11 (0.29)	25 (0.86)	127 (0.40)	10 (1.00)	12 (1.00)	107 (0.35)	36 (0.20)	**473** **(1.00)**	51 (0.33)
Solomon Islands	38 (0.19)	6 (0.16)	10 (0.34)	31 (0.10)	7 (0.70)	11 (0.92)	25 (0.08)	35 (0.20)	51 (0.11)	**156** **(1.00)**

Table 4.S3 – Selected metadata for *M. bouillonii* samples

DateCollected	Latitude :Longitude	Collection Region	ShrimpV Nonshrimp	CultureV Collection
7/13/2014	14°16'52"S:170°38'19"W	American Samoa	unknown	collection
7/12/2014	14°20'04"S:170°48'07"W	American Samoa	shrimp	collection
7/15/2014	14°16'59"S:170°43'23"W	American Samoa	shrimp	collection
7/3/2017	13°26'40"N:144°38'11"E	Guam	nonshrimp	collection
7/3/2016	13°26'40"N:144°38'11"E	Guam	shrimp	collection
3/21/2016	13°26'38"N 144°38'36"E	Guam	nonshrimp	collection
3/21/2016	13°26'57"N 144°39'27"E	Guam	shrimp	collection
5/31/2018	13°26'40"N:144°38'11"E	Guam	shrimp	collection
6/1/2018	13°26'40"N:144°38'11"E	Guam	nonshrimp	collection
6/10/2016	13°26'57"N:144°39'27"E	Guam	nonshrimp	collection
6/14/2016	13°19'01"N:144°39'12"E	Guam	shrimp	collection
6/13/2016	13°26'11"N:144°37'37"E	Guam	shrimp	collection
8/13/2020	13°26'40"N:144°38'11"E	Guam	shrimp	culture
8/13/2020	13°26'40"N:144°38'11"E	Guam	nonshrimp	culture
6/16/2017	13°39'02"N:144°51'08"E	Guam	shrimp	collection
6/16/2017	13°39'02"N:144°51'08"E	Guam	nonshrimp	collection
6/21/2017	13°36'11"N:144°50'06"E	Guam	shrimp	collection
6/13/2016	13°26'40"N:144°38'11"E	Guam	shrimp	collection
6/13/2016	13°26'40"N:144°38'11"E	Guam	nonshrimp	collection
6/10/2016	13°26'57"N:144°39'27"E	Guam	shrimp	collection
6/14/2016	13°15'02"N:144°39'10"E	Guam	shrimp	collection
5/24/2019	13°36'11"N:144°50'06"E	Guam	shrimp	collection
5/23/2019	13°26'40"N:144°38'11"E	Guam	shrimp	collection
6/11/2016	13°16'12"N:144°39'42"E	Guam	shrimp	collection
4/7/2018	10°34'10"N 72°37'53"E	Lakshadweep Islands	shrimp	collection
4/8/2018	10°32'54"N 72°37'18"E	Lakshadweep Islands	shrimp	collection
4/8/2018	10°34'28"N 72°38'04"E	Lakshadweep Islands	shrimp	collection
2/6/2016	10°34'28"N 72°38'04"E	Lakshadweep Islands	shrimp	collection
08/02/2009	5°53'21"N 162°05'54"W	Palmyra Atoll	unknown	collection
8/16/2008	5°52'51"N 162°05'46"W	Palmyra Atoll	shrimp	collection
8/3/2009	5°53'47"N 162°07'04"W	Palmyra Atoll	shrimp	collection
8/3/2009	5°53'47"N 162°07'04"W	Palmyra Atoll	shrimp	collection
4/8/2009	5°53'39"N 162°03'38"W	Palmyra Atoll	shrimp	collection
5/2/2007	10°15'50"S 150°46'12"E	Papua New Guinea	shrimp	collection
7/12/2007	5°25'34"S 150°05'45"E	Papua New Guinea	shrimp	collection
08/10/2010	4°16'04"S 152°20'16"E	Papua New Guinea	shrimp	culture
4/21/2006	10°32'46"S 151°02'31"E	Papua New Guinea	shrimp	collection
12/12/2003	10°24'50"S 150°37'60"E	Papua New Guinea	shrimp	collection
5/19/2005	4°16'04"S 152°20'16"E	Papua New Guinea	shrimp	collection
8/13/2020	4°16'04"S:152°20'16"E	Papua New Guinea	shrimp	culture
05/16/2017	16°51'06"N 112°20'56"E	Paracel Islands	shrimp	collection
05/19/2017	16°51'06"N 112°20'56"E	Paracel Islands	shrimp	collection

05/19/2017	16°51'06"N 112°20'56"E	Paracel Islands	shrimp	collection
05/31/2007	27°21'09"N 33°54'28"E	Red Sea	unknown	culture
2/1/2013	15°09'18"N:145°42'20"E	Saipan	shrimp	collection
1/31/2013	15°09'18"N:145°42'20"E	Saipan	shrimp	collection
1/29/2013	15°09'37"N 145°41'43"E	Saipan	shrimp	collection
4/14/2018	na	Solomon Islands	shrimp	collection
4/14/2018	na	Solomon Islands	nonshrimp	collection

Table 4.S4 – Stepping function used for fragmenting ions during MS acquisition

Time	Collision RF	Transfer Time	Collision
0	450.0	70.0	125
25	550.0	75.0	100
50	800.0	90.0	100
75	1100.0	95.0	75

Table 4.S5 - CID energies for MS/MS data acquisition

Type	Mass	Width	Collision	Charge State
Base	100.00	4.00	25.00	1
Base	100.00	4.00	20.00	2
Base	300.00	5.00	30.00	1
Base	300.00	5.00	25.00	2
Base	500.00	6.00	35.00	1
Base	500.00	6.00	30.00	2
Base	1000.00	7.00	45.00	1
Base	1000.00	7.00	40.00	2
Base	1500.00	8.00	60.00	1
Base	1500.00	8.00	50.00	2
Base	2000.00	10.00	60.00	1
Base	2000.00	10.00	60.00	2

4.5: References

Adpressa DA, Loesgen S (2016) Bioprospecting Chemical Diversity and Bioactivity in a Marine Derived *Aspergillus terreus*. Chem Biodivers 13:253-259. https://doi.org/10.1002/cbdv.201500310

Aron AT, Gentry EC, McPhail KL, Nothias LF, Nothias-Esposito M, Bouslimani A, Petras D, Gauglitz JM, Sikora N, Vargas F, van der Hooft JJJ, Ernst M, Kang KB, Aceves CM, Caraballo-Rodríguez AM, Koester I, Weldon KC, Bertrand S, Roullier C, Sun K, Tehan RM, Boya P CA, Christian MH, Gutiérrez M, Ulloa AM, Mora JAT, Mojica-Flores R, Lakey-Beitia J, Vásquez-Chaves V, Zhang Y, Calderón AI, Tayler N, Keyzers RA, Tugizimana F, Ndlovu N, Aksenov AA, Jarmusch AK, Schmid R, Truman AW, Bandeira N, Wang M, Dorrestein PC (2020) Reproducible molecular networking of untargeted mass spectrometry data using GNPS. Nat Protoc 15:1954–1991. https://doi.org/10.1038/s41596-020-0317-5

Bode HB, Bethe B, Höfs R, Zeeck A (2002) Big effects from small changes: possible ways to explore nature's chemical diversity. Chembiochem 3:619-627. https://doi.org/ 10.1002/1439-7633(20020703)3:7<619::AID-CBIC619>3.0.CO;2-9

Cantrell TP, Freeman CJ, Paul VJ, Agarwal V, Garg N (2019) Mass Spectrometry-Based Integration and Expansion of the Chemical Diversity Harbored Within a Marine Sponge. J Am Soc Mass Spectrom 30:1373–1384. https://doi.org/10.1021/jasms.8b06062

Chanana S, Thomas CS, Braun DR, Hou Y, Wyche TP, Bugni TS (2017) Natural Product Discovery Using Planes of Principal Component Analysis in R (PoPCAR). Metabolites 7:34. https://doi.org/10.3390/metabo7030034

Chevrette MG, Gutiérrez-García K, Selem-Mojica N, Aguilar-Martínez C, Yañez-Olvera A, Ramos-Aboites HE, Hoskisson PA, Barona-Gómez F (2020) Evolutionary dynamics of natural product biosynthesis in bacteria. Nat Prod Rep 37:566-599. https://doi.org/10.1039/C9NP00048H

Crnkovic CM, May DS, Orjala J (2018) The impact of culture conditions on growth and metabolomic profiles of freshwater cyanobacteria. J Appl Phycol 30:375–384. https://doi.org/10.1007/s10811-017-1275-3

da Silva RR, Wang M, Nothias LF, van der Hooft JJJ, Caraballo-Rodríguez AM, Fox E, Balunas MJ, Klassen JL, Lopes NP, Dorrestein PC (2018) Propagating annotations of molecular networks using in silico fragmentation. PLOS Comput Biol 14:e1006089. https://doi.org/10.1371/journal.pcbi.1006089

Engene N, Rottacker EC, Kaštovský J, Byrum T, Choi H, Ellisman MH, Komárek J, Gerwick WH (2012) *Moorea producens* gen. nov., sp. nov. and *Moorea bouillonii* comb. nov., tropical marine cyanobacteria rich in bioactive secondary metabolites. Int J Syst Evol Microbiol 62:1171–1178. https://doi.org/10.1099/ijs.0.033761-0

Ernst M, Kang KB, Caraballo-Rodríguez AM, Nothias LF, Wandy J, Chen C, Wang M, Rogers S, Medema MH, Dorrestein PC, van der Hooft JJJ (2019) MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools. Metabolites 9:144. https://doi.org/10.3390/metabo9070144

Fernie AR, Stitt M (2012) On the Discordance of Metabolomics with Proteomics and Transcriptomics: Coping with Increasing Complexity in Logic, Chemistry, and Network Interactions Scientific Correspondence. Plant Physiol 158:1139–1145. https://doi.org/10.1104/pp.112.193235

Hoffmann T, Krug D, Bozkurt N, Duddela S, Jansen R, Garcia R, Gerth K, Steinmetz H, Müller R (2018) Correlating chemical diversity with taxonomic distance for discovery of natural products in myxobacteria. Nat Commun 9:803. https://doi.org/10.1038/s41467-018-03184-1

Leber CA, Naman CB, Keller L, Almaliti J, Caro-Diaz EJE, Glukhov E, Joseph V, Sajeevan TP, Reyes AJ, Biggs JS, Li T, Yuan Y, He S, Yan X, Gerwick WH (2020) Applying a Chemogeographic Strategy for Natural Product Discovery from the Marine Cyanobacterium *Moorena bouillonii*. Mar Drugs 18:515. https://doi.org/10.3390/md18100515

Luesch H, Yoshida WY, Moore RE, Paul VJ, Corbett TH (2001) Total Structure Determination of Apratoxin A, a Potent Novel Cytotoxin from the Marine Cyanobacterium *Lyngbya majuscula*. J Am Chem Soc 123:5418–5423. https://doi.org/10.1021/ja010453j

Luzzatto-Knaan T, Garg N, Wang M, Glukhov E, Peng Y, Ackermann G, Amir A, Duggan BM, Ryazanov S, Gerwick L, Knight R, Alexandrov T, Bandeira N, Gerwick WH, Dorrestein PC (2017) Digitizing mass spectrometry data to explore the chemical diversity and distribution of marine cyanobacteria and algae. eLife 6:e24214. https://doi.org/10.7554/eLife.24214.001

Maloney KN, Botts RT, Davis TS, Okada BK, Maloney EM, Leber CA, Alvarado O, Brayton C, Caraballo-Rodríguez AM, Chari JV, Chicoine B, Crompton JC, Davis SR, Gromek SM, Kurnianda V, Quach K, Samples RM, Shieh V, Sultana CM, Tanaka J, Dorrestein PC, Balunas MJ, McFadden CS (2020) Cryptic Species Account for the Seemingly Idiosyncratic Secondary Metabolism of *Sarcophyton glaucum* Specimens Collected in Palau. J Nat Prod 83:693-705. https://doi.org/10.1021/acs.jnatprod.9b01128

Mohimani H, Gurevich A, Shlemov A, Mikheenko A, Korobeynikov A, Cao L, Shcherbin E, Nothias LF, Dorrestein PC, Pevzner PA (2018) Dereplication of microbial metabolites through database search of mass spectra. Nat Commun 9:4035. https://doi.org/10.1038/s41467-018-06082-8

Moss NA, Seiler G, Leão TF, Castro-Falcón G, Gerwick L, Hughes CC, Gerwick WH (2019) Nature's Combinatorial Biosynthesis Produces Vatiamides A–F. Angew Chem Int Ed 58:9027–9031. https://doi.org/10.1002/anie.201902571

Myers OD, Sumner SJ, Li S, Barnes S, Du X (2017) One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks. Anal Chem 89:8696–8703. https://doi.org/10.1021/acs.analchem.7b00947

Naman CB, Rattan R, Nikoulina SE, Lee J, Miller BW, Moss NA, Armstrong L, Boudreau PD, Debonsi HM, Valeriote FA, Dorrestein PC, Gerwick WH (2017) Integrating molecular networking and biological assays to target the isolation of a cytotoxic cyclic octapeptide, samoamide A, from an American Samoan marine cyanobacterium. J Nat Prod 80:625–633. https://doi.org/10.1021/acs.jnatprod.6b00907

Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, De Los Santos ELC, Yeong M, Cruz-Morales P, Abubucker S, Roeters A, Lokhorst W, Fernandez-Guerra A, Cappelini LTD, Goering AW, Thomson RJ, Metcalf WW, Kelleher NL, Barona-Gomez F, Medema MH (2019) A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol 16:60–68. https://doi.org/10.1038/s41589-019-0400-9

NOAA's environmental sensitivity index: Guam and the Northern Mariana Islands shapefile. (2005) National Oceanic and Atmospheric Administration (NOAA), National Ocean Service, Office of Response and Restoration, Emergency Response Division, Seattle, Washington. https://response.restoration.noaa.gov/maps-and-spatial-data/download-esi-maps-and-gis-data.html

Nothias LF, Petras D, Schmid R, Dührkop K, Rainer J, Sarvepalli A, Protsyuk I, Ernst M, Tsugawa H, Fleischauer M, Aicheler F, Aksenov AA, Alka O, Allard PM, Barsch A, Cachet X, Caraballo-Rodriguez AM, Da Silva RR, Dang T, Garg N, Gauglitz JM, Gurevich A, Isaac G, Jarmusch AK, Kameník Z, Kang KB, Kessler N, Koester I, Korf A, Le Gouellec A, Ludwig M, Martin H C, McCall LI, McSayles J, Meyer SW, Mohimani H, Morsy M, Moyne O, Neumann S, Neuweger H, Nguyen NH, Nothias-Esposito M, Paolini J, Phelan VV, Pluskal T, Quinn RA, Rogers S, Shrestha B, Tripathi A, van der Hooft JJJ, Vargas F, Weldon KC, Witting M, Yang H, Zhang Z, Zubeil F, Kohlbacher O, Böcker S, Alexandrov T, Bandeira N, Wang M, Dorrestein PC (2020) Feature-based Molecular Networking in the GNPS Analysis Environment. Nat Methods 17:905–908. https://doi.org/10.1038/s41592-020-0933-6

Noyer C, Thomas OP, Becerro MA (2011) Patterns of Chemical Diversity in the Mediterranean Sponge *Spongia lamella*. PLOS ONE 6:e20844. https://doi.org/10.1371/journal.pone.0020844

Okada BK, Wu Y, Mao D, Bushin LB, Seyedsayamdost MR (2016) Mapping the Trimethoprim-Induced Secondary Metabolome of *Burkholderia thailandensis*. ACS Chem Biol 11:2124–2130. https://doi.org/10.1021/acschembio.6b00447

Pandya C, Farelli JD, Dunaway-Mariano D, Allen KN (2014) Enzyme Promiscuity: Engine of Evolutionary Innovation. J Biol Chem 289:30229-30236. https://doi.org/10.1074/jbc.R114.572990

Pluskal T, Castillo S, Villar-Briones A, Orešič M (2010) MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11:395. https://doi.org/10.1186/1471-2105-11-395

Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Linington RG (2016) Retrospective analysis of natural products provides insights for future discovery trends. Proc Natl Acad Sci USA 114:5601–5606. https://doi.org/10.1073/pnas.1614680114

Rutledge PJ, Challis GL (2015) Discovery of microbial natural products by activation of silent biosynthetic gene clusters. Nat Rev Microbiol 13:509–523. https://doi.org/10.1038/nrmicro3496

Selander E, Kubanek J, Hamberg M, Andersson MX, Cervin G, Pavia H (2015) Predator lipids induce paralytic shellfish toxins in bloom-forming algae. Proc Natl Acad Sci USA 112:6395-6400. https://doi.org/10.1073/pnas.1420154112

Serrano R, González-Menéndez V, Rodríguez L, Martín J, Tormo JR, Genilloud O (2017) Co-culturing of Fungal Strains Against *Botrytis cinerea* as a Model for the Induction of Chemical Diversity and Therapeutic Agents. Front Microbiol 8:649. https://doi.org/10.3389/fmicb.2017.00649

Shannon P, Markiel1 A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski1 B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–504. https://doi.org/10.1101/gr.1239303

Singh SB, Pelaez F (2008) Biodiversity, chemical diversity and drug discovery. In: Petersen F, Amstutz R (eds.) Natural Compounds as Drugs Volume I. Progress in Drug Research, Birkhäuser Basel vol 65. https://doi.org/10.1007/978-3-7643-8117-2_4

van der Hooft JJJ, Wandy J, Barrett MP, Burgess KEV, Rogers S (2016) Topic modeling for untargeted substructure exploration in metabolomics. Proc Natl Acad Sci USA 113:13738-13743. https://doi.org/10.1073/pnas.1608041113

Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu WT, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu CC, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw CC, Yang YL, Humpf HU, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, Boya P CA, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodríguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard PM, Phapale P, Nothias LF, Alexandrov T, Litaudon M, Wolfender JL, Kyle JE, Metz TO, Peryea T, Nguyen DT, VanLeer D, Shinn P, Jadhav A, Müller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BØ, Pogliano K, Linington RG, Gutiérrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34:828–837. https://doi.org/10.1038/nbt.3597

Chapter 5: Improving short read assemblies for comparative genomic insights in *Moorena bouillonii*

5.0: Abstract

The cyanobacterium *Moorena bouillonii* is known to be a prolific producer of natural products. More can be learned about the chemical diversity that it has the potential to produce by studying the biosynthetic gene clusters (BGCs) present in its genome. Long-read sequencing of a *M. bouillonii* culture from Papua New Guinea, followed by de novo assembly with Flye yielded the first complete genome for the species. Short-read sequences of nineteen potential *M. bouillonii* samples were assembled and iteratively scaffolded using a newly developed bioinformatics software called iTerator. iTerator significantly improved draft genome contiguity and BGC completeness. Biosynthetic diversity analyses revealed that *M. bouillonii* chemodiversity appears, from a genomic perspective, to be more highly concentrated in non-ribosomal peptide synthetase (NRPS) and polyketide synthetase (PKS) pathways, than in terpene and ribosomally synthesized and post-translationally modified peptides (RiPPs) pathways.

5.1: Introduction

Cyanobacteria are a prolific source of biologically active and chemically diverse natural products; as of this writing (Fall 2020), they are reported to be the source of over 800 natural products (MarinLit). These natural products are not evenly distributed across the numerous taxa that make up the Cyanobacteria phylum. In fact, the majority of these have been attributed to what used to be morphologically defined as the genus *Lyngbya* (Engene et al 2011; Engene et al 2012). The outlying secondary metabolite productivity of *Lyngbya* was made even more worthy of skepticism in that natural products were predominantly being discovered in marine *Lyngbya* spp., particularly *L. majuscula* (Engene et al 2011). After the incorporation of phylogenetics into cyanobacterial taxonomy (Hoffmann, Komárek, and Kaštovský 2005), it was determined that *Lyngbya* as well as two other natural product rich genera, *Oscillatoria* and *Symploca*, were indeed polyphyletic groups (Engene et al 2011; Engene et al 2013a), and researchers set about reorganizing cyanobacterial taxa to align with the new phylogenetic insights. Their efforts have led to the reassignment of several clades previously considered a part of *Lyngbya* into the new genera *Moorea* (Engene et al 2012), *Okeania* (Engene et al 2013b), and *Dapis* (Engene, Tronholm, and Paul 2018). The genus *Moorea*, which contains the species *Moorena bouillonii* (L.Hoffmann & Demoulin) Engene & Tronholm 2019 (Oscillatoriaceae) and *M. producens*, has since been renamed to *Moorena* (Tronholm and Engene 2019).

Even with the remarkable advances that have been made applying phylogenetic classification to bring more clarity to cyanobacterial taxonomy and the continued interest in *Moorena* spp. as significant producers of natural products, only a very small amount of research has been dedicated to *Moorena* genomics, *M. bouillonii* in particular. To date, only one high quality draft genome of *M. bouillonii* has been published (Leão et al 2017), and it comes from the reference strain of the species (Engene et al 2012); genomic information for *M. producens* is only slightly more available, with two high quality draft genomes and one complete genome (Leão et al 2017). Much of the *M. bouillonii* draft genome is contained in a linear scaffold of 8.23 Mb consisting of 291 contigs with approximately 32,000 Ns (Leão et al 2017). The remainder of the genome is comprised of 12 small unmapped scaffolds (1.6-16.7 kb) (Leão et al 2017). The *M. bouillonii* genome contains 31 biosynthetic gene clusters (BGCs), accounting for about 15% of its genome. The products of three of these BGCs have been identified. They include the BGCs for lyngbyabellin A (Monroe et al 2012), the apratoxins (Grindberg et al 2011), and the columbamides (Kleigrewe et al 2016). Beyond additional work focused on the structural biology and biosynthetic mechanism of the polyketide loading module that initiates the biosynthesis of apratoxin (Skiba et al 2017; Skiba et al 2018), there is little more known about the genomics or biosynthetic machinery of *M. bouillonii*.

Learning more about *M. bouillonii* genomics necessarily has a number of challenges and complications. First and foremost, *M. bouillonii* is a relatively newly described species, having only been defined in 1991 under the name *L. bouillonii* (Hoffmann and Demoulin 1991). Along with the aforementioned taxonomic issues with *Lyngbya* and the multiple name changes, this has resulted in the fairly small body of literature, and its associated phylogenetic data, that can be reasonably attributed to *M. bouillonii* to be riddled with misnomers, adding confusion. An additional contributor to the misidentification of *Moorena* spp. in the scientific literature is a lack of phylogenetic resolution between the species. Neither the 16S rRNA gene, nor the 16S–23S ITS region offers the degree of divergence necessary to reliably differentiate *M. bouillonii* from *M. producens* (Engene et al 2012; Curren and Leong 2019), an issue made even more complex by the multiple, heterogenous copies of the 16S rRNA gene in these cyanobacteria

(Engene et al 2010; Engene and Gerwick 2011). One might consider applying multilocus sequence typing (MLST) to overcome this issue, however forthcoming work suggests that even this may not be sufficient for resolving *M. bouillonii* from *M. producens* (Leão et al 2020 [*in prep*]). Beyond all of this, there remains technical challenges for the extraction, sequencing, and assembly of *Moorena* genomes. *Moorena* filaments are surrounded by a thick polysaccharide sheath that is colonized by a diversity of heterotrophic bacteria (Engene et al 2012), including seemingly obligate symbionts (Cummings et al 2016). This means that beyond isolating singular cells from *Moorena* filaments (Grindberg et al 2011), any DNA extraction from *Moorena* spp., regardless of whether the biomass was obtained from the environment or from a mono-cyanobacterial culture, is likely to be a metagenomic sample.

A large-scale sequencing effort was conducted by the Gerwick Lab, in collaboration with the Knight Lab at UCSD, acquiring short-read data for hundreds of cyanobacterial samples, including samples of *M. bouillonii*. As is often the case with such short-read sequencing efforts, as well as with metagenomic samples (Navarro-Muñoz et al 2019), many of the resultant assemblies were fragmented. In order to improve the contiguity of these genomes and therefore facilitate deeper insights into *M. bouillonii* genomics, a bioinformatic pipeline called iTerator (iT) was developed which optimizes assembly outcomes, and iteratively scaffolds and extends contigs while directly removing gaps. The motivation in this study for endeavoring to increase genome contiguity was to provide a clearer view of the biosynthetic potential of *M. bouillonii* by increasing the number of complete BGCs present in each sample. These development efforts are described herein, along with biosynthetic gene cluster (BGC) distributional analyses performed upon the outputted draft genomes. As a result of these investigations, a collection of draft genomes was significantly improved in terms of contiguity and BGC completeness, and analyses of these genomes revealed *M. bouillonii* chemodiversity to be concentrated in non-ribosomal peptide synthetase (NRPS) and polyketide synthase classes of natural products.

5.2: Results & Discussion

5.2.1: Complete *Moorena bouillonii* reference genome

Biomass from a Papua New Guinea culture of *M. bouillonii* (referred to throughout by the sample code PNG), the same organism from which the aforementioned high-quality draft genome was assembled, was extracted and sequenced with PacBio RS to produce long reads. Assembly of these long reads with Flye (Lin et al 2016; Kolmogorov et al 2019) resulted in a contiguous circularized chromosome 8.77 Mb in size, along with three circularized plasmids sized 21 kb, 714 bp, and 664 bp. This represents the second complete genome from *Moorena* spp. (Leão et al 2017) and the first for *M. bouillonii*, significantly improving upon the previous best draft genome for the species.

5.2.2: Multi-assembly of genomes for optimized outcomes

In module 1 of iT (Figure 5.1), each set of sample reads was run through a bioinformatics workflow involving the generation of twelve SPAdes (Bankevich et al 2012; Nurk et al 2017; Prjibelski et al 2020) assemblies based on combinations of three assembly modes (standard, isolate, and meta) and four kmer sets (1st: 21, 33, 55; 2nd: 21, 33, 55, 77; 3rd: 21, 33, 55, 77, 99; 4th: 21, 33, 55, 77, 99, 121). The best of the twelve assemblies was selected for moving forward based on which produced the maximum longest contig, with the option to select based on alternative metrics such as number of contigs, total length, N50, L50, etc. Across the 16 sets of reads that were successfully run through this assembly workflow, 8 different assembly parameter combinations were found to be most optimal (Table 5.S1). This illustrates that different samples benefit from different parameter sets and suggests that there is benefit to tuning assembly parameters on a per sample basis. The most successful assembly parameter sets were standard and isolate assembly modes paired with the 4th kmer set, each accounting for 4 samples. With another sample producing its maximum longest contig when the 4th kmer set was paired with the meta assembly mode, 9 of the 16 samples produced optimal assemblies with the 4th kmer set. Another 3 samples produced optimal assemblies with the 3rd kmer set, indicating that using larger k values, though computationally more expensive, can improve assembly outcomes. Seven optimal assemblies were generated with standard SPAdes, five with isolate mode, and four with meta mode. This variability is likely driven by the proportion of heterotrophic bacterial DNA captured during extraction, and, in the case of field samples, the degree to which the sample was dominated by *M. bouillonii* or part of a more diverse assemblage. Three samples (sample codes 4A5, 4G3, and CL2) failed to progress through the assembly workflow, triggering computational issues with memory allocation. It is unclear whether those issues stemmed from memory limitations of the hardware used for computation or represent issues intrinsic to SPAdes version 3.14.0 in allocating memory resources under certain circumstances. Regardless, a single assembly was produced for each of these three samples using SPAdes version 3.13.1 in standard mode with the 2nd kmer set.

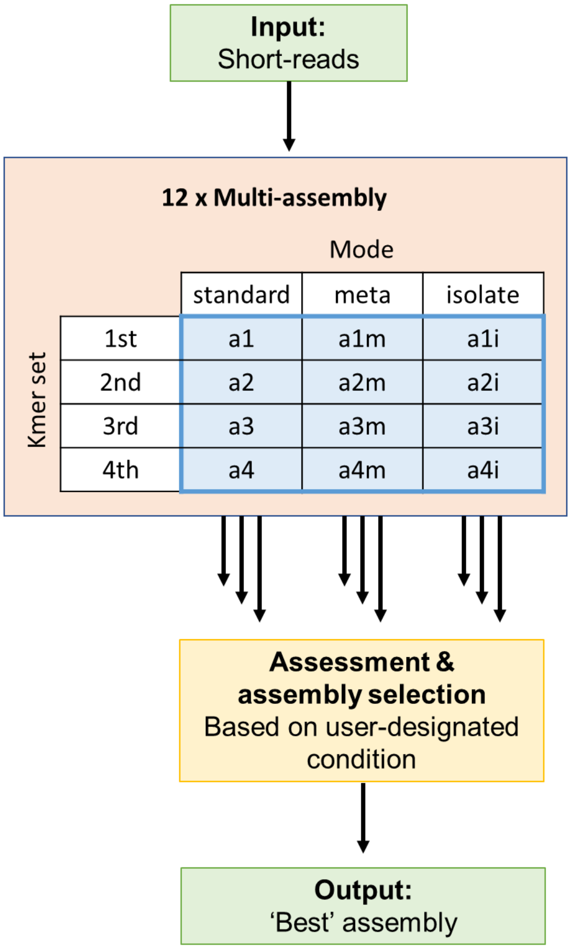


Figure 5.1 – Diagram of iT Module 1 workflow

5.2.3: Improvements to contiguity and BGC completeness via iterative scaffolding and gap filling

Module 2 of iT (Figure 5.2) involves four steps: contig extension and de novo scaffolding with SSPACE (Boetzer et al 2011), multi-reference scaffolding with Multi-CSAR (Chen, Shen, and Lu 2018), gap filling with GapFiller (Boetzer and Pirovano 2012), and the removal of any remaining gaps to make all scaffolds contiguous. These four steps are iterated through until the number of contigs in the draft assembly stops decreasing and stabilizes. In running Multi-CSAR, two reference genomes were used: PNG and a complete *Moorena* genome previously found to be very similar to *M. bouillonii* (Leão et al 2017; GenBank accession: CP017600.1). This reference genome is referred to as PAL. Due to time constraints, the process was truncated for 6 samples after 19 iterations, prior to contig count stabilization. Of these 6 samples, 4 contained too many contigs less than 1000 bp to progress through module 2, and thus required filtering of contigs to only include those at least 1000 bp in length.

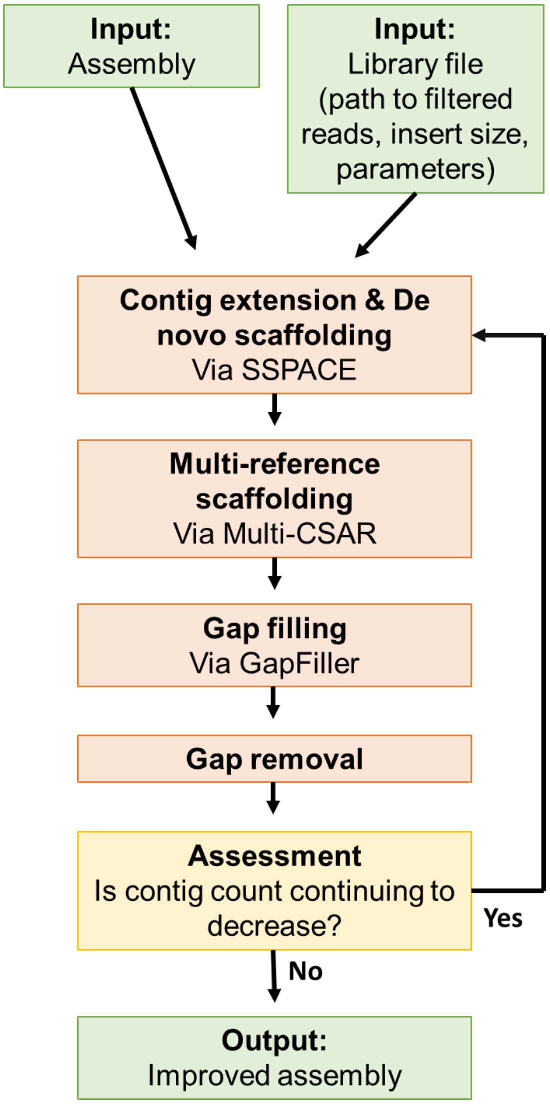

Input:
Assembly
Input:
Library file
(path to filtered reads, insert size, parameters)
Contig extension & De novo scaffolding
Via SSPACE
Multi-reference scaffolding
Via Multi-CSAR
Gap filling
Via GapFiller
Gap removal
Assessment
Is contig count continuing to decrease?
Yes
No
Output:
Improved assembly

Figure 5.2 – Diagram of iT Module 2 workflow

Module 2 of iT proved successful in significantly improving assembly contiguity. Eighteen out of the 19 samples saw reductions in the number of contigs ranging from 33.6% to 99.9% (Table 5.1). Excluding the 6 samples whose runs were ended after 19 iterations, reductions in the number of contigs range from 96.5% to 99.9%. A substantial proportion of contig consolidation and reduction occurs during the first iteration; amongst samples that were allowed to run until they reached a stable contig count, decreases in contig count ranged from 72.0% to 98.9%. Even after sizable initial reductions in contig count, subsequent iterations, sometimes numbering into the hundreds, continued to consolidate contigs with consistent efficacy amongst samples. Decreases in contig count between the 1st iteration and the last iteration, considering only samples that proceeded past 19 iterations, ranged from 74.0 % to 95.1%.

Table 5.1 - Contig counts for each original assembly, after 1 iteration, and after the final iteration.

sample code	original assembly	after 1 iteration	after final iteration	number of iterations	% decrease [original to final]	% decrease [original to iteration 1]	% decrease [iteration 1 to final]
1C3	48985	8769	984	305	98.0	82.1	88.8
1C9	264291	12659	1778	288	99.3	95.2	86.0
1D4	243533	3535	3043	16	98.8	98.5	13.9
1F2	126149	1387	185	159	99.9	98.9	86.7
3A2	20843	576	150	89	99.3	97.2	74.0
3F7	29534	977	237	129	99.2	96.7	75.7
3H8	38291	8565	992	230	97.4	77.6	88.4
3I6	66182	6591	803	139	98.8	90.0	87.8
3I7	40629	10121	898	178	97.8	75.1	91.1
3I8	36777	5136	250	137	99.3	86.0	95.1
3I9	42376	6341	1485	272	96.5	85.0	76.6
4A1	104137	14700	1374	181	98.7	85.9	90.7
4A3[1,2]	8746	6672	5599	19	36.0	23.7	16.1
4A4	30289	8496	919	218	97.0	72.0	89.2
4A5[1,2]	6275	5697	4166	19	33.6	9.2	26.9
4G2[1,2]	4432	6829	6706	19	-51.3	-54.1	1.8
4G3[1]	144820	4396	2486	19	98.3	97.0	43.4
CL1[1]	25410	9971	4196	19	83.5	60.8	57.9
CL2[1,2]	6573	2335	1887	19	71.3	64.5	19.2

[1] iT run truncated to 19 iterations
[2] 1000 bp minimum contig length filter applied to assembly

Along with improvements in contiguity, iT also helped in increasing BGC completeness. 13 out of the 19 samples increased in the number of complete BGCs that they were annotated to have (Figure 5.3). For all samples, the number of partial BGCs decreased from the original assembly to the final iteration (Table 5.S2), suggesting that even when no complete BGCs were revealed through iT, there was consolidation of partial BGCs occurring. This is further supported by comparing counts of partial BGCs of which one edge is captured within a contig (referred to here as 'edge BGCs') (Figure 5.S1). Out of the 6 samples which saw no increases in complete BGCs from original assembly to final iteration, 3 did show increases in the number of edge BGCs.

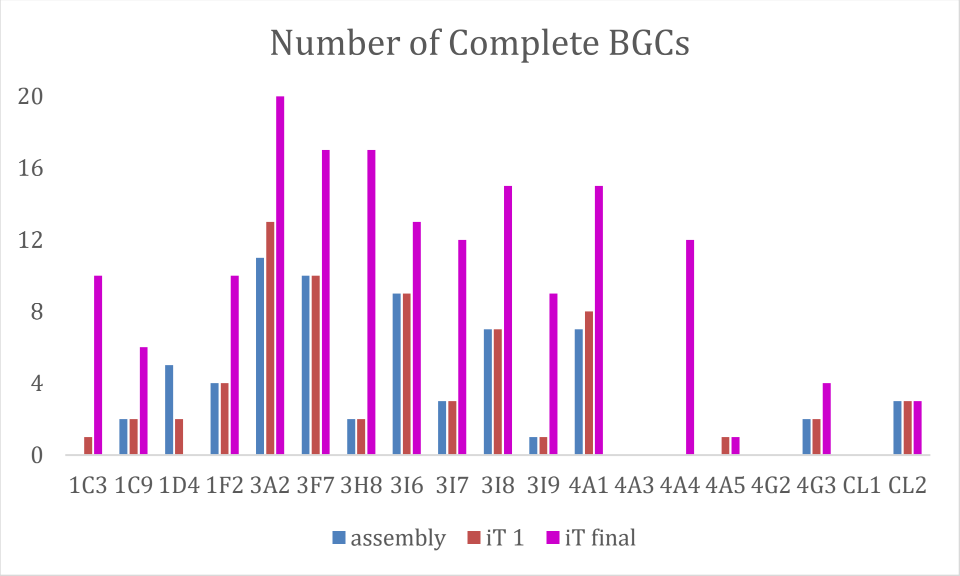


Figure 5.3 – Changes in the number of complete BGCs from original assemblies to final iterations.

iT was not uniformly successful, but deficient results can be explained by factors external to its functioning, namely quality of sample sequences. Sample 1D4 reported a decrease in complete BGCs, minimal improvements in contiguity from its first iteration to its final, and it terminated after only 16 iterations. Its 'best' assembly was produced using 'meta' mode in SPAdes, and an inspection of its assembly statistics revealed a GC content of 53.8%, the highest of all 19 samples. Furthermore, many of the BGCs detected in the assembly, including all 5 complete BGCs, were annotated as having strong similarity to BGCs from various heterotrophic bacterial genera, including *Labrenzia*, *Hyphomonas*, and *Methylovulum*. This suggests that 1D4 was a complex metagenomic sample, and failure of a large portion of its reads to map to the *Moorena* reference genomes used in this study resulted in a suboptimal product draft genome. Implementing more flexible parameters for determining contig count stabilization and so allowing more iterations may improve the draft assembly, but 1D4 may also be exposing a general limitation of iT in handling complex metagenomes. This could be addressed by adding a binning step prior to module 2, and could be implemented using DarkHorse (Podell and Gaasterland 2007) or Autometa (Miller et al 2019)

Samples 4A3, 4A5, 4G2, and CL2 reported the four worst percentage decreases in contig counts between assembly and final iteration, with 4G2 actually increasing in contig count. This is partially a function of these four assemblies requiring a 100 bp minimum contig length filter applied prior to iT module 2. Filtering was necessary due to the highly fragmented nature of the assemblies (Table 5.S3) very large number of small contigs included in these assemblies causing colossal memory usage that made proceeding without filtration computationally intractable. It is likely that allowing for more iterations with these samples, along with the other two samples that terminated after 19 iterations, would result in additional improvements in contiguity and BGC completeness. It is also possible that these four genomes illustrate an upper bound in fragmentation that can be effectively handled by iT.

The results reported above illustrate iT's success as a proof-of-concept; iterative scaffolding and gap filling with the removal of extraneous gaps produces more contiguous draft genomes with more complete BGCs. iT is, however, far from optimized, computationally speaking. SSPACE (Boetzer et al 2011) and GapFiller (Boetzer and Pirovano 2012) are older pieces of software that require relatively large amounts of memory and make poor use of parallel processing. Blossom V (Kolmogorov 2009), which is used by Multi-CSAR (Chen, Shen, and Lu 2018) to merge scaffolds from different references, requires substantially increased amounts of time with increasing numbers of contigs. These inefficiencies, and the large numbers of iterations often required to reach contig count stabilization, result in a very slow process. As it is clear that iT creates measurable improvements to short-read derived draft genomes, its computational deficiencies represent opportunities for future development of a potentially valuable bioinformatic tool.

5.2.4: BGC distributions

Across the 19 *M. bouillonii* draft genomes and the complete PNG genome, along with 3 high-quality *M. producens* genomes (PAL, JHB [GenBank accession: CP017708.1], and 3L[GenBank accession: MKZR00000000.1]), 1214 partial and complete BGCs were clustered using BiGSCAPE (Navarro-Muñoz et al 2019) into 160 families (Table 5.2). A total of 528, or approximately 43% of the BGCs, were singletons, meaning that they did not cluster with any other BGCs in the dataset. This number is likely inflated by the large number of partial BGCs included in the dataset, but is nevertheless striking, as it suggests a significant degree of biosynthetic diversity in this predominantly *M. bouillonii* genomic dataset.

Biosynthetic diversity is not equally distributed across the dataset. Terpene BGCs appear to be heavily conserved; out of 80 terpene BGCs, only 12 are singletons. The remaining 68 BGCs are split across only 9 families, with the largest family including 14 BGCs from 14 individual samples. A closely related family considered to be in the same 'clan' includes another 7 BGCs. Another way to consider the level of conservation among clusters is to inspect the ratio of links-to-families. Links are connections between BGCs based on their pairwise distance scores being below a designated cutoff (0.30, in this case). A high links-to-families ratio would indicate that when BGCs are clustered into families, many BGCs are similar to many other BGCs in those families. The global links-to-families ratio for the entire dataset is 10.73; the links-to-families ratio for terpenes is 43.67, providing yet another metric by which terpenes BGCs appear to be very well conserved. Not quite as dramatic as terpenes, but similarly well-conserved in terms of similarity are the RiPP BGCs. A total of 275 BGCs are split across 39 families, with only 70 singletons.

Table 5.2 - Summary of BiGSCAPE output

	total BGCs	**max BGCs per family**	**singletons**	**num families**	**links**	**links/families**	**singletons/total BGCs**
NRPS	397	10	216	49	293	5.98	0.54
Others	134	11	29	23	194	8.43	0.22
PKS type I	80	3	60	9	13	1.44	0.75
PKS other	146	11	83	20	102	5.10	0.57
Terpene	80	14	12	9	393	43.67	0.15
PKS-NRPS hybrids	102	9	58	11	78	7.09	0.57
RiPPs	275	15	70	39	644	16.51	0.25
Total	1214	73	528	160	1717	10.73[1]	0.43[2]

[1] Total links divided by total number of families
[2] Total number of singletons divided by total number of BGCs

At the other end of the biosynthetic diversity spectrum lies the type I PKS BGCs. A total of 60 out of 80, or a full 75% of type I PKS BGCs captured in this dataset, were found to be singletons. With the remaining 20 BGCs divided amongst 9 families, and a links-to-families ratio of 1.44, connectivity indicative of similarity is very poor. Notably, while there were 80 BGCs of both the terpene and type I PKS classes in the dataset, PNG was not detected to contain any pure type I PKS BGCs but contained 3 terpene BGCs. This relative scarcity, along with the general scarcity of type I PKS natural products described from *M. bouillonii*, suggests that the high degree of diversity in type I PKS BGCs could be, in part, artificial. Partial type I PKS BGCs may have related BGCs classified as NRPS-PKS hybrids. The same can be said for other PKS and NRPS BGCs, the classes with the second and third lowest links-to-families ratios, respectively. While not nearly as apparently diverse as the classes from which it is derived, the NRPS-PKS hybrid class does boast a 57% singleton rate and is composed of 44 BGCs spread across 11 families.

In partial refutation of concerns of artificial diversity amongst PKS and NRPS classes, one can look to how BGCs from PAL and PNG are distributed in families. As the only two complete genomes in the dataset, their genome completeness indicates BGC completeness. Complete BGCs are necessarily the most likely to be included in clusters, as they contain the most information; this provides increased opportunities for partial BGCs to align with complete BGCs, and so initiate a cluster. For type I PKS, PAL has two BGCs classified as such, both of which fall into families. For other PKS classes, 8 out of 11 PAL BGCs and 4 out of 5 PNG BGCs are singletons. A total of 5 of 18 and 5 of 12 NRPS BGCs for PAL and PNG, respectively, are singletons, with several others found in small families of only two or three BGCs. This propensity for PNG and PAL complete BCGs to be found as singletons provides an orthogonal line of evidence suggesting that biosynthetic diversity is more prominent among NRPS and PKS related classes of natural products in *M. bouillonii*, and perhaps *Moorena* more generally. Studying the distribution of BGCs across samples of *M. bouillonii* presents a paradoxical pair of trends. As discussed in the introduction, *M. bouillonii* and *M. producens* are very similar phylogenetically; 16S rRNA sequences and even MLST fail to resolve the two species. This is reflected, to some extent in the distribution of BGCs, as numerous *M. bouillonii* BGCs can be found clustered into families with BGCs from the three *M. producens* genomes included in this study. Opposingly, as discussed in detail in Chapter 4 of this dissertation, *M. bouillonii* is very diverse in its chemical composition, with high variability between, and in some cases even within, geographic locations. This too is reflected in the distribution of BGCs, with NRPS and PKS classified BGCs in particular showing a high degree of diversity.

5.3: Methods

5.3.1: Sample collection and selection

Cyanobacterial biomass was collected via snorkeling or scuba diving in tropical marine coral reef environments (for metadata, please see Table 5.S4). Biomass was preserved in RNA-later and/or transported back to the laboratory in a culture flask of seawater, in order to establish non-axenic mono-cyanobacterial cultures (Moss et al 2018).

Twenty samples of cyanobacteria, both field-collected and laboratory-cultured, were selected for study. Samples were selected based on collection location, putative field identification, woven ‘cobweb’ morphology, and shrimp association; some or all of these factors were used in each case for tentatively identifying each sample as *Moorena bouillonii*. In some cases, 16S rRNA gene sequence information and/or Mash (Ondov et al 2016; Leão, unpublished) taxonomic classification were available, these were considered only in a secondary sense, as 16S rRNA sequencing lacks the phylogenetic resolution necessary to differentiate *M. bouillonii* from its congener *Moorena producens* (Engene et al 2012), and Mash taxonomic classification went only to the genus level. The dataset was compiled via an inclusive approach, opting to retain edge cases rather than discard them.

5.3.2: DNA extraction

RNA-later preserved samples and cultured biomass were frozen with liquid nitrogen, and ground with mortar and pestle. DNA extraction was conducted according to the “QIAGEN Bacterial Genomic DNA Extraction Kit” protocol with modification; incubation times were increased by 1 h and 10 µL of Proteinase K was used (10 mg/mL). For more information on the extraction protocol, see Moss et al (2018).

5.3.3: Library preparation and sequencing

For the majority of samples sequenced, libraries were generated using a miniaturized version of the Kapa HyperPlus Illumina-compatible library prep kit (Kapa Biosystems®). Extracted DNA was normalized to 5 ng total input per sample in an Echo 550 acoustic liquid handler (Labcyte Inc). A Mosquito HTS liquid-handler (TTP Labtech Inc) was used for 1/10 scale enzymatic fragmentation, end-repair, and adapter-ligation reactions. Sequencing adapters were added according to the iTru protocol (Glenn et al 2019) in which short universal adapter segments are ligated, followed by the addition of sample-specific barcoded sequences in a subsequent PCR step. Amplified and barcoded libraries were quantified using the PicoGreen assay and were pooled in approximately equimolar ratios and were sequenced on an Illumina HiSeq 4000 instrument to >30X metagenomic coverage. PNG was sequenced via PacBio RS® using a 10kb library prep. Two samples from Guam, referred to by sample codes CL1 and CL2 were separately prepared and sequenced as follows: 500 ng of genomic DNA from each sample was fragmented by Adaptive Focused Acoustics (E220 Focused Ultrasonicator, Covaris; Woburn, MA) to produce an average fragment size of 350 bp. Fragmented DNA was purified using the Agencourt AMPure XP beads (Beckman Coulter; Fullerton, CA) and sequencing libraries were generating using the KAPA Hyper Prep Kit (KAPA Biosystems, Wilmington, MA) following manufacturer's instructions and using 9 cycles of amplification. Library quality was assessed using High Sensitivity D1000 kit on a 2200 TapeStation instrument (Agilent Technologies, Santa Clara, CA). Sequencing was performed using an Illumina MiSeq System (Illumina, San Diego, CA), generating 150 bp paired end reads.

5.3.4: Assembly of *M. bouillonii* reference genome

PNG PacBio long-reads were assembled using Flye (Lin et al 2016; Kolmogorov et al 2019). The PacBio .bam source file was converted to fastq using reformat.sh from bbmap (Bushnell 2014), and submitted uncorrected to Flye. The reads were designated pacbio-raw and the genome size was set to 8m.

5.3.5: iTerator bioinformatics pipeline

iT is written in Python 3 (Van Rossum and Drake 1995) and is built off of pandas (0.25.2) (McKinney 2010; Pandas-Dev/Pandas 2018). It is available as a Jupyter notebook (Pérez and Granger 2007; Kluyver et al 2016) at https://github.com/c-leber/iTerator. Module one of iT generates 12 assemblies using different parameter sets (Table 5.S5) in SPAdes (Bankevich et al 2012; Nurk et al 2017; Prjibelski et al 2020), followed by assembly assessments, and selection of the optimal assembly based on a user-selected parameter (maximum longest contig, in the case of this study). All assemblies are annotated with Prokka (Seemann 2014) and the number of open reading frames (ORFs) is counted for each (Podell, unpublished). QUAST (Gurevich et al 2013) is used to generate statistics for each assembly, and bbmap (Bushnell 2014) is used to check assembly coverage. The results of all of these assembly assessments are compiled and saved for future reference.

Module two of iT involves iterating through contig extension and de novo scaffolding with SSPACE (Boetzer et al 2011), multi-reference scaffolding with Multi-CSAR (Chen, Shen, and Lu 2018), gap filling with GapFiller (Boetzer and Pirovano 2012), and gap removal. The library files needed for SSPACE and GapFiller were manually prepared for each sample, with BWA set as the aligner, minimum allowed error for insert size set to 0.75, and read orientation indicated to be 'FR'. The library files included paths to sample reads trimmed to remove adapters in Trimmomatic (Bolger, Lohse, and Usadel 2014) with the following parameters: ILLUMINACLIP:Trimmomatic-0.38/adapters/TruSeq3-PE.fa:2:30:10:8:true, ILLUMINACLIP:Trimmomatic-0.38/adapters/TruSeq3-PE-2.fa:2:30:10:8:true. Insert sizes were determined by running each sample through an independent SSPACE run with a library file assigning the arbitrary insert size of 200, followed by retrieval of the mean insert size calculations from the output summary file. SSPACE parameters are as follows: x = 1; o = 10; r = 0.75. Multi-CSAR was run with the NUCmer flag, and the complete PNG genome produced with Flye (Lin et al 2016; Kolmogorov et al 2019) was used as a reference genome. The *Moorena producens* genome PAL (GenBank accession: CP017600.1) was also used as a reference genome, due to its high similarity to PNG (Leão et al 2017). GapFiller was run with 1,000,000 iterations (orders of magnitude more than necessary), assuring that gaps would be filled until no longer possible with available data, rather than ending after an arbitrary number of iterations.

5.3.6: Analyses of BGCs

The 19 post-iT draft genomes, the 2 *Moorena* reference genomes used with Multi-CSAR (PNG and PAL [GenBank accession: CP017600.1]), and two more high-quality *M. producens* draft genomes (JHB [GenBank accession: CP017708.1] and 3L [GenBank accession: MKZR00000000.1]) were submitted to antiSMASH (Blin et al 2019) for annotation of BGCs. Samples were run with ‘relaxed’ detection strictness, and with all extra annotation features turned on. AntiSMASH results were downloaded, compiled and submitted to BigSCAPE (Navarro-Muñoz et al 2019) with default parameters.

Chapter 5 is coauthored with Leão, Tiago; Podell, Sheila; Moss, Nathan A.; Whitner, Syrena; Glukhov, Evgenia; Sanders, Jon G.; Reyes, Andres Joshua; Biggs, Jason S.; Humphrey, Gregory; Zhu, Qiyun; Belda-Ferre, Pedro; Allen, Eric E.; Knight, Rob; Gerwick, Lena; Gerwick, William H. The dissertation author was the primary author of this chapter. The dissertation author co-conceived of the work, performed a portion of the necessary DNA extractions and sample preparations for sequencing, wrote and implemented code for the genome assembly module and the draft genome improvement module, conducted analyses of results, and was the primary author of this work.

5.4 Appendix: Supplemental Information

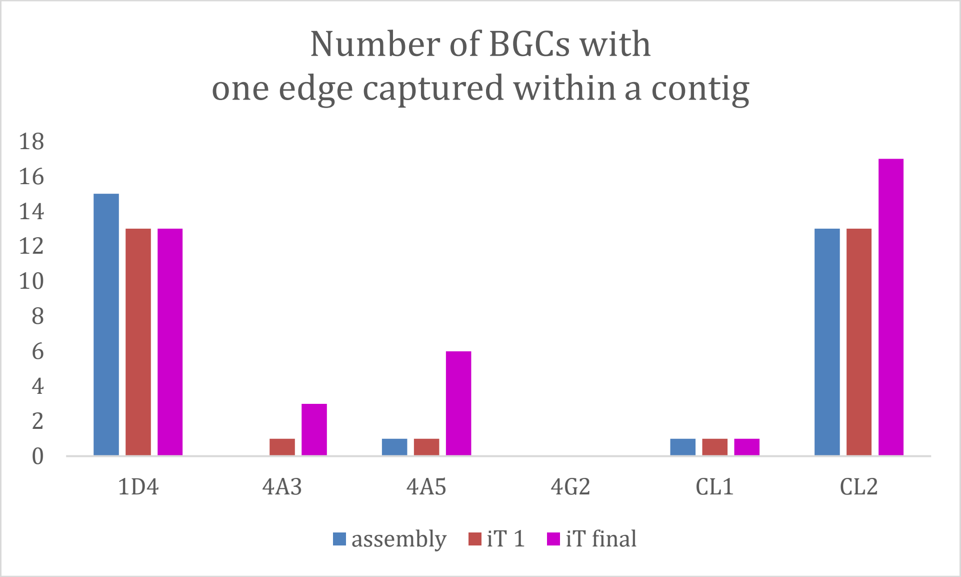


Figure 5.S1 – Counts of partial BGCs with one edge captured within a contig

Table 5.S1 – Assembly identifier for each sample. See Table 5.S5 for assembly parameters

sample code	assembly identifier
1C3	a4
1C9	a3m
1D4	a2m
1F2	a3
3A2	a3
3F7	a4i
3H8	a4i
3I6	a4i
3I7	a4i
3I8	a4
3I9	a1i
4A1	a4m
4A3	a1
4A4	a4
4A5	33
4G2	a2m
4G3	33
CL1	a4
CL2	33

Table 5.S2 – Counts of partial, complete, and total BGCs for samples

	assembly BGCs			iT 1 BGCs			iT final BGCs		
	partial	**complete**	**total**	**partial**	**complete**	**total**	**partial**	**complete**	**total**
1C3	79	0	79	65	1	66	35	10	45
1C9	59	2	61	56	2	58	27	6	33
1D4	112	5	118	94	2	96	82	0	82
1F2	17	4	21	15	4	19	8	10	18
3A2	27	11	38	26	13	39	17	20	37
3F7	44	10	54	32	10	42	35	17	52
3H8	87	2	89	77	2	77	40	17	57
3I6	45	9	54	38	9	47	21	13	34
3I7	53	3	56	49	3	52	19	12	31
3I8	16	7	23	14	7	21	5	15	20
3I9	60	1	61	56	1	57	17	9	26
4A1	33	7	40	25	8	33	10	15	25
4A3	88	0	88	78	0	78	66	0	66
4A4	72	0	72	62	0	62	30	12	42
4A5	68	0	68	65	1	66	61	1	61
4G2	40	0	40	34	0	34	17	0	17
4G3	69	2	71	70	2	72	63	4	67
CL1	41	0	41	39	0	39	28	0	28
CL2	91	3	94	69	3	72	63	3	66

Table 5.S3 – Draft genome assembly statistics

Sample	# contigs	Largest contig	Total length	GC (%)	N50	N75	L50	L75	N's per 100 kb
1C3	48985	52423	17468868	44.84	4063	1453	1000	2898	0
1C9	264291	101434	21443038	45.61	1259	626	3160	9841	0
1D4	243533	1699961	39621843	53.85	3826	1175	1203	6510	0
1F2	126149	119761	16558283	53.71	1762	666	1013	4982	0
3A2	20843	226891	8829058	44.28	45675	22813	60	129	0
3F7	29534	199188	20025612	50.4	2574	1139	827	3926	0
3H8	38291	73969	19179189	43.89	5223	1887	910	2466	0
3I6	66182	136302	18447793	46.44	2792	721	753	4443	0
3I7	40629	97288	15519789	43.7	5730	1087	484	2220	0
3I8	36777	165107	11507732	45.89	25712	878	101	959	0
3I9	42376	42858	10343463	43.64	4131	1640	658	1676	0
4A1	104137	131294	24256264	48	845	599	4056	12869	0
4A3	192416	73614	28435805	53.38	1380	784	5440	12450	0.01
4A4	30289	72950	16697226	43.87	4708	1803	811	2294	0.01
4A5	177445	41228	25118982	50.16	1262	691	4015	10735	0.04
4G2	721045	74008	23326670	52.7	734	590	10010	18972	0
4G3	144820	40995	11714778	43.62	8296	3305	409	943	0.01

CL1	25410	73295	9461306	44.37	4847	2012	511	1276	0
CL2	756685	62536	35334048	45.17	1052	639	5898	17081	0

Table 5.S4 – Sample metadata

sample ID	**field ID**	**collection region**	**latitude: longitude**	**collection date**	**culture vs environment**
1D4	Black *Lyngbya*	Palmyra Atoll	5°53'05"N: 162°05'06" W	5/24/2013	culture
3F7	*Lyngbya* sp.	Papua New Guinea	10°17'16"S : 151°00'23" E	4/22/2006	culture
1C3	*Moorea bouillonii*	American Samoa	14°20'04"S : 170°48'07" W	7/12/2014	environment
1F2	*Moorea* possibly *bouillonii*	American Samoa	14°16'52"S : 170°38'19" W	7/13/2014	environment
4G2	*Moorea bouillonii*	Guam	13°26'57"N : 144°39'27" E	3/21/2016	environment
4G3	*Moorea bouillonii*	Guam	13°26'38"N : 144°38'36" E	3/21/2016	environment
3I6	*Lyngbya bouillonii*	Papua New Guinea	5°25'34"S: 150°05'45" E	7/12/2007	environment
3I7	*Lyngbya bouillonii*	Papua New Guinea	5°25'34"S: 150°05'45" E	7/12/2007	environment
3I8	*Lyngbya bouillonii*	Papua New Guinea	5°25'34"S: 150°05'45" E	7/12/2007	environment
3I9	*Lyngbya bouillonii*	Papua New Guinea	5°25'34"S: 150°05'45" E	7/12/2007	environment
3H8	*Lyngbya bouillonii*	Papua New Guinea	5°17'36"S: 150°06'14" E	7/13/2007	environment
4A1	*Lyngbya bouillonii*	Papua New Guinea	5°17'36"S: 150°06'14" E	7/13/2007	environment
4A3	*Lyngbya bouillonii*	Papua New Guinea	5°17'36"S: 150°06'14" E	7/15/2007	environment
4A4	*Lyngbya bouillonii*	Papua New Guinea	5°26'56"S: 150°47'54" E	7/16/2007	environment
3A2	*Moorea bouillonii*	Saipan	15°09'18"N :	1/31/2013	environment

			145°42'20" E		
1C9	*Moorea bouillonii*	Saipan	15°09'37"N : 145°41'43" E	1/29/2013	environment
4A5	*Lyngbya bouillonii*	Papua New Guinea	5°26'56"S: 150°47'54" E	7/16/2007	environment
CL1	*Moorea bouillonii*	Guam	13°26'40"N : 144°38'11" E	7/3/2017	environment
CL2	*Moorea bouillonii*	Guam	13°26'40"N : 144°38'11" E	7/3/2017	environment
PNG	*Lyngbya cobweb*	Papua New Guinea	4°16'04"S: 152°20'16" E	5/19/2005	culture

Table 5.S5 – SPAdes assembly parameter sets

assembly	mode	kmers	error correct	careful
a1	standard	21,33,55	yes	yes
a2	standard	21,33,55,77	yes	yes
a3	standard	21,33,55,77,99	yes	yes
a4	standard	21,33,55,77,99,127	yes	yes
a1m	meta	21,33,55	no	no
a2m	meta	21,33,55,77	no	no
a3m	meta	21,33,55,77,99	no	no
a4m	meta	21,33,55,77,99,127	no	no
a1i	isolate	21,33,55	no	no
a2i	isolate	21,33,55,77	no	no
a3i	isolate	21,33,55,77,99	no	no
a4i	isolate	21,33,55,77,99,127	no	no
33[a]	standard	21,33,55,77	yes	yes

[a] SPAdes version 3.13.1

5.5: References

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 19:455-477. https://doi.org/10.1089/cmb.2012.0021

Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T (2019) antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res 47:W81–W87. https://doi.org/10.1093/nar/gkz310

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27:578–579. https://doi.org/10.1093/bioinformatics/btq683

Boetzer M, Pirovano W (2012) Toward almost closed genomes with GapFiller. Genome Biol 13:R56. https://doi.org/10.1186/gb-2012-13-6-r56

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bushnell B (2014) BBMap: A Fast, Accurate, Splice-Aware Aligner. In: 9th Annual Genomics of Energy & Environment Meeting, Walnut Creek, CA, March 17-20, 2014.

Chen KT, Shen HT, Lu CL (2018) Multi-CSAR: a multiple reference-based contig scaffolder using algebraic rearrangements. BMC Syst Biol 12:139. https://doi.org/10.1186/s12918-018-0654-y

Cummings SL, Barbé D, Leão TF, Korobeynikov A, Engene N, Glukhov E, Gerwick WH, Gerwick L (2016) A novel uncultured heterotrophic bacterial associate of the cyanobacterium *Moorea producens* JHB. BMC Microbiol 16:198. https://doi.org/10.1186/s12866-016-0817-1

Curren E, Leong SCY (2019) Global phylogeography of toxic cyanobacteria *Moorea producens* reveals distinct genetic partitioning influenced by Proterozoic glacial cycles. Harmful Algae 86:10-19. https://doi.org/10.1016/j.hal.2019.04.010

Engene N, Choi H, Esquenazi E, Rottacker EC, Ellisman MH, Dorrestein PC, Gerwick WH (2011) Underestimated biodiversity as a major explanation for the perceived rich secondary metabolite capacity of the cyanobacterial genus *Lyngbya*. Environ Microbiol 13:1601-1610 https://doi.org/10.1111/j.1462-2920.2011.02472.x

Engene N, Coates RC, Gerwick WH (2010) 16S rRNA heterogeneity in the filamentous marine cyanobacterial genus *Lyngbya*. J Phycol 46:591-601. https://doi.org/10.1111/j.1529-8817.2010.00840.x

Engene N, Gerwick W (2011) Intra–genomic 16S rRNA gene heterogeneity in cyanobacterial genomes. Fottea 11:17-24. https://doi.org/10.5507/fot.2011.003

Engene N, Gunasekera SP, Gerwick WH, Paul VJ (2013a) Phylogenetic Inferences Reveal a Large Extent of Novel Biodiversity in Chemically Rich Tropical Marine Cyanobacteria. Appl Environ Microbiol 79:1882–1888. https://doi.org/ 10.1128/AEM.03793-12

Engene N, Paul VJ, Byrum T, Gerwick WH, Thor A, Ellisman MH (2013b) Five chemically rich species of tropical marine cyanobacteria of the genus *Okeania* gen. nov. (Oscillatoriales, Cyanoprokaryota). J Phycol 49:1095-1106. https://doi.org/10.1111/jpy.12115

Engene N, Rottacker EC, Kaštovský J, Byrum T, Choi H, Ellisman MH, Komárek J, Gerwick WH (2012) *Moorea producens* gen. nov., sp. nov. and *Moorea bouillonii* comb. nov., tropical marine cyanobacteria rich in bioactive secondary metabolites. Int J Syst Evol Microbiol 62:1171–1178. https://doi.org/10.1099/ijs.0.033761-0

Engene N, Tronholm A, Paul VJ (2018) Uncovering cryptic diversity of *Lyngbya*: the new tropical marine cyanobacterial genus *Dapis* (Oscillatoriales). J Phycol 54:435-446. https://doi.org/10.1111/jpy.12752

Glenn TC, Nilsen RA, Kieran TJ, Sanders JG, Bayona-Vásquez NJ, Finger Jr JW, Pierson TW, Bentley KE, Hoffberg SL, Louha S, García-De León FJ, Del Río-Portilla MA, Reed KD, Anderson JL, Meece JK, Aggrey SE, Rekaya R, Alabady M, Bélanger M, Winker K, Faircloth BC (2019) Adapterama I: Universal stubs and primers for 384 unique dual-indexed or 147,456 combinatorially-indexed Illumina libraries (iTru &amp; iNext). bioRxiv 049114. https://doi.org/10.1101/049114

Grindberg RV, Ishoey T, Brinza D, Esquenazi E, Coates RC, Liu W, Gerwick L, Dorrestein PC, Pevzner P, Lasken R, Gerwick WH (2011) Single Cell Genome Amplification Accelerates Identification of the Apratoxin Biosynthetic Pathway from a Complex Microbial Assemblage. PLOS One 6:e18565. https://doi.org/10.1371/journal.pone.0018565

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hoffmann L, Komárek J, Kaštovský J (2005) System of cyanoprokaryotes (cyanobacteria) – state in 2004. Algol Stud 117:95-115. https://doi.org/10.1127/1864-1318/2005/0117-0095

Hoffmann L, Demoulin V (1991) Marine Cyanophyceae of Papua New Guinea. II. Lyngbya bouillonii Sp. Nov., A remarkable tropical reef-inhabiting blue-green alga. Belj J Bot 124:82-88

Kleigrewe K, Almaliti J, Tian IY, Kinnel RB, Korobeynikov A, Monroe EA, Duggan BM, Di Marzo V, Sherman DH, Dorrestein PC, Gerwick L, Gerwick WH (2015) Combining Mass Spectrometric Metabolic Profiling with Genomic Analysis: A Powerful Approach for Discovering Natural Products from Cyanobacteria. J Nat Prod 78:1671–1682. https://doi.org/10.1021/acs.jnatprod.5b00301

Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, Kelley K, Hamrick J, Grout J, Corlay S, Ivanov P, Avila D, Abdalla S, Willing C, Jupyter Development Team (2016) Jupyter Notebooks—a publishing format for reproducible computational workflows, In: Loizides F, Scmidt B (eds.) Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press, Amsterdam, The Netherlands. pp. 87–90. https://doi.org/10.3233/978-1-61499-649-1-87

Kolmogorov A (2009) Blossom V: a new implementation of a minimum cost perfect matching algorithm. Math Prog Comp 1:43–67. https://doi.org/ 10.1007/s12532-009-0002-8

Kolmogorov M, Yuan J, Lin Y, Pevzner P (2019) Assembly of Long Error-Prone Reads Using Repeat Graphs. Nat Biotechnol 37:540–546. https://doi.org/10.1038/s41587-019-0072-8

Leão T, Castelão G, Korobeynikov A, Monroe EA, Podell S, Glukhov E, Allen EE, Gerwick WH, Gerwick L (2017) Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus *Moorea*. Proc Natl Acad Sci USA 114:3198-3203. https://doi.org/10.1073/pnas.1618556114

Leão T, Wang M, Moss N, da Silva R, Sanders J, Nurk S, Gurevich A, Humphrey G, Reher R, Zhu Q, Belda-Ferre P, Glukhov E, Whitner S, Alexander KL, Rex R, Pevzner P, Dorrestein PC, Knight R, Bandeira N, Gerwick WH, Gerwick L (2020) A multi-omics description of the natural product potential of tropical marine cyanobacteria. *In prep*

Lin Y, Yuan J, Kolmogorov M, Shen MW, Chaisson M, Pevzner PA (2016) Assembly of Long Error-Prone Reads Using de Bruijn Graphs. Proc Natl Acad Sci USA 113:E8396-E8405. https://doi.org/10.1073/pnas.1604560113

McKinney W (2010) Data Structures for Statistical Computing in Python. In: van der Walt S, Millman J (eds.) Proceedings of the 9th Python in Science Conference (SciPy2010), Austin, TX, USA, June 28 - July 3, 2010. pp. 51–56. https://doi.org/10.25080/Majora-92bf1922-00a

Miller IJ, Rees ER, Ross J, Miller I, Baxa J, Lopera J, Kerby RL, Rey FE, Kwan JC (2019) Autometa: automated extraction of microbial genomes from individual shotgun metagenomes. Nucleic Acids Res 47:e57. https://doi.org/10.1093/nar/gkz148

Monroe EA, Choi H, Lesin V, Sirotkin A, Dvorkin M, Pevzner P, Gerwick WH, Gerwick L (2012) Genomic insights into secondary metabolism of the natural product-rich cyanobacterium *Moorea bouillonii*. Planta Med 78:CL52. https://doi.org/10.1055/s-0032-1320287

Moss NA, Leão T, Glukhov E, Gerwick L, Gerwick WH (2018) Collection, Culturing, and Genome Analyses of Tropical Marine Filamentous Benthic Cyanobacteria. In: Tawfik DS (ed.) Methods in Enzymology, 1st ed. Academic Press, Cambridge, MA, USA. vol 604, pp. 3–43. https://doi.org/10.1016/bs.mie.2018.02.014

Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, De Los Santos ELC, Yeong M, Cruz-Morales P, Abubucker S, Roeters A, Lokhorst W, Fernandez-Guerra A, Cappelini LTD, Goering AW, Thomson RJ, Metcalf WW, Kelleher NL, Barona-Gomez F, Medema MH (2019) A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol 16:60–68. https://doi.org/10.1038/s41589-019-0400-9

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA (2017) metaSPAdes: a new versatile metagenomic assembler. Genome Res 27:824-834. https://doi.org/ 10.1101/gr.213959.116

Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM (2016) Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17:132. https://doi.org/10.1186/s13059-016-0997-x

Pandas-dev/pandas, Version v0.25.2 (2018) Zenodo. https://doi.org/10.5281/zenodo.3509135

Pérez F, Granger BE (2007) IPython: A System for Interactive Scientific Computing, Comput Sci Eng 9:21–29. https://doi.org/10.1109/MCSE.2007.53

Podell S, Gaasterland T (2007) DarkHorse: a method for genome-wide prediction of horizontal gene transfer. Genomic Biol 8:R16. https://doi.org/10.1186/gb-2007-8-2-r16

Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A (2020) Using SPAdes De Novo Assembler. Curr Protoc Bioinformatics 70:e102. https://doi.org/10.1002/cpbi.102

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Skiba MA, Sikkema AP, Moss NA, Tran CL, Sturgis RM, Gerwick L, Gerwick WH, Sherman DH, Smith JL (2017) A Mononuclear Iron-Dependent Methyltransferase Catalyzes

Initial Steps in Assembly of the Apratoxin A Polyketide Starter Unit. ACS Chem Biol 12:3039–3048. https://doi.org/10.1021/acschembio.7b00746

Skiba MA, Sikkema AP, Moss NA, Lowell AN, Su M, Sturgis RM, Gerwick L, Gerwick WH, Sherman DH, Smith JL (2018) Biosynthesis of t-Butyl in Apratoxin A: Functional Analysis and Architecture of a PKS Loading Module. ACS Chem Biol 13:1640–1650. https://doi.org/10.1021/acschembio.8b00252

Tronholm A, Engene N (2019) *Moorena* gen. nov., a valid name for “*Moorea* Engene & al.” nom. inval. (Oscillatoriaceae, Cyanobacteria). Notulae Algarum 122:1–2

Van Rossum G, Drake Jr. FL (1995) Python reference manual, Release 2.0.1. PythonLabs, Amsterdam, The Netherlands. https://docs.python.org/2.0/ref/ref.html

Conclusions, Reflections & Future Perspectives

During the five years that I have been studying *Moorena bouillonii* (L.Hoffmann & Demoulin) Engene & Tronholm 2019 (Oscillatoriaceae) and its symbiosis with *Alpheus frontalis* H. Milne Edwards 1837 (Alpheidae), our knowledge of this charismatic partnership and its associated natural product arsenal has greatly expanded. Our work has demonstrated that *M. bouillonii – A. frontalis* colonies in the Mariana Islands inhabit the interstices of scleractinian coral reefs, in contrast to their reported overgrowth of gorgonians (Yamashiro, Isomura, Sakai 2014; Leber et al 2020b [*under review*]; Chapter 1). We also measured an increased nutrient environment within shrimp-woven cyanobacterial structures and displayed that removal of shrimp causes cyanobacteria accustomed to shrimp-inhabitation to respond negatively. This suggests that nutrients are one dimension by which *M. bouillonii* benefits from its arrangement with *A. frontalis*, even as it endures reductions in biomass due to *A. frontalis* feeding (Leber et al 2020b [*under review*]; Chapter 1). Our analyses of patterns in the distribution of *M. bouillonii* molecular features, what we refer to as chemogeography, led us to describe a new regionally-specific family of compounds with unusual synergistic cytotoxicity with lipopolysaccharide (LPS) (Leber et al 2020a; Chapter 2). These analyses also taught us about how known compound families were distributed across geographically disparate samples of *M. bouillonii* and improved our understanding of how chemical compositional similarity varies through different spatial scales (Chapter 4). The resultant insights are directly translatable into prioritized molecular features ripe for characterization via discovery efforts. By developing a bioinformatics approach that greatly improved contiguity and biosynthetic gene cluster (BGC) completeness, we were able to learn more about *M. bouillonii* biosynthetic diversity. Specifically, our efforts revealed that there appears to be much more potential for chemodiversity in polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) classes of natural products produced by *M. bouillonii*, as compared to terpenes and ribosomally synthesized and post-translationally modified peptides (RiPPS).

Starting from an understated description in the literature of *M. bouillonii* growing without *A. frontalis* (Matthew, Schupp, and Luesch 2008) that captured my curiosity, we progressed through a series of breakthroughs in elucidating a fascinating story in chemical ecology. We determined the mechanism of shrimp distribution across *M. bouillonii* growths with similar chemical compositions to likely not be induction, but to instead be driven by shrimp selection for a particular chemotype (Chapter 3). We found that shrimp could navigate a Y-tube with stunning accuracy in order to locate their preferred chemotype of *M. bouillonii*, indicating that they relied upon a waterborne cue. By running experiments to constrain the characteristics of that elusive cue, paying close attention to complex shrimp behavioral responses to extract information not always evident by simple metrics, and leveraging comparative analyses of the appropriate subset of *M. bouillonii* natural products, we were able to identify a promising lead molecule for the chemical cue (Chapter 3).

Besides the expansion of knowledge, one also finds in our work promising opportunities to further expand scientific understanding and so push farther into the unknown. The growth of *M. bouillonii – A. frontalis* colonies in coral reef interstices, in contrast with its reported overgrowth of gorgonians (Yamashiro, Isomura, Sakai 2014; Leber et al 2020b [*under review*]; Chapter 1), begs the question of why? Why were colonies growing in such an exposed fashion, so out of character from what we documented on the reefs of Guam and Saipan? What factors, either intrinsic to the cyanobacterium or shrimp, or stemming from environmental externalities, allowed this to occur? We know that *M. bouillonii* chemical composition varies over geographical space (Leber et al 2020a; Chapter 2; Chapter 4), suggesting that the gorgonian-overgrowing colonies in Aka Jima could be quite chemically different than those growing in the cracks and crevices of coral reefs on Guam. Perhaps a particular blend of secondary metabolites specific to southern Japan allows for the gorgonian overgrowth to occur. Incorporating these samples into chemogeographical analyses could provide better understanding of how these samples may or may not differ chemically from other populations of *M. bouillonii*. Expanding on the topic of chemogeography, our efforts not only laid a foundation for numerous isolation efforts prioritized based on regional-specificity (Chapter 4), they also encourage the application of a chemogeographic strategy for drug discovery in other taxa.

Without biological validation, the compound we identified remains as a potential, but not yet verified, shrimp attractant waterborne chemical cue (Chapter 3). By comparison to an analytical standard and collection of data from additional replicates, we plan to ascertain the identity of our differential molecular feature as isopropyl myristate. By applying isopropyl myristate to cotton balls and providing those cotton balls to shrimp both in the glass dish and Y-tube experimental assay formats, we hope to solidify this finding. Even when we do, we still will not know what other chemistry may be involved in mediating shrimp-cyanobacterial interactions. Are there other cues acting in conjunction, such as contact cues? Is there a chemical difference between shrimp-associated and non-shrimp-associated *M. bouillonii* that makes shrimp-associated more beneficial, or non-shrimp-associated more harmful? These would be fascinating questions to answer and could build off the foundation of our work.

The bioinformatic software that I created, iTerator, was extremely successful in not only improving the contiguity of short read assemblies, but also in consolidating partial BGCs for more complete BGCs (Chapter 5). Use of this tool in a high-throughput manner, however, is limited by the non-optimal computational foundation upon which it is built. The software elements that comprise iTerator could be upgraded for more effective use of memory and better leveraging of parallel processing. This represents an opportunity for further developing a promising proof-of-concept to a much more practical, and so much more efficacious bioinformatic approach.

As the specifics of knowledge gained, the questions answered and left unanswered, and the opportunities for future discovery cross the mind-manuscript barrier onto this page, I am left with one omnipresent concept; this concept seems to me to be the common thread throughout my PhD work. And this concept is **complexity**.

Through the studies reported herein, we now understand that the *M. bouillonii – A. frontalis* association does not exist as a simple monolithic mutualism, but rather, with variation along numerous dimensions. Colonies were reported in the literature with morphologies suggestive of gorgonian overgrowth (Yamashiro, Isomura, Sakai 2014); however, we documented *M. bouillonii – A. frontalis* colonies in three different growth morphologies, none of which were suggestive of overgrowth of scleractinian corals (Leber et al 2020b [under review]; Chapter 1). Our isolation and characterization of doscadenamide A led to a putative family of ten natural products with even more analogs still to be annotated and described, predominantly produced only by samples from Guam and Saipan (Leber et al 2020a; Chapter 2). As our broader chemogeographic analyses revealed, samples from regions lacking in the doscadenamides contained their own regionally-specific compounds (Chapter 4), indicating that while *A. frontalis* shrimp can be found across the Indian Ocean and western tropical Pacific weaving *M. bouillonii*, they are living in very different chemical environments. And yet, when shrimp are presented with two chemical environments with many similarities, such as the shrimp-associated and non-shrimp-associated *M. bouillonii* on the reefs of Guam, they detect a waterborne chemical cue that allows them to consistently select on over the other (Chapter 3).

Complexity is necessarily difficult to comprehend, and so a large portion of my PhD work was aimed at thoughtfully simplifying the complexity through the identification and leveraging of patterns: patterns in cyanobacterial growth morphology, patterns in nutrient concentration, patterns in chemogeography, patterns in shrimp distribution and behavior, patterns in BGCs. These patterns, however, must be wielded with caution because they are indeed simplifications. Understanding *M. bouillonii* to display patterns in chemical composition related to location of collection is only correct when adequately accounting for differences in shrimp-associated and geographic scale (Chapter 4). Similarly, the Y-tube served as a useful model for measuring shrimp response to the *M. bouillonii* waterborne cue, but was only useful when not confounded by the various externalities that shrimp also had behavioral responses to, such as light, sound, or an incoming plankton bloom (Chapter 3). This tension, between striving to understand complexity through pattern identification and being thwarted in efforts to find patterns by previously unknown (or under-considered) complexity, mandates that one approach complex biological phenomena with an appreciation for the nuances that can and will arise.

And so, I am left humbled by nature's complexity, and committed to a continued thoughtful exploration of the patterns that help to make sense of that complexity. However, nature's complexity, particularly in the form of biodiversity, is under threat. Overexploitation and habitat degradation, compounded by global climate change, are currently accelerating rates of extinction by orders of magnitude (Naman, Leber, and Gerwick 2018). This matters to natural products researchers, who depend on biodiversity for uncovering novel chemical entities, and patients, who might one day receive treatment originating from an obscure cyanobacterium, but also to all living things on our planet, who benefit every day from healthy and stable ecosystems. Therefore, I must amend my commitment to not only continue to study, explore, and appreciate nature’s vast and curious complexities, but also to protect them. And I urge you to do the same.

References

Leber CA, Naman CB, Keller L, Almaliti J, Caro-Diaz EJE, Glukhov E, Joseph V, Sajeevan TP, Reyes AJ, Biggs JS, Li T, Yuan Y, He S, Yan X, Gerwick WH (2020a) Applying a Chemogeographic Strategy for Natural Product Discovery from the Marine Cyanobacterium *Moorena bouillonii*. Mar Drugs 18:515. https://doi.org/10.3390/md18100515

Leber CA, Reyes AJ, Biggs JS, Gerwick WH (2020b) Cyanobacteria-Shrimp Colonies in the Mariana Islands. *Under Review*.

Matthew S, Schupp PJ, Luesch H (2008) Apratoxin E, a Cytotoxic Peptolide from a Guamanian Collection of the Marine Cyanobacterium *Lyngbya bouillonii*. J Nat Prod 71:1113–1116. https://doi.org/10.1021/np700717s

Naman CB, Leber CA, Gerwick WH (2017) Chapter 5 - Modern Natural Products Drug Discovery and Its Relevance to Biodiversity Conservation. In: Kurtböke I (ed.) *Microbial Resources*. Academic Press, pp 103-120

Yamashiro H, Isomura N, Sakai K (2014) Bloom of cyanobacterium *Moorea bouillonii* on the gorgonian coral *Annella reticulata* in Japan. Sci Rep 4:6032. https://doi.org/10.1038/srep06032