Title

Comparative genomics uncovers the prolific and distinctive metabolic potential of the cyanobacterial genus Moorea.

Journal

Proceedings of the National Academy of Sciences, 114(12)

Authors

Leao, Tiago

Castelão, Guilherme

Korobeynikov, Anton

et al.

Publication Date

2017-03-21

DOI

10.1073/pnas.1618556114

Peer reviewed

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

Tiago Leao[a], Guilherme Castelão[b], Anton Korobeynikov[c,d], Emily A. Monroe[e], Sheila Podell[a], Evgenia Glukhov[a], Eric E. Allen[a], William H. Gerwick[a,f], and Lena Gerwick[a,1]

[a]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92093; [b]Climate, Atmospheric Sciences, and Physical Oceanography, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92093; [c]Department of Statistical Modelling, St. Petersburg State University, Saint Petersburg 198504, Russia; [d]Center for Algorithmic Biotechnology, St. Petersburg State University, Saint Petersburg 198504, Russia; [e]Department of Biology, William Paterson University, Wayne, NJ 07470; and [f]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093

Edited by Robert Haselkorn, University of Chicago, Chicago, IL, and approved February 6, 2017 (received for review November 11, 2016)

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 a complete genome of a filamentous tropical marine cyanobacterium, *Moorea producens* PAL, which reveals that about one-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. 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.

tropical marine cyanobacteria | genome comparison | biosynthetic gene clusters | heterocyst glycolipids | gene cluster network

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 (1, 2). 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 (3). Filamentous diazotrophic cyanobacteria from subsection VIII, such as *Nostoc* and *Anabaena*, fix nitrogen within specialized cells called heterocysts (4).

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 (5). One genus of cyanobacteria in particular, *Moorea*, has been an exceptionally rich source of novel bioactive NPs (6). This taxonomic group, previously identified as "marine *Lyngbya*" but recently reclassified on the basis of genetic data as *Moorea*, consists of large, nondiazotrophic filaments that are mostly found growing benthically in shallow tropical marine environments (7). 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 (8). 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) (9). Genomic analyses of these filamentous cyanobacteria have revealed that even well-studied strains possess additional genetic capacity to produce novel and chemically unique NPs (10), and suggest that bottom-up approaches (11) would be productive; a recent example is given by the discovery and description of the columbamides from *Moorea bouillonii* (12). 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 a 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

Significance

The genus *Moorea* has yielded more than 40% of all reported marine cyanobacterial natural products. Preliminary genomic data suggest that many more natural products are yet to be discovered. However, incomplete genomic information has hampered the discovery of novel compounds using genome-mining approaches. Here, we report a complete genome of a filamentous marine tropical cyanobacterium, *Moorea producens* PAL, along with the improvement of other three *Moorea* draft genomes. Our analyses revealed a vast and distinctive natural product metabolic potential in these strains, highlighting that they are still an excellent source of unique metabolites despite previous extensive studies.

Author contributions: T.L., W.H.G., and L.G. designed research; T.L. performed research; E.G. cultured organisms; T.L., G.C., A.K., E.A.M., and E.G. contributed new reagents/analytic tools; T.L., G.C., A.K., E.A.M., S.P., E.E.A., W.H.G., and L.G. analyzed data; and T.L., S.P., E.E.A., W.H.G., and L.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The genomes of PAL, JHB, 3L, and PNG have been deposited at DNA Data Bank of Japan/European Nucleotide Archive/GenBank (accession nos. GCA_001767235.1, GCA_000211815.1, MKZR00000000, and MKZS00000000, respectively).

[1]To whom correspondence should be addressed. Email: lgerwick@ucsd.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1618556114/-/DCSupplemental.

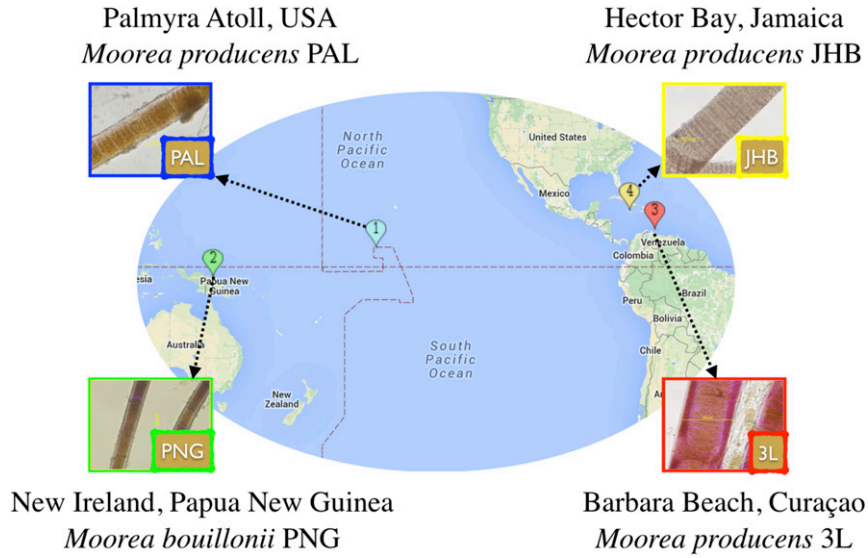

Fig. 1. Geographical location and microscopy images (using 40× magnification) of the four investigated *Moorea* strains.

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 NP pathways, providing support and raising expectations for future genome-guided isolation efforts.

Results and Discussion

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* (Fig. 1): *M. producens* PAL 15AUG08-1, *M. producens* JHB 22AUG96-1, *M. producens* NAK12DEC93-3L, 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 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 (7). The exterior of the sheath material is richly populated with various heterotrophic bacteria, some of which may exist in obligate commensal relationships (13). 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 1993, and it produces barbamide, dechlorobarbamide, carmabins A and B, curacins A–C, and curazole. These two Atlantic strains 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. (8), the compounds cited above are produced via enzymes encoded by unique biosynthetic genes, some of which 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.

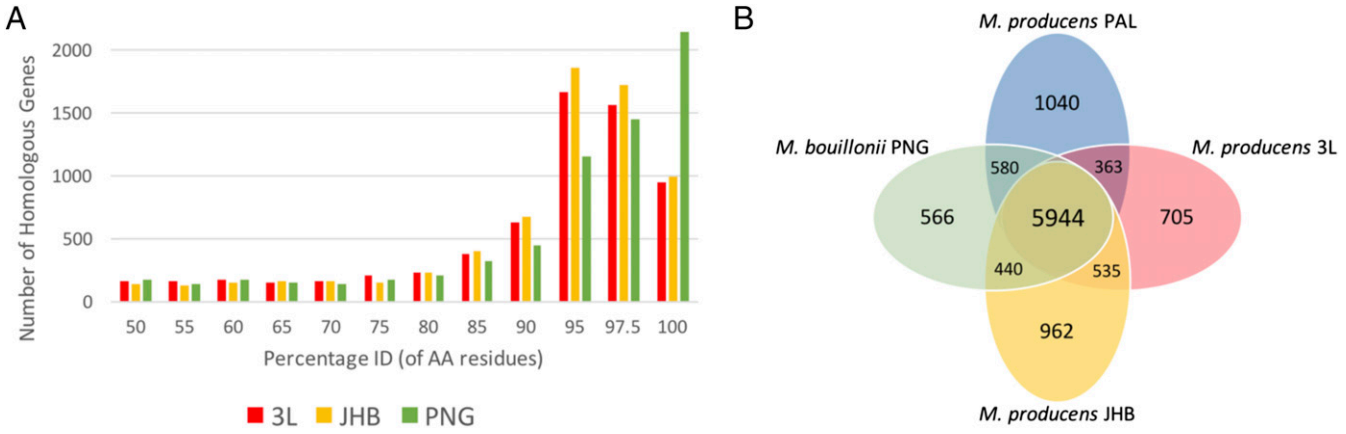

Fig. 2. (*A*) A histogram of percent amino acid identity for all shared homologous genes with the PAL genome (bidirectional best BLAST hit, minimum ID (identity) of 50%). (*B*) Venn diagram for the shared homologous genes and strain-specific genes among the four *Moorea* strains. AA, amino acid.

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.

The Use of Hybrid Assembly and Long-Reads Scaffolding to Obtain a 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 nonaxenic 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 (14), 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 used to improve overall draft genome quality; however, in this case, it was still fragmented because large repeated regions remained unresolved (15). To resolve these regions and close the genome, we developed an 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 (16). 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, a complete genome of a tropical filamentous marine cyanobacterium (Table S1). To assure that no cyanobacterial contigs were left out of the assembly, especially in light of the fact that the sequenced culture was nonaxenic, 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 1 for the circular plasmid) from the PAL genome were the only cyanobacterial contigs in the sample (confirming that the culture was monocyanobacterial). 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 α-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 (*Supporting Information*) (17, 18). In the case of *M. producens* JHB, this reference assembly procedure resulted in a linear chromosomal scaffold of 9.6 Mb

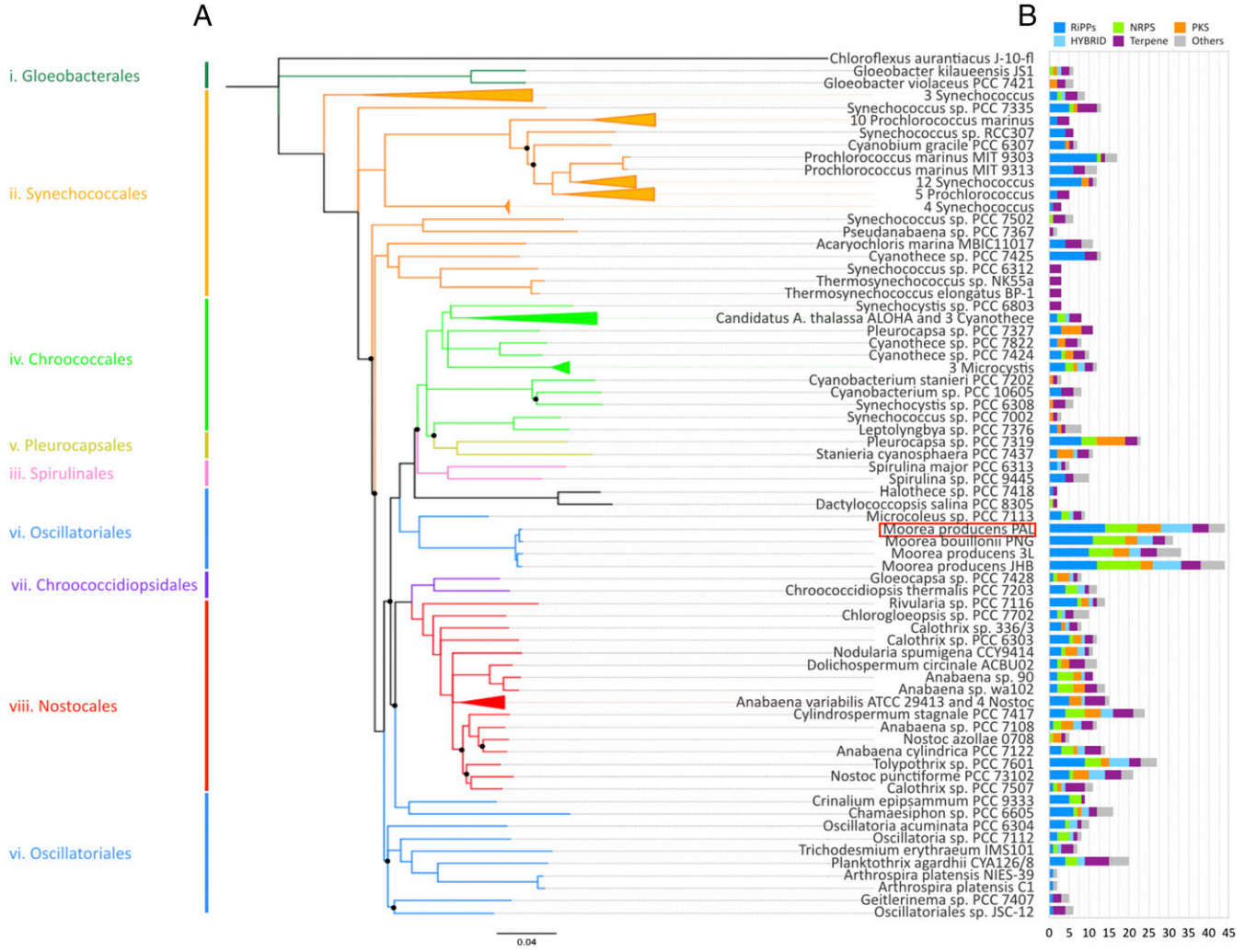

Fig. 3. (*A*) Phylogenomic analyses of completed cyanobacterial genomes using 29 conserved genes from Calteau et al. (19). Branches are colored according to cyanobacterial subsections (except by PCC 7418 and PCC 8305, which are not yet classified). All bootstrap values are higher than 85, except those 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.

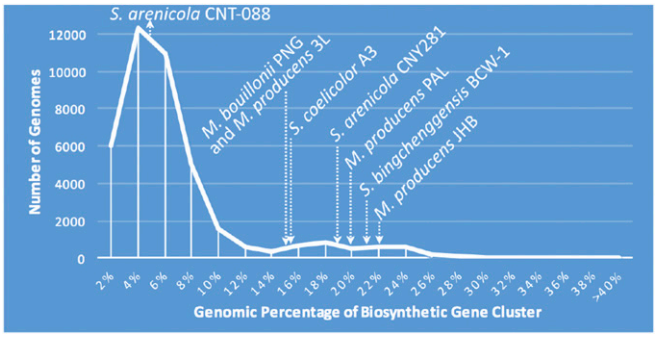

Fig. 4. Distribution of bacterial genomes from JGI/IMG database in terms of genomic percentage dedicated to secondary metabolism (NP biosynthesis). Several prolific NP producers are identified in the figure, including *Streptomyces coelicolor* A3, *Streptomyces bingchenggensis* BCW-1, and two *Salinispora* strains (highest and lowest genomic percentages from this genus). The total number of genomes interrogated was 40,532.

consisting of 205 contigs with ∼26,000 Ns that connect the contigs, along with two small plasmid scaffolds of 9.5 and 2 kb. The final draft genome of *M. bouillonii* PNG consisted of a linear chromosomal scaffold of 8.23 Mb (291 contigs and ∼32,000 Ns) and 12 unmapped scaffolds from 1.6 to 16.7 kb. The *M. producens* 3L final draft genome consisted of a linear chromosomal scaffold of 8.15 Mb (205 contigs and ∼20,000 Ns) and 78 unmapped scaffolds from 0.5 to 9.4 kb. Additional features of these four genomes are presented in Table S1.

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 Joint Genome Institute (JGI)/Integrated Microbial Genomes (IMG) database (total of 107 genomes, Dataset S1, worksheet 1)]. Our reference genome, *M. producens* PAL, contained all 195 housekeeping genes, reinforcing its completeness. The other three draft genomes were compared with 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 (missing genes listed on the Dataset S1, worksheet 2). 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 S1, such as GC content, number of genes, and percentage of annotated genes, are consistent with other cyanobacterial genomes (19).

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 km 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 (20). 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% (7). This is visualized as a circular map that compares the reference and draft genomes (Fig. S1) and bar graphs that depict the number and the percent identities between homologous genes in the different genomes (Fig. 2*A*). In Fig. S1, 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 (Fig. 2*A*). 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; Fig. 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. Last, the MUMmer plots in Fig. S2 indicate that these *Moorea* genomes are also highly syntenic with one another (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 5,944 homologous genes as identified by BLAST analysis (Fig. 2*B*). Therefore, only 8–13.5% 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 cluster of orthologous groups (COG) categories of all four genomes is remarkably similar (Table S2). 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 homologs 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), nondiazotrophic (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 cofactors including CoA, cobalamin, biotin, flavin, NAD, heme, and thiamine. Additionally, the number of

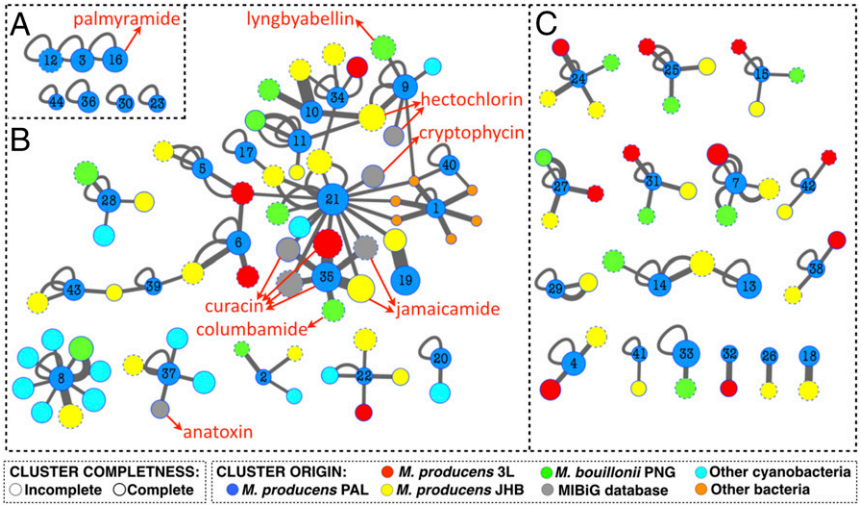

Fig. 5. Gene cluster networking of PAL versus gene clusters from PNG, 3L, JHB, the MiBIG database, completed cyanobacterial genomes from JGI/IMG, and their closest homologs from the National Center for Biotechnology Information (NCBI) database (according to antiSMASH results). *A* represents only orphan 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 subclusters. 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. For more information regarding *Moorea* clusters, see Dataset S2, worksheet 2 (numbers on nodes refer to tabulated data in Dataset S2).

MICROBIOLOGY

Table 1. Summary table listing number of known (K), "cryptic" (C), and "orphan" (O) NP pathways according to Fig. 5

	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.

specialized sigma factors in the genomes of these four filamentous marine cyanobacteria strains, as previously discussed in Jones et al. (21), are virtually the same (five 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 (see comparison of COG categories in *Supporting Information*).

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 (22, 23). These genes are commonly found in diazotrophic cyanobacteria from subsections VIII but are 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 VI, 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 Fig. S3, 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 one another in the 3L genome. Two new genes at this position that encode for hypothetical proteins have apparently replaced the *hgl* cluster in the 3L genome (red box in Fig. S3). Despite the presence of the *hgl* cluster, filaments cultured in nitrogen-deficient medium (up to 8 d 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) (24). The only regulatory homolog for heterocyst development located in *Moorea* was *hetR* (~70% nucleotide identity, located about 1.7–2.2 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 (Fig. S3) 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 study reports 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) (24). 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 unclustered 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 nondiazotrophic cyanobacteria (25). The fact that *Moorea* strains survive up to 8 d 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 (26). Of note, our genomic analysis revealed that each of the *Moorea* genomes contained one cyanophycin synthetase and at least one cyanophycinase gene.

Uncovering the Metabolic Potential of the Genus *Moorea*. A phylogenomic analysis (Fig. 3*A*) confirmed that these four *Moorea* strains are monophyletic, supporting the findings of high genomic synteny. However, based on phylogeny (Fig. 3*A*) and the occurrence of the *hgl* cluster, this genus may be misplaced within section VI of the cyanobacteria. Another highly prominent feature that distinguishes *Moorea* from other cyanobacteria (Fig. 3*B*) 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. Although *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 one-half this number. As such, *Moorea* spp. are "superproducers" among cyanobacteria, and on average 18% of their genome is dedicated to secondary metabolism (Table S1), nearly four times the average of other cyanobacteria (1). In comparison with all other bacterial genomes (Fig. 4), *Moorea* are among the most prolific producers of NPs with only some actinobacterial strains being more endowed (27). The discrepancy between our analyses and that performed previously by Jones et al. (21) on the draft genome of *M. producens* 3L is due to the fact that the BGC-mining tool antiSMASH (28) 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.

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 (29). Because 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 biocompass.net/. The output can be displayed as a network diagram using Cytoscape, version 3.2.1 (Fig. 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 nonadjacent

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 (Fig. 5 and Table 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 unique NPs. Interestingly, 26 of the PAL clusters (about 59%, Fig. 5*C*) only have homology to other *Moorea* pathways, confirming previous chemical investigations that indicated they possess a unique secondary metabolite profile compared with other bacteria (8). Moreover, these findings suggest that *M. producens* PAL is not only a source of unique NPs but that these NPs will likely be composed of unique 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 seven clusters, ~16%) (Fig. 5*A*).

As previously reported, accurate prediction of BGC borders is a common challenge for the field (27, 30). This issue can have an effect on the estimated percentage of the genome dedicated to NP 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% (Dataset S2, worksheet 2), confirming the validity of the relationships shown in Fig. 4. Further analyses of various features of *Moorea*'s BGCs (Dataset S2, worksheet 2), such as G+C content, few mobile elements within clusters, and encoding of relatively rare structural moieties (8), suggest that these strains have vertically acquired these biosynthetic pathways, consistent with previous reports for cyanobacteria (19). However, a larger sample size and better-characterized pathway products are needed to fully understand the evolution and distribution of *Moorea*'s NP pathways.

In summary, 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. Development of a reference genome for *M. producens* PAL has increased understanding of the genomic capacities of three related strains of filamentous cyanobacteria, providing fresh insights into this important source of NPs. Using gene cluster networking, we were able to demonstrate that many of the *Moorea* BGCs are rare among bacterial genomes, and suggest future directions for productive genome-guided isolation efforts of unique NPs from this genus.

Materials and Methods

See *SI Materials and Methods* for details of sampling, culturing methods, DNA extraction, sequencing, assembly, genome comparison, and other bioinformatic analyses.

ACKNOWLEDGMENTS. This research was supported by National Institutes of Health Grants CA108874 and GM107550 (to W.H.G. and L.G.) and by Russian Science Foundation Grant 14-50-00069 (to A.K.). We thank the CAPES Foundation for Research Fellowship 13425-13-7 (to T.L.).

1. Shih PM, et al. (2013) Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc Natl Acad Sci USA* 110(3):1053–1058.
2. Flores E, López-lozano A, Herrero A (2015) Nitrogen fixation in the oxygenic (cyanobacteria): The fight against oxygen. *Biol Nitrogen Fixat* 2:879–889.
3. Zehr JP (2011) Nitrogen fixation by marine cyanobacteria. *Trends Microbiol* 19(4): 162–173.
4. Komarek J, Kastovsky J, Mares J, Johansen JR (2014) Taxonomic classification of cyanoprokaryotes (cyanobacterial genera) 2014, using a polyphasic approach. *Preslia* 86(4):295–335.
5. Newman DJ, Cragg GM (2016) Natural products as sources of new drugs from 1981 to 2014. *J Nat Prod* 79(3):629–661.
6. Dittmann E, Gugger M, Sivonen K, Fewer DP (2015) Natural product biosynthetic diversity and comparative genomics of the cyanobacteria. *Trends Microbiol* 23(10):642–652.
7. Engene N, et al. (2012) *Moorea producens* gen. nov., sp. nov. and *Moorea bouillonii* comb. nov., tropical marine cyanobacteria rich in bioactive secondary metabolites. *Int J Syst Evol Microbiol* 62(Pt 5):1171–1178.
8. Kleigrewe K, Gerwick L, Sherman DH, Gerwick WH (2016) Unique marine derived cyanobacterial biosynthetic genes for chemical diversity. *Nat Prod Rep* 33(2):348–364.
9. Wang M, et al. (2016) Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat Biotechnol* 34(8):828–837.
10. Moss NA, et al. (2016) Integrating mass spectrometry and genomics for cyanobacterial metabolite discovery. *J Ind Microbiol Biotechnol* 43(2-3):313–324.
11. Luo Y, Cobb RE, Zhao H (2014) Recent advances in natural product discovery. *Curr Opin Biotechnol* 30:230–237.
12. Kleigrewe K, et al. (2015) Combining mass spectrometric metabolic profiling with genomic analysis: A powerful approach for discovering natural products from cyanobacteria. *J Nat Prod* 78(7):1671–1682.
13. Cummings SL, et al. (2016) A novel uncultured heterotrophic bacterial associate of the cyanobacterium *Moorea producens* JHB. *BMC Microbiol* 16(1):198.
14. Nurk S, et al. (2013) Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J Comput Biol* 20(10):714–737.
15. Utturkar SM, et al. (2014) Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics* 30(19):2709–2716.
16. Boetzer M, Pirovano W (2014) SSPACE-LongRead: Scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* 15(1):211.
17. Pop M, Phillippy A, Delcher AL, Salzberg SL (2004) Comparative genome assembly. *Brief Bioinform* 5(3):237–248.
18. Galardini M, Biondi EG, Bazzicalupo M, Mengoni A (2011) CONTIGuator: A bacterial genomes finishing tool for structural insights on draft genomes. *Source Code Biol Med* 6(1):11.
19. Calteau A, et al. (2014) Phylum-wide comparative genomics unravel the diversity of secondary metabolism in cyanobacteria. *BMC Genomics* 15(1):977.
20. Ziemert N, et al. (2014) Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proc Natl Acad Sci USA* 111(12):E1130–E1139.
21. Jones AC, et al. (2011) Genomic insights into the physiology and ecology of the marine filamentous cyanobacterium *Lyngbya majuscula*. *Proc Natl Acad Sci USA* 108(21):8815–8820.
22. Campbell EL, Cohen MF, Meeks JC (1997) A polyketide-synthase-like gene is involved in the synthesis of heterocyst glycolipids in *Nostoc punctiforme* strain ATCC 29133. *Arch Microbiol* 167(4):251–258.
23. Zhang CC, Laurent S, Sakr S, Peng L, Bédu S (2006) Heterocyst differentiation and pattern formation in cyanobacteria: A chorus of signals. *Mol Microbiol* 59(2):367–375.
24. Stucken K, et al. (2010) The smallest known genomes of multicellular and toxic cyanobacteria: Comparison, minimal gene sets for linked traits and the evolutionary implications. *PLoS One* 5(2):e9235.
25. Lee H-M, Vázquez-Bermúdez MF, de Marsac NT (1999) The global nitrogen regulator NtcA regulates transcription of the signal transducer PII (GlnB) and influences its phosphorylation level in response to nitrogen and carbon supplies in the cyanobacterium *Synechococcus* sp. strain PCC 7942. *J Bacteriol* 181(9):2697–2702.
26. Berg H, et al. (2000) Biosynthesis of the cyanobacterial reserve polymer multi-L-arginyl-poly-L-aspartic acid (cyanophycin). *J Biochem* 267:5561–5570.
27. Cimermancic P, et al. (2014) Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* 158(2):412–421.
28. Weber T, et al. (2015) AntiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res* 43(W1):W237–W243.
29. Doroghazi JR, et al. (2014) A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat Chem Biol* 10(11):963–968.
30. Medema MH, Fischbach MA (2015) Computational approaches to natural product discovery. *Nat Chem Biol* 11(9):639–648.
31. Taniguchi M, et al. (2010) Palmyramide A, a cyclic depsipeptide from a Palmyra Atoll collection of the marine cyanobacterium *Lyngbya majuscula*. *J Nat Prod* 73(3):393–398.
32. Marquez BL, et al. (2002) Structure and absolute stereochemistry of hectochlorin, a potent stimulator of actin assembly. *J Nat Prod* 65(6):866–871.
33. Grindberg RV, et al. (2011) Single cell genome amplification accelerates identification of the apratoxin biosynthetic pathway from a complex microbial assemblage. *PLoS One* 6(4):e18565.
34. Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RY (1979) Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *Microbiology* 111(1):1–61.
35. Albertsen M, et al. (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* 31(6):533–538.
36. Podell S, Gaasterland T (2007) DarkHorse: A method for genome-wide prediction of horizontal gene transfer. *Genome Biol* 8(2):R16.
37. Kurtz S, et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5(2):R12.
38. Alikhan N-F, Petty NK, Ben Zakour NL, Beatson SA (2011) BLAST ring image generator (BRIG): Simple prokaryote genome comparisons. *BMC Genomics* 12(1):402.
39. Medema MH, Takano E, Breitling R (2013) Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Mol Biol Evol* 30(5):1218–1223.
40. Langille MGI, Hsiao WWL, Brinkman FSL (2010) Detecting genomic islands using bioinformatics approaches. *Nat Rev Microbiol* 8(5):373–382.
41. Weiling H, Xiaowen Y, Chunmei L, Jianping X (2013) Function and evolution of ubiquitous bacterial signaling adapter phosphopeptide recognition domain FHA. *Cell Signal* 25(3):660–665.
42. Singh SP, Montgomery BL (2011) Determining cell shape: Adaptive regulation of cyanobacterial cellular differentiation and morphology. *Trends Microbiol* 19(6):278–285.
43. Kuo YC, et al. (2013) Characterization of putative class II bacteriocins identified from a non-bacteriocin-producing strain *Lactobacillus casei* ATCC 334. *Appl Microbiol Biotechnol* 97(1):237–246.