PNAS

Main Manuscript for

Short macrocyclic peptides in sponge genomes

Zhenjian Lin[a], Vinayak Agarwal[b], Ying Cong[a], Shirley A. Pomponi[c], and Eric W. Schmidt[a]

[a]Department of Medicinal Chemistry, University of Utah, Salt Lake City, UT, 84112, USA;

[b]School of Chemistry and Biochemistry, and School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, 30332, USA

[c]Harbor Branch Oceanographic Institute, Florida Atlantic University, Fort Pierce, FL, 34946, USA

*Eric W. Schmidt; Vinayak Agarwal

Email: ews1@utah.edu; vagarwal@gatech.edu

Author Contributions: E.S. and Z.L. designed research; Z.L., V.A., Y.C. and S.P. performed research; E.S. and Z.L. wrote the paper.

Competing Interest Statement: The authors declare no competing interests.

Classification: natural products, biosynthesis, bioinformatics.

Keywords: sponges, cyclic peptide, circular peptide, RIPP, metazoan biosynthesis, stylissamide, phakellistatin

This PDF file includes:

Main Text

Figures 1 to 6

Abstract

Sponges (Porifera) contain many peptide specialized metabolites with potent biological activities and significant roles in shaping marine ecology. It is well established that symbiotic bacteria produce bioactive “sponge” peptides, both on the ribosome (RiPPs) and nonribosomally (NRPs). Here, we demonstrate that sponges themselves also produce many bioactive macrocyclic peptides, such as phakellistatins and related proline rich macrocyclic peptides (PRMPs). Using the *Stylissa carteri* sponge transcriptome, methods were developed to find sequences encoding 46 distinct RiPP-type core peptides, of which ten encoded previously identified PRMP sequences. With this basis set, the genome and transcriptome of the sponge *Axinella corrugata* was interrogated to discover 35 PRMP precursor peptides encoding 31 unique core peptide sequences. At least 11 of these produced cyclic peptides that were present in the sponge and could be characterized by mass spectrometry, including stylissamides A-D and seven previously undescribed compounds. Precursor peptides were encoded in the *A. corrugata* genome, confirming their animal origin. The peptides contained signal peptide sequences and highly repetitive recognition sequence-core peptide elements with up to 25 PRMP copies in a single precursor. In comparison to sponges without PRMPs, PRMP sponges are incredibly enriched in potentially secreted polypeptides, with >23,000 individual signal peptide encoding genes found in a single transcriptome. The similarities between PRMP biosynthetic genes and neuropeptides in terms of their biosynthetic logic suggests a fundamental biology linked to circular peptides, possibly indicating a widespread and underappreciated diversity of signaling peptide post-translational modifications across the animal kingdom.

Significance Statement

The translation from genomes to biology and medicine is not straightforward, in part because many gene products are modified after the resulting proteins are synthesized. Short peptides found in the animal kingdom such as hormones, neurotransmitters, and venom toxins sometimes have post-translational modifications essential for their activity. Here, we define a class of short peptides, produced by sponge animals, that are joined head-to-tail into large families of cyclic peptides. While previously circularization of short peptides was mostly associated with microbial and plant metabolism, it is also a major modification in animals.

Introduction

Nature applies many different strategies to synthesize macrocyclic peptides, which are intensely active in the clinic and are also important in the ecology of producing organisms. Among these, N-C terminal macrocyclic peptides (sometimes called “circular peptides”) are widespread in nature and are often associated with potent biological activities (1, 2). N-C cyclization is also thought to confer important pharmacological properties, such as improved serum stability (3). Consequently, N-C cyclic peptides are commonly sought motifs in pharmaceutical research and development. Among these, much interest has centered on proline rich macrocyclic peptides (PRMPs) from marine sponges (4). Ninety-three structurally distinct sponge PRMPs are described in the Marinlit database (https://marinlit.rsc.org/). Metabolomics data indicate that PRMP numbers are grossly underestimated, and that most PRMPs remains to be discovered (5). Cyclization confers a much different shape in the PRMPs than might be anticipated from short linear peptides, reinforcing the value and structural diversity resulting from N-C macrocyclization (6–8).

Biosynthetic approaches are crucial to the macrocyclic peptide field because they enable the genomics-based discovery of hidden chemical and biological diversity. In addition, the biosynthetic enzymes are useful in synthesis of constrained macrocycles and in generating gene-encoded cyclic peptide libraries. While the biosynthesis of several classes of N-C cyclic peptides has been well studied (9, 10), the biosynthetic origin of the PRMPs was unknown. Indeed, amide macrocyclic peptides from animals are essentially unknown outside of the θ-defensins from primates (11). When their biosynthesis is understood, animal cyclic peptides generally originate in symbiotic bacteria, and not in the animals themselves (12). Here, we set out to determine how PRMPs are made in sponge holobionts.

N-C cyclic peptides are sometimes biosynthesized by nonribosomal peptide synthetases (NRPSs), but frequently they are made on the ribosome, where they are often referred to as “ribosomally synthesized and posttranslationally modified peptide” (RiPP) natural products (13). During ribosomal biosynthesis, N-C macrocycles are first synthesized as precursor peptides. Maturation usually involves two proteases, one that cleaves off the leader peptide to release a free N-terminus, and a second that cleaves off a C-terminal sequence and performs the circularization (transpeptidation) step. The part of the precursor peptide that ends up in the mature, macrocyclic peptide is usually referred to as the “core peptide”. The presence of highly conserved N- and C-terminal recognition sequences flanking the core peptide enables the core to be hypervariable, leading to a wide variety of biochemical products made by simple and highly conserved enzymatic machinery (13).

In smaller N-C macrocyclic peptides <15 amino acids found in bacteria, circularization often requires a constrained heterocycle, such as proline or thiazoline (a cyclized cysteine residue), at the C-terminus (14) These residues promote turn formation in proteins and likely help to catalyze a productive conformation for internal cyclization. This is not universal, as some N-C cycles found in bacteria, plants, fungi and the θ-defensins in primates, do not follow this pattern; cyclization may sometimes be nucleated around cysteine knots (14).

The sponge PRMPs are <15 amino acids and contain multiple proline residues. We thus envisioned that they would be synthesized via a RiPP-like mechanism similar to the short peptides found in bacteria and fungi. Here, we describe evidence strongly supporting this hypothesis. Numerous short peptides are encoded in sponge genomes, and not in bacterial symbionts, demonstrating that short macrocyclic peptides are found throughout the kingdom of life. The precursor peptides resemble those found in bacteria and fungi, including signal peptides, recognition sequences, and multiple cores encoded in a single precursor peptide. Overall, these motifs are very similar to those found in neuropeptides, except that they lead to N-C cyclization.

Results

Identification of PRMP precursor genes

In the sponge holobiont, either bacterial symbionts or the sponge itself might be responsible for making macrocycles. We first investigated the hypothesis that PRMPs are produced by the microbes using a *Stylissa massa* sponge specimen from Guam, which contains PRMPs (15). We applied our standard workflow (16) that has successfully identified many biosynthetic pathways in the associated microbial genomes of marine animals, yet we could not find any evidence that bacterial symbionts

produce PRMPs. In those methods, we deeply sequence the metagenome and perform a variety of automated and manual analyses. For example, in this specific case such an analysis would include obtaining bacterial contigs using metagenome binning and looking for all possible RiPP core peptide sequences that might encode the desired products, among many other steps that are used. No candidate RiPP or NRPS sequences were identified, suggesting that the PRMPs might instead be synthesized by the sponge host.

To investigate this hypothesis, we focused on the sponge species *Stylissa carteri*, as it is rich in PRMPs (17), and its transcriptome data are available in NCBI. Although we believed it likely that these compounds are RiPPs, we exhaustively searched for viable NRPS candidates in the transcriptome but did not identify any candidate proteins. To search for RiPPs, we aimed to find the precursor peptide sequences within the transcriptomes. We obtained the PRMPs from sponge family Axinellidae and genus *Stylissa* in the Comprehensive Marine Natural Products Database (18). The cyclic peptide sequences were linearized, considering all possibilities, and then used as a query with 15 *S. carteri* translated transcriptome assemblies using blastp. Forty of the query sequences had exact matches in 79 predicted protein sequences from *S. carteri*, providing candidate PRMP precursor peptides. However, the use of short query sequences resulted in nonspecific hits, which were challenging to distinguish because of the multiple possible linear sequences encoded in each cyclic peptide. Thus, we required additional methods to determine which of the 79 hits represented valid precursor peptides.

Precursor peptides for cyclic RiPPs contain conserved (often, identical) recognition sequences that direct the proteolysis and cyclization steps. However, absent prior knowledge of what those sequences might be, such repetitive sequences, or *motifs*, are difficult to discover. Tools such as MEME-ChIP (19) are highly useful in this regard since they enable discovery of repeated memes/motifs without prior knowledge. We applied MEME-ChIP (19) to identify whether any of the *S. carteri* hit precursor peptides share similar motifs. Sixteen of the PRMP precursor candidates shared at least two conserved motifs. Notably, multiple copies of the same motif were observed within a single precursor, which is a key characteristic of cyclic RiPP precursors across kingdoms of life (20) (Figure S1). The predicted motifs were manually aligned and corrected, leading to the identification of a highly conserved motif, "FMPDEVKKQ". Between these motifs, putative core sequences were identified, some of which were identical to sequences known to encode chemically characterized PMRPs, including those for axinastatins 4 and 5, hymenamide H, phakellistatins 2, 13, and 16, stylissamides C and X, and stylissatins B and C (Figure 1A). In total, 22 contigs were identified in the *S. carteri* transcriptome that contained the "FMPDEVKKQ" motif with variable core peptide sequences. The *S. carteri* transcriptomes encoded 46 unique core peptides, ten of which were identical to the sequences of previously isolated, purified, and chemically characterized PRMPs. While the recognition sequences were highly conserved, the core peptides were hypervariable, another feature that is commonly observed in cyclic RiPPs across kingdoms. In addition, the N-termini of the precursor peptides contained signal peptide sequences, easily identifiable by standard signal peptide tools. Signal peptides are found in fungal RiPP precursor peptides(21). Together, these factors led us to propose the hypothesis that the identified RiPP precursor peptides encoded sponge PRMPs.

Genome-guided discovery of novel PRMPs

To test the biosynthetic hypothesis, we aimed to use the precursor peptides predictively in the forward discovery of novel compounds. This method has been applied previously in providing validation for biosynthetic pathways in complex, non-model organisms (12, 22) We obtained *Axinella corrugata* from Florida; although axinellid sponges often contain PRMPs, they have not been previously reported from this species. We sequenced and assembled the metagenome and transcriptome of *A. corrugata*. The precursor peptides identified in *S. carteri* were truncated into "recognition-core-recognition" sequence subunits, which were then aligned and used to construct a profile hidden Markov model (HMM) (Supporting Information TEXT1 and TEXT2). Hmmsearch (23) using this model led to the identification of 35 precursor peptides, including 31 unique core sequences, in the *A. corrugata* transcriptome. We then obtained the chemical extract of *A. corrugata* and interrogated it by mass spectrometry, using a previously reported method (5). We observed 11 of the predicted core peptides, the sequences of which were confirmed by MS^2 fragmentation (Figure 2C and S2). Four of these peptides (cyclic-VYPYKPP, IYPFPPP, FIPYPFP, and FIPYPLP) are compounds stylissamides A-D (7, 24–26), respectively, which were previously isolated from *Stylissa caribica*. The remaining seven were previously unreported compounds, demonstrating a high chemical novelty discovered by genomic prediction. These known compounds also have very similar MS^2 spectra to those reported in the literature, further supporting the chemical methodology.

When these compounds have been previously isolated from sponges, they have been found in ~0.005-0.05% of dry weight. We wondered why only 11 PRMPs were observed, while 31 core sequences were found. The number of core peptide reads in the transcriptomes was compared with the MS counts for each peptide. PRMP compounds were identified for the ten most abundant transcripts, and the MS counts correlated with reads count (Pearson 0.7, *p*-value 0.0078) (Figure 3). Therefore, the compounds observed by MS represented the most abundantly transcribed core peptides in the sponge. The excellent correlation between highly expressed RiPP genes in the sponge and the discovery of the resulting previously undescribed compounds in the sponge extracts strongly supports the hypothesis that the compounds are sponge genome encoded. The relative amounts of cyclic peptides observed by MS are similar to those from previously reported PRMP-containing sponges (27).

A hypothesis for PRMP biosynthesis

Like other RiPP cyclic peptide precursor peptides, PRMP precursors contain core sequences flanked by recognition sequences; like many eukaryotic peptides, they also contain signal sequences (Figures 2 and 4). Based upon these features, it can be anticipated that the precursor is synthesized on the ribosome and targeted to endoplasmic reticulum (ER), where the signal peptide is removed by signal peptidase. The identified signal peptide sequences contain the Ser-Arg motif that is cleaved by the signal peptidase (28). It is likely that a second protease would then recognize the Lys-Lys-Gln motif and related sequences found in the recognition sequence, releasing free N-termini that can be cyclized. A second protease would recognize the Phe-Met-Pro motif and related recognition sequence elements, performing the C-terminal cleavage and transpeptidation/cyclization to afford the natural product. Recognition sequence-facilitated steps are known to occur in bacteria, plants, and fungi, and their expansion here to the animal kingdom reveals that they are ubiquitous biochemical elements. A Pro residue is found at the C-terminus of all core peptides for which PRMPs could be found. Proline

residues are known to facilitate peptide macrocyclization in the same manner in which they lead to turn formation in proteins (29). Finally, PRMP production in the ER would facilitate downstream processes, such as secretion in response to an external signal. Eukaryotic neuropeptides also share several of these features, including the presence of multiple hormone sequences on a single precursor, the presence of protease recognition elements, and the signal peptide (30).

Genetic architecture and evolution of sponge PRMP precursor peptides

Transcriptome analysis revealed that the diversity of PRMPs encoded in sponge genomes greatly exceeds the number of PRMP natural products that have been reported to date. Because of the repetitive nature of the precursor peptides and their encoding genes, our analysis of GenBank sequences of axinellid sponges revealed what appeared to be fragments of the full-length precursor peptides. To obtain high-quality, full-length genes, the metagenome of *A. corrugata* was sequenced using MinION Nanopore. MinION reads (10x coverage) were corrected using trimmed Illumina short reads to obtain an assembly (294 Mbp bp, total number of sequences: 6975; N50:141,632 bp, GC %: 44.75 %). The eukaryotic contigs were selected using the Autometa taxonomy identification pipeline and annotated using AUGUSTUS with the transcriptome data as training data. The resulting sequences confirmed the repetitive nature of the PRMP genes identified in *A. corrugata*.

In total, 10 *A. corrugata* contigs were found to contain eleven intact PRMP precursor peptide genes (Figure 5A), which encode 26 different core peptides (Figure 5B). Each precursor peptide encoded between 2 and 6 discrete core peptides. Because in most cases identical core peptides were repeated more than once, the number of cores observed in the 10 precursor peptides spanned 2 to 25 (average number of cores per precursor: 10.4). Thus, the precursor peptides were highly complex and highly repetitive.

Intriguingly, cores sharing similar sequences are exclusively found within the same precursor (Figure 5B). This suggests that tandem gene duplication may play a significant role in the diversification of PRMP precursor genes. The introns in these genes also exhibit a highly similar sequence pattern, characterized by an "AT" rich region at the 5'-end and "AC" repeats at the 3'-end (Figure S3). These conserved repeats within the introns may serve as recombination sites during tandem duplication events, leading to the generation of highly diverse core peptides. The introns are positioned between the signal peptide and the first recognition sequence, as well as prior to the last "P" within each core peptide sequence (Figure 5B). This positioning highlights the importance of maintaining a proline residue at the C-terminus of the core peptides during recombination, which are found in all of the previously described PRMPs. The structural feature of PRMP cores flanked by conserved introns would make it possible to generate diverse peptides by the production of multiple mRNA isoforms from a single gene through alternative splicing.

Confirmation that PRMP precursor genes are encoded in the sponge genome

The identified precursor peptide genes contained long introns, consistent with their origins in the sponge genomes rather than in bacteria. The precursor genes were also assembled into sponge-specific contigs, suggesting an origin in the sponge genome. To further confirm the origin of the PRMP precursor genes in *A. corrugata*, two approaches were utilized: 1) blastp-based Lowest Common Ancestor (LCA) analysis; and 2) metagenome binning.

LCA analysis used contigs that encoded PRMP precursor peptides. The contigs were analyzed using AUGUSTUS to predict 340 genes present on the contigs (31). For each predicted gene, LCA was calculated using the "autometa-taxonomy-lca" command in the Autometa tool (32). The number of available sponge reference genes is very limited, so many LCAs could not be called with certainty; nonetheless, 113 of the predicted genes were identified with a marine sponge LCA (Figure S4).

We also applied a binning pipeline (33) to classify contigs resulting from the metagenome sequence. Initially, the taxonomy of each contig was determined using Autometa (32). Binning was performed using t-SNE algorithm based on tetramer distribution and sequencing coverage. A large bin was distinct from several, much smaller bins (Figure 5). The large bin contained the majority of contigs from the metagenome assembly. It was identified as a sponge bin due to the presence of all sponge-derived contigs identified by Autometa, such as those containing 18S rRNA genes and the sponge mitochondrial genome. All contigs containing PRMP precursors were also found within the sponge bin. Conversely, bacterial contigs identified by Autometa comprised the smaller bins and contained only a small portion of the sequence. These results are consistent with previous reports that describe *Axinella* sp. as having a low abundance of microbial communities (34). Consequently, the origin of the PRMPs was unambiguously assigned to the sponge.

Signal peptide containing repeat proteins in marine sponges

PRMP precursor peptides have two distinctive features: a signal peptide and highly repetitive core sequences. Based on the characteristics, we bioinformatically surveyed the distribution of RiPP-like precursor peptides in marine sponge transcriptome data sets. We selected 83 sponge transcriptome sequence read archive (SRA) data sets, which included 80 different species. The protein sequences from each SRA assembly were first filtered by length (20-400 amino acids), and then submitted to SignalP 6.0 (35) for prediction of signal peptide containing proteins. The sponges are rich in signal peptides: hundreds to thousands of signal peptides were detected in each species, leading to identification of 304,407 signal peptides in the 80 species (Table S1).

Sequences encoding signal peptides were then analyzed for the presence of tandem repeats using XSTREAM (36). Three criteria were applied to the analysis to enrich for RiPP-like peptides: the tandem repeat is not in the signal peptide region; the length of the tandem repeat is at least 7 amino acids; and the number of tandem repeats is greater than 2. In the 80 species, 1,158 peptides met these criteria, comprising <1% of signal peptides in most of the sponge species.

In our *A. corrugata* sequence, 23,577 peptides contained a signal sequence, while 135 peptides met the criteria to be RiPP-like, including 20 PRMP precursors (Table S1).

Peptides containing signal sequences and tandem repeats are widely distributed in marine sponges: 96% of the species analyzed here contained such sequences. However, these peptides lack significant sequence similarity to each other. Protein similarity network analysis by EFI-EST (37) (Figure S5) showed that, among the 1,293 detected signal peptides, only 261 of them are in clusters that are shared by 59 different sponge species. Each cluster likely represents a family of related precursor peptides that may share similar functions. For example, the 21 detected PRMP precursors

are clustered together in the network analysis. This reveals that, in addition to the PRMP precursors, there are other families of precursor peptides detected in *A. corrugata*, observed as singlet nodes in the network (Figure S5 and S6).

Discussion

Neuropeptides are an ancient and widely distributed form of neuronal communication found in both cnidarian and bilaterian animals (38). While sponges lack neurons, interestingly, certain common proneuropeptides found in other animals, such as phoenixin (PNX) and nesfatin (39), are also detected in sponge transcriptomes. However, it is important to note that, to date, the peptides encoded by such transcripts have not been detected in sponges (39). An exception to this is the discovery of ribosomally synthesized linear peptides, barrettides (40), which were found in the deep-sea sponge *Geodia barretti*. Intriguingly, the PRMP precursors show no similarity to any of the proneuropeptide families reported in previous studies (38), not even among most of the sponge species with publicly available genomic data. PRMP precursor genes exhibit significantly higher transcription levels, approximately ten times more than the transcription of the phoenixin gene in *Axinella*. The structures of PRMPs are distinct from those of barrettides or other predicted neuropeptides in sponges, as PRMPs undergo post-translational modifications in the form of C-N macrocyclization. The diversity of core peptides within PRMPs suggests a wide range of biological functions, which might therefore play crucial ecological roles in sponges. In contrast to the core peptides, the signal peptides and recognition sequences exhibit a high degree of conservation among different PRMP precursor peptides, even across two sponge genera (Figure 1). This observation suggests a biosynthetic plasticity in the biosynthetic machinery in marine sponge, where the essential enzymes involved in signal recognition and peptide cleavage and cyclization are maintained across sponge species.

PRMPs are distributed across various marine sponges with diverse geographical and phylogenetic backgrounds (5). The finding of similar PRMPs in both *S. caribica* and *A. corrugata* further supports their proposed phylogenetic relationship (41). Both *Axinella* and *Stylissa* are polyphyletic genera (42, 43). *A. corrugata* (Order Axinellida, Family Axinellidae) and some species of *Stylissa* (Order Scopalinida, Family Scopalinidae) are proposed to be more closely related to species in the order Agelasida than to other species of Axinellidae (41, 43). Other studies have reported the occurrence of similar compound classes in *S. caribica* and *A. corrugata* (44) . The two species are very similar in appearance and can only be distinguished after careful microscopical study of the architecture of the sponge. Further morphological and molecular systematic studies are required to confirm the phylogenetic relationship of these species to each other and to the order Agelasida and, indeed, whether they are two distinct species.

PRMPs are relatively low in abundance compared to other metabolites found in the same species. The lower abundance coupled with relatively high transcription and the presence of secretion signals suggests a role for PRMPs as secreted peptides, in analogy to neuropeptides or secreted peptide hormones/venoms. Further research is required to determine the potential physiological and ecological roles of these diverse sponge peptides, which are currently unknown.

The low abundance and high sequence diversity of PRMPs makes their isolation and structure elucidation challenging. Recent advancements in mass spectrometry-based methods have revealed a much greater diversity of PRMPs in marine sponges than previously recognized (5). Here, we identified the biosynthetic pathway of these cyclic peptides, specifically the RiPPs type. Through the survey of transcriptome data from a single sponge species, we achieved a significant increase in the diversity of PRMPs. This exciting discovery has expanded our understanding of the potential variety and complexity of cyclic peptides in marine sponges. Such a genomic strategy in discovering peptidic metabolites has the potential to address the biomass challenges associated with natural product discovery from marine animals. This work along with recent advances (45, 46) is leading to a greater appreciation that the sponge animal itself is a natural product synthesis factory, in addition to the well validated contribution of the microbiome. This is especially true for low microbial abundance sponges, wherein the rarified abundance of the microbiome likely is not a major contributor to natural products detected in the holobiont extracts.

Materials and Methods

RNA extraction and transcriptome sequencing

Live specimens of *Axinella corrugata* were collected by scuba on June 17, 2022 off the coast of Fort Lauderdale, Florida (latitude 26.15N, longitude -80.09W) at a depth of 20 m. The collection was permitted under Florida Fish and Wildlife Conservation Commission Special Activity License SAL-20-2233-SR. The samples were processed using our previously described pipeline (47). Briefly, <2 mm^2 tissue slices are homogenized and processed to obtain polyA-selected cDNA, and then sequenced at ~450 bp insert size to 100 M read pairs. Data were assembled as previously described (47).

Genome sequencing

A. corrugata gDNA from the homogenized tissue was extracted using Qiagen DNeasy Blood & Tissue Kit. Illumina library preparation and sequencing was performed at the HCI-HTG. Sequencing library preparation was performed using an NEBNext Ultra II DNA Library Prep Kit with a 450 bp mean insert size. Sequencing used an Illumina NovaSeq 6000 sequencer with 2 x 150 bp runs. Raw reads were trimmed and adaptors removed by trimmomatic. Long reads sequencing library was prepared following the protocol of Genomic DNA by Ligation (SQK-LSK110) and sequenced in nanopore R9.4.1 flow cell.

The raw long reads from nanopore were corrected using Ratatosk (48) using the short Illumina reads, and then assembled using Flye (49). The animal genes were predicted using AUGUSTUS 3.3 (50) with the transcriptome assembly as training data.

Metagenome binning

The metagenome contigs were filtered by length (≥3 kb) and annotated by the module make_taxonomy_table.py in autometa (v2.0.0). The tetranucletotide composition of each contig was calculated the Perl script tetramers.pl described in the YAMB package (51) (v2.1.0.0). The coverage for each contig was determined using bbwrap with Illumina reads. The t-distributed stochastic neighbor embedding (t-SNE) dimensionality reduction and sequential DBSCAN data clusterization were performed. The clusters were examined by the taxonomy annotation result from Autometa.

SRA data preparation

SRA fastq raw reads for sponge *S. carteri* (SRR1038495, SRR1038496, SRR1038497, SRR1738063, SRR1738064, SRR1738065, SRR1738069, SRR1738070, SRR1738071, SRR1738092,

SRR1738093, SRR1738094, SRR1738097, SRR1738098, SRR1738099) were obtained from NCBI. Raw reads were processed as described above.

Transcriptome mining for RIPPs precursor genes: First, known PRMPs from sponges *Stylissa* and *Axinella* were obtained from CMNPD (18) All possible linear sequences for each macrocyclic peptide were prepared. All possible linear sequences were pooled and used as query sequences. Query sequences were searched using the blastp algorithm against the SRA assembly database, with an e-value threshold of 1e-10. Hits that met the criteria (qcov=100, perc_identity=100) were selected and considered as precursor gene candidates. The hit sequences were analyzed with meme-chip (19) using the following parameters: "-meme-minw 4 -meme-maxw 11 -meme-mod anr -meme-nmotifs 3". Sequences with common repetitive motifs were manually checked for their potential to be precursor genes.

Signal peptide sequence prediction

Protein sequences from each transcriptome assembly were filtered with a maximum length of 300 amino acids. The filtered protein sequences were analyzed using a standalone version SignalP6.0 (35). the predicted signal peptide containing proteins sequences were examined for repeat sequence using XSTREAM (36) with command "java -Xmx1000m -Xms1000m -jar xstream.jar -i1 -N -g0 -L6 -z".

Sponge chemical analysis

Sponge specimens were frozen and lyophilized to dryness. Dry sponge samples were soaked in 1:1 methanol:methylene chloride (1 mL solvent per 50 mg sponge dry weight) for 48 h. The organic solvent was withdrawn, centrifuged to remove debris, and analyzed directly by liquid chromatography/mass spectrometry (LC/MS) using an Agilent 1290 ultra high-performance liquid chromatography instrument coupled to a Bruker ImpactII high resolution time of flight mass spectrometer operated using an electrospray ionization source. Samples were analyzed in the positive ionization mode. Methods and data analysis proceeded as previously described (5) without deviation.

Data Availability. All data used in this study are present in the Manuscript and Supporting Information. *A. corrugata* raw sequencing data were deposited in GenBank under Bioproject number PRJNA1001903, and accession numbers are provided for all other sequences used.

Acknowledgments

This work was funded by NSF CHE 2003756 to EWS and CHE 2004030 to VA. The authors are thankful to N. Garg at Georgia Institute of Technology for use of mass spectrometry instrumentation.

References

1. R. Ribeiro, E. Pinto, C. Fernandes, E. Sousa, Marine Cyclic Peptides: Antimicrobial Activity and Synthetic Strategies. *Mar Drugs* **20** (2022).
2. S. H. Joo, Cyclic Peptides as Therapeutic Agents and Biochemical Tools. *Biomol Ther (Seoul)* **20**, 19 (2012).
3. C. Ngambenjawong, H. H. Gustafson, M. Sylvestre, S. H. Pun, A Facile Cyclization Method Improves Peptide Serum Stability and Confers Intrinsic Fluorescence. *Chembiochem* **18**, 2395–2398 (2017).

4. W. Y. Fang, R. Dahiya, H. L. Qin, R. Mourya, S. Maharaj, Natural Proline-Rich Cyclopolypeptides from Marine Organisms: Chemistry, Synthetic Methodologies and Biological Status. *Mar Drugs* **14** (2016).
5. I. Mohanty, *et al.*, Enzymatic Synthesis Assisted Discovery of Proline-Rich Macrocyclic Peptides in Marine Sponges. *ChemBioChem* **22**, 2614–2618 (2021).
6. G. R. Pettit, *et al.*, Isolation and Structure of the Marine Sponge Cell Growth Inhibitory Cyclic Peptide Phakellistatin 1. *J Nat Prod* **56**, 260–267 (1993).
7. R. Mohammed, J. Peng, M. Kelly, Mark. T. Hamann, Cyclic Heptapeptides from the Jamaican Sponge *Stylissa caribica*. *J Nat Prod* **69**, 1739–1744 (2006).
8. P. Hosseinzadeh, *et al.*, Comprehensive computational design of ordered peptide macrocycles. *Science (1979)* **358**, 1461–1466 (2017).
9. S. Sarkar, W. Gu, E. W. Schmidt, Expanding the chemical space of synthetic cyclic peptides using a promiscuous macrocyclase from prenylagaramide biosynthesis. *ACS Catal* **10**, 7146 (2020).
10. B. Franke, J. S. Mylne, K. J. Rosengren, Buried treasure: biosynthesis, structures and applications of cyclic peptides hidden in seed storage albumins. *Nat Prod Rep* **35**, 137–146 (2018).
11. L. Leonova, *et al.*, Circular minidefensins and posttranslational generation of molecular diversity. *J Leukoc Biol* **70**, 461–4 (2001).
12. M. Morita, E. W. Schmidt, Parallel lives of symbionts and hosts: chemical mutualism in marine animals. *Nat Prod Rep* **35**, 357–378 (2018).
13. M. Montalbán-López, *et al.*, New developments in RiPP discovery, enzymology and engineering. *Nat Prod Rep* **38**, 130 (2021).
14. L. Cascales, D. J. Craik, Naturally occurring circular proteins: distribution, biosynthesis and evolution (2010) https:/doi.org/10.1039/c0ob00139b (July 4, 2023).
15. J. Sun, W. Cheng, N. J. de Voogd, P. Proksch, W. Lin, Stylissatins B–D, cycloheptapeptides from the marine sponge Stylissa massa. *Tetrahedron Lett* **57**, 4288–4292 (2016).
16. N. A. Nguyen, *et al.*, An Obligate Peptidyl Brominase Underlies the Discovery of Highly Distributed Biosynthetic Gene Clusters in Marine Sponge Microbiomes. *J Am Chem Soc* **143**, 10221–10231 (2021).
17. A. H. Afifi, *et al.*, Carteritins A and B, cyclic heptapeptides from the marine sponge Stylissa carteri. *Tetrahedron Lett* **57**, 1285–1288 (2016).
18. C. Lyu, *et al.*, CMNPD: a comprehensive marine natural products database towards facilitating drug discovery from the ocean. *Nucleic Acids Res* **49**, D509–D515 (2021).
19. P. Machanick, T. L. Bailey, MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).
20. G. M. Rubin, Y. Ding, Recent advances in the biosynthesis of RiPPs from multicore-containing precursor peptides. *J Ind Microbiol Biotechnol* **47**, 659–674 (2020).

21. M. Umemura, *et al.*, Characterization of the biosynthetic gene cluster for the ribosomally synthesized cyclic peptide ustiloxin B in Aspergillus flavus. *Fungal Genet Biol* **68**, 23–30 (2014).
22. J. A. McIntosh, Z. Lin, M. D. B. Tianero, E. W. Schmidt, Aestuaramides, a natural library of cyanobactin cyclic peptides resulting from isoprene-derived Claisen rearrangements. *ACS Chem Biol* **8**, 877–883 (2013).
23. J. Mistry, R. D. Finn, S. R. Eddy, A. Bateman, M. Punta, Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* **41**, e121 (2013).
24. X. Wang, B. I. Morinaka, T. F. Molinski, Structures and Solution Conformational Dynamics of Stylissamides G and H from the Bahamian Sponge *Stylissa caribica*. *J Nat Prod* **77**, 625–630 (2014).
25. C. Cychon, M. Köck, Stylissamides E and F, Cyclic Heptapeptides from the Caribbean Sponge *Stylissa caribica*. *J Nat Prod* **73**, 738–742 (2010).
26. S. Scarpato, *et al.*, New Tricks with an Old Sponge: Feature-Based Molecular Networking Led to Fast Identification of New Stylissamide L from Stylissa caribica. *Mar Drugs* **18**, 443 (2020).
27. C. Cychon, G. Schmidt, M. Köck, Sequencing of cyclic peptides by NMR and MS techniques demonstrated on stylissamides A–F. *Phytochemistry Reviews* **12**, 495–505 (2013).
28. S. M. Auclair, M. K. Bhanu, D. A. Kendall, Signal peptidase I: Cleaving the way to mature proteins. *Protein Sci* **21**, 13 (2012).
29. W. Gu, S. H. Dong, S. Sarkar, S. K. Nair, E. W. Schmidt, The biochemistry and structural biology of cyanobactin biosynthetic enzymes. *Methods Enzymol* **604**, 113 (2018).
30. A. Corbière, *et al.*, Strategies for the Identification of Bioactive Neuropeptides in Vertebrates. *Front Neurosci* **13** (2019).
31. M. Stanke, B. Morgenstern, AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* **33**, W465 (2005).
32. I. J. Miller, *et al.*, Autometa: automated extraction of microbial genomes from individual shotgun metagenomes. *Nucleic Acids Res* **47**, e57 (2019).
33. P. D. Scesa, Z. Lin, E. W. Schmidt, Ancient defensive terpene biosynthetic gene clusters in the soft corals. *Nat Chem Biol* **18**, 659–663 (2022).
34. J. R. White, *et al.*, Pyrosequencing of Bacterial Symbionts within Axinella corrugata Sponges: Diversity and Seasonal Variability. *PLoS One* **7**, e38204 (2012).
35. F. Teufel, *et al.*, SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology 2022 40:7* **40**, 1023–1025 (2022).
36. A. M. Newman, J. B. Cooper, XSTREAM: A practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics* **8**, 1–19 (2007).

37. J. A. Gerlt, *et al.*, Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim Biophys Acta* **1854**, 1019 (2015).
38. G. Jékely, Global view of the evolution and diversity of metazoan neuropeptide signaling. *Proceedings of the National Academy of Sciences* **110**, 8702–8707 (2013).
39. L. A. Yañez-Guerra, D. Thiel, G. Jékely, Premetazoan Origin of Neuropeptide Signaling. *Mol Biol Evol* **39** (2022).
40. K. Steffen, *et al.*, Barrettides: A Peptide Family Specifically Produced by the Deep-Sea Sponge *Geodia barretti*. *J Nat Prod* **84**, 3138–3146 (2021).
41. C. C. Morrow, *et al.*, Congruence between nuclear and mitochondrial genes in Demospongiae: A new hypothesis for relationships within the G4 clade (Porifera: Demospongiae). *Mol Phylogenet Evol* **62**, 174–190 (2012).
42. E. Gazave, *et al.*, Polyphyly of the genus Axinella and of the family Axinellidae (Porifera: Demospongiaep). *Mol Phylogenet Evol* **57**, 35–47 (2010).
43. B. Alvarez, MichaeL. D. Crisp, F. Driver, J. N. A. Hooper, RoB. W. M. Van Soest, Phylogenetic relationships of the family Axinellidae (Porifera: Demospongiae) using morphological and molecular data. *Zool Scr* **29**, 169–198 (2000).
44. A. Galitz, Y. Nakao, P. J. Schupp, G. Wörheide, D. Erpenbeck, A Soft Spot for Chemistry–Current Taxonomic and Evolutionary Implications of Sponge Secondary Metabolite Distribution. *Mar Drugs* **19**, 448 (2021).
45. K. Wilson, *et al.*, Terpene biosynthesis in marine sponge animals. *Proceedings of the National Academy of Sciences* **120** (2023).
46. E. P. Stout, Y.-G. Wang, D. Romo, T. F. Molinski, Pyrrole Aminoimidazole Alkaloid Metabiosynthesis with Marine Sponges Agelas conifera and Stylissa caribica. *Angewandte Chemie International Edition* **51**, 4877–4881 (2012).
47. Z. Lin, F. Li, P. J. Krug, E. W. Schmidt, The polyketide to fatty acid transition in the evolution of animal lipid metabolism. Research Square (2023).
48. G. Holley, *et al.*, Ratatosk: hybrid error correction of long reads enables accurate variant calling and assembly. *Genome Biol* **22**, 1–22 (2021).
49. M. Komolgorov*, et al*., metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* **17**, 1103-1110 (2020).
50. O. Keller, M. Kollmar, M. Stanke, S. Waack, A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**, 757–763 (2011).
51. A. Korzhenkov, YAMB: metagenome binning using nonlinear dimensionality reduction and density-based clustering. *BioRxiv*, 521286 (2019).
52. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

Figures

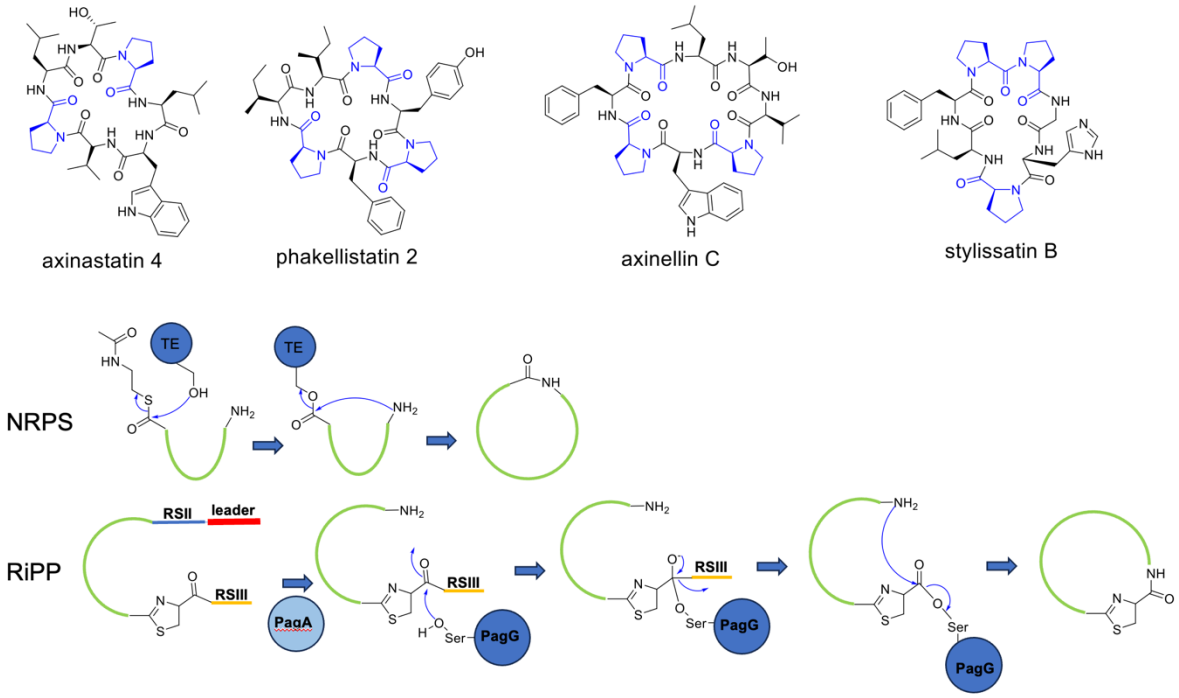

Figure 1. **Structures of marine sponge derived PRMPs and two common enzymatic mechanisms of peptide macrocyclization.** NRPS: nonribosomal peptide synthetase. TE: thioesterase domain responsible for macrocyclizing some nonribosomal peptides. RiPPs: ribosomally synthesized and post-translationally modified peptides. RSII and RSIII: recognition sequences that are used by protease such as PagA and PagG in bacteria to N-C cyclize short peptides.

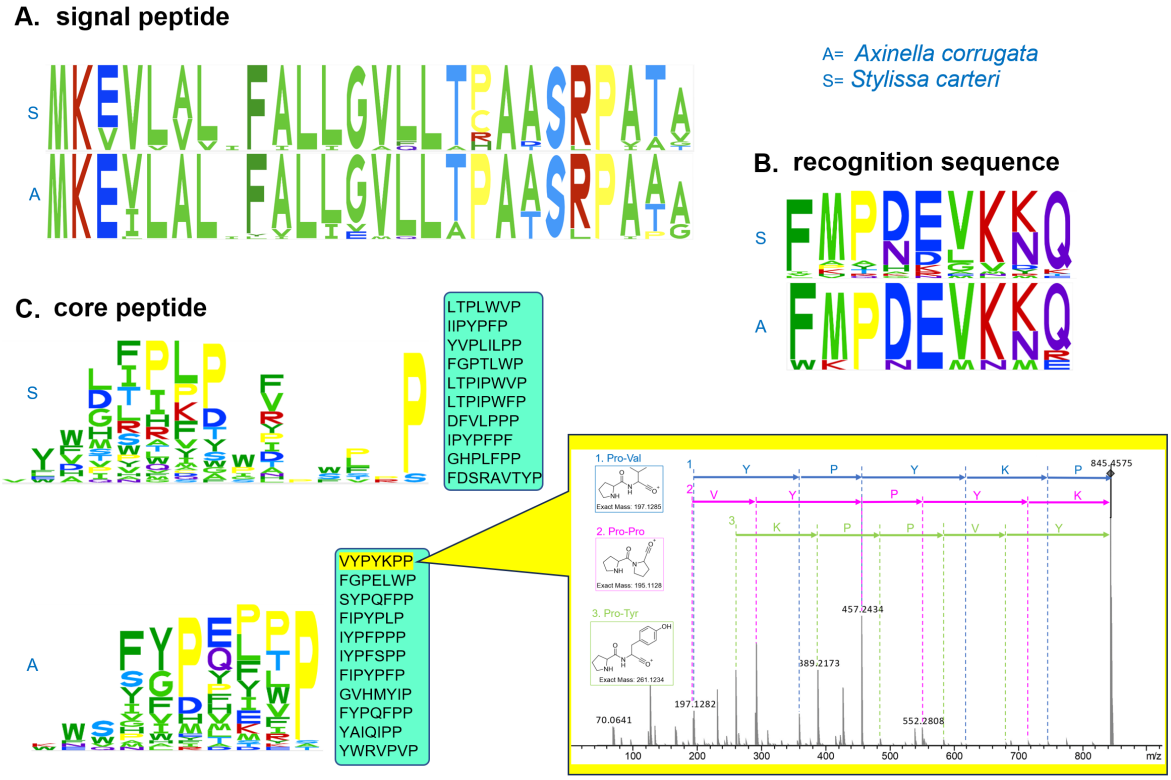

Figure 2. Sequence logos of **A**) identified signal peptides, **B**) recognition sequences and **C**) core sequences from PRMPS precursor peptides. The core sequences in the box for *S. carteri* are identical to those of known sponge PRMPs, the core sequences in the box for *A. corrugata* are experimentally identified cyclic peptides by MS^2 sequencing, which was examplified by the identification of cyclic (VYPYKPP). Starting from each of the three different dipeptide fragments, ^{N}Pro-Val^{C} (1, MS2 *m/z* 197), ^{N}Pro-Pro^{C} (2, MS2 *m/z* 195), and ^{N}Pro-Tyr^{C} (3, MS2 *m/z* 261), the same cyclic sequence, cyclo(VYPYKPP), was recovered.

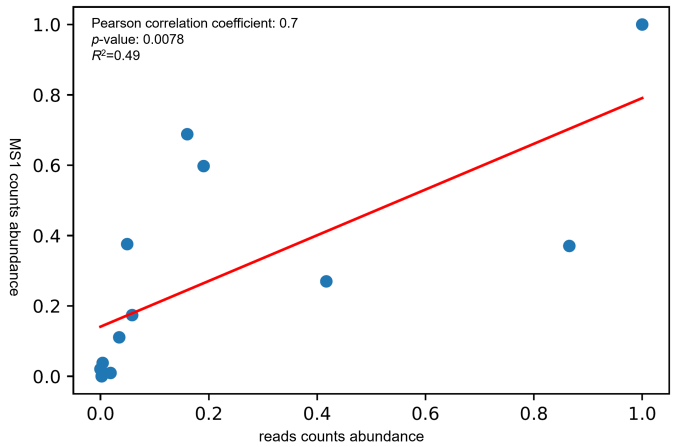

Figure 3. Relative transcriptional abundance of the top 13 transcribed cores and their correlation with the corresponding counts detected in LC-MS analysis. Each dot represents a core peptide sequence (see Table S2).

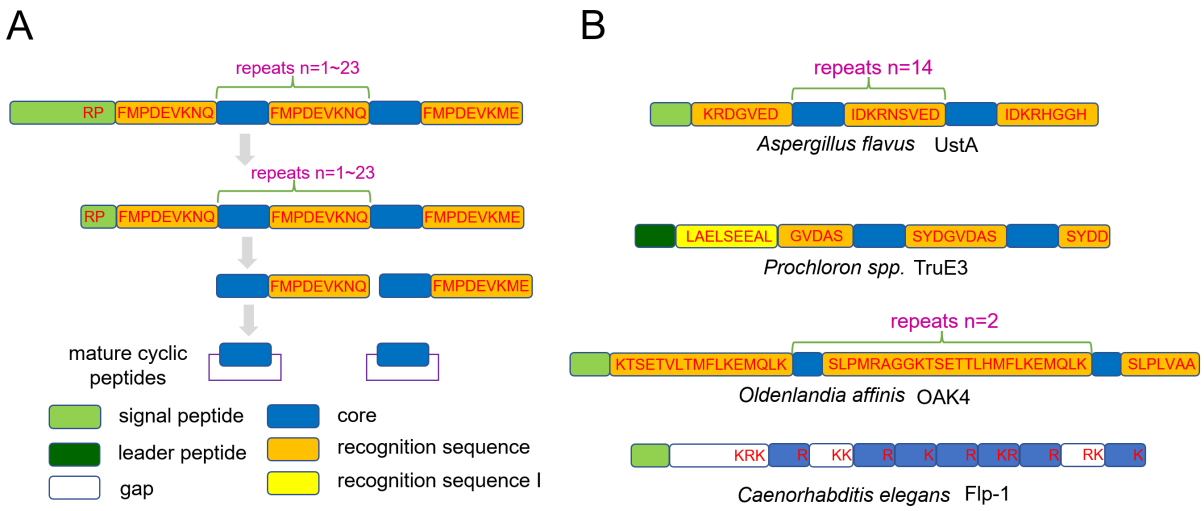

Figure 4. **Architecture of multicore-containing RiPP precursor peptides. A)** General architecture of sponge PRMP precursor peptides. **B)** For comparison, representative multicore-containing RiPPs from fungal (*Aspergillus*), bacterial (*Prochloron*), and plant (*Oldenlandia*) kingdoms, and a neuropeptide from animals.

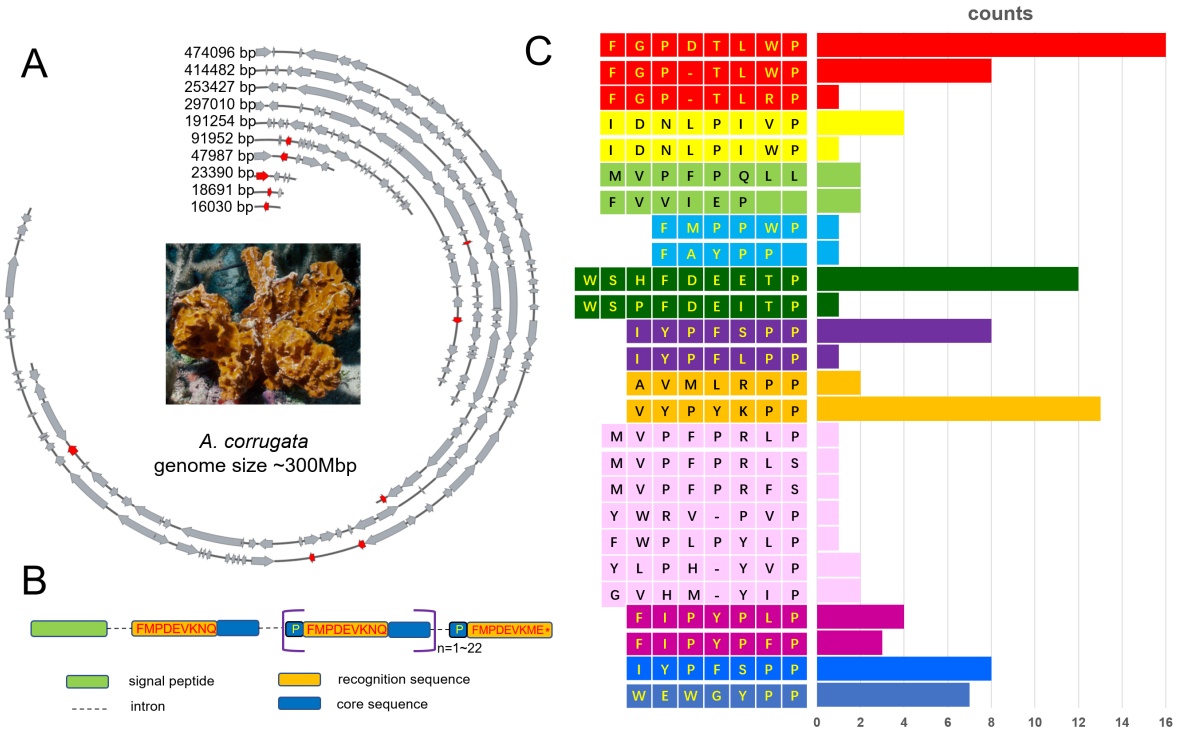

Figure 5. PRMP biosynthetic genes in the *A. corrugata* genome. A) PRMP precursor distribution in *A. corrugata* contigs. The red ORFs represent PRMP precursor genes, gray ORFs represent other genes. **B**) Diagram of PRMP precursor peptide gene architecture in *A. corrugata*. Repetitive sequences are also observed in introns (Figure S3). **C)** The core peptides found in the intact precursor peptides from genome assembly. Each different color represents a single precursor peptide. The numbers on the x-axis indicate the counts for each core peptide in the corresponding precursor peptide.

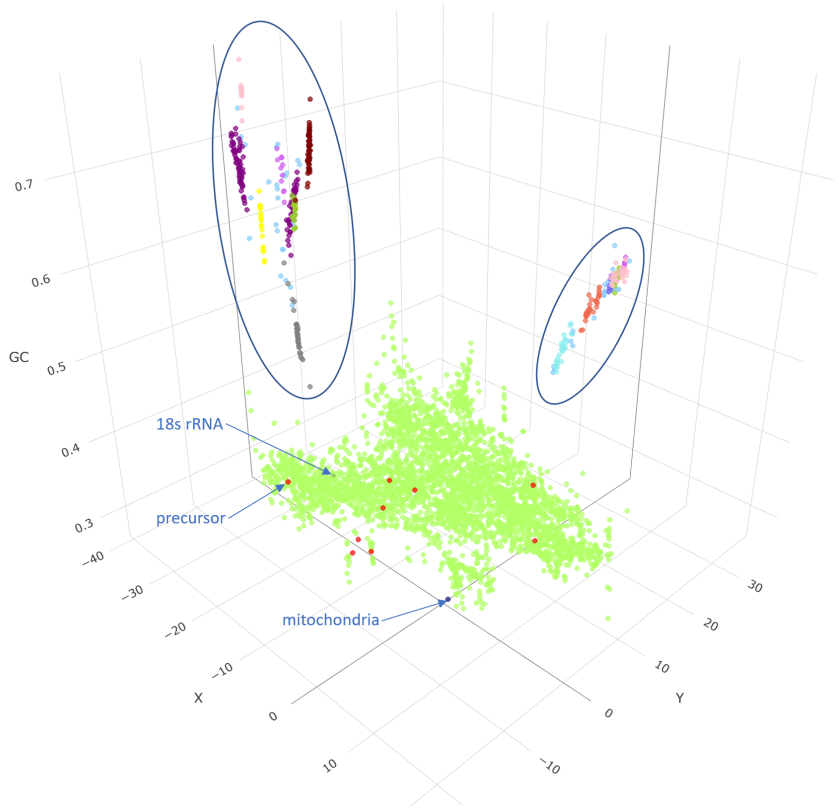

Figure 6. 3D plot of bins for *A. corrugata* metagenome. Each point represents a contig in the metagenome. They are plotted on the two dimensions that result from dimension-reduction by BH-tSNE, versus the GC content of the contig in the third dimension. The two circled bins are comprised of bacterial contigs, while the light green dots represent contigs from the sponge genome. The sponge contigs containing PRMP precursor genes are shown in red.