UNIVERSITY OF CALIFORNIA SAN DIEGO

Comparative Genomics and Genome Mining Insights into Natural Product Rich Marine Cyanobacteria

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

Doctor of Philosophy

in

Marine Biology

by

Tiago Ferreira Leao

Committee in charge:

William H. Gerwick, Chair
Eric Allen, Co-chair
Lena Gerwick, Co-chair
Pieter Dorrestein
Paul Jensen

2019

SIGNATURE PAGE

The dissertation of Tiago Ferreira Leao is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-chair

Co-chair

Chair

University of California San Diego
2019

DEDICATION

This dissertation is dedicated to my incredible parents, Jorge Leao and Maria de Nazaré Ferreira, my beloved sister Camila and my inspiring brother Gabriel.

EPIGRAPH

"Education never ends
It is a series of lessons with the greatest for the last."

– Sir Arthur Conan Doyle, *His Last Bow*

"We are all smart. Distinguish yourself by being kind."

– Charles Gordon

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

A – Adenylation domain

AA – Amino acid

ACP – Acyl carrier protein

antiSMASH – antibiotics & Secondary Metabolite Analysis Shell

AT – Acyltransferase domain

BGC – Biosynthetic gene cluster

BLAST – Basic local alignment search tool

C – Condensation domain

COG – Clusters of Orthologous Groups

CRISPR – Clustered regularly interspaced short palindromic repeats.

DMAPP – Dimethylallyl pyrophosphate

FDA – U.S. food and drug administration

GARLIC – global alignment for natural products cheminformatics

GCF – Gene cluster family

GNPS – Global natural products social molecular networking

HMMs – Hidden Markov models

IMG – Integrated Microbial Genomes & Microbiomes

IPP – Isopentenyl pyrophosphate

JGI – Departmant of energy joint genome institute

KS – Ketosynthase

MAAs – Mycosporine-like amino acids

LC-MS/MS – Liquid chromatography coupled with tandem mass spectrometry

NMR – Nuclear magnetic ressonance

NPs – Natural products

NRPS – Nonribosomal peptide synthetases

PCP – Peptidyl-carrier protein

PCR – Polymerase chain reaction

PKS – Polyketide synthases

RAST – Rapid Annotation using Subsystem Technology

RiPPs – Ribosomally synthesized and post-translationally modified peptides

rRNA – ribosomal ribonucleic acid

TAR – Transformation-Associated Recombination

UV – Ultraviolet

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The work in this dissertation was facilitated by several past and current members of the Gerwick Lab, especially: Nathan Moss, Bohan Ni, Karin Kleigrewe, Raphael Reher, Christopher Leber, Ben Naman and Evgenia Glukhov.

I would like to thank my advisors William H Gerwick and Lena Gerwick for providing all the support so I could execute the research here presented. I would like to thank you for the priceless advices during my Ph.D. research career. I would like to thank my committee members as well, Professors Paul Jensen, Eric Allen, and Pieter Dorrestein for their time and guidance towards meaningful projects and outcomes.

I also am thankful for the invaluable contribution of collaborators at other institutions: Anton Korobeynikov, Sergey Nurk and Alexey Gurevich.

I'm also very thankful for all my friends, for the special moments that made my stay in San Diego extremely enjoyable and fun. It is challenging to describe the importance of the support of friends and colleagues during my Ph.D. studies.

Last, I would like to thank my beloved parents for giving me a strong foundation based in love, kindness, respect and for teaching me the value of hard work and dedication. To my sister Camila and my brother Gabriel for always being available when I needed the most.

Chapter 1, the topics "overview of the genetic diversity from cyanobacteria" and "Biosynthesis of cyanobacterial natural products" are reproduced, with permission, of the material as it appears in *Taylor and Francis Publishers*, 2017, Lena Keller, Tiago Leao and William H. Gerwick.

Chapter 2, in full, is a reprint, with permission, of the material as it appears in *BMC Microbiology*, 2016, Susie L. Cummings*, Debby Barbé*, Tiago Ferreira Leao*, Anton Korobeynikov, Niclas Engene, Evgenia Glukhov, William H. Gerwick and Lena Gerwick. The

dissertation author is one of the primary investigator (*shared authorship) and author of this material.

Chapter 3, in full, is a reprint, with permission, of the material as it appears in *Proc. Natl. Acad. Sci.*, 2017, Tiago Leao, Guilherme Castelão, Anton Korobeynikov, Emily A. Monroe, Sheila Podell, Evgenia Glukhov, Eric E. Allen, William H. Gerwick and Lena Gerwick. The dissertation author is the primary investigator and author of this material.

Chapter 4, in full, currently being prepared for submission for publication. Tiago Leão, Ricardo Silva, Nathan Moss, Mingxun Wang, Jon Sanders, Sergey Nurk, Alexey Gurevich, Gregory Humphrey, Raphael Reher, Qiyun Zhu, Pedro Belda-Ferre, Pieter Dorrestein, Rob Knight, Pavel Pevzner, William H. Gerwick and Lena Gerwick. “Genomic insights into an expanded diversity of filamentous marine cyanobacteria reveals the extraordinary biosynthetic potential of *Moorea* and *Okeania*”. The dissertation author is the primary investigator and author of this material.

VITA

ACADEMIC

AUGUST 2010 – JUNE 2014 BACHELOR OF SCIENCE	University of Pará, Belém, Brazil Biotechnology (GPA: 9.19/10.00)
AUG 2012 – AUGUST 2013 EXCHANGE STUDENT	University of Michigan, Ann Arbor, MI, USA Biology (GPA: 3.60/4.00)
JULY 2014 – JULY 2019 DOCTOR OF PHILOSOPHY	University of California San Diego, CA, USA Marine Biology

SELECTED PUBLICATIONS

Moss NA, Seiler G, **Leao TF**, Castro-Falcòn G, Gerwick L, Hughes C, Gerwick WH. Nature's combinatorial biosynthesis produces vatiamides A-F. *Angewandte*. In press.

Moss NA, **Leao T**, Glukhov E, Gerwick L, Gerwick WH (2018). Collection, Culturing, and Genome Analyses of Tropical Marine Filamentous Benthic Cyanobacteria. *Methods in enzymology* 604, 3-43

Moss NA, **Leao T,** Rankin MR, McCullough TM, Qu P, Korobeynikov A, Smith JL, Gerwick L, Gerwick WH (2018) Ketoreductase Domain Dysfunction Expands Chemodiversity: Malyngamide Biosynthesis in the Cyanobacterium *Okeania* hirsute. *ACS Chem. Biol*., 13, 3385–3395

Keller L, **Leao T**, Gerwick WH (2017). Chemical Biology of Cyanobacteria. Chemical Biology of Natural Products. *Taylor and Francis Publishers* v.1. p.43-87

Kinnel RB, Esquenazi E, **Leao T**, Moss N, Mevers E, Pereira A, Monroe EA, Korobeynikov A, Murray T, Sherman D, Gerwick L, Dorrestein P, Gerwick WH (2017). A Maldiisotopic Approach to Discover Natural Products: Cryptomaldamide, a Hybrid Tripeptide from the Marine Cyanobacterium *Moorea producens*. *J Nat Prod*. 80 (5):1514–1521.

Leao T, Castelão G, Korobeynikov A, Monroe EA, Podell S, Glukhov E, Allen E, Gerwick WH, Gerwick L. Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus *Moorea* (2017). *Proc Natl Acad Sci*. 114(12): 3198–3203.

Cummings SL*, Barbé D*, **Leao 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(1):198 *shared authorship

Moss NA, Bertin MJ, Kleigrewe K, **Leão TF**, Gerwick L, Gerwick WH (2015). Integrating mass spectrometry and genomics for cyanobacterial metabolite discovery. *J Ind Microbiol Biotechnol*. (43):313–24.

GRANTS AND FELLOWSHIPS

2014-2018 – Brazilian Science Without Borders Fellowship
2016 – Indo-USA Knowledge Exchange Program Travel Grant
2016 – Gordon Natural Products Research Conference Travel Grant

AWARDS

2018 – Claude E. Zobell Fellowship, Marine Biology Division, SIO Department

ABSTRACT OF THE DISSERATION

Comparative Genomics and Genome Mining Insights into
Natural Product Rich Marine Cyanobacteria

by

Tiago Ferreira Leao

Doctor of Philosophy in Marine Biology

University of California San Diego, 2019

William H. Gerwick, Chair
Eric Allen, Co-chair
Lena Gerwick, Co-chair

Numerous advantages of systematic genomics have been leveraged in other areas, such as transcriptomics, proteomics and human microbiome studies, but it has yet to be effective in stimulating an increase in the discovery of natural products (secondary metabolites).[1] Since the discovery of the first microbial genome (2002), the surprisingly high percentage of cryptic biosynthetic pathways per genome (genes not connected to any known secondary metabolite, aka a natural product) has continued as a consistent theme.[2] Cyanobacteria from benthic environments are prolific producers of natural products, however genomes from these organisms remain scarce in public repositories. Hence, this dissertation

focused on sequencing and performing genomic analysis of cyanobacteria collected from several tropical benthic ecosystems around the planet. First, we analyzed the relationship between a cyanobacterium host and its associated heterotrophic microbiome. By evaluation of the non-axenic culture of *Moorea producens* JHB, we characterized a novel uncultured acidobacteria heterotroph living in association with its cyanobacterial host.[3] This heterotroph was only 85% similar to the closest cultured representative, and it presented a large number of genes encoding for transcriptional regulators. Further, it was found to be auxotrophic for several proteogenic amino acids. Next, we expanded our genome comparison by deeper analysis of four *Moorea* genomes. Our genome comparison revealed that *Moorea*, already a known prolific producer of secondary metabolites, harbored an even richer metabolic potential, four times above the cyanobacterial average. We observed that *Moorea* conserved its primary metabolism while evolved an intricate secondary metabolism machinery, accounting for up to 20% of its genomic content. These findings are very promising for future genome mining efforts in those four strains. We further expanded our sequencing efforts by sampling 165 metagenomes of filamentous marine cyanobacteria collected from around the globe. Our new metagenomic pipeline was able to generate 81 high-quality genomes, including 26 *Moorea* and 22 *Okeania* strains. Our genome comparison highlighted that these two genera are among the most diverse and prolific producers of natural products in our dataset (comparing pairwise 506 cyanobacterial genomes, 425 from the NCBI database). Gene networking revealed the abundance of unique natural product biosynthetic gene cluster (only encountered in a single strain) as well as “extended families” (found in several strains).

CHAPTER 1: Introduction

1.1. The importance of marine cyanobacteria for ecology and drug discovery

Cyanobacteria are found in almost every niche on the planet, oftentimes performing vital processes like photosynthesis and nitrogen fixation. Cyanobacterial photosynthesis is essential for producing the oxygen that allowed the evolution of complex organisms, marine invertebrates and other marine and terrestrial animals.[4] However, oxygen inhibits the activity of nitrogenases (nitrogen fixing enzymes), requiring cyanobacteria to develop alternative strategies for conciliating oxygenic photosynthesis and nitrogen fixation. Some cyanobacteria utilize temporal separation which allows for nitrogen to be fixed during the dark period. In filamentous cyanobacteria, spatial separation is another strategy to overcome the incompatibility of nitrogen and CO_2 fixation. Spatial separation can be obtained by the formation of specialized cells named heterocysts around which are accumulated specialized glycolipids in order to maintain an internal anoxic environment.[5] Given the richness of nutrients produced by filamentous cyanobacteria, they tend to be surrounded by many associated heterotrophs. In exchange, these heterotrophs can produce vitamins and other co-factors that are important for cyanobacterial growth.[6]

About two thirds of all drugs on the market today are derived from the secondary metabolites of animals, plants and microorganisms.[7] These secondary metabolites, also known as natural products, have been major inspirational sources of therapeutic agents used to treat cancer, bacterial infections, parasitic infections and many other disease states.[7] Marine cyanobacteria are especially interesting because they produce a dizzying array of natural products (NPs) with many unusual functional groups and atom arrangements.[1] These cyanobacterial natural products have powerful biological properties and work through distinct and oftentimes-unique pharmacological mechanisms, making them important tool compounds

as well as drug leads. For example, apratoxin A (Figure 1.1), a low nM cytotoxic natural product (from the marine cyanobacterium *Moorea bouillonii* PNG), inhibits the Sec61 complex in the secretory pathway, a potentially new molecular target for anticancer therapy.[8] Dolastatin 10 (Figure 1.1), a marine cyanobacterial metabolite produced by *Symploca* spp., has a strong antitubulin activity[9] and its synthetic analog serves as the 'warhead' of an FDA approved Antibody-Drug Conjugate known as ADCETRIS (aka brentuximab vedotin).

Historically, NPs have been discovered via chemical and/or bioactivity-guided approaches, referred as "top-down approaches". However, the rediscovery of known chemical backbones has highlighted the need for new methods in the NP discovery toolbox. In parallel, the genomics revolution has revealed that the number and chemical diversity of NP discovery to date is less than a quarter of the total microbial potential,[10,11] a phenomenon often referred to as natural product "dark matter".[12–14] Therefore, new approaches have been developed to complement traditional methods, termed "bottom-up approaches".[15] 'Bottom-up approaches consist of evaluating the natural product potential of a given microbe by first studying its biosynthetic gene clusters (BGCs), thereby providing the initial directions in the search for potentially novel bioactive metabolites. This methodology is founded on observations that most natural products are formed by many enzymes working together in an assembly line-like configuration. These enzymes derive from biosynthetic genes, which are conveniently grouped together into biosynthetic gene clusters (BGCs) in microbial genomes.[15]

In 2000, Challis and Ravel made one of the earliest bottom-up discoveries by investigating the biosynthetic genes in *Streptomyces coelicolor* A3(2).[16] By predicting the amino acid (AA) building blocks that should be found in the natural product (NP) coelichelin (Figure 1.1), they were able to up-regulate and isolate this new UV-active iron siderophore (the common term for a small molecule that binds iron with high affinity). Many more "reversely discovered" NPs (compounds found using genetic insights) were reported from *Streptomyces* spp., as well

as other members of the phylum actinobacteria that could use genome mining approaches to elucidate cryptic NPs (natural products that initially can only be perceived via DNA analysis).[1] In 2007, a new approach combining genomic predictions of building blocks and isotope-labeled feeding experiments lead to the discovery of the bioactive orfamides A-C (orfamide A illustrated in Figure 1.1).[17] In 2010, the application of Transformation-Associated Recombination (TAR) cloning for NP discovery[18] represented a major advance for pursuing unexplored microbial NPs. TAR cloning allows capture of full BGCs and heterologous expression in yeast/bacterial host, thereby enabling the access to metagenomic samples and "silent" BGCs in cultured microbes (silent refers to gene clusters that are under-expressed in standard culture conditions). In addition, the development of long range PCR and Gibbs assembly methods have also facilitated the cloning of BGCs into expression vectors.[19] More recently, the successful blend of TAR-cloning with CRISPR/Cas9-directed recombination (a cutting edge technology for high-efficiency DNA editing) represents another promising solution for accessing novel and uncharacterized BGCs.[20] However, there are only a few examples that illustrate the success of applying these technologies for mining microbes from the phylum actinobacteria.[12,21]

Reports of the use of these genome-based technologies in other microbes are very scarce. For example, there are very few reports of bottom-up approaches in the NP-rich marine tropical filamentous cyanobacteria. One example is the discovery of the columbamides, which combined genomics and metabolomics in the marine cyanobacterium *Moorea bouillonii* PNG to isolate a novel class of chlorinated acyl amides.[22] Another more recent example is represented by the vatiamide class of NP and their biosynthetic pathway. Vatiamides[23] were identified via genomics by analyzing a BGC fairly similar to jamaicamide, however, it contained three different terminations for the same alkylated fatty acid tail. Because of the successful discoveries of the columbamides and vatiamides (Figure 1.1), and the fact that preliminary genomic data has suggested that many more natural products are yet to be discovered in the genus *Moorea* [24], I

decided to concentrate my investigations on tropical filamentous marine cyanobacteria. I focused on the use of omics approaches to answer the following questions: what is the composition of the bacterial community associated with *Moorea*? What is its potential for inter-species interactions? How large is the secondary metabolic potential of these cyanobacteria? How distinctive is this potential compared to other bacteria? How can we automatically annotate cryptic NPs from these strains using 'bottom-up' approaches?

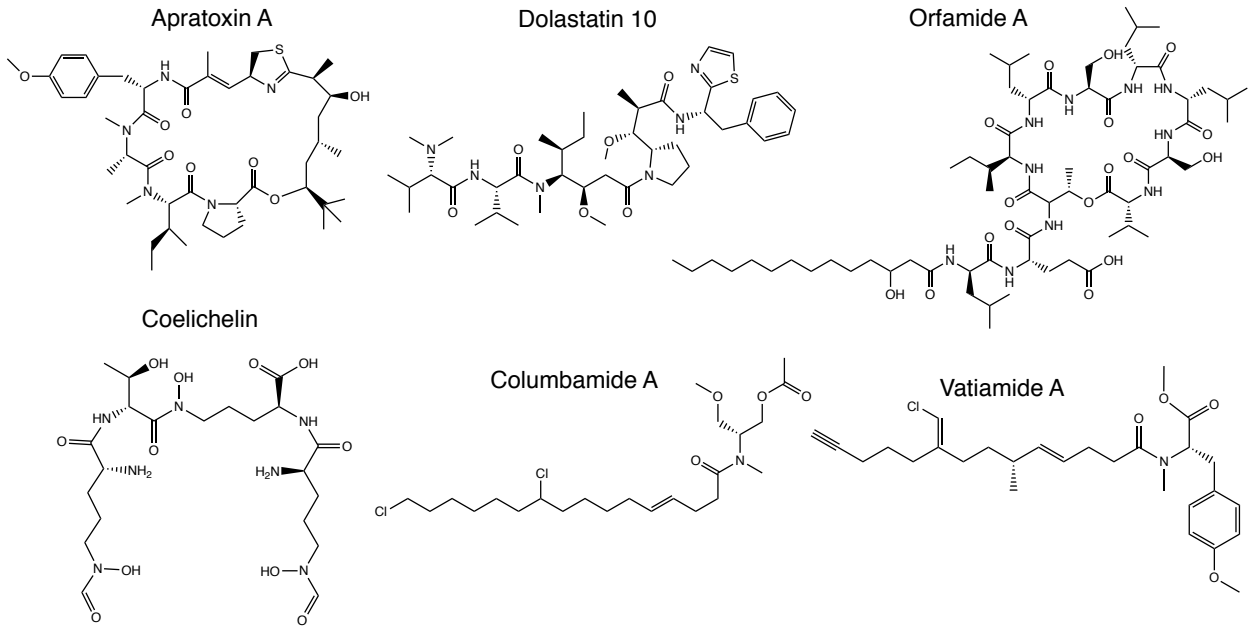


Figure 1.1. Structures of highly bioactive compounds and key compounds discovered via 'bottom-up' approaches.

1.2. Overview of the genetic diversity from cyanobacteria

Despite the diversity of cyanobacteria and their importance for both primary and secondary metabolism, genomic investigations are unevenly distributed through this phylum. Table 1 illustrates the number of cyanobacterial genomes available in the IMG/JGI databases and organized according to the phylogenetic subsections. Because cyanobacteria were included in the biological code (in 1978), their taxonomy has been under constant review and restructuring. Initially, cyanobacteria were classified in five subsections that were subsequently extended into eight subsections due to the inclusion of genetic data as a tool for taxonomic

classification.[25] This revised classification relies mainly on phylogenetic analysis of multiple conserved protein sequences in combination with key morphological traits, such as thylakoid patterns, the formation of filaments and the presence of heterocysts. Despite the extensive work to correctly classify cyanobacteria, two organisms, *Halothece* sp. and *Rubidibacter lacune*, do not fit in any of the current subsections and therefore remain unclassified, indicating that further analysis will be required in the future.

Table 1.1. Summary of publically available genomes at JGI/IMG database.

Subsection	# Genera	# Strains
Unknown	N/A	14
Unclassified	2	2
Melainabacteria	4	6
I. Gloeobacterales	1	2
II. Synechococcales	14	181
III. Spirulinales	1	2
IV. Chroococcales	7	54
V. Pleurocapsales	4	5
VI. Oscillatoriales	11	55
VII. Chroococcidiopsidales	3	5
VIII. Nostocales	22	67
Total	**69**	**393**

As observed in Table 1.1, the subsection *Synechococcales* (II) is the only group with a large number of genomes. Many of these genomes are fairly similar to one another and their size tend to be the smallest among the phylum (less than 5 Mb), facilitating genome sequencing and assembly. Additionally, *Synechococcales* are very important for biogeochemical cycles in the open ocean. *Synechococcus* and *Prochlorococcus* are among the major sources of carbon and nitrogen in oligotrophic environments.[60] Nevertheless, these genera tend to produce few secondary metabolites, hence, representing a less interesting target for drug discovery. These cyanobacteria contain only few BGCs for ribosomally synthesized and post-translationally modified peptides (RiPPs) as well as a few terpenes; however, key discoveries have been made

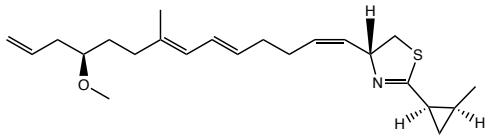
from these including the patellamides and prochlorosins (further discussed in the upcoming section of this chapter).[26]

The major secondary metabolite producers among cyanobacteria explored to date are from the subsections *Oscillatorialles* (VI) and *Nostocales* (VIII). In 2015, it has been estimated that these two subsections account for two thirds of all isolated cyanobacterial NPs.[7,62] Most of the secondary metabolites elucidated from these two subsections derive from classes of polyketide synthases (PKSs), nonribosomal peptide synthetases (NRPSs), hybrid PKS/NRPS and indole-alkaloids. The subsection *Chroococcales* (IV) was consider the third most productive with emphasis on cyanotoxins commonly produced by the freshwater genus *Microcystis*. The other subsections (I, III, V and VII) are underrepresented in terms of sequenced genomes and in terms of the number of reported NPs; there are few to any studies reporting novel natural products from these subsections to date.

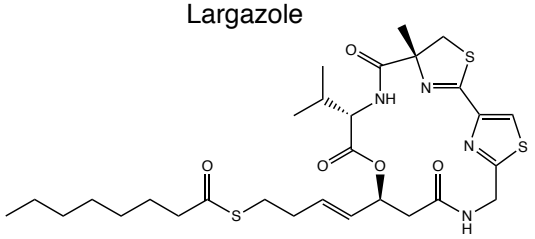
1.3. Biosynthesis of cyanobacterial natural products

As mentioned above, cyanobacteria produce a dizzying array of interesting structures with unusual methylations, halogenations, and oxidations. Several of these NPs belong to the class of hybrid PKS/NRPS metabolites, also known as lipopeptides. For example, the lipopeptide curacin A[27] (Figure 1.2) features a fatty acid chain with beta-branching and a thiazolene ring adjacent to the cyclopropyl ring. In addition, this interesting structure also shows cytotoxicity at the low nM range by targeting intracellular tubulins. Largazole (Figure 1.2),[28] a cyclic depsipeptide containing an unusual thioester in the fatty acid moiety, has potent antiproliferative activity against cancer cells. Nostophycin[29] (Figure 1.2) is a cyanotoxin that contains a β-amino acid residue. These three examples illustrate the abundant class of hybrid PKS/NRPS natural products in cyanobacteria, a large group of compounds that tend to possess potent biological activities.

Curacin A



Largazole



Nostophycin

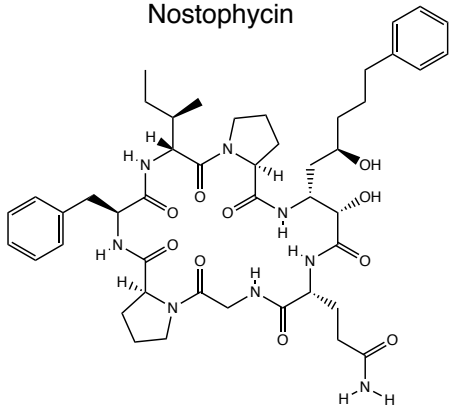


Figure 1.2. Structures for compounds with intriguing biosynthesized moieties.

In terms of biosynthetic machinery, the minimal composition of a polyketide synthase (PKS) module is comprised of a ketosynthase (KS) domain, an acyltransferase (AT) domain and an acyl carrier protein (ACP) domain. PKSs can be divided in three classes: type 1, consisting of a large mega-synthase compartmentalized in modules containing a series of catalytic domains; type 2, consisting of a multi-enzymatic complex of several small and interactive enzymes with particular function; type 3, consisting of a PKS enzymes that form homodimers with a single catalytic site that performs iterative extensions of the polyketide product.[30] NRPSs have a similar architecture, containing a condensation (C) domain, an adenylation (A) domain and a peptidyl-carrier protein (PCP).[31] Combinatorial arrangement of these modules can produce a variety of natural product scaffolds that can be further decorated with epimerization

(inverting the AA stereochemistry) and other modifications by enzymes such as P450 oxidases and methyltransferases.

Another abundant biosynthetic scheme in cyanobacteria involves ribosomally synthesized and post-translationally modified peptides (RiPPs). In this class of BGC, a short precursor peptide is produced via ribosomal processes and then undergoes a number of modifications such as heterocyclization, methylation, oxidation, prenylation, and then macrocyclization to generate a final mature peptide. Cyanobactins, lathionines and microviridines are the most predominant BGCs encoding for RiPPs in cyanobacteria. However, it is intriguing that only one lanthionine has been isolated to date, known as prochlorosin. Cyanobactins are more abundant in the genera *Oscillatoria*, *Arthrospira*, and *Microcystis*, but are scarce in the genera *Prochlorococcus* and *Synechococcus*. No cyanobactins have been characterized outside the phylum cyanobacteria. Microviridins are a small family of cyclic N-acetylated trideca- or tetradeca-peptides containing lactam and lactone rings. This class of RiPPs generally possess potent serine protease inhibitory activity.

Genome mining is a powerful tool to prospect for RiPPs pathways because of the predictability of the precursor peptide and its post-translational modifications. About a third of all RiPPs characterized till date have been isolated via genome-guided NP discovery.[26] Examples include trichamide, the prochlorosins, the aeruginosamides, and viridamide (Figure 1.3). Trichamide represented one of the early efforts of bottom-up discovery in cyanobacteria,[32] preceding the development of the popular genome mining tool named antiSMASH (2011) for the automatic annotation of BGCs. The first cyanobactin discovered was patellamide, elucidating the biosynthetic logic behind cyanobactin pathways and serving as the “hook” for mining other cyanobactins. By using the patellamide genes, it was possible to identify a homologous and small BGC of 12.5 kb in the genome of toxic bloom-forming cyanobacterium, *Trichodesmium erythraeum* IMS101. The precursor peptide for trichamide contained similar post-translational

modification sites as the patellamide pathway, allowing for a relatively high-quality structure prediction for the NP using genomic information. The predictions were confirmed via LC MS/MS by linking the predicted amino-acids to mass fragments, demonstrating that RiPPs can exhibit a good parallel between genomic predictions and mass fragmentation experiments. Similarly, Li *et al.*[33] probed several marine *Prochlorococcus* and *Synechococcus* strains, and found a single *lanM*-like homolog for several *lanA*-like homologs. LanM cyclizes a single precursor LanA peptide for its maturation. In vitro evidence indicates that this LanM homolog has low substrate specificity, being capable of processing multiple different precursor peptides. Therefore, the cyanobacterium benefits from this enzyme promiscuity to create NP libraries, a hypothesis that subsequently led to the discovery of the prochlorosins. The first linear cyanobactins, aeruginosamides and viridamide, were detected via a large genome-mining effort of 126 cyanobacterial genomes. This mining identified 31 cyanobactin BGCs, which were present in 24% of the genomes. LC-MS/MS guided isolation led to the discovery of the first linear prenylated cyanobactins, the structures of which were confirmed via NMR and Marfey's analysis.

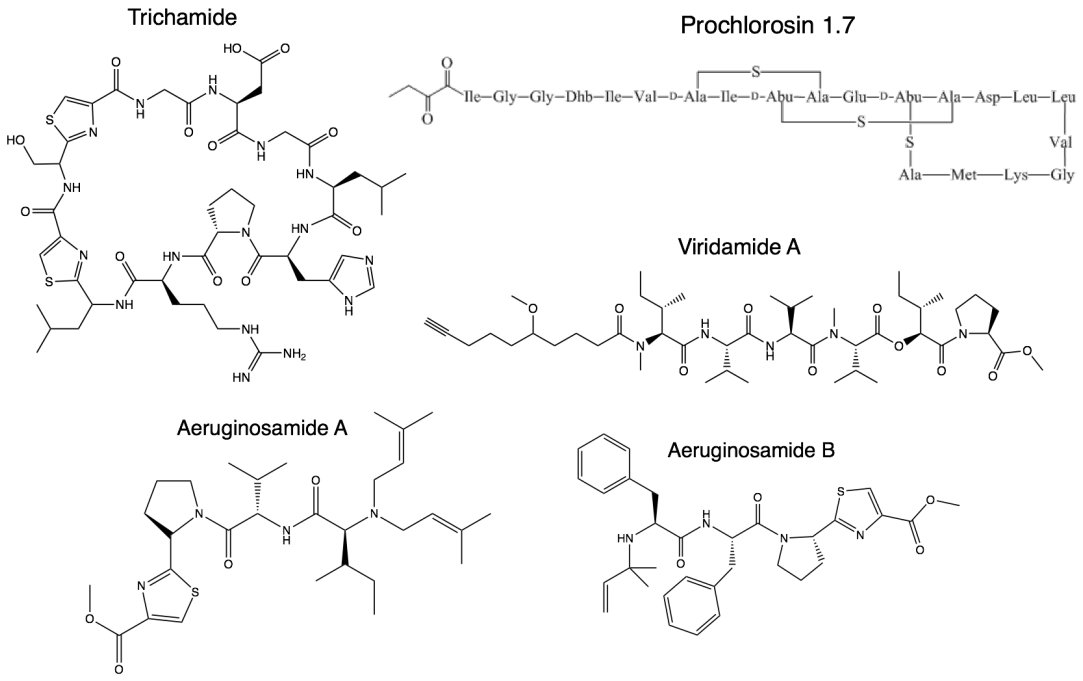


Figure 1.3. Structures for ribosomally synthesized and post-translationally modified peptides from cyanobacteria.

Lastly, two other NP classes are noteworthy: the indole-alkaloids and the mycosporine-like amino acids (MAAs). The order *Stigonematales* (subsection VIII) contains cyanobacteria known for the production of structurally complex indole alkaloids, including hapalindole A, welwitindolinone A isonitrile, and 12-epi-fisherindole G (Figure 1.4). The indole moiety is produced by tryptophan biosynthetic genes and the terpene moiety is a product of isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP) building blocks. The indole-alkaloid backbones possess several modifications that give rise to the current diversity in structures, including chlorination, oxidation, cyclization, prenylation, and methylation reactions.[34] A novel halogenation enzyme (WelO5) was discovered from the biosynthesis of an indole-alkaloid named welwitindolinone A. This non-heme iron-dependent enzyme is capable of chlorinating the free substrate without the intervention of carrier proteins.[35] Mycosporine and MAAs are a family of UV-absorbing compounds (310–360 nm) that can be divided into two different classes according to the number of AA attached to the cyclohexenone core. A single substitution defines the class of mycosporines, including mycosporine serinol and mycosporine

glycine (Figure 1.4), whereas a double substitution defines the MAA class, such as shinorine (Figure 1.4). Despite the conservation of the core biosynthetic genes, two different mechanisms can be used by cyanobacteria in order to attach the serine residue in the NP shinorine. *Anabaena variabilis* ATCC 29413 uses a NRPS gene for the AA condensation whereas *Nostoc punctiforme* ATCC 29133 uses a ATP-grasp ligase for catalyzing this same biosynthetic step.[36]

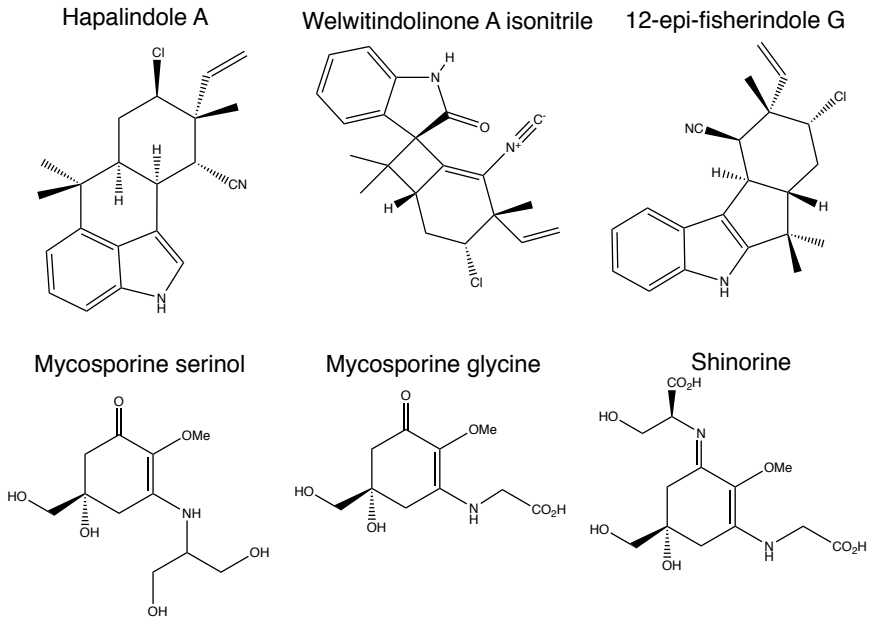


Figure 1.4. Structure for prenylated indole-alkaloids and mycosporine-like amino acids (MAAs) from cyanobacteria

1.4. Tool and strategies for analyzing biosynthesis of natural products

As previously mentioned, biosynthetic genes encoding for natural products tend to cluster together in microbial genomes. Hence, the annotation of a given biosynthetic gene cluster (BGC) can be achieved by identifying a "hallmark" gene within the queried genome and then inspecting its gene neighborhood for additional biosynthetic genes. This search procedure can be efficiently executed by the tool antiSMASH.[37] AntiSMASH uses Hidden Markov Models (HMMs) for annotation of BGCs. HMMs are built from alignments of experimentally validated proteins, with emphasis on the conserved residues from the protein's catalytic site. BLAST (especially DELTA-BLAST) can be another alternative tool to identify and annotate BGCs. BLAST works well for sequences that share a lot of homology. MultigeneBLAST and

CORASON can provide good visualization for BGC alignments, exhibiting the similarity between every homologous gene shared among two or more BGCs. For high-throughput analysis, BiG-SCAPE can generate similarity scores between pairs of BGCs, scores that can be displayed in a gene cluster network. This kind of network visualization displays BGCs being grouped together in gene cluster families (GCFs). Families that are rare (containing very few family members) tend to be of higher interest for genome mining. Challenges in this networking approach include: 1) the uncertainty of the correct positions for the BGC's borders, defining which genes would be included in the BGC, 2) the need to select a cutoff in order to define how the BGCs will group together, and 3) the presence of "superfamilies" that are connected to each other by weak links, incorrectly grouping together two or more families into one. Once BGCs are annotated, the predicted domains can provide clues concerning the structure of the biosynthesized natural product. The specificity of adenylation domains can be predicted, identifying which amino acid (AA) is probably incorporated, and epimerization domains can be used to predict the AA stereochemistry. Of particular note, ClusterFinder is another BGC detection tool that works by using HMMs from domains found at validated biosynthetic gene clusters (BGC state) and distinguishes them from domains found at random genome regions (non-BGC state). This later algorithm tends to annotate more BGCs, however, it also presents a higher false discovery rate.

Not many approaches can successfully connect a putative BGC to its biosynthetic product. For peptides with a high degree of co-linearity between NRPS genes and the final molecular scaffold, one can use the peptidogenomics[38] approach, predicting the AA chain and searching for MS2 fragments that match the predicted peptide. The same approach can be performed on RiPPs by analyzing the precursor peptide and predicting the post-translational modifications. Another approach by GARLIC[39] uses rules for biochemical reactions, converting complex scaffolds into a sequence of monomers, followed by an alignment of monomers from

the BGC versus monomers from characterized structures. This approach can be very efficient for BGC annotation, linking a DNA sequence to a natural product. However, this software requires the final structure and oftentimes BGCs can be useful earlier in the structure elucidation process.

On the metabolomics side, GNPS[40] allows for cosine comparison between MS2 fragmentation spectra. GNPS can analyze a given sample against a database with thousands of library spectra. If the cosine score is high enough, a given spectrum receives a library annotation, facilitating the analysis for untargeted metabolomics projects. Once cosine similarities are generated and a cutoff is selected, the molecules can be displayed in a network using CytoScape.[41] This visualization can be key for identifying the presence of a compound in some types of experiments (e.g. siderophores produced in high concentrations due to low iron in the media and absent in a second set of experiments due to high iron in the media).

In summation, the increasing quantity of genomics and metabolomics data will allow the development of more precise and efficient tools, in addition to enabling novel types of analyses in the natural product field, such as the creation of a systematic and automated approach for connecting cryptic metabolites to their respective biosynthetic gene clusters.

1.5. Dissertation contents

Given the number of structurally unique natural products isolated from filamentous marine cyanobacteria, it is surprising that there is a lack of genomic information for these prolific microbes. Hence, I undertook research to expand the genomic coverage of the phylum cyanobacteria in order to include dozens of new genomes from the NP-prolific genera *Moorea* and *Okeania*. More specifically, this endeavor involved the genetic description of *Moorea producens* JHB and an uncultured associated acidobacteria, the expansion of *Moorea* by obtaining one complete and two more draft genomes for a comparative genomics analysis, and

finally, a high-throughput effort to obtain several *Moorea* and *Okeania* genomes with the purpose of developing an automated genome mining approach.

Chapter 2 describes the 6 Mb complete genome from an acidobacteria associated with a non-axenic culture of *Moorea producens* JHB. The heterotrophic acidobacteria Mor1 share only 85% identity to the closest cultured bacteria. Mor1 lacks some of the genes for the biosynthesis of proteogenic amino acids such as Phe, Tyr, Trp, Lys, Arg, His, Gly and Ala. This acidobacteria has an abundance of transcriptional regulators, harboring over two times more regulatory genes than found in the genome of JHB. Transposases were completely absent in the Mor1 genome. In contrast, JHB contains several transposases, especially nearby NP biosynthetic genes, similar to other cyanobacteria. A high number transposases are common in the early stages of intracellular symbioses, hence, Mor1 is likely to be an extracellular associate, a hypothesis supported via semi-quantitative PCR and electron microscopy. This putative symbiont was found exclusively in cultures from *Moorea producens* and we were unable to transfer Mor1 via co-cultures of JHB along with other non-*Moorea producens* cyanobacteria.

Chapter 3 describes how we obtained the first complete genome from the genus *Moorea* and how we used this complete genome to improve three other draft genomes. A phylum-wide analysis (including only complete genomes) highlighted *Moorea* as one of the highest metabolic potential among cyanobacteria. We observed that nearly a fifth of *Moorea*'s genome tends to be dedicated to the production of secondary metabolites, one of the highest percentages among all bacteria. Our comparative analysis indicated that these four *Moorea* genomes were remarkably similar to one another and they diverged from the closest relative, *Microcoleus* sp. PCC7113. Lastly, our gene cluster networking approach indicated that the numerous BGCs from *Moorea* are also fairly unique and rarely found among other cyanobacteria.

Chapter 4 introduces a 'bottom-up' approach named pattern-based genome mining using Jaccard Index as an indicator of beta-diversity. First, we sequenced and assembled 165 draft

genomes from 143 environmental samples of potentially prolific marine cyanobacteria. A total of 85 of these genomes passed the quality control and exhibited excellent completeness regarding housekeeping genes for primary metabolism. These draft genomes were annotated and analyzed for biosynthetic gene clusters.

Lastly, Chapter 5 summarizes the genetic discoveries made in the previous chapters, as well as it highlights future directions for the projects described herein. This summary includes the impact of my research on our genetic understanding of cyanobacteria as well as their prolific capacity to produce natural products.

Chapter 1, the topics “overview of the genetic diversity from cyanobacteria” and “Biosynthesis of cyanobacterial natural products” are reproduced, with permission, of the material as it appears in *Taylor and Francis Publishers*, 2017, Lena Keller, Tiago Leao and William H. Gerwick.

CHAPTER 2: A Novel Uncultured Heterotrophic Bacterial Associate of the Cyanobacterium *Moorea producens* JHB

2.1. Abstract

Filamentous tropical marine cyanobacteria such as *Moorea producens* strain JHB possess a rich community of heterotrophic bacteria on their polysaccharide sheaths; however, these bacterial communities have not yet been adequately studied or characterized. Through efforts to sequence the genome of this cyanobacterial strain, the 5.99 MB genome of an unknown bacterium emerged from the metagenomic information, named here as Mor1. Analysis of its genome revealed that the bacterium is heterotrophic and belongs to the phylum acidobacteria, subgroup 22; however, it is only 85% identical to the nearest cultured representative. Comparative genomics further revealed that Mor1 has a large number of genes involved in transcriptional regulation, is completely devoid of transposases, is not able to synthesize the full complement of proteogenic amino acids and appears to lack genes for nitrate uptake. Mor1 was found to be present in lab cultures of *M. producens* collected from various locations, but not other cyanobacterial species. Diverse efforts failed to culture the bacterium separately from filaments of *M. producens* JHB. Additionally, a co-culturing experiment between *M. producens* JHB possessing Mor1 and cultures of other genera of cyanobacteria indicated that the bacterium was not transferable. Thus, these data support a specific relationship between this novel uncultured bacterium and *M. producens*.

2.2. Introduction

Filamentous cyanobacteria, bathed in seawater and often growing in nutrient-rich environments, are surrounded by diverse communities of heterotrophic bacteria. The heterotrophic bacteria closely associated with cyanobacteria likely consume released nutrients, but may also produce vitamins and other factors useful to cyanobacterial growth, as well as

assisting in cycling of CO_2 and phosphate, or lowering O_2 levels for oxygen-sensitive processes such as nitrogen fixation.[6] Various studies have classified some of the taxa of heterotrophic bacteria that live in close proximity to cyanobacterial blooms, including common aquatic phyla such as *Proteobacteria, Bacteroidetes, Actinobacteria*, and *Planctomycetes.*[42,43] Some potentially new species or genera were also located within these samples, which could suggest that some bacteria may have specific relationships with cyanobacteria.[42] However, many of these latter bacteria are also found living independently of cyanobacteria,[43] and the makeup of cyanobacterial-associated communities varies based on the location, type of cyanobacteria, and environmental conditions including nutrient availability and temperature.[43–45] The heterotrophic bacterial community around cyanobacterial blooms appears to be directly influenced by the bloom in that the community structure changes over its progression.[45] In fact, if the cyanobacteria are eliminated by a viral infection, the heterotrophic bacterial community drastically shifts.[46] Conversely, heterotrophic bacteria have also been shown to affect the growth of cyanobacteria. Various strains of bacteria found living with *Nodularia spumigena* were co-cultured with the cyanobacterium, and several were found to either increase or decrease the growth of the cyanobacterium compared to axenic cultures.[47] Additional studies of heterotrophic bacteria associated with cyanobacterial blooms have verified that co-cultures can increase or decrease cyanobacterial growth.[48] This is likely due to specific interactions of carbon and nutrient exchange.[49] However, the interactions between cyanobacteria and natural assemblages of heterotrophic bacteria involve a large number and variety of interfaces, and are certainly more complex than a specific symbiosis involving two specific partners. Thus, it becomes clear that gaps exist in our knowledge of the microbial communities surrounding cyanobacteria.

Cyanobacteria have been a rich source of bioactive natural products (secondary metabolites and/or toxins), and their biosynthesis has been studied at the chemical, biochemical and genomic levels.[50–52] There is some evidence that the bacterial communities associated with

cyanobacteria may affect these biosynthetic processes in different ways. Heterotrophic bacterial communities surrounding cyanobacteria may not only change cyanobacterial growth characteristics, but also have the potential to break down cyanobacterial toxins[53] or modulate toxin production. For example, toxic *Microcystis* blooms with different heterotrophic bacteria produce altered microcystins of varying toxicity.[54]

Additionally, in some cases there are uncertainties about which organism is the true producer of a natural product, or if a collaborative biosynthetic effort is required between the cyanobacterium and a heterotrophic bacterium. Considering the complexity of these metabolites and their assembly pathways, it is highly unlikely that the pathways separately evolved in such divergent organisms as cyanobacteria and heterotrophic bacteria. For example, the lyngbyatoxins, a class of potent skin irritants and tumor promoters isolated from field collections of *M. producens,* show high structural and pharmacological similarity to teleocidin, a metabolite which is produced by *Streptomyces* species.[53,55] Similarly, an extract from an assemblage of the cyanobacteria *Moorea producens* and *Tolypothrix* sp. yielded the toxin kalkipyrone; this metabolite is closely related to the *Streptomyces* metabolites actinopyrone and iromycin.[56,57] Another example is given by swinholide A, an actin-binding toxin originally isolated from the sponge *Theonella swinhoei*, but subsequently shown to originate from a member of the complex community of heterotrophic bacteria growing within the sponge.[58] However, swinholide A was also isolated from field collections of a marine cyanobacterium along with a glycosylated derivative,[59] initially creating some confusion about the true metabolic source of this complex polyketide. Nevertheless, recent characterization of closely related gene clusters for swinholide-like molecules from a heterotrophic bacterial symbiont of the sponge, *Entotheonella* sp., and several cultured cyanobacteria, reveals a complex evolutionary relationship between these pathways, and suggest a mixture of vertical inheritance and convergent evolutionary processes.[60] As these examples illustrate, there is considerable uncertainty concerning the true

biosynthetic source of secondary metabolites isolated from cyanobacteria that possess natural assemblages of heterotrophic bacteria. Overall, the study of these cyanobacterial-associated heterotrophic bacterial communities is important as it relates to the ecology and physiology of these organisms as well as the roles and production of their secondary metabolites.

Moorea producens (previously *Lyngbya majuscula*)[61] is a filamentous tropical marine cyanobacterium capable of photosynthesis but unable to fix atmospheric nitrogen. Members of this genus are known to be prolific producers of natural products; around 200 secondary metabolites have been isolated from this organism, and the genomes of various strains contain many polyketide synthase and non-ribosomal peptide synthetase genes.[61–63] *M. producens* has been observed to possess a large community of bacteria on its filaments.[61] However, very little is known about this bacterial community or its inter-relationships and interactions with the cyanobacterial host. One such strain, *M. producens* JHB, a known producer of the natural products hectochlorin,[64] the jamaicamides,[65] and cryptomaldamide [unpublished], was originally collected from a shallow habitat in Hector's Bay, Jamaica in 1996. It has been maintained in uni-cyanobacterial culture since this time along with its associated heterotrophic bacterial community. The metagenome of this *M. producens* JHB strain was sequenced and assembled, followed by extensive binning for cyanobacterial versus heterotrophic bacterial DNA. This process yielded a draft genome of the cyanobacterium along with the essentially complete 5.99 MB genome of a *M. producens* JHB-associated bacterium. Analysis of this latter bacterial genome, along with experiments to determine its identity and potential function as an associate of *M. producens* JHB, is the focus of this current report.

2.3. Methods

2.3.1. Cyanobacterial cultures

Moorea producens JHB (GenBank: FJ151521.1) was collected in Hector's Bay, Jamaica in August 1996.[65] *M. producens* 3L (NR116539) was collected in December of 1993 at Las

Palmas Beach near the CARMABI Research Station in Curaçao, Netherland Antilles, N12 07.387' W68 58.157'.[62] 3L *Oscillatoria* (EU244875), identified as *Oscillatoria nigroviridis,* was isolated as a contaminant of the 3L *M. producens* strain. *M. bouillonii* (FJ041298) was collected in May of 2005 near Pigeon Island in Papua New Guinea, S4 16.063' E152 20.266'. *Leptolyngbya* sp. (ISBN3Nov94-8, KC207938.1) was collected in November of 1994 near Sulawesi, Indonesia. PAP25Jun12-3 was collected in June of 2012 near Portobello, Panama. All of these were established and maintained as uni-cyanobacterial cultures using standard microbiological isolation techniques.[65,66] The cultures were grown under static conditions at 28°C under uniform illumination (4.67 µmol photon s^{-1} m^{-2}) with a 16hr/8hr light/dark cycle provided by 40 W cool white fluorescent lights. SWBG-11 media contained 35 g/L Instant Ocean (Aquarium Systems Inc.).

2.3.2. DNA extraction and sequencing

DNA was extracted from the harvested biomass of cultures of *M. producens* JHB, along with its microbiome of heterotrophic bacteria, using the JGI phenol-chloroform protocol. The metagenomic DNA was sequenced using the Illumina HiSeq system, paired end library of 2 x 100bp. Approximately 12 GB of data were obtained.

2.3.3. Assembly and other bioinformatics

The metagenomic reads were assembled using SPAdes version 3.0.0.[67] The contigs were binned by GC content, coverage, tetranucleotide fingerprint, and phylogenetic classification of 107 single copy genes. This binning strategy strongly suggested that the six largest non-cyanobacterial contigs most likely belonged to the same taxon. Using Geneious De Novo Assembler (Geneious®), an isolated reassembly of these six contigs resolved the repeated regions and generated a circular scaffold comprised of a single contig that only lacked part of the 16S-ITS-23S rRNA operon. However, previous PCR experiments (as described below) had already provided a single and complete 16S rRNA sequence. The final circular

scaffold, including the complete 16S rRNA gene, was submitted for automatic annotation using RAST.[68] Numbers of copies of this complete 16S rRNA gene were confirmed by comparing the coverage of a single copy gene found only in this genome (*selA*) versus the coverage of the 16S rRNA gene, confirming that a single 16S rRNA gene was present, and most likely, only a single 16S-ITS-23S operon as well. The identified *selA* gene, which is unique to the Mor1 genome, was used for further experiments as a marker for presence or absence of the Mor1 bacterium. In addition, a more detailed annotation of the Mor1 genome was obtained by submitting the genome to the expert reviewed annotation at JGI (Joint Genome Institute) IMG/ER web platform and to antiSMASH[69] for identification/annotation of secondary metabolite biosynthetic gene clusters.

The assembly of the *M. producens* JHB genome was performed using a combination of assembly utilizing SPAdes along with a reference assembly to a closed *Moorea* genome (unpublished). This assembly generated a single scaffold of 9.37 Mb with a 43.5% GC content. The *M. producens* JHB genome was submitted to the same annotation tools as Mor1, and comparative genomics and statistics between Mor1, *M. producens* JHB, and other genomes were developed using Genome Statistics, Search Pathways, COG Homology and the Abundance Profile tools from the JGI (Joint Genome Institute) IMG/ER database.

The relative abundance of *M. producens* JHB and Mor1 in the overall sequenced metagenomic sample was estimated by the percentage of reads recruited to each of these draft genomes compared to the total number of metagenomic reads (approximately 16 Mb). The genomes were normalized by the average genome size of 3.6 Mb (average size of all 3,777 complete bacterial genomes currently available at JGI database). Similarly, the same percentage was calculated for 421 contigs (maximum size of 10,422 bp, total size of 225,489 bp) not assembled into a genome and not belonging to neither JHB or Mor1 (representing other JHB associates), The recruitment of reads was performed by using Bowtie2 mapping with the option end-to-end and disregarding pair end reads to minimize the exclusion of reads from small

contigs. Gene calling using Prodigal was performed and the predicted open reading frames (proteins) were submitted to DarkHorse[70] in order to infer phylogenetic classification for these open reading frames.

2.3.4. 16S rRNA gene location and analysis

The full length16S rRNA sequence was obtained using PCR. DNA was extracted from *M. producens* JHB cultures using the QIAGEN Genomic-tip 20/G kit and following its standard protocol, and PCR was performed using 25 µL volumes, containing 12.5 µL of 2x *Taq* Master Mix, 0.5 µL $MgCl_2$ (25 mM), 1.0 µL of each primer (10 µM), 1.0 µL of DNA template, and 9 µL sterile water. The amplification conditions were as follows: initial denaturation at 95ºC for 4 min, followed by 30 cycles of 95ºC for 30 s, 56ºC with 1 Fw + 1451 Rv or 61ºC with 1 Fw + 899 Rv/1151 Rv for 30 s, and 72ºC for 30 s, followed by a final extension step at 72º for 1 min. Primer sequences are shown in Table 2.1, and were designed based on the sequence of the 16S rRNA gene from *Escherichia coli* strain K-12**.** The ensuing PCR product was cloned into the pCR 4-TOPO Vector (Invitrogen TOPO TA Cloning Kit) using the standard protocol, followed by sequencing. The full 16S rRNA sequence was analyzed by BLASTn and RDP Classifier[71] to gain more insights into the phylogenetic characteristics of the unknown organism Mor1.
A phylogenetic tree based on this 16S rRNA gene was created, incorporating 16S rRNA sequences from acidobacteria, proteobacteria, and cyanobacteria. 16S rRNA sequences were obtained from GenBank, then aligned using MUSCLE aligner with 5 iterations, gap open score -1 and word size of 5 bp. The tree was built using Geneious Tree Builder, with the Jukes-Cantor genetic distance model, Neighbor-joining tree build method, 100 bootstraps, and *Anabaena variabilis* ATCC 29413 as the out-group.

Table 2.1. Primers designed for use in this study

Primer name	Primer sequence	T_m in °C
1 Fw	5′ -AAGGAGGTGATCCAGCCGCAGG- 3′	66.0
899 Rv	5′ -TGAGAGGGTGACCGGCCACACT- 3′	67.0
1151 Rv	5′ -AGGCGACGATGGGTAGCCGACC- 3′	68.0
1451 Rv	5′ -CTGGAGAGTTTGATCCTGGCTCAG- 3′	61.0
selA Fw 428	5′ -ACTATCGCAAGGCGATCAACAAGA- 3′	58.6
selA Rv 1180	5′ -CTAGCTCATCGCTCCTATCAG- 3′	58.3

2.3.5. Culturing attempts of the associated bacterial community from *Moorea* producens JHB

Efforts to culture Mor1 separately from filaments of *M. producens* JHB used a variety of solid media containing 2% agar, as listed in Table 2.2. Intact or cut filaments of *M. producens* JHB were placed onto each media type. For some culturing trials, *M. producens* JHB filaments were freeze-dried and ground up and then added to the media. Additionally, associated bacteria were washed from the surface of *M. producens* JHB filaments using the following protocol: 2 g of biomass was placed into 10 mL of 0.45 M NaCl, 10 mM KCl, 7 mM Na_2SO_4, 0.5 mM $NaHCO_3$, and 10 mM EDTA. Added to this was 0.1 mL filter-sterilized Rapid Multienzyme Cleaner (3M). The sample was then incubated for 2 h at room temperature while shaking at 80 rpm. The sample was vortexed and then centrifuged at 300 x g for 15 minutes. An aliquot of the supernatant (50-100 μL) containing associated bacteria was then plated onto the various types of media.

Table 2.2. Solid media utilized for culturing heterotrophic bacteria associated with *M. producens* JHB sheaths

Media name	Media content	Enrichment
SWBG-11	SWBG-11	N/A
Enriched SWBG-11	SWBG-11	0.4 % glucose
MA	Difco Marine Agar	N/A
SSS	3 % Sigma Sea Salt + 0.4 % mannose + 0.3 % casamino acids	N/A
Enriched SSS	3 % Sigma Sea Salt + 0.4 % mannose + 0.3 % casamino acids	0.5 µM or 2 µM Ferric Ammonium Citrate
A1	1 % starch + 0.2 % yeast extract + 0.4 % peptone	N/A
SWBG-11	SWBG-11	Media mixed with cut up filaments of *Moorea producens* JHB
SWBG-11	SWBG-11	Media mixed with freeze dried and ground up filaments of *Moorea producens* JHB

Bacterial colonies that grew on these plates were isolated and grown overnight in liquid media. DNA was extracted from the overnight cultures using the Wizard Genomic DNA Purification Kit (Promega). PCR was performed on the DNA samples using the *selA* primers (Table 2.1), and the 16S rRNA 27F and 1492R primers.[71] PCR was carried out in 25 μL volumes, containing 12.5 μL of 2x *Taq* Master Mix, 0.5 μL $MgCl_2$ (25 mM), 1.0 μL of each primer (10 μM), 1.0 μL of DNA template, and 9 μL sterile water. The amplification conditions were as follows: initial denaturation at 95 ºC for 4 min, followed by 30 cycles of 95 ºC for 30 s, 55 ºC for 30 s, and 72 ºC for 30 s, followed by a final extension step at 72 ºC for 1 min. The 16S rRNA PCR products were then cloned into the pCR 4-TOPO Vector (Invitrogen TOPO TA Cloning Kit)

using the standard protocol, followed by sequencing. The obtained sequences were analyzed using BLASTn.

2.3.6. Electron microscopy

Samples for TEM were prepared using 2% glutaraldehyde in saltwater (1 h), 2 x 5 min rinses in saltwater, 1% osmium tetroxide (1 h), 1 x 5 min rinse in 0.15 M cacodylate buffer, 2 x 5 min rinses in ddH_20 and 2% uranyl acetate overnight. This was followed the next day with 2 x 5 min ddH_20 rinses. Dehydration was achieved with a graded (20%, 50%, 70%, 90%, 100%, 100%) EtOH series. The samples were then embedded in 50/50 mixture of Spurr's/EtOH overnight. The next day the samples were incubated in 100% Spurr's for 24 h, after which the samples were placed in 100% fresh Spurr's for 2 x 1 h and left to polymerize for 48 h. Thin sections (70 nm) were obtained using an Ultracut E microtome (Reichert-Jung, Vienna, Austria) and then placed on 200 mesh fine bar copper grids. The grids were subsequently stained with uranyl acetate and Sato lead. A 1200FX TEM (JEOL, Tokyo, Japan) was used to view the samples.

2.3.7. Semi-quantitative PCR of DNA from washed and unwashed filaments

One sample of JHB filaments was prepared according to the wash protocol specified above in "Culturing trials of the associated bacterial community of *Moorea producens* JHB". After centrifugation, the supernatant was removed and the cyanobacterial filaments were used as a "washed filament" sample.

DNA was extracted from the washed filament sample as well as an equivalent mass of unwashed JHB filaments using the Wizard Genomic DNA Purification Kit (Promega) with the standard protocol. PCR was then performed on both samples using the *selA* primers (Table 2.1). PCR was carried out in 25 μL volumes, containing 12.5 μL of 2x *Taq* Master Mix, 0.5 μL $MgCl_2$ (25 mM), 1.0 μL of each primer (10 μM), 1 μL (18.6 ng/μL) of DNA template, and 9 μL sterile water. The amplification conditions were as follows: initial denaturation at 95ºC for 4 min,

followed by 28 cycles of 95ºC for 30 s, 55ºC for 30 s, and 72ºC for 30 s, followed by a final extension step at 72º for 1 min. After PCR, the samples were run on a 1% agarose gel and the intensities of the bands quantified using Gel Quant Express (Life Technologies).

2.3.8. Examination for the presence of Mor1 in other cultures

In order to examine the extent of Mor1 in other of our laboratory cultures, a possible indication of cross-contamination between cultures, the following were tested for the presence of the *selA* gene (present in Mor1 but not *M. producens*; see Results): *Moorea producens* JHB, *Moorea producens* 3L, *Moorea bouillonii*, 3L *Oscillatoria*, *Scytonema hoffmani* "2846 axenic and xenic," *Leptolyngbya* sp. (coded ISBN3Nov94-8), and PAP25Jun12-2. In each case, DNA was extracted from several grams of wet biomass using the QIAGEN Genomic-tip 20/G kit with its standard protocol. PCR was performed on the extracted DNA samples using the *selA* primers (sequences indicated in Table 2.1). PCR was carried out in 25 µL volumes, containing 12.5 µL of 2x *Taq* Master Mix, 0.5 µL $MgCl_2$ (25 mM), 1.0 µL of each primer (10 µM), 1.0 µL of DNA template, and 9 µL sterile water. The amplification conditions were as follows: initial denaturation at 95ºC for 4 min, followed by 30 cycles of 95ºC for 30 s, 55ºC for 30 s, and 72ºC for 30 s, followed by a final extension step at 72º for 1 min.

2.3.9. Co-culturing of the JHB strain with other cyanobacteria

To determine whether Mor1 could be transferred from *M. producens* JHB to other cyanobacteria, co-culturing experiments were performed. The cyanobacteria chosen for co-culture were *M. producens* 3L, *Oscillatoria*, *Leptolyngbya* sp. (coded ISBN3Nov94-8), and PAP25Jun12-2. The co-cultures and controls were set up as diagrammed in Table 2.3. Each co-culture or control was grown in duplicate, in 250 mL of SWBG-11 each. After 2 weeks, the co-cultures were separated under sterile conditions using a dissecting microscope, and grown in SWBG-11 until several grams of wet biomass could be obtained for DNA extraction.

Table 2.3. Matrix of co-culturing experiments involving M. producens JHB and various other filamentous tropical marine cyanobacteria

	M. producens JHB	3 L *Oscillatoria*	*Leptolyngbya* sp.	PAP25Jun12-2
M. producens JHB	Single culture control	Mor1 transfer experiment	Mor1 transfer experiment	Mor1 transfer experiment
3 L *Oscillatoria*	X	Single culture control	Mor1-absent control	Mor1-absent control
Leptolyngbya sp.	X	X	Single culture control	Mor1-absent control
PAP25Jun12-2	X	X	X	Single culture control

The entries indicate the intent of each co-culturing combination

DNA extraction was performed for each sample utilizing JGI's phenol-chloroform protocol or the QIAGEN Genomic-tip 20/G kit using the standard protocol. Each DNA sample was then tested for the presence of the 16S rRNA gene (to indicate sample quality) and the *selA* gene using PCR. Primers 27F and 781R for 16S rRNA,[72] and *selA*Fw 428 and *selA*Rv 1180 were used (Table 2.1). PCR was carried out in 20 μL volumes, containing 10 μL of 2x *Taq* Master Mix, 0.5 μL $MgCl_2$ (25 mM), 1.0 μL of each primer (10 μM), 1.0 μL of DNA template, and 6.5 μL sterile water. The amplification conditions were as follows: initial denaturation at 95ºC for 4 min, followed by 30 cycles of 95ºC for 30 s, 50ºC with 16S rRNA or 55ºC with *selA* for 30 s, and 72ºC for 60 s, followed by a final extension step at 72ºC for 7 min.

2.3.10. Accession number

The genome of the bacterium Mor1 has been deposited in GenBank under the accession number CP011806.

2.4. Results and discussion

2.4.1. Genome assembly and annotation

The non-axenic uni-culture of the cyanobacterium *M. producens* JHB, originally collected in Hector's Bay, Jamaica, was sequenced along with its associated heterotrophic bacterial community by Illumina HiSeq sequencing, assembly with SPAdes, binned and reassembled with the Geneious *De Novo* Assembler. In addition to a single scaffold for the cyanobacterial genome (to be reported separately), this process yielded a 5.99 Mb contig from an associated bacterium. Average coverage of this bacterial contig was 33.6 fold and it possessed a 66.8%

GC content, very different from the 43.5% GC content of the *M. prod*ucens JHB genome. Further analysis revealed that the scaffold was circular and lacked only a fraction of the 16S-ITS-23S operon (partial 16S and 23S rRNA genes were present but the full ITS region was absent) comprising 2820 nucleotides between the 5' and 3' ends of the circular scaffold. Continued search of the raw sequencing data was unsuccessful to resolve this region. However, sequence data from PCR amplification of the complete 16S rRNA gene was incorporated into the scaffold (hereafter named the Mor1 chromosome).

The assembled complete genome of Mor1 was submitted for rapid automatic annotation through RAST (http://rast.nmpdr.org/), and using RAST's SEED Viewer, revealed that the genes for photosynthesis or carbon fixation were lacking, thus indicating that it was heterotrophic. Preliminary comparison of the genomes of *M. producens* JHB and Mor1, again using RAST, revealed several genes present in the bacterium but not in JHB. Of these, the L-seryl-tRNA selenium transferase gene (*selA*) was selected for use as a specific genetic marker of Mor1, useful for examining the presence of this organism in other cyanobacterial strains. This gene encodes for the tRNA incorporation of selenium-containing cysteine residues in proteins and is not common in cyanobacteria.[73] Additional inspection of the sequenced metagenome revealed that the *selA* gene was present only in the Mor1 chromosome and as a single copy. Moreover, BLASTN analysis revealed that none of the 30,622 bacterial genomes in the JGI database contains a single sequence with more than 50% coverage and 90% identity to the Mor1 *selA* gene, identifying this gene as an excellent specific genetic marker of this bacterium. Consequently, specific primers were created from the sequence of the *selA* gene and utilized in later experiments as described below.

The antiSMASH program was used to identify secondary metabolite pathways within the Mor1 genome,[69] and revealed one polyketide type III pathway and one terpene biosynthetic pathway. The polyketide synthase (PKS) type III pathway has high amino acid identity as well as open reading frame organization with the alkylresorcylic acid pathway in *Myxococcus*

xanthus,[74] as shown in Figure 2.1. However, efforts to identify alkylresorcylic acid from the chemical extract of the *M. producens* JHB consortium using the GnPS mass spectral network were not successful.[40] Thus, from gene sequence and MS analyses, Mor1 is not a major producer of recognizable secondary metabolites (e.g. PKS, NRPS or hybrid natural products typical of cyanobacteria).

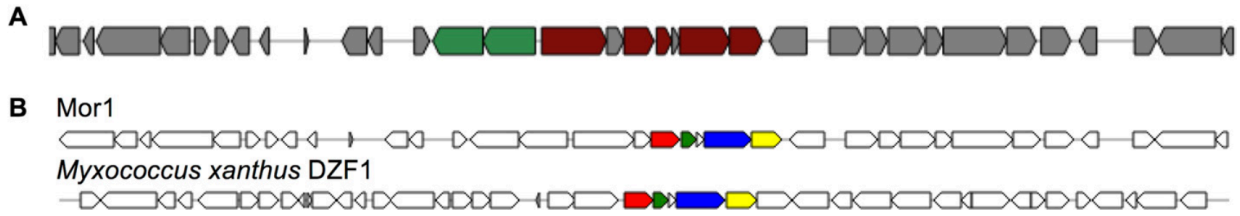


Figure 2.1. Schematic of the polyketide synthase (PKS) type III pathway discovered within the Mor1 genome using antiSMASH. A) The entire Mor1 type III PKS gene cluster with maroon genes corresponding to biosynthetic genes, green genes corresponding to regulatory genes, and grey genes corresponding to other uncharacterized genes. This locus is located between nucleotides 1238773 – 1279822 of the genome. B) The comparison of the Mor1 gene cluster with homologous areas within the alkylresorcylic acid pathway gene cluster in the genome of Myxococcus xanthus DZF1. The red arrow corresponds to a stilbene synthase gene with 58% identity and 99% coverage; the green arrow corresponds to a methyltransferase gene with 54% identity and 85% coverage; the blue arrow corresponds to an AMP-dependent synthetase with 54% identity and 96% coverage; the yellow arrow corresponds to a monooxygenase gene with 47% identity and 69% coverage. Images generated by antiSMASH.

Comparison of the 16S rRNA sequence of Mor1 to NCBI's database via BLAST revealed an identity of less then 95% to an unknown, uncultured bacteria from marine environmental sediment samples. The closest match for a cultured bacterium was *Desulfobacca acetoxidans* (GenBank: NC_015388.1), with an 85% identity and 100% coverage. The 16S rRNA sequence was submitted to RDP Classifier for further phylogenetic characterization.[71] This resulted in identification of the organism as belonging to the phylum acidobacteria, subgroup 22 with a 100% confidence threshold. Further taxonomic classification of acidobacteria subgroup 22 does not currently exist;[75,76] thus, this bacterium belongs to an, as yet, unnamed and unidentified genus and species. A phylogenetic tree comparing the 16S rRNA sequence of Mor1 with those of other bacteria, including members of acidobacteria, cyanobacteria, and proteobacteria, is depicted in Figure 2.2**.** Although the phylum acidobacteria is not currently well-classified, members of the phylum have been discovered living within the associated communities of

marine sponges and zoanthids, suggesting that marine acidobacteria are capable of complex interactions and symbioses with other organisms.[77,78]

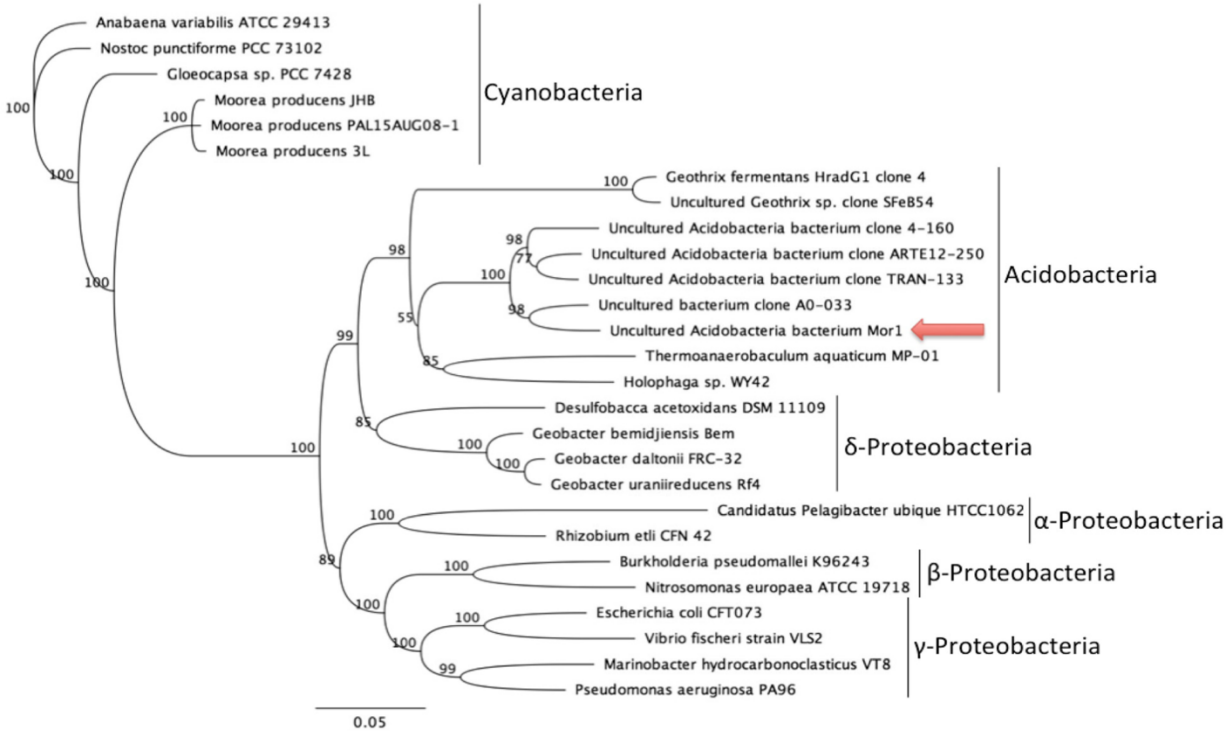


Figure 2.2. Phylogenetic tree comparing the 16S rRNA sequence of Mor1 to those of other bacteria. *Anabaena variabilis* ATCC 29413 was used as the outgroup. Mor1, indicated by an arrow, clusters with uncultured Acidobacteria strains, indicating that it likely belongs to a novel clade of phylum Acidobacteria.

2.4.2. Relative abundance and estimated consortium composition

The relative abundance of *M. producens* JHB, Mor1, and other associates was estimated from the recruitment of raw reads. As expected, *M. producens* JHB was the most abundant taxon represented by 64% of the reads in the metagenomic sample. This was followed by Mor1, with 19% of the reads. The total relative abundance of all other associates was 17%, implying that Mor1 is more abundant than the sum of all the other associated heterotrophic bacteria. Phylogenetic classification of these other associates using DarkHorse allowed for a qualitative assessment of the consortium composition (Figure 2.3). Unfortunately, the relative abundance of each taxon designated by the DarkHorse analysis cannot be precisely

quantified, due to the short lengths of the contigs, which dramatically increase the chances of a highly repetitive gene skewing the read estimation by up to an order of magnitude.

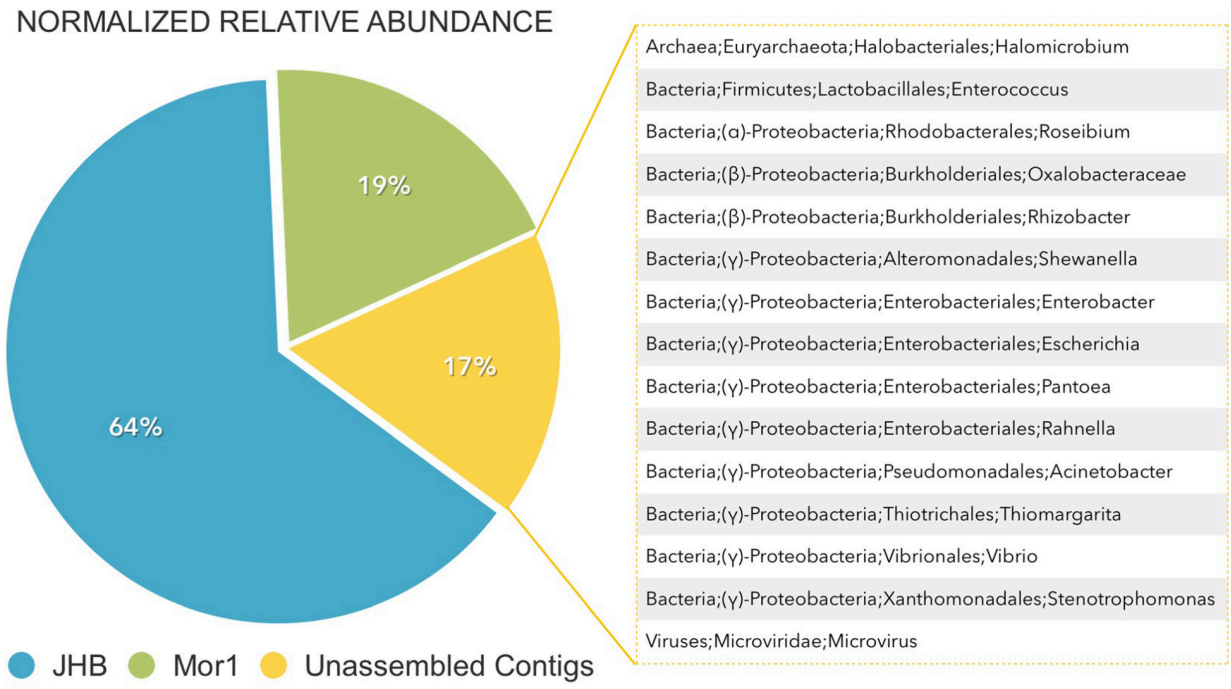


Figure 2.3. Relative abundance of raw reads belonging to *Moorea producens* JHB, Mor 1 and all the reads unassembled, thus, not belonging to either *Moorea producens* JHB or Mor1

2.4.3. Efforts to culture Mor1 free of *M. producens* JHB

A number of different kinds of media, as described in the Methods and Table 2.2, were evaluated for Mor1 cultivation. These culture attempts included numerous nutrient combinations as well as enrichments with iron. In order to provide potentially required growth factors for Mor1 culture that might be found in *M. producens* JHB, filaments of JHB were cut into short pieces, freeze-dried and ground with a mortar and pestle, and then added to nutrient agar for culturing experiments. Source bacteria for these culture attempts were obtained from the cyanobacterial sheaths by a wash procedure described in the Methods, and bacteria were cultured from the wash buffer. As a result of these trials, dozens of different bacterial cultures were obtained. By 16S rRNA analysis, these included species of the genera *Muricauda*,

Alteromonas, *Rhodovulum*, and *Alcinovorax*, as well as *Marinobacter salsuginis* and a *Rhodobacteraceae* strain. However, none of the cultured strains were found to possess the s*elA* gene by PCR analysis, and hence, Mor1 was not among the culturable bacteria from *M. producens* JHB. We propose that Mor1 has nutrient requirements not met by any of these supplemented media types.

2.4.4. Indicatives that Mor1 exists mainly on the exterior of the *M. producens* JHB sheath

TEM images of cross-sections of JHB filaments are shown in Figure 2.4. Bacteria are evident on the outside of the polysaccharide sheath, but the space between the cyanobacterial cell and the sheath appears free of bacteria, and intracellular bacteria are also not in evidence. Thus, Mor1 is likely located on the exterior of the cyanobacterial sheaths. To further explore this hypothesis, we examined two samples of *M. producens* JHB using semi-quantitative PCR of the *selA* gene. One sample contained the intact external bacterial community (unwashed) whereas the second sample was subjected to a wash protocol (described in Methods) designed to remove a substantial fraction of the externally attached bacteria. The *selA* signal was decreased in the washed sample by 56.6%, indicating that a majority of Mor1 was removed by the washing procedure. Thus, we believe that Mor1 exists predominantly on the outside surface of JHB sheaths (Figure 2.5).

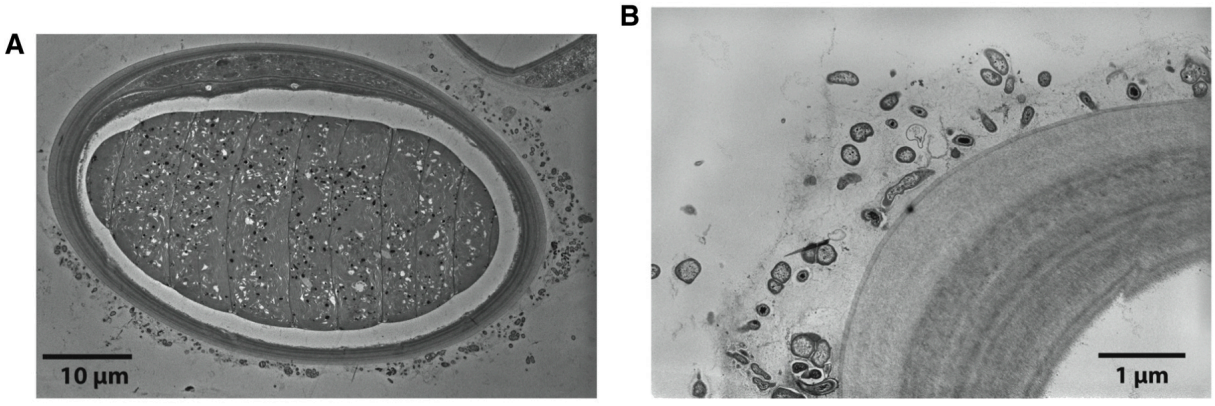


Figure 2.4. Transmission electron microscopy **(**TEM) images of 1 Moorea producens JHB, showing its polysaccharide sheath and the location of its associated bacterial community. A) Cross section of a filament of M. producens JHB,

showing the cell centrally, surrounded by the intermembrane space and polysaccharide sheath. Note that the bacterial growth appears outside the polysaccharide sheath, and not within the intermembrane space. B) Close-up of bacterial growth on the outside of the polysaccharide sheath.

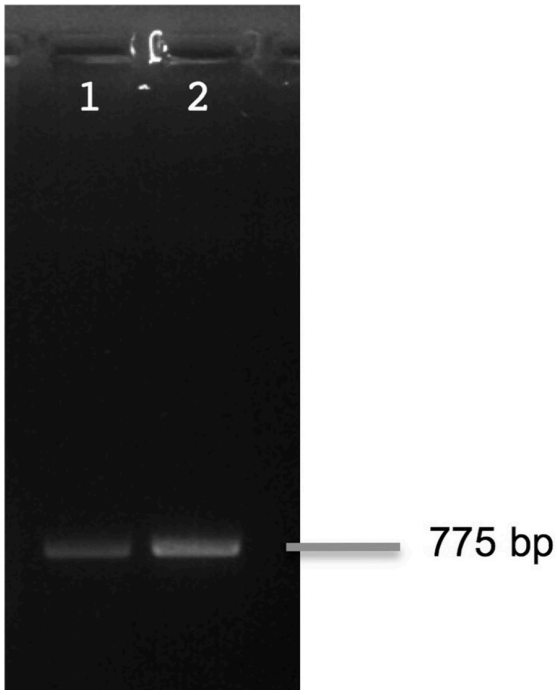


Figure 2.5. *Semi-quantitative PCR of washed and unwashed M. producens JHB samples, visualized on a 1% agar gel. Lane 1: Washed Moorea producens JHB sample. Lane 2:* Unwashed Moorea producens JHB sample. External washing of the filaments reduced the incidence of the Mor1 selA gene by an estimated 56.63% using densitometry.

2.4.5. Examination of the specificity of Mor1 on *Moorea* spp.

To explore whether Mor1 is a specific associate of *Moorea* and not generally a microbial constituent of our laboratory cyanobacterial cultures, seven different genera/species were tested for the presence of the *selA* gene (see Methods and Figure 2.6)**.** The *selA* gene only appeared in cultures of *M. producens* 3L collected in Curaçao and in *M. producens* JHB from Jamaica, and not in any other of our cyanobacterial cultures, including *Moorea bouillonii* from Papua New Guinea. On the basis of this observation, Mor1 was deduced to not be a general laboratory bacterial contaminant in our cultures, and thus we speculated that it is a highly specific associate of *M. producens*.

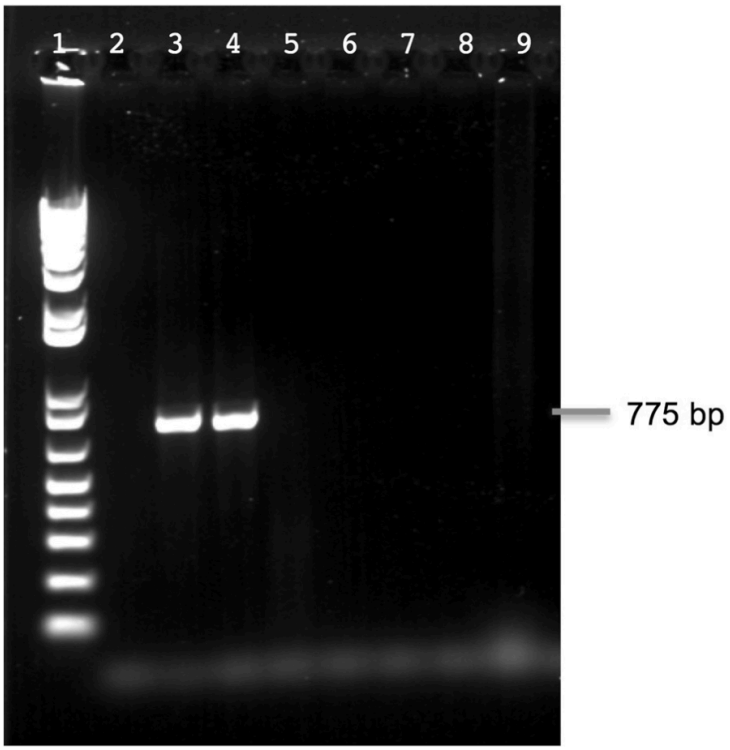


Figure 2.6. Evaluation of laboratory cyanobacterial cultures for the presence of Mor1. The figure shows the results of PCR with selA primers of various laboratory cyanobacterial cultures, run on a 1% agarose gel. Lane 1: Molecular weight marker (Invitrogen 1 kb Plus DNA Ladder). Lane 2: Negative control, sterile water. Lane 3: *M. producens* JHB. Lane 4: *M. producens* 3L. Lane 5: *M. bouillonii*. Lane 6: 3L *Oscillatoria*. Lane 7: PAP25Jun12-2. Lane 8: *Leptolyngbya* sp. Lane 9: *Scytonema hoffmani* "2846 axenic and xenic."

To further explore this hypothesis and to characterize the specificity of the relationship, a set of co-culturing experiments were performed (Table 2.3). The aim of these was to examine whether Mor1 could be transferred to different genera of cyanobacteria by growing them in co-culture with *M. producens* JHB. Initial PCR screening for the *selA* gene in the "acceptor" species verified that Mor1 was absent and thus exclusive to the *M. producens* JHB culture. *M. producens* JHB was then co-cultured in intimate contact with the strains listed in Table 2.3 for 2 weeks. The individual strains were then separated and cultured for a variable period to obtain sufficient biomass for DNA extraction and PCR analysis. Two different PCRs were performed on each of the co-culture samples, as shown in Figure 2.7. The 16S rRNA gene was used as a positive control that verified that each sample had similar amounts of high-quality DNA (Figure 2.7A). Indeed, each sample showed a strong 16S rRNA band of essentially equal intensity.

When the same samples were tested for the presence of the *selA* gene, the *selA* gene signal only appeared in the *M. producens* JHB samples and was absent in all of the "acceptor" cyanobacterial cultures that had been co-cultured with JHB (Figure 2.7B). From these experiments, we conclude that Mor1 was not transferrable to these other strains, and thus constitutes a specific associate of *M. producens*.

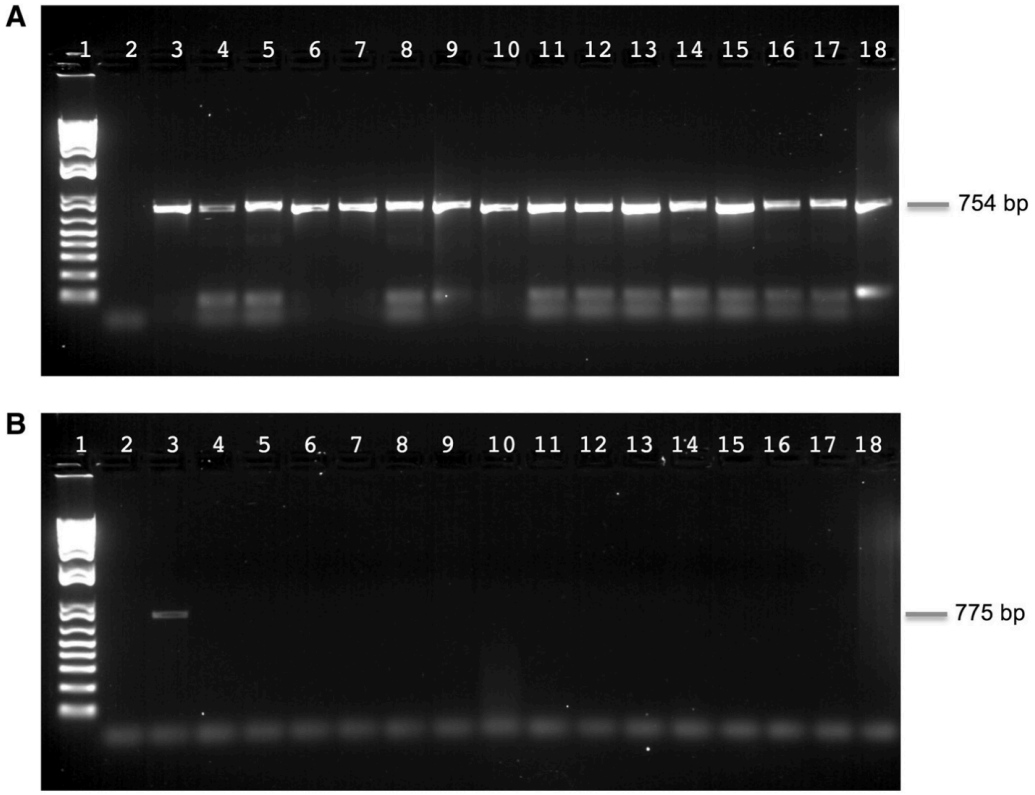


Figure 2.7. Results of co-culturing *M. producens* JHB with various other laboratory cultures to evaluate the transferability of Mor1. Shown are 16S rRNA and selA PCR reactions using DNA from co-culturing of *M. producens* JHB with other cyanobacteria, visualized on 1% agar gels. (A) The 16S rRNA PCR of all samples. (B) The selA PCR of all samples. Both gels were loaded with the same order of samples for each respective PCR reaction. Lane 1: Molecular weight marker (Invitrogen 1 kb Plus DNA Ladder). Lane 2: Negative control, sterile water. Lane 3: *M. producens* JHB. Lane 4: 3L *Oscillatoria*. Lane 5: PAP25Jun12-2. Lane 6: *Leptolyngbya* sp. Lanes 7 and 8: 3L *Oscillatoria* from co-culture with JHB, duplicate co-cultures. Lanes 9 and 10: PAP25Jun12-2 from co-culture with JHB, duplicate co-cultures. 1 Lanes 11 and 12: *Leptolyngbya* sp. from co-culture with JHB, duplicate co-cultures. Lane 13: PAP25Jun12-2 from co-culture with *Leptolyngbya* sp. Lane 14: PAP25Jun12-2 from co-culture with 3L *Oscillatoria*. Lane 15: *Leptolyngbya* sp. from co-culture with PAP25Jun12-2. Lane 16: *Leptolyngbya* sp. from co-culture with 3L *Oscillatoria*. Lane 17: 3L *Oscillatoria* from co-culture with PAP25Jun12-2. Lane18: 3L *Oscillatoria* from co-culture with *Leptolyngbya* sp.

2.4.6. Genome comparison between *M. producens* JHB and Mor1

To explore the potential metabolic interactions between *M. producens* JHB and Mor1, the gene abundance profiles of these two organisms were calculated using the online

Abundance Profile tool from the IMG/ER website (https://img.jgi.doe.gov/cgi-bin/mer/main.cgi). The two bacteria share 939 clusters of orthologous genes (COGs). The number of COGs exclusive to *M. producens* JHB and exclusive to Mor1 are 549 and 495, respectively. All orthologous genes (OG) are clustered and classified by category in Figure 2.8, where substantial differences are highlighted in red between the gene counts and the corresponding cell functions (categories A, G, K, Q, R, X and Z). The categories A and Z represent “RNA processing and modification” and “Cytoskeleton”, respectively, and these categories are highlighted because *M. producens* JHB lacks OG in these categories. Those same genes are missing in other *Moorea* sp. (unpublished) genomes, as well as missing in 90% of the 345 cyanobacteria from JGI/IMG (larger than 1Mb) for category Z and around 84% are missing similar genes in category A. Because these categories of genes appear not to perform essential cell functions in cyanobacteria, they are not considered further in this analysis. Next, categories G, Q and R represent “Carbohydrate transport and metabolism”, “Secondary metabolites biosynthesis, transport and catabolism” and “General function prediction only”, respectively. The number of genes in these categories would be expected to be more numerous in the *M. producens* JHB genome, given it is a larger genome that it also contains many more biosynthetic gene clusters (predicted by antiSMASH to be an astounding 43 biosynthetic gene clusters which account for approximately 22% of the *M. producens* JHB genome).

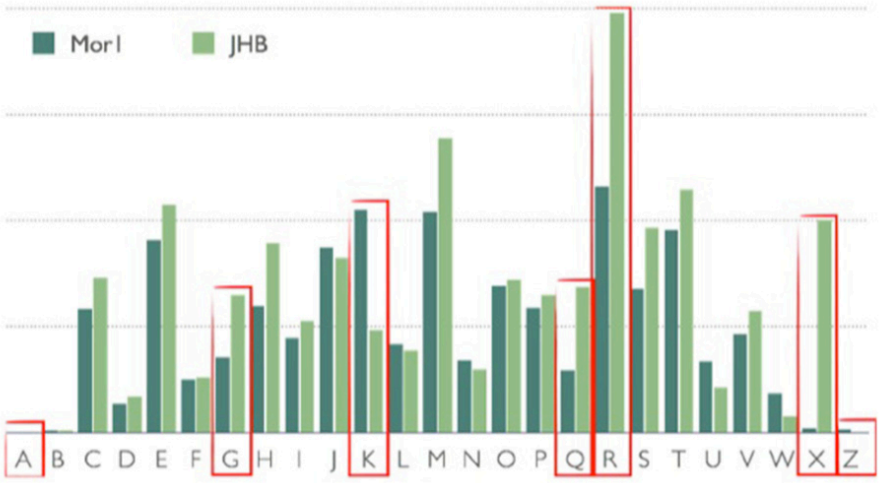


Figure 2.8. Function category comparison of COGs between Mor1 (dark green) and JHB (light green). Categories that are indicated as different are: A = RNA processing and modification, G = Carbohydrate transport and metabolism, K = Transcription, Q = Secondary metabolites biosynthesis, R = General function prediction only, X = Mobilome (prophages and transposons) and Z = Cytoskeleton.

Strikingly, category K, "Transcription", is represented by more than double the OGs in the Mor1 genome versus the JHB genome. This is unexpected as it was previously reported that *M. producens* 3L (aka *L. majuscula* 3L) contains a large number of genes involved in transcription and signal transduction. According to this report, "the numbers and diversity of sigma factors that are global regulators of gene expression in bacteria appear higher in *L. majuscula* 3L than in most other cyanobacteria".[79] Indeed, *Moorea producens* JHB has the same number of annotated sigma factors as 3L, 15 in total. Remarkably, the associated Mor1 genome contains 90 sigma-factor genes (the average number of sigma-factor genes among the acidobacteria genomes currently available at JGI/IMG is 26), 55 of which are annotated as "RNA polymerase sigma-70 factor [a Extra Cytoplasmic Function (ECF) subfamily of factors]. The most important mode of action for ECF sigma in Gram negative bacteria (such as the acidobacteria) is through Cell Surface Signaling.[80] The other proteins involved in cell surface signaling via the sigma-70 factor are the anti-sigma factor TonB-dependent outer membrane receptor and the TonB-ExbB-ExbD system. The only other receptors found that can be associated with iron metabolism in Mor1 are annotated as "Outer membrane receptor proteins,

mostly Fe transport". *M. producens* JHB lacks this ECF Cell Surface Signaling system, thereby suggesting that the iron acquisition and regulation systems are much more sophisticated in Mor1 than *M. producens* JHB. This hypothesis was explored by adding Ferric Ammonium Citrate to the enriched Sigma Sea Salt media in an attempt to culture Mor1; however, this was unsuccessful.

Lastly, category X [Mobilome (prophages and transposons)] has the most notable difference in gene count between the two bacterial species. The *M. producens* JHB genome possesses 199 transposases whereas none are found in the Mor1 genome. It has been hypothesized that intracellular bacteria have a tendency to accumulate transposases in early stages of intracellular symbiosis.[81,82] Therefore, the lack of transposases suggests that Mor1 is not an intracellular symbiont; rather, it appears to be extracellular, which is supported by the previously mentioned decrease in signal of the *selA* gene signal when semi-quantitative PCR was performed (Figure 2.5). Intriguingly, while analyzing the COGs of 26 other acidobacteria genomes available at IMG/JGI, it was observed that only one other acidobacterial genome [JGI GOLD ID: Ga0001215] lacks transposases, indicating that this is an uncommon feature within this phylum. Transposases are important for giving genomes the ability to adapt to evolutionary pressures by facilitating horizontal gene transfer or rearranging of the genome.[83] However, obligate pathogens and endosymbionts have lower numbers of transposases.[82] Hence, the absence of transposases in Mor1 suggests that the potential symbiotic relationship with the *M. producens* JHB strain precludes the need for horizontal gene transfers or rearrangement of the Mor1 genome.[84] Lastly, both *M. producens* JHB and Mor1 harbor a gene from category X known as ParE. This gene is responsible for plasmid stabilization, thus indicating that both organisms may harbor plasmids, even though contigs encoding for plasmids were only found in association with the *M. producens* JHB genome (on the basis of similar GC content). However,

it was not possible to completely assemble any plasmids from the metagenomic data due to the fragmented nature of the assembly.

With regards to primary metabolism, Mor1 is only prototrophic for the biosynthesis of L-alanine, L-aspartate, L-glutamate, L-glycine, and L-glutamine, as well as for common co-factors such as flavin, coenzyme A, NAD, heme and thiamine. The lack of biosynthetic genes for a number of key primary metabolites, including several essential amino acids, suggests that Mor1 is adapted to thrive in a consortium with other bacteria, such as with *M. producens* JHB and its microbiome. Specifically, Mor1 lacks biosynthetic genes for several important amino acids: the aromatic amino acids Phe, Tyr and Trp, the positively charged amino acids Lys, Arg and His, and all non-polar amino acids except glycine and alanine. Complementing this, however, is the occurrence in the Mor1 genome of several transporters that are annotated as "amino acid/polyamine/organocation transporter (APC superfamily)", "amino acid/amide ABC transporter substrate-binding protein (HAAT family)", and "amino acid transporter".

Furthermore, it is not capable of cobalamin or biotin biosynthesis, a metabolic insufficiency clearly revealing its dependency on other organisms for survival. Interestingly, a transporter for the uptake of cobalamin was identified in both Mor1 and *M. producens* JHB, which also lacks the capacity for biotin synthesis; this indicates that other bacteria in the consortium are likely providing this key co-factor. Because Mor1 possesses the genes for the biotin carboxyl carrier protein and biotin ligase, biotin is clearly required, but is presumably acquired through uptake from the environment.

In general, the genus *Moorea* is unable to fix nitrogen,[79] and by genome analysis of *M. producens* JHB and Mor1, neither of these bacteria possess the required nitrogen fixation genes. However, three interesting Orthologous Groups (OGs) were identified in Mor1 that might be aiding in nitrogen metabolism. The first OG consists of a "uncharacterized protein, possibly involved in nitrogen fixation" (COG3197) whereas the latter two are predicted to be "signal

transduction histidine kinases involved in nitrogen fixation and metabolism regulation" (COG5000). Comparison of the nitrogen metabolism KEGG pathway from Mor1 and *M. producens* JHB revealed that they share very few genes (marked in blue, Figure 2.9) and that they appear to assimilate nitrogen from very different sources. Whereas *M. producens* JHB possesses the genetic capacity for the uptake of extracellular nitrate and ensuing assimilatory nitrate reduction to produce ammonia and thereby incorporate nitrogen into its amino acids, Mor1 lacks both the nitrate uptake and assimilatory pathways. Rather, Mor1 possesses genes possibly involved in the uptake of ammonium, suggesting that it may rely on acquiring nitrogen from this source as well as by the uptake and recycling of amino acids.

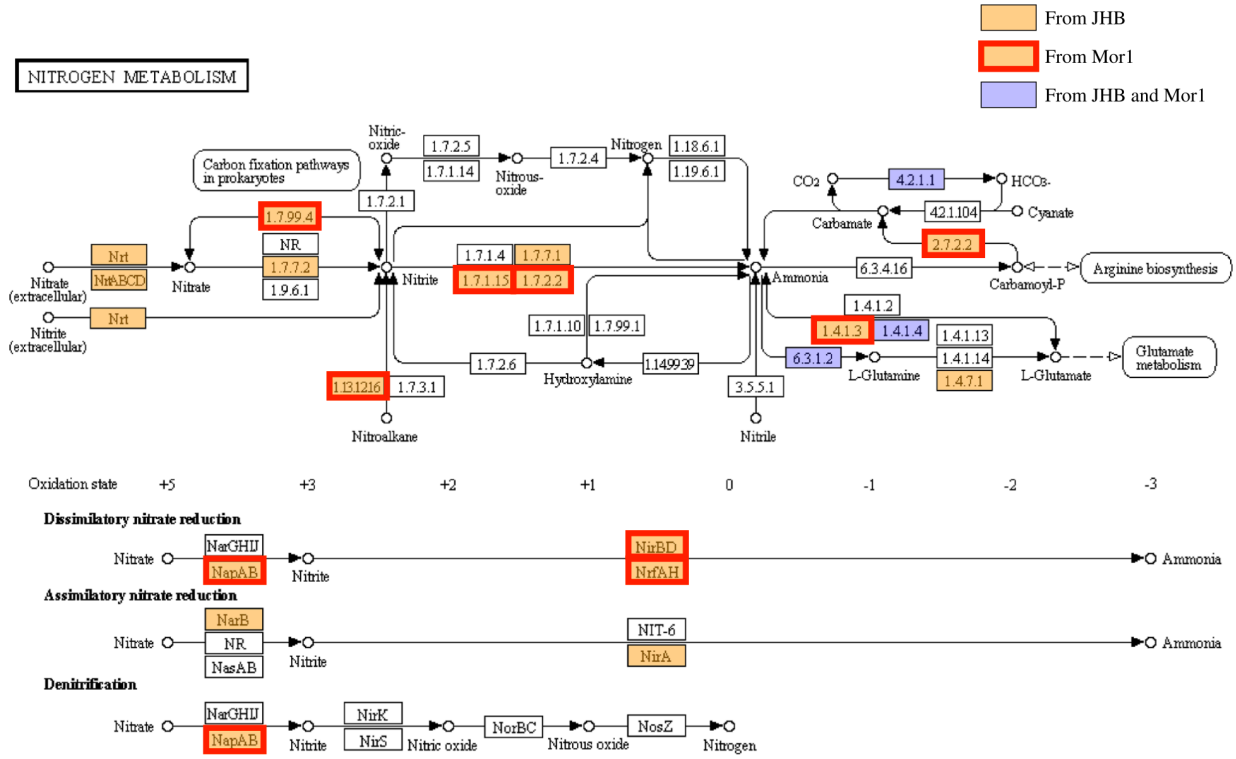


Figure 2.9. Nitrogen KEGG pathway comparison between A. *Moorea producens* JHB and B. Mor1. White boxes represent absent genes, orange represent genes from JHB, Orange box with red line represent genes found in Mor1 and blue boxes represent homologs of the same gene found in both JHB and Mor1 (http://www.genome.jp/kegg/pathway.html)

2.5. Acknowledgement

The work was supported by the NIH grants CA108874 and GM107550. A.K. was supported in part by the Russian Science Foundation (grant 14-50-00069). T.F.L. was funded by a CAPES Foundation Fellowship, nº 13425137, Ministry of Education of Brazil.

Chapter 2, in full, is a reprint, with permission, of the material as it appears in BMC Microbiology, 2016, Susie L. Cummings*, Debby Barbé*, Tiago Ferreira Leao*, Anton Korobeynikov, Niclas Engene, Evgenia Glukhov, William H. Gerwick and Lena Gerwick. The dissertation author is one of the primary investigator (*shared authorship) and author of this material.

CHAPTER 3: Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus *Moorea*

3.1. Abstract

Cyanobacteria are major sources of oxygen, nitrogen and carbon in nature. In addition to the importance of their primary metabolism, some cyanobacteria are prolific producers of unique and bioactive secondary metabolites. Chemical investigations of the cyanobacterial genus *Moorea* have resulted in the isolation of over 190 compounds in the last two decades. However, preliminary genomic analysis has suggested that genome-guided approaches can enable the discovery of novel compounds from even well studied *Moorea* strains, highlighting the importance of obtaining complete genomes. We report the first complete genome of a filamentous tropical marine cyanobacterium, *Moorea producens* PAL, which reveals that about a fifth of its genome is devoted to production of secondary metabolites, an impressive four times the cyanobacterial average. Moreover, possession of the complete PAL genome has allowed improvement to the assembly of three other *Moorea* draft genomes. Comparative genomics revealed that they are remarkably similar to one another, despite their differences in geography, morphology and secondary metabolite profiles. Additionally, *Moorea* species dedicate on average 18% of their genomes to natural products biosynthesis, nearly four times the average for cyanobacteria. Gene cluster networking highlights that this genus is distinctive among cyanobacteria, not only in the number of secondary metabolite pathways, but also in the content of many pathways, which are potentially distinct from all other bacterial gene clusters to date. These findings portend that future genome-guided secondary metabolite discovery and isolation efforts should be highly productive.

3.2. Introduction

Cyanobacteria are carbon-fixing, oxygenic photosynthetic prokaryotes that play essential roles in nearly every biotic environment. Moreover, the development of oxygenic photosynthesis in cyanobacteria was responsible for creating Earth's oxygen rich atmosphere, thereby stimulating evolution of the extraordinary species diversity currently present.[5,85] In the open ocean, nitrogen-fixing (N_2-fixing) cyanobacteria are the major source of biological nitrogen, and this can be a limiting factor to productivity in these oligotrophic environments.[86] Filamentous diazotrophic cyanobacteria from subsection VIII, such as *Nostoc* and *Anabaena*, fix nitrogen within specialized cells called heterocysts.[25]

Apart from their importance in biogeochemical cycles because of their primary metabolism, cyanobacteria are also a prolific source of secondary metabolites known as natural products (NPs). NPs from diverse life forms have been major inspirational sources of therapeutic agents used to treat cancer, infections, inflammation and many other disease states.[7] One genus of cyanobacteria in particular, *Moorea*, has been an exceptionally rich source of novel bioactive NPs.[26] This taxonomic group, previously identified as "marine *Lyngbya*" but recently reclassified on the basis of genetic data as *Moorea*, consists of large, non-diazotrophic filaments that are mostly found growing benthically in shallow tropical marine environments.[61] This genus has already yielded over 190 new NPs in the past two decades, accounting for more than 40% of all reported marine cyanobacterial NPs.[24] The discovery of these NPs was mostly driven by classical isolation approaches, although this has been accelerated by the recent development of Mass Spectrometry (MS)-based molecular networking (groups metabolites according to their MS fragmentation fingerprints, simplifying the search for new NPs or their analogs)[40]. Genomic analyses of these filamentous cyanobacteria have revealed that even well studied strains possess additional genetic capacity to produce novel and chemically unique NPs,[87] and suggests that bottom-up approaches[15] would be productive; a recent example is given by the

discovery and description of the columbamides from *M. bouillonii*.[88] Additionally, and despite the growing interest and importance of genome-guided isolation of NPs as well as the vast biosynthetic potential of these tropical filamentous marine cyanobacteria, not a single complete genome is available in the public databases. Such a complete genome is essential to serve as a reference for other sequencing projects and thereby improve our understanding of their full biosynthetic capacity to produce NPs.

In the present project, we applied a variety of computational and assembly methods to obtain the first complete genome of a tropical filamentous marine cyanobacterium (the genome of *Moorea producens* PAL). This knowledge was applied to three other draft genomes by reference assembly (*Moorea producens* JHB, *Moorea producens* 3L and *Moorea bouillonii* PNG), thereby greatly improving their assemblies as well as the ensuing evaluation of their metabolic and NP-producing capabilities. Comparisons between these genomes demonstrated that these four strains are remarkably similar, despite their differences in geographical site, morphology and NP chemistry. Additionally, the presence in *Moorea* spp. of glycolipid biosynthetic genes associated with heterocyst formation, the site of nitrogen fixation in some filamentous cyanobacteria, suggests that this genus evolved from one that was capable of fixing atmospheric nitrogen. Moreover, we observed that these four *Moorea* strains are metabolically distinct from all previously described cyanobacteria, both in number and content of their natural product pathways, providing support and raising expectations for future genome-guided isolation efforts.

3.3. Methods

3.3.1. Sampling, culturing, microscopy and previous sequencing efforts

PAL was collected at ~ 1 meter from a remote island in the Northern Pacific Ocean, Palmyra Atoll, in August of 2008, as previously described by Taniguchi *et al.*, 2010.[89] Its DNA was

previously extracted (using the protocol below, except by the SDS pre-treatment step) and submitted for MiSeq Illumina® sequencing, using a 300bp paired end library. JHB was collected from Jamaica, at 2 meters depth in Hector Bay, in August of 1996, as described by Marquez *et al.*, 2002.[90] Its DNA was extracted and submitted for HiSeq Illumina® sequencing, using a 100bp paired end library. PNG was collected from Pigeon island at 10 meters depth in Papua New Guinea in May 2005, as described by Grindberg *et al.*, 2011.[91] It was extracted and sequenced similarly to JHB. Last, 3L was collected near Carmabi station at ~2 meters depth from Curaçao in December of 1993 and it was extracted and sequenced by Sanger® technology, as described by Jones *et al.*, 2011.[79] All four genomes were highly fragmented and due to those circumstances, only the 3L genome was published (access number GCA_000211815.1). 3L is the only *Moorea* genome published to date. All of these strains were established as uni-cyanobacterial cultures using standard microbiological isolation methods (plating, dilution and microscopy analysis) and they have been maintained as live cultures in salt water BG-11 media[92] since their isolation. Experiments with nitrogen depleted media used the same BG-11 formulation, except that ferric ammonium citrate ferric was substituted for ferric citrate and $NaNO_3$ was substitute for NaCl, both in equimolar concentrations. Filaments from nitrogen depleted cultures were subjected to 100 μl of Alcian blue staining (1% in 3% Acetic Acid) for 2 seconds, followed by washing with sterile distilled water and then observed under light microscopy using an Olympus IX51 epifluorescent microscope and Olympus U-CMAD3 camera. Control filaments were grown in regular salt water BG-11 media, and also stained using Alcian blue.

3.3.2. DNA extraction, PacBio sequencing of *M. producens* PAL and *de novo* assembly

In order to reduce the amount of heterotrophic contaminant bacteria, a pre-treatment was performed with SDS 10%, followed by BG-11 rinsing for SDS removal. DNA extraction was performed using a “QIAGEN Bacterial Genomic DNA Extraction Kit” optimized for cyanobacteria

by carefully grinding filaments using mortar and pestle under liquid nitrogen before extracting using the standard kit protocol. The quality of the genomic DNA (gDNA) was evaluated by Nanodrop, 1% agarose gel electrophoresis and Genomic DNA Screen Tape® analysis. Post-quality control, the gDNA was sequenced using a PacBio RS II plataform (Pacific Biosciences®) at the Institute of Genomic and Medicine, UC San Diego, using a 10 kb fragment library and two Smart Cells to obtain high coverage. Both short and long reads were assembled together into contigs using SPAdes version 3.5, using default settings with automatic coverage cutoff. Scaffolds were generated by using SSPACE-LongReads, with default settings, and gaps were closed using long reads by Geneious® 8.1, with a minimum read coverage of 98 fold. The binning pipeline was adapted from Albertsen *et al.,*[93] using coverage versus GC content to produce bins, indicating that all scaffolded contigs most likely belong to the same taxon of cyanobacteria. In this pipeline described by Albertsen *et al.*, the minimum length threshold for the contigs was 500bp and other parameters evaluated were the phylogenetic designations of 107 single copy genes conserved in bacteria and tetranucleotide fingerprint.

3.3.3. Reference assembly of draft genomes

Moorea producens JHB and *Moorea bouillonii* PNG reads were both assembled by SPAdes 3.5, using default settings with automatic coverage cutoff, generating 2,435 and 908 contigs, respectively. These contigs were scaffolded using SSPACE-ShortReads with a minimal link number of 20 reads, in order to ensure more accurate scaffolding. The complete genome of *M. producens* PAL was used as a reference template to order the contigs of the other strains into scaffolds with the software CONTIGuator. The scaffolds were submitted to the tool GapFiller, in order to reduce the number of Ns and gaps. Scaffolds not included in the reference assembly step (because of low nucleotide similarity to the reference template) were binned and named either unmapped scaffolds or plasmid scaffolds, depending on gene content annotations and best BLAST matches to NCBI/NR database. The taxonomic origins of binned contigs were verified

using DarkHorse software version 1.5[70] to determine closest phylogenetic matches in Genbank nr for each individual gene in the contigs submitted. *Moorea producens* 3L was previously assembled and binned by Jones and Monroe *et al.* (2011) using similar methods, therefore, we proceeded straight to reference assembly. The genomes of PAL, JHB, 3L and PNG have been deposited at DDBJ/ENA/GenBank under the accessions GCA_001767235.1, GCA_000211815.1, MKZR00000000, MKZS00000000, respectively.

3.3.4. Comparative genomics

The complete reference genome and the three draft genomes were submitted to the JGI/IMG (Joint Genome Institute, Integrated Microbial Genomes database) expert review pipeline for annotation. Comparative genomics and statistics analyses were obtained using the Genome StatisticsPairwise average nucleotide identity, Genome Gene Best Homologs (Figure 3.2), COG Homology and Abundance Profile tools from the JGI/IMG webserver. The cutoff between two genes to be considered homologs was 50% amino acid identity. Synteny plots were generated by MUMmer 3.0,[94] with a maximum gap of 500 bp and minimum cluster length of 100 bp between *Moorea* genomes. Circular maps were generated by BRIG software,[95] using *Moorea producens* PAL as a reference chromosome. MultiGeneBLAST[96] was used to compare *hgl* core genes and vicinities among the four *Moorea* genomes. Traditional BLASTN and BLASTP searches versus the NCBI/NR database were performed to investigate some specific genes (example *hetR*, *ntcA*, *patS* and so on).

3.3.5. Phylogenomics

Phylogenomic analysis was performed using 29 different conserved genes reviewed in Calteau *et al.*,[97] from all finished cyanobacterial genomes at JGI up to July 13th, 2016 (so as to increase tree resolution) plus our four *Moorea* genomes, totalizing 107 genomes. The genes were aligned by MUSCLE and the tree was built by the program Geneious Tree Builder, using

the model Jukes-Cantor genetic building model, and the Neighbor-Joining method, with 1000 bootstrap repetitions and one out group genome (*Chloroflexus auranticus* J-10). The tree image was edited in FigTree 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/), with branches colored according cyanobacterial subsections as previously reported in.[25]

3.3.6. Biosynthetic gene clusters

Biosynthetic gene clusters (BGCs) were identified using antiSMASH 3.0,[98] standard configurations, ClusterFinder option off. All BGCs from *Moorea* strains were manually verified, in order to reduce false positives. All 103 currently available finished cyanobacterial genomes were also submitted to antiSMASH, but not manually verified. The number of BGCs per genome was generated using a python script to parse antiSMASH results and to incorporate counts into Figure 3.3B (gene clusters containing only RiPP precursor peptides were excluded from the count). Of particular note, despite the use of antiSMASH to identify the BGCs in these cyanobacterial genomes (including our strains), we used the statistics for Figure 3.4 (plot of genomic percentage devoted to BGCs in bacteria) directly from JGI-IMG database, using all bacterial genomes in the July 2016 version of the database. We acknowledge that antiSMASH represents a more accurate tool for gene cluster identification than ClusterFinder,(30) the standard tool used by JGI/IMG, however, submitting all bacterial genomes to antiSMASH was not feasible to our analyses. Nevertheless, as observed in *Moorea producens* PAL, the use of antiSMASH, manual inspection and border refinement of PAL's BGCs did not significantly differ from JGI/IMG statistics.

3.3.7. Gene cluster networking

For the dereplication and initial analysis of biosynthetic gene clusters in *Moorea* spp. genomes, we developed a custom pipeline called Comparative Synteny Software for Biosynthetic Gene Clusters (BioCompass). Source code and detailed usage instructions are

provided at http://biocompass.net/ (along with the instructions to reproduce this exact analysis). Analogously to Doroghazi *et al.*,[10] we grouped the BGCs into gene cluster families based on synteny and homology. A similarity matrix was used to divide each given BGC into subclusters based on synteny at best MultiGeneBLAST hits (obtained using antiSMASH 3.0) and the functional annotation of each gene in the queried cluster. This information was then incorporated into a query-specific database to search for the best matches for each subcluster. The newly created database included microbial BGCs identified by antiSMASH (downloaded from NCBI database, Genbank NR August 2016), the MiBIG repository (version 1.2), and BGCs from all finished JGI-IMG cyanobacterial genomes (same included in phylogenomic analysis). Final similarity scores were calculated via MultiGeneBLAST for each subcluster, and the output was displayed as a network diagram using Cytoscape v3.2.1 (as shown in Figure 3.5). The cutoff parameters used were 25% of matching genes per subcluster, with matching genes defined as having a minimum of 50% amino acid identity, 80% alignment coverage, a cumulative BLAST bit score of 1000, and a MultiGeneBLAST score of 5.

3.4. Results and discussion

3.4.1. Geographical, morphological and chemical features of four filamentous marine cyanobacteria

The present study analyzed and compared four strains of tropical filamentous marine cyanobacteria of the genus *Moorea* (Figure 3.1): *M. producens* PAL 15AUG08-1, *M. producens* JHB 22AUG96-1, *M. producens* NAK12DEC93-3La and *M. bouillonii* PNG 19MAY05-8 (abbreviated as PAL, JHB, 3L and PNG, respectively). All of these strains were laboratory cultured in saltwater BG-11 media since the time of their original collection. PAL was collected from a remote island in the Northern Pacific Ocean, Palmyra Atoll, in August of 2008, and it produces the NPs palmyramide A and curacin D. PNG was collected from Papua New Guinea

in May 2005 and it produces columbamide A-C, apratoxins A-C and lyngbyabellin A. These two Pacific Ocean strains have similar morphologies comprised of discoid cells that are arranged into large isopolar filaments, present as trichomes covered by thick mucilaginous sheaths.[61] The exterior of the sheath material is richly populated with various heterotrophic bacteria, some of which may exist in obligate commensal relationships.[3] However, *M. bouillonii* PNG has a lighter coloration and thinner filaments (around 20-40 μm instead of 80-100 μm in PAL). The other two strains described here, JHB and 3L, are from the Caribbean Sea and hence constitute Atlantic species. JHB was collected from Hector's Bay, Jamaica, in August 1996 and it produces hectoramide, hectochlorin A-D and jamaicamide A-F. The 3L strain was collected from Curaçao in December of 1993 and it produces barbamide, dechlorobarbamide, carmabins A and B, curacins A–C and curazole. These two Atlantic strain have a similar morphology to PAL, with the exception that 3L has an overall red coloration caused by larger relative proportion of the pigment phycoerythrin. As recently reviewed by Kleigrewe *et al.* (2016)[24], the compounds cited above are produced via enzymes encoded by unique biosynthetic genes that are almost exclusive to filamentous marine tropical cyanobacteria. Moreover, it is interesting to observe that some of the unique structural features in *Moorea* NPs (e.g. terminal olefins, *t*-butyl groups, *gem*-dichloro groups) are shared among different cyanobacterial metabolites that have different structural backbones. This suggests the likelihood of combinatorial repurposing of these genetic elements during the evolution of their pathways. Given the divergent geographical locations of their collection, differences in morphology, and variations in NP chemistry, a comparative genomic study of these four strains was undertaken.

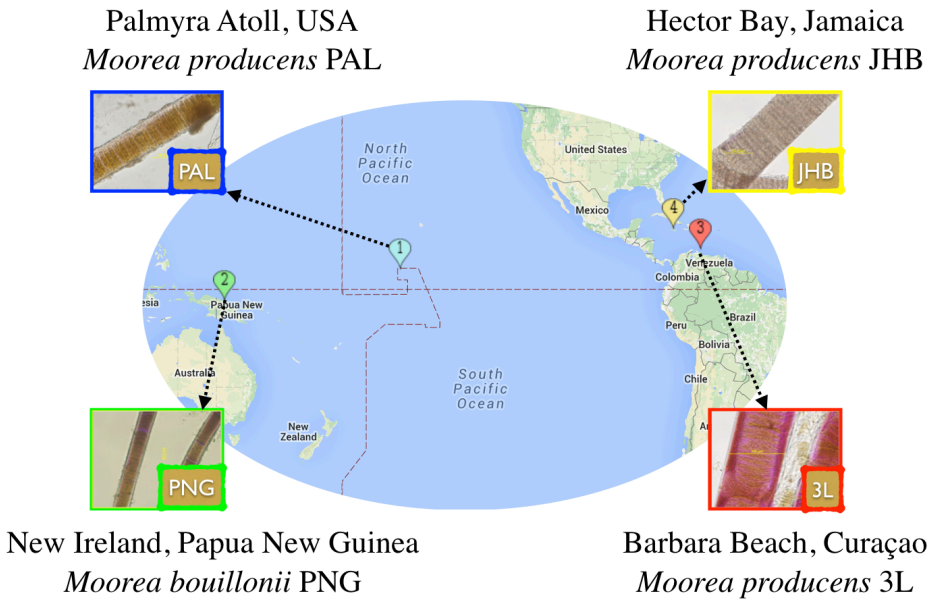


Figure 3.1. Geographical location and microscopy images of the four investigated *Moorea* strains.

3.4.2. The Use of Hybrid Assembly and Long Reads Scaffolding to Obtain the First Complete Genome of Tropical Filamentous Cyanobacterium

The genus *Moorea* currently lacks a reliable reference genome. This would be invaluable for the relative placement of fragmented genomic data from sequencing projects of other *Moorea* strains. Therefore, to obtain a high quality genome sequence, two different methods were used, Illumina MiSeq and PacBio, using DNA from a non-axenic laboratory culture of *Moorea producens* PAL. Both the short and long reads were assembled together (described as 'hybrid assembly') using standard settings of SPAdes 3.5,[99] and yielded 47 linear contigs larger than 500 bp along with one circular contig of 35.5 kb (a candidate cyanobacterial plasmid). Hybrid assembly has previously been employed to improve overall draft genome quality, however, in this case it was still fragmented because large repeated regions remained unresolved.[100] In order to resolve these regions and close the genome, we developed a new approach that involved trimming the repetitive edges from the assembled contigs (which tend to have assembly mistakes) and then submitting these trimmed contigs to SSPACE-LongReads

scaffolding with the standard settings.[101] Fourteen of the contigs assembled into a single circular scaffold of 9.67 Mb, and gaps were closed, again using the long reads. The minimum coverage was 98-fold, and together with the 35.5 kb circular plasmid, it constitutes the complete *M. producens* PAL genome, the first complete genome of a tropical filamentous marine cyanobacterium (Table A3.1). To assure that no cyanobacterial contigs were left out of the assembly, especially in light of the fact that the sequenced culture was non-axenic, we performed a binning procedure using multiple features (GC content, coverage, phylogenetic identification of conserved genes, tetranucleotide fingerprint). This analysis confirmed that all 15 contigs (14 comprising the circular chromosome and one for the circular plasmid) from the PAL genome were the only cyanobacterial contigs in the sample (confirming that the culture was mono-cyanobacterial). Moreover, the binning procedure identified a fully assembled large contig of 3.63 Mb that represents a draft genome of a *Hyphomonas* sp. strain "Mor2" (Genbank: CP017718), an uncultured alpha-proteobacteria associated with *M. producens* PAL.

Possession of this reference genome for *M. producens* PAL enabled a substantial improvement in the assemblies of several other *Moorea* genomes via standard referencing procedures.[102,103] In the case of *M. producens* JHB, this reference assembly procedure resulted in a linear chromosomal scaffold of 9.6 Mb consisting of 205 contigs with approximately 26,000 Ns that connect the contigs, along with two small plasmid scaffolds of 9.5 kb and 2 kb. The final draft genome of *M. bouillonii* PNG consisted of a linear chromosomal scaffold of 8.23 Mb (291 contigs and approximately 32,000 Ns) and 12 unmapped scaffolds from 1.6 kb to 16.7 kb. The *M. producens* 3L final draft genome consisted of a linear chromosomal scaffold of 8.15 Mb (205 contigs and approximately 20,000 Ns) and 78 unmapped scaffolds from 0.5 to 9.4 kb. Additional features of these four genomes are presented in Table A3.1.

The completeness of the four genomes was estimated by the presence and absence of ubiquitous cyanobacterial housekeeping genes (e.g. present in single copy in nearly all finished

cyanobacterial genomes from JGI/IMG database (total of 107 genomes). Our new reference genome, *M. producens* PAL, contained all 195 housekeeping genes, reinforcing its completeness. The other three draft genomes were compared to the same 195 single copy gene dataset, and revealed that the assemblies of 3L, PNG and JHB contained 98.97%, 98.46% and 99.49% of these genes, respectively. These percentages are close to the reference genome and thus indicative of their relative completeness and the excellent quality of their assembly. Other parameters from Table A3.1 such as GC content, number of genes, and percentage of annotated genes, are consistent with other cyanobacterial genomes.[97]

3.4.3. Genome Comparison Among *Moorea* Strains Reveals Significant Synteny

Given the wide geographical range from which the four *Moorea* strains were obtained, spanning some 16,000 kilometers and existing in two distinct oceans, one could expect that they might show considerable sequence divergence. However, a precedent set from the genus *Salinispora* indicates that genomic conservation is in some cases observed for geographically divergent species.[104] The four genomes investigated here were found to be remarkably similar with a very high average nucleotide identity (minimum of 94.6%), consistent with previously reported 16S rRNA gene identities of more than 99%.[61] This is visualized as a circular map that compares the reference and draft genomes (Figure A3.1), and bar graphs that depict the number and the percent identities between homologous genes in the different genomes (Figure 3.2A). In Figure A3.1, the high nucleotide identity between the *Moorea* genomes indicates that the reference assembly approach was a good solution for improving the quality of these three draft genomes. This high nucleotide identity translates to a high amino acid similarity, confirming their close evolutionary relationship (Figure 3.2A). It is remarkable that *M. producens* PAL has higher similarity to *M. bouillonii* PNG than to other *M. producens* strains (also observed in the phylogenomic tree, Figure 3.3), suggesting that it may require reclassification at the

species level. These phylogenetic relationships may reflect the degree of separation between Pacific (PAL and PNG) and Atlantic strains (3L and JHB), however, a larger genome dataset will be required to substantiate this hypothesis. Lastly, the MUMmer plots in Figure A3.2 indicate that these *Moorea* genomes are also highly syntenic with one other (similar genomic regions are present in the same order), yet are very distinct from the genome of *Microcoleus* sp. PCC 7113, the closest sequenced relative to *Moorea*.

These four *Moorea* genomes share 5944 homologous genes as identified by BLAST analysis (Figure 3.2B). Therefore, only 8 to 13.5% of the total genes per genome are strain-specific. Unfortunately, the great majority of the strain-specific genes lack detailed annotation (e.g. hypothetical proteins). On average, the largest number of annotated orthologous genes (OG) belong to categories “R: General function prediction only” (13%), “M: Cell wall biogenesis” (9%), “T: Signal transduction mechanisms” (7%), “E: Amino acid transport and metabolism” (7%) and “X: Mobilome” (7%). As expected by the high synteny and average nucleotide identity, the gene counts in most COG categories of all four genomes is remarkably similar (Table A3.2). Moreover, most of these categories possess a very similar OG content among the strains, represented by the normalized D-rank. When the D-rank is close to zero, the genes in the category have higher similarity to the homologues in the reference genome. In the categories related to primary metabolism, all four strains are nearly identical. All are annotated as photosynthetic (atmospheric carbon dioxide as primary carbon source), non-diazotrophic (absence of nitrogenase genes), capable of the biosynthesis of all proteinogenic amino acids (except for tyrosine and phenylalanine), and possessing the biosynthetic genes for important co-factors including coenzyme A, cobalamin, biotin, flavin, NAD, heme and thiamine. Additionally, the number of specialized sigma factors in the genomes of these four filamentous marine cyanobacteria strains, as previously discussed in Jones *et al.*[79] are virtually the same (5

specialized sigma factors per genome). Despite the significant similarity between the four genomes, some COG categories were indicative of a number of subtle genetic differences.

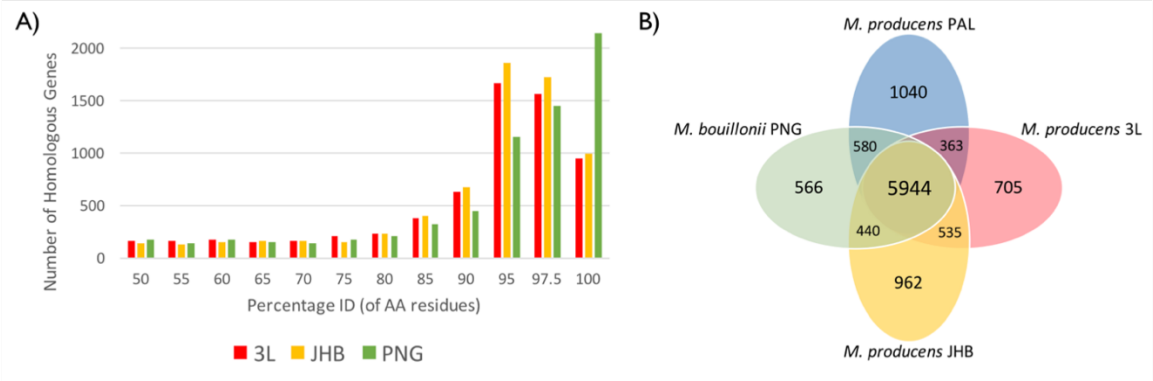


Figure 3.2. A) A histogram of percent amino-acid identity for all shared homologous genes (bidirectional best BLAST hit, minimum ID of 50%). B) Venn diagram for the shared homologous genes and strain-specific genes among the four *Moorea* strains.

3.4.4. The Evolved Loss of Nitrogen Fixation in the Genus *Moorea*?

The gene cluster for heterocyst envelope glycolipid biosynthesis (*hgl*) has been identified and characterized in the filamentous diazotrophic cyanobacteria *Anabaena* sp. PCC 7120 and *Nostoc punctiforme* ATCC 29133.[105,106] These genes are commonly found in diazotrophic cyanobacteria from subsections VIII, but lacking in the other subsections. BLAST analysis of 267 cyanobacterial genomes from JGI/IMG confirmed the absence of these four core genes in subsections I-VII. As expected, *M. producens* 3L, a filamentous non-heterocyst forming cyanobacterium from subsection VII, does not possess the *hgl* cluster. Surprisingly, the other three *Moorea* genomes described herein (PAL, PNG and JHB) contain the complete *hgl* cluster. As depicted in Figure A3.3, it appears that *M. producens* 3L recently lost the *hgl* cluster. Homologs of the genes upstream and downstream of the *hgl* cluster in PNG, JHB and PAL are adjacent to another in the 3L genome (red box in Figure A3.3). Two new genes at this position that encode for hypothetical proteins have apparently replaced the *hgl* cluster in the 3L genome. Despite the presence of the *hgl* cluster, filaments cultured in nitrogen deficient medium (up to eight days at which time the cells start to rapidly die) did not develop heterocysts nor did they

visibly produce heterocyst glycolipids (e.g. they were not reactive to Alcian blue staining, a dye used for acidic polysaccharides such as heterocyst glycolipids).[84] The only regulatory homolog for heterocyst development located in *Moorea* was *hetR* (approximately 70% nucleotide ID, located about 1.7 to 2.2 Mb apart from the *hgl* cluster); the *ntcA* and *patS* genes were absent. An additional four predicted regulatory elements in the immediate vicinity of the *hgl* core (Figure A3.3) suggest that its regulation may be different and perhaps more complex than previously reported in *Nostocales*. Future transcriptomic experiments may provide insights into the regulation of this cluster.

This is the first report of a cyanobacterium from outside subsection VIII that possesses the *hgl* cluster. To the best of our knowledge, the only other cyanobacterium capable of forming heterocyst glycolipids and not fixing nitrogen (the *nif* cluster is absent) is *Raphidiopsis brookii* D9 (*Nostocales*, subsection VIII).[84] Here we propose an analogous situation where the retention of the *hgl* cluster (except by 3L) and a selective loss of the *nif* cluster has occurred. However, because there are no close relatives of *Moorea* that possess *nif* genes, we are unable to draw specific conclusions regarding the position or timing of this loss. Interestingly, several un-clustered genes are present in these four genomes with predicted functions as “global nitrogen regulator”, “nitrogen fixation proteins of unknown function” and “nitrogen regulatory protein P-II 1”; nonetheless, these genes have also been reported in non-heterocyst forming and non-diazotrophic cyanobacteria.[107] The fact that *Moorea* strains survive up to eight days under nitrogen deprivation can likely be attributed to the presence of cyanophycin, a multi-L-arginyl-poly-L-aspartate nitrogen storage reserve material typical of cyanobacteria.[108] Of note, our genomic analysis revealed that each of the *Moorea* genomes contained one cyanophycin synthetase and at least one cyanophycinase gene.

3.4.5. Uncovering the Metabolic Potential of the Genus *Moorea*

A phylogenomic analysis (Figure 3.3A) confirmed that these four *Moorea* strains are monophyletic, supporting the findings of high genomic synteny. However, based on phylogeny (Figure 3.3A) and the occurrence of the *hgl* cluster, this genus may be misplaced within section VII of the cyanobacteria. Another highly prominent feature that distinguishes *Moorea* from other cyanobacteria (Figure 3.3B) is the large number of biosynthetic gene clusters (BGCs). The average number of BGCs in this clade is dramatically larger than any other radiation of cyanobacteria. While *Moorea* harbors an average of 38 per genome, some of the closest relatives (e.g. *Microcoleus* sp. PCC 7113, *Dactylococcopsis. salina* PCC 8305, *Gleocapsa sp*. PCC 7428) contain less than half this number. As such, *Moorea* spp. are "superproducers" among cyanobacteria, and on average 18% of their genome is dedicated to secondary metabolism (see Table A3.1), nearly four times the average of other cyanobacteria.[85] In comparison to all other bacterial genomes (Figure 3.4), *Moorea* are among the most prolific producers of NPs with only some actinobacterial strains being more endowed.[109] The discrepancy between our analyses and that performed previously by Jones *et al.*[79] on the draft genome of *M. producens* 3L is due to the fact that the BGC-mining tool antiSMASH[98] was not yet available in the earlier analysis. In the previous study, BGCs in the 3L genome were identified primarily by BLAST searching for NRPS and PKS genes, and this resulted in an underestimation of the resident biosynthetic pathways.

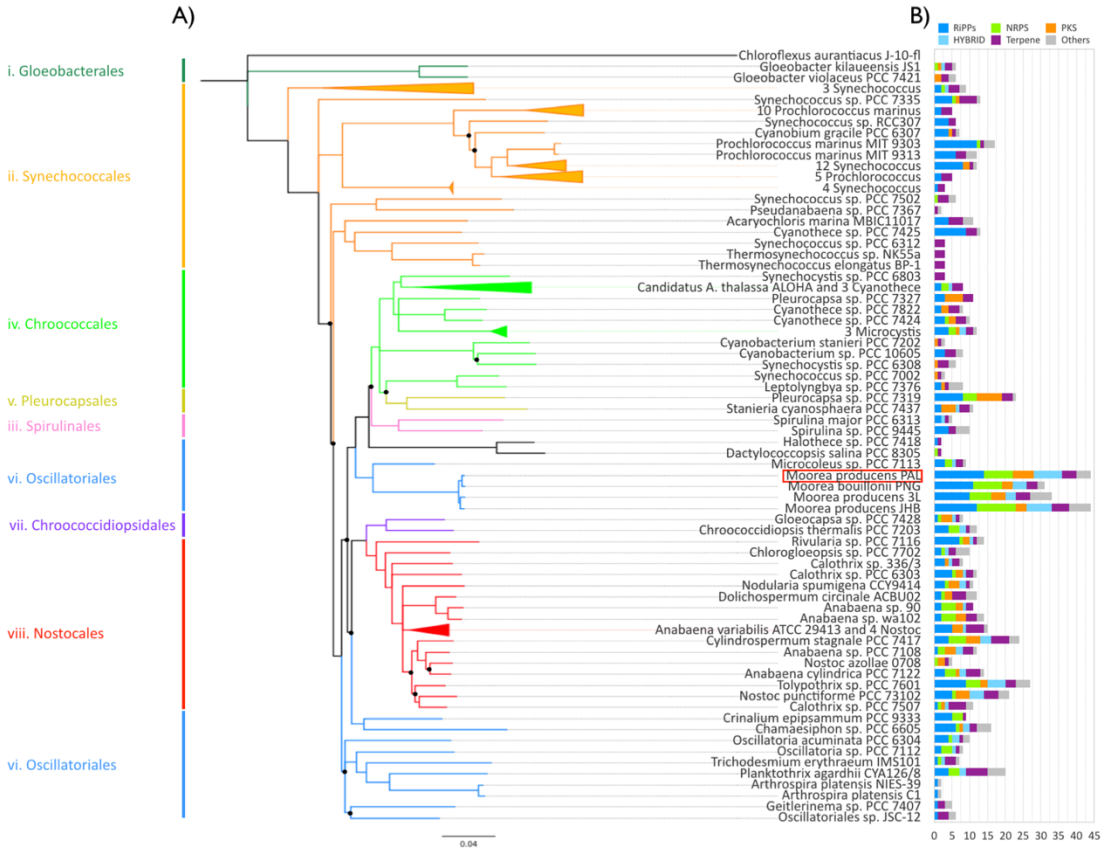


Figure 3.3. A) Phylogenomic analyzes of completed cyanobacterial genomes using 29 conserved genes from Calteau and collaborators.[97] Branches are colored according to cyanobacterial subsections (except by PCC 7418 and PCC 8305, which are not yet classified). All bootstrap values are higher than 85, except those marked by a circle (minimum bootstrap value is 52). B) The number of biosynthetic gene clusters as deduced by antiSMASH analysis and colored by antiSMASH NP categories. For branches with more than one genome (triangular tips), the number of BGCs correspond to the most prolific genome.

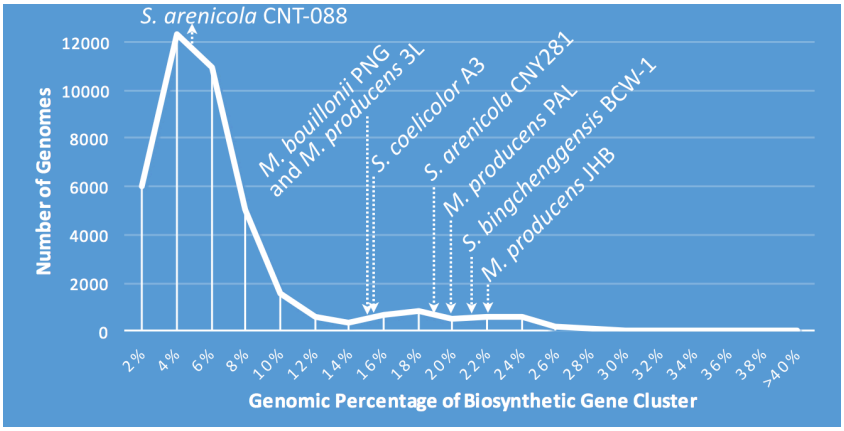


Figure 3.4. Distribution of bacterial genomes from JGI/IMG database in terms of genomic percentage dedicated to secondary metabolism (natural products biosynthesis). Several prolific NP producers are underlined in the figure, including *Streptomyces coelicolor* A3, *Streptomyces bingchenggensis* BCW-1 and two *Salinispora* strains (highest and lowest genomic percentages from this genus). The total number of genomes interrogated was 40,532.

In order to investigate the novelty of these numerous *Moorea* BGCs, we decided to group these BGCs into families according to sequence homology at the gene level. This 'gene cluster networking' procedure has been applied to explore the biosynthetic capacity of 830 actinobacterial genomes.[10] Since the code to the aforementioned networking approach is not publicly available, we adapted our own strategy for the discovery of gene cluster families (as described in Supporting Information). We refer to this workflow as BioCompass, found at http://biocompass.net/. The output can be displayed as a network diagram using Cytoscape v3.2.1 (Figure 3.5). BioCompass predictions were verified to match well known previously characterized pathways. For uncharacterized pathways, all BioCompass predictions were manually examined to confirm consistency between the multigene alignments within members of the same family. Nodes in the network signify gene clusters whereas edges represent shared subclusters or subunits of the gene cluster. Subclusters indicate groups of adjacent and/or non-adjacent genes that share synteny and predicted function. Self-loops represent unique subclusters (not shared with any other pathway).

As depicted by the gene cluster network (Figure 3.5 and Table 3.1), the great majority of gene clusters from PAL (40 out of 44 clusters, around 91%) match only cryptic gene clusters in other organisms (gene clusters not assigned to known NPs), suggesting that they likely encode the biosynthesis of novel NPs. Interestingly, 26 of the PAL clusters (about 59%, Figure 3.5C) only have homology to other *Moorea* pathways, confirming previous chemical investigations that indicated they possess a unique secondary metabolite profile compared to other bacteria.[24] Moreover, these findings suggest that *M. producens* PAL is not only a source of novel NPs, but that these NPs will likely be comprised of new chemical backbones. Finally, given the level of synteny between *Moorea* genomes, it is intriguing to observe a significant number of orphan gene clusters (gene clusters only found in PAL, a total of 7 clusters, approximately 16%)(Figure 3.5A).

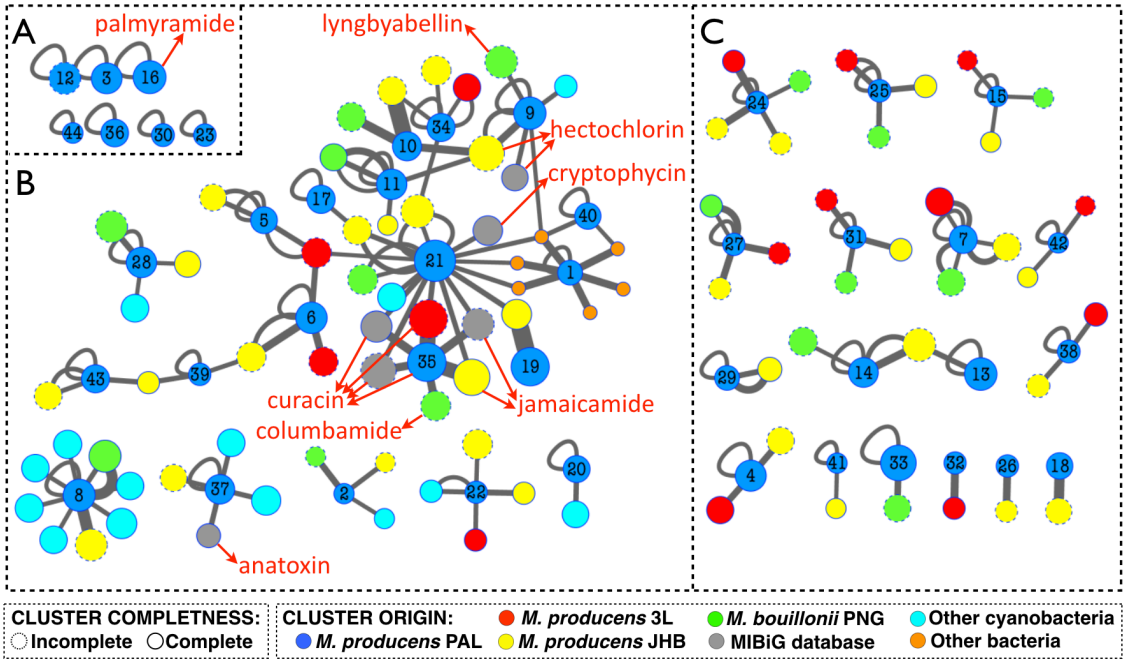


Figure 3.5. Gene cluster networking of PAL versus gene clusters from PNG, 3L, JHB, the MiBIG database, completed cyanobacterial genomes from JGI-IMG and their closest homologs from the NCBI database (according to antiSMASH results). A) represents only orphan gene clusters from the PAL genome. B) contains known and cryptic gene clusters from cyanobacteria and C) contains only *Moorea*-specific cryptic gene clusters. Nodes represent clusters and edges represent sub-clusters. Node size is proportional to gene cluster size. Incomplete gene clusters are sequences that contain undefined nucleotides and therefore require further validation. Known gene clusters are named in red.

Table 3.1. Summary table listing number of known (K), "cryptic" (C), and "orphan" (O) NP pathways.

	PKS			NRPS			PKS-NRPS			RiPP			Terpene			Others			Sum per strain		
Annotation	K	C*	O†	K	C	O	K	C	O	K	C	O	K	C	O	K	C	O	K	C	O
PAL	—	5	1	—	7	1	2‡	5	1	—	10	4	—	4	—	—	4	—	2	35	7
JHB	—	3	—	—	11	—	2§	5	—	—	12	—	—	5	—	—	6	—	2	42	—
PNG	—	3	—	—	8	—	3¶	1	—	—	10	1	—	3	—	—	2	—	3	27	1
3L	—	4	—	2#	4	—	1‖	2	—	—	9	1	—	4	—	—	6	—	3	29	1
Subtotal	—	15	1	2	30	1	8	13	1	—	41	6	—	16	—	—	18	—	10	106	9
Total		16			33			22			47			16			18			152	

Pathways are divided by biosynthetic category. Zeroes were replaced with dashes to improve data visualization. NRPS, nonribosomal peptide synthetase; PKS, polyketide synthase; RiPP, ribosomally synthesized and posttranslationally modified peptides.
*Cryptic: A gene cluster not assigned to any known NP.
†Orphan: A cryptic gene cluster only found in one strain (no matches to any sequence in the NCBI database).
‡Palmyramide and curacin.
§Hectochlorin and jamaicamide.
¶Lyngbyabellin, columbamide, and apratoxin.
#Carmabin and barbamide.
‖Curacin.

As previously reported, accurate prediction of BGC borders is a common challenge for the field.[109,110] This issue can have an effect on the estimated percentage of the genome dedicated to natural product biosynthesis. However, the homology alignment feature of BioCompass allowed us to refine the BGC borders by removing unshared genes of unknown

function, excluding from the analysis predicted proteins most likely representing genes adjacent rather than integral to BGCs. This more conservative approach to estimating cluster sizes had only a small effect on the percentage of the *M. producens* PAL genome allocated to secondary metabolism, reducing it from 19.89% (JGI) to 18.02%, confirming the validity of the relationships shown in Figure 3.4. Further analyses of various features of *Moorea*'s BGCs, such as G+C content, low of co-localization with genomic islands, and encoding of relatively rare structural moieties[24] suggests that these strains have vertically acquired these biosynthetic pathways, consistent with previous reports for cyanobacteria.[97] However, a larger sample size and better-characterized pathway products are needed to fully understand the evolution and distribution of *Moorea*'s NP pathways.

3.5. Acknowledgements

This research was supported by National Institute of Health grants CA108874 and GM107550, to W.H.G. and L.G. Also supported by the Russian Science Foundation, grant 14-50-00069 to A.K. We thank CAPES Foundation for research fellowship to TFL (13425-13-7).

Chapter 3, in full, is a reprint, with permission, of the material as it appears in Proc. Natl. Acad. Sci., 2017, Tiago Leao, Guilherme Castelão, Anton Korobeynikov, Emily A. Monroe, Sheila Podell, Evgenia Glukhov, Eric E. Allen, William H. Gerwick and Lena Gerwick. The dissertation author is the primary investigator and author of this material.

3.6. Appendix

Table A3.1. Genomic features of *Moorea* spp. genomes (one complete and three drafts). Statistics obtained from JGI-IMG annotation, unless marked with * for statistic from IslandViewer3 and ** for antiSMASH. GI = Genomic Islands and BGCs = Biosynthetic Gene Clusters.

	M. producens PAL	*M. producens* JHB	*M. bouillonii* PNG	*M. producens* 3L
Main scaffold size (chromosome)	9.67 Mb	9.35 Mb	8.23 Mb	8.15 Mb
Total genome size	9.71 Mb	9.38 Mb	8.32 Mb	8.37 Mb
Contigs that constitute the main scaffold (chromosome)	1	205	291	204
Unmapped scaffolds	0	0	12 (0.09Mb)	78 (0.19 Mb)
Plasmid scaffolds	1 (circular)	2 (linear contigs)	0	0
G+C content	43.52%	43.67%	43.63%	43.68%
N50 of scaffolds (besides plasmids)	NA	NA	8,262,658	8,171,464
tRNA genes	60	54	56	56
rRNA genes (5S, 16S, 23S)	6	3	6	7
Total genes	7571	7517	6982	7080
Functions assigned	62.17%	62.22%	62.33%	61.5%
Chromosomal GI*	27	30	20	24
Number of BGCs**	44	44	31	33
Genomic % of BGC	19.89	21.96	14.96	14.99

Table A3.2. COGs comparison by category. In red, the highest D-ranks, highlighting differences between categories from draft genomes compared to the reference (PAL). Yellow represents the most common categories (in average percentage of genes). * indicates D-ranks with P value higher that 0.05 (not statically significant).

		Reference	Draft Genomes					
Category	Category Description	PAL	D-rank	PNG	D-rank	3L	D-rank	JHB
B	Chromatin structure and dynamics	2	0*	2	0*	2	0*	2
C	Energy production and conversion	157	0.03	155	0.05	151	0.02	171
D	Cell cycle control, cell division, chromosome partitioning	43	0.01	40	0.05	35	0.03	39
E	Amino acid transport and metabolism	245	0.05	237	0.03	236	0.02	252
F	Nucleotide transport and metabolism	69	0.01	62	0.01	63	0.03	61
G	Carbohydrate transport and metabolism	141	0.05	155	0.04	137	0.04	153
H	Coenzyme transport and metabolism	210	0.03	206	0.01	194	0.01	210
I	Lipid transport and metabolism	115	0.01	96	0.05	92	0.08	123
J	Translation, ribosomal structure and biogenesis	194	0*	190	0*	188	0*	194
K	Transcription	114	0.03	112	0.09	101	0.08	113
L	Replication, recombination and repair	104	0.09	107	0*	91	0.03	91
M	Cell wall/membrane/envelope biogenesis	296	0.09	285	0.23	293	0.16	326
N	Cell motility	69	0.09	63	0.12	59	0.07	69
O	Posttranslational modification, protein turnover, chaperones	163	0.04	149	0.01	145	0.04	169
P	Inorganic ion transport and metabolism	153	0.04	148	0.01	141	0.02	153
Q	Secondary metabolites biosynthesis, transport and catabolism	171	0.58	118	0.57	112	0.08	161
R	General function prediction only	480	0.02	425	0.04	409	0.03	465
S	Function unknown	211	0.01	203	0.01	200	0.04	226
T	Signal transduction mechanisms	266	0.21	256	0.05	237	0.03	269
U	Intracellular trafficking, secretion, and vesicular transport	62	0.27	39	0.31	42	0.2	49
V	Defense mechanisms	135	0.03	125	0.16	115	0.08	135
W	Extracellular structures	17	0*	16	0*	19	0*	18
X	Mobilome: prophages, transposons	294	0.73	183	0.49	226	0.48	235
B-X	Total	3711		3372		3288		3684

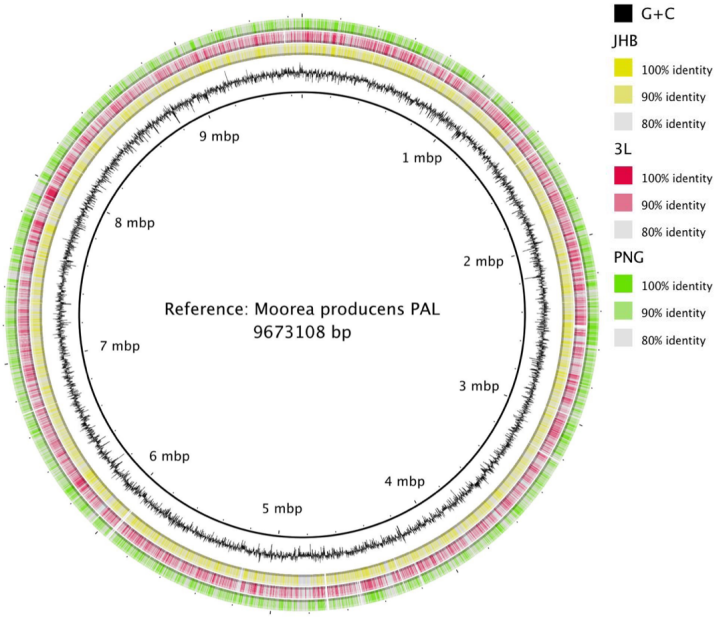


Figure A3.1. Circular map showing three linear draft genomes (PNG, JHB, and 3L) aligned to the reference PAL circular chromosome. Each one of the three outer rings represents a main scaffold/chromosome, with the color code representing percent nucleotide identity. Fourth ring represents G+C content of the reference.

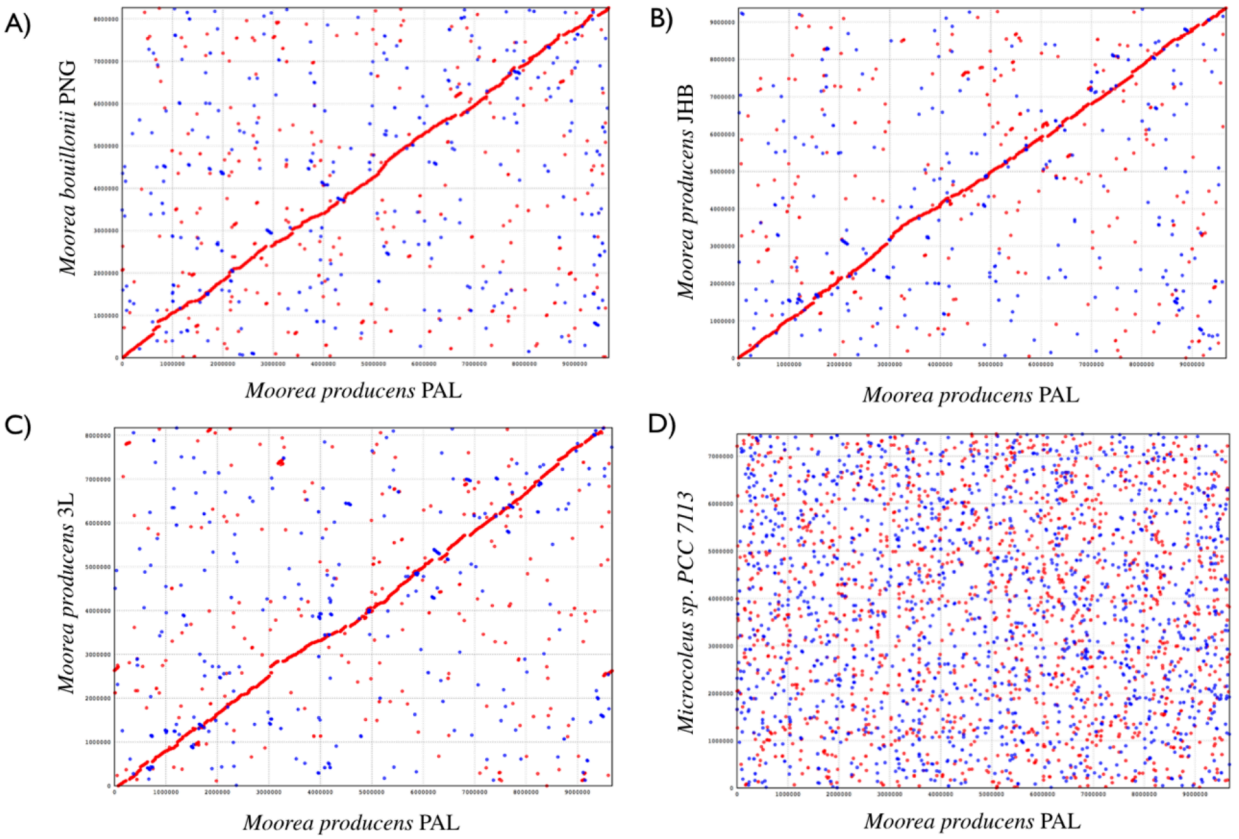


Figure A3.2. MUMmer plots comparing colinearity of *Moorea* producens PAL to (A) *Moorea* bouillonii PNG; (B) *Moorea* producens JHB; (C) *Moorea* producens 3L; and (D) Microcoleus sp. PCC 7113, the next closest phylogenomic relative.

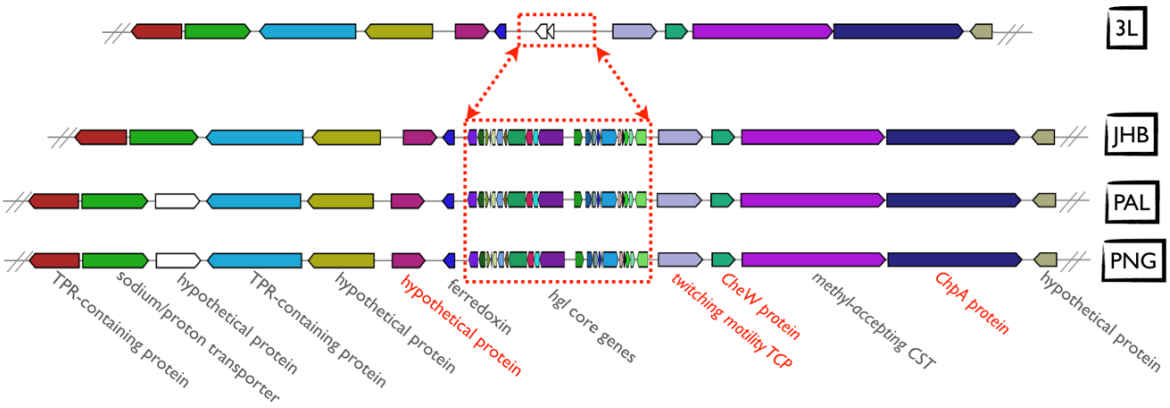


Figure A3.3. Schematic of synteny within the vicinity of *hgl* core genes (same colors represent homologous genes); red boxes and arrows represent the dis- placement of those genes by mobile elements in *M. producens* 3L genome. Minimum identity is 92%, and all coverages are 100%; *hgl* core represents 19–21 genes. CST, chemotaxis sensory transducer; TCP, two-component system; TPR, tetratricopeptide repeat. Gene product predictions were retrieved from JGI expert-reviewed annotation. According to antiSMASH, genes in red are response regulators.

CHAPTER 4: Genomic insights into an expanded diversity of filamentous marine cyanobacteria reveals the extraordinary biosynthetic potential of *Moorea* and *Okeania*

4.1. Abstract

Microbial secondary metabolites (natural products) have been an important and inspirational source for numerous pharmaceutical drugs. However, despite the advent of high throughput DNA sequencing and new analytical techniques, the number of newly discovered natural products per year has remained relatively constant over the past decade. In principal, DNA sequence information can be used to predict new natural product structures in a process known as genome mining. This perceived genetic potential is often referred to as “biosynthetic dark matter”, and recent studies indicate that filamentous marine cyanobacteria harbor an exceptional metabolic potential, portending a promising future for genome-mining from these microbes. However, cyanobacterial genomes are relatively rare in public databases, accounting for 0.6% of all microbial genomes published to date. Natural product rich filamentous marine cyanobacteria are even scarcer; only four *Moorea* and two *Okeania* genomes are currently available in the NCBI database. Here, we utilize a rapid, efficient and reproducible pipeline for genome assembly and initial mining of cyanobacterial biosynthetic pathways. The pipeline generated 81 high-quality genomes from filamentous marine cyanobacteria collected pantropically, including 26 *Moorea* and 22 *Okeania* strains. Genome comparisons highlighted that these two genera are among the most diverse and prolific producers of natural products in our dataset. Furthermore, networking of biosynthetic gene clusters revealed many that were either ‘genera-specific’ or present as ‘extended families’ (found in several strains).

4.2. Introduction

Cyanobacteria are photosynthetic microbes that are abundant in a diverse range of habitats, and support critical life processes in oligotrophic environments via photosynthesis and/or nitrogen fixation. Apart from their well-known importance in biogeochemical cycles because of their primary metabolic characteristics, cyanobacteria are also a prolific and sometimes distinctive source of secondary metabolites known as natural products (NPs). Natural products have been a major inspirational source of new pharmaceutical agents.[7] Despite the advent of new bioprospecting techniques to mine NPs,[111] the number of NPs discovered per year has remained relatively constant over the past decade.[112] The number of novel structures isolated has also remained constant,[112] despite continuing investigation of under-explored habitats and microbial sources of NPs. As a complement to new microbial sources, genome mining can be a powerful tool for prioritizing and directing the isolation of novel chemical entities from these organisms.

Despite the vast genetic diversity of cyanobacteria[25] and their different roles in human and planetary health, genomic investigations have been unevenly distributed throughout this phylum.[113] The current phylogenetic coverage of 425 publically available cyanobacterial genomes from the NCBI RefSeq database is heavily biased toward subsection II. About half (212 out of 425) of the NCBI strains are from the unicellular genera *Prochlorococcus* and *Synechococcus* (subsection II), probably due to their ease of culturing, small genome sizes and their importance in oceanic nitrogen fixation and photosynthesis.[5] Moreover, organisms from these genera usually are less contaminated by associated heterotrophs in laboratory cultures and have a small number of biosynthetic gene clusters (BGCs, known for encoding the NP biosynthetic machinery in microbes); these features tend to facilitate genome sequencing and assembly.[2,114] However, these genera are less relevant for genome mining and drug discovery efforts due to the low quantity and diversity of their BGCs (as opposed to the genus *Moorea*

which, on average, has a four-fold higher biosynthetic potential).[26] Reasons for the relative scarceness of sequenced NP rich cyanobacterial genomes may result from difficulty in culturing these types of filamentous marine cyanobacteria, the repetitive elements found in their genomes, their larger genome size, and the lack of an efficient and automated genome assembly pipeline. Here, we provide a rapid, efficient and fully reproducible pipeline for genome assembly and initial mining of cyanobacterial biosynthetic pathways. This pipeline produced 81 high quality draft genomes from complex environmental and cultured marine cyanobacterial samples, and allowed us to perform comparative genomics, searching for patterns in the presence or absence of BGCs that can be used in automated genome mining.

4.3 Methods

4.3.1 Collection, DNA extraction and sequencing

Samples were collected via scuba diving or snorkeling in shallow benthic environments (no deeper than 20 meters) from different shores around the globe. The collected biomass was preserved in RNAlater for subsequent sequencing or 1:1 isopropanol:seawater for posterior mass spectrometry analysis. Given that *Moorea* and *Okeania* can form macroscopic tufts in seawater, we focused on collecting samples that matched the morphology from these two genera. When possible, we obtained purified cultures using standard microbiology and microscopy techniques, generating 22 non-axenic uni-cyanobacterial cultures. RNAlater samples were processed via freezing with liquid nitrogen, followed by grinding and extraction according to the “QIAGEN Bacterial Genomic DNA Extraction Kit” protocol, incubation times extended for 1h and the volume of Proteinase K used was 10μl (10 mg/ml). For details on the extraction procedure, please refer to the following book chapter: Collection, Culturing, and Genome Analyses of Tropical Marine Filamentous Benthic Cyanobacteria, by Moss and collaborators (2018).[115]

The library was generate using a miniaturized version of the Kapa HyperPlus Illumina-compatible library prep kit (Kapa Biosystems®). DNA extracts were normalized to 5 ng total input per sample in an Echo 550 acoustic liquid handling robot (Labcyte Inc). Next, we used a Mosquito HTS liquid-handling robot (TTP Labtech Inc) for 1/10 scale enzymatic fragmentation, end-repair, and adapter-ligation reactions. Sequencing adapters were based on the iTru protocol, in which short universal adapter stubs are ligated first and then sample-specific barcoded sequences added in a subsequent PCR step. Amplified and barcoded libraries were then quantified by the PicoGreen assay and pooled in approximately equimolar ratios before being sequenced on an Illumina HiSeq 4000 instrument to >30X metagenomic coverage. Two samples were complemented with long read sequence. *Leptolyngbya* sp. SIOISBB was sequenced with Nanopore MinION® using 1D² Sequencing Kit (R9.5) (ligation-based). *Moorea* sp. SIOASIH was sequenced PacBio RS® using a 10kb library prep.

4.3.2. Genome assembly pipeline

187 sequenced metagenomic samples were assembled with metaSPAdes 3.12.0. Assembled contigs were annotated with Prokka 1.11 and the phylogenetic assignment for each annotated gene was predicted via DarkHorse 2.0. Only contigs that followed minimal requirements were binned in: one or more cyanobacterial genes, and; GC content smaller or equal to 58% (determined via QUAST analysis of high GC draft metagenomes). 165 draft genomes were successfully scaffolded using MEDUSA (http://combo.dbe.unifi.it/medusa) and high quality reference genomes from the NCBI database. Once scaffolded, the quality control via CheckM (http://ecogenomics.github.io/CheckM/) approved 81 high-quality draft genomes (over 90% completeness). *Leptolyngbya* sp. SIOISBB long reads were process for base calling with Albacore 2.0.2, trimmed with PoreChop 1.0 and assembled with Canu 1.6. *Moorea* sp.

SIOASIH long reads were part of a hybrid assembly with metaSPAdes 3.12.0, followed by binning and quality control as described above for the short reads.

4.3.3. Phylogenomics

We selected the same set of conserved 29 housekeeping genes used in Leao and collaborators (2017)[2] and Calteau and collaborators (2014)[97]. The homologs of these housekeeping genes were identified in our strains via Diamond search (v0.8.31.93). Once identified and extracted, 29 set of genes were aligned using MUSCLE v3.8 and trimmed via trimAl v1.2 (both in standard settings). Once the alignments were complete, we concatenated the housekeeping genes. A phylogenetic tree was reconstructed based on the concatenated ribosomal protein sequences extracted from the cyanobacterial genomes, using maximum likelihood (ML) implemented in IQ-TREE 1.6.10. Amino acid substitution model was determined using ModelFinder as part of IQ-TREE, which chose LG+R10 (LG substitution matrix, plus FreeRate model with 10 rate categories) as the best model. Phylogenetic reconstruction was performed using this model and IQ-TREE default settings. Branch supports were provided using 100 replicates of classical bootstrap, the out-group was Melainabacteria SM1 D11. The overall shape and clades of the tree were consistent with previous studies. All genomes but *Merismopedia* sp. 2A8 contained a complete set of the 29 selected housekeeping genes.

4.3.4. Gene cluster networking and diversity analysis

187 metagenomic drafts (not fully assembled like the 81 high-quality scaffolded drafts) were investigated for their biosynthetic gene clusters (BGCs). Their predicted BGCs via antiSMASH v3.0.5 were networked using BiG-SCAPE (https://bigscape-corason.secondarymetabolites.org/index.html) to obtain pairwise similarity scores between the BGCs. The similarity cutoff calibration was determined by checking the networking performance for annotated BGCs from the MIBiG database (Minimum Information about a Biosynthetic Gene

cluster database). The best cutoffs included 0.7 similarity (0.3 distance), which corresponded to 99% structural similarity between the annotated metabolites in MIBiG. For the de-noising density-based scanning (DBScan) algorithm, the best outcome corresponded to the epsilon score of 0.6, which presented one of the highest similarities between annotated structure (over 80%) and retained the most families. By selecting a cutoff, we converted a similarity metric into presence/absence of gene families (group of homologous BGCs). We built a pairwise Brays-Curtis beta-diversity dissimilarity matrix among all tested strains (using skbio.diversity package in python), including the 425 previously published NCBI genomes. Subsequently, we calculated the average Brays-Curtis beta-diversity per strain and built a Principal Coordinate Analysis (PCoA) plot to highlight samples with high diversity score (over 95% average Brays-Curtis beta-diversity). For analyzing the extended families, we decided not to include a cutoff threshold so we could evaluate all BGCs, including incomplete ones. Lastly, we subset the gene network for gene cluster families containing the newly sequenced *Moorea* and *Okeania* samples, plotting this network using Cytoscape 3.2.1.

4.4. Results and discussion

4.4.1. Description of samples

The Scripps cyanobacterial sample collection consists of environmental samples collected and stored in RNA-later as well as roughly 70 cultures of tropical filamentous marine cyanobacteria. For the current project, we sequenced 143 environmental samples from genetically underexplored cyanobacteria, found as macroscopic tufts in sub-tidal tropical ecosystems throughout the globe, (Figure A4.1A), along with 22 purified non-axenic cultures (total of 165 strains); three of these non-axenic cultures were cyanobacteria derived from sub-tidal stromatolites. By applying the genome assembly steps outlined below to this combination of environmental and cultured samples, we obtained 81 high quality draft genomes (over 90%

CheckM completeness, Figure A4.1B); 67 of these derived from environmental samples and 14 from non-axenic uni-cyanobacterial cultures.

4.4.2. New genome assembly pipeline improves genomic diversity of natural product rich cyanobacteria

The current phylogenetic distribution of 425 publically available cyanobacterial genomes from NCBI RefSeq database (Dec 2018) is heavily biased toward subsection II. Strains from this subsection tend to be less promising for drug discovery due to the low abundance and diversity of BGCs in their genomes.[26] The lack of an efficient and automated genome assembly pipeline has hindered the expansion of NP rich cyanobacterial genomes from other subsections, and consequently, has hampered genome mining approaches from these promising sources of structurally diverse NPs. Therefore, we developed a rapid, automated and reproducible pipeline for assembly and mining of cyanobacterial genomes. This pipeline sequentially performs assembly by metaSPAdes,[67] taxonomic binning using GC content and DarkHorse,[70] and quality control analysis to yield high-quality genomes that can be used for more sophisticated investigation, such as biodiversity assessment and genome mining for new NP scaffolds.

Our pipeline was able to expand the public repository of cyanobacterial genomes by 20%, including several of the natural product rich filamentous cyanobacteria from subsection VI (for example, 26 *Moorea* and 22 *Okeania* strains), that are currently underrepresented in databases. A histogram (Figure A4.1B) represents the distribution of these newly assembled genomes according to their CheckM[116] completeness, where the 81 high quality drafts are highlighted in red (from 90 to 100% completeness). Figure A4.1C indicates a weak significant correlation (Pearson's rho of 0.56, p-value of 3.4E-09) between the final draft quality and the number of reads from metagenomes that were attributable to cyanobacteria. As noted in Figure A4.1C, the assembly of a high quality draft genome is expected if greater than 1.25 million

cyanobacterial reads are obtained. For one of our cultures, *Leptolyngbya* sp. SIOISBB (collection code ISB3NOV948BCUL), which is currently under development as a heterologous host for expression of marine BGCs, we complemented the short read sequencing with long reads from Nanopore MinION® in order to obtain the complete genome. For a second culture, *Moorea* sp. SIO1ASIH (collection code ASI16JUL142CUL), which is the producer of a related suit of compounds given the common names of 'vatiamides A-F', we complemented the short read sequencing with PacBio RS® in order to reduce the draft genome to 24 contigs. Therefore, we were able to preferentially collect, sequence and extract from complex metagenomes several genetically distinct marine cyanobacteria, thereby significantly expanding knowledge of natural product rich filamentous cyanobacteria and identifying promising strains for further exploration and comparison at the genomic level.

Cyanobacterial phylogeny has been and continues to be challenging, mainly due to previous assignments that used morphological characteristics and lacked a genetic basis. Figure 4.1 illustrates the phylogenetic assignments for the 81 high quality draft genomes, including 26 *Moorea*, 22 *Okeania*, 12 *Symploca*, 7 *Leptolyngbya*, 3 *Cyanothece*, 2 *Sphaerospermopsis*, 2 *Kamptonema*, 1 *Desertifilum*, 1 *Merismopedia*, 1 *Microcoleus*, 1 *Nostoc*, 1 *Oscillatoria*, 1 Planktothrix and 1 *Spirulina*. Moreover, the three cyanobacterial cultures isolated from stromatolites consisted of two *Leptolyngbya* and one *Kamptonema*. As can be observed in Figure 4.1, the *Moorea* clade (in red) is tightly defined and it remains fairly distant from the closest non-*Moorea* reference genome (*Microcoleus* sp. PCC7113). In contrast, the *Okeania* clade (in blue) appears to be more diverse (higher degree and number of branches) and its clade includes the closest relative, *Trichodesmium erythraeum* IMS101. Red arrows in the phylogenomic tree highlight the 33 remaining genomes (Figure 4.1). The 81 draft genomes ranged from 3.3 up to 20 Mb, with the average size was 9.3 Mb. The GC content ranged from 36% to 53.6% and the average GC content was 42.75%. The number of scaffolds ranged from

3 up to 6,587, and the average number of scaffolds per genome was 1,474. The total number of BGCs ranged from a genome with a single BGC up to an exceptional genome with 68 BGCs.

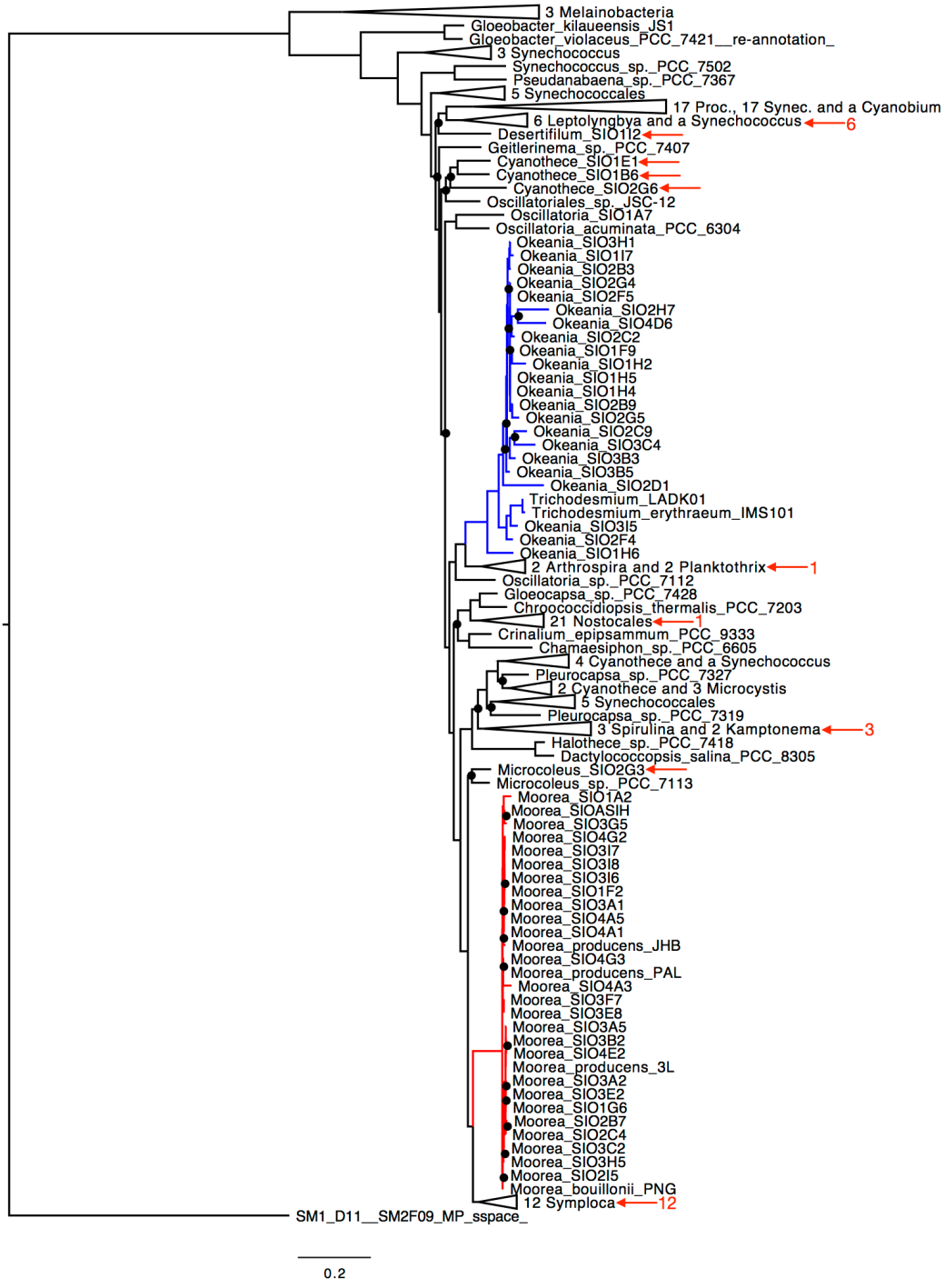


Figure 4.1. Phylogenomic analyses of completed cyanobacterial genomes using 29 conserved genes from Calteau et al. [97]. Tips were labeled either according to phylogenomic cladding and 16S identity. Major clades, *Moorea* and *Okeania*, are colored in red and blue, respectively. The remaining genomes sequenced in this study are identified by red arrows. Proc. = *Prochlorococcus*; Synec. = *Synechococcus*. Bootstrap values lower than 70 are displayed as dots.

4.4.3. Phylum-wide BGC analysis highlights the most diverse genomes

As described above, an automated workflow was developed for 165 different cyanobacterial environmental assemblages and cultures that utilized short reads provided by Illumina HiSeq4000®. This generated 81 high-quality draft genomes with 90-100% completeness. Subsequently, an expanded public repository of 506 cyanobacterial genomes, including the 81 new draft genomes reported here, were scanned for BGC potential using antiSMASH v3.0;[98] this identified a total of 1,979 BGCs. These BGCs were compared via domain similarity using BiG-SCAPE v1.0[117] which groups homologous BGCs into gene cluster families (GCFs). However, the frequency of incorrectly grouped BGCs per family (hereafter referred to as "noise") is heavily dependent on the number of incomplete pathways and the similarity cutoff selected in the analysis. Hence, by performing a cutoff calibration for families containing two or more previously characterized BGCs from the MIBiG database (currently containing 1,416 BGC entries), these could be well differentiated into discreet GCFs. Using Tanimoto scoring,[112] we compared the annotated structures for structural similarity between the MIBiG gene pathways predicted to be homologous with the goal of evaluating homology between chemical structures and biosynthetic genes. However, different cutoffs resulted in different levels of 'noise'. The more 'noise' present in GCFs results in less average structural similarity. Therefore, we evaluated the most efficient cutoff by analyzing the validated part of our dataset (e.g. families that contain two or more MIBiG BGCs). However, only 9% of the GCFs that are identifiable from the 506 genomes analyzed here have representative BGCs that are annotated in the MIBiG database. Thus, many of the families are annotated by code number rather than compound family name. Next, we performed a Density-Based Scanning for Applications with Noise (DBSCAN) de-noising step in order to minimize the problem introduced by incomplete BGCs. The cutoff selected during the MIBiG validation was then applied to the

remainder of the dataset, generating a gene cluster similarity network. In this case, the gene cluster network contained 1,979 BGCs from cyanobacteria plus 1,416 BGCs from MIBiG, yielding 341 gene cluster families under the best cutoff outcome of 0.6, as illustrated in Figure 4.2A (for details, see Supporting Information page S3).

Next, using the gene cluster similarity network, we accessed the distribution of GCFs in these samples and calculated diversity scores, such as beta-diversity via Jaccard similarity. In this analysis, beta-diversity illustrates the percentage difference between pairs of samples in terms of their harbored BGCs (pairwise measurement defined as the intersection over the union of two cyanobacterial strains A and B). In a phylum-wide analysis (506 genomes), cyanobacteria that exhibited low average beta-diversity and a small number of BGCs in their genomes appeared to be the most likely to yield the rediscovery of previously characterized NP scaffolds. Conversely, highly diverse samples that displayed many different biosynthetic gene families were the most likely to yield novel structures. A Principal Coordinate beta-diversity analysis (PCoA plot from Figure 4.2B) clearly identified a group that harbors the most diverse distribution of GCFs, containing an average dissimilarity score of over 95%. This group is comprised of 49 genomes with the top five most diverse being 9 *Moorea*, 8 *Okeania*, 6 *Planktothrix*, 4 *Microcystis* and 2 *Symploca* samples. Of note, this group also included several *Prochlorococcus* and *Synechococcus* strains; however, we kept them out of this analysis because most of these latter genomes are over-represented and only harbored a few BGCs (e.g. their diversity scores were relatively inaccurate), as commonly observed for this clade. Summarizing, this beta-diversity analysis confirmed that *Moorea* has the most prolific metabolic potential among cyanobacteria, and also revealed that a few other marine strains are highly diverse and rich in natural product BGCs. For example, the genus *Okeania* is similarly rich in these regards to *Moorea*, and has subsequently been targeted for new natural product discovery.[118,119]

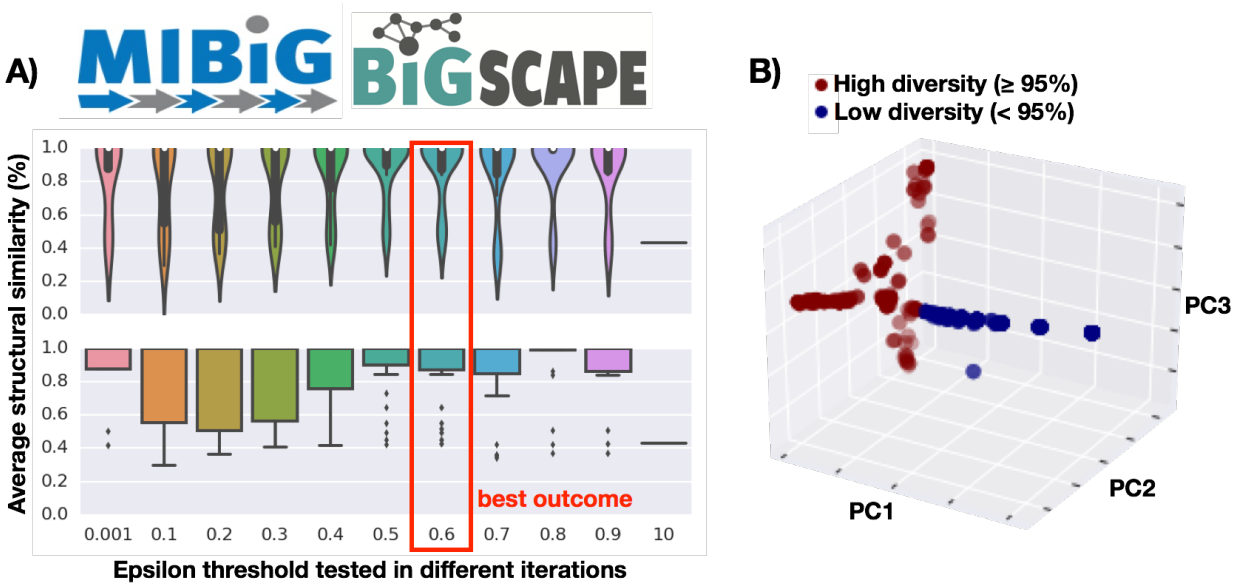


Figure 4.2. A) Cutoff selection via networking expert-annotated biosynthetic gene clusters from MIBIG database. When networks were created with a cutoff of 0.6, gene clusters (and the molecules they produce) tended to share more than 80% structural similarity between their final molecular products. B) Principal Coordinate Analysis (PCoA) for beta-diversity scores from 506 draft genomes (including the 81 marine cyanobacteria generated in this project). The red group represents the most prolific natural product rich samples in our dataset (over average 95% beta-diversity) and the blue group represents samples with average score below 95%.

4.4.4. Comparing the diverse secondary metabolism of *Moorea* and *Okeania*

This sequencing effort has resulted in the addition of many new BGCs to the gene cluster network (Figure 4.3A), and highlight the importance of the newly sequence genomes to the previously reported secondary metabolism in *Moorea* (in red) and *Okeania* (in blue). The genus *Moorea*, represented here by 26 high quality draft genomes, contains 733 BGCs that fall into 63 GCFs (Figure 4.3B). Twenty-three high quality draft genomes were obtained for the genus *Okeania*, and this contained 270 BGCs that described 45 GCFs (Figure 4.3C). Thus, *Moorea* contains more “extended gene families” (e.g. belonging to the same gene cluster family and found in many strains) than *Okeania*. The gene families in *Moorea* have a larger number of BGCs that share an overall average identity of 40-60% at the nucleotide level. In *Moorea*, each family possessed an average of 18.3 BGCs (Figure 4.4A, top), and seven of these 63 families were unique to the genera *Moorea*, confirming the previously reported prolific metabolic potential of this genus.[2] Surprisingly, eight of the families harbored nearly 80% of the BGCs

present in *Moorea* (hence, they were considered to be 'extended families'). This analysis further revealed that there are three classes of GCFs; unique GCFs, small GCFs and 'extended families' (GCFs found in most strains, highlighted in Figure 4.4B). These 'extended families' include putative housekeeping BGCs such as phytoene-like terpenes, olefin synthases and heterocyst glycolipids (*hgl*) clusters. They also included the curacin A-D gene family, 2 cryptic cyanobactin families, a cryptic dipeptide family and a fully cryptic pathway annotated as "other" (according to antiSMASH 3.0). The abundance of 'extended families' in *Moorea* may result from this genus generally occupying similar ecological niches on tropical reef systems that are subject to comparable environmental pressures. Interestingly, within a given "extended BGC family" (Figure 4.4B, rows indicated with red arrows), the sequences are not very similar to one another (40-60%), indicating that these have been subject to evolutionary pressures to shuffle and mutate the component biosynthetic genes. Similar trends are observed in *Okeania* (Figure 4.4C) wherein the average number of BGCs per family is 12.8 and three of the 45 families are unique to the genus *Okeania*, thus also representing an under-explored NP potential. In total, seven of the *Okeania* 'extended families' harbored 78% of *Okeania*'s BGCs, and include GCFs likely encoding for phytoene-like terpenes and the lipopeptide family of NPs known as 'malyngamides'.[118] Cryptic 'extended family' pathways in *Okeania* included a fully cryptic BGC, a few cyanobactins and two small NRPS/PKS gene clusters, the products of which remain unknown.

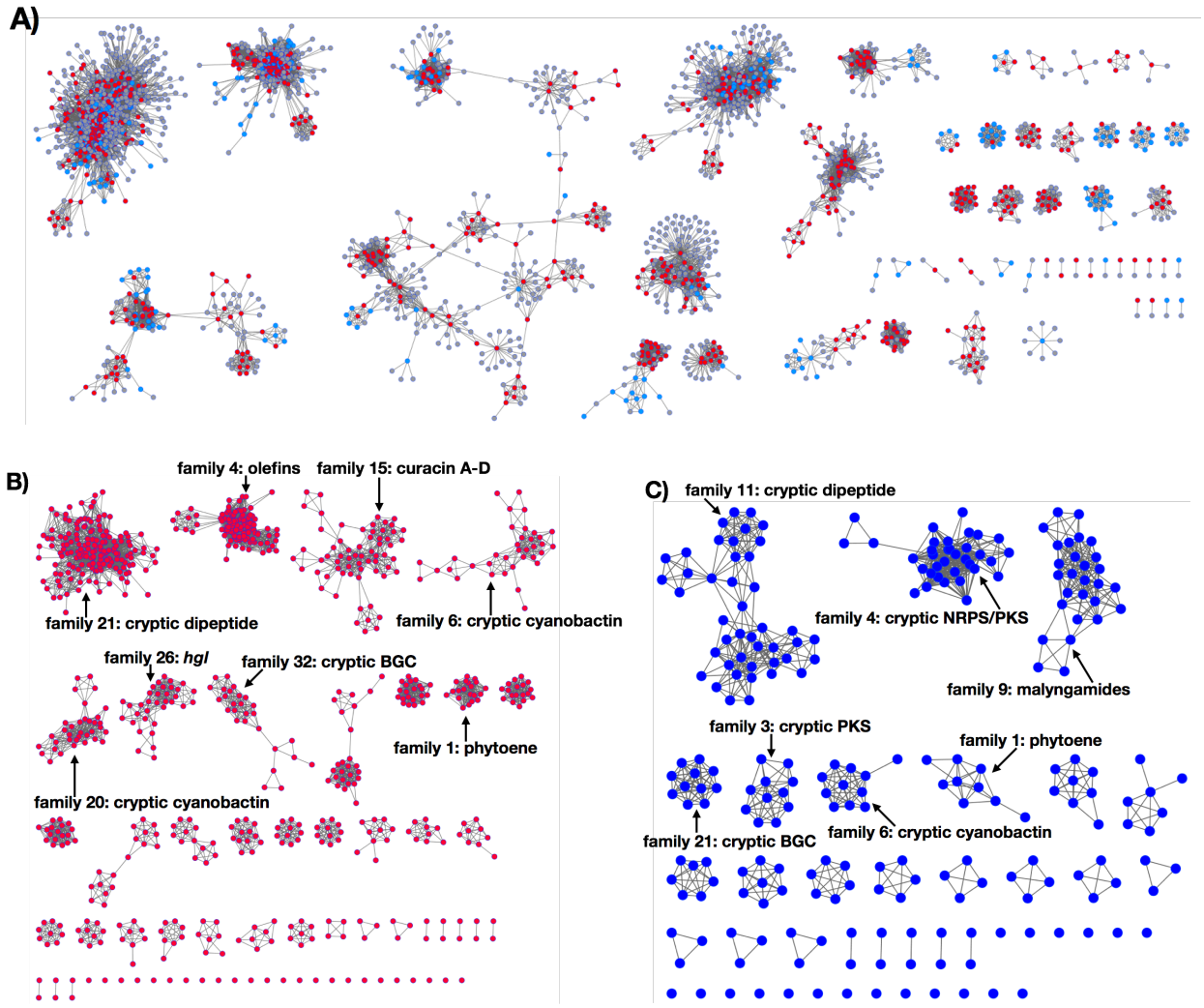


Figure 4.3. A) BiG-SCAPE similarity network for biosynthetic gene clusters (BGCs) from *Moorea* (red) and *Okeania* (blue). B) BiG-SCAPE similarity network for BGCs only from *Moorea*. C) BiG-SCAPE similarity network for BGCs only from *Okeania*. Connected nodes belong to the same gene families. In panels B and C, "extended families" (families with total number of BGCs higher than the average) are annotated with 'family' designation and predicted natural product type, when known.

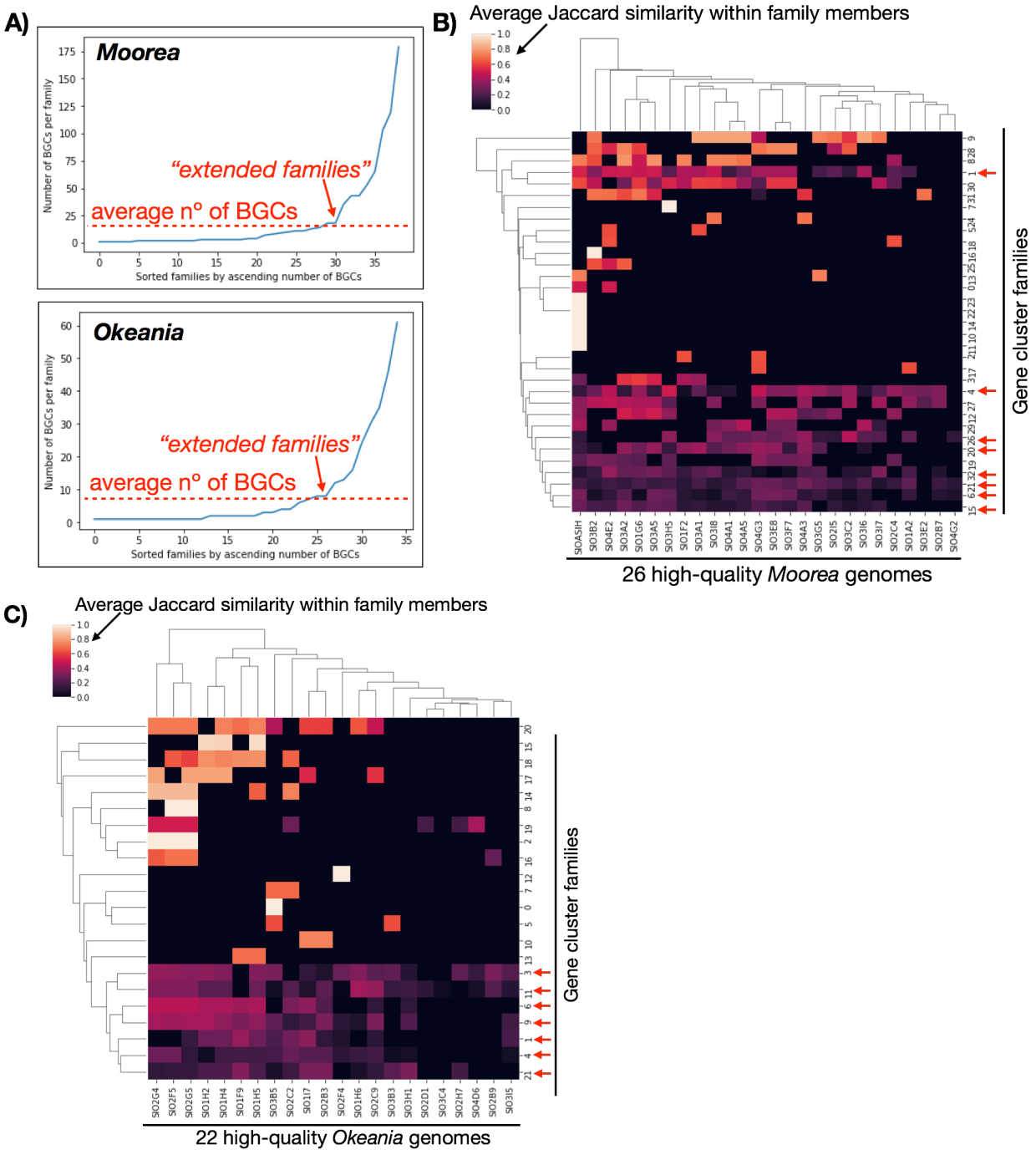


Figure 4.4. A) Number of biosynthetic gene clusters (BGCs) found in each gene cluster family (GCF) in ascending number of BGCs. B) *Moorea* heatmap for BiG-SCAPE similarity scores within members of the same GCF. In *Moorea*, these are: cryptic cyanobactins (#6 and #20); olefin synthase (#4); phytoene-like terpene (#1); curacins family (#15); associated with the hgl cluster (#26); a cryptic and small dipeptide (#21); a fully cryptic GCF (#32). C) *Okeania* heatmap for BiG-SCAPE similarity scores within members of the same GCF. In *Okeania*, the "extended families" were annotated as: a putative phytoene (#1); cryptic cyanobactin (#6); cryptic small NRPS/PKS (#4); malyngamides (#9); a small cryptic PKS (#3); a fully cryptic BGC (#21) (labeled as "other" by antiSMASH). In the heatmaps, unique BGCs are represented by 1.0 similarity. Red arrows highlight "extended families" (families containing more than the average number of BGCs in the given genus).

Chapter 4, in full, currently being prepared for submission for publication. Tiago Leão, Ricardo Silva, Nathan Moss, Mingxun Wang, Jon Sanders, Sergey Nurk, Alexey Gurevich, Gregory Humphrey, Raphael Reher, Qiyun Zhu, Pedro Belda-Ferre, Pieter Dorrestein, Rob Knight, Pavel Pevzner, William H. Gerwick and Lena Gerwick. “Genomic insights into an expanded diversity of filamentous marine cyanobacteria reveals the extraordinary biosynthetic potential of *Moorea* and *Okeania*”. The dissertation author is the primary investigator and author of this material.

4.5. Appendix

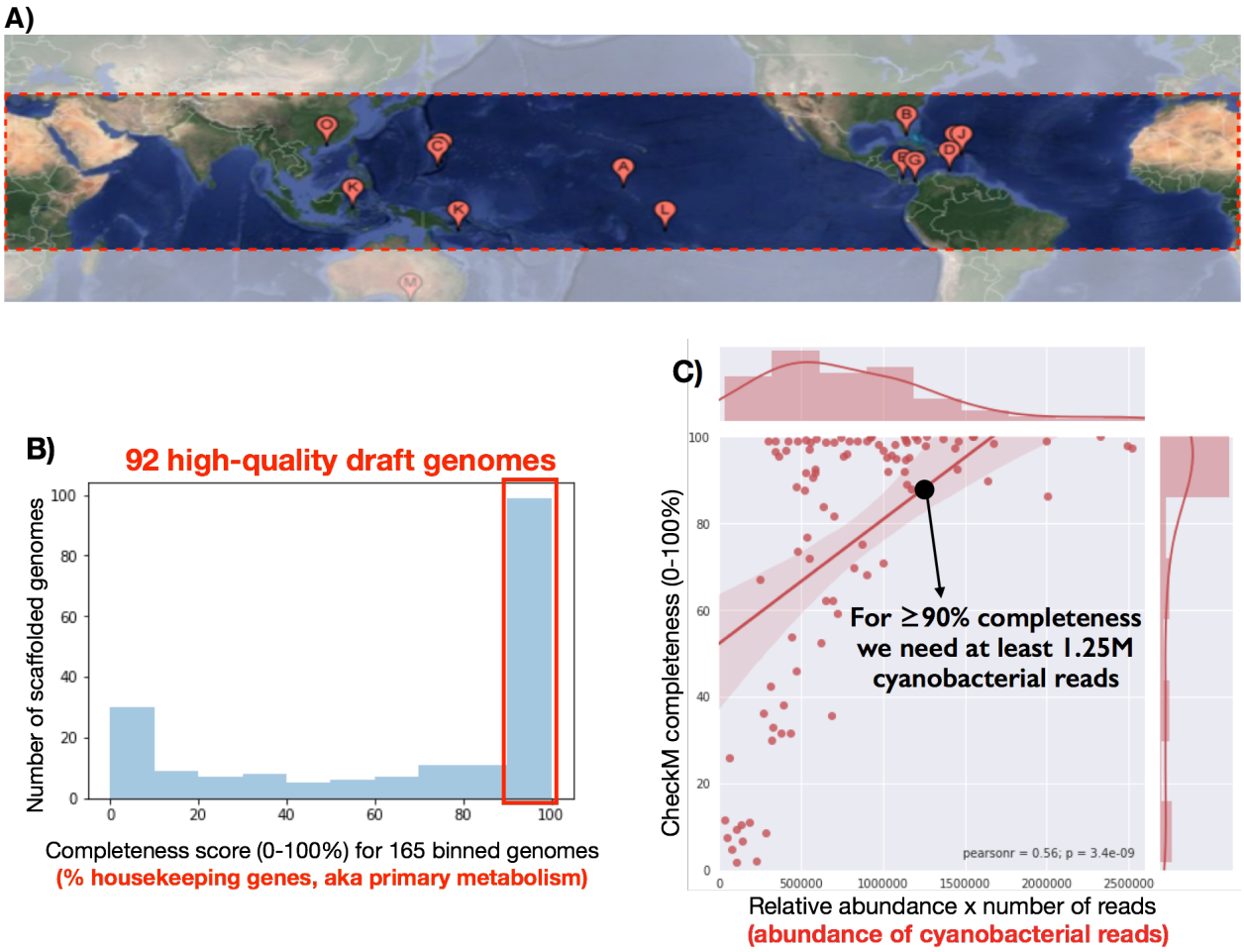


Figure A4.1. A) Geography locations for 165 collected metagenomic samples B) Quality scores for assembled 165 draft genomes, highlighting the 85 high-quality draft genomes obtained by the Cyanobiome project C) Correlation between completeness scores and abundance of cyanobacterial reads in the original metagenomic sample.

CHAPTER 5: Conclusion

Cyanobacteria are one of the three most prolific phyla for the discovery of new microbial natural products,[112] the other two being actinobacteria and myxobacteria. As exemplified in Chapter 1 (introduction), much is known about the chemistry and biochemistry of cyanobacterial natural products. A growing body of knowledge also exists about how these natural products are biosynthesized, in particular hybrid PKS/NRPS, RiPPs, indole-alkaloids and MAAs. However, many more natural products are yet to be identified from these prokaryotes, and further exploration of unusual enzymatic transformations are warranted. In addition, not as much is known about other general microbial traits. For example, despite the fact that the cyanobacterial genus *Moorea* is one of the major producers of natural products, only one draft genome of this genus existed in public databases before this thesis. *Okeania*, another prolific natural product producing cyanobacterial genus, had no genomes in public databases before this work. Thus, this present work greatly expands our knowledge on the genetics behind the prolific producers of secondary metabolites from the genera *Moorea* and *Okeania*. Summarizing, we were able to sequence and bioinformatically investigate several different strains from these two genera, providing many conclusions that improved our understanding of their microbiology, genetics and natural products potential.

5.1. Insights into the microbiology of *Moorea producens* JHB and an associated uncultured heterotroph

In chapter 2, the sequencing effort of the *M. producens* JHB metagenome revealed a complete 5.99 Mb genome of an unknown, uncultured heterotrophic bacterium, named here as Mor1. The organism belongs to the phylum acidobacteria, subgroup 22, but is unable to be taxonomically further classified, and as of yet, has not been successfully cultured. A comparative genomics study generated several hypotheses regarding the potential relationship

between the *M. producens* JHB and Mor1. Four main areas of interest emerged from this analysis: transcriptional regulation, iron metabolism, nitrogen cycling between the two microbial species, and a complete lack of transposases in the Mor1 genome. Examination for the presence of Mor1 in various laboratory cultures, along with co-culturing experiments to evaluate the transferability of Mor1 to other cyanobacteria, support the idea that it is a specific associate of some strains of *M. producens*.

Because all natural products investigations of *M. producens* JHB to date have occurred using either field collected or cultured non-axenic biomass, the latter comprised of cyanobacterial filaments and associated microbiome, there remains the possibility that some of the identified secondary metabolites (e.g. jamaicamide, hectochlorin) are actually produced by Mor1 or another associated heterotrophic bacterium. Because we have not been able to cultivate Mor1 or the cyanobacterium independent of one another, we are not able to answer this question using chemical methodologies. However, several lines of evidence support the conclusion that these natural products are of cyanobacterial origin: 1) their chemical structures are consistent with pathways known for cyanobacterial natural products, 2) they are produced in relatively high yield per unit of biomass, 3) their biosynthetic pathways use motifs, codons, and GC content consistent with cyanobacterial pathways, and 4) new to this reported work, the major associated heterotrophic bacterium, Mor1, lacks the genes for these metabolic pathways. In fact, only two secondary metabolite pathways were detected in Mor1 based on an antiSMASH analysis of its assembled genome, and neither of these is predicted to produce a compound thought to be of cyanobacterial origin.

To further characterize the potential symbiotic relationship between Mor1 and *M. producens*, additional effort is needed to culture *M. producens* and Mor1 independently of one another. Growth rates of *M. producens* with and without Mor1 might infer a symbiotic interaction of these two microbial species. Interaction between these two organisms involving nutrient

exchange or signaling molecules could be examined via a transcriptomic analysis, imaging mass spectrometry, or further chemical analyses. Overall, this work reveals that niche environments such as the sheaths of tropical marine cyanobacteria may be rich locations in which to prospect for novel microbial species with potentially useful biotechnology applications.

5.2. Insights into the genetics of four *Moorea* strains

In chapter 3, the development of a reference genome for *M. producens* PAL was achieved and has increased our understanding of the genomic capacities of three related strains of filamentous cyanobacteria, providing fresh insights into this important source for natural products. Analysis of the genetic constitution and relationship of *Moorea* to other cyanobacteria suggests that the genus is distinctive among known cyanobacteria, especially in its exceptional capacity for production of secondary metabolites. Phylogenomic analysis of all complete cyanobacterial genomes demonstrated that *Moorea* contained at least twice the number of biosynthetic gene clusters (BGCs) compared to its closest neighbors. This comparison was expanded to all bacteria available in the JGI database and these four *Moorea* ranked among the richest in terms of metabolic potential, dedicating about one fifth of its genome for the production of secondary metabolites.

Using gene cluster networking, we were able to demonstrate that many of the *Moorea* BGCs are rare among bacterial genomes, and suggests future directions for productive genome-guided isolation efforts of novel NPs from this genus. To accomplish the gene cluster networking, the program BIOCOMPASS was developed. In this networking program, similar BGCs like jamaicamides, curacins and carmabins grouped together to the same biosynthetic family. In contrast, we identified other fairly unique BGCs like cryptomaldamide and palmyramide which were distinct from all other families (also named orphan BGCs). It was interesting to observe the abundance of orphan BGCs, providing indication of the potential to

discover more unique metabolites from this genus. This potential was further confirmed in Chapter 4, by expanding the genomic analysis into more *Moorea* genomes.

5.3. Insights into the expanded diversity of *Moorea* and *Okeania* and their natural product potential

In chapter 4, we were able to enrich the diversity of genomic information for natural product rich cyanobacteria by providing 81 new high-quality draft genomes. We demonstrated via a phylum-wide analysis how prioritization of samples using beta-diversity can highlight "taxonomic chemodiversity hotspots" in a given dataset. This was demonstrated using the combination of the 81 new cyanobacterial genomes reported here along with 425 cyanobacterial genomes obtained from NCBI. Using principal coordinate analysis, we observed a split between samples with high beta-diversity (over 95% average pairwise dissimilarity score) and low beta-diversity (less than 95%). The most abundant samples in the high diversity group included *Moorea* and *Okeania*, confirming our previous results indicating that these two genera have one of the largest secondary metabolite potential among all cyanobacteria. Additionally, we observed that two kinds of BGC families were abundant in these two genera: genus-specific, families only found within a particular species, and; "extended families", families found in multiple species and sometimes across different genera. 'Extended families' presented overall lower gene similarity, allowing room for the production of dissimilar natural products, such as the curacins and the malyngamides, two lipopeptides that differ in the incorporated amino acid and the substitutions on the lipid chain.

Lastly, we concluded that the genera *Moorea* and *Okeania* present a vast potential for genomic-driven natural product discovery using such techniques as pattern-based genome mining or the genomisotopic approach. Given the large number of paired genomic-mass spectrometry data sets, future efforts can search for patterns in the presence/absence of BGCs and compounds, isolating peaks for which the distribution in samples matches the distribution of

certain BGCs of interest. For some samples of particular interest, one could predict the building blocks required by a given queried BGC and attempt to feed labeled substrates that would shift the molecular ion in the mass spectrum of a molecule that incorporates these substrates. This could be used to target the molecule for isolation and structure elucidation. The genetic resources, analysis and modeling of the present data have provided an initial framework for correlating biosynthetic genes to their secondary metabolite counterparts, a promising lead for future annotation of cryptic metabolites from the prolific genera *Moorea* and *Okeania*.

5.4. Future perspective

Natural product discovery has evolved considerably with the advance of new approaches for discovering novel molecules that might be promising leads for drug development. In particular, bottom-up approaches (approaches that use the genetic information to infer clues that can help isolate new natural products) can aid the identification of new classes of compounds, avoiding the common issue of rediscovery of known molecules. However, the field of natural product discovery still lacks a systematic approach for genome mining. Nevertheless, genome mining has been successfully used for identifying under-explored sources of natural products, although it is rarely used for systematically connecting cryptic BGCs to detected metabolites (e.g. ions detected via LC-MS/MS). Given the vast knowledge of marine filamentous cyanobacterial metabolomics,[120] this thesis focused on expanding the knowledge on the genetics of these organisms in the hope that these two broad data sets might be connected in future efforts.

Many systematic analyses have been developed for such applications such as phylogenetics, transcriptomics, proteomics and primary metabolomics; however, these have not yet been effective in improving the discovery rate of natural products.[1] Since the sequencing of the first microbial genome in 2002, a surprisingly high percentage of cryptic biosynthetic

pathways per genome have been found (genes not connected to any known secondary metabolite).[2] In addition, the number of new genomes being sequenced has continued to steadily increase,[121] highlighting the need for new techniques that can help annotate this ever increasing natural product dataset. At the present time, there are over 1 million cryptic BGCs that are not annotated.[122] Previous efforts envisioned that presence/absence of BGCs and metabolites could be used to connect genes to molecules. However, such presence/absence approach (named pattern-based genome mining) can produce very sparse matrices such that BGCs can happen to be connected to a molecule just by chance, representing incorrect assignments generated by this simplistic correlative approach. Therefore, we envision that a future direction for the systematic analysis of such large datasets would be a “supervised machine learning” approach, using artificial intelligence to order label data so as to then predict new labels for unstructured data. If such an approach is successful, natural product researchers would then be able to predict the correct BGCs for each cryptic metabolite peak identified via mass spectrometry, portending a bright future for genome mining of novel natural products.

REFERENCES

1. Deane, C. D. & Mitchell, D. a. Lessons learned from the transformation of natural product discovery to a genome-driven endeavor. *J. Ind. Microbiol. Biotechnol.* **41**, 315–331 (2014).

2. Leao, T., Castelão, G., Korobeynikov, A., Monroe, E. A., Podell, S., Glukhov, E., Allen, E. E., Gerwick, W. H. & Gerwick, L. Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus *Moorea*. *Proc. Natl. Acad. Sci.* 201618556 (2017). doi:10.1073/pnas.1618556114

3. Cummings, S. L., Barbé, D., Leao, T. F., Korobeynikov, A., Engene, N., Glukhov, E., Gerwick, W. H. & Gerwick, L. A novel uncultured heterotrophic bacterial associate of the cyanobacterium Moorea producens JHB. *BMC Microbiol.* **16**, 198 (2016).

4. Schirrmeister, B. E., Gugger, M. & Donoghue, P. C. J. Cyanobacteria and the Great Oxidation Event: evidence from genes and fossils. *Palaeontology* **58**, 769–785 (2015).

5. Flores, E., López-lozano, A. & Herrero, A. Nitrogen Fixation in the Oxygenic (Cyanobacteria): The Fight Against Oxygen. *Biol. Nitrogen Fixat.* **2**, 879–889 (2015).

6. Eiler, A. & Bertilsson, S. Composition of freshwater bacterial communities associated with cyanobacterial blooms in four Swedish lakes. *Environ. Microbiol.* **6**, 1228–1243 (2004).

7. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *J. Nat. Prod.* **79**, 629–61 (2016).

8. Huang, K.-C., Chen, Z., Jiang, Y., Akare, S., Kolber-Simonds, D., Condon, K., Agoulnik, S., Tendyke, K., Shen, Y., Wu, K.-M., Mathieu, S., Choi, H. -w., Zhu, X., Shimizu, H., Kotake, Y., Gerwick, W. H., Uenaka, T., Woodall-Jappe, M. & Nomoto, K. Apratoxin A Shows Novel Pancreas Targeting Activity Through The Binding of Sec61. *Mol. Cancer Ther.* **15**, 1–10 (2016).

9. Luesch, H., Moore, R. E., Paul, V. J., Mooberry, S. L. & Corbett, T. H. Isolation of dolastatin 10 from the marine cyanobacterium Symploca species VP642 and total stereochemistry and biological evaluation of its analogue symplostatin 1. *J. Nat. Prod.* **64**, 907–910 (2001).

10. Doroghazi, J. R., Albright, J. C., Goering, A. W., Ju, K., Haines, R. R., Tchalukov, K. A., Labeda, D. P., Kelleher, N. L. & Metcalf, W. W. A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat. Chem. Biol.* **10**, 963–8 (2014).

11. Duncan, K. R., Crüsemann, M., Lechner, A., Sarkar, A., Li, J., Ziemert, N., Wang, M., Bandeira, N., Moore, B. S., Dorrestein, P. C. & Jensen, P. R. Molecular Networking and Pattern-Based Genome Mining Improves Discovery of Biosynthetic Gene Clusters and their Products from Salinispora Species. *Chem. Biol.* 460–471 (2015). doi:10.1016/j.chembiol.2015.03.010

12. Ziemert, N., Alanjary, M. & Weber, T. The evolution of genome mining in microbes – a review. *Nat. Prod. Rep.* **0**, 1–18 (2016).

13. Charlop-Powers, Z., Owen, J. G., Reddy, B. V. B., Ternei, M., Guimaraes, D. O., De Frias, U. A., Pupo, M. T., Seepe, P., Feng, Z. & Brady, S. F. Global biogeographic sampling of bacterial secondary metabolism. *Elife* **2015**, 1–10 (2015).

14. Lok, C. Mining the microbial dark matter. *Nature* **522**, 270–273 (2015).

15. Luo, Y., Cobb, R. E. & Zhao, H. Recent advances in natural product discovery. *Curr. Opin. Biotechnol.* **30**, 230–237 (2014).

16. Challis, G. L. & Ravel, J. Coelichelin, a new peptide siderophore encoded by the *Streptomyces coelicolor* genome: structure prediction from the sequence of its non-ribosomal peptide synthetase. *FEMS Microbiol. Lett.* **187**, 111–114 (2000).

17. Gross, H., Stockwell, V. O., Henkels, M. D., Nowak-Thompson, B., Loper, J. E. & Gerwick, W. H. The Genomisotopic Approach: A Systematic Method to Isolate Products of Orphan Biosynthetic Gene Clusters. *Chem. Biol.* **14**, 53–63 (2007).

18. Kim, J. H., Feng, Z., Bauer, J. D., Kallifidas, D., Calle, P. Y. & Brady, S. F. Cloning large natural product gene clusters from the environment: Piecing environmental DNA gene clusters back together with TAR. *Biopolymers* **93**, 833–844 (2010).

19. Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A. & Smith, H. O. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).

20. Kang, H.-S., Charlop-Powers, Z. & Brady, S. F. Multiplexed CRISPR/Cas9 and TAR-mediated promoter engineering of natural product biosynthetic gene clusters in yeast. *ACS Synth. Biol.* acssynbio.6b00080 (2016). doi:10.1021/acssynbio.6b00080

21. Bachmann, B. O., Van Lanen, S. G. & Baltz, R. H. Microbial genome mining for accelerated natural products discovery: Is a renaissance in the making? *J. Ind. Microbiol. Biotechnol.* **41**, 175–184 (2014).

22. Kleigrewe, K., Almaliti, J., Tian, I. Y., Kinnel, R. B., Korobeynikov, A., Monroe, E. A., Duggan, B. M., Di Marzo, V., Sherman, D. H., Dorrestein, P. C., Gerwick, L. & Gerwick, W. H. Combining Mass Spectrometric Metabolic Profiling with Genomic Analysis: A Powerful Approach for Discovering Natural Products from Cyanobacteria. *J. Nat. Prod.* **78**, 1671–1682 (2015).

23. Moss, N. A., Seiler, G., Leão, T. F., Castro-Falcón, G., Gerwick, L., Hughes, C. C. & Gerwick, W. H. Nature's combinatorial biosynthesis produces vatiamides A-F. *Angew. Chemie Int. Ed.* (2019). In Press. doi:10.1002/anie.201902571

24. Kleigrewe, K., Gerwick, L., Sherman, D. H. & Gerwick, W. H. Unique marine derived cyanobacterial biosynthetic genes for chemical diversity. *Nat. Prod. Rep.* **0**, 1–17 (2016).

25. Komarek, J., Kastovsky, J., Mares, J. & Johansen, J. R. Taxonomic classification of cyanoprokaryotes (cyanobacterial genera) 2014, using a polyphasic approach. *Preslia* **86**, 295–335 (2014).

26. Dittmann, E., Gugger, M., Sivonen, K. & Fewer, D. P. Natural Product Biosynthetic Diversity and Comparative Genomics of the Cyanobacteria. *Trends Microbiol.* **23**, 642–652 (2015).

27. W.H. Gerwick, M.A. Roberts, P.J. Proteau, J.-L. C. Structure of curacin A, a novel antimitotic, antiproliferative, and brine shrimp toxic natural product from the marine cyanobacterium Lyngbya majuscula. *J. Org. Chem.* **59**, 1243–1245

28. Taori, K., Paul, V. J. & Luesch, H. Structure and Activity of Largazole , a Potent Antiproliferative Agent from the Floridian Marine Cyanobacterium Symploca sp . 1806–1807 (2008). doi:10.1021/ja7110064

29. Fujii, K., Sivonen, K., Kashiwagi, T. & Hirayama, K. Nostophycin , a Novel Cyclic Peptide from the Toxic Cyanobacterium Nostoc sp . 152. 5777–5782 (1999). doi:10.1021/jo982306i

30. Yu, D., Xu, F., Zeng, J. & Zhan, J. Critical Review Type III Polyketide Synthases in Natural Product Biosynthesis. **64**, 285–295 (2012).

31. Strieker, M. & Tanovic, A. Nonribosomal peptide synthetases : structures and dynamics ´ and Mohamed A Marahiel. (2010). doi:10.1016/j.sbi.2010.01.009

32. Sudek, S., Haygood, M. G., Youssef, D. T. A. & Schmidt, E. W. Structure of trichamide, a cyclic peptide from the bloom-forming cyanobacterium Tnchodesmium erythraeum, predicted from the genome sequence. *Appl. Environ. Microbiol.* **72**, 4382–4387 (2006).

33. Li, B., Sher, D., Kelly, L., Shi, Y., Huang, K., Knerr, P. J., Joewono, I., Rusch, D., Chisholm, S. W. & van der Donk, W. A. Catalytic promiscuity in the biosynthesis of cyclic peptide secondary metabolites in planktonic marine cyanobacteria. *Proc. Natl. Acad. Sci. USA* **107**, 10430–5 (2010).

34. Hillwig, M. L., Fuhrman, H. A., Ittiamornkul, K., Sevco, T. J., Kwak, D. H. & Liu, X. Identification and characterization of a welwitindolinone alkaloid biosynthetic gene cluster in the stigonematalean cyanobacterium Hapalosiphon welwitschii. *ChemBioChem* **15**, 665–669 (2014).

35. Hillwig, M. L. & Liu, X. A new family of iron-dependent halogenases acts on freestanding substrates. *Nat. Chem. Biol.* **10**, 6–10 (2014).

36. Sinha, R. P., Klisch, M., Walter Helbling, E. & Häder, D. P. Induction of mycosporine-like amino acids (MAAs) in cyanobacteria by solar ultraviolet-B radiation. *J. Photochem. Photobiol. B Biol.* **60**, 129–135 (2001).

37. Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. a, Weber, T., Takano, E. & Breitling, R. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **39**, W339-46 (2011).

38. Kersten, R. D., Yang, Y.-L., Xu, Y., Cimermancic, P., Nam, S.-J., Fenical, W., Fischbach, M. A., Moore, B. S. & Dorrestein, P. C. A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat. Chem. Biol.* **7**, 794–802 (2011).

39. Dejong, C. A., Chen, G. M., Li, H., Johnston, C. W., Edwards, M. R., Rees, P. N., Skinnider, M. A., Webster, A. L. H. & Magarvey, N. A. Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. *Nat. Chem. Biol.* (2016). doi:10.1038/nchembio.2188

40. Wang, M., Carver, J. J., Phelan, V. V, Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A. V, Meehan, M. J., Liu, W.-T., Crüsemann, M., Boudreau, P. D., Esquenazi, E., Sandoval-Calderón, M., Kersten, R. D., Pace, L. A., Quinn, R. A., Duncan, K. R., Hsu, C.-C., Floros, D. J., Gavilan, R. G., Kleigrewe, K., Northen, T., Dutton, R. J., Parrot, D., Carlson, E. E., Aigle, B., Michelsen, C. F., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B. T., Gerwick, L., Liaw, C.-C., Yang, Y.-L., Humpf, H.-U., Maansson, M., Keyzers, R. A., Sims, A. C., Johnson, A. R., Sidebottom, A. M., Sedio, B. E., Klitgaard, A., Larson, C. B., Boya P, C. A., Torres-Mendoza, D., Gonzalez, D. J., Silva, D. B., Marques, L. M., Demarque, D. P., Pociute, E., O'Neill, E. C., Briand, E.,

Helfrich, E. J. N., Granatosky, E. A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J. J., Zeng, Y., Vorholt, J. A., Kurita, K. L., Charusanti, P., McPhail, K. L., Nielsen, K. F., Vuong, L., Elfeki, M., Traxler, M. F., Engene, N., Koyama, N., Vining, O. B., Baric, R., Silva, R. R., Mascuch, S. J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P. G., Dai, J., Neupane, R., Gurr, J., Rodríguez, A. M. C., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B. M., Almaliti, J., Allard, P.-M., Phapale, P., Nothias, L.-F., Alexandrov, T., Litaudon, M., Wolfender, J.-L., Kyle, J. E., Metz, T. O., Peryea, T., Nguyen, D.-T., VanLeer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K. M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P. R., Palsson, B. Ø., Pogliano, K., Linington, R. G., Gutiérrez, M., Lopes, N. P., Gerwick, W. H., Moore, B. S., Dorrestein, P. C. & Bandeira, N. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).

41. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. 2498–2504 (2003). doi:10.1101/gr.1239303.metabolite

42. Ã, A. E. H., Heyduck-s, B. & Fischer, U. Phylogenetic classification of heterotrophic bacteria associated with filamentous marine cyanobacteria in culture. **32**, 256–265 (2009).

43. Brauer, V. S., Stomp, M., Bouvier, T., Fouilland, E., Leboulanger, C., Confurius-Guns, V., Weissing, F. J., Stal, L. & Huisman, J. Competition and facilitation between the marine nitrogen-fixing cyanobacterium Cyanothece and its associated bacterial community. *Front. Microbiol.* **5**, 1–14 (2015).

44. Shi, L., Cai, Y., Yang, H., Xing, P., Li, P., Kong, L. & Kong, F. Phylogenetic diversity and specificity of bacteria associated with Microcystis aeruginosa and other cyanobacteria. *J. Environ. Sci.* **21**, 1581–1590 (2009).

45. Tuomainen, J., Hietanen, S., Kuparinen, J., Martikainen, P. J. & Servomaa, K. Community structure of the bacteria associated with Nodularia sp. (cyanobacteria) aggregates in the Baltic Sea. *Microb. Ecol.* **52**, 513–522 (2006).

46. Van Hannen, E. J., Zwart, G., Agterveld, M. P. V. A. N., Gons, H. J., Ebert, J. & Laanbroek, H. J. Changes in Bacterial and Eukaryotic Community Structure after Mass Lysis of Filamentous Cyanobacteria Associated with Viruses. **65**, 795–801 (1999).

47. Salomon, P. S., Janson, S. & Granéli, E. Molecular identification of bacteria associated with filaments of Nodularia spumigena and their effect on the cyanobacterial growth. *Harmful Algae* **2**, 261–272 (2003).

48. Berg, K. a, Lyra, C., Sivonen, K., Paulin, L., Suomalainen, S., Tuomi, P. & Rapala, J. High diversity of cultivable heterotrophic bacteria in association with cyanobacterial water blooms. *ISME J.* **3**, 314–325 (2009).

49. Beliaev, A. S., Romine, M. F., Serres, M., Bernstein, H. C., Linggi, B. E., Markillie, L. M., Isern, N. G., Chrisler, W. B., Kucek, L. a, Hill, E. a, Pinchuk, G. E., Bryant, D. a, Steven Wiley, H., Fredrickson, J. K. & Konopka, A. Inference of interactions in cyanobacterial-heterotrophic co-cultures via transcriptome sequencing. *ISME J.* **8**, 1–13 (2014).

50. Wiegand, C. & Pflugmacher, S. Ecotoxicological effects of selected cyanobacterial

secondary metabolites a short review. **203**, 201–218 (2005).

51. Gerwick, W. H. & Moore, B. S. Lessons from the past and charting the future of marine natural products drug discovery and chemical biology. *Chemistry and Biology* **19**, 85–98 (2012).

52. Jones, A. C., Gu, L., Sorrels, C. M., Sherman, D. H. & Gerwick, W. H. New tricks from ancient algae: natural products biosynthesis in marine cyanobacteria. *Curr. Opin. Chem. Biol.* **13**, 216–223 (2009).

53. Hirota Fujiki & Mooret, R. E. Indole alkaloids : Dihydroteleocidin B, teleocidin, and lyngbyatoxin A as members of a new class of tumor promoters. **78**, 3872–3876 (1981).

54. Dziallas, C. & Grossart, H.-P. Microbial interactions with the cyanobacterium Microcystis aeruginosa and their dependence on temperature. *Mar. Biol.* **159**, 2389–2398 (2012).

55. Edwards, D. J. & Gerwick, W. H. Lyngbyatoxin biosynthesis: sequence of biosynthetic gene cluster and identification of a novel aromatic prenyltransferase. *J. Am. Chem. Soc.* **126**, 11432–3 (2004).

56. Graber, M. A. & Gerwick, W. H. Kalkipyrone , a Toxic γ -Pyrone from an Assemblage of the Marine Cyanobacteria Lyngbya majuscula and Tolypothrix sp . **16**, 677–680 (1998).

57. Surup, F., Wagner, O., Frieling, J. Von, Schleicher, M., Oess, S. & Grond, S. The Iromycins , a New Family of Pyridone Metabolites from Streptomyces sp . I . Structure , NOS Inhibitory Activity , and Biosynthesis. 5085–5090 (2007). doi:10.1021/jo0703303

58. Bewley, C. A., Holland, N. & Faulkner, D. J. Two classes of metabolites from Theonella swinhoei are localized in distinct populations of bacterial symbionts. 716–722 (1996).

59. Andrianasolo, E. H., Gross, H., Goeger, D., Musafija-girt, M., Mcphail, K., Leal, R. M., Mooberry, S. L. & Gerwick, W. H. Isolation of Swinholide A and Related Glycosylated Derivatives from Two Field Collections of Marine Cyanobacteria. 6225–6228 (2005). doi:10.1021/ol050188x

60. Ueoka, R., Uria, A. R., Reiter, S., Mori, T., Karbaum, P., Peters, E. E., Helfrich, E. J. N., Morinaka, B. I., Gugger, M., Takeyama, H., Matsunaga, S. & Piel, J. Metabolic and evolutionary origin of actin-binding polyketides from diverse organisms. *Nat. Chem. Biol.* **11**, 705–712 (2015).

61. Engene, N., Rottacker, E. C., Kaštovský, J., Byrum, T., Choi, H., Ellisman, M. H., Komárek, J. & Gerwick, W. H. 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 (2012).

62. Chang, Z., Sitachitta, N., Rossi, J. V., Roberts, M. A., Flatt, P. M., Jia, J., Sherman, D. H. & Gerwick, W. H. Biosynthetic pathway and gene cluster analysis of curacin A, an antitubulin natural product from the tropical marine cyanobacterium Lyngbya majuscula. *J. Nat. Prod.* **67**, 1356–1367 (2004).

63. Flatt, P. M., O'Connell, S. J., McPhail, K. L., Zeller, G., Willis, C. L., Sherman, D. H. & Gerwick, W. H. Characterization of the initial enzymatic steps of barbamide biosynthesis. *J. Nat. Prod.* **69**, 938–44 (2006).

64. Ramaswamy, A. V, Sorrels, C. M. & Gerwick, W. H. Cloning and Biochemical Characterization of the Hectochlorin Biosynthetic Gene Cluster from the Marine

Cyanobacterium Lyngbya majuscula. 1977–1986 (2007).

65. Edwards, D. J., Marquez, B. L., Nogle, L. M., McPhail, K., Goeger, D. E., Roberts, M. A. & Gerwick, W. H. Structure and biosynthesis of the jamaicamides, new mixed polyketide-peptide neurotoxins from the marine cyanobacterium Lyngbya majuscula. *Chem. Biol.* **11**, 817–33 (2004).

66. Sitachitta, N., M??rquez, B. L., Thomas Williamson, R., Rossi, J., Ann Roberts, M., Gerwick, W. H., Nguyen, V. A. & Willis, C. L. Biosynthetic pathway and origin of the chlorinated methyl group in barbamide and dechlorobarbamide, metabolites from the marine cyanobacterium Lyngbya majuscula. *Tetrahedron* **56**, 9103–9113 (2000).

67. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. a., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. a. & Pevzner, P. a. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

68. Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F. & Stevens, R. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **42**, D206-14 (2014).

69. Blin, K., Medema, M. H., Kazempour, D., Fischbach, M. a, Breitling, R., Takano, E. & Weber, T. antiSMASH 2.0--a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res.* **41**, W204-12 (2013).

70. Podell, S. & Gaasterland, T. DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol.* **8**, R16 (2007).

71. Wang, Q., Garrity, G. M., Tiedje, J. M., Cole, J. R. & Al, W. E. T. Naı ¨ ve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy □ †. **73**, 5261–5267 (2007).

72. Nubel, U., Garcia-pichel, F. & Muyzer, G. PCR Primers To Amplify 16S rRNA Genes from Cyanobacteria. **63**, 3327–3332 (1997).

73. Romero, H., Zhang, Y., Gladyshev, V. N. & Salinas, G. Evolution of selenium utilization traits. **6**, (2005).

74. Hayashi, T., Kitamura, Y., Funa, N. & Ohnishi, Y. Fatty Acyl-AMP Ligase Involvement in the Production of Alkylresorcylic Acid by a Myxococcus xanthus Type III Polyketide Synthase. 2166–2176 (2011). doi:10.1002/cbic.201100344

75. Barns, S. M., Cain, E. C., Sommerville, L. & Kuske, C. R. Acidobacteria Phylum Sequences in Uranium-Contaminated Subsurface Sediments Greatly Expand the Known Diversity within the Phylum □ †. **73**, 3113–3116 (2007).

76. Navarrete, A. A., Kuramae, E. E., Hollander, M. De, Pijl, A. S., Veen, J. A. Van & Tsai, S. M. Acidobacterial community responses to agricultural management of soybean in Amazon forest soils. **83**, 607–621 (2013).

77. Webster, N. S., Luter, H. M., Soo, R. M., Botté, E. S., Simister, R. L., Abdo, D. & Whalan, S. Same , same but different : symbiotic bacterial associations in GBR sponges. **3**, 1–11 (2013).

78. Connor-sánchez, A. O., Rivera-domínguez, A. J., Santos-briones, C. D. L., López-aguiar, L. K., Peña-ramírez, Y. J. & Prieto-davo, A. Acidobacteria appear to dominate the microbiome of two sympatric Caribbean Sponges and one Zoanthid. 1–6 (2014).

79. Jones, A. C., Monroe, E. a, Podell, S., Hess, W. R., Klages, S., Esquenazi, E., Niessen, S., Hoover, H., Rothmann, M., Lasken, R. S., Yates, J. R., Reinhardt, R., Kube, M., Burkart, M. D., Allen, E. E., Dorrestein, P. C., Gerwick, W. H. & Gerwick, L. Genomic insights into the physiology and ecology of the marine filamentous cyanobacterium Lyngbya majuscula. *Proc. Natl. Acad. Sci. USA* **108**, 8815–8820 (2011).

80. Llamas, M. a., Imperi, F., Visca, P. & Lamont, I. L. Cell-surface signaling in Pseudomonas: Stress responses, iron transport, and pathogenicity. *FEMS Microbiol. Rev.* **38**, 569–597 (2014).

81. Moran, N. A. & Plague, G. R. Genomic changes following host restriction in bacteria. doi:10.1016/j.gde.2004.09.003

82. Lackner, G., Moebius, N., Partida-martinez, L. P., Boland, S. & Hertweck, C. Evolution of an endofungal Lifestyle : Deductions from the Burkholderia rhizoxinica Genome. (2011).

83. Aziz, R. K., Breitbart, M. & Edwards, R. a. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* **38**, 4207–4217 (2010).

84. Stucken, K., John, U., Cembella, A., Murillo, A. A., Soto-Liebe, K., Fuentes-Valdés, J. J., Friedel, M., Plominsky, A. M., Vásquez, M. & Glöckner, G. The smallest known genomes of multicellular and toxic cyanobacteria: Comparison, minimal gene sets for linked traits and the evolutionary implications. *PLoS One* **5**, e9235 (2010).

85. Shih, P. M., Wu, D., Latifi, A., Axen, S. D., Fewer, D. P., Talla, E., Calteau, A., Cai, F., Tandeau de Marsac, N., Rippka, R., Herdman, M., Sivonen, K., Coursin, T., Laurent, T., Goodwin, L., Nolan, M., Davenport, K. W., Han, C. S., Rubin, E. M., Eisen, J. a, Woyke, T., Gugger, M. & Kerfeld, C. a. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc. Natl. Acad. Sci. USA* **110**, 1053–8 (2013).

86. Zehr, J. P. Nitrogen fixation by marine cyanobacteria. *Trends Microbiol.* **19**, 162–173 (2011).

87. Moss, N. A., Bertin, M. J., Kleigrewe, K., Leão, T. F., Gerwick, L. & Gerwick, W. H. Integrating mass spectrometry and genomics for cyanobacterial metabolite discovery. *J. Ind. Microbiol. Biotechnol.* 313–324 (2015).

88. Kleigrewe, K., Almaliti, J., Tian, I. Y., Kinnel, R. B., Korobeynikov, A., Monroe, E. A., Duggan, B. M., Di Marzo, V., Sherman, D. H., Dorrestein, P. C., Gerwick, L. & Gerwick, W. H. Combining Mass Spectrometric Metabolic Profiling with Genomic Analysis: A Powerful Approach for Discovering Natural Products from Cyanobacteria. *J. Nat. Prod.* **78**, 1671–1682 (2015).

89. Taniguchi, M., Nunnery, J. K., Engene, N., Esquenazi, E., Byrum, T., Dorrestein, P. C. & Gerwick, W. H. Palmyramide a, a cyclic depsipeptide from a palmyra atoll collection of the marine cyanobacterium lyngbya majuscula. *J. Nat. Prod.* **73**, 393–398 (2010).

90. Marquez, B. L., Watts, K. S., Yokochi, A., Roberts, M. A., Verdier-Pinard, P., Jimenez, J. I., Hamel, E., Scheuer, P. J. & Gerwick, W. H. Structure and absolute stereochemistry of hectochlorin, a potent stimulator of actin assembly. *J. Nat. Prod.* **65**, 866–871 (2002).

91. Grindberg, R. V, Ishoey, T., Brinza, D., Esquenazi, E., Coates, R. C., Liu, W., Gerwick, L., Dorrestein, P. C., Pevzner, P., Lasken, R. & Gerwick, W. H. Single cell genome amplification accelerates identification of the apratoxin biosynthetic pathway from a complex microbial assemblage. *PLoS One* **6**, e18565 (2011).

92. Rippka, R., Deruelles, J., Waterbury, J. B., Herdman, M. & Stanier, R. Y. Generic Assignments, Strain Histories and Properties of Pure Cultures of Cyanobacteria. *J. Gen. Microbiol.* **111**, 1–61 (1979).

93. Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W. & Nielsen, P. H. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–8 (2013).

94. Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C. & Salzberg, S. L. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).

95. Alikhan, N.-F., Petty, N. K., Ben Zakour, N. L. & Beatson, S. a. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* **12**, 402 (2011).

96. Medema, M. H., Takano, E. & Breitling, R. Detecting Sequence Homology at the Gene Cluster Level with MultiGeneBlast. **30**, 1218–1223 (2013).

97. Calteau, A., Fewer, D. P., Latifi, A., Coursin, T., Laurent, T., Jokela, J., Kerfeld, C. a, Sivonen, K., Piel, J. & Gugger, M. Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. *BMC Genomics* **15**, 977 (2014).

98. Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., Lee, S. Y., Fischbach, M. a., Muller, R., Wohlleben, W., Breitling, R., Takano, E. & Medema, M. H. antiSMASH 3.0--a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* 1–7 (2015). doi:10.1093/nar/gkv437

99. Nurk, S., Bankevich, A., Antipov, D., Gurevich, A. A., Korobeynikov, A., Lapidus, A., Prjibelski, A. D., Pyshkin, A., Sirotkin, A., Sirotkin, Y., Stepanauskas, R., Clingenpeel, S. R., Woyke, T., McLean, J. S., Lasken, R., Tesler, G., Alekseyev, M. A. & Pevzner, P. A. Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.* **20**, 714–737 (2013).

100. Utturkar, S. M., Klingeman, D. M., Land, M. L., Schadt, C. W., Doktycz, M. J., Pelletier, D. A. & Brown, S. D. Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics* **30**, 2709–2716 (2014).

101. Boetzer, M. & Pirovano, W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15**, 211 (2014).

102. Pop, M., Phillippy, A. & Delcher, A. . Comparative genome assembly. *Bioinformatics* **5**, 237–248 (2004).

103. Galardini, M., Biondi, E. G., Bazzicalupo, M. & Mengoni, A. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source Code Biol. Med.* **6**, 11 (2011).

104. Ziemert, N., Lechner, A., Wietz, M., Millán-Aguiñaga, N., Chavarria, K. L. & Jensen, P. R. Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. *Proc. Natl. Acad. Sci. USA* **111**, E1130-9 (2014).

105. Campbell, E. L., Cohen, M. F. & Meeks, J. C. A polyketide-synthase-like gene is involved in the synthesis of heterocyst glycolipids in Nostoc punctiforme strain ATCC 29133. *Arch. Microbiol.* **167**, 251–258 (1997).

106. Zhang, C. C., Laurent, S., Sakr, S., Peng, L. & Bédu, S. Heterocyst differentiation and pattern formation in cyanobacteria: A chorus of signals. *Mol. Microbiol.* **59**, 367–375 (2006).

107. Lee, H.-M., Vazquez-Bermudez, M. F. & Tandeau de Marsac, N. The Global Nitrogen Regulator NtcA Regulates Transcription of the Signal Transducer P II (GlnB) and Influences Its Phosphorylation Level in Response to Nitrogen and Carbon Supplies in the. *J. Bacteriol.* **181**, 2697–2702 (1999).

108. Berg, H., Ziegler, K., Piotukh, K., Baier, K., Lockau, W. & Volkmer-Engert, R. Biosynthesis of the cyanobacterial reserve polymer multi- L -arginyl-poly- L -aspartic acid (cyanophycin). *J. Biochem.* **267**, 5561–5570 (2000).

109. Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Wieland Brown, L. C., Mavrommatis, K., Pati, A., Godfrey, P. A., Koehrsen, M., Clardy, J., Birren, B. W., Takano, E., Sali, A., Linington, R. G. & Fischbach, M. A. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).

110. Medema, M. H. & Fischbach, M. a. Computational approaches to natural product discovery. *Nat. Chem. Biol.* **11**, 639–648 (2015).

111. Blin, K., Kim, H. U., Medema, M. H. & Weber, T. Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. 1–11 (2017). doi:10.1093/bib/bbx146

112. Pye, C. R., Bertin, M. J., Lokey, R. S., Gerwick, W. H. & Linington, R. G. Retrospective analysis of natural products provides insights for future discovery trends. *Proc. Natl. Acad. Sci.* 201614680 (2017). doi:10.1073/pnas.1614680114

113. Keller, L., Leão, T. & Gerwick, W. H. Chemical Biology of Marine Cyanobacteria. in *Chemical Biology of Natural Products* 43–87 (2017).

114. Alvarenga, D. O., Fiore, M. F. & Varani, A. M. A Metagenomic Approach to Cyanobacterial Genomics. *Front. Microbiol.* **8**, 1–16 (2017).

115. Moss, N., Leao, T., Glukhov, E., Gerwick, L. & Gerwick, W. H. Collection, Culturing, and Genome Analyses of Tropical Marine Filamentous Benthic Cyanobacteria. **xx**, 1–44 (2018).

116. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM : assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

117. Navarro-muñoz, J. C., Selem-mojica, N., Mullowney, M. W., Kautsar, S., Abubucker, S., Roeters, A., Lokhorst, W., Fernandez-guerra, A., Barona-gomez, F. & Medema, M. H. A computational framework for systematic exploration of biosynthetic diversity from large-scale genomic data. *BioRxiv* (2018).

118. Nathan A. Moss, Tiago Leao, Michael R. Rankin, Tyler M. McCullough, Pingping Qu, Anton Korobeynikov, Janet L. Smith, L. G. and W. H. G. Ketoreductase Domain Dysfunction Expands Chemodiversity: Malyngamide Biosynthesis in the Cyanobacterium

Okeania hirsuta. (2018). doi:10.1021/acschembio.8b00910

119. Engene, N., Paul, V. J., Byrum, T., Gerwick, W. H., Thor, A. & Ellisman, M. H. Five chemically rich species of tropical marine cyanobacteria of the genus Okeania gen. nov. (Oscillatoriales, Cyanoprokaryota). *J. Phycol.* **49**, 1095–1106 (2013).

120. Luzzatto-Knaan, T., Garg, N., Wang, M., Glukhov, E., Peng, Y., Ackermann, G., Amir, A., Duggan, B. M., Ryazanov, S., Gerwick, L., Knight, R., Alexandrov, T., Bandeira, N., Gerwick, W. H., Dorrestein, P. C., Dorrestein, P., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R., Scalbert, A., Hsu, C., Floros, D., Gavilan, R., Kleigrewe, K., Northen, T., Dutton, R., Parrot, D., Carlson, E., Aigle, B., Michelsen, C., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B., Gerwick, L., Liaw, C., Yang, Y., Humpf, H., Maansson, M., Keyzers, R., Sims, A., Johnson, A., Sidebottom, A., Sedio, B., Klitgaard, A., Larson, C., P, C. B., Torres-Mendoza, D., Gonzalez, D., Silva, D., Marques, L., Demarque, D., Pociute, E., O'Neill, E., Briand, E., Helfrich, E., Granatosky, E., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J., Zeng, Y., Vorholt, J., Kurita, K., Charusanti, P., McPhail, K., Nielsen, K., Vuong, L., Elfeki, M., Traxler, M., Engene, N., Koyama, N., Vining, O., Baric, R., Silva, R., Mascuch, S., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P., Dai, J., Neupane, R., Gurr, J., Rodríguez, A., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B., Almaliti, J., Allard, P., Phapale, P., Nothias, L., Alexandrov, T., Litaudon, M., Wolfender, J., Kyle, J., Metz, T., Peryea, T., Nguyen, D., VanLeer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P., Palsson, B., Pogliano, K., Linington, R., Gutiérrez, M., Lopes, N., Gerwick, W., Moore, B., Dorrestein, P., Bandeira, N., Seipke, R., Shen, B., Sherman, D., Sivonen, K., Smanski, M., Sosio, M., Stegmann, E., Süssmuth, R., Tahlan, K., Thomas, C., Tang, Y., Truman, A., Viaud, M., Walton, J., Walsh, C., Weber, T., Wezel, G. van, Wilkinson, B., Willey, J., Wohlleben, W., Wright, G., Ziemert, N., Zhang, C., Zotchev, S., Breitling, R., Takano, E. & Glöckner, F. Digitizing mass spectrometry data to explore the chemical diversity and distribution of marine cyanobacteria and algae. *Elife* **6**, 1686–1699 (2017).

121. Mukherjee, S., Seshadri, R., Varghese, N. J., Eloe-Fadrosh, E. A., Meier-Kolthoff, J. P., Göker, M., Coates, R. C., Hadjithomas, M., Pavlopoulos, G. A., Paez-Espino, D., Yoshikuni, Y., Visel, A., Whitman, W. B., Garrity, G. M., Eisen, J. A., Hugenholtz, P., Pati, A., Ivanova, N. N., Woyke, T., Klenk, H. P. & Kyrpides, N. C. 1,003 Reference Genomes of Bacterial and Archaeal Isolates Expand Coverage of the Tree of Life. *Nat. Biotechnol.* **35**, 676–683 (2017).

122. Hadjithomas, M., Chen, I. A., Chu, K., Huang, J., Ratner, A., Palaniappan, K., Andersen, E., Markowitz, V., Kyrpides, N. C. & Ivanova, N. IMG-ABC : new features for bacterial secondary metabolism analysis and targeted biosynthetic gene cluster discovery in thousands of microbial genomes. **45**, 560–565 (2017).