A Global Approach to Estimating the Abundance and Duplication of Polyketide Synthase Domains in Dinoflagellates

Evolutionary Bioinformatics
Volume 17: 1–24

DOI: 10.1177/11769343211031871
SAGE

Ernest P Williams, Tsvetan R Bachvaroff and Allen R Place

Institute of Marine and Environmental Technologies, University of Maryland Center for Environmental Science, Baltimore, MD, USA.

ABSTRACT: Many dinoflagellate species make toxins in a myriad of different molecular configurations but the underlying chemistry in all cases is presumably via modular synthases, primarily polyketide synthases. In many organisms modular synthases occur as discrete synthetic genes or domains within a gene that act in coordination thus forming a module that produces a particular fragment of a natural product. The modules usually occur in tandem as gene clusters with a syntenic arrangement that is often predictive of the resultant structure. Dinoflagellate genomes however are notoriously complex with individual genes present in many tandem repeats and very few synthetic modules occurring as gene clusters, unlike what has been seen in bacteria and fungi. However, modular synthesis in all organisms requires a free thiol group that acts as a carrier for sequential synthesis called a thiolation domain. We scanned 47 dinoflagellate transcriptomes for 23 modular synthase domain models and compared their abundance among 10 orders of dinoflagellates as well as their co-occurrence with thiolation domains. The total count of domain types was quite large with over thirty-thousand identified, 29 000 of which were in the core dinoflagellates. Although there were no specific trends in domain abundance associated with types of toxins, there were readily observable lineage specific differences. The Gymnodiniales, makers of long polyketide toxins such as brevetoxin and karlotoxin had a high relative abundance of thiolation domains as well as multiple thiolation domains within a single transcript. Orders such as the Gonyaulacales, makers of small polyketides such as spirolides, had fewer thiolation domains but a relative increase in the number of acyl transferases. Unique to the core dinoflagellates, however, were thiolation domains occurring alongside tetratricopeptide repeats that facilitate protein-protein interactions, especially hexa and hepta-repeats, that may explain the scaffolding required for synthetic complexes capable of making large toxins. Clustering analysis for each type of domain was also used to discern possible origins of duplication for the multitude of single domain transcripts. Single domain transcripts frequently clustered with synonymous domains from multi-domain transcripts such as the BurA and ZmaK like genes as well as the multi-ketosynthase genes, sometimes with a large degree of apparent gene duplication, while fatty acid synthesis genes formed distinct clusters. Surprisingly the acyl-transferases and ketoreductases involved in fatty acid synthesis (FabD and FabG, respectively) were found in very large clusters indicating an unprecedented degree of gene duplication for these genes. These results demonstrate a complex evolutionary history of core dinoflagellate modular synthases with domain specific duplications throughout the lineage as well as clues to how large protein complexes can be assembled to synthesize the largest natural products known.

KEYWORDS: Dinoflagellate, Polyketide, Toxin, Hidden Markov Model, Clustersing, Gene Duplication

RECEIVED: October 11, 2020. **ACCEPTED:** June 23, 2021.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded in part by a grant from NOAA-NOS-NCCOS-2012-2002987 to A.R.P. This paper is Contribution No. 5963 from the University of Maryland Center for Environmental Science and No. 21-004 from the Institute of Marine and Environmental Technology.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Ernest P Williams, Institute of Marine and Environmental Technologies, University of Maryland Center for Environmental Science, 701 East Pratt Street, Baltimore, MD 21202, USA. Email: williamse@umces.edu

Introduction

Dinoflagellates are unicellular aquatic eukaryotes with an interesting and complicated evolutionary history.[1,2] Generally speaking, they can be divided into 2 main groupings with the heterotrophic, often parasitic syndiniales at the base of the dinoflagellate lineage and the often mixotrophic "core dinoflagellates" extending out into the distal branches.[2] All of the core dinoflagellates have a chloroplast or evidence of a lost chloroplast with multiple symbiotic events occurring throughout the lineages.[3,4] Although many core dinoflagellates are mixotrophic,[5,6] the majority of dinoflagellate "algae" that form harmful algal blooms are photosynthetic, *Noctiluca scintillans* (Macartney) being the exception. Toxic dinoflagellates are exclusively photosynthetic and there is evidence that toxin synthesis may initiate in the chloroplast,[7,8] indicating a potential relationship between photosynthesis and natural product synthesis in dinoflagellates. *Amphidinium carterae* (Hulbert) is a basal, photosynthetic dinoflagellate that makes the toxin amphidinol as well as many derivatives termed amphidinolides,[9,10] indicating that the acquisition of a plastid and toxicity are early events in the evolution of the core dinoflagellates. Many dinoflagellate toxins pose human health concerns by a variety of mechanisms[11] as well as ecological and trophic impacts.[12,13]

The toxins themselves are almost universally polyketides, that is, they are formed from sequentially added acetate subunits that are modified prior to the addition of the next acetate subunit.[14] The workhouse enzymatic domain in the synthesis of polyketides is the ketosynthase (KS) domain, a condensation domain that incorporates malonyl-CoA into an existing acyl chain as acetate with the release of CO_2 driving the reaction.[15,16] Analogous to this reaction are non-ribosomal peptide

synthases that also perform a condensation type reaction with substrate specificity provided by an adenylation domain that binds a carboxylic acid, often an amino acid, and passes it to the condensation domain for incorporation.[17,18] The enzymes that incorporate each building block work along with modifying domains to form synthetic modules that can create very complex biomolecules and are responsible for many known naturally occurring compounds including antibiotics.[19-21] Labeling studies have shown that dinoflagellate toxins exclusively incorporate acetate from malonyl-CoA,[10,22-25] unlike bacteria that often incorporate similar subunits such as propionate or butyrate,[26] and the toxins often start with (or are occasionally extended by) small amino acids like glycine or other carboxylic acids like glycolate.[24,27] There is also evidence for alkylation by methionine or acetate as well as side-by-side "alpha" carbons from acetate explained by the deletion of carbon by a theorized Favorskii rearrangement removing the beta carbon from one acetate.[28] Toxins range in complexity and size from the 31 carbon Gymnodimine[29] to the 164 carbon maitotoxin that has 98 stereocenters.[30]

In spite of their complexity, synthesis of the backbone of dinoflagellate toxins utilizes the same core machinery as lipid synthesis. Lipids as a secondary metabolite are differentiated from natural products in that they are fully saturated, highly regulated, and usually synthesized and modified in the chloroplast, mitochondrion, and cytosol.[31-33] Thus, there is frequently a segregation of genes, phylogenetically and physically, involved in lipid synthesis from those involved in secondary metabolite synthesis, including in dinoflagellates.[34] In terms of acetate incorporation, all dinoflagellate toxins and lipids rely on the aforementioned ketosynthase domains along with several biologically universal modification domains: ketoreductases (KRs), dehydratases (DHs), and enoyl reductases (ERs) to form a sequentially reduced backbone structure and acyl transferases (ATs) and thioesterases (TEs) to move and terminate growing acyl chains. These enzymatic domains interact with the substrate and each other via a reaction center created by transferring the phosphopantetheinyl arm of coenzyme A onto a carrier protein.[35,36] One key difference between lipid and other secondary metabolite synthesis is that lipid synthesis is iterative, utilizing a single carrier protein called the acyl carrier protein while natural products are made with multiple modules with a homologous carrier domain called a thiolation domain. Whether the particular chemistry of each module is a PKS, an NRPS, or a hybrid system; a thiolation domain acts as the reaction center for all of these modular synthases. Likewise, a thiolation domain would be the reaction center for each module involved in toxin synthesis in dinoflagellates. This is useful when dealing with dinoflagellates since the type I multi-domain polyketide synthases and non-ribosomal peptide synthases found in fungi[37] and usually associated with eukaryotes are relatively uncommon in dinoflagellate transcriptomes with most transcripts containing one or rarely a few domains that would have to be combined into a multi-enzyme synthetic complex,[38] similar to the type II polyketide synthases and non-ribosomal peptide synthases usually found in prokaryotes.[17,39] This is not surprising since dinoflagellates often encode genes as tandem repeats of gene copies rather than gene clusters of common metabolic function,[40] but this also makes phylogenetic reconstruction difficult even for single domains due to a high copy number of very similar sequences.

The exceptions to the multitude of single domain transcripts in dinoflagellates are several multi-domain genes that have conserved domain arrangement and sequence. Two of these are the BurA and ZmaK-like genes[41] that contain both adenylation and ketosynthase domains in what appears to be a single module. There is also a multi-module gene usually containing at least 3 consecutive ketosynthase containing modules, here referred to as the triple KS.[38,42] Phylogenies of dinoflagellate modular synthase domains usually form a robust set of dinoflagellate clades but with poor support placing these clades among eukaryotic outgroups, as well as no obvious reflection of relationships within core dinoflagellates.[34,43,44] This not only reveals a gap in annotated sequences that can function as outgroups to dinoflagellates but also indicates that at least some of these modular synthases are likely of bacterial origin, specifically BurA and ZmaK, which have only been described in prokaryotes and seem to have been transferred in their entirety.[45,46] Thus, with the exceptions of the conserved fatty acid biosynthetic genes and the corresponding acyl carrier protein, phylogenetic comparisons to model eukaryotes or prokaryotes are generally uninformative when trying to deduce the roles of dinoflagellate modular synthases in toxin production. Likewise, the traditional nomenclature of polyketide synthases that relies on single versus multi-domain and prokaryote versus eukaryote fails to describe the domains in dinoflagellates in a useful manner.

The primary aim of this study was to survey genes that may be involved in dinoflagellate natural product synthesis, specifically toxins, without prejudice from what has been described in prokaryotes or distantly related model eukaryotes. *Amphidinium carterae* was used as a model because it is a basal toxic dinoflagellate[2] and has the BurA and ZmaK like genes as well as the triple KS gene in their apparent entirety and single copy.[41,42] The domains selected were taken from these previously annotated multi-domain dinoflagellate transcripts resulting in several unexpected discoveries such as a large number of adenylation domains seemingly without the traditional condensation domains as well as scaffolding domains associated with specific synthetic domains. This global approach was also able to describe the relative copy number of each synthetic domain revealing several atypical relationships. One example is a large number of enoyl reductases compared to dehydratases, which is very strange since enoyl reductases theoretically act downstream and should be less abundant than dehydratases. The second portion of this survey was to place the retrieved domains

into theoretical functional bins based on sequence similarity using a method that is not hampered by gene duplication and horizontal gene transfer. Although many of these synthetic domains can and have been given hypothetical function using phylogenetic inference with model systems as outgroups, the results presented here demonstrate many novel sequence clusters that are difficult to resolve phylogenetically as well as some very atypical gene expansions including acyl transferases and ketoreductases involved in lipid synthesis that were largely overlooked in previous studies. The results of this study demonstrate another way in which dinoflagellates defy the paradigms established by model systems, in this case in terms of the mechanisms of natural product (toxin) synthesis, and are presented here as a framework to be used in future biochemical experiments to validate the hypothetical functions of PKS and NRPS genes in dinoflagellates.

Materials and Methods

Transcriptome preparation and analysis

A total of 61 initial transcriptomes were selected for domain searches with the majority of dinoflagellate transcriptomes taken from the CAMERA database, originally published in (http://camera.calit2.net/mmetsp/list.php,[47] and now hosted at https://www.imicrobe.us/#/projects/104 NCBI project #PRJNA231566, as assembled contigs using Trinity. In addition, data for cultures of *Karenia brevis* (C.C. Davis), *Karlodinium veneficum* (D. Ballantine), and *Akashiwo sanguinea* (K. Hirasaka) that were infected with the syndinean parasite of the genus *Amoebophyra* were collected from previous phylogenetic studies.[1,48,49] For *A sanguinea* the transcriptome was done with and without infection and for the *K. veneficum* parasite there is a genome available for comparison.[50] In addition to these transcriptomes the deep sequencing transcriptomes (using Hi-Seq) for *K. brevis*,[38] and 2 *Gambierdiscus* species,[51] *G. excentricus* (S. Fraga), and *G. polynesiensis* (Chinain and M. Faust) that were assembled using CLC (595M, 118M, 884M reads, respectively) were included. Unfortunately the transcriptomes from the 2 *Gambierdiscus* species in the transcriptome sequence archive were incomplete with about 70 PKS genes identified in the initial study deposited separately in Genbank. These were added back into the total domain count following domain searches. Each transcriptome was translated in all 6 frames using a Perl script and Genbank translation Table 1 (standard eukaryotic) prior to analysis.

Benchmarking Universal Single-Copy Orthologs (BUSCO)[52] was used to determine transcriptome quality using the Eukaryota odb9 dataset with final scores (single and multiple copy orthologs) ranging from 1.7% for *Perkinsus chesapeaki* to 86.1% for *Alexandrium tamarense* (Lebour) and a median of 77.4% (Table 1). The eukaryote database was chosen over the protist database after initial tests with the protist database gave very low scores (approximately 30% maximum, data not shown). This study is intentionally specific to the "core" dinoflagellates so only closely related outgroups were used including *Perkinsus marinus* (Levine), *Chromera velia* (R.B. Moore et al), and *Triceratium dubium* (Brightwell), as well as the syndinean parasite of crustaceans *Hematodinium* sp. and the other aforementioned syndiniales with dinoflagellate hosts. No pertinent domains were found in the transcriptomes of syndinean parasites with dinoflagellate hosts except for 3 transcripts from the *K. veneficum* parasite and were thus excluded from further analyses giving a total of 46 transcriptomes with a BUSCO score 64% or greater that were included in the final tabulations following domain searches. The 2 *K. brevis* transcriptomes using different assembly methods had similar BUSCO scores and so the Trinity assembled transcriptome was selected for the final tabulations to make comparisons with other transcriptomes, the majority of which were assembled using Trinity, more informative. *Oxyrrhis* and all outgroup species were given their own taxonomic bin. The forty remaining ingroup transcriptomes were placed into 7 taxonomic bins at approximately the ordinal level including the Gonyaulacales (10 species), the Thoracosphaerales (*Brandtodinium*), the Prorocentrales (3 species), the Peridiniales (10 species), the Dinophysiales (2 species), the Noctilucales (*Noctiluca*), and the Gymnodiniales (8 species) with the Suessiales (5 additional species) as a subgrouping of the Gymnodiniales. The 64% cutoff was chosen as a natural observed breakpoint for transcriptomes that had a full repertoire of domains relative to other closely related species (Supplemental Figure S1). Some of the outgroup species had lower BUSCO scores (*P. marinus* 30%, *C. velia* 54.8%, *T. dubium* 41.5%) than the 64% cutoff. Although the scores were low, these transcriptomes were included since most of the tabulations are based on ratios and the domain searches successfully recovered transcripts with synthetic modules, for example, 183 domain hits for *T. dubium* and 104 domain hits for *C. velia*. Also, BUSCO analysis of the *P. marinus* genome (Genbank Bioproject PRJNA12737) yielded a completeness score of 53.3% indicating that the alveolate sequences may not be well represented in the BUSCO database and/or that parasitism has resulted in gene reduction. A lack of sequence representation in the BUSCO database is also supported by maximum BUSCO scores of approximately 80%, even for deeply sequenced transcriptomes showing that the BUSCO scores could be used as a guide but were not quantitative.

HMM assembly and domain searches

Amphidinium carterae (Hulbert) was used to create dinoflagellate specific hidden Markov models (HMMs) of domains from modular synthases. Although many robust models exist for model species, protists in general are poorly sampled and with almost no experimental verification, predictions based on those models are difficult. Four transcripts of multi-domain synthases from the *A. carterae* transcriptome were used (Figure 1). Each is readily found in other dinoflagellate taxa with the

Table 1. Transcriptome BUSCO scores and domain content.

SPECIES	COMBINED § (%)	SINGLE§ (%)	DUPLICATE § (%)	DOMAINS	DOMAIN TYPES
Alexandrium andersonii	33.30	25.70	7.60	1155	541
Alexandrium catenella	68.30	54.10	14.20	1244	584
Alexandrium margalefi	73.00	59.10	13.90	1555	801
Alexandrium minutum	27.40	22.10	5.30	169	91
Alexandrium monilatum	80.20	57.40	22.80	1292	678
Alexandrium tamarense	86.10	49.80	36.30	1739	956
Karlodinium veneficum	78.90	55.80	23.10	1227	638
Amphidinium carterae	78.60	66.70	11.90	727	388
Amphidinium klebsii	80.20	66.70	13.50	720	393
Amphidinium massartii	77.30	67.70	9.60	614	314
Akashiwo sanguinea	83.10	46.50	36.60	1497	848
Azadinium spinosum	80.20	57.80	22.40	2122	1108
Brandtodinium nutriculum	64.60	52.10	12.50	859	420
Ceratium fusus	81.50	60.40	21.10	1066	653
Chromera velia	54.80	47.50	7.30	104	66
Crypthecodinium cohnii	79.20	63.40	15.80	1231	716
Dinophysis acuminata	71.00	53.80	17.20	1590	787
Durinskia baltica	81.10	45.50	35.60	801	387
Gambierdiscus excentricus	74.30	63.70	10.60	847	448
Gyrodinium instriatum	80.90	53.50	27.40	2110	1147
Glenodinium foliaceum	81.80	46.20	35.60	1078	532
Gonyaulax spinifera	50.90	38.00	12.90	883	412
Gambierdiscus polynesiensis	68.30	59.70	8.60	1378	766
Gymnodinium catenatum	83.20	63.70	19.50	579	325
Hematodinium sp.	77.20	33.30	43.90	724	471
Heterocapsa arctica	64.00	51.80	12.20	797	398
Heterocapsa rotundata	65.70	57.10	8.60	712	343
Karenia brevis_CLC	80.90	64.70	16.20	2006	1197
Karenia brevis_Trinity	83.50	61.40	22.10	1526	939
Kryptoperidinium foliaceum	84.80	36.00	48.80	1699	823
Lingulodinium polyedra	80.80	58.40	22.40	2407	1304
Noctiluca scintillans	77.50	65.00	12.50	625	324
Oxyrrhis marina (LB1974)	75.20	60.70	14.50	399	240
Oxyrrhis marina (unknown)	79.50	62.00	17.50	461	290
Pelagodinium beii	68.00	53.10	14.90	945	462
Peridinium aciculiferum	78.80	58.70	20.10	846	443

(Continued)

Table 1. (Continued)

SPECIES	COMBINED § (%)	SINGLE§ (%)	DUPLICATE § (%)	DOMAINS	DOMAIN TYPES
Perkinsus chesapeaki	1.70	1.70	0.00	16	9
Perkinsus marinus	30.00	23.10	6.90	75	40
Prorocentrum hoffmanianum	79.60	58.10	21.50	1178	581
Prorocentrum micans	78.30	47.90	30.40	1210	696
Prorocentrum minimum (1329)	76.90	48.50	28.40	678	370
Prorocentrum minimum (2233)	39.20	30.00	9.20	276	155
Polarella glacialis (2088)	55.80	44.20	11.60	594	298
Protoceratium reticulatum	68.30	54.10	14.20	1247	592
Pyrodinium bahamense	73.30	61.40	11.90	1138	598
Scrippsiella hangoei	81.90	57.10	24.80	1234	724
Scrippsiella hangoei_like	82.50	62.00	20.50	1002	532
Scrippsiella trochoidea	77.80	55.40	22.40	1554	906
Symbiodinium (B1)	80.90	70.00	10.90	676	385
Symbiodinium (C1)	84.80	62.00	22.80	199	109
Symbiodinium (C15)	35.30	32.00	3.30	828	418
Symbiodinium (2430)	57.10	49.50	7.60	568	278
Symbiodinium (421)	66.30	30.00	36.30	1352	659
Symbiodinium (D1a)	46.90	28.40	18.50	558	307
Symbiodinium (Mp)	81.60	70.00	11.60	766	368
Symbiodinium (A)	61.40	52.50	8.90	752	381
Triceratium dubium	41.50	32.30	9.20	183	98

§ Percentages shown are the fraction of BUSCO genes retrieved by one (Single), multiple (Duplicate), or any number (Combined) of transcripts in each transcriptome.

same domain arrangement. The first (comp6001_c0_seq1) is a hybrid PKS/NRPS, the BurA-like gene described in the bacterial genus *Burkholderia* that participates in the synthesis of burkholderic acid.[45] It has an unusual domain order containing 2 thioesterase, 2 thiolation, an adenylation (described by the NCBI conserved domain database as an acyl-CoA ligase), a ketosynthase, a ketoreductase, and an acyl-transferase domain. The second (comp26075_c0_seq1) is also a hybrid PKS/NRPS that is most similar on a sequence basis to the ZmaK gene described in *Bacillus cereus* to act in the synthesis of zwittermicin.[46] It contains 2 thiolation, an adenylation, an acyl-transferase, a ketosynthase, a ketoreductase, a dehydratase, an enoyl reductase, and a FSH1 serine hydrolase domain. While the BurA-like gene has the same domain content and arrangement as the genes from *Burkholderia*, the domain arrangement of the ZmaK-like gene is not similar to the ZmaK gene from *B. cereus* making predictions about substrates or function in dinoflagellates unreliable. The third multi-domain transcript is a straightforward multiple ketosynthase-containing set of overlapping transcripts (comp305_c0_seq1 and comp32615_c0_seq1) that have a total of 4 thiolation domains and 3 possible modules, each with a ketosynthase domain as well as a ketoreductase; a ketoreductase and a dehydratase; and a ketoreductase, a dehydratase, an enoyl reductase, and a thioesterase. This triple-KS transcript has a ketosynthase in the third module described as an acyl-transferase containing ketosynthase by the NCBI conserved domain database. Thus, an acyl-transferase may or may not be detected depending on the software used and database queried. A final *A. carterae* transcript (comp14261_c0_seq1) used to make HMMs is herein termed TeCATe due to the flanking thioesterase domains and repeating adenylation and condensation domains as well as a GCN5-associated N-acetyl transferase (GNAT) domain that transfers acetate from acetyl CoA to a substrate,[53] but conservation of this sequence in other dinoflagellate species is low. One additional sequence is

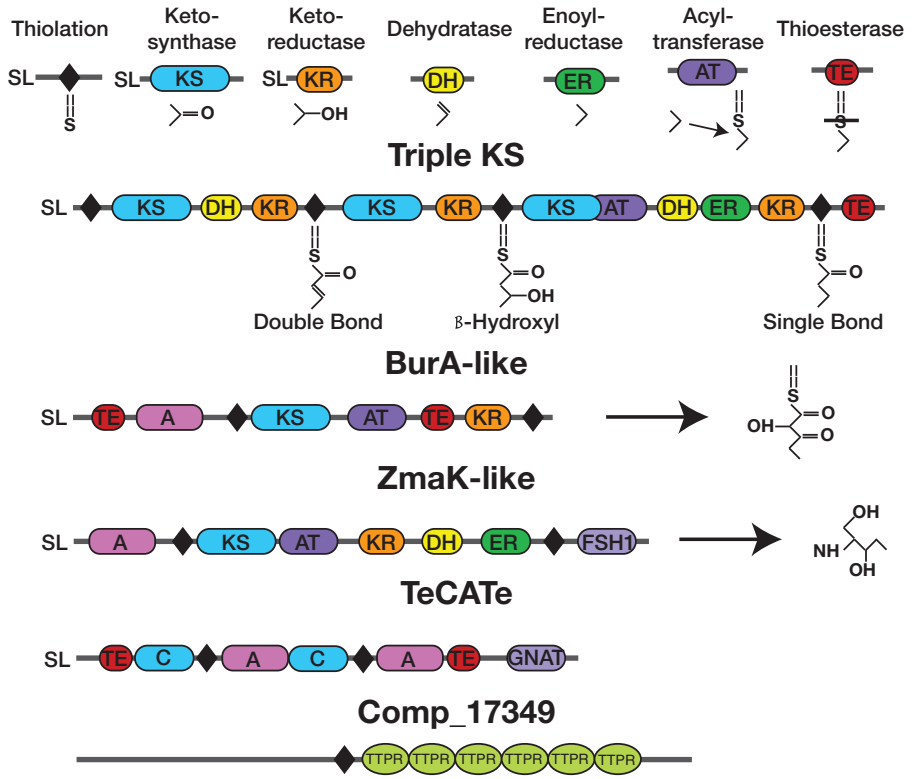

Figure 1. Domain arrangement of *A. carterae* transcripts used in hidden Markov model creation.
Individual modular synthase domains are shown at the top with example products for their reaction. In addition Adenylation (A), FSH1 serine hydrolases (FSH1), GCN5-associated N-acetyl transferase (GNAT), and tetratrico peptide repeats (TTPR) are shown for the multi-domain transcripts with examples of potential products included. "SL" refers to the dinoflagellate spliced leader sequence and is present if a spliced leader sequence has been verified.

comp17349_c0_seq1 that contains a thiolation domain and a tetratricopeptide repeat that is used in protein-protein interactions across all life in a variety of process and configurations.[54] This combination was first described in *K. brevis*[38] and was included to determine the prevalence and association of this repeat domain with other modular synthase domains. It is unclear if any of these transcripts participate in toxin synthesis but they are readily identifiable and the domain arrangement of the triple-KS, BurA and ZmaK like genes in *A. carterae* is conserved in other dinoflagellates indicating that the function is also likely conserved.

A total of 22 domains were used for HMM creation with sequence boundaries based on InterPro[55] annotations as implemented in Macvector V16.0.1. These included the adenylation, the ketosynthase, the ketoreductase, and the acyl-transferase domain as well as thioesterase domain 1 from BurA; the adenylation, dehydratase, enoyl reductase, and serine hydrolase domains from ZmaK; ketoreductase domain 2, thiolation domain 3, ketosynthase domain 3, dehydratase domain 3, enoyl reductase domain 3, and the thioesterase domain from the triple-KS; adenylation domain 1, thiolation domain 1, both condensation domains, and the GNAT domain from TeCATe; and finally the thiolation and tetratricopeptide repeat domains from comp_17349_c0_seq1 (Figure 1). These domains were chosen to provide replicative sampling of each domain across multiple sequences when possible.

The protein translation from the *A. carterae* sequence of each domain was used as the query sequence for a BLAST search across all possible protein translations of the *A. carterae* transcriptome with no cutoff to give as broad a sampling as possible. The aligned region of each BLAST hit was then compiled into a single file for each query domain in fasta format and aligned using Muscle V3.8.31.[56] These alignments were then each used to generate an *A. carterae* specific hidden Markov model (HMM) for each domain using hmmbuild in the HMMER V3.3 package.[57] Each HMM was then compressed with hmmpress and used by hmmsearch with an e-value cutoff of 1e−10 across the protein translations of all 58 transcriptomes with the results given in tab-delimited format for processing. An e-value cutoff was given for the HMM search and not the BLAST search with the assumption that spurious BLAST hits would be represented in the HMM as aligned characters with very low bit scores and that the e-value cutoff in the HMM search would prevent propagation of these errors while maximizing sensitivity. A Perl script was then used to tabulate the data from the HMM search giving a count of each HMM for a given transcript (Supplemental Table S1). The tabulated results were summarized graphically in R V3.3.2 using the GGplot package. For redundant domains the HMM with the highest number of counts for a given transcript was used to maximize sensitivity, for example, if the 3 ketosynthase HMMs returned counts of 1, 2, and 1 the transcript would be

counted as having 2 ketosynthase domains. This differentiates the domains that correspond to a specific HMM, from a domain type that equals the functional classification such as ketosynthase

Domain clustering

Protein sequences for each domain were retrieved from the 6-frame translation of each transcriptome using a Perl script and the output from hmmsearch giving the translation frame and position of the alignments for each HMM. Multiple hits within a transcript were denoted as the number out of the maximum number of that domain in the transcript, for example, Ketosynthase 1_3 for ketosynthase domain 1 out of a total of 3 along with the transcript and host identification. The sequence files were dereplicated prior to clustering to remove redundant protein sequences via a Perl script. The extracted protein sequences were output in fasta format and sequences for adenylation, ketosynthase, thiolation, acyl-transferase, and thioesterase domains were each clustered using CLANS.[58] This software uses an all by all BLAST search and the subsequent e-values are used as attraction values to group sequences. This is not as robust a method as phylogenetic inference for finding closest ancestors but is able to group replicated sequences with their ancestor in a 3-dimensional way and is useful when trying to compare and visualize many very similar sequences. Clusters were visually identified based on a high relative number of internal edges and the sequence names of each node within each cluster were exported to a text file. The sequence list of each cluster was then compared to the master list of domain counts for each transcript to determine the content of each cluster that could be annotated. The vast majority of sequences were single domain transcripts. However, if all of a particular domain from a multi-domain transcript was encompassed by a single cluster then that cluster was labeled for that multi-domain transcript. For example, if the ketosynthase domain from every BurA transcript was found in a single cluster then that cluster was labeled "BurA."

The thiolation domains were a special case in that almost all of the retrieved domains formed just one cluster. In order to provide resolution, the acyl carrier protein sequences (presumably involved in lipid synthesis) from *A. carterae* (comp649_c0_seq2, comp2819_c0_seq1 and comp3690_c0_seq1) were used as BLAST queries against the other transcriptomes and a separate HMM search was performed. The thiolation domains from these sequences were then added back in to the clustering analysis. This was not necessary for the other domains where either the genes involved in lipid synthesis were retrieved in the initial HMM search or there was sufficient resolution of clusters to make the inclusion of fatty acid biosynthesis genes unnecessary. For the smaller datasets of acyl-transferase and thioesterase domains, verification of the clustering results were attempted by maximum likelihood based phylogenetic inference using RAxML[59] using rapid bootstrapping of 100 replicates and seed values of 11111 for both the bootstrapping and parsimony steps.

Results

BUSCO scores

The scores from the BUSCO analysis ranged from 1.6% for *P chesapeaki* to 86.1% for *A. tamarense* (Table 1). There was also frequent duplication with up to 48.8% of the orthologs used for testing present in multiple copies in *Kryptoperidinium foliaceum* (F. Stein). Despite deep sequencing of several of the transcriptomes, the highest BUSCO score would not be considered a complete transcriptome, indicating that many of the "common" eukaryotic orthologs are not present or were not detected. Deep sequencing also did not guarantee a higher than average score with the *G. polynesiensis* transcriptome analysis resulting in a score of 68.3%. Several of the transcriptomes had very low scores such as the *A. andersonii* (33.3%) and *P. minimum* strain 2233 (39.2%) that correlated to a lower number of assembled contigs (1 M and 500k, respectively) compared to those with high scores (1.8 M for *A. tamarense*)

Domain tabulation

The "core" dinoflagellates were shown to have many more synthase modules relative to the syndinales and outgroups. *Lingulodinium polyedra* (F. Stein) possessed the most domains (total HMM hits) and domain types (unique functional group hits, eg, "ketosynthase") with 2407 and 1304, respectively (Table 1). In total there were 55 818 HMM hits with sufficient scores (<1e−10) across all transcriptomes (*including those with low BUSCO scores*) with a median value of 859 per transcriptome, although around 40% of these were the tetratricopeptide repeats predominantly occurring in the core dinoflagellates. When the number of modular synthase hits (excluding tetratricopeptide repeats) was reduced to functional domains by taking the maximum score across all HMMs for the same domain there were 27 424 domains in the core dinoflagellates compared to 1332 for the outgroup species, or an average of 669 and 222 per transcriptome, respectively (Table 2). When the 2 dinoflagellate outgroup species *Hematodinium* and *Oxyrrhis* are removed and the remaining outgroup alveolates are taken separately the average drops even further to 79. The largest difference was in the thioesterase domains that were thirteen times more abundant in the core dinoflagellates while often found in single copy in the outgroup species. Thiolation and ketosynthase domains were also much more abundant in the core dinoflagellates with a more than 6-fold increase indicating that core dinoflagellates possess a higher synthetic capacity than other dinoflagellates and alveolates on average.

The GNAT and condensation domains from the TeCATe transcript were poorly represented in the core dinoflagellates and absent from the outgroup species (Table 2). This is likely due to the low number of BLAST results (5 for GNAT, 6 for

Table 2. Summary of Domain Types.

SPECIES	ADENYLATION	KETOSYNTHASE	KETOREDUCTASE	DEHYDRATASE	ENOYL REDUCTASE
Akashiwo_sanguineum	238	168	123	10	194
Alexandrium_catenella	97	172	74	0	57
Alexandrium_margalefi	129	238	117	10	84
Alexandrium_monilatum	161	163	113	9	62
Alexandrium_tamarense	241	188	165	6	117
Amphidinium_carterae	109	76	61	27	46
Amphidinium_klebsii	115	63	74	27	38
Amphidinium_massartii	85	68	67	9	32
Azadinium_spinosum	163	342	148	25	93
Brandtodinium_nutriculum	128	112	68	7	53
Ceratium_fusus	277	74	123	9	90
Crypthecodinium_cohnii	226	72	113	27	179
Dinophysis_acuminata	119	248	96	3	70
Durinskia_baltica	150	74	70	6	45
Gambierdiscus_excentricus	101	35	92	4	71
Gambierdiscus_polynesiensis	125	177	106	12	90
Glenodinium_foliaceum	237	92	89	4	71
Gymnodinium_catenatum	138	36	51	3	46
Gyrodinium_instriatum	559	83	157	29	84
Heterocapsa_arctica	71	129	66	11	25
Heterocapsa_rotundata	66	120	50	6	28
Karenia_brevis	234	166	163	9	110
Karlodinium_veneficum	260	95	117	5	82
Kryptoperidinium_foliaceum	370	129	120	5	98
Lingulodinium_polyedrum	217	371	239	3	97
Noctiluca_scintilans	85	67	75	6	50
Pelagodinium_beii	145	102	75	6	60
Peridinium_aciculiferum	144	79	76	4	56
Prorocentrum_hoffmanianum	143	169	92	12	63
Prorocentrum_micans	252	64	158	13	91
Prorocentrum_minimum_1329	127	42	95	3	49
Protoceratium_reticulatum	147	182	68	3	61
Pyrodinium_bahamense	133	145	95	6	62
Scrippsiella_hangoei	202	122	128	14	104
Scrippsiella_hangoei_like	181	92	95	4	63
Scrippsiella_trochoidea_CCMP3099	313	141	152	14	105
Symbiodinium_sp_B1	96	53	88	7	46
Symbiodinium_sp_C1	125	70	91	7	58
Symbiodinium_sp_CCMP421	187	147	101	13	81
Symbiodinium_sp_cladeA	78	74	54	24	33
Symbiodinium_sp_Mp	120	66	76	2	42
SUM (core dinoflagellates)	7094	5106	4181	404	2986
Chromera_velia	24	1	20	0	22
Hematodinium_sp	119	61	91	18	106
Oxyrrhis_marina	129	21	71	0	60
Oxyrrhis_marina_LB1974	105	25	48	0	43
Perkinsus_marinus	16	0	9	13	15
Triceratium_dubium	46	9	28	0	25
SUM (outgroups)	439	117	267	31	271
TOTAL	7533	5223	4448	435	3257

Abbreviations: FSH1, fission yeast serine hydrolase 1; GNAT, GCN5-associated N-acetyl transferase; TTPR, tetratricopeptide repeat.
Values shown represent the count of each domain type in each transcriptome or the sum when designated.

THIOESTERASE	THIOLATION	ACYL_ TRANSFERASE	FSH1	CONDENSATION	GNAT	TTPR	SUM	SUM (NO TTPR)
11	74	58	41	5	0	561	1483	922
13	77	104	36	0	0	767	1397	630
22	86	129	46	1	0	715	1577	862
17	87	79	41	1	0	723	1456	733
23	95	145	69	1	0	785	1835	1050
17	61	17	15	10	5	304	748	444
25	60	17	14	8	3	320	764	444
8	42	16	15	6	2	285	635	350
29	144	166	84	0	0	1217	2411	1194
5	51	30	9	1	0	299	763	464
49	33	48	19	2	0	148	872	724
6	52	77	19	3	0	278	1052	774
14	127	108	55	1	0	1013	1854	841
4	39	29	8	0	0	294	719	425
14	64	63	42	0	0	657	1143	486
20	106	169	34	0	0	655	1494	839
6	49	46	11	2	0	271	878	607
7	36	21	14	0	0	39	391	352
12	200	81	57	1	0	359	1622	1263
6	58	54	9	2	0	288	719	431
9	45	36	7	4	0	269	640	371
9	168	75	71	6	2	601	1614	1013
4	89	41	23	4	0	436	1156	720
14	90	68	21	3	0	489	1407	918
25	137	210	82	6	0	1192	2579	1387
4	26	23	9	0	0	192	537	345
5	69	32	15	0	0	321	830	509
7	59	34	22	1	0	375	857	482
8	55	51	31	1	0	705	1330	625
36	94	33	26	0	0	181	948	767
16	31	21	18	2	0	167	571	404
11	78	66	26	2	0	650	1294	644
10	67	104	22	0	0	538	1182	644
15	120	41	38	1	0	461	1246	785
7	83	37	25	1	0	422	1010	588
19	154	63	32	7	0	487	1487	1000
7	95	17	19	1	0	349	778	429
7	65	25	18	1	0	362	829	467
13	117	49	23	0	0	586	1317	731
6	33	21	8	7	1	124	463	339
9	62	23	21	0	0	400	821	421
549	3278	2527	1195	91	13	19285	46709	27424
1	1	1	2	0	0	13	85	72
1	52	25	27	3	0	8	511	503
3	11	21	9	0	0	56	381	325
1	14	24	6	0	0	58	324	266
0	0	0	2	0	0	0	55	55
0	0	1	1	1	0	165	276	111
6	78	72	47	4	0	300	1632	1332
555	3356	2599	1242	95	13	19585	48341	28756

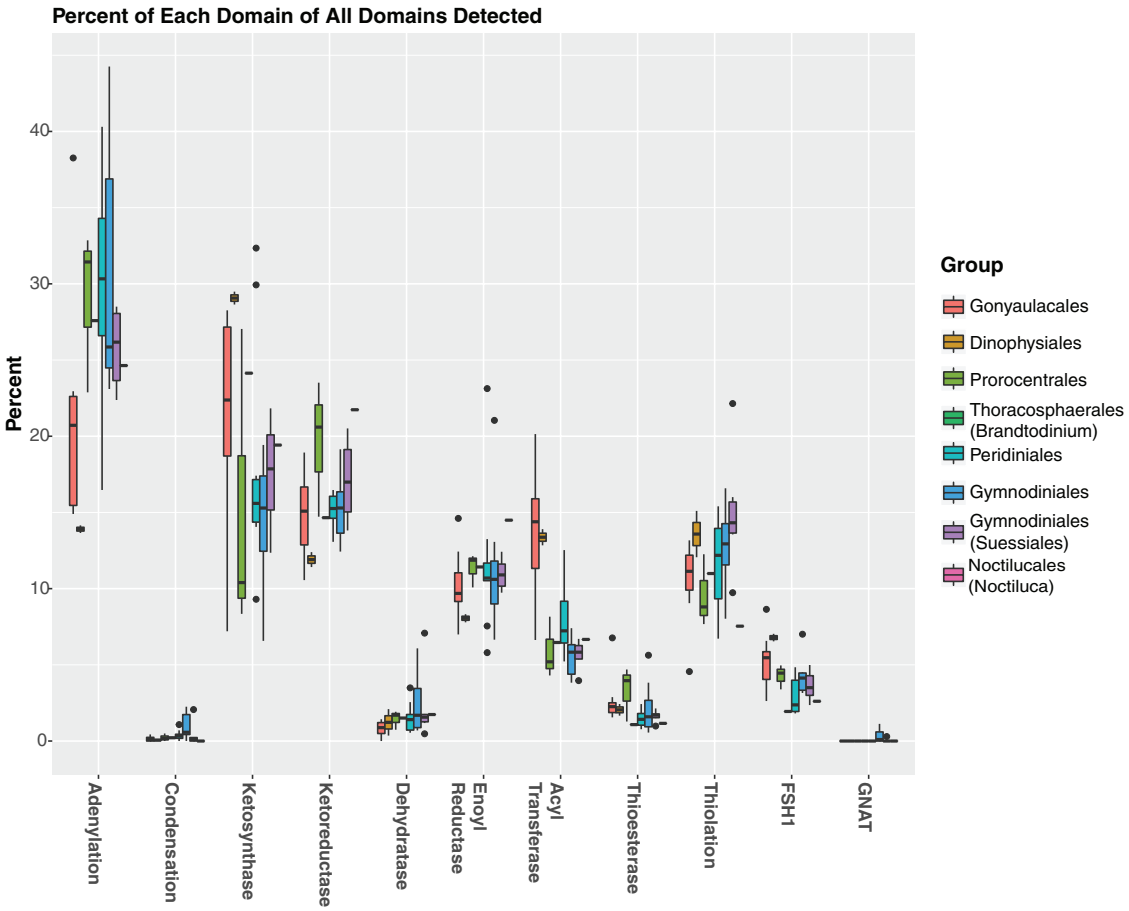

Figure 2. The percent of each domain relative to all domains detected in dinoflagellates using the dinoflagellate HMMs.
The relative dinoflagellate domain abundance for each domain is shown with the percent shown on the *Y*-axis and boxplots of the values when more than one species was present in each group with black circles denoting outlier values. Dinoflagellates were grouped taxonomically by their order and colored according to the legend on the right. The domains are shown on the *X*-axis excluding the tetratricopeptide repeat domains that were used in the calculation but were frequently not associated with any of the modular synthase domains.

condensation domain 1, and 9 for condensation domain 2) from the primary search in *A. carterae* that limited the creation of a robust HMM for GNAT and condensation domains. By contrast the adenylation domain from the TeCATe transcript resulted in 41 BLAST hits. The HMM search was still more sensitive than BLAST alone with a total of 13 transcripts detected with a GNAT domain using the HMM with an e-value cutoff of 1e−10 versus 8 using BLAST with no cutoff among all transcriptomes (data not shown). Still, these domains may have been under-sampled especially for taxa more distantly related to *A. carterae*. The GNAT and condensation domains were the least represented among all other domains with adenylation and ketosynthase having the highest relative abundance across taxonomic groups (Figure 2). Thiolation, acyl-transferase, enoyl reductase, and ketoreductase domains were also usually well represented while dehydratase, thioesterase, and the FSH1 serine hydrolase were in relatively low abundance.

This picture changes when looking at multi-domain transcripts (transcripts with more than one domain type, not including multiple domain hits of the same domain type), where roughly a third of dehydratase domains and half of thiolation domains are found in multi-domain transcripts while adenylation, ketosynthase, ketoreductase, and enoyl reductase domains are predominantly found as single domains (Figure 3). These trends frequently held across taxonomic groupings except for thioesterases, which were found 10% to 15% of the time in multi-domain transcripts for the Gonyaulacales and Dinophysiales and a quarter to a third of the time as multi-domain transcripts in the other taxonomic groups. Multi-domain transcripts were the exception in core dinoflagellates accounting for 8.34% of all domain types (excluding tetratricopeptide repeats) with an average of 13.84% for each species and domain type combination.

The relative abundance of modular synthase domains was similar across species with no obvious differences in species with a known toxin. The principal components plot based on domain counts and colored by toxin type was used to demonstrate this (Figure 4(A)) and showed a general clustering of all species, irrespective of toxin type except for 3 species: *Gyrodinium instriatum*, that does not make a known toxin and has a higher proportional number of adenylation domains (Axis 1 outlier on the far left), and *Lingulodinium polyedra* and *Azadinium spinosum* that make yessotoxin and azaspiracids,

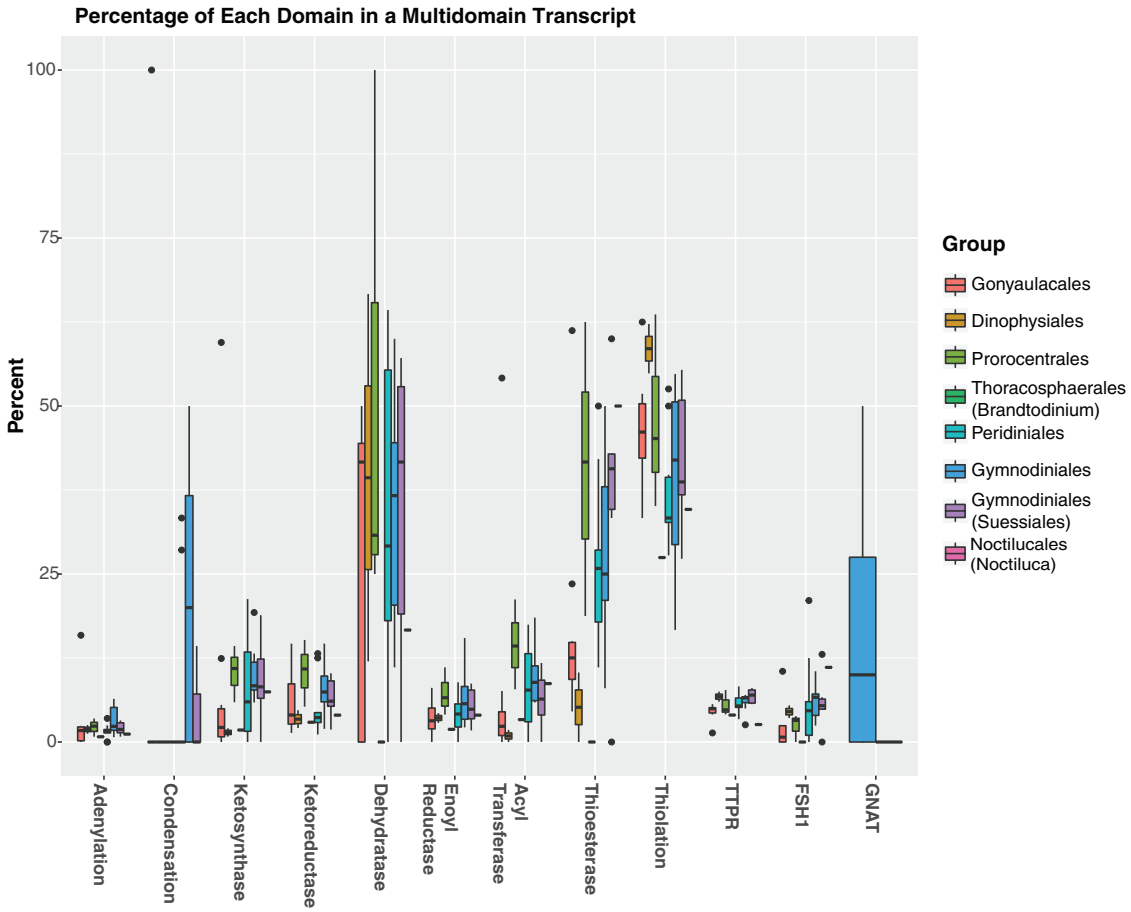

Figure 3. The percent of each domain in a multidomain transcript relative to the total number of each domain found in dinoflagellates.
The relative abundance of each domain type out of the total number of that domain (excluding tandemly repeated domains without other domain types) is shown as a percent on the *Y*-axis and a box plot when multiple species are present with black circles denoting outlier values. Dinoflagellates were grouped taxonomically by their order and colored according to the legend on the right. The GNAT domain was only present in 2 taxonomic orders, the Gymnodiniales and the Suessiales, and the boxplots are thus drawn with a different width.

respectively, and have a proportionally higher number of ketosynthase domains (Axis 2 outliers on the bottom). There were however, lineage specific differences for specific domain types. Thiolation domains were more relatively abundant in the more basal Gymnodiniales compared to acyl transferase in a decreasing trend to the more distal Gonyaulacales (Figure 5). This is also visible in the plot of domains as percentages. The acyltransferase domains make up a much higher percentage in the Gonyaulacales and Dinophysiales versus other taxonomic groups (Figure 2), although this is less obvious for the thiolation domains. There is also a high average number of thiolation domains in a transcript when comparing the Gymnodiniales to the other taxonomic groups (Figure 6).

Although tetratricopeptide repeats were found in almost every transcriptome, the abundance was much higher in the core dinoflagellates (19 285 in core dinoflagellates vs 300 in outgroups or a 4-fold increase on average per transcriptome) and the combination of this repeat with thiolation domains was only found in the core dinoflagellates (Figure 7). The number of repeats varied within a transcript from 1 to 20 and the distribution of repeat number is approximately log normal in shape with low numbers of repeats being very frequent (Figure 8(A)). This distribution changes dramatically when looking at repeat numbers in transcripts with a thiolation domain where 6 and 7 member repeats are very frequent approximating a t-distribution (Figure 8(B)). The relative number of repeats among each taxonomic group did not vary greatly and approximated the relative total number of domains found.

Domain clustering and gene duplication

The protein sequence clustering results in a 3-dimensional relationship of the domains where more similar sequences are spaced more closely together and are shown as points in the data space. If the points are close enough to pass a threshold of a calculated probability then a line is drawn denoting significant similarity and a group of points with interconnecting lines was denoted as a cluster. While these data do not tell us the inferred ancestry of the sequences like a phylogenetic tree would, the relationships are not forced into a bifurcating arrangement. This is helpful in visualizing many very similar sequences such as dinoflagellate domains where there is an abundance of gene duplication and strict orthology is difficult to ascertain. The results for each domain retrieved by HMM

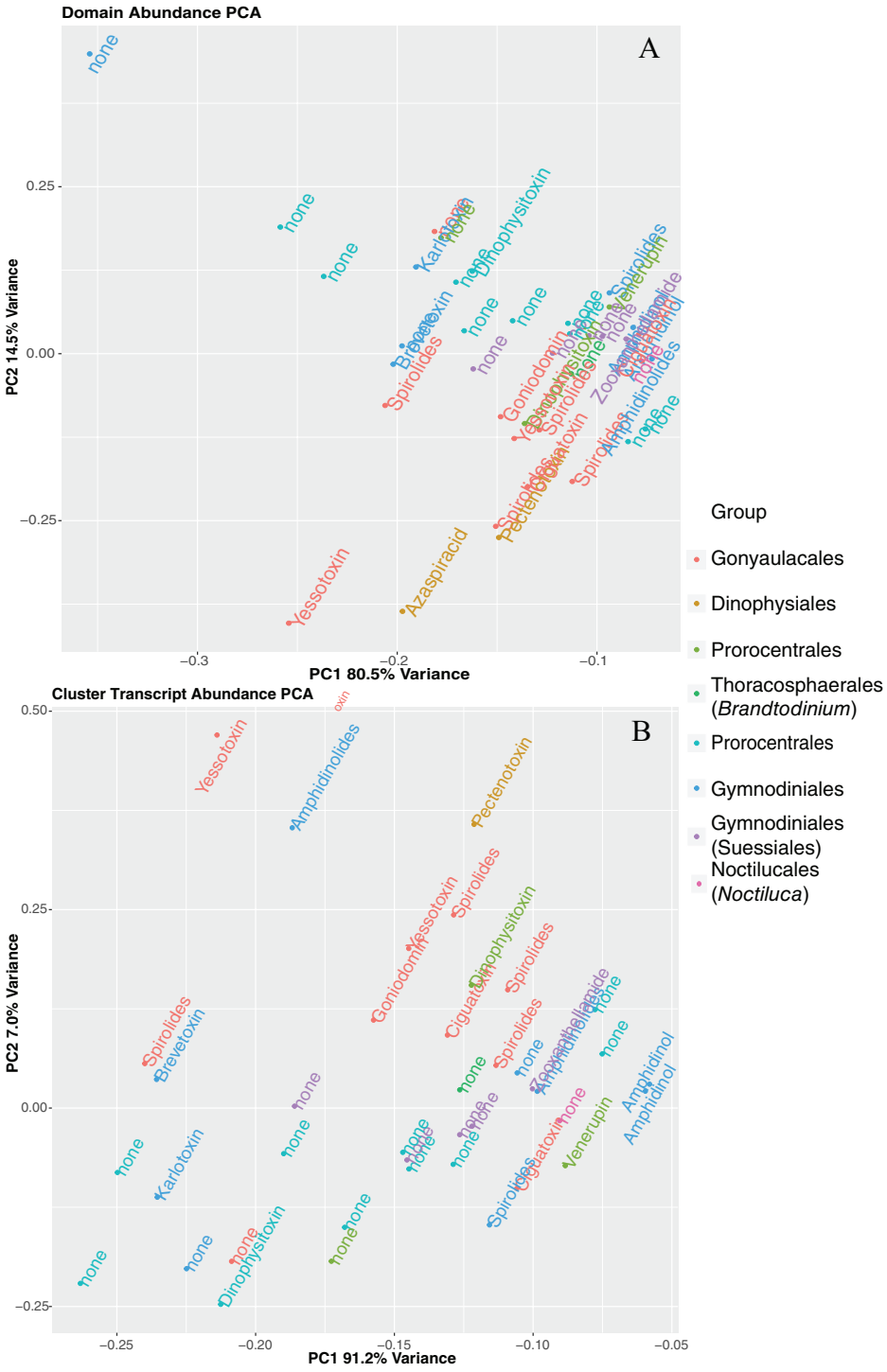

Figure 4. Principal component plots of the overall domain counts within dinoflagellates (A) and the breakdown of those counts within clusters of highly similar sequences (B).

Principal components are shown for the total domain counts among taxonomic groups of dinoflagellates (A) as well as the count of individual transcript within clusters of very similar domain sequences (B). Principal component 1 is shown on the *X*-axis and component 2 is shown on the *Y*-axis with the relative contribution of each component shown next to each axis. The individual points represent an individual transcriptome that is colored for the order level taxonomy of each species from which the transcriptome was samples shown on the legend on the right. Each point is also labeled with a widely recognized toxin that is made by that species.

searches were compared based on the number of clusters, where a large number of clusters denotes a high relative degree of inferred functional diversity, and the size of the clusters that is an indication of the amount of gene duplication. Clusters were also searched for the annotated multi-domain transcripts to allow for comparisons of clusters across and within each domain. The number of sequences used for clustering varied substantially between domains with 15 865 adenylation; 10 118 ketoreductase; 9832 ketosynthase; 7854 thiolation; 7025 enoyl-reductase; 3324 dehydratase; 2492 acyl transferase; and 1085 thioesterase domains, following dereplication of sequences. There are also likely some false positives from the HMM

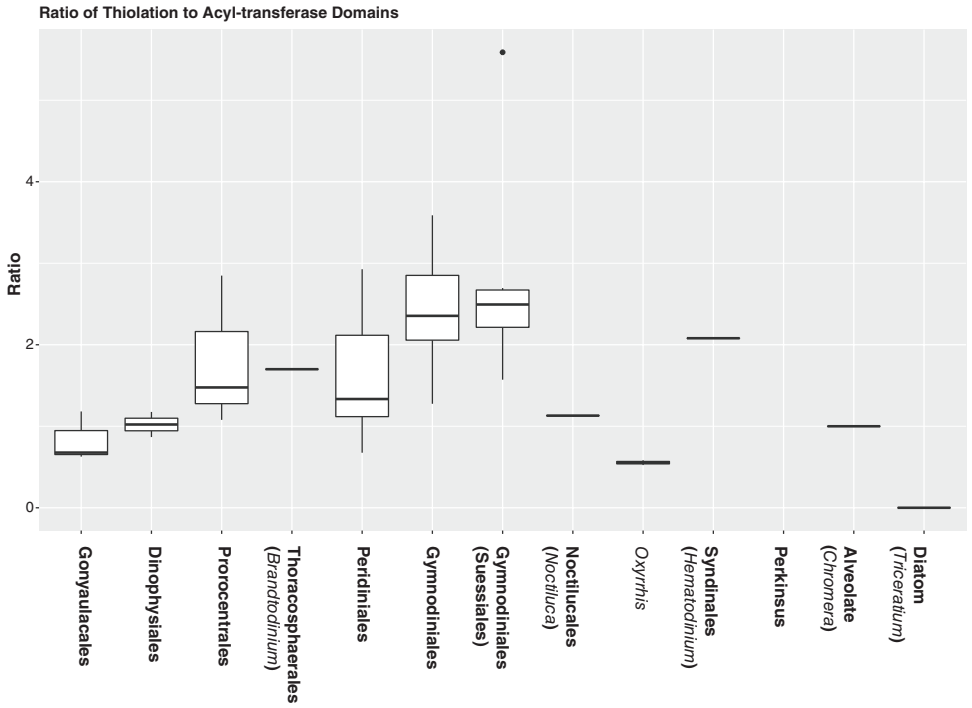

Figure 5. The ratio of thiolation to acyl-transferase domains in dinoflagellate and outgroup taxa.
The relative abundance of thiolation and acyl-transferase domains are shown as a boxplot with the ration of thiolation to acyl-transferases on the *Y*-axis. The *X*-axis shows the taxonomic grouping with dinoflagellates grouped by Order and single species given as their genus in parentheses with the exception of *Oxyrrhis*. The Syndiniales and other outgroup taxa are also shown represented by a single individual.

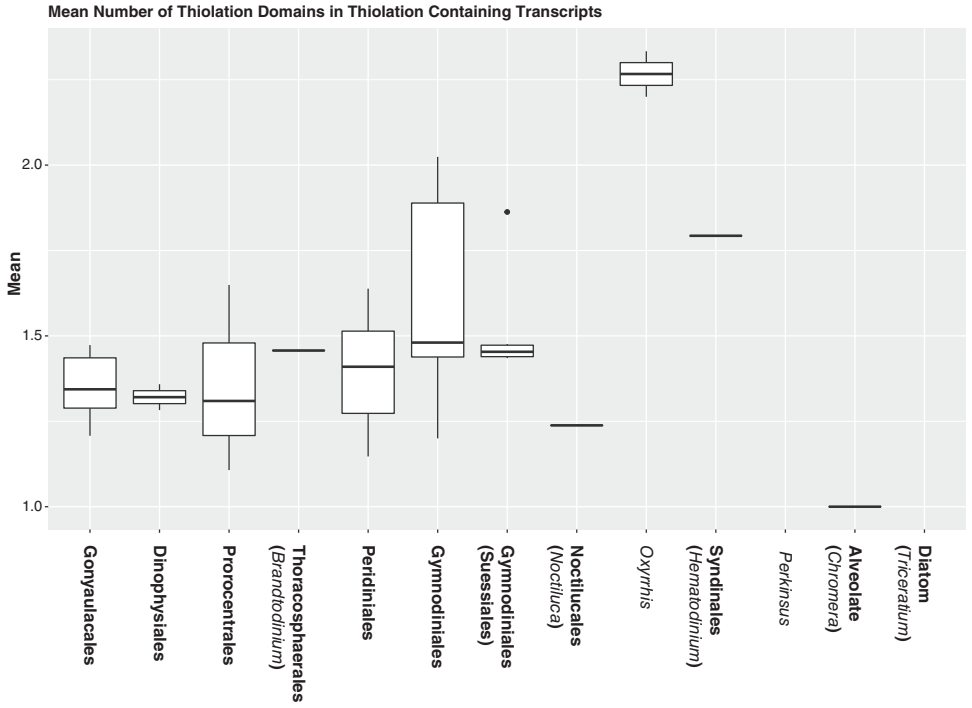

Figure 6. The mean number of thiolation domains in all thiolation domain containing transcripts.
The average number of thiolation domains per transcript in all thiolation domain containing transcripts is shown on the *Y*-axis while the *X*-axis shows the taxonomic grouping with dinoflagellates grouped by Order and single species given as their genus in parentheses with the exception of *Oxyrrhis*. Transcripts did not have to have any other domain type in order to be counted and many transcripts contained thiolation domains exclusively.

search with 7887 of the 202 024 total sequences containing internal stop codons that may be from the translation of a spurious open reading frame that was coincidently similar to the HMM. These false positives as well as some truncated sequences appear in the clusterings as outlying spots. Also, sequence depth may artificially inflate or deflate the size of each cluster. The BUSCO scores for each transcriptome used in clustering were similar so this is not likely to be a dramatic

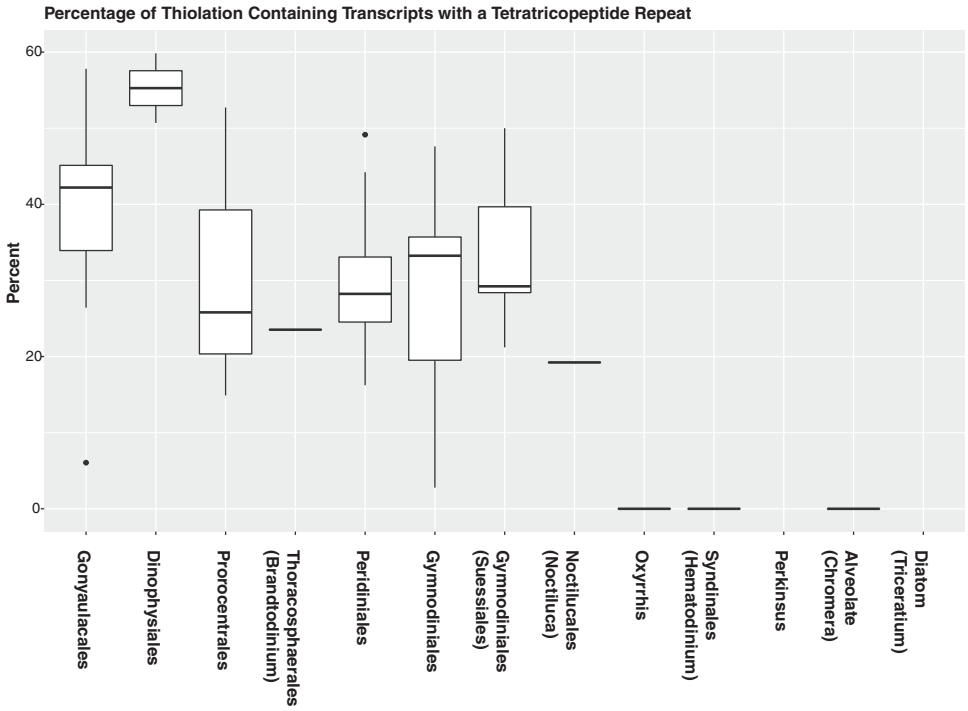

Figure 7. Percentage of transcripts containing a thiolation domain and tetratricopeptide repeats in dinoflagellates and outgroup species.
The percentage of thiolation domain containing transcripts that also contain a tetratricopeptide repeat are shown with the percentage on the *Y*-axis as boxplots when more than one species is present. The *X*-axis shows the order level taxonomy of the dinoflagellate species with the exception of *Oxyrrhis* and *Perkinsus* where the phylogenetic placement is less certain. The remaining outgroup species are described as their common phylum name followed by the genus level taxonomy in parentheses.

effect. Likewise, this is unlikely to affect the number of clusters, only the size, since only a relatively large number of similar sequences would generate a cluster.

The clustering of thiolation domains represent an example of low diversity and low copy number with only 2 clusters formed when the acyl carrier protein was added into the dataset and a small trail off of the largest cluster containing one of the ZmaK thiolation domains (Figure 9). The main cluster has several subclusters, one containing BurA transcripts, a second containing the triple KS transcripts, a third that has several transcripts with adenylation and ketosynthase domains that are not 1 of the 4 annotated transcripts from Figure 1. There is some resolution of the subclusters separating the annotated transcripts from each other but they are very tightly linked with many internal edges. The other main cluster exclusively contains acyl carrier protein sequences from each of the transcriptomes. These differences can be seen when viewing an alignment of the binding sites from *A. carterae* for the phosphopantetheinyl transferase that activates the thiolation domain (Figure 9 insert). Most of the domains have similar positively charged residues following and negatively charged residues preceding the invariant serine that serves as the site of phosphopantetheinate attachment. For one of the ZmaK sites the negatively charged residue is instead positively charged and for the acyl carrier protein there is a methionine. The acyl carrier protein from *Escherichia coli* also has a methionine showing how conserved this residue is in the acyl carrier protein making this gene easy to distinguish from other modular synthases. Thus, there is a clear segregation of fat synthesis from other small molecule synthesis in the thiolation domain clusters that is irrespective of any implied gene origin such as horizontal gene transfer from bacteria in the case of BurA and ZmaK. This can also be used to validate the data to some degree as the ACP cluster contains approximately 230 sequences averaging to 4 per transcriptome. *A. carterae* has 3 readily identifiable acyl carrier proteins indicating that a reasonably expected number of sequences are present in the ACP cluster.

Other domains with limited clusters include the dehydratase, enoyl reductase, and thioesterase domains (Figure 10). The dehydratase domains (Figure 10(A)) form a single cluster of sequences including those from the triple-KS and ZmaK transcripts as well as an ancillary cluster that is annotated as a "Domain of Unknown Function" by the NCBI COG database and is similar to dehydratases involved in tyrosine metabolism by BLAST. The low copy number of the dehydratase domain is in contrast to the enoyl reductase clusters that are numerous (Figure 10(B)). Diversity is still low with a single cluster containing the ZmaK transcript and 2 other clusters that do not contain annotated transcripts but there are also many sequence fragments that form satellite points and do not cluster. The thioesterase domain clustering (Figure 10(C)) contains 3 low abundance clusters all containing annotated transcripts.

For the ketosynthase and adenylation domains the domain count was very high with several unknown clusters (Figure 11), similar to the enoyl reductase clustering but with a larger number of clusters. For the adenylation domain clustering (Figure

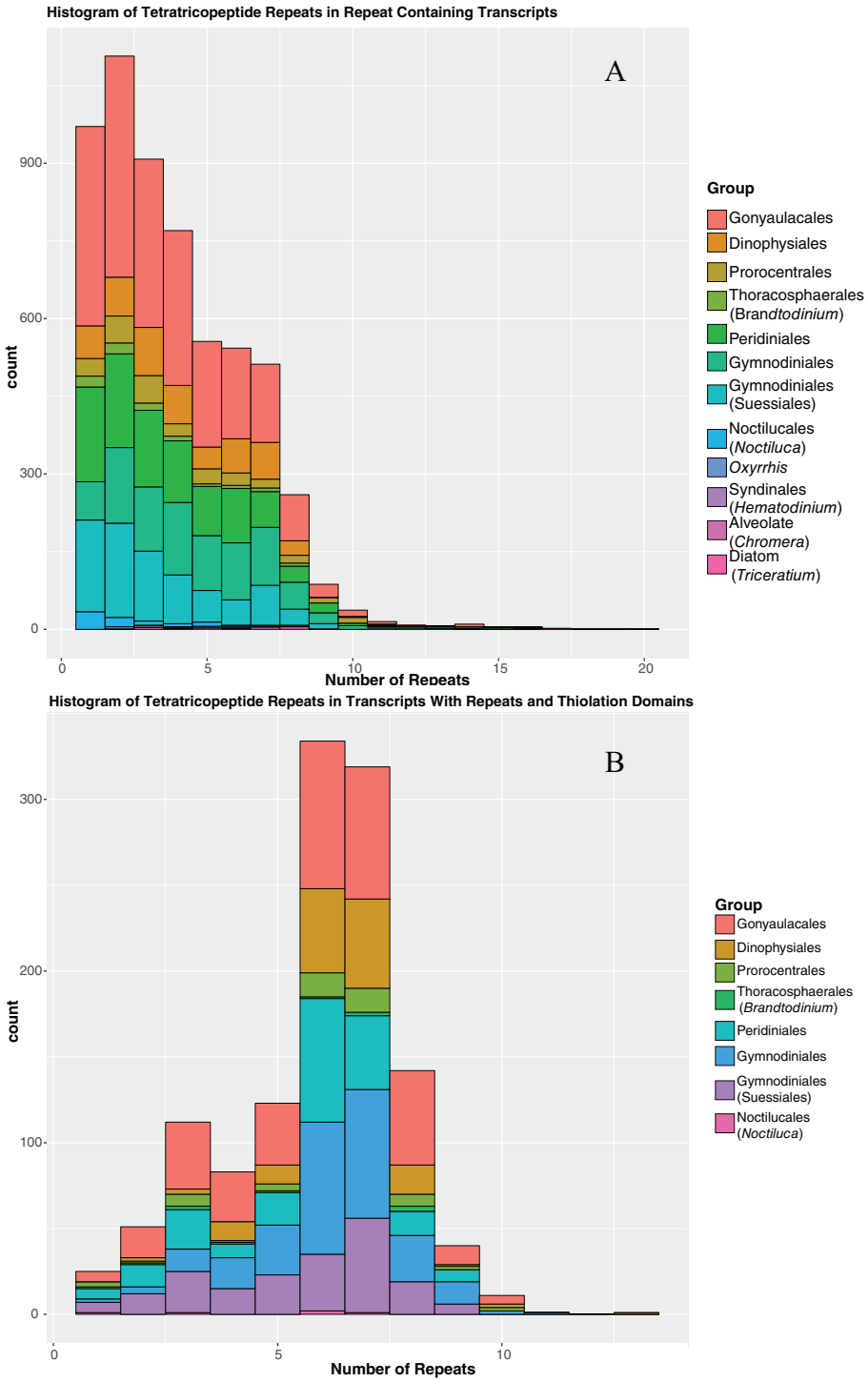

Figure 8. Histogram of the number of tetratricopeptide repeats in repeat containing transcripts.
The count of the tetratricopeptide containing transcripts is shown on the *Y*-axis while the number of repeats in those transcripts is shown on the *X*-axis. The upper panel (A) shows the histogram of tetratricopeptide repeats in all repeat containing transcripts while the lower panel (B) shows the histogram for transcripts that also contains a thiolation domain. This combination of tetratricopeptide repeats was only observed in core dinoflagellates and thus the legends for the 2 panels are not identical. The upper legend for panel A includes dinoflagellates grouped by order with individual specimens given as their genus in parentheses along with outgroup species while the lower legend for panel B only includes core dinoflagellates.

11(A)) the number of sequences was the largest with over 15 000 unique sequences. The adenylation domains from the ZmaK and BurA transcripts were found in separate clusters but for the ketosynthase domain (Figure 11(B)) the annotated transcripts all occupy a single cluster with poor resolution of subclusters and include the condensation domains from the TeCATe transcript. Both domains produced several clusters that do not contain annotated transcripts and the ketosynthase domains involved in fat synthesis labeled "FabB" form a distinct low abundance cluster.

This pattern of large clusters of single domain transcripts that are similar to domains from the annotated transcripts and very small conserved clusters of fat synthesis genes appears to reverse for the acyl transferases and ketoreductases (Figure 12).

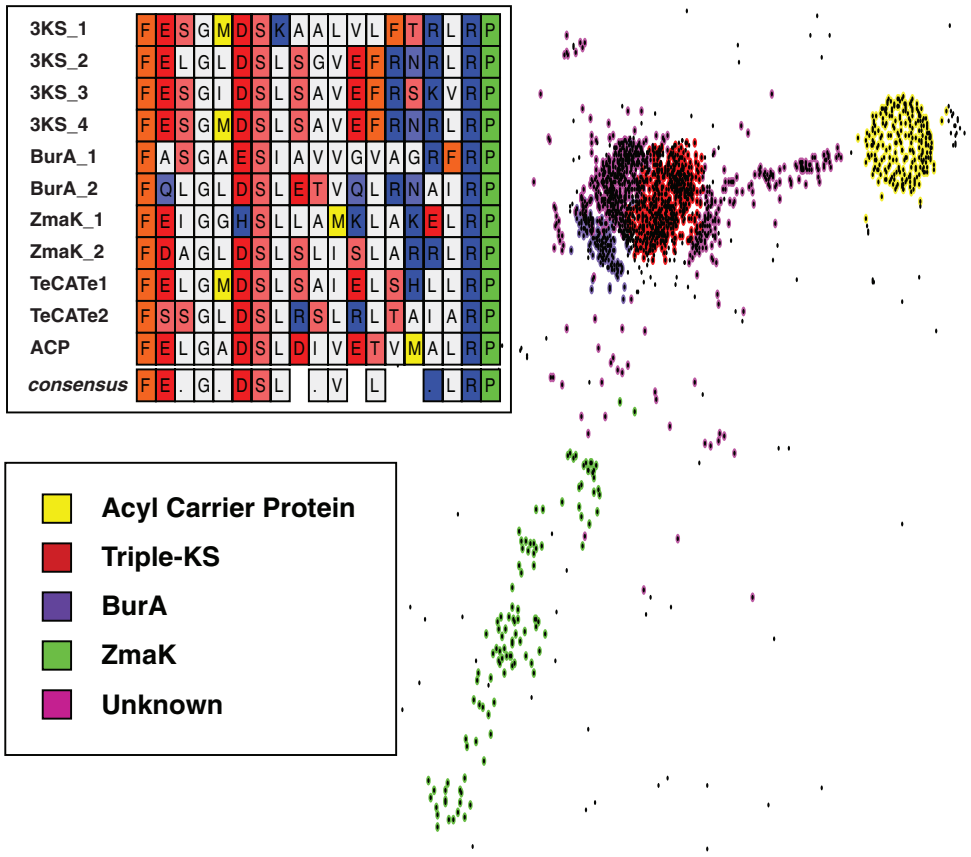

Figure 9. Cluster plot of the thiolation domains.
A clustering of the protein sequences for dinoflagellate thiolation domains is shown. The acyl carrier protein was added back into the analysis since this was not recovered in the hidden Markov model search and is colored yellow. Thiolation domain clusters containing the Triple-KS and ZmaK_2 (red), BurA (purple), and other unidentified thiolation domains form the central cluster while a cluster containing the ZmaK_1 thiolation domain (green) is shown extending down and to the left of the central cluster. An alignment of the reference thiolation domains from *Amphidinium carterae* is shown in the upper left.

Despite acyl transferases being one of the lowest abundance domains and ketoreductases one of the highest, both clusterings contain large clusters of sequences similar to the fat synthesis genes FabD (acyl transferase) and FabG (ketoreductase). There are small clusters of acyl transferase domains from the triple KS and BurA transcripts with very few single domain transcripts. The cluster containing FabD like transcripts is quite large with many of these transcripts containing ankyrin repeats that promote and regulate protein-protein interactions.[60] There is also a small cluster of single domain transcripts that are not similar to the annotated transcripts. Likewise the large ketoreductase clusters (Figure 12(B)) are comprised of the annotated transcripts contained in a single cluster, 3 clusters that do not contain annotated transcripts and a final very large cluster containing the FabG gene.

Principal components plots of the domain counts for each cluster (Figure 3(B)) gave similar results to the overall domain counts with no association between toxins produced or phylogenetic group to principal component positions. Principal component axis 1, which accounted for 91.2% of the variance differed mainly in the expansion of BurA-like domains with species on the left portion of the graph possessing large numbers of BurA-like domains while those on the right had very few. This did not correlate to intact BurA transcripts and the level of expansion was not always consistent, *e.g. Karenia brevis* was found to have 512 BurA-like adenylation domains but only 13 BurA-like thioesterase domains (Supplemental Table S2). In both cases this count was much higher than other species, but not equivalent.

Phylogenetic inference was attempted on the 2 smallest datasets, acyl transferases and thioesterases, to determine ancestry and compare the results to the clustering results. The resultant trees (Supplemental files 1 and 2) had zero or near zero bootstrap support for all major and minor branches up to the final bifurcations indicating that determination of ancestry was not possible for these datasets and methods. The highest scoring trees were able to replicate the clustering results to some degree with major clades mirroring the clusters formed indicating that there was some phylogenetic signal present. Also, very similar sequences or assembly variants were visible with high bootstrap support at the distal branches.

Discussion

Modular synthases are abundant in the core dinoflagellates

The goal of this study was to investigate the abundance, diversity, domain arrangement, and evolution of the suite of enzymes

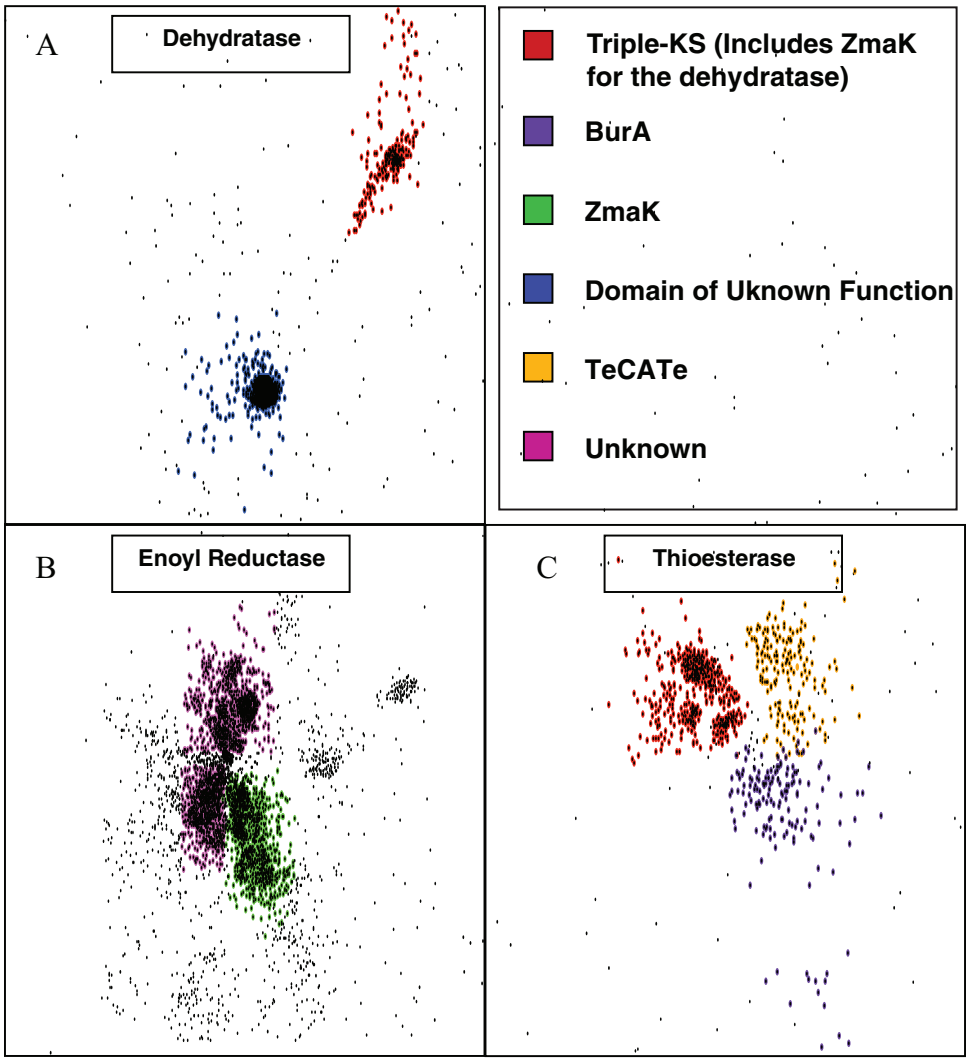

Figure 10. Protein sequence clusters of dehydratase, enoyl reductase, and thioesterase domains.
The clusterings of the dehydratase (A), enoyl reductase (B), and Thioesterase (C) protein sequences are shown. All domains exhibit relatively low levels of duplication. The dehydratase domain has 2 clusters including 1 of "unknown function" that is similar to amino acid dehydratases according to the NCBI ortholog database. The enoyl reductase domain has 3 clusters including ZmaK and 2 of unknown similarity while domains from the triple KS transcript are found in several different clusters depending on the species. The simplest domain is thioesterase with clusters containing the BurA, triple KS, and TeCATe transcripts from *Amphidinium carterae*.

likely to participate in the synthesis of dinoflagellate toxins by focusing on dinoflagellates and the synthetic domains found consistently within dinoflagellates independent of existing model frameworks. There is an inherent need to study these synthetic pathways since dinoflagellate toxins are the largest known natural products with high potential toxicity,[11,61] are synthesized using many non-canonical and interesting chemistries,[25,28] and have potential therapeutic uses;[61,62] but we have very little understanding of how they are synthesized. Based on isotopic labeling studies it is clear that they are predominantly made of acetate units incorporated by polyketide synthases[23,25] with the occasional amino acid or other carboxylic acid used as starter and extender units via non-ribosomal peptide synthases.[27] In both cases a condensation reaction incorporates a chemical unit into a growing molecule that is subsequently modified either during elongation or following the synthesis of a large portion of the molecule.[16] In order to facilitate discussion we have combined polyketide synthases and non-ribosomal peptide synthases into the term "modular synthases" to encompass both condensation reactions and the general similarities of their chemistry and genetics. For polyketide synthases, the condensation reaction is performed by a ketosynthase that usually incorporates acetate from malonyl CoA but can also facilitate the addition of other short carbohydrates[26] with the release of carbon dioxide. Non-ribosomal peptide synthases use adenylation domains to pass a specific substrate, often an amino acid, to the condensation domain with microcystin being a common example.[63] Adenylation domains can also be found in the same module as ketosynthases in a hybrid system as is the case for BurA and ZmaK that participate in the synthesis of burkholderic acid in *Burkholderia* species and zwittermicin in *Bacillus* species, respectively.[45,46] The BurA and ZmaK synthetic pathways are also important to mention because the module has been fragmented in their respective bacterial genomes with separate modules occurring on distant regions of the chromosome while BurA additionally serves an unusual role in bridging pathways. Due to the processive nature of these modular synthases the pathway is generally encoded as

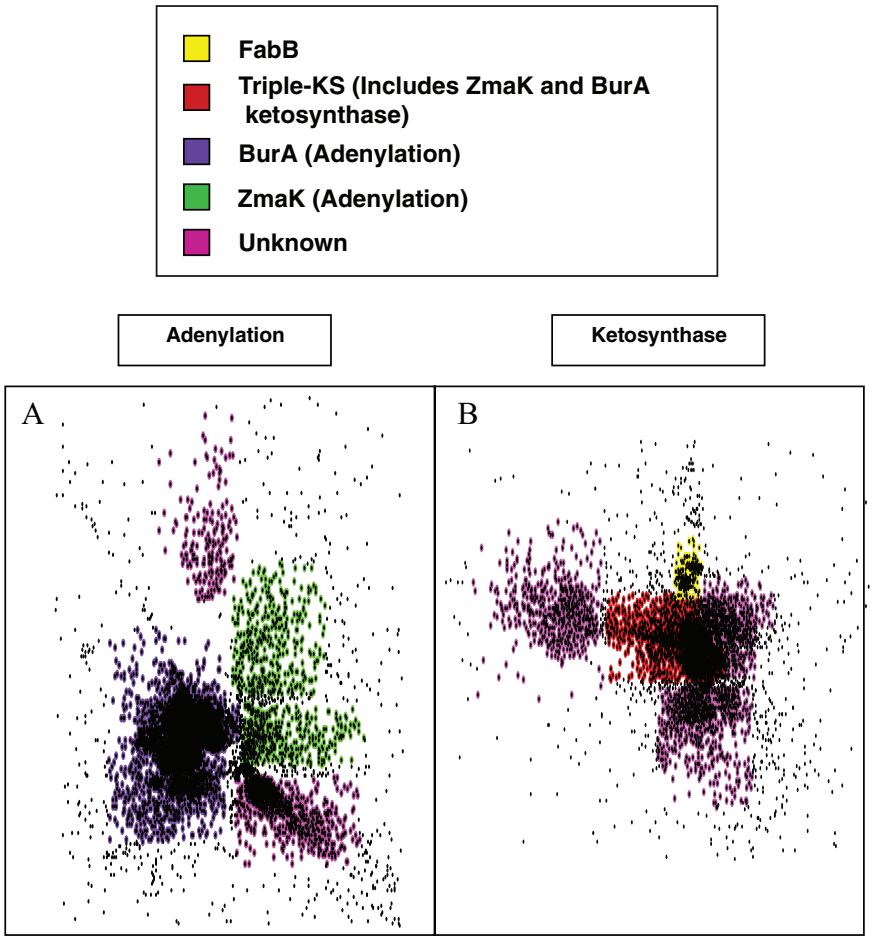

Figure 11. Adenylation and Ketosynthase protein sequence clusters.
The clusterings of the adenylation (A) and ketosynthase (B) protein sequences are shown. Both exhibit gene duplication of domains from the triple KS, BurA, and ZmaK transcripts forming 2 large clusters of adenylation domains and 1 very dense cluster of ketosynthase domains. Several clusters of unknown similarity are also apparent as well as the FabB ketosynthase cluster that is involved in lipid synthesis and does not appear to be heavily duplicated.

syntenic modules with domains in a more or less linear fashion that can be used to predict the final product of synthesis.[64] Although many domains can come into play in a trans fashion,[39] the most common trans-acting domains are acyl transferases and thioesterases. These domains also do not need to be collinear with other synthetic modules and have been shown to be synthetically active when whole genomic sections have been cloned.[65] A cursory BLAST analysis of the published *Polarella glacialis*[66] genome shows that domains are commonly found in tandem repeats of the same domain with different domains found on different scaffolds with the exception of common multi-domain transcripts such as the triple KS (Supplemental Table S4).

Dinoflagellates regulate their gene expression largely post-transcriptionally[67,68] making linear encoding of the modular synthase domains obsolete. Unsurprisingly, the vast majority of modular synthase domains have been fragmented and duplicated, similar to what has been shown for other gene families in dinoflagellates such as actin and translation initiation factors.[40,69] In this study adenylation, ketosynthase, and ketoreductase domains were frequently observed as single domain transcripts, as has been observed previously.[7,38,43] Both single and multi-domain transcripts of modular synthases occurred in high abundance in all core dinoflagellates and their distribution was not correlated with taxonomy, toxicity, or toxin type (Figure 4). This apparently ubiquitous synthetic capacity argues that secondary metabolite synthesis is a common feature of all core dinoflagellates, a theory supported by observations that polyketide synthesis genes are found in species that do not produce known polyketide toxins.[70] Similarly, the only phosphopantetheinyl transferase, the enzyme required to activate thiolation domains and initiate secondary metabolite synthesis, found in all core dinoflagellates was able to activate a NRPS based reporter system indicative of natural product rather than lipid synthesis.[71] This is in contrast to syndinean dinoflagellates and other alveolates that had a much lower abundance of synthetic domains with all domains in similar abundance (Table 2). This is likely due to serial duplication that is a hallmark of core dinoflagellate evolution[72] and has been shown to affect the evolution of the synthetic pathway for saxitoxin in particular.[73]

Single domain transcripts exhibit domain specific patterns of duplication

Multidomain transcripts were observed in all core dinoflagellates. The triple KS and BurA-like transcripts are more or less

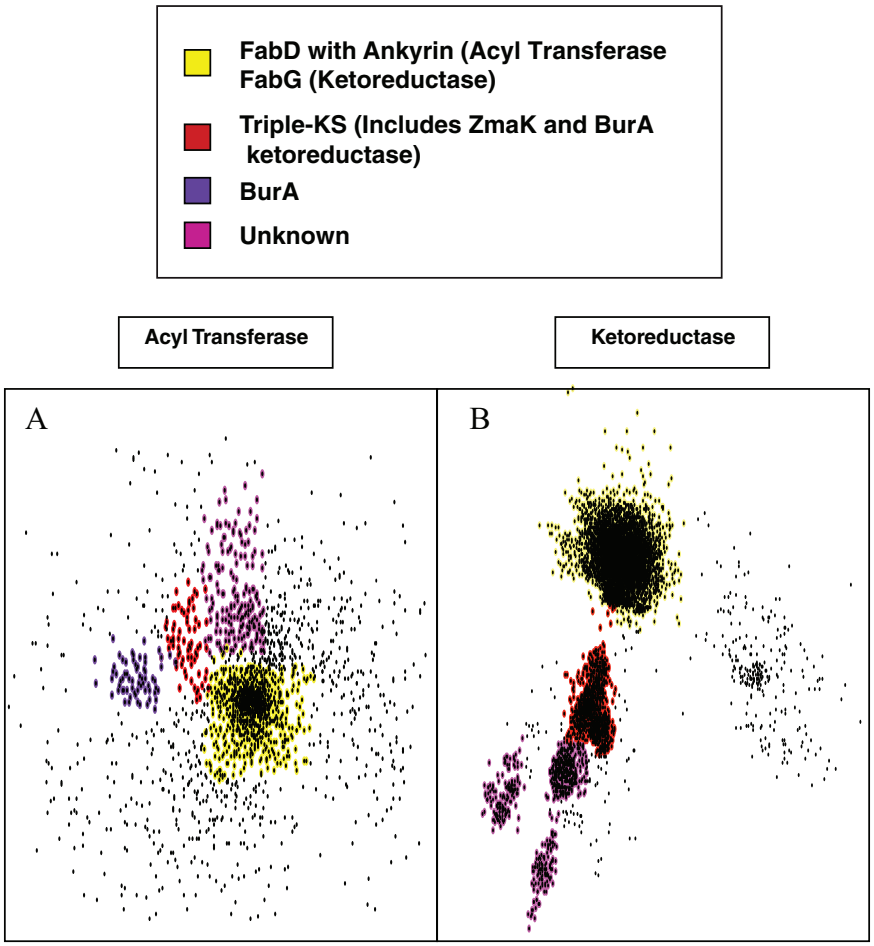

Figure 12. Acyl transferase and ketoreductase protein sequence clusters.
The clusterings of the acyl transferase (A) and ketoreductase (B) protein sequences are shown. Both exhibit gene duplication of apparent genes involved in fat synthesis (yellow) with FabD for the acyl transferases (often with ankyrin domains) and the FabG gene for ketoreductases. Other clusters include the Triple-KS genes (also including the ZmaK and BurA ketoreductases), a separate BurA acyl transferase cluster, and several clusters of unknown similarity.

intact across almost all of the core dinoflagellates and can be readily found by simple domain counting (Supplemental Table S3) or by looking for large transcripts with ketosynthase domains. The ZmaK-like and TeCATe transcripts are less robust and are often truncated or have missing domains but are still readily recognizable. It appears that the BurA-like and ZmaK transcripts were horizontally transferred from bacteria since they are largely absent in the syndiniales but are present in a number of bacterial species as part of conserved synthetic pathways. Although modular synthases were almost entirely absent from *Amoebophyra* species, multi-domain polyketide synthases were found in *Hematodinium* in this study (Table 2) as well as another separate transcriptomics study that determined them to be cytosolic in nature.[74] The sequence arrangement is very similar between the *A. carterae* transcriptome and *Hematodinium* genome polyketide synthases. Similarly in *Toxoplasma* and *Cryptosporidium* there are multi-module PKS genes similar to the dinoflagellate triple KS used as a model here that are theorized to process fatty acids.[75,76] Dinoflagellates are known to make many poly-unsaturated fatty acids[77,78] and these triple KS genes may be involved. *Hematodinium*, unlike the alveolate *Chromera velia* and diatom *Triceratium dubium*, has adenylation domains and condensation domains similar to the TeCATe transcript. Thus, it is possible that the triple KS and TeCATe transcripts were present in some form in the dinoflagellate common ancestor. They could either have been modified or lost such as in the *Amoebophyra* species that infect dinoflagellates and parasitize essential fatty acids from their host, or kept intact in species like *Hematodinium* that infects crustacean hosts not known to make the polyunsaturated fatty acids found in dinoflagellates.

The origin of the single domain transcripts is much harder to ascertain due simply to the sheer number of very similar sequences. Previous studies have focused on phylogenies of adenylation and ketosynthase domains that could be annotated using traditional nomenclature.[34,43,79] This makes sense considering that these 2 domains represent the workhorse enzymes of modular synthesis and there is precedent for the gain or loss of a domain being diagnostic for toxicity.[80] Traditional nomenclature of polyketide synthases, however, is largely based on whether a gene is eukaryotic or prokaryotic and whether it is multi-domain or an assemblage of single domains, *i.e.* type I and type II,[16] and dinoflagellates have been shown to possess genes similar to both eukaryotic and prokaryotic models that

are both single and multi-domain.[34] This nomenclature combined with distantly related model organisms such as humans and yeast is therefore not very useful when trying to unravel the mechanisms underlying dinoflagellate modular synthases. Traditional datasets and nomenclature can also be misleading when trying to annotate sequences since protists are notoriously under-sampled in public databases relative to yeast and vertebrate models as well as prokaryotes. In Kohli et al,[51] 264 ketosynthase and ketoreductase single-domain transcripts as well as 24 multi-domain PKS transcripts were found in *G. excentricus* and *G. polynesiensis* transcriptomes using the BLAST2GO pipeline and HMMs based on annotations from previous studies. The present study using HMMs created from BLAST searches in *A. carterae* of the aforementioned domains in known multi-domain transcripts yielded 156 additional ketosynthase and ketoreductase domains in single and multi-domain transcripts for the 2 *Gambierdiscus* species, although more is not necessarily better, especially when confirming predictions experimentally is still out of reach. All possible genes could play a role in toxin synthesis ignoring bias from annotations in model organisms since it is clear that what is atypical chemistry for model organisms appears to be the norm for toxin synthesis in dinoflagellates.[28] This is also true of the BurA and ZmaK genes themselves that are atypical for prokaryote polyketide synthesis modules but appear to have been successfully transferred and retained in dinoflagellates.

The clustering analysis is in some ways more informative than phylogeny and annotation in that it gives an indication of the level of gene duplication for a domain within dinoflagellates and allows visualization for a large number of sequences without the preconception of a bifurcating style of evolution during speciation. It is also important to include as many domain types and not focus on ketosynthases alone since the loss of an acyl transferase or thioesterase can result in a truncated structure as been hypothesized based on chemical comparison of pinnatoxin and gymnodimine to spirolides.[28] The underlying hypothesis is that domains with sequence similarity may perform similar functions or be involved in similar pathways since this is often the primary constraint on evolution but neofunctionalization is also a possibility. Thus, the clusters were colored according to the presence of domains from the known multi-domain transcripts as a way of binning the clusters and begin to ascertain the functions of the many single domain transcripts. One case is the large number of single adenylation and ketosynthase domain transcripts that are similar to the annotated transcripts as well as other clusters of unknown similarity (Figure 11). Two reasonable explanations for this diversity are that the domains themselves were serially duplicated and fragmented from parent multidomain transcripts resulting in gene expansion, or that there was a functional constraint forcing domains acquired by other means as single domain transcripts to evolve convergently and form multidomain transcripts. It is possible that both are happening, for example, gene duplication for the discrete cluster of BurA-like adenylation domains and convergent evolution for the ZmaK-like adenylation domains that are linked to another cluster of adenylation domains of unknown origin. It is unclear if the ketosynthase domains from the central cluster containing multi-domain transcripts are performing similar functions and the outlying clusters are performing different functions such as chain length factors, or if ketosynthase domains are being acquired faster than convergent evolution is acting. For the thiolation domains convergent evolution is more likely considering that a single cluster encompasses domains from several different multi-domain transcripts while the acyl carrier proteins have their own cluster (Figure 9). This would make sense considering that dinoflagellates only have between 1 and 3 phosphopantetheinyl transferases that can activate these thiolation domains.[71]

The acyl transferase and ketoreductase domain clusterings are especially interesting as the only case where the gene for fat synthesis is present in very large clusters with small clusters containing domains from the multi-domain transcripts (Figure 12). This was first described in the Symbiodiniaceae and described as FabD-like Trans ATs.[43] While horizontal transfer and duplication of entire fat synthesis gene clusters has been shown,[81,82] the extensive gene duplication suggested by the data presented here for the acyl transferase and ketoreductase genes would be unprecedented. Convergent evolution is unlikely given that fat synthesis is usually tightly regulated and none of the other fat synthesis genes show this type of clustering. It is possible that FabD and FabG like genes were coopted for some other function following an initial duplication and gene expansion followed. This would also indicate that these domains are performing a function separate from the multi-domain transcripts given that that they almost always have intact acyl transferase or ketoreductase domains. The triple KS is a special case here since the second ketosynthase is annotated as an acyl transferase containing ketosynthase by the conserved domain database of NCBI and the acyl transferase HMM only detected a domain in some transcripts but not others. This means that a Trans-acting acyl transferase is possible for some of the triple KS modules if the ketosynthase has lost the acyl transferase functionality, but this is speculation given these data. In general the acyl transferase and ketoreductase clusters of unknown similarity were probably acquired later or gene duplication occurred in early dinoflagellates since they are in very low abundance in the basal species, for example, *A. carterae* only has acyl transferases that are BurA-like (8 copies) and FabD-like (10 copies) and only 2 of 31 ketoreductases are found in the unknown cluster (Supplemental Table S2).

The thioesterase and dehydratase clusterings paint a very different picture than the other abundant and diverse domains as one of the few cases where the domain count is consistently low with small clusters (Figure 10). There is still some gene expansion such as the Bur-A like thioesterases in *K. brevis* and

G. spinifera that appear to have been duplicated along with other BurA-like domains, just to a lesser degree (Supplemental Table S2). This small number of thioesterases in most species indicates that for very large toxins the number of synthetic complexes is low or that synthesis is highly iterative since a thioesterase is usually necessary to terminate each portion of synthesis.[16] However, it is important to remember that the "low abundance" of thioesterases is a relative description since thioesterases are more than 9-fold more abundant in the core dinoflagellates than in the outgroup species (Table 2). The dehydratases, like the thioesterases, are usually encountered in multi-domain transcripts (Figure 3). Ketosynthase and ketoreductase domains on the other hand are abundant as single domain transcripts. When looking at the chemical structure of many dinoflagellate toxins the acetate units are frequently hydroxylated, indicating that the ketone has been modified by a ketoreductase but not a dehydratase.[28] These hydroxyls then frequently form epoxide bonds resulting in the "zipped up" structures of brevetoxin and yessotoxin. This makes the abundance of enoyl reductases strange since they would theoretically act after the dehydratases to further saturate the polyketide but the enoyl reductases are much more abundant than the dehydratases (Figure 10). It may be that many of the enoyl reductases have been coopted to operate on a substrate other than polyketides or that the dehydratases act as a chokepoint in synthesis and that their abundance is under tighter regulation or selection pressure. Given the number of enoyl reductase fragments (Figure 10(B)) it may also be that this gene is subject to a much higher level of gene duplication but that not all of the transcripts are being translated. Either way, the large number of enoyl reductases relative to dehydratases in dinoflagellates is in stark contrast to what is frequently described in prokaryote and fungal models where regulation of gene expression is much better understood and domain abundance directly correlates to the structure of the final product.

The phylogenetic analyses attempted on the thioesterase and acyl transferase domains had no bootstrap for all major nodes in spite of being able to produce clades with similar structure to the clustering output in the highest scoring trees (Supplemental Files). The only nodes with bootstrap support above 70% were those containing sequence variants from a single species or assembly variants from a single transcriptome. This is not surprising since gene copy number has made sequence phylogeny difficult for dinoflagellates in the past.[1,2,40] Also, given the amount of horizontal gene transfer the concept of orthology become difficult to prove in general,[3] and in this case a functional approach is more useful if the goal is to extend hypothesis to biochemical characterization.

In general there was a lack of condensation domains despite a large number of adenylation domains in all the core dinoflagellates. Although the condensation domains in the TeCATe transcript used to construct the HMM are similar to canonical condensation domains it is quite possible that there are other condensation domains not associated with multi-domain transcripts. It is certainly true that condensation domains can have their own specificity in natural product synthesis forming both amide and epoxide bonds without the aid of adenylation domains.[83] Condensation domains are unlikely to play a large role in toxin synthesis in dinoflagellates given their almost ubiquitous use of acetate and general lack of amino acids although the frequent use of glycolate as a starter is conspicuous,[28] and an unknown trans-acting condensation domain may be critical in initiating toxin synthesis.

Scaffolding domains and single domain transcripts are associated with toxin synthesis

Given their abundance, one could speculate that it is largely the single domain genes that are responsible for toxin synthesis with multi-domain genes like the triple KS responsible for the synthesis of poly-unsaturated fatty acids or portions of toxins like the acyl chains. It is also possible that these multi-domain genes or modules within them act on specific segments of toxin synthesis either individually or iteratively as has been proposed several times.[38,42,80,84] If it is mostly single domain genes involved in toxin synthesis then the thiolation domains with tetratricopeptide repeats may be important in scaffolding protein domains and providing reaction centers for the large complexes necessary to synthesize toxins.[85] The fact that the fusion of a thiolation domain and a tetratricopeptide repeat is never found in conjunction with another domain and only present in the core dinoflagellates also correlates to toxin synthesis via single domain genes since none of the syndiniales or other alveolates make large polyketide toxins. Also the acyl transferase domains may be involved in their own scaffolding or reaction center bridging given the occurrence of ankyrin repeats in many of the FabD like acyl transferase containing transcripts.[60] The interplay between scaffolding by thiolation domains and reaction center bridging by trans-acting acyl transferases may be a driving force in the evolution of modular synthesis in dinoflagellates. Specifically, the number of acyltransferases relative to thiolation domains increases as one moves from the most basal Gymnodiniales to the more distal Gonyaulacales (Figure 5). This shift is also evident in the decrease in the mean number of thiolation domains in a transcript (Figure 6). The occurrence of multiple thiolation domains in tandem within a transcript was first observed in the *K. brevis* transcriptome and appears to be a hallmark of the Gymnodiniales that include species that make sterolysins and brevetoxin[10,22,86-88] as well as the Suessiales that can make zooxanthellatoxin and zooxanthellamide.[61] Unfortunately this is not diagnostic with many species that do not have a described toxin such as *Pelagodinium beii* having a higher average (1.47) than *Karlodinium veneficum* (1.28) that makes karlotoxin. The *Gambierdiscus* species had the highest average number of thiolation domains among the Gonyaulacales (1.58 and 1.68) that

otherwise had low averages indicating that multiple thiolation domains may be one strategy in synthesizing long polyketides such as ciguatoxin.[89] Unfortunately, many polyketides in dinoflagellates are likely undescribed if they do not act as a toxin in humans making it difficult to correlate the synthesis of dinoflagellate polyketides to any molecular results.

Conclusion

In general there was no overarching signal relating domain count or domain expansion to toxin production as shown in the principal components plots (Figure 4). This is largely due to the abundance of modular synthase genes among all core dinoflagellates investigated. So if core dinoflagellates are all making polyketides, what is their purpose? *Karlodinium veneficum* is the only case where an ecological role has been identified, *i.e.* prey capture,[12] but it is also the only case where the toxin is found readily outside the cell. Another proposed role for a toxin is mediating redox potential by brevetoxin in the chloroplast of *Karenia brevis*.[90] This helps explain why complex polyketides are found in photosynthetic species as well as the apparent association between function and synthesis in the chloroplast since it is a major source of redox stress. Thus, focusing on "toxin" synthesis may not be advantageous in the long run versus understanding the modular synthases in dinoflagellates as a whole and their biological role within dinoflagellates. Just as subtle differences in the availability of a thioesterase or acyl transferase can radically alter the final structure of a polyketide, assays that identify known toxins can falsely label a species or strain as being non-toxic despite that organism making polyketides that are only subtly different than the toxin standards.

The data presented here shows long-term evolution along the entire scope of dinoflagellate history with the acquisition of tetratricopeptide repeats fused to thiolation domains in the core dinoflagellates and the increase in acyl transferase domains as a major component of the synthetic domain population, specifically the FabD-like acyl transferases. It also shows short-term evolution with rapid increases in the copy number of certain domains that was first shown in *Symbiodinium* species,[43] and appears to be a universal feature of dinoflagellate evolution that could also explain why many of the larger toxins are unique to certain lineages. While it seems like a natural progression to use molecular datasets from dinoflagellates to make predictions about the functionality of synthetic domains, existing datasets have been validated with species very distantly related to dinoflagellates, and protists in general, making these predictions unlikely to be realistic. For example the Beedessee et al paper from 2019[43] used up to date methods to predict the substrates of adenylation domains in dinoflagellates resulting in tryptophan, phenylalanine, and glycine. This is unlikely to be true since tryptophan and phenylalanine are not found in described dinoflagellate natural products that would utilize adenylation domains. Also, using the same method as the Beedessee et al paper to predict the substrate for the *A. carterae* adenylation domain of BurA, as well as from the original BurA sequence from *Burkholderia*, similarly results in phenylalanine, but this was shown to actually be a methionine modified to a propanal by radioisotopic labeling in the bacterium.[45] Thus, while abundant molecular data for many dinoflagellate species is certainly a boon in the study of dinoflagellate biology and toxin synthesis in particular, it must be tempered with biochemical validation to determine functionality and to start the process of unraveling synthetic pathways.

ORCID iD

Ernest P Williams https://orcid.org/0000-0001-8727-4968

Supplemental Material

Supplemental material for this article is available online.

REFERENCES

1. Bachvaroff TR, Gornik SG, Concepcion GT, et al. Dinoflagellate phylogeny revisited: using ribosomal proteins to resolve deep branching dinoflagellate clades. *Mol Phylogenet Evol*. 2014;70:314-322.
2. Janouškovec J, Gavelis GS, Burki F, et al. Major transitions in dinoflagellate evolution unveiled by phylotranscriptomics. *Proc Natl Acad Sci U S A*. 2017;114:E171-E180.
3. Keeling PJ. The endosymbiotic origin, diversification and fate of plastids. *Philos Trans R Soc Lond B Biol Sci*. 2010;365:729-748.
4. Schnepf E, ElbräChter M. Dinophyte chloroplasts and phylogeny – a review. *Grana*. 1999;38:81-97.
5. Jacobson DM, Anderson DM. Widespread phagocytosis of ciliates and other protists by marine mixotrophic and heterotrophic thecate dinoflagellates1. *J Phycol*. 1996;32:279-285.
6. Jeong HJ, Yoo YD, Kim JS, Seong KA, Kang NS, Kim TH. Growth, feeding and ecological roles of the mixotrophic and heterotrophic dinoflagellates in marine planktonic food webs. *Ocean Sci J*. 2010; 45:65-91.
7. Monroe EA, Johnson JG, Wang Z, Pierce RK, Van Dolah FM. Characterization and expression of nuclear-encoded polyketide synthases in the brevetoxin-producing dinoflagellate Karenia brevis1. *J Phycol*. 2010;46:541-552.
8. Van Dolah FM, Zippay ML, Pezzolesi L, et al. Subcellular localization of dinoflagellate polyketide synthases and fatty acid synthase activity. *J Phycol*. 2013;49:1118-1127.
9. Kobayashi J. Amphidinolides and its related macrolides from marine dinoflagellates. *J Antibiot*. 2008;61:271-284.
10. Meng Y, Van Wagoner RM, Misner I, Tomas C, Wright JL. Structure and biosynthesis of amphidinol 17, a hemolytic compound from Amphidinium carterae. *J Nat Prod*. 2010;73:409-415.
11. Wang DZ. Neurotoxins from marine dinoflagellates: a brief review. *Mar Drugs*. 2008;6:349-371.
12. Sheng J, Malkiel E, Katz J, Adolf JE, Place AR. A dinoflagellate exploits toxins to immobilize prey prior to ingestion. *Proc Natl Acad Sci U S A*. 2010;107:2082-2087.
13. Van Wagoner RM, Deeds JR, Tatters AO, Place AR, Tomas CR, Wright JL. Structure and relative potency of several karlotoxins from Karlodinium veneficum. *J Nat Prod*. 2010;73:1360-1365.
14. Bentley R, Bennett JW. Constructing polyketides: from collie to combinatorial biosynthesis. *Ann Rev Microbiol*. 1999;53:411-446.
15. Jenke-Kodama H, Dittmann E. Evolution of metabolic diversity: insights from microbial polyketide synthases. *Phytochemistry*. 2009;70:1858-1866.
16. Khosla C. Structures and mechanisms of polyketide synthases. *J Org Chem*. 2009;74:6416-6420.
17. Izoré T, Cryle MJ. The many faces and important roles of protein-protein interactions during non-ribosomal peptide synthesis. *Nat Prod Rep*. 2018;35:1120-1139.
18. Rausch C, Hoof I, Weber T, Wohlleben W, Huson DH. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol Biol*. 2007;7:78.
19. Gurney R, Thomas CM. Mupirocin: biosynthesis, special features and applications of an antibiotic from a gram-negative bacterium. *Appl Microbiol Biotechnol*. 2011;90:11-21.
20. Lim SK, Ju J, Zazopoulos E, et al. Iso-Migrastatin, migrastatin, and dorrigocin production in Streptomyces platensis NRRL 18993 is governed by a single

biosynthetic machinery featuring an acyltransferase-less type I polyketide synthase. *J Biol Chem*. 2009;284:29746-29756.
21. McDaniel R, Thamchaipenet A, Gustafsson C, Fu H, Betlach M, Ashley G. Multiple genetic modifications of the erythromycin polyketide synthase to produce a library of novel "unnatural" natural products. *Proc Natl Acad Sci U S A*. 1999;96:1846-1851.
22. Houdai T, Matsuoka S, Murata M, et al. Acetate labeling patterns of dinoflagellate polyketides, amphidinols 2, 3 and 4. *Tetrahedron*. 2001;57:5551-5555.
23. Lee MS, Qin G, Nakanishi K, Zagorski MG. Biosynthetic studies of brevetoxins, potent neurotoxins produced by the dinoflagellate Gymnodinium breve. *J Am Chem Soc*. 1989;111:6234-6241.
24. Macpherson GR, Burton IW, LeBlanc P, Walter JA, Wright JL. Studies of the biosynthesis of DTX-5a and DTX-5b by the dinoflagellate Prorocentrum maculosum: regiospecificity of the putative Baeyer-Villigerase and insertion of a single amino acid in a polyketide chain. *J Org Chem*. 2003;68:1659-1664.
25. Wright JLC, Hu T, McLachlan JL, Needham J, Walter JA. Biosynthesis of DTX-4: confirmation of a polyketide pathway, proof of a Baeyer-villiger oxidation step, and evidence for an unusual carbon deletion process. *J Am Chem Soc*. 1996;118:8757-8758.
26. Moore BS, Hertweck C. Biosynthesis and attachment of novel bacterial polyketide synthase starter units. *Nat Prod Rep*. 2002;19:70-99.
27. Rasmussen SA, Binzer SB, Hoeck C, et al. Karmitoxin: an amine-containing polyhydroxy-polyene toxin from the marine dinoflagellate *Karlodinium armiger*. *J Nat Prod*. 2017;80:1287-1293.
28. Van Wagoner RM, Satake M, Wright JL. Polyketide biosynthesis in dinoflagellates: what makes it different. *Nat Prod Rep*. 2014;31:1101-1137.
29. Seki T, Satake M, Mackenzie L, Kaspar HF, Yasumoto T. Gymnodimine, a new marine toxin of unprecedented structure isolated from New Zealand oysters and the dinoflagellate, Gymnodinium sp. *Tetrahedron Lett*. 1995;36:7093-7096.
30. Sasaki M, Matsumori N, Maruyama T, et al. The complete structure of Maitotoxin, part I: configuration of the C1-C14 side chain. *Angew Chem Int Ed Engl*. 1996;35:1672-1675.
31. Buhman KK, Chen HC, Farese RV. The enzymes of neutral lipid synthesis. *J Biol Chem*. 2001;276:40369-40372.
32. Marechal E, Block MA, Dorne A-J, Douce R, Joyard J. Lipid synthesis and metabolism in the plastid envelope. *Physiol Plant*. 1997;100:65-77.
33. Tatsuta T, Scharwey M, Langer T. Mitochondrial lipid trafficking. *Trends Cell Biol*. 2014;24:44-52.
34. Kohli GS, John U, Van Dolah FM, Murray SA. Evolutionary distinctiveness of fatty acid and polyketide synthesis in eukaryotes. *ISME J*. 2016;10:1877-1890.
35. Beld J, Sonnenschein EC, Vickery CR, Noel JP, Burkart MD. The phosphopantetheinyl transferases: catalysis of a post-translational modification crucial for life. *Nat Prod Rep*. 2014;31:61-108.
36. Wang H, Fewer DP, Holm L, Rouhiainen L, Sivonen K. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *Proc Natl Acad Sci U S A*. 2014;111:9259-9264.
37. Schümann J, Hertweck C. Advances in cloning, functional analysis and heterologous expression of fungal polyketide synthase genes. *J Biotechnol*. 2006;124:690-703.
38. Van Dolah FM, Kohli GS, Morey JS, Murray SA. Both modular and single-domain Type I polyketide synthases are expressed in the brevetoxin-producing dinoflagellate, Karenia brevis (Dinophyceae). *J Phycol*. 2017;53:1325-1339.
39. Hertweck C, Luzhetskyy A, Rebets Y, Bechthold A. Type II polyketide synthases: gaining a deeper insight into enzymatic teamwork. *Nat Prod Rep*. 2007;24:162-190.
40. Bachvaroff TR, Place AR. From stop to start: tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate Amphidinium carterae. *PLoS One*. 2008;3:e2929.
41. Bachvaroff TR, Williams EP, Jagus R, Place AR. A cryptic noncanonical multimodule PKS/NRPS found in dinoflagellates. The 16 International Conference on Harmful Algae, 27-30 October 2014, Wellington, New Zealand; 2015:101-104.
42. Van Dolah FM, Morey JS, Milne S, Ung A, Anderson PE, Chinain M. Transcriptomic analysis of polyketide synthases in a highly ciguatoxic dinoflagellate, Gambierdiscus polynesiensis and low toxicity Gambierdiscus pacificus, from French Polynesia. *PLoS One*. 2020;15:e0231400.
43. Beedessee G, Hisata K, Roy MC, Van Dolah FM, Satoh N, Shoguchi E. Diversified secondary metabolite biosynthesis gene repertoire revealed in symbiotic dinoflagellates. *Sci Rep*. 2019;9:1204.
44. Meyer JM, Rödelsperger C, Eichholz K, et al. Transcriptomic characterisation and genomic glimps into the toxigenic dinoflagellate Azadinium spinosum, with emphasis on polykeitde synthase genes. *BMC Genomics*. 2015;16:27.
45. Franke J, Ishida K, Hertweck C. Genomics-driven discovery of burkholderic acid, a noncanonical, cryptic polyketide from human pathogenic Burkholderia species. *Angew Chem Int Ed Engl*. 2012;51:11611-11615.
46. Kevany BM, Rasko DA, Thomas MG. Characterization of the complete zwittermicin A biosynthesis gene cluster from Bacillus cereus. *Appl Environ Microbiol*. 2009;75:1144-1155.
47. Sun S, Chen J, Li W, et al. Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic Acids Res*. 2011;39:D546-D551.
48. Bachvaroff TR, Handy SM, Place AR, Delwiche CF. Alveolate phylogeny inferred using concatenated ribosomal proteins. *J Eukaryot Microbiol*. 2011;58:223-233.
49. Williams E, Place A, Bachvaroff T. Transcriptome analysis of core dinoflagellates reveals a universal bias towards "GC" rich codons. *Mar Drugs*. 2017;15:125.
50. Bachvaroff TR. A precedented nuclear genetic code with all three termination codons reassigned as sense codons in the syndinean Amoebophrya sp. Ex Karlodinium veneficum. *PLoS One*. 2019;14:e0212912.
51. Kohli GS, Campbell K, John U, et al. Role of modular polyketide synthases in the production of polyether ladder compounds in ciguatoxin-producing gambierdiscus polynesiensis and G. excentricus (Dinophyceae). *J eukaryot Microbiol*. 2017;64:691-706.
52. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 2015;31:3210-3212.
53. Favrot L, Blanchard JS, Vergnolle O. Bacterial GCN5-relatedN-acetyltransferases: from resistance to regulation. *Biochemistry*. 2016;55:989-1002.
54. Zeytuni N, Zarivach R. Structural and functional discussion of the tetra-tricopeptide repeat, a protein interaction module. *Structure*. 2012;20:397-405.
55. Hunter S, Apweiler R, Attwood TK, et al. InterPro: the integrative protein signature database. *Nucleic Acids Res*. 2009;37:D211-D215.
56. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792-1797.
57. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res*. 2013;41:e121.
58. Frickey T, Lupas A. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*. 2004;20:3702-3704.
59. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 2014;30:1312-1313.
60. Mosavi LK, Cammett TJ, Desrosiers DC, Peng ZY. The ankyrin repeat as molecular architecture for protein recognition. *Protein Sci*. 2004;13:1435-1448.
61. Fukatsu T, Onodera K, Ohta Y, et al. Zooxanthellamide D, a polyhydroxy polyene amide from a marine dinoflagellate, and chemotaxonomic perspective of the Symbiodinium polyols. *J Nat Prod*. 2007;70:407-411.
62. Javed F, Qadir MI, Janbaz KH, Ali M. Novel drugs from marine microorganisms. *Crit Rev Microbiol*. 2011;37:245-249.
63. Fewer DP, Rouhiainen L, Jokela J, et al. Recurrent adenylation domain replacement in the microcystin synthetase gene cluster. *BMC Evol Biol*. 2007;7:183.
64. Khosla C, Kapur S, Cane DE. Revisiting the modularity of modular polyketide synthases. *Curr Opin Chem Biol*. 2009;13:135-143.
65. Piel J. A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of Paederus beetles. *Proc Natl Acad Sci U S A*. 2002;99:14002-14007.
66. Stephens TG, González-Pech RA, Cheng Y, et al. Genomes of the dinoflagellate Polarella glacialis encode tandemly repeated single-exon genes with adaptive functions. *BMC Biol*. 2020;18:56.
67. Lidie KB, Ryan JC, Barbier M, Van Dolah FM. Gene expression in Florida red tide dinoflagellate Karenia brevis: analysis of an expressed sequence tag library and development of DNA microarray. *Mar Biotechnol*. 2005;7:481-493.
68. Morse D, Milos PM, Roux E, Hastings JW. Circadian regulation of bioluminescence in Gonyaulax involves translational control. *Proc Natl Acad Sci U S A*. 1989;86:172-176.
69. Jones GD, Williams EP, Place AR, Jagus R, Bachvaroff TR. The alveolate translation initiation factor 4E family reveals a custom toolkit for translational control in core dinoflagellates. *BMC Evol Biol*. 2015;15:14.
70. Snyder RV, Gibbs PDL, Palacios A, et al. Polyketide synthase genes from marine dinoflagellates. *Mar Biotechnol*. 2003;5:1-12.
71. Williams EP, Bachvaroff TR, Place AR. The phosphopantetheinyl transferases in dinoflagellates. *The 18th International Conference on Harmful Algae. Nantes, France, 21-26 October, 2018. Harmful Algae 2018-from Ecosystems to Socioecosystems*; 2020:176-180.
72. Shoguchi E, Shinzato C, Kawashima T, et al. Draft assembly of the Symbiodinium minutum nuclear genome reveals dinoflagellate gene structure. *Curr Biol*. 2013;23:1399-1408.
73. Murray SA, Diwan R, Orr RJ, Kohli GS, John U. Gene duplication, loss and selection in the evolution of saxitoxin biosynthesis in alveolates. *Mol Phylogenet Evol*. 2015;92:165-180.
74. Gornik SG, Cassin AM, MacRae JI, et al. Endosymbiosis undone by stepwise elimination of the plastid in a parasitic dinoflagellate. *Proc Natl Acad Sci U S A*. 2015;112:5767-5772.
75. Mazumdar J, Striepen B. Make it or take it: fatty acid metabolism of apicomplexan parasites. *Eukaryot Cell*. 2007;6:1727-1735.

76. Zhu G, Li Y, Cai X, Millership JJ, Marchewka MJ, Keithly JS. Expression and functional characterization of a giant Type I fatty acid synthase (CpFAS1) gene from Cryptosporidium parvum. *Mol Biochem Parasitol*. 2004;134:127-135.
77. Leblond JD, Evans TJ, Chapman PJ. The biochemistry of dinoflagellate lipids, with particular reference to the fatty acid and sterol composition of a Karenia brevis bloom. *Phycologia*. 2003;42:324-331.
78. Mansour MP, Volkman JK, Jackson AE, Blackburn SI. The fatty acid and sterol composition of five marine dinoflagellates. *J Phycol*. 1999;35:710-720.
79. John U, Beszteri B, Derelle E, et al. Novel insights into evolution of protistan polyketide synthases through phylogenomic analysis. *Protist*. 2008;159:21-30.
80. Kohli GS, John U, Figueroa RI, et al. Polyketide synthesis genes associated with toxin production in two species of Gambierdiscus (Dinophyceae). *BMC Genomics*. 2015;16:410.
81. Chan CX, Baglivi FL, Jenkins CE, Bhattacharya D. Foreign gene recruitment to the fatty acid biosynthesis pathway in diatoms. *Mob Genet Elements*. 2013;3:e27313.
82. Hutcheon C, Ditt RF, Beilstein M, et al. Polyploid genome of Camelina sativa revealed by isolation of fatty acid synthesis genes. *BMC Plant Biol*. 2010;10:233.
83. Lin S, Van Lanen SG, Shen B. A free-standing condensation enzyme catalyzing ester bond formation in C-1027 biosynthesis. *Proc Natl Acad Sci U S A*. 2009;106:4183-4188.
84. Beedessee G, Kubota T, Arimoto A, et al. Integrated omics unveil the secondary metabolic landscape of a basal dinoflagellate. *BMC Biol*. 2020;18:139.
85. Clairfeuille T, Norwood SJ, Qi X, Teasdale RD, Collins BM. Structure and membrane binding properties of the endosomal tetratricopeptide repeat (TPR) domain-containing sorting nexins SNX20 and SNX21. *J Biol Chem*. 2015;290:14504-14517.
86. Ishida H, Nozawa A, Totoribe K, et al. Brevetoxin B1, a new polyether marine toxin from the New Zealand shellfish, Austrovenus stutchburyi. *Tetrahedron Lett*. 1995;36:725-728.
87. Peng J, Place AR, Yoshida W, Anklin C, Hamann MT. Structure and absolute configuration of karlotoxin-2, an ichthyotoxin from the marine dinoflagellate Karlodinium veneficum. *J Am Chem Soc*. 2010;132:3277-3279.
88. Van Wagoner RM, Deeds JR, Satake M, Ribeiro AA, Place AR, Wright JL. Isolation and characterization of karlotoxin 1, a new amphipathic toxin from Karlodinium veneficum. *Tetrahedron Lett*. 2008;49:6457-6461.
89. Satake M, Morohashi A, Oguri H, et al. The absolute configuration of ciguatoxin. *J Am Chem Soc*. 1997;119:11325-11326.
90. Chen W, Colon R, Louda JW, Del Rey FR, Durham M, Rein KS. Brevetoxin (PbTx-2) influences the redox status and NPQ of Karenia brevis by way of thioredoxin reductase. *Harmful Algae*. 2018;71:29-39.