Origin of Chemical Diversity in Prochloron-Tunicate Symbiosis

Zhenjian Lin, Joshua P. Torres, M. Diarey Tianero, Jason C. Kwan, Eric W. Schmidt

Department of Medicinal Chemistry, University of Utah, Salt Lake City, Utah, USA

ABSTRACT

Diversity-generating metabolism leads to the evolution of many different chemicals in living organisms. Here, by examining a marine symbiosis, we provide a precise evolutionary model of how nature generates a family of novel chemicals, the cyanobactins. We show that tunicates and their symbiotic Prochloron cyanobacteria share congruent phylogenies, indicating that Prochloron phylogeny is related to host phylogeny and not to external habitat or geography. We observe that Prochloron exchanges discrete functional genetic modules for cyanobactin secondary metabolite biosynthesis in an otherwise conserved genetic background. The module exchange leads to gain or loss of discrete chemical functional groups. Because the underlying enzymes exhibit broad substrate tolerance, discrete exchange of substrates and enzymes between Prochloron strains leads to the rapid generation of chemical novelty. These results have implications in choosing biochemical pathways and enzymes for engineered or combinatorial biosynthesis.

IMPORTANCE

While most biosynthetic pathways lead to one or a few products, a subset of pathways are diversity generating and are capable of producing thousands to millions of derivatives. This property is highly useful in biotechnology since it enables biochemical or synthetic biological methods to create desired chemicals. A fundamental question has been how nature itself creates this chemical diversity. Here, by examining the symbiosis between coral reef animals and bacteria, we describe the genetic basis of chemical variation with unprecedented precision. New compounds from the cyanobactin family are created by either varying the substrate or importing needed enzymatic functions from other organisms or via both mechanisms. This natural process matches successful laboratory strategies to engineer the biosynthesis of new chemicals and teach a new strategy to direct biosynthesis.

Secondary metabolites are specialized molecules that are often directed outward at other organisms (1, 2). For example, under the sea, soft-bodied animals defend themselves using a diverse array of specialized chemicals, which are required for their survival (3, 4). Symbiotic bacteria, and not the host animal, synthesize many secondary metabolites (5), providing a link between symbiosis, chemistry, the survival of the animal and bacterial associates, and effects on predators and other organisms on coral reefs (6–8). The chemicals underlying these interactions are structurally diverse, comprising families of related compounds that are useful in drug discovery. Slight changes to chemical structure can drastically alter function, yet many marine natural products families are extremely diverse. Compounds such as marine animal defensive chemicals are critical to interactions between organisms (3) so that structural changes have consequences that potentially ripple through the environment. An open question has been, given these many constraints, how can chemical diversity arise? The evolution of novel chemistry likely relies in part on diversity-generating (DG) metabolism, which features pathways comprised of exceptionally broad-substrate enzymes (1, 9–11). These metabolic pathways natively accept many different substrates, enabling generation of chemical diversity. In turn, this provides organisms with rapidly evolving chemistry to face new challenges.

The tunicate-Prochloron symbiosis provides a model system to understand diversity-generating metabolism. Prochloron cyanobacteria are abundant obligate symbionts found in many tunicates of the family Didemnidae (12). Among the ubiquitous Prochloron compounds are the cyanobactins, a group of ribosomally synthesized and posttranslationally modified peptide (RiPP) secondary metabolites (13–18). Cyanobactins in Prochloron are hypervariable in sequence and structure, and their biosynthetic pathways provide a canonical example of diversity generation (9). Cyanobactins originate from simple ribosomally transcribed precursor peptides, exemplified by PatE and TruE, which are then modified by a series of enzymes. Cyanobactins are also widely found in free-living cyanobacteria (13, 15, 19, 20).

Despite previous observations of related cyanobactin pathways in Prochloron (13, 15), the genetic basis underlying chemical diversity was not well understood. Two major hurdles stood in the way: only two basic cyanobactin pathway sequences were known in Prochloron, making it difficult to construct evolutionary hypotheses, and the evolution of the host-symbiont symbiosis was not well characterized. Previously, based upon small subunit (SSU) rRNA gene sequences, the relationships between Prochloron and tunicates appeared stochastic (21–23). However, at least in the tunicate Lissoclinum patella, Prochloron chemistry correlated with host phylogeny (24, 25), indicating that there might be a genetic relationship between symbiont and host. Indeed, a recent phylogeny constructed from 120 available cyanobacterial genomes showed that Prochloron from samples L2 and L3 shared a
branch, while those from L1 and L4 shared another (26). This phylogenetic relationship reflected a phylogeny of the host tunicates obtained from 18S RNA and cytochromes oxidase I genes (24). In many didemnid species, Prochloron is transmitted from parent to offspring through generations (27), which further suggested the possibility of a more specific relationship between symbiont and host phylogeny.

Here, we analyzed sufficient further samples of the Prochloron-tunicate symbiosis to show that, in contrast to previous results, Prochloron and hosts have congruent phylogenies, indicating a specific relationship between symbiont and host. Two new cyanobactin pathways were discovered, which in combination with phylogenetic studies enabled us to finely define the evolutionary trajectory of cyanobactin pathways in tunicates. As a result, we provide a real-world example of how diversity-generating metabolism solves the problem of adapting to new environmental conditions. The evolutionary trajectory closely resembles idealized strategies for combinatorial biosynthesis in the laboratory, setting the stage for the rational engineering of desired chemicals.

MATERIALS AND METHODS

General. All PCRs were carried using either Platinum Taq High-Fidelity (Invitrogen) or Phusion High-Fidelity DNA Polymerase (NEB), containing 0.1 μl of DNA polymerase (for 10-μl-scale reactions), the supplied buffer (1X), 2 μM each primer, and 0.2 mM each deoxynucleoside triphosphate (dNTP; Invitrogen). Platinum Taq reaction conditions consisted of a hot start (95°C for 5 min), followed by 35 cycles of 95°C for 30 s, variable annealing temperatures for 30 s, and 68°C for variable extension times, with a final extension step of 68°C for 10 min. Phusion reaction conditions consisted of a hot start (95°C for 5 min), followed by 35 cycles of 95°C for 30 s, variable annealing temperatures for 30 s, and 68°C for variable extension times, with a final extension step of 72°C for 10 min. All cloning was carried out by using a TOPO TA cloning kit (with TOP10 chemically competent cells; Invitrogen) after gel extraction using a QIAquick gel extraction kit (Qiagen) or by direct PCR purification using a QIAquick PCR purification kit (Qiagen). Transformed TOP10 cells were grown on LB agar plates containing 5-bromo-4-chloro-3-indolyl-β-d-galactopyranoside (X-Gal; 40 μg/ml) and ampicillin or kanamycin (both, 50 μg/ml) at 30°C, and individual colonies were grown overnight in LB medium containing the same antibiotics at 30°C, with shaking. Plasmids were isolated using a QIAprep Spin Miniprep kit (Qiagen). Liquid chromatography-mass spectrometry (LC-MS) analysis was carried out using a Micromass Q-TOF Micro and 2795 HT high-performance liquid chromatography (HPLC) system (both, Waters).

Sample collection and processing. Samples were collected by scuba diving or snorkeling from Papua New Guinea (PNG), the Solomon Islands, Fiji, and Palau, with appropriate legal agreements in place as previously described (25). Freshly collected samples were rinsed with sterile artificial seawater, diced, and placed in RNAlater for later processing in the lab. Frozen samples of the same colonies were also kept for chemical analysis. Diplosoma sp. from Okinawa was kindly provided by Katsuhiro Ueda. Table S1 in the supplemental material describes sample numbers and the location of collection for each sample. RNAlater-preserved tissue was subjected to an established method for DNA extraction from tunicates, followed by phenol-chloroform extraction and isopropanol precipitation (28). The resulting DNA samples were purified a second time with a Genomic DNA Clean & Concentrator kit (Zymo Research) and further subjected to Illumina sequencing or PCR analysis. Table S2 in the supplemental material describes which samples were used for metagenome sequencing and distinguishes between which samples were sequenced in this study and which were sequenced in previous studies.

We used 19 Prochloron-containing tunicates from different branches of the Didemnidae. Metagenome sequences were analyzed from 11 different tunicate-Prochloron associations, including the didemnid tunicates Lissoclinum patella, Lissoclinum bistratum, Didemnum molle, and Diplosoma sp., from widely different geographic locations. In this study, we specifically extracted and analyzed Prochloron whole-genome sequences and host 18S RNA gene sequences from the metagenomic sequence. Metagenomes were previously reported for L. patella samples L1 to L4 (25), obtained from Palau, the Solomon Islands, Fiji, and the north coast of PNG. Here, we obtained a further seven Prochloron-tunicate metagenome sequences and partially assembled the Prochloron genomes: L. patella (samples L5 and L6 from southern PNG) L. bistratum (LV5 from eastern PNG), D. molle (samples E11-036 and E11-037 from adjacent locations in PNG), and Diplomospora sp. (one from Okinawa; sample E11-016 from PNG).

Illumina sequencing, assembling, binning, and annotation. The Illumina HiSeq 2000 data sets of L5 and L6 were previously reported (29). Illumina libraries were prepared for Okinawa, E11-037, E11-036, E11-016, and LV5. Libraries were sequenced using an Illumina HiSeq 2000 sequencer in 101-bp/125-bp paired-end runs (see Table S3 in the supplemental material). The library for E11-036 was sequenced using an Illumina MiSeq sequencer in 251-bp paired-end runs. Raw fasta files were assembled using IDBA_ud (30) (parameters: mink = 20, maxk = 100, step = 20, inner_mink = 10, inner_step = 5, prefix = 3, min_count = 2, min_support = 1, num_threads = 0, seed_kmer = 30, min_contig = 200, similar = 0.95, max_mismatch = 3, min_pairs = 3) in Futuregrid (http://futuresystems.org). The Prochloron genome was recovered from the metagenome by MetaAnnotator (31) in DIAG (http://diagnostic.org/). Thirty-one bacterial housekeeping genes were used to verify the completeness of each recovered genome. Prochloron genomes were annotated using the RAST server (32-34). Prochloron genomes P1 to P4 (from previously reported L. patella samples L1 to L4 [25]) were obtained from GenBank (accession numbers AGRF00000000, AFSJ00000000, AFSK00000000, and AGGA00000000).

Prochloron genomes were very similar between both D. molle samples and between L5 and L6 although only L6 and one D. molle sample were used in the analysis because of higher quality. Thus, out of 11 metagenomes, 9 were used for full analysis.

Selection of genes for phylogenetic analysis. Proteins encoded in Prochloron genomes were obtained using the RAST annotation output. Widespread genes present in a single copy in all genomes were identified by performing a blastp search using genome P1 against the eight other Prochloron genomes. Hits were defined as identity of >80%, query coverage of >0.8, and subject coverage of >0.8. Genes that met these criteria and were present in all nine samples were selected by the “regular expression” command (awk 'b[!$1ARGIND]+{i[f++a$1]=2=print1}'). The orthologous group for each marker was retrieved by a PerlScript (fasta_extract_mult.pl) (see the supplemental material) and used for phylogenetic analysis.

Prochloron 16S rRNA gene sequences were extracted from raw Illumina reads and amplified from genomic DNA extracted from ascidian samples. Primers AsCF2new (5′-CAAGGAAAGCGAGCGCCCGACAAAT) and AsCR5new (5′-GGCTGTTGTCACAAAGGGACGGGA) (28) were used to amplify 18S rRNA genes (see Table S4 in the supplemental material). The PCR product was cloned into a TOPO TA Zero Blunt vector (Invitrogen). Individual clones bearing the correct insert sizes were sequenced. Because of deep branching in the 18S sequences, species-specific primers were used in some cases: Diplosoma spp. (DiplosomaF, 5′-GGCTTCAGAACAGACTTG; DiplosomaR, 5′-GATGCCCTTTTACGGCG) (Lissoclinum spp. (LissoclinumF, 5′-TACCGACACTGCGAAAGGC; LissoclinumR, 5′-GCTCGATCCCAGAGGAC); and Didemnum spp. (DidemnumF, 5′-GCTACAGGTCGTCGGCGCCAACG; DidemnumR, 5′-TTCACGAGATTGAGTTCGGC).
RESULTS

Prochloron and host tunicates share congruent phylogenies. Host SSU 18S rRNA gene sequences provided a phylogenetic tree that reflected the known taxonomic relationship between dinemids (Fig. 1B) (23). In contrast, Prochloron 16S rRNA genes from the same samples led to a tree that substantially deviated from the 18S tree topology (see Fig. S1 in the supplemental material). Previously, this lack of congruence was interpreted as a lack of phylogenetic relationship between host and symbiont (23). However, while we were annotating genomes, we noticed trends suggesting a possible evolutionary relationship between host and symbiont at the whole-genome level. We proposed that the apparent stochastic distribution found in 16S genes could be an artifact resulting from the extremely close relationship between Prochloron genomes and consequent lack of variation in the 16S rRNA data set. We compared the sequenced metagenomes to test the hypothesis that Prochloron genomes reflect host phylogeny. Alternatively, the genomes could reflect geography, or there may be no relationship.

We concatenated 1,851 Prochloron genes that could be found in all nine samples and performed phylogenetic analysis. The resulting tree was congruent with the host 18S tree, providing support for our hypothesis (Fig. 1; see also Table S5 in the supplemental material). Trees were reconciled using Jane, version 4, providing strong support for the observed congruency (see Fig. S2). Jane, version 4, also provided evidence for a coevolution or cospeciation model, but the available evidence is also consistent with a host-switching model (see Discussion).

Statistical analyses of trees derived from individual functional categories of genes also supported phylogenetic congruency in each category (see Fig. S3 and Table S6 in the supplemental material). Most individual genes showed the same trend, but because of the extremely close relationships between Prochloron strains, insufficient variation was observed in most genes to draw firm conclusions. A limitation is the presence of certain classes of repetitive genes in Prochloron genomes (25). These genes were identified as described in Materials and Methods and eliminated from the analysis.

Gene content further reinforced the congruent phylogeny. Samples L1 and L4 are from clade A of L. patella; they share 5,175 genes (Fig. 2A). In contrast, L1 and L2 from different L. patella clades share only 3,393 genes. When representatives of the L. patella clades, D. molle, L. bistratum, and Diplosoma sp. were compared, 1,912 genes were shared in common (Fig. 2B).

A weakness of the metagenomic analysis was that we had insufficient sampling from the genera Diplosoma and Didemnum. DNA was obtained from an additional eight samples of Diplosoma and D. molle. We selected a Prochloron marker gene that was reliably found in the metagenomes, phylogenetically informative, and essential: psbU (coordinates 31698 to 32078 in contig 2414, GenBank accession number HQ407369.1) encoding an extrinsic protein of the photosystem II complex. The phylogenetic congruency hypothesis predicted that the phylogeny of psbU would be congruent with that of the tunicate 18S rRNA genes. Indeed, these trees were very similar, lending further support to the idea that Prochloron and host phylogenies are congruent (Fig. 3). In contrast, the observed phylogenies did not reflect differences in collection location or external environment. Taken together, phylogenetic evidence shows that Prochloron divergence is related to host divergence and not to other factors.
**Discovery of two new cyanobactin pathways.** All seven samples of *Lissoclinum* spp. examined in this study contained cyanobactin clusters. As previously described, L1 and L4 contained a total of at least three described *pat* pathways (with two in different locations in the L1 chromosome), while L2 and L3 contained a total of two *tru* pathways (25). We also found two new cyanobactin clusters, described below.

Using LC-MS, we previously showed that tunicates L5 and L6 contained patellins 3 and 5, while *L. bistratum* contained bistratamides A and E (28). We therefore predicted that these animals should contain specific types of cyanobactin pathways. The pathways should encode all of the enzymes needed to perform the posttranslational modifications present in the natural products, and the precursor peptide sequences should exactly match the amino acid sequences of the products. Indeed, each animal contained only a single cyanobactin biosynthetic pathway encoding the correct combination of enzymes, and the precursor peptides contained only amino acid sequences that encoded patellins 3 and 5 and bistratamides A and E. Therefore, L5 and L6 contained a novel *tru*-like cluster, *trf* (*tru* from the eastern fields of PNG; GenBank accession number KX100577), encoding patellins 3 and 5 (41), while *L. bistratum* contained a *pat*-like cluster, *bis* (bistratamides cluster; GenBank accession number KX100576), encoding bistratamides A and E (42, 43).

The *bis* genes and predicted proteins were highly similar to those from the *pat* pathway (44), with the exception of several with
more similarity to homologs from the ten pathway (Table 1). ten was previously found in free-living cyanobacteria from a different genus (21), so that the ten section could conceivably originate in a ten-like pathway in other Prochloron strains or from free-living cyanobacterial relatives. Similarity between bis and ten was observed at the predicted protein level, but significant similarity was not detected by BLAST searches at the nucleotide level.

The bis cluster encoded a single precursor peptide, BisE, that contained core peptides corresponding to bistratamides A and E. Interestingly, three core peptides were present, including two copies of the bistratamide A core and one copy of the bistratamide E core, while most tunicate cyanobactin precursors contain two core peptides. The L5/L6 trf cluster encoded a precursor peptide for patellins 3 and 5 and was otherwise nearly identical to tru, with one exception (Table 2).

Comparative analysis of tunicate cyanobactin clusters. All Prochloron cyanobactin pathways are 96 to 100% identical at the DNA and protein sequence levels in the 5' end of the pathway, encoding PatA, -B, -C, the substrate-binding domain of PatD, and intergenic sequences (Fig. 4; Tables 1 and 2). The L. patella pathways are about 99 to 100% identical in this region, even when samples are obtained from widely different regions of the Pacific Ocean, while the L. bistratum sample is 96% identical to the others. Similarly, in the extreme 3' end of the pathway, in the DUF domain of patG, pathways are 96 to 100% identical.

Between these extreme ends, significant recombination has occurred, generating many different functional classes of cyanobactin natural products (Fig. 5). (Here and elsewhere, the concept of recombination, and not a biochemical mechanism, is implied.) The second half of PatD, the YcaO domain, is either tru-like or pat-like; these two groups are about 74% identical to each other in the YcaO domain. The difference in the YcaO domains dictates whether cysteine and serine/threonine are heterocyclized (pat and bis) or whether only cysteine is heterocyclized (tru and trf) (45).

The leader peptides encoded by E genes are nearly identical among all pathways (Fig. 5A). The major differences between E genes are found in the core peptides, which are hypervariable, leading to many different compounds (21). Multiple E peptides

![FIG 2 Venn diagrams showing shared gene content between Prochloron samples. (A) Analysis of Prochloron from L. patella clades I (L1 and L4) and II (L2 and L3). (B) Analysis of representatives from L. patella clades (L1), D. molle (E11-036), L. bistratum (LV5), and Diplosoma sp. (E11-016). Values represent the numbers of genes.]

![FIG 3 Maximum-likelihood (ML) tree from analysis of the Prochloron psbU gene. The trees were generated by MEGA, version 7.0. Shapes indicate collection location as defined in Fig. 1C. Maximum-likelihood bootstrap values and Bayesian clade credibility values are indicated at the nodes (bootstrap values/clade credibility values).]

TABLE 1 bis cluster

<table>
<thead>
<tr>
<th>Locus</th>
<th>Protein homolog(s) or organism (% identity)</th>
<th>Gene homolog(s) (% identity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BisA</td>
<td>TruA (97), PatA (97)</td>
<td>truA* (98)</td>
</tr>
<tr>
<td>BisB</td>
<td>TruB (96), PatB (96)</td>
<td>truB (98), patB (98)</td>
</tr>
<tr>
<td>BisC</td>
<td>TruC (96), PatC (96)</td>
<td>truC (98), patC (98)</td>
</tr>
<tr>
<td>BisD</td>
<td>PatD (96), TruD (88)</td>
<td>patD (96), truD (87)</td>
</tr>
<tr>
<td>BisE</td>
<td>TenE (76)</td>
<td>NA</td>
</tr>
<tr>
<td>Leader only</td>
<td>TruE (100), PatE (97)</td>
<td>truE (96), patE (96)</td>
</tr>
<tr>
<td>BisF1</td>
<td>TenF (73)</td>
<td>NA</td>
</tr>
<tr>
<td>BisF2</td>
<td>PatP (58)</td>
<td>NA</td>
</tr>
<tr>
<td>BisG</td>
<td>TenG (68)</td>
<td>NA/rearrangements</td>
</tr>
<tr>
<td>Oxidase</td>
<td>Microcystis aeruginosa (78), TenG (73), PatG (73)</td>
<td></td>
</tr>
<tr>
<td>Pro tease</td>
<td>PatG (89), TruG (83)</td>
<td>patG* (85), truG (83)</td>
</tr>
<tr>
<td>DUF</td>
<td>PatG (96), TruG (94)</td>
<td>patG (96), truG (96)</td>
</tr>
<tr>
<td>Preoxidase region</td>
<td>PatG (85)</td>
<td>patG (85)</td>
</tr>
</tbody>
</table>

* Gene identity less than expected because of length differences.

b NA, not available; nucleotide sequences were not significantly identical in BLAST searches.

c Thirty-five amino acids.

d Domain boundaries as called by the Conserved Domains Database (CDD)/NCBI.

e Pseudogene.
containing different core sequences can exist within single animals. These different E peptides apparently exist as single copies within a set of otherwise complete, identical cyanobactin gene clusters. For example, the genome of L1 contains two identical cyanobactin biosynthetic pathways in separate regions of the chromosome, but with different core peptides. Because the E genes are sometimes present in differing quantities, it is clear that animals sometimes contain more than one strain of Prochloron. This provides fertile ground for generation of E variants.

Downstream of the E genes, the comparison becomes more complicated. pat from L1/L4 and tru from L2/L3 contain one and two different F genes, respectively. bis and trf each contain genes for two F proteins, one of which is a pseudogene in each case (Fig. 5). Only a single F protein is required for high yield in recombinant expression experiments, even when two such genes are present, so that a pseudogene second copy is consistent with the known functions of the pathways (9).

The F proteins are all different from each other and exhibit a complex pattern of substitution. In bis, the potentially functional F protein is more closely related to TenG from the tenuecyclamide pathway (free-living cyanobacteria) than to any Prochloron gene. In trf, TrfF1 is identical to TruF1, the prenylating protein in the tru pathway (46). Indeed, the trf products are prenylated, according to MS analysis (28). The F proteins have functional consequences in that TruF1 from L2/L3 and TrfF1 prenylate cyclic peptides, while the other F proteins have unknown (nonenzymatic) functions (9, 21). Potentially, they are required to bind or chaperone the cyclic peptides.

Finally, the G proteins contain macrocyclase domains (47). In pat and bis pathways in which the thiazoline moiety and potentially the oxazoline moiety are oxidized to thiazole/oxazole moieties, the G proteins also contain an oxidase domain. The G proteins have undergone a complex series of rearrangements where the oxidase domain has crossed in several times. In particular, the bis oxidase protein is more similar to that of ten in tenuecyclamides (48). Although PatG and TruG are able to macrocyclize hexapeptides in vitro, perhaps the efficiency is too low for in vivo biosynthesis, and a ten-like protein is therefore required.

<table>
<thead>
<tr>
<th>TABLE 2 trf cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locus</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>TrfA</td>
</tr>
<tr>
<td>TrfB</td>
</tr>
<tr>
<td>TrfC</td>
</tr>
<tr>
<td>TrfD</td>
</tr>
<tr>
<td>TrfE leader</td>
</tr>
<tr>
<td>TrfF1</td>
</tr>
<tr>
<td>TrfF2</td>
</tr>
<tr>
<td>TrfGb</td>
</tr>
<tr>
<td>DUF</td>
</tr>
</tbody>
</table>

a Consisting of 35 amino acids.

b Domain boundaries as called by the Conserved Domains Database (CDD)/NCBI.

c Pseudogene.

d NA, not available; nucleotide sequences were not significantly identical in BLAST searches.

e Except for the first 194 nucleotides.

FIG 4 Similarity of Prochloron cyanobactin pathways. Pathways obtained in this study were compared using blastn. Colors indicate percent identity between the DNA sequences, from red (65% identical) to green (100% identical), as indicated on the figure.
In phylogenetic trees, the A, B, and C proteins and the heterocyclase domain of D and DUF domain of G proteins are congruent with those of the tunicate hosts (Fig. 6). However, the swapped regions (YcaO domain of D, F proteins, and oxidase/macrocyclase domain of G) (Fig. 4) do not follow a phylogenetic pattern consistent with their horizontal origin but, instead, reflect chemical differences found in products.

**DISCUSSION**

Diversity-generating metabolism enables organisms to rapidly generate highly diverse chemical repertoires, which then undergo selection in response to challenge. Here, we obtained additional cyanobactin biosynthetic pathways from marine animals and demonstrated a phylogenetic congruence between symbiont and host. Using these data, we propose the evolutionary trajectory of cyanobactin biosynthesis. Two types of changes are readily apparent. First, substrates change by modifications to the core peptides encoded by the E genes. Second, enzymes are exchanged between otherwise identical pathways that expand the available posttranslational chemistry. Taken together, these changes lead to natural libraries of small molecules.

Here, using whole-genome data and a larger sample size, we show that there is a clear relationship between symbiont and host phylogeny. In contrast, there is no relationship between *Prochloron* and environmental habitat since adjacent animals found in the same habitats contain quite different *Prochloron* strains. In fact, several of the samples collected in this study were found in adjac-

**FIG 5** Recombination in the pathway generates many different functional classes of cyanobactin natural products. (A) Precursor peptide sequence alignments show that the leader peptide portion is nearly constant. At bottom, consensus (Con) and alternative substitutions (Alt) are shown. Important residues for recognition sequences I, II, and III are shown in yellow, blue, and gray, respectively. Hypervariable core peptide sequences are underlined in red. Note that each precursor encodes 2 or 3 core peptides. Total consensus and mutations for each position in the core peptide are shown in the box at right. (B) Pathway comparison in variable regions using tblastx. This represents an informative subset of alignments from the sequences shown in Fig. 4. (C) Cyanobactin structures synthesized by various pathways. Functional groups are indicated by the same colors as found in gene regions shown in panel B. Genetic changes shown in panels A and B are responsible for the resulting chemical differences.

**FIG 6** Comparison of the maximum-likelihood (ML) tree of the core cyanobactin genes with *Lissoclinum* spp. hosts. The trees were generated by MEGA, version 7.0. Maximum likelihood bootstrap values and Bayesian clade credibility values are indicated at the nodes (bootstrap values/clade credibility values). (A) ML tree of concatenated nucleotide sequences of the A, B, and C genes and heterocyclase domain of D and DUF domain of G genes. (B) ML tree of 18S rRNA nucleotide sequences from the same samples.
cent positions on the reef, yet *Prochloron* had more in common with samples from the same species collected thousands of kilometers away than with samples from the neighboring different animal species. This clearly indicates strong host specificity rather than specificity resulting from abiotic factors.

As a caveat, here we studied a subset of tunicates where the relationship with *Prochloron* is stable and probably obligate; there are other, more complex cases where the relationships may be different (27, 49). Exceptions to the rule may be found in further sampling, and species other than those studied here may exhibit different patterns.

The literature is replete with examples in which phylogenetic congruency and other biological data are used to infer coevolution or cospeciation (50). Some evidence obtained in this study suggests that *Prochloron* and hosts are coevolving or codiverging. For example, analysis with Jane, version 4, strongly supported the coevolution model. However, we do not believe that our data support coevolution for the following reasons. First, *Prochloron* secondary metabolite pathways are known to occur sporadically in a manner that is consistent with exchange between populations, rather than gene loss (51). Second, the host exhibits more sequence divergence than the symbionts. Normally, in coevolution one would expect the reverse to be true due to the faster doubling time of bacteria (52). Finally, *Prochloron* is absent from many branches of Didemmiidae that fall inside the 18S rRNA tree shown in Fig. 1, as well as from the most ascidian species in Didemmiidae in general. Common examples include *Didemnum vexillum*, *Lissoclinum perforatum*, and many others. (This may indicate symbiont loss as well.) An additional problem is that there is no fossil record for tunicates, making timing of potential coevolution events highly problematic and limiting the methods that can be employed.

Thus, it is unlikely that the phylogenetic congruency between symbiont and host reflects coevolution. Instead, we propose that the phylogenetic congruency reflects a host-switching model (53), in which specific *Prochloron* genotypes prefer specific host types, or vice versa. This preference could be mediated by cell surface proteins, which are extremely varied in or vice versa. This preference could be mediated by cell surface

The recombinational hypothesis is most obvious in substrate evolution in cyanobactin pathways. In substrate evolution, the core peptides diversify, while the enzymes remain constant. In several cyanobactin pathways of free-living organisms, multiple E genes exist in a tandem array, including several that are obviously nonfunctional (lacking core sequences) (14). Only a small subset of these is actually converted into products. This enables an accordion-like strategy of evolution, wherein E genes are copied and then recomposed to produce diversified core peptides within a single precursor. There is significant evidence of this in that *Prochloron* E genes have been identified encoding one, two, and three (bis) core peptides, indicating facile recombination.

A computational comparison of biosynthetic gene clusters across sequenced genomes has revealed that different families of gene clusters take different evolutionary trajectories (54). Among these, there are many examples of concerted evolution, in which genes are more similar to their orthologs within a strain than to the paralogs between strains. It was suggested that clusters or portions of clusters using concerted evolution may be a good starting point for biosynthetic engineering. The cyanobactin *put* and *tru* pathways similarly have several features that might be explained by concerted evolution, such as the recombination of precursor peptides and high similarities between specific portions of the biosynthetic clusters. These cyanobactin pathways are highly plastic and are well known for their ability to be engineered, providing some validation for the idea that pathways subject to concerted evolution are good for engineering. A special feature of the tunicate-derived pathways is their extremely close genetic identities, which allow the evolutionary pathways to be precisely defined (13, 21).

As a relevant contrast, several elegant studies show how closely related pathways diverge by swapping functional motifs (35–59). In these studies, the proteins are not as closely related to each other as those described here so that multiple mutations may be involved in creating functional swaps. Future work will determine whether these differences complicate engineering efforts or whether concerted evolution is a general indication of a diversity-generating metabolic pathway that is immediately useful for engineering.

In summary, here we provide evidence of the existence of diversity-generating metabolic pathways and propose their mode of evolution. Absolutely essential to this evolution, all enzymes in the cyanobactin pathways so far characterized are exceptionally broad-substrate tolerant (9). By swapping core peptides and enzymes, cyanobactin pathways rapidly (on an evolutionary time scale) access compounds with different functions. In turn, the resulting compounds are bioactive and are present as a major proportion of the extractable metabolites of the animals. This mechanism circumvents a presumed and widely cited limitation in the evolution of secondary metabolites, wherein structural changes usually lead to nonfunctional compounds and are thus proposed not to be evolutionarily favored. The presence of broad-
substrate biosynthetic pathways enables populations to access many different types of compounds without sacrificing the original, active metabolites. New compounds that confer a selective advantage may then become fixed, and in some circumstances the pathway may lose its diversity-generating character.

ACKNOWLEDGMENTS

Computer time was obtained on FutureGrid (developed and supported by NSF 0910812 and NSF DBI-0959894), National Science Foundation-funded MRI-R2 project DBI-0959894 (DIAG), and the University of Utah Center for High Performance Computing.

FUNDING INFORMATION

This work, including the efforts of Zhenjian Lin, Joshua Torres, Diarey Tianero, Jason Kwan, and Eric Schmidt, was funded by HHS | NIH | National Institute of General Medical Sciences (NIGMS) (GM107557 and GM102602).

REFERENCES


21. Kwan JC, Tianaero MDB, Donia MS, Wyche TP, Bugni TS, Schmidt EW. 2014. Host control of symbiotic natural product chemistry in cryptic pop-
49. McIntosh.
50. Nielsen.
51. Perez.

3460 aem.asm.org June 2016 Volume 82 Number 12

Applied and Environmental Microbiology

Lin et al.

46. Sardar, Ravel J.
50. Nicholson, Carmeli S.