Genomic insights into the physiology and ecology of the marine filamentous cyanobacterium *Lyngbya majuscula*

Adam C. Jones,1,2 Emily A. Monroe,1,3 Sheila Podell,2 Wolfgang R. Hess,2 Sven Klages,4 Eduardo Esquenazi,5 Sherry Niessen,5 Heather Hoover,6 Michael Rothmann,7 Roger S. Lasken,5 John R. Yates III,6 Richard Reinhardt,3 Michael Kube,5 Michael D. Burkart,2 Eric E. Allen,4 Pieter C. Dorrestein,4,5 William H. Gerwick,6,7 and Lena Gerwick1,2

*Center for Marine Biotechnology and Biomedicine, *6Marine Biology Research Division, Scripps Institution of Oceanography, *7Department of Chemistry and Biochemistry, Division of Biological Sciences, and *5Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California at San Diego, La Jolla, CA 92093; *3Institute of Biology III, University of Freiburg, D-79104 Freiburg, Germany; *4Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany; *5Center for Physiological Proteomics, The Scripps Research Institute, La Jolla, CA 92037; *6Microbial and Environmental Genomics, J. Craig Venter Institute, San Diego, CA 92121; and *7Genome Centre Cologne at MPI for Plant Breeding Research, D-50829 Cologne, Germany

Edited by Jerrold Meinwald, Cornell University, Ithaca, NY, and approved April 7, 2011 (received for review January 26, 2011)

Filamentous cyanobacteria of the genus *Lyngbya* are important contributors to coral reef ecosystems, occasionally forming dominant cover and impacting the health of many other co-occurring organisms. Moreover, they are extraordinarily rich sources of bioactive secondary metabolites, with 35% of all reported cyanobacterial natural products deriving from this single pantropical genus. However, the true natural product potential and life strategies of *Lyngbya* strains are poorly understood because of phylogenetic ambiguity, lack of genomic information, and their close associations with heterotrophic bacteria and other cyanobacteria. To gauge the natural product potential of *Lyngbya* and gain insights into potential microbial interactions, we sequenced the genome of *Lyngbya majuscula* 3L, a Caribbean strain that produces the tubulin polymerization inhibitor curacin A and the molluscicide barbamide, using a combination of Sanger and 454 sequencing approaches. Whereas ~293,000 nucleotides of the draft genome are putatively dedicated to secondary metabolism, this is far too few to encode a large suite of *Lyngbya* metabolites, suggesting *Lyngbya* metabolites are strain specific and may be useful in species delineation. Our analysis revealed a complex gene regulatory network, including a large number of sigma factors and other regulatory proteins, indicating an enhanced ability for environmental adaptation or microbial associations. Although *Lyngbya* species are reported to fix nitrogen, nitrogenase genes were not found in the genome or by PCR of genomic DNA. Subsequent growth experiments confirmed that *L. majuscula* 3L is unable to fix atmospheric nitrogen. These unanticipated life history characteristics challenge current views of the genus *Lyngbya*.

diazotrophy | polyketide synthase/nonribosomal peptide synthetase | mass spectrometry | harmful algal bloom | marine biology

Among the oldest life forms on Earth, cyanobacteria are well recognized for their global ecological importance and ubiquitous distribution across virtually all ecosystems (1). In the marine realm, some species of cyanobacteria contribute significantly to nitrogen fixation and global carbon flux (2), whereas others are prevalent as benthic constituents of tropical coral reefs (3). Over the past several decades, cyanobacteria have become recognized as an extremely rich source of novel, bioactive secondary metabolites (= natural products), with ~700 different compounds having been isolated and characterized (4). These compounds have gained considerable attention due to their pharmaceutical and biotechnology potential (5), but also notoriety for their environmental toxicity and threats to humans, wildlife, and livestock (6).

Marine strains of the genus *Lyngbya* are some of the most prolific producers of natural products.Nearly 240 compounds are reported from this genus, and 76% of these are attributed to a single species, *Lyngbya majuscula* (Harvey ex Gomont), which is found globally in shallow tropical and subtropical environments (4). *Lyngbya* bloom events pose a significant challenge to coral reefs, as *Lyngbya* can negatively impact coral larvae recruitment (7), quickly colonize available substrate, and persist in the presence of herbivores because of their chemical defenses (8). In the last 10 y, focused investigations into the biosynthesis of *L. majuscula* natural products have revealed gene clusters that encode the molecular assembly of several of these compounds, including the antitumor agent curacin A (9–11), the UV-sunscreen pigment scytonemin (12), and the lyngbyatoxins (14), dermatotoxic agents responsible for “swimmer’s itch.” Most of the gene clusters encode modular, mixed polyketide synthase (PKS)/nonribosomal peptide synthetase (NRPS) assembly lines, with several using highly unusual mechanisms to incorporate other functional groups into the resultant molecules (15).

Despite these advances in compound identification and biosynthesis, comparatively little is known about *Lyngbya* evolution or the full potential of specific *L. majuscula* strains to produce the natural products attributed to this species. Recent reassessment of phylogenetic diversity in the genus *Lyngbya* using the 16S rRNA gene has shown that *Lyngbya* appears to occupy three distinct clades: a halophilic/brackish/freshwater lineage, a lineage more closely related to the genus *Oscillatoria*, and a marine lineage (16). Moreover, metabolites attributed to *L. majuscula* have typically been isolated from field collections, which poses two problems: most taxonomic classifications have been based on morphological characteristics and not genetic evidence, and this cyanobacterium typically grows in close association with other microorganisms. Thus, it is possible that the total number of natural products associated with the species *L. majuscula* has been overestimated.

To determine the capacity for natural products biosynthesis in a specific strain of *L. majuscula*, we sequenced the genome of *L. majuscula* 3L, a strain that falls within the marine lineage described earlier, and has also recently been referred to as *Lyngbya sordida* 3L (16). *L. majuscula* 3L was originally isolated in Curacao, Netherlands Antilles, and has been maintained in stable culture.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The entire genome shotgun project reported in this paper has been deposited in the DDBJ/EMBL/GenBank database (accession no. AEFP00000000).

1A.C.J. and E.A.M. contributed equally to this work.

2To whom correspondence should be addressed: lgerwick@ucsd.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1101137108/-/DCSupplemental.
for ~15 y (17). This strain produces curacin A (18), the molluscicidal compound barbamide (19), and the lipopeptide carmabine A (20) (Fig. S1A and B). A draft genome was obtained from an integrated strategy involving Sanger sequencing of DNA from cultured filaments in combination with 454 sequencing of DNA generated via multiple displacement amplification (MDA) from single L. majuscula 3L cells; the latter approach was necessary to overcome the inability to create or maintain axenic cultures of L. majuscula. From this sequencing effort, we aimed to confirm the presence of the gene clusters encoding each of these molecules, search for other unknown (orphan) natural product biosynthetic pathways, and gain insights into the physiological ecology of L. majuscula in tropical environments, including possible interactions with other microorganisms and the ability of L. majuscula to fix atmospheric nitrogen (21, 22).

Results and Discussion

Genome Assembly and Annotation. L. majuscula 3L sequence reads were obtained from two independent, nonaxenic cultures, using two different DNA isolation procedures and two different sequencing technologies (Sanger and 454 approaches). The reads from both the Sanger and 454 libraries were pooled and treated as a single metagenomic dataset to identify core sequences of the Lyngbya genome that were common to both datasets. This strategy enabled evaluation of whether scaffolds assembled from sequences in both datasets contained constituent reads from one or both library sources to assist in separating consensus Lyngbya sequences from non-Lyngbya contaminants. Coassembly of 712,948 Sanger and 454 reads produced 6,217 scaffolds, ranging in size from 1,000 to 59,782 nucleotides, G + C content between 25 and 76%, and coverage depth from 1- to 62-fold. Classification of 16S rRNA genes in the combined assembly, taxonomic heterogeneity of the scaffolds, and details regarding the binning procedure are available in Table S1 and Fig. S2. A total of 161 scaffolds were identified as likely originating from L. majuscula, on the basis of a combination of 16S rRNA genes, predicted protein matches to GenBank nr sequences, percent G + C nucleotide composition, and assembly coverage depth. Detailed properties for all Lyngbya-associated scaffolds are provided in Table S2. The combined scaffolds total ~8.5 Mb, a total genome size consistent with other filamentous cyanobacteria, such as Nostoc punctiforme (8.2 Mb) and Trichodesmium erythraeum (7.8 Mb). It is uncertain whether this draft assembly represents the entire L. majuscula genome, but several lines of evidence suggest it is nearly complete. A survey of 102 housekeeping genes identified as nearly universal in bacteria (25) indicates that 101 of these are present in the Lyngbya draft genome. Copy numbers for these housekeeping genes correlate well with other sequenced cyanobacterial genomes, including those expected to have single copies (Dataset S1). In addition, previously known, independently sequenced L. majuscula genes for the curacin A (9) and barbamide (24) pathways are present and complete, despite not being used to guide any aspect of the assembly. The L. majuscula 3L draft genome was submitted to the Joint Genome Institute Integrated Microbial Genome’s (IMG) expert review for automated annotation of putative ORFs. Within the 8.5-Mb genome (44% G + C content), 56 tRNAs, 2 rRNA operons, and 7,479 protein-encoding genes were identified, with 54% of these protein-encoding genes having predicted functions. This number is higher than for N. punctiforme (6,086 genes) and T. erythraeum (4,451 genes). The largest percentage of annotated genes (based on clusters of orthologous groups categories, COGs) appears to be involved in replication, recombination, and DNA repair (9%), cell wall biogenesis (8%), and signal transduction mechanisms (7%). Despite previous reports that L. majuscula strains are karyotypic, no nitrogenase genes were found in this draft genome.

Secondary Metabolism Genes in L. majuscula 3L Draft Genome. Despite the large number of natural products attributed to L. majuscula, only 126 genes (3%, 293 kb) of the L. majuscula 3L draft genome are predicted to be involved in secondary metabolite biosynthesis, transport, and catabolism. The majority of these are modular NRPS- and/or PKS-related genes (44%, 199 kb). Eight biosynthetic gene clusters were identified that likely encode natural products (Fig. 1 and Fig. S1). The two most apparent clusters were those of the previously characterized natural products curacin A (HQ696500) and barbamide (HQ696501) (Fig. S1A). The sequences for both pathways were complete and consistent with the sequences previously reported (9, 24). Two separate scaffolds contain genes putatively involved in carminab biosynthesis (Fig. S1B) on the basis of predictions of adenylation domain substrate specificity (Materials and Methods) from the NRPS ORFs in each partial gene cluster. Five additional biosynthetic gene clusters were found in the L. majuscula 3L genome; however, they do not appear to encode natural products previously detected from this species (Fig. 1). The largest of these is an apparently intact 29-kb NRPS-dominated gene cluster on scaffold 52116 that is flanked by transposase genes on both sides (HQ696495). The adenylation domain active sites of the bimodular NRPS protein are predicted to activate and incorporate proline and arginine. Surrounding the NRPS are genes for an arginosuccinate lyase, which may provide arginine for the NRPS adenylation domain, and a GCN5-related N-acetyltransferase (GNAT), which may acylate arginine similarly to what has been reported (23). The adenylation domain is nearly universal in bacteria (23), indicating that 101 of these are present in the L. majuscula genome (10). The gene context of this motif in the current cluster appears to be different and is thus more likely to be involved in acetylation of a basic amino acid. Additionally, two adjacent phytanoyl-CoA dioxygenase (phyH)/L-proline 4-hydroxylase genes immediately precede the NRPS gene and possibly are involved in hydroxylation or halogenation of the proline residue (Fig. 1).

A second potentially complete orphan gene cluster in the L. majuscula 3L genome, located on scaffold 52118, is ~20 kb in size and is flanked on the 5′ side by a transposase gene (HQ696496) (Fig. 1). Three NRPS ORFs, one of which is bimodular, were predicted to encode isoleucine, lysine, tyrosine, and proline. The cluster also contains a sulfotransferase, suggesting the amino acid chain could be sulfated. An epimerase domain is present in the module incorporating lysine, and thus, as in almost all known cyanobacterial metabolites containing this basic amino acid, it is likely of D configuration (4). Another separate, mixed NRPS/PKS ORF is located 13 kb downstream of the last NRPS ORF on the scaffold and appears to have an adenylation domain specific for either phenylalanine or tyrosine (HQ696497) (Fig. 1). From catalytic activities predicted in the PKS portion, the amino acid is likely extended with acetate, the intermediate ketone reduced to an alcohol, and then released from the enzyme by a thioesterase. The majority of the genes surrounding this stand-alone NRPS/PKS gene appear to be involved in primary metabolism, and it is unclear whether they are involved in modifications of the NRPS/PKS product.

The remaining orphan clusters are on scaffolds 52120 and 52117 (Fig. 1). Scaffold 52120 has bimodular and single module NRPS genes that have predicted adenylation specificities for α-aminoacidic acid, glutamine, and proline, respectively (HQ696498). A predicted thioesterase is present at the terminus of the second NRPS gene. These are flanked by genes encoding the hypothetical proteins and proteins predicted to be involved in cytochrome c biosynthesis. The single NRPS ORF on scaffold 52117 encodes two modules (proline and threonine adenylation specificity), and as with the NRPS/PKS on scaffold 52118, the surrounding genes appear to be related to primary metabolism (HQ696499).

To determine whether any of the above predicted “cryptic metabolites” were expressed in cultures of L. majuscula 3L, we first profiled water soluble and organic extracts by LC/MS (Fig. S3), matrix-assisted laser desorption ionization (MALDI)/MS, and Fourier transform (FT)/MS. Curacin A and carmabine were readily
observed using all three techniques, and barbamide was detected using FT/MS. We did not detect any mass/charge values ascribable to the unknown metabolites predicted above. Using *L. majuscula* 3L soluble protein extracted from cultured biomass, we also performed a proteomic analysis to determine relative expression levels of secondary metabolite biosynthetic proteins under normal culture conditions. Multidimensional protein identification analysis (MudPIT; ref. 26) yielded spectral counts from at least two of four technical replicates for 1,043 proteins (Dataset S2), which represented ∼14% of the encoded proteins annotated in the *L. majuscula* 3L genome. The most readily detected proteins using MudPIT were pigment-associated proteins, including phycocyanin subunits and phycobilisome proteins (∼2,000 spectral counts per protein, Table S3A). Spectral counts for nearly all of the proteins in the curacin A and barbamide pathways were quantified, as were several proteins predicted to be involved in carmabin biosynthesis (Table S3B). Spectral counts for nearly all of the proteins in the curacin A and barbamide pathways were quantified, as were several proteins predicted to be involved in carmabin biosynthesis (Table S3B). Only one probable secondary metabolite protein from an orphan pathway was detected in at least two technical replicates (the free-standing proline/threonine NRPS from scaffold 52,117). None of the remaining orphan pathway proteins discovered during the genome annotation process were expressed to a measurable level by MudPIT analyses. Together with the absence of any of the predicted compounds in both the water soluble and organic extracts from *L. majuscula* 3L, these data suggest that these orphan gene clusters are either expressed at low, nearly undetectable levels or not expressed at all under normal culture conditions.

### Complex Regulatory Gene Network of *L. majuscula* 3L.

In light of the considerable number of natural products attributed to various strains of *L. majuscula* (nearly 200 reported metabolites), it was unexpected that *L. majuscula* 3L dedicates only 3% of its genome to secondary metabolism, which is significantly lower than that observed in marine actinobacteria such as *Salinispora* (9.9%; ref. 27), and that only three NRPS/PKS-type biosynthetic pathways larger than 30 kb were present. A larger component of the *L. majuscula* 3L genome is devoted to regulatory genes involved in transcription and signal transduction. Marine *Lyngbya* strains grow in shallow tropical areas with frequent exposure to diverse environmental stress factors such as desiccation during low tide or exposure to high fluxes of UV light. As noted previously, *Lyngbya* can usually be found living in close association with other cyanobacteria and heterotrophic bacteria. Even when growing separately from macroscopic assemblages, *L. majuscula* filaments retain a large number of associated bacterial cells on their polysaccharide sheath that are visible using DAPI staining (28). Therefore, a more careful evaluation of the *L. majuscula* 3L transcription and transduction genes was performed to evaluate the capacity of this organism for environmental adaptation and microbial communication.

*L. majuscula* 3L contains an unusual assortment of regulatory genes compared with other cyanobacteria. Comparison of the 15 sigma factor genes annotated in *L. majuscula* 3L against the nine well-characterized type I, II, and III sigma factors of *Synechocystis* sp. PCC 6803 (29) revealed that *L. majuscula* has precisely five matching sigma factor for each of the five type I and II σ70 factors SigA–SigE. In addition, it possesses another five factors belonging to the type III class (Fig. 2). Of this latter group, two most closely resemble SigF of *Synechocystis* sp. PCC 6803, whereas the remaining three are distinct from all other known type III factors. However, the most striking observation is the presence of 5 additional sigma factors, which have no close homolog in any of the previously sequenced model cyanobacteria *Synechocystis* sp. PCC 6803, *Anabaena* sp. PCC 7120, or *Synechococcus* sp. PCC 7942. These sigma factors are between 257 and 563 residues in length and have an unusual domain structure. A domain with pronounced similarity to σ24-type factors of the extracytoplasmic function (ECF) subfamily is located in the N-terminal half of the proteins. This is intriguing because the large and diverse group of ECF sigma factors plays a key role in...
adaptation to environmental conditions (30). The other two related proteins occur as single-copy genes in the marine filamentous cyanobacteria *Trichodesmium* and *Lyngbya* sp. PCC 8106 (Fig. 2) and are annotated as SigW and Sig24 ECF-type sigma factors. However, the fact that *L. majuscula* 3L possesses five such factors suggests that a multitude of regulatory mechanisms could exist in this organism for potential interaction with the marine environment or associated microorganisms.

Moreover, 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 [i.e., *Anabaena* PCC 7120 has 11 sigma factors, whereas *Anabaena variabilis*, American Type Culture Collection (ATCC) 29413, and *T. erythraeum* each have 7]. In the longest ORF (HQ962799), this domain is preceded by and partially overlaps an SpvB domain (closest homolog: *Salmonella* virulence plasmid 65 kDa B protein, pfam03534). The C-terminal halves of these proteins have no close homologs in the National Center for Biotechnology Information (NCBI) or pfam databases, and they contain two 48-residue-long repeats. The similarity among four of these five proteins in their C-terminal component suggests the presence of a novel protein domain that is presently uncharacterized. ECF sigma factors are frequently cotranscribed with one or more downstream negative regulators, which function as antisigma factors that bind and inhibit the cognate sigma factor (30). The *L. majuscula* 3L ECF-type sigma factors appear to belong to a class of sigma factors in which a regulatory domain has been fused to the protein. The recently identified sigma factor PhyR in *Methyllobacterium extorquens* provides a possible paradigm for such a possibility (31). In PhyR, an amino terminal ECF sigma factor-like domain is fused to a carboxyterminal receiver domain of a response regulator, suggesting PhyR can respond by sensing changes in the environment directly. A number of predicted short microRNA sequences were also found throughout the draft genome, including a cluster of mir-569 microRNA genes, suggesting that *L. majuscula* 3L use some level of post-transcriptional regulation. Although regulation of secondary metabolism in filamentous cyanobacteria has not been extensively evaluated, we did find that homologs of two proteins to those possibly involved in jamaicamide biosynthetic regulation (32) were expressed to detectable levels according to MupTFT analysis. Whether these latter proteins are involved in the regulation of curacin A or other secondary metabolite pathways in *L. majuscula* 3L remains to be determined.

**Absence of Nitrogen Fixation in *L. majuscula* 3L.** Perhaps the most unexpected finding in the *L. majuscula* 3L genome analysis was the lack of any genes involved in nitrogen fixation. Nitrogen availability is thought to be a major factor regulating primary production in shallow marine environments, and fixation of atmospheric nitrogen (N2) by some prokaryotes, including cyanobacteria, is a critical source of bioavailable nitrogen for marine ecosystems worldwide (2). Several genera of cyanobacteria have been shown to fix nitrogen, including *Lyngbya* species (21). *L. majuscula* nitrogen fixation has been detected previously by acetylene reduction assays (21, 22), and a dinitrogenase reductase (*nifH*) has been characterized from *L. majuscula* collected near Zanzibar in the Indian Ocean (22).

To independently investigate the capacity of *L. majuscula* 3L to fix nitrogen, the presence of nitrogenase genes was evaluated by PCR approaches as well as in several growth experiments performed in the absence of nitrate in the culture media. Using primers previously published to amplify *nifH* from *L. majuscula* (22), we successfully amplified a PCR product from genomic DNA isolated from *Oscillatoria nigro-viridis* 3LOSC, a cyanobacterial strain found growing in association with *L. majuscula* 3L in the field, but failed to amplify a product from *L. majuscula* 3L genomic DNA (Fig. S4). Additional experiments were performed to determine whether *L. majuscula* 3L could grow and survive in the absence of a fixed nitrogen source. Single filaments grown in nitrate-free (98%) media were comparable in length to filaments grown in normal SW BG-11 after 1 wk of growth; however, cell morphology and pigmentation were significantly altered in the nitrate-free samples. The length of cells grown in the nitrate-free media visibly increased and the filaments changed from dark red to light green and became colorless upon extended culture. Similar phenotypes were observed when this experiment was repeated using larger-scale 50-mL batch cultures (Fig. 3A). To assess nitrogen accumulation by *L. majuscula* 3L in nitrate-free media, soluble protein was isolated from 50-mL nitrate-free cultures after 1 wk of growth and compared with control cultures grown in control media for the same duration. During the course of the two independent experiments performed, control cultures significantly increased in protein content (P = 0.0077), whereas the nitrate-free cultures showed no increase in protein content (P = 0.358), indicating that *L. majuscula* 3L was unable to actively assimilate nitrogen from atmospheric dinitrogen (Fig. 3B).

We also explored the ability of *L. majuscula* 3L to assimilate atmospheric nitrogen through 15N isotope feeding experiments. *L. majuscula* 3L filaments were grown in media containing 15N-labeled sodium nitrate for ~21 d until nitrogen-containing compounds were fully labeled with this heavy isotope, as assessed by MALDI/TOF mass spectrometry of the metabolome (*Materials and Methods*). The fully 15N-labeled filaments were then grown in nitrate-free media for 10 d, and the incorporation of the prevailing natural 14N isotope from atmospheric N2 into nitrate-containing compounds was evaluated using MALDI/MS. The shift from 15N to 14N was calculated for pheophytin ∼0.358), indicating that *L. majuscula* 3L in nitrate-free (98%) media shifted to lighter mass by incorporation of 14N. Controls grown in regular 14N nitrate media shifted to lighter mass by 99% (±6.6%).
The 19% shift observed in the nitrate-free media may be due to other trace amounts of nitrogen in the media or may represent recycling of internal nitrogen stores that were not labeled during the incubation with 15N nitrate. A recent study examining proteomic changes in the cyanobacterium Synechocystis sp. 6803 in response to various environmental stresses, including low nitrogen, found that in addition to switching to alternative carbon and nitrogen assimilation pathways, Synechocystis can access internal carbon and nitrogen stores based on up-regulation of proteins associated with cyanophycin breakdown and downstream arginine catabolism (34). To provide nitrogen and carbon to the cell, cyanophycin, a storage polymer of L-aspartic acid and L-arginine, is broken down into arginine and aspartic acid by cyanophycinase. Arginine and aspartic acid can be subsequently broken down by arginine decarboxylases and agmatinase and/or arginase (34). The L. majuscula 3L genome contains genes for a cyanophycin synthetase (HQ692807), cyanophycinase (HQ692806), two arginine decarboxylases (HQ692805 and HQ692804), and one agmatinase (HQ692805). The second arginine decarboxylase is in close proximity to the agmatinase on scaffold 52022, supporting their suggested role in cyanophycin recycling. The presence of these genes in the L. majuscula 3L genome provides evidence that L. majuscula is capable of obtaining nitrogen from cellular storage, and this capacity to use internal nitrogen stores in low nitrogen environments could explain the 19% 15N shift observed in the MALDI growth experiments. The loss of pigmentation observed under nitrate-free conditions is also consistent with the observations of down-regulation of photosystem proteins in response to environmental stresses (34). Collectively, these phenotypic and growth assessments strongly suggest that L. majuscula is unable to fix atmospheric nitrogen, and that under nitrate-free growth conditions, recycles nitrogen from storage proteins such as cyanophycin.

The apparent discrepancy between our investigations with L. majuscula 3L and past demonstrations of nitrogen fixation in L. majuscula may reflect strain differences or, possibly, the different criteria used in assigning the taxonomy of these cyanobacteria (21, 22). A 16S rRNA phylogenetic assessment of L. majuscula Lyngbya lineage places this strain in the marine Lyngbya lineage as recently reported (16), which is phylogenetically distinct from freshwater Lyngbya strains. It is also conceivable that previous nitrogen fixation experiments with L. majuscula, wherein the organism was identified solely by morphology, may have actually investigated other morphologically similar but unrelated genera. For example, Oscillatoria, another cyanobacterial genus reported to fix nitrogen (35), is morphologically very similar to Lyngbya and can be easily misidentified without the taxonomic support provided by phylogenetics.

Conclusions
Since the first evaluation of their natural products 40 years ago, tropical filamentous marine cyanobacteria are now established as rich sources of novel bioactive molecules. We selected L. majuscula 3L for genome sequencing because it is a strain that has been successfully cultivated in our laboratory for 15 y and has been studied extensively for its natural products and biosynthetic pathways. The genome sequence contained intact gene clusters for curacin A and barbamide, consistent with our previous reports (9, 24), as well as genes in good agreement with carmabin biosynthesis. However, no other gene clusters above 30 kb were evident in the draft genome, and only five other PKS and/or NRPS pathways were detected. Organic extracts from L. majuscula 3L and expression analysis of the soluble proteome revealed that these unknown pathways are either not expressed or expressed at undetectable levels under typical culture conditions. Thus, the more than 200 metabolites reported from this species are likely due to a very large number of different chemical strains or chemotypes. Moreover, the processes of horizontal gene transfer or evolutionary pathway divergence, as suggested for the curacin A and the jamaicamide pathways from L. majuscula (11), or the apratoxins from strains of Lyngbya bouillonii (36), are likely responsible for this impressive molecular diversity.

The discovery of regulatory genes conferring an enhanced ability for microbial interactions and/or environmental adaptation and lack of traditional nitrogen fixation pathways were additional unexpected findings in the L. majuscula 3L genome. Lyngbya is almost always found growing in close association with other cyanobacteria, diverse microorganisms, and invertebrates in the field. A wide variety of heterotrophic bacteria remain on the surface of the polysaccharide sheath even after extensive purification of field isolates. The relationship between Lyngbya and these associated organisms remains unclear, but the possibility of complex interactions taking place among them is a fascinating focus for future research and may also explain the large number and variety of ECF sigma factors and other regulatory

**Fig. 3.** Absence of nitrogen fixation in L. majuscula 3L. (A) Phenotypic changes of L. majuscula 3L when grown in nitrate-free media. Microscopic images are 400×. (B) Total soluble protein in L. majuscula 3L grown with (control, diamonds) and without (squares) nitrate for 8 d compared with day 0. Error bars represent SEM between replicate experiments (n = 6 per treatment). (C) % 15N incorporation into pheophytin a assayed by MALDI/TOF. L. majuscula 3L was grown in 15N-labeled nitrate until nitrogen-containing molecules were fully labeled. Filaments were then grown in SW BG-11 media with 15N-labeled nitrate, nitrate-free SW BG-11 media, and control 14N nitrate SW BG-11 media (n = 8) for 10 d and incorporation of 14N into pheophytin a was measured by MALDI/TOF. Error bars are SEM.
genes described above. Our finding that nitrogen fixation does not occur in this L. majuscula strain is in direct contrast to previous reports, but may be another indication that fine-scale phylogenetic relationships of marine filamentous cyanobacteria need to be better defined. Among nitrogen-fixing organisms, cyanobacteria form a monophyletic group; however, within cyanobacteria, the capacity to fix nitrogen appears to be polyphyletic, suggesting multiple gene losses of nitrogen-fixation genes over time, horizontal gene transfer causing independent introduction of nitrogen-fixation genes, or a combination of both (37, 38). Recent examination of nitrogen-fixation gene evolution implies a complex history of both gene loss and horizontal gene transfer events (37, 38). The current L. majuscula 3L genomic data, showing loss events (via the presence of pseudogenes or regions in the genome where there was possible loss of the entire nif gene cluster), but as additional sequence data for diazotrophic and nondiazotrophic marine filamentous cyanobacteria become available, the evolutionary history of nitrogen fixation in this group can be better understood. Because L. majuscula strains previously found to fix nitrogen were identified using morphological techniques, it is difficult to determine how closely related these may be to L. majuscula 3L and whether the ability to fix nitrogen is more of an exception than a rule for this genus.

The first sequencing of a marine Lyngbya species presented here clearly accentuates the need for genomic study of additional Lyngbya strains. Recent phylogenetic assessment has revealed a significant degree of ambiguity in Lyngbya taxonomy (16). At least three separate lineages have been described from different environments. Specific natural products isolated from Lyngbya may be a more effective means of delineating these cyanobacterial strains, as has been proposed for marine actinomycetes (39). Additional genome sequencing of other Lyngbya collections will be required to better understand how the traits described here compare between species, strains, and across geographic locations.

Materials and Methods

L. majuscula 3L was originally collected in 1993 near CARMABI Research Station in Curacão, Netherlands Antilles, and live cultures have since been maintained (17) in SW BG-11 media (40). See SI Materials and Methods for details of fosmid library construction, single cell isolation, sequencing methods, genome assembly, binning techniques, and all other experiments. This whole genome shotgun project has been deposited at DDBJ/EMBL/GenBank under accession no. AEPP0000000. The version described in this paper is the first version, AEP010000000.

ACKNOWLEDGMENTS. We thank Dr. Hyukjae Choi for assistance with LC/MS analysis and Dr. Roland Korol for assistance with FT/MS. This work was supported by the Network of Excellence in Marine Genomics Europe, National Institutes of Health (NIH) Grants CA108874 (to W.H.G. and L.G.) and R21HG005107-02 (to E.E.A.) and California Sea Grant R/NMP-103EPD. E.E.A. is an Institutional Research and Academic Career Development Award (IRACDA) Postdoctoral Fellow supported by National Grant GM085776.

17. Swingley WD, Blankenship RE, Raymond J (2010) Integrating Markov clustering and R21HG005107-02 (to E.E.A.) and California Sea Grant R/NMP-103EPD. E.E.A. is an Institutional Research and Academic Career Development Award (IRACDA) Postdoctoral Fellow supported by National Grant GM085776.
Supporting Information

SI Materials and Methods

Culturing Techniques. The original collection of *Lyngbya majuscula* 3L in 1993 was conducted near the CARMAPI Research Station on the island of Curaçao, Netherlands Antilles. Pan (10 L) and Erlenmeyer flask (1 L) cultures have since been maintained (1, 2) at Oregon State University and Scripps Institution of Oceanography, University of California at San Diego in SW BG-11 (3) at 28 °C, under 16 h light/8 h dark cycles at ~5 μE m⁻² s⁻¹.

For Sanger sequencing, DNA from laboratory-cultured *L. majuscula* 3L was extracted and used in creation of two separate fosmid libraries (CopyControl Fosmid Library Production kit, Epicentre) using standard protocols, and these libraries were sequenced at the Max Planck Institute for Molecular Genetics. For 454 sequencing, single cells from a *L. majuscula* 3L were liberated from filaments using a razor blade and captured with a capillary tube using a microscope. The cells were lysed and their DNA amplified using the Repli-g minikit (Qiagen). Random primers were added (4) and multiple displacement amplification (MDA) of genomic DNA was performed. The amplified DNA was screened for purity by PCR using cyanobacterial 16S rRNA gene primers (5) and cloning (TOPO-TA, Invitrogen) and submitted for 454 sequencing (J. C. Venter Institute, La Jolla, CA). A total of 223 Mb were sequenced and the average length of the sequence reads was 387 bp.

Genome Assembly, Binning, and Annotation. A total of 136,560 Sanger sequences and 576,388 pyrosequencing (FLX 454) reads were combined into a single hybrid assembly using Celera Assembler software, version 5.4 (6). The mer overlapper program setting was used to achieve optimal integration of Sanger and 454 datasets (7). Relative numbers of Sanger and 454 reads incorporated into each scaffold produced by the combined assembly were parsed from Celera Assembler output using a custom perl script. ORFs and amino acid sequences were predicted from all scaffolds using the gene finding program MetaGene (8). Predicted proteins were evaluated for phylogenetic relatedness to known sequences in National Center for Biotechnology Information GenBank or using the DarkHorse program, version 1.3, with a threshold filter setting of 0.1 (9). Only matches with alignments covering at least 70% of total query length, and blastp e-value scores of 1e⁻⁵ or better were included in the DarkHorse analysis.

Assembled scaffolds were placed in one of three phylogenetic categories (*Lyngbya*, non-*Lyngbya*, or ambiguous) on the basis of manual assessment of several different, independent parameters. For each scaffold, number and alignment quality of predicted protein matches to GenBank sequences associated with phylum Cyanobacteria were compared with proteins from other taxonomic groups on the basis of DarkHorse analysis of the GenBank nr blastp search. 16S rRNA matches to known microbial sequences were identified by blastn search against the GreenGenes reference database, requiring a minimum alignment length of 200 nucleotides, and e value of 1e⁻⁷ or better (10). For scaffolds longer than 5,000 nucleotides, inclusion of reads from both Sanger and 454 libraries, percent G + C nucleotide composition, and coverage depth were also considered in determining scaffold origin. Scaffolds classified as most likely belonging to *Lyngbya* were annotated using the Integrated Microbial Genome Expert Review (IMG-ER) service of the Joint Genome Institute (11). Further annotation was performed using IMG-ER database tools.

Annotation of Secondary Metabolite Pathways. To identify secondary metabolite pathways, the genome was queried for genes annotated as polyketide synthase or nonribosomal peptide synthetase-related genes in addition to BLAST queries using previously sequenced secondary metabolite pathways. Molecule predictions for the orphan pathways were based on adenylation domain specificities predicted using NRPspredictor (12) and other domains present in the pathways. Secondary metabolite pathway figures were generated using Geneious bioinformatics software (13).

Organic Extraction of *L. majuscula* 3L Filaments for Secondary Metabolite Profiling. *L. majuscula* 3L filaments (140 mg dry weight) were obtained from a pan culture that had been growing under standard culture conditions for ~4 mo. The filaments were extracted twice in 2:1 dichloromethane (DCM):methanol (MeOH), and these extractions were filtered through cheesecloth and Whatman filter paper and combined. This crude extract was then extracted with MilliQ water in a separatory funnel to generate organic and aqueous fractions. The *L. majuscula* filaments were extracted again in 1:1 ethanol (EtOH):water with stirring, and this extract was also filtered through cheesecloth and filter paper. Portions of each of these three extracts (“organic,” “aqueous,” and “1:1 EtOH:H₂O biomass extraction”) were evaporated to dryness before being resuspended in MeOH and passed through a 0.2-μm syringe filter. The extracts were profiled using direct injection (positive ion mode) on a Thermo-Finnigan LCQ Advantage Max mass spectrometer at 1 or 10 mg/mL, as well as by matrix-assisted laser desorption ionization MALDI/MS as previously described (14). Separate organic *L. majuscula* 3L extracts were profiled using Fourier transform (FT)/MS (15).

Proteomic Analysis. Preparation of soluble protein. *L. majuscula* 3L tissue was harvested from a pan culture using forceps and measured by displacement in a 50-mL Falcon tube (4 mL of Falcon tube (4 mL of cultured biomass) containing lysis buffer (10 mM Tris HCl, pH 7.5, 150 mM NaCl). Before cell lysis, the buffer was treated with a broad range protease inhibitor (cOmplete, EDTA-free; Roche). The *L. majuscula* filaments were sonicated repeatedly on ice, and the resulting material was centrifuged at 4 °C at 14,000 rpm to pellet cellular debris and insoluble protein. The supernatant containing soluble protein was collected after centrifugation, and flash frozen in liquid N₂ before use in proteomic analysis.

Liquid chromatography/mass spectrometry (MudPIT) analysis. A total of 100 μg of protein from the preparation described above were TCA precipitated. Precipitated proteins were resuspended in 8 M urea, 50 mM Tris (pH 8.0) and digested in the presence of ProteasMAX using the suggested protocol (Promega). Peptides were acidic to a final concentration of 5% formic acid and 25 μg of peptides were bomb loaded onto a biphasic (strong cation exchange/reverse phase) capillary column for multidimensional protein identification analysis (MudPTP). Peptides were separated and analyzed by 2D-LC separation in combination with tandem MS as previously described (16). Peptides were eluted in an 11-step salt gradient and data were collected in an ion trap mass spectrometer (ThermoFisher; LTQ) set in a data-dependent acquisition mode with dynamic exclusion turned on (90 s). Each full MS survey scan was followed by 7 MS/MS scans. Spray voltage was set to 2.75 kV and the flow rate through the column was 0.20 μL/min.
Growth of single L. majuscula and 22,935 human International Protein Index sources, 2,161 scaffolds were categorized as most likely Ampli (PN1: 5 L. majuscula 3L and 22,935 human International Protein Index protein sequences) that were concatenated to a decoy database in which the sequences for each entry in the original database was reversed (17). In total the search database contained 60,828 protein sequence entries (30,414 real sequences and 30,414 decoy sequences). SEQUEST searches allowed for oxidation of methionine residues (16.0 Da), static modification of cysteine residues (57.0 Da, due to alkylation), no enzyme specificity, and a mass tolerance set to ±1.5 Da for precursor mass and ±0.5 Da for product ion masses. The resulting ms² spectra matches were assembled and filtered using DTASelect2 (version 2.0.27). For this analysis, tryptic, half-tryptic, and fully-tryptic peptides were each individually evaluated using the DTASelect2 software (18). In each of these subgroups the distribution of Xcorr and deltaCN values for a direct (to the direct database) and decoy (reversed database) were separated by quadratic discriminant analysis. Outlier hits in the two distributions were removed. Spectral matches were retained with Xcorr and deltaCN values that produced a maximum peptide false positive rate of 1%, which was the maximum frequency to the decoy database (number of decoy database hits/number of filtered peptides identified × 100). This value is calculated by the DTASelect2 software. In addition, a minimum peptide length of seven amino acid residues was imposed and protein identification required the matching of at least two peptides per protein. Such criteria resulted in the elimination of most decoy database hits. In our dataset, the identification of nontryptic peptides included half-tryptic peptides from the N- and C termini of the identified proteins. Other nontryptic peptides that were identified may represent endogenous activities of cellular proteases or peptides generated by in-source fragmentation.

Phylogenetic Analysis of Regulatory Proteins. Fifteen L. majuscula sigma factors were compared with the 9 sigma factors from Synechocystis PCC 6803. The phylogenetic relationships were inferred using the minimum evolution method and 1,000 bootstrap replicates. The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site. The ME tree was searched using the close-neighbor interchange (CNI) algorithm at a search level of 1. The neighbor-joining algorithm was used to generate the initial tree. All positions containing gaps and missing data were eliminated from the dataset (complete deletion option). There were a total of 126 positions in the final dataset. Phylogenetic analyses were conducted in MEGA4 (19).

Nitrogen-Fixation Methods. Amplification and sequencing of nifH. Genomic DNA was isolated from L. majuscula 3L and Oscillatoria nigro-viridis 3L-OSC using the Promega Wizard genomic DNA purification kit and quantified using an ND-1000 NanoDrop spectrophotometer (Thermo Scientific). Primers previously used to identify and sequence nifH from L. majuscula (PN1: 5′-CGTCAGGTCAAAAGATTAC-3′; PN2: 5′-ACACCACTAGCATGACAT-3′) (20) were used in 25-μL PCR reactions containing 1× Promega PCR master mix, 0.4 μM of each primer, and 1 μL of genomic DNA template. The following cycling parameters were used for PCR: Initial denaturation at 95 °C for 1 min, 30 cycles of 95 °C for 30 s, 48 °C for 30 s, and 72 °C for 1.5 min, and a final extension at 72 °C for 5 min. General 16S primers for cyanobacteria (5) were used as a positive control. PCR products from O. nigro-viridis 3L-OSC were purified using Qiagen Minelute PCR purification kit, cloned into a TOPO-TA cloning vector (Invitrogen) for sequencing, and sequenced by Seqxcel (San Diego).

Growth of L. majuscula in Nitrate-Free Media. Growth of single filaments. Single L. majuscula 3L filaments were grown in 24-well plates in SW BG-11 media at 28 °C on a 16:8 light:dark cycle. Twelve wells contained regular SW BG-11 media, and 12 wells contained nitrate-free SW BG-11 media. Filaments were measured using ImageJ (21) at day 0 and day 7. Filament lengths at day 7 were compared with their respective lengths at day 0 to approximate growth and cell size.

Growth of batch cultures. A total of 16 50-μL batch cultures were set up in either control SW BG-11 or nitrate-free SW BG-11 media (8 per set) with equal amounts of starting material (~0.05 g wet weight) under the same culturing conditions as above. Triplicate samples of the starting material were obtained to represent day 0. To assess protein content as a metric for growth, samples were placed in 2 mL of protein lysis buffer (10 mM Tris pH 7.5, 150 mM NaCl, and Roche Complete EDTA-free protease inhibitors), and homogenized using a probe sonicator. Samples were centrifuged at 13,000 × g for 10 min and soluble protein was recovered. Total soluble protein was calculated using the Pierce BCA assay according to the manufacturer’s protocols and using BSA as protein standard. At day 8, total biomass was harvested from triplicate control and nitrate-free cultures, and protein measurements were done in the same manner. Statistical analyses were performed on the two replicate experiments for a total of six biological replicates (three per experiment). Student’s t test was used to analyze significant differences between the experimental treatments using JMP software version 9 (SAS Institute). The SE between experimental replicates was calculated by the following equation: \( \frac{\sqrt{sd_x^2 + sd_y^2}}{n} \). Microscopy was performed using an Olympus IX51 microscope.

N\text{2}, Assimilation Studies. L. majuscula 3L filaments were grown in SW BG-11 media containing 15N-labeled nitrate for ~21 d until the majority of nitrogen-containing molecules contained labeled nitrogen as assessed by MALDI/TOF (14). Filaments were then grown in 24-well plates in either SW BG-11 media containing 15N-labeled nitrate (negative control), regular SW BG-11 media (positive control), or nitrate-free SW BG-11 media (eight wells per treatment). Filaments were sampled at day 0 and day 10 and frozen until subsequent MALDI/TOF analysis. Approximately 1 μL of MALDI matrix solution [per 1 mL: 35 mg α-cyano-4-hydroxycinnamic acid (CHCA), 35 mg 2,5-dihydroxybenzoic acid (DHB) (universal MALDI matrix; Sigma Aldrich), 750 μL acetonitrile, 245 μL milliQ H2O, 2 μL TFA] per 0.1 μg of biomass was mixed in a tube or well. After 20–30 s, 1 μL of this crude matrix solution was deposited on a well (spot) of a Bruker Microflex MSP 96 stainless steel target plate. After each spot had dried at room temperature, the plate was analyzed using a Bruker Microflex MALDI/TOF mass spectrometer equipped with flexControl 3.0 as described (14).

SI Results and Discussion

16S rRNA Genes in the Assembly and Taxonomic Heterogeneity of Scaffolds. 16S rRNA sequences identified in the assembled scaffolds include several representatives from Alpha-proteobacteria, Gamma-proteobacteria, and Bacteriodetes, but only Lyngbya from phylum Cyanobacteria (Table S1). Most of the non-Lyngbya matches are composed exclusively of reads from only one of the two libraries (100% or 0% from 454), suggesting they were present in only one of the two cultures used to obtain sequencing data. The exception to this pattern was scaffold 46429, which most closely matched an uncultured Alpha-proteobacterium isolated from a diseased Caribbean coral (22).

After the binning procedure that identified 161 scaffolds from Lyngbya sources, 2,161 scaffolds were categorized as most likely
derived from non-
Lyngbya sources. These scaffolds contain multiple predicted proteins matching noncyanobacteria GenBank entries. Most of these scaffolds are composed exclusively of reads from either Sanger or 454 libraries, but not both types of reads coassembled in the same scaffold. A total of 3,896 scaffolds were left unclassified. Most of these sequences are very short, with coverage too low to determine whether unusual G + C composition or absence of reads from one of the two sequencing libraries might be due to random statistical variation. Few of the unclassified scaffolds are large enough to provide even one ORF with reliable database matches. Some contain a single predicted peptide or peptide fragment closely matching a single database protein, but this evidence alone was not considered sufficient to determine scaffold origin, given the strong possibility of horizontal gene transfer between unrelated bacteria in mixed cultures.

Overall taxonomic heterogeneity of scaffolds generated in the combined assembly was assessed by binning scaffolds according to G + C content in 1% increments, then adding together the lengths of all scaffolds in each bin to obtain total assembled nucleotides (Fig. S2). Longer sequences originating from the same organism are expected to have a fairly uniform G + C percentage, consistent with species-specific nucleotide composition generally characteristic of genomic DNA (23). This pattern is observed in Fig. S2, as raw reads (Fig. S2A) are first recruited to contigs (Fig. S2B), then to slightly longer scaffolds by joining adjacent contigs with gaps (Fig. S2C). The most dominant peak in Fig. S2C, at 43% G + C, is associated with Lyngbya, containing coassembled reads from both Sanger and 454 libraries. Most scaffolds in the 47% G + C peak are composed almost exclusively of 454 reads, in many cases without any contribution from Sanger reads. On the other hand, most scaffolds in the broad 59% G + C peak contain exclusively Sanger reads, and lack 454 reads. These results, in combination with 16S rRNA data in Table S1, support the premise that the two cultures used to obtain these sequencing libraries contained different types of impurities, and that a core Lyngbya genome can be recovered from a metagenomic environment by combining sequence data from biological replicates.

**Fig. S1.** Secondary metabolite gene clusters for known compounds in the *L. majuscula* 3L draft genome. (A) Intact gene clusters for curacin A and barbamide, both of which are entirely consistent with previous descriptions. (B) Two scaffolds (51,865 and 52,117) containing probable portions of the carmabion biosynthetic gene cluster, based on adenylation domain substrate specificity. Figure constructed using Geneious 5.1 (13).
Fig. S2. Percent G + C content of combined Sanger–454 assembly for (A) raw reads, (B) assembled contigs, and (C) scaffolds created by joining contigs with gaps.
Fig. S3. Organic extraction of L. majuscula 3L filaments for secondary metabolite profiling. Mass/charge (m/z) values from direct injection of L. majuscula 3L (A) organic extract, (B) aqueous extract, and (C) 1:1 EtOH:H$_2$O biomass extract. Curacin A (m/z 374.2) was present in all three extracts, while carmabin A (726.4) was present in the organic and biomass extracts. Mass/charge values of the molecules depicted in Fig. 1 [gene cluster 1 (329.4), cluster 2 (519.6), cluster 3 (209.2), cluster 4 (374.4), and cluster 5 (216.2)] were not identified.

Fig. S4. Independent investigation of nifH in L. majuscula 3L. To independently assess the presence of nifH genes in L. majuscula 3L, primers previously published for L. majuscula nifH genes were used in PCR to amplify nifH from genomic L. majuscula 3L DNA. O. nigo-viridis 3L-OSC was used as a positive control for the nifH PCR, and 16S primers were used as an overall positive control. Lanes: M, 1 kb marker; 1–3, O. nigo-viridis 3L-OSC; 4–6, L. majuscula 3L; 7, no template control.

Other Supporting Information Files

Table S1 (DOC)
Table S2 (DOC)
Table S3 (DOC)
Dataset S1 (XLS)
Dataset S2 (XLS)