

**Okinawa Institute of Science and Technology
Graduate University**

**Thesis submitted for the degree
Doctor of Philosophy**

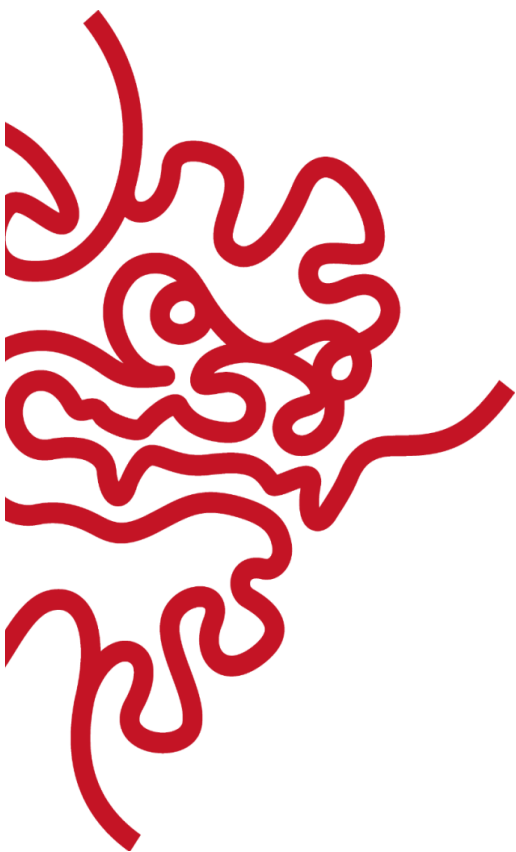
**Genomic insights on secondary metabolism
in symbiotic dinoflagellates**

by

Girish Beedessee

Noriyuki Satoh

April 2019



Declaration of Original and Sole Authorship

I, Girish Beedessee, declare that this thesis entitled “**Genomic insights on secondary metabolism in symbiotic dinoflagellates**” and the data presented in it are original and my own work.

I confirm that:

- This work was done solely while a candidate for the research degree at the Okinawa Institute of Science and Technology Graduate University, Japan.
- No part of this work has previously been submitted for a degree at this or any other university.
- References to the work of others have been clearly attributed. Quotations from the work of others have been clearly indicated, and attributed to them.
- In cases where others have contributed to part of this work, such contribution has been clearly acknowledged and distinguished from my own work.
- None of this work has been previously published elsewhere, with the exception of the following:

1. **Beedessee G**, Hisata K, Roy MC, van Dolah F, Satoh N, Shoguchi E. (2019) Diversified secondary metabolite biosynthesis gene repertoire revealed in symbiotic dinoflagellates. *Sci Reports* 9:1204

2. **Beedessee G**, Hisata K, Roy MC, Satoh N, Shoguchi E. (2015) Multifunctional polyketide synthase genes identified by genomic survey of the symbiotic dinoflagellate, *Symbiodinium minutum*. *BMC Genomics* 16:941

Signature

A handwritten signature in blue ink, appearing to read 'G. Beedessee', is written on a light yellow rectangular background.

Date: 04/19/2019

ABSTRACT

Dinoflagellates (division Pyrrhophyta, class Dinophyceae) are an important group of phytoplankton found in a wide range of environment reflecting a remarkable diversity in form and nutrition styles. They are typically unicellular, photosynthetic, free-swimming and form part of freshwater, brackish and marine phytoplankton communities. Dinoflagellates also produce a wide variety of secondary metabolites including toxins that are dangerous to man, marine animals, fish and other member of food chains. At present, the only available genomes of dinoflagellates are that of the family Symbiodiniaceae. Decoding higher order dinoflagellates remains a challenge because of their large nuclear genomes (up to 250 Gbp). Dinoflagellates highlight the extent of divergence that has taken place in the evolution of eukaryotic life. Taking together the economical, ecological and evolutionary importance of dinoflagellates, undertaking their genome sequencing is a valuable venture. For these reasons, this dissertation aims at understanding how the chemical diversity arises in the family Symbiodiniaceae and explain what evolutionary drivers contribute to this diversity. Next, I decode the genome of a basal dinoflagellate, *Amphidinium gibossum*, known to produce interesting small molecules of biological importance. The purpose of this new genome was to investigate if *A. gibossum* secondary metabolism differs from that of the family Symbiodiniaceae. I found that the underlying chemistry is similar, and I attempt to explain how specialized enzymes generate unique chemical diversity in them. Lastly, I focus on how nutrient starvation affect secondary metabolism in *A. gibossum*. In several dinoflagellates, phosphate and nitrate stress are known to increase or decrease toxin production, but the underlying transcriptomic mechanism remains limited. During such stress conditions, expression of membrane transporters for import of specific ions is upregulated and expression of secondary metabolism is correlated with nutrient availability, involving the action of miRNAs.

Acknowledgments

I wish to thank my supervisor, Prof. Noriyuki Satoh. He gave me the freedom to drive this project the way I wanted and provided all the resources that I could imagine for completing this thesis work. His fresh eyes helped a lot during proofreading of this thesis. I am grateful to Dr. Eiichi Shoguchi, who introduced the amazing world of dinoflagellates to me. He has been very patient throughout this work and has polished my scientific skills.

I also thank all the members of the Marine Genomics Unit, who provide important advice on technical and analysis aspects of this project. Special thanks to Dr. Asuka Arimoto and Dr. Koki Nishitsuji for their help in troubleshooting during experimental and computational analysis.

I would like to acknowledge assistance received from the Dr. Miyuki Kanda (DNA Sequencing Section), Dr. Koji Koizumi (OIST Imaging), Dr. Micheal Roy (Instrumental Analysis Section) and Scientific Computing Section for technical support.

I greatly appreciate the constant support for the OIST graduate school; all the members of this team have taken care of my student life at OIST, allowing me to focus exclusively on my research work. Finally, I would like to thank my wife, Ashmika, who has been very supportive over the past four years, allowing me to lead my work style.

ABBREVIATIONS

HAB	harmful algal bloom
PCP	peridinin-chlorophyll a-protein
AZP	azaspiracid poisoning
ASP	amnesic shellfish poisoning
CFP	ciguatera fish poisoning
DSP	diarrhetic shellfish poisoning
NSP	neurotoxic shellfish poisoning
PSP	paralytic shellfish poisoning
DNA	deoxyribonucleic acid
EST	expressed sequenced tags
PKS	polyketide synthase
ACP	acyl carrier protein
KS	ketosynthase
AT	acyl transferase
KR	ketoreductase
DH	dehydratase
ER	enoylreductase
TBP	TATA binding protein
NRPS	non-ribosomal peptide synthetase
ORF	Open reading frame
DMF	N, N-dimethylformamide

Table of contents

1 Introduction	1
1.1 General features of dinoflagellates	1
1.2 Dinoflagellate genome organization	3
1.3 Transcription in dinoflagellates	4
1.4 Mitochondrial and Chloroplast genomes	5
1.5 Toxin biosynthesis in dinoflagellates	5
1.6 Biotechnological applications of dinoflagellates	9
1.7 Aims of this thesis	9
2 Secondary metabolite genes in Symbiodiniaceae	10
2.1 Introduction	10
2.2 Materials and methods	12
2.2.1 Symbiodiniaceae <i>cultures</i>	12
2.2.2 Data retrieval	12
2.2.3 Phylogenetic analysis	13
2.2.4 Genomics locations and <i>in silico</i> analysis of <i>PKS</i> and <i>NRPS</i> Genes	14
2.2.5 Polyol extraction and mass spectrometry analysis of Symbiodiniaceae cultures.	15
2.2.6 KS protein localization	15
2.3 Results	16
2.3.1 Phylogenetic analyses of ketosynthase and acyltransferase domains.....	16
2.3.2 Phylogenetic analysis of adenylation and condensation domain in NRPS	22
2.3.3 Identification of biosynthetic gene clusters from Symbiodiniaceae	24
2.4 Discussion	25
2.4.1 Evolution of modularity within Symbiodiniaceae <i>genomes</i>	25

2.4.2 Evolution of polyketide biosynthesis	27
2.4.3 Evolution of non-ribosomal peptide biosynthesis	28
2.4.4 Secondary metabolic pathways are conserved in the family Symbiodiniaceae	29
3 Genome of <i>Amphidinium gibossum</i>	31
3.1 Introduction	31
3.2 Materials and methods	33
3.2.1 Biological sample and genome size estimation	33
3.2.2 Genome size estimation	34
3.2.3 DNA sample preparation and sequencing	34
3.2.4 Evaluation of genome completeness and removal of bacterial/viral sequences	35
3.2.5 Transcriptome assembly for generating gene models	35
3.2.6 cDNA construction, Iso-seq sequencing and data processing	36
3.2.7 Annotation of repetitive elements and gene models generation	36
3.2.8 Pfam and KEGG pathway analysis	37
3.2.9 Phylogenetic analysis of PKS and NRPS proteins	37
3.2.10 PKS protein immunolocalization	38
3.3 Results	39
3.3.1 Genomic features of <i>A. gibossum</i>	39
3.3.2 Evidence of multifunctional PKS transcripts in <i>A. gibossum</i>	42
3.3.3 Features of abundant domains, pathway and repetitive elements analysis	43
3.3.4 Analyses of ketosynthase, acyltransferase, adenylation and condensation domains	44
3.4 Discussion	49
3.4.1 The advances of genomic findings of <i>A. gibossum</i>	49
3.4.2 Biochemistry of secondary metabolism in dinoflagellates	49

3.4.3 Secondary metabolism machinery is conserved in dinoflagellates	50
4 Transcriptome of <i>Amphidinium gibossum</i>	53
4.1 Introduction	53
4.2 Materials and methods	54
4.2.1 Biological sample	54
4.2.2 Culture and nutrient treatment	54
4.2.3 Transcriptome analysis, annotation and differential gene expression	55
4.2.4 Bioinformatic analysis of small RNA	56
4.2.5 Identification of key proteins in microRNA biogenesis pathways	57
4.2.6 Mass spectrometry	58
4.2.7 NanoLC-MS analysis of the <i>Amphidinium</i> extract	58
4.3 Results	59
4.3.1 Transcriptome assembly and functional annotation	59
4.3.2 Differential expression analysis under nitrogen starvation	60
4.3.3 Differential expression analysis under phosphate starvation	61
4.3.4 Identification of miRNAs, differential expression and target prediction	65
4.3.5 Metabolomics analysis	66
4.4 Discussion	70
4.4.1 Nitrogen metabolism	70
4.4.2 Phosphate metabolism	70
4.4.3 Secondary metabolism during nutrient starvation	71
4.4.4 <i>Amphidinium gibossum</i> RNAi pathway and its role in nutrient starvation	72

5 Final Conclusion	74
5.1 Symbiodiniaceae genomes generate chemical diversity by expanding its secondary metabolism genes	74
5.2 <i>A. gibossum</i> genome illuminates a conserved secondary metabolism in dinoflagellates	74
5.3 Transcriptome approaches to understand <i>A. gibossum</i> secondary metabolism	75
5.5 Concluding remarks	75
6 References	76
Appendices	

List of Figures

Figure 1.1 Diagrammatic cross-section of a dinoflagellate and phylogenetic relationship of dinoflagellates and acquisition of special characters.....	2
Figure 1.2 Simplified scheme of PKS and NRPS subtypes.....	8
Figure 2.1 Phylogenetic analysis of ketosynthase (KS) domains of eukaryotic and prokaryotic polyketide and fatty acid synthases.....	18
Figure 2.2 Phylogenetic analysis of acyltransferase (AT) domain of eukaryotic and prokaryotic polyketide and fatty acid synthases.....	20
Figure 2.3 Pathway duplication and conservation within and across <i>Symbiodiniaceae</i>	21
Figure 2.4 Phylogenetic comparison of adenylation (A) and condensation (C) domains of prokaryotic and eukaryotic NRPS.....	23
Figure 2.5 Multifunctional PKS genes in <i>Symbiodiniaceae</i>	25
Figure 3.1 General features of <i>Amphidinium gibossum</i>	40
Figure 3.2 PKS transcripts recovered from Iso-Seq	42
Figure 3.3 KEGG pathway analysis in <i>A. gibossum</i>	44
Figure 3.4 Phylogenetic analysis of ketosynthase (KS) and acyltransferase (AT) domains	46
Figure 3.5 Phylogenetic comparison of adenylation (A) and condensation (C) domains	47
Figure 3.6 Immunofluorescent staining of <i>Amphidinium</i> cells with anti-KS and anti-KR antibodies	48
Figure 3.7 Biosynthesis of specialized metabolites from <i>Symbiodiniaceae</i> and <i>A. gibossum</i> dinoflagellates	52
Figure 4.1 Gene annotation of <i>Amphidinium gibossum</i> unigenes using gene ontology (GO)...	60
Figure 4.2 Global expression profile of differentially expressed genes under nitrogen starvation	62

Figure 4.3 Global expression profile of differentially expressed genes under phosphate starvation	64
Figure 4.4 NanoLC-MS profile of the methanol extract of <i>Amphidinium gibossum</i> at three time points	67
Figure 4.5 Summary of cellular overview of the main differential expressed genes during nitrogen and phosphate starvation	68
Figure 4.6 Alignment of functional domains of the <i>A. gibossum</i> homolog.....	69

List of Tables

Table 3.1 Genome statistics of <i>Amphidinium gibossum</i> and other Symbiodiniaceae.....	41
Table 3.2 Top 30 abundant domains in <i>A. gibossum</i>	41
Table 4.1 Significantly enriched KEGG pathways upregulated under N starvation	63
Table 4.2 Significantly enriched KEGG pathways downregulated under N starvation	63
Table 4.3 Significantly enriched KEGG pathways upregulated under P starvation	63

Appendix

Appendix A | Figure showing GC plots of 4 scaffolds associated with dinoflagellate PKS-I

Appendix B | Table showing features of LTR-retrotransposons identified from PKS and NRPS-associated scaffolds

Appendix C | Figure showing nanoLC-MS profile and mass spectrum of methanol fraction of clade A3, B1 and C.

Appendix C | Figure showing similarity profile of methanol extract of clade B1 and C

Appendix E | Figure showing immunofluorescent staining of *Cladocopium* sp. (clade C) cells

Appendix F | Phylogenetic analysis of alignment of *Amphidinium* partial LSU rDNA

Appendix G | Figure showing comparison with FACS of *A. gibossum*

Appendix H | Table showing details of genome assembly and annotation statistics

Appendix I | Figure showing recovery of BUSCO and CEGMA genes in *A. gibossum*

Appendix J | Table showing *A. gibossum* repeat content

Appendix K | Table showing examples of some potent amphidinolides

Appendix L | Figure showing NMR profile of methanol extract of *A. gibossum*

Appendix M | Figure showing physiological parameters of *A. gibossum* under N-P depletion

Appendix N | Figure showing top 10 represented KEGG pathways

Appendix O | Figure showing length and distribution of microRNAs detected

Appendix P | Figure showing gene ontology of predicted target unigenes of 1 differentially expressed miRNA under nitrate stress

Appendix Q | Figure showing gene ontology of predicted target unigenes of 4 differentially expressed miRNA under phosphate stress

1 General features of dinoflagellates

1.1 Introduction

Dinoflagellates are a phylum of unicellular eukaryotes, mostly 10-100 μm in size, living in diverse ecosystems. They are characterized by two flagella and a unique cell-covering called the theca (Lin, 2011). Dinoflagellates belong to the group Alveolata, which also contains two other phyla, Ciliata and Apicomplexa. The ciliates are mostly unicellular heterotrophs or parasitic while apicomplexans are mostly animal parasites and contain a nonphotosynthetic plastid (apicoplast) (Wisecaver & Hackett, 2011). Dinoflagellates are important eukaryotic producers in the ocean and play important roles as symbionts in reef-forming corals (Coffroth & Santos, 2005). They also produce a wide range of secondary metabolites that have significant impact on the fisheries and marine ecosystems (Wang, 2008). Based on theca, two different cell types can be seen: (1) fragile and naked unarmored cells that have an outer plasmalemma surrounding a single layer of flattened vesicles and (2) rigid armored dinoflagellates that have cellulose or other polysaccharides within vesicles (Hackett *et al.*, 2004).

The two flagella facilitate motility; one is rooted in the sulcus (longitudinal groove) and directs the cell while the second is found in the cingulum (transverse groove) and is involved on propelling (Figure 1.1a). In alveolates, dinoflagellates form a monophyletic group and are closely related to apicomplexans, having diverged 800-900 million years ago (Hackett *et al.*, 2007; Bhattacharya *et al.*, 2007). Dinoflagellates consist of eight major classes, namely Gonyaulacales, Prorocentrales, Gymnodiniales, Peridinales, Suessiales, Noctilucales, Syndiniales and Blastodinales. The basal lineages and evolutionary relationships among the classes still remain debatable (Hoppenrath & Leander, 2010; Janouskovec *et al.*, 2017) (Figure 1.1b)

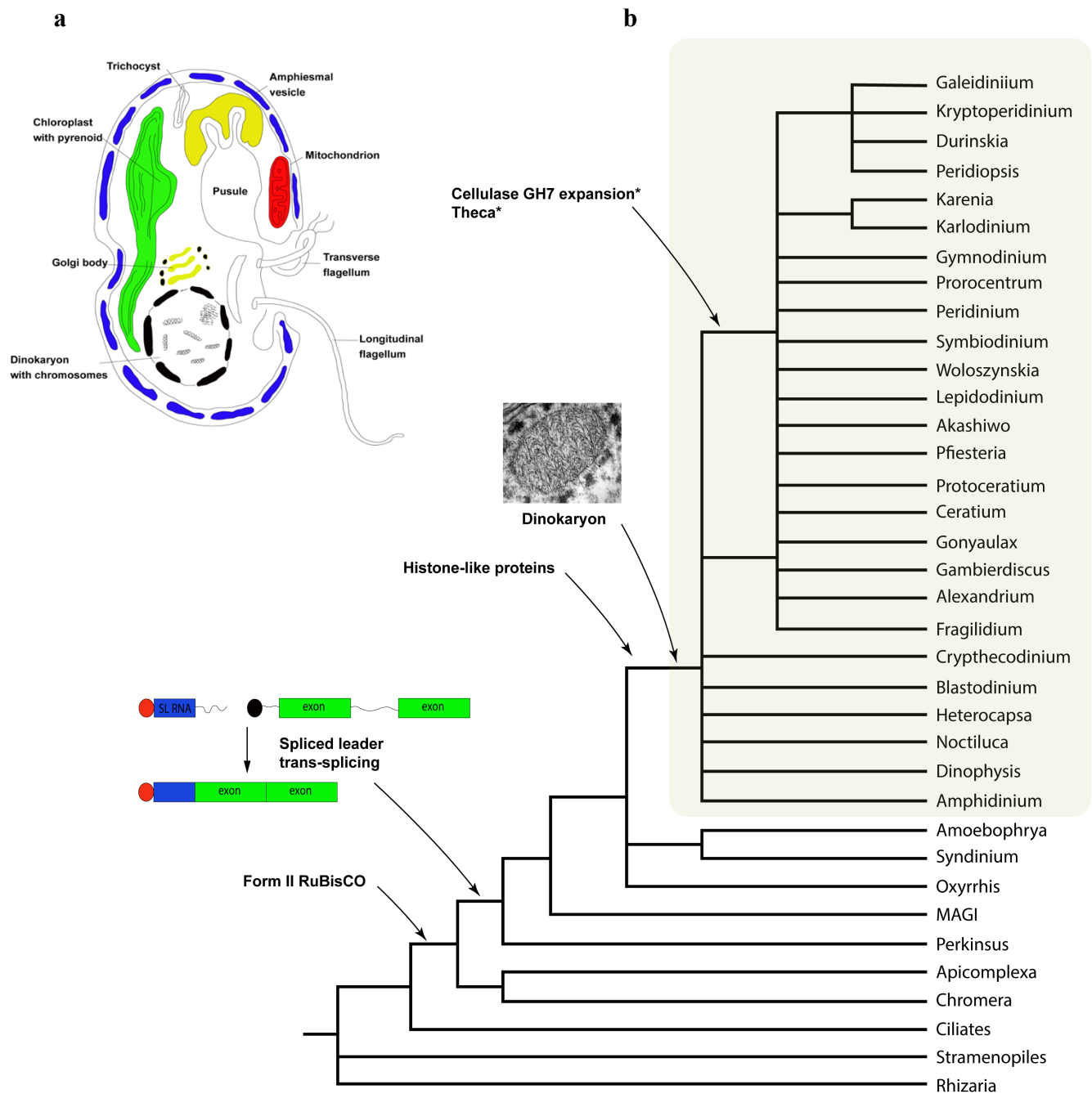


Figure 1.1 | (a) Diagrammatic cross-section of a dinoflagellate. (Redrawn from Taylor, 1980) **(b) Phylogenetic relationship of dinoflagellates and acquisition of special characters during evolution.** The shaded box represents the core dinoflagellates. (Modified from Wisecaver & Hackett (2011)).

1.2 Dinoflagellates genome organization

Dinoflagellates have a number of unique features that distinguish them from other eukaryotes, namely large amount of DNA (LaJeunesse, 2005), unusual bases (Rae, 1976), and absence of nucleosomes (Rizzo, 1972; Haapala, 1973). The occurrence of these characteristics justifies the need to elucidate structure and composition of dinoflagellate genomes. A 616-Mbp gene-rich nuclear DNA assembly from an estimated 1.5-Gbp of the coral symbiont, *Symbiodinium minutum* was the first dinoflagellate genome decoded (Shoguchi *et al.*, 2013). In the past few years, several other *Symbiodinium* genomes have been decoded (Lin *et al.*, 2015; Aranda *et al.*, 2016; Shoguchi *et al.*, 2018; Liu *et al.*, 2018). These reports showed the uniqueness and divergent characteristics of dinoflagellates genomes when compared to other eukaryotes.

Symbiodinium spp. are reported to possess the smallest genomes in dinoflagellates, ranging from 1.5-4.8 pg DNA per haploid genome (LaJeunesse, 2005) while the largest genome is found in *Prorocentrum micans* (250 pg DNA per haploid genome) (Veldhuis, 1997). Genes usually occur in multiple copies in tandem arrays, with the number of copies varying between 20-10,000 (e.g. protein kinases in *L. polyedrum*, actin in *A. carterae* and rDNA in *Alexandrium* spp., respectively) (Salois & Morse, 1997; Bachvaroff & Place, 2008; Galluzzi *et al.*, 2009). Using a regression model (Hou and Lin, 2009), a recent estimate of 34,156 and 75,461 genes was proposed for small and large dinoflagellates, respectively (Murray, 2016). To accommodate such large amount of genetic material, dinoflagellate nuclei contain large numbers of chromosomes, up to 270 (Rizzo, 2003).

Nuclear DNA in dinoflagellates occurs in liquid crystalline form (Bouligand, 2001; Chow *et al.*, 2010) and chromosomes are permanently condensed and appear as “bands” under the electron microscope (Rizzo, 2003). Dinoflagellate nuclear DNA is found to be extensively methylated; up to 70% of the thymine is replaced by 5-hydroxymethyluracil (Rae, 1978). A potential gene involved in methylation regulation, S-adenosylmethionine (SAM) has been

associated with saxitoxin synthesis (Harlow, 2007). Dinoflagellate introns are also unusual and have been found not to obey any known splice site consensus sequence MAG | GTRAGT at the 5' splice site and CAG | G at the 3' splice-site (Mount, 1992; Zhang, 1998). The *Symbiodinium* genomes have been shown that GC and GA are also present 5' splice site, in addition to GT. Additional features include the unusual arrangement of genes, namely a unidirectionally aligned gene and a cluster-like gene organization (Shoguchi *et al.*, 2013; Lin *et al.*, 2015; Aranda *et al.*, 2016; Shoguchi *et al.*, 2018; Liu *et al.*, 2018).

1.3 Transcription in dinoflagellates

One major feature of dinoflagellate transcription is the addition of conserved sequence, spliced leader (SL) at the 5' end of mRNA molecules. The presence of the this 22-nt leader sequence on the end of 5' end of transcripts was revealed in expressed sequenced tags (ESTs) from several dinoflagellates (Zhang *et al.*, 2007b; Lidie & Van Dolah, 2007). The role of SL trans-splicing is to convert polycistronic mRNA to monocistronic mRNA, and this might regulate gene expression (Zhang *et al.*, 2007b). *cis*-regulatory elements such as TATA box appear to be absent in dinoflagellate genomes; however, a new class of transcription initiation factor with strong homology to TATA box-binding proteins (TBP) has been found in dinoflagellates (Guillebault *et al.*, 2002). Recent data identified TTTT and TTTG as the most represented and conserved motifs in *S. kawagutii*, suggesting the possibility of replacement of TATA box conserved position with TTTT in dinoflagellates (Lin *et al.*, 2015).

Transcriptional regulation in dinoflagellates is a feature that differs from other eukaryotes; lesser genes (~5-30%) appear to be regulated at the transcription level compared to post-translational stage (Johnson *et al.*, 2012). MicroRNAs (miRNAs) are likely involved in controlling gene expression post-transcriptionally. In recent years, relatively few studies have reported the presence of miRNAs in dinoflagellates (Baumgarten *et al.*, 2013; Gao *et al.*, 2013;

Lin *et al.*, 2015). In *S. kawagutii*, miRNAs are believed to control 6026 genes, mostly linked with metabolic processes, and interestingly, some target genes in the coral host *Acropora digitifera* (Lin *et al.*, 2015). During phosphorus limitation in *Prorocentrum donghaiense*, miRNA sequencing revealed 17 miRNAs, possibly regulating 3268 protein-coding genes (Shi *et al.*, 2017).

1.4 Mitochondrial and Chloroplast genomes

In comparison to their nuclear genomes, organelle genomes of dinoflagellates are smaller in terms of number of genes. Dinoflagellate mitochondrial genomes are highly reduced with only three protein-coding genes (*cob1*, *cox1* and *cox3*) and two highly fragmented rRNAs (Jackson *et al.*, 2007; Kamikawa *et al.*, 2009; Nash *et al.*, 2007). No tRNAs have been found in the mitochondrial genomes, suggesting the total dependence on imported tRNAs for protein translation (Waller & Jackson, 2009). Dinoflagellate mitochondrial and chloroplast mRNAs undergo extensive and diverse editing compared to the largely limited A → G and C → U changes that occur in other eukaryotes. Nine types of editing have been reported in dinoflagellates (Lin, 2008). RNA editing is absent from ciliates and apicomplexans and has evolved independently in dinoflagellates, acting mainly at protein-coding and rRNA gene level. Many chloroplast and mitochondrial genes have been transferred to the nucleus (Zhang, 1999; Hackett *et al.*, 2004; Howe *et al.*, 2008). Once these transferred genes are transcribed and translated, their protein products are imported into their respective organelles (Jackson *et al.*, 2007; Nash *et al.*, 2008; Slamovits *et al.*, 2007).

1.5 Toxin biosynthesis in dinoflagellates

Marine algal toxins have been grouped in relation to six human illnesses: azaspiracid poisoning (AZP), amnesic shellfish poisoning (ASP), ciguatera fish poisoning (CFP), diarrhetic shellfish

poisoning (DSP), neurotoxic shellfish poisoning (NSP), and paralytic shellfish poisoning (PSP), respectively. Four of these are caused by dinoflagellate-derived polyketide toxins (Rein & Snyder, 2006). Some toxins are small heterocyclic guanidinium alkaloids while others are derivatives of polyketides. Polyketides are biosynthesized by specific enzymes called polyketide synthases (PKSs) via the sequential Claisen condensations of small carboxylic acid subunits in a fashion similar to fatty acid biosynthesis. Traditionally, polyketide synthases have been classified into three types (Type I, II and III); however, there have been suggestions to reconsider this classification scheme (Shen, 2003). Dinoflagellate-derived polyketides are grouped based on their structural type; (i) polyether ladders, (ii) macrocycles (including macrolides and non-macrolides), and (iii) linear polyethers (Rein, 1999). Polyketide synthase (PKS) and non-ribosomal peptide synthase (NRPS) are two important classes of modular enzymes involved in secondary metabolite biosynthesis, where modules integrate building blocks into a growing chain like an assembly line. As shown in Figure 1.2a, the core enzymes of PKSs include ketosynthase (KS), acyl transferase (AT), and acyl carrier protein (ACP) (PP-binding) domains. In addition, polyketide synthesis may involve three optional domains: ketoreductase (KR), dehydratase (DH), and enoylreductase (ER) (Figure 1.2a). Type I PKSs are large multifunctional enzymes in which several domains are found in a single protein (Figure 1.2c). Type II PKSs are multiprotein complexes of several individual enzymes. Type III PKSs are mainly involved in flavonoid biosynthesis in plants.

On the other hand, NRPSs are modular multi-enzyme complexes that synthesize a diverse array of biologically active peptides or lipopeptides (Schwarzer *et al.*, 2003). Biosynthesis of non-ribosomal peptides occurs via the action of catalytic modules within NRPS, that are composed of three compulsory domains; adenylation (A), thiolation (T) and condensation (C). The process involves recognition of amino acid (or hydroxyl acid) by the A-domain, covalent attachment of the adenylated amino acid to a phosphopantetheine carrier of

the T-domain, and finally peptide bond formation between two consecutively bound amino acids to a growing peptide chain by the C-domain. These core domains are often supported by domains such as an epimerization (E) domain, a dual/epimerization (E/C) domain, a reductase (R) domain, a methylation (MT) domain, and a cyclization (C) domain or an oxidation (Ox) domain, respectively (Marahiel *et al.*, 1997). Finally, PKSs and NRPSs have a fourth common domain, the thioesterase (TE) domain, that releases the assembled polypeptide and polyketide chains from the enzyme complex (Figure 1.2b). PKS and NRPS pathways often cross-talk such that a polyketide product is elongated by NRPS or *vice versa* to produce hybrid natural products. The role of several transcriptionally regulated genes during the subphase stage of cell cycle has even been linked to toxin biosynthesis in the dinoflagellate *Alexandrium fundyense* (Taroncher-Oldenburg, G & Anderson, 2000).

Type I PKS genes were first identified using a PCR approach in several dinoflagellates and several experiments supported a dinoflagellate origin for most of the PKS genes (Snyder *et al.*, 2003). Over the years there have been reports of monofunctional PKS genes being characterized from several dinoflagellates (Monroe & Van Dolah, 2008; Eichholz *et al.*, 2012; Salcedo *et al.*, 2012; Pawlowicz *et al.*, 2014; Meyer *et al.*, 2015; Kohli *et al.*, 2015). However, recent surveys have started to reveal the presence of multifunctional PKS domains within dinoflagellates along with the commonly found monofunctional domains (Beedessee *et al.*, 2015; Kohli *et al.*, 2017; Van Dolah *et al.*, 2017).

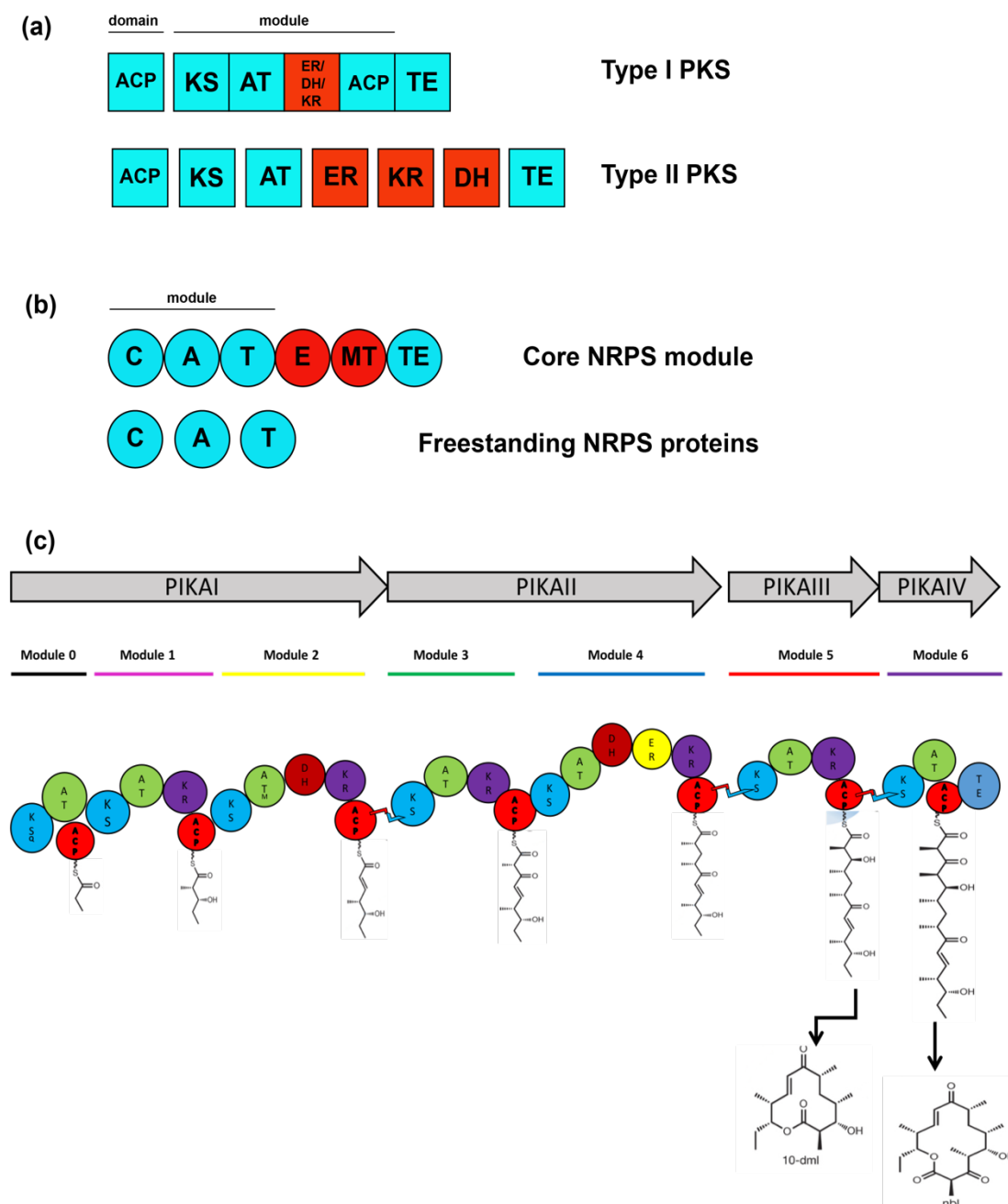


Figure 1.2 | Simplified scheme of PKS (a) and NRPS subtypes (b). Blue shapes are compulsory domains while red shapes are optional domains. (c) An example of a modular polyketide synthase for pikromycin, consisting of 6 modules made of PIKAI-IV polypeptides for polyketide biosynthesis (Modified from Dutta *et al.*, 2014).

1.6 Biotechnological applications of dinoflagellates

Dinoflagellate toxins have gained increasing interest for biotechnology and potential medical applications. Okadaic acid, causative agent for DSP, was linked to several health risks and been useful for understanding cellular role of phosphatases (Tunez, 2003). It is also a model potent neurotoxin for studying changes in schizophrenia and other neurodegenerative diseases (He *et al.*, 2005). Okadaic acid can behave as an inhibitor of protein phosphatase 2A and thus has been used to investigate mechanisms of anti-tumor agents on breast cancer (Liu & Sidell, 2005).

Compounds known as zooxanthellatoxins (ZTs) and zooxanthellamides (ZADs) with potent vasoconstrictive and cytotoxic activity have been isolated from several strains of cultured dinoflagellate *Symbiodinium* sp. (Nakamura *et al.*, 1995a; Nakamura *et al.*, 1995b; Onodera, 2005; Fukatsu, 2007). Symbioimine, obtained from the same dinoflagellate is a potential drug for prevention and treatment of osteoporosis in postmenopausal women and maybe useful in development of anti-inflammatory drugs against cyclooxygenase-2-associated diseases (Kita *et al.*, 2005). Antifungal agents, gambieric acids A-D, have been isolated from the marine dinoflagellate *Gambierdiscus toxicus* (GIII strain) and have been found to display significant activity against filamentous fungi, in some cases 2000-fold more active than amphotericin B (Nagai, 1992; Nagai, 1993).

1.7 Aims of this thesis

Based on the background mentioned above, this thesis aims to address three questions, namely (1) how chemical diversity arises in the late-branching dinoflagellate family Symbiodiniaceae; (2) whether the genome of the early-branching dinoflagellate, *Amphidinium gibossum*, follows the same metabolic code as Symbiodiniaceae; and (3) does nutrient stress affect secondary metabolism in *Amphidinium gibossum*

2 Secondary metabolite genes in Symbiodiniaceae

2.1 Introduction

Dinoflagellates of the family Symbiodiniaceae (LaJeunesse *et al.*, 2018) have symbiotic associations with many invertebrates, such as corals and clams. This invertebrate-Symbiodiniaceae relationship appears to provide a competitive advantage (Trench, 1979), causing the production and exchange of metabolites by members of this mutualism (Lewis & Smith, 1971). This genus is known to be sources of unusual, large, polyhydroxyl and polyether compounds or “super-carbon-chain compounds (SCC),” made of long-chain scaffolds functionalized by oxygen (Uemura, 1971). Molecular phylogenetic analysis has also classified diverse members of this family into nine clades (A to I) by molecular phylogenetic analysis (Pochon & Gates, 2010). Zooxanthellatoxins (ZTs) and zooxanthellamides (ZADs) are some of these compounds that have been isolated from numerous clades and a clade-to-metabolite connection has been suggested and experimentally supported, in which specific Symbiodiniaceae can produce particular metabolites (Fukatsu *et al.*, 2007). Nakamura *et al.* (1998) proposed the existence of common biogenetic processes, such as the polyketide pathway, that generates products similar to palytoxins and zooxanthellatoxins. Several other secondary metabolites have been characterized from these clades, but their ecological functions and biosynthetic pathways are yet to be identified (Gordon & Leggat, 2010).

A genomic survey revealing how secondary metabolite genes are organized in *Breviolum minutum*, added much information to prior transcriptomic analyses (Beedessee *et al.*, 2015). New Symbiodiniaceae genomes are now available that permit us to survey and compare genes involved with metabolite biosynthesis (Shoguchi *et al.*, 2013; Lin *et al.*, 2015; Aranda *et al.*, 2016; Shoguchi *et al.*, 2018). However, the question of how chemical diversity arises in Symbiodiniaceae remains unanswered. The evolution of novel chemistry is depended on diversity-generating metabolism, which encompasses broad-substrate enzymes (Williams

et al., 1989). Metabolic pathways can accept several different substrates, producing diverse chemical products and this offers organisms a unique chemistry to face environmental challenges (Murray *et al.*, 2016). There are two main classes of modular enzymes that are involved in secondary metabolite biosynthesis, namely polyketide synthase (PKS) and non-ribosomal peptide synthase (NRPS), that function like an assembly line where modules incorporate building blocks into a growing chain (Wang *et al.*, 2014). PKS and NRPS pathways often cross-talk where a polyketide product can be elongated by NRPS or vice versa to make hybrid natural products, thereby increasing structural diversity (Du *et al.*, 2001).

Pathways that play a role in secondary metabolite biosynthesis are among the most fast evolving genetic elements (Fischbach *et al.*, 2008). Many processes such as gene loss, duplication, and horizontal gene transfer (HGT) have played important roles in spreading of PKSs in fungi and bacteria (Kroken *et al.*, 2003; Jenke-Kodama *et al.*, 2005). Within *PKS* and *NRPS* genes, mutations, domain rearrangements, and module duplications are known to generate novel, diverse small-molecules (Fischbach *et al.*, 2008). Several entry points exist where combinatorial potential arises. The AT domain in PKS shows specificity for malonyl-CoA, methylmalonyl-CoA, or other malonyl-CoAs, while the KR domain can produce two stereoisomers (Caffrey, 2003). On the contrary, NRPS can accept 500 different monomers such as nonproteinogenic amino acids, fatty acids and α -hydroxyl acids (Caboche *et al.*, 2008; Strieker *et al.*, 2010). Different tailoring enzymes such as glycosyltransferases, halogenases, methyltransferases, and oxidoreductases can additionally modify the chemical structure of secondary metabolites by adding various functional groups (Rix *et al.*, 2002).

To probe the existence of shared biosynthetic pathways, three Symbiodiniaceae (clades A3, B1, and C) were investigated, these being known to synthesize different metabolites, and I surveyed their genomes for genes implicated in polyketide and non-ribosomal peptide biosynthesis. I further examined how these genomes are armed to enlarge their gene catalogue

for biosynthesis of complex secondary metabolites and propose possible diversification strategies that have contributed to such chemical diversity.

2.2 Material and methods

2.2.1 Symbiodiniaceae cultures

Symbiodinium tridacnidorum (Clade A3) and *Cladocopium* sp. (clade C) were collected from the clam *Tridacna crocea* and bivalve *Fragum* sp., respectively, by late Dr. Terufumi Yamasu (University of the Ryukyus, Okinawa, Japan). *Breviolum minutum* (Clade B1) was collected from the stony coral, *Montastraea faveolata* by Dr. Mary Alice Coffroth (University of New York, Buffalo, USA). The cultures were grown in autoclaved, artificial seawater containing 1X Guillard's (F/2) marine-water enrichment solution (Sigma-Aldrich: G0154), complemented with antibiotics (ampicillin (100 µg/mL), kanamycin (50 µg/mL), and streptomycin (50 µg/mL). The protocol of Shoguchi *et al.* (2013) was followed for culturing and sampling of the dinoflagellates.

2.2.2 Data retrieval

PKS (KS & AT), FAS (FabB-KASI, FabF-KASII & FabD) and NRPS (A & C) sequences for the clades A3, B1, C, and *Fugacium kawagutii* were accessed from two genome browser (<http://marinegenomics.oist.jp/genomes/gallery/>, http://web.malab.cn/symka_new/genome.js) (Koyanagi *et al.*, 2013; Lin *et al.*, 2015). Additionally, transcriptome data for several dinoflagellates, apicomplexans, stramenopiles, and haptophytes were retrieved from the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (<http://datacommons.cyverse.org/browse/iplant/home/shared/imicrobe/camera>) and reviewed for comparative analysis (Keeling *et al.*, 2014). Amino acid sequences of PKS and NRPS domains of other animals, prokaryotes, fungi, and chlorophytes were obtained from Genbank

with bonus sequences from dinoflagellates (Eichholz *et al.*, 2012; Kohli *et al.*, 2017). Supplementary NRPS sequences from *Proteobacteria*, *Firmicutes*, and *Cyanobacteria* were retrieved from Wang *et al.* (2014). Conserved active-site residues and functional prediction in sequences were identified using Pfam (Punta *et al.*, 2012). PKS, FAS, and NRPS sequences with full domains and conserved active sites were used. Throughout this chapter, gene models from the three Symbiodiniaceae genomes (A3, B1 and C) are tagged with the letters A, B, and C to improve the readability and interpretation.

2.2.3 Phylogenetic analysis

For Bayesian inference and maximum likelihood analysis, Type I and II PKS/FAS and condensation (C) and adenylation (A) domain sequences representing different taxa were used. Domain sequence datasets were aligned separately using the MUSCLE algorithm, which consisted of 233 KS sequences (226 aa), 96 AT sequences (208 aa), 117 A-sequences (400 aa), and 110 C-sequences (260 aa) (Edgar *et al.*, 2004). Unaligned regions (e.g. large insertions and deletions) were removed before phylogenetic analyses. Maximum likelihood phylogenetic analysis was conducted using RaxML with 1000 bootstraps using the GAMMA and Le-Gasquel amino acid replacement matrix (Stamatakis *et al.*, 2014). Bayesian inference was implemented with MrBayes v.3.2 using the same replacement model (maximum of six million generations and four chains or until the posterior probability approached 0.01) (Ronquist *et al.*, 2012). Trees and statistics were summarized using a 25% burn-in of the data. The two methods estimate phylogeny based on different assumptions and algorithms. Figtree (<http://tree.bio.ed.ac.uk/software/figtree/>) was used to edit trees.

2.2.4 Genomic locations and *in silico* analysis of *PKS* and *NRPS* genes

The Latent Semantic Indexing of the LSI-based A-domain predictor was used to determine the specificity of the A-domain (Baranašić *et al.*, 2014). In order to determine C-domain types, NaPDos was used (Ziemert *et al.*, 2012). Symbiodiniaceae AT sequences were compared to the Hidden Markov Model-based ensemble (HMM) of Khayatt *et al.* (2013). Additional information on possible substrate specificity was predicted using I-TASSER (Zhang, 2008). To identify NRPS and PKS gene clusters within given scaffold regions, AntiSMASH (Antibiotics & Secondary Metabolite Analysis SHell) version 4.1.0 was used with default settings using nucleotides sequences as queries (Blin *et al.*, 2017). The subcellular localization of PKS proteins (e.g. chloroplast and mitochondria) and the presence of signal peptide or membrane anchor were predicted using ChloroP 1.1 and TargetP 1.1 (cut-off score of ≥ 0.50) and the subcellular localization predictor, DeepLoc, respectively (Emanuelsson *et al.*, 1997; Emanuelsson *et al.*, 2007; Armenteros *et al.*, 2017). To align and visualize syntenic relationships between the three genomes, NUCmer operation of SyMap v4.2 (Synteny Mapping and Analysis Program) was used (Soderlund *et al.*, 2011). GFFs (General Feature Files) containing scaffold information and descriptions of these genomes were imported into SyMap. An all-against-all BLAST search of *PKS*-coding scaffolds of one genome against itself was conducted at a BLAST bit score cutoff of ≥ 100 and e-value $\leq e^{-20}$, so as to determine orthologs. Outputs were parsed, and orthologous pair detection was completed using custom perl scripts. All possible segmental duplications were visualized using Circos (Krzywinski *et al.*, 2009). GC-profile was used to analyse GC content variations in *PKS*-coding scaffolds using a halting parameter of 100 (Gao & Zhang, 2006). Long terminal repeat (LTR) retrotransposon-specific features were detected using LTR Finder 1.05 with defaults parameters (Xu & Wang, 2007).

2.2.5 Polyol extraction and mass spectrometry analysis of Symbiodiniaceae cultures

Cultured cells were collected by centrifugation (9000 xg, 14,000 g, 10 min, 10°C) and extracted with methanol (3 times at RT). Subsequent extraction was conducted following Beedessee *et al.* (2015). All crude extracts were lyophilized and stored at -30 °C. MS data was acquired using A Thermo Scientific hybrid (LTQ Orbitrap) mass spectrometer, and high-resolution MS spectrum was collected at 60,000 resolution in FTMS mode (Orbitrap), at full mass range (m/z 400-2,000 Da) with spray voltage (1.9 kV), capillary temperature (200 °C), and both negative and positive ion modes. Crude extract was diluted (1:50) and separated on a capillary ODS column. A 20-min gradient was used for polyol separation.

2.2.6 Immunofluorescence

KS proteins were visualized using a modified protocol of Berdieva *et al.* (2018). Cells were prefixed in methanol: F/2 medium (1:1) at RT for 15 min. After overnight fixation in methanol at -20 °C, cells were washed in PBS, followed by permeabilization (1% Triton X-100 for 15 min except for 5 min for clade B1). Cells were then washed with PBS and blocked with 5% normal goat serum-PBST (1h). After overnight incubation at 4°C with primary anti-KS antibodies (provided by Dr. Frances Van Dolah, College of Charleston, USA) (1:100 dilution in blocking solution), primary antibody solution was removed, followed by 3 x 5-min PBS washes. Cells were then incubated with Alexa Fluor 488 (Abcam Cat #ab150077) secondary antibody (1h at RT in a 1:100 dilution with blocking solution) ending with several PBS washes. Cells were visualized using a Zeiss Axio-Observer Z1 LSM780 confocal microscope under a Plan-APOCHROMAT 63X/1.4 oil DIC objective lens. Primary antibodies were omitted for negative controls. ImageJ was used to analyzed Z-stacks profiles (Schindelin *et al.*, 2012).

2.3 Results

2.3.1 Syntenic and phylogenetic analyses of ketosynthase and acyltransferase domains

In order to understand molecular evolution and diversification of PKS and FAS, an extensive search for *PKS* (*KS* and *AT*) and *FAS* (*FabB-KASI*, *FabF-KASII* and *FabD*) genes within three Symbiodiniaceae genomes was conducted since these domains are conserved (Kroken *et al.*, 2003). The sequences were integrated into a dataset of well-characterized sequences from multiple taxa and subjected to phylogenetic analysis. Majority KS domains clustered according to their domain organization types under a reliable node (Bayesian Inference posterior probability, 0.79 and maximum likelihood bootstrap support, 99%) (Figure 2.1).

Recently, contigs encoding multiple PKS domains were reported in the dinoflagellates, *Gambierdiscus excentricus* and *Gambierdiscus polynesiensis* (Kohli *et al.*, 2017). The dataset also included those sequences and they clustered into three dinoflagellate groups (Dinoflagellate PKS I, II and III clades; blue highlighted inset of Figure 2.1). 25 KS sequences each from clades A3, B1 and C were confirmed. The present analysis showed only one gene model, B1030341.t1, to be associated with Type II fatty acid synthesis (FabF-KASII) and one gene model, B1027279.t1, in the FabB-KASI group. There is a clear separation between Type I PKS / FAS and Type II FAS, an observation in agreement to that reported by Kohli *et al.* (2016).

Additionally, the present analysis exposes the expanded nature of KS genes into nine PKS groups (Dinoflagellate PKS I-III and Symbiodiniaceae PKS I-VI) associated with either multi- or monofunctional domains (Figure 2.1). Interestingly, one clade (Dinoflagellate PKS-I) was found to be closely related to cyanobacterial KS sequences. The GC profile of PKS-I clade scaffolds from clade C showed some regions of higher GC content (45-46.5%), in comparison to the average genomic GC content of 43.0%, suggestive of gene transfer event

(Appendix Figure A). cTP (chloroplast transit peptide) signal was detected in ~3% (3/83) of the sequences while 12% (10/83) of sequences contained mitochondrial targeting peptide (mTP) or secretory signal each (Figure 2.1).

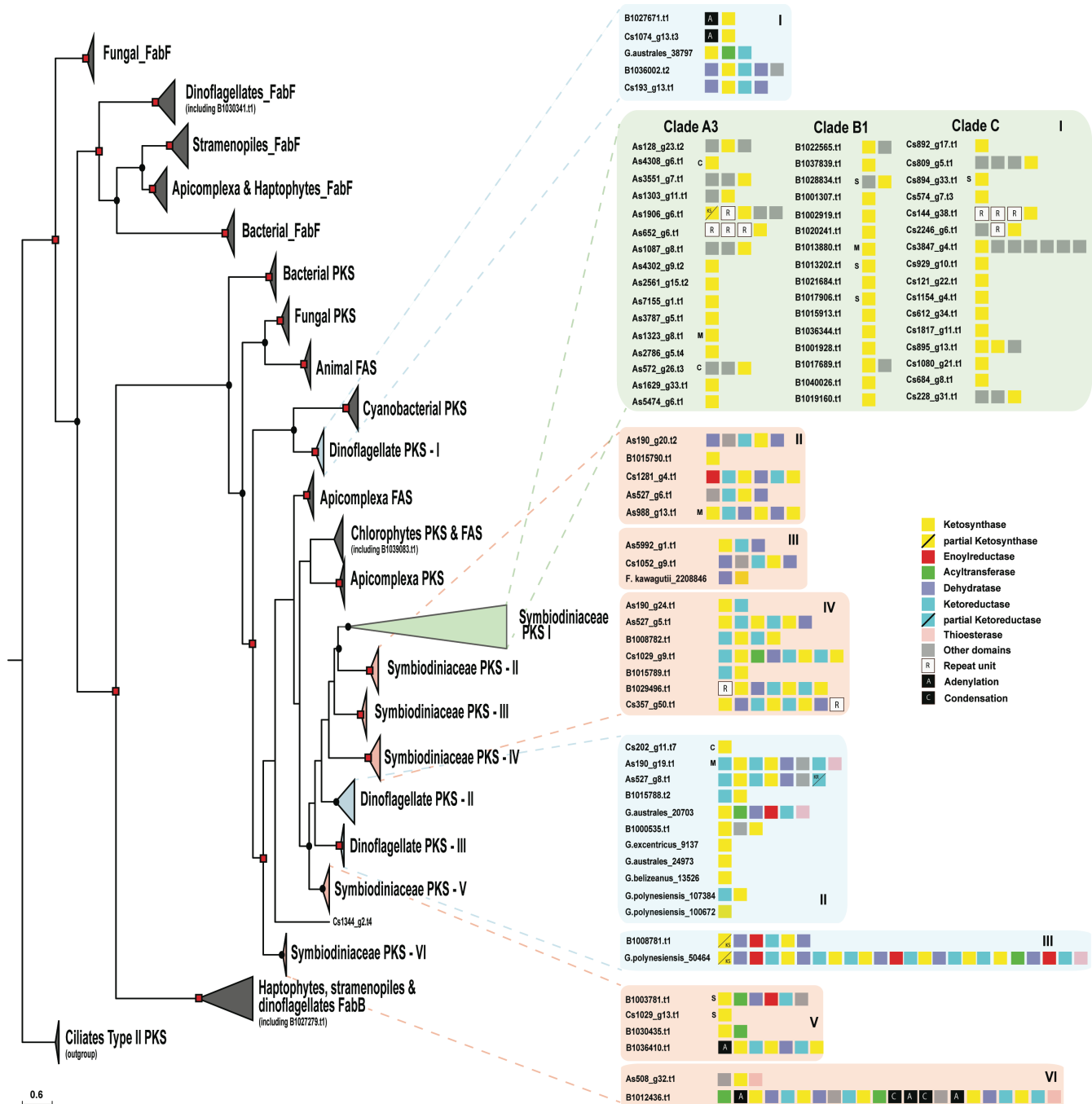


Figure 2.1 | Phylogenetic analysis of ketosynthase (KS) domains of eukaryotic and prokaryotic polyketide and fatty acid synthases. Analysis of ketosynthase, FabB-KASI, and FabF-KASII domains displays extensive diversification of these domains into nine groups. Posterior probabilities generated by Bayesian inference are indicated by dots (0.70-0.89) and squares (0.9-1.0). M, S, and C denote mitochondria, secretory, and chloroplasts signal peptide, respectively

An unusual feature among the three genomes is the high number (26) of *trans*-AT genes in contrast to *cis*-AT (4). A phylogenetic tree of the AT domain consisted of two main nodes, *cis*-AT and *trans*-AT (Bayesian Inference posterior probability, 1.00 and maximum likelihood probability, 81%) (Figure 2.2), deviating from the classical substrate-based clustering (Khayatt *et al.*, 2013). Alignment of the *trans*-AT motif revealed a deviation from the usual GHSxG conserved motif to GLSxG where x can be any residue; thus, a change from a basic amino acid (histidine) to an aliphatic one (leucine) while *cis*-AT maintained their GHSxG motif. The implication of His→Leu remains to be investigated (Figure 2.2). Use of the HMMs by Khayatt *et al.* (2013) did not suggest any clear distinction regarding which substrates are being incorporated into biosynthetic pathways. However, I-TASSER predicted that most *Symbiodinium* AT sequences pertain to the family of malonyl-CoA ACP transferase. Downstream of the active site serine, a motif (YASH or HAFH) is involved in the choice of either methylmalonyl-CoA or malonyl-CoA, respectively (Tang *et al.*, 2006). The motif, GAFH, present in most *Symbiodinium* sequences reflects the prediction of I-TASSER. ~9 % (3/33) of AT gene models contained the cTP or mTP signals (Figure 2.2).

Comparative visualization of *PKS*-containing scaffolds from the three genomes showed extensive duplication events in the three clades between genes associated with polyketide biosynthetic clusters (Figure 2.3a). Genomic synteny was observed between clades B1 and A3 (8 syntenic blocks), clades B1 and C (10 syntenic blocks), and clades A3 and C (7 syntenic blocks) (Figure 2.3b-d), respectively while only four *PKS*-containing gene clusters were found to be shared among all the three clades (green boxes in Figure 2.3b-d). The observed rearrangements within the syntenic scaffolds included mainly deletions. Transposons were found on scaffolds carrying *PKS*- and *NRPS*-encoding genes, suggesting that these genes can be influenced by transposable elements. 47% (52/110) of *PKS*- and 34% (14/41) *NRPS*-containing scaffolds possessed LTR signatures (Appendix Table B). Taken together, these

results indicate that *PKS* genes have diversified in each *Symbiodinium* clade by several evolutionary processes.

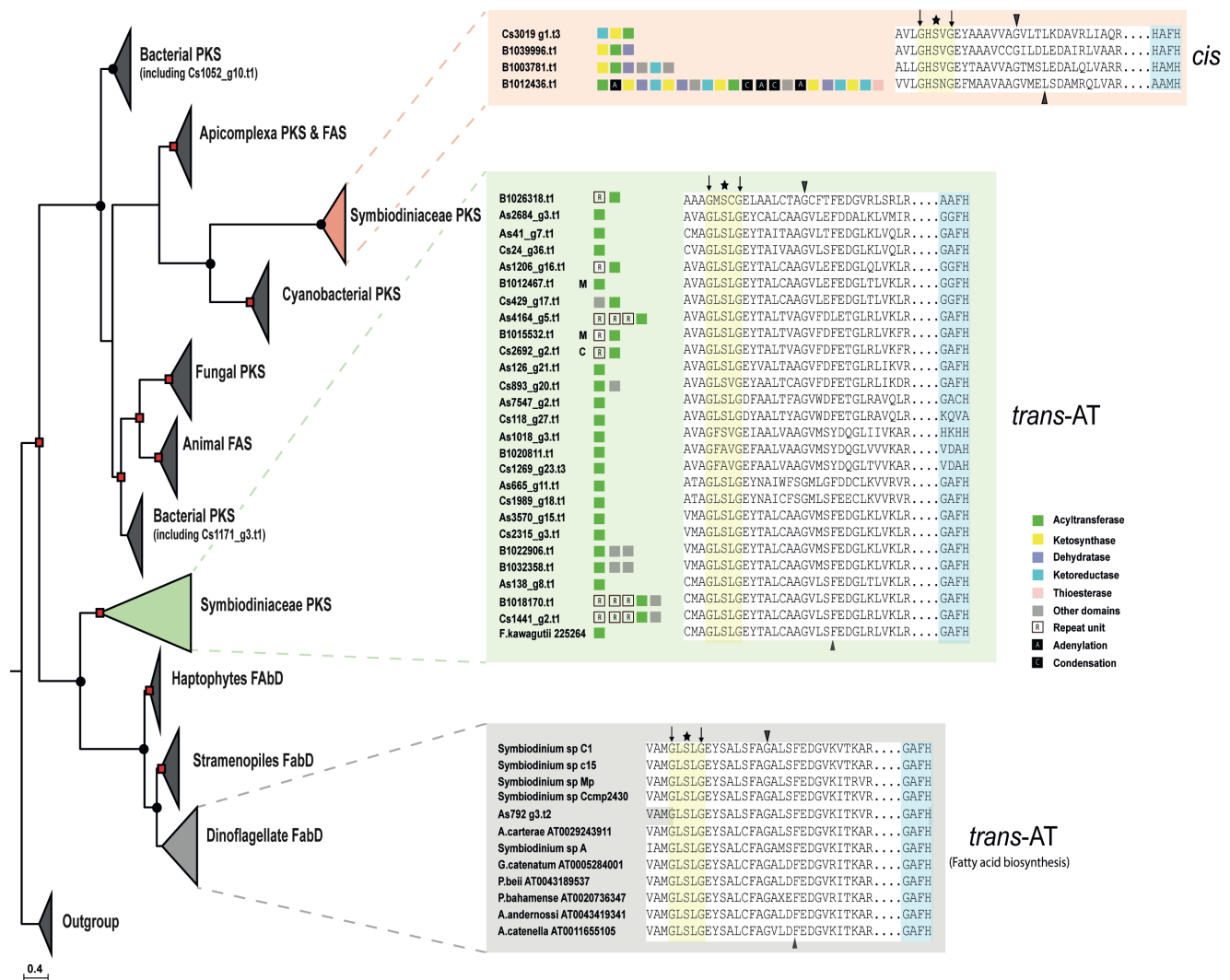


Figure 2.2 | Phylogenetic analysis of acyltransferase (AT) domain of eukaryotic and prokaryotic polyketide and fatty acid synthases. A clear demarcation between *cis*- and *trans*-AT is detectable. Bayesian inference posterior probability are shown by dots (0.70-0.89) and square (0.9-1.0). Black triangles show conserved residues characteristic to specific substrate groups, asterisk indicates active site residue, and black arrows indicate conserved residues used by HMM (Khayatt *et al.*, 2013). C, M and S depict chloroplast, mitochondria, and secretory signal peptide, respectively.

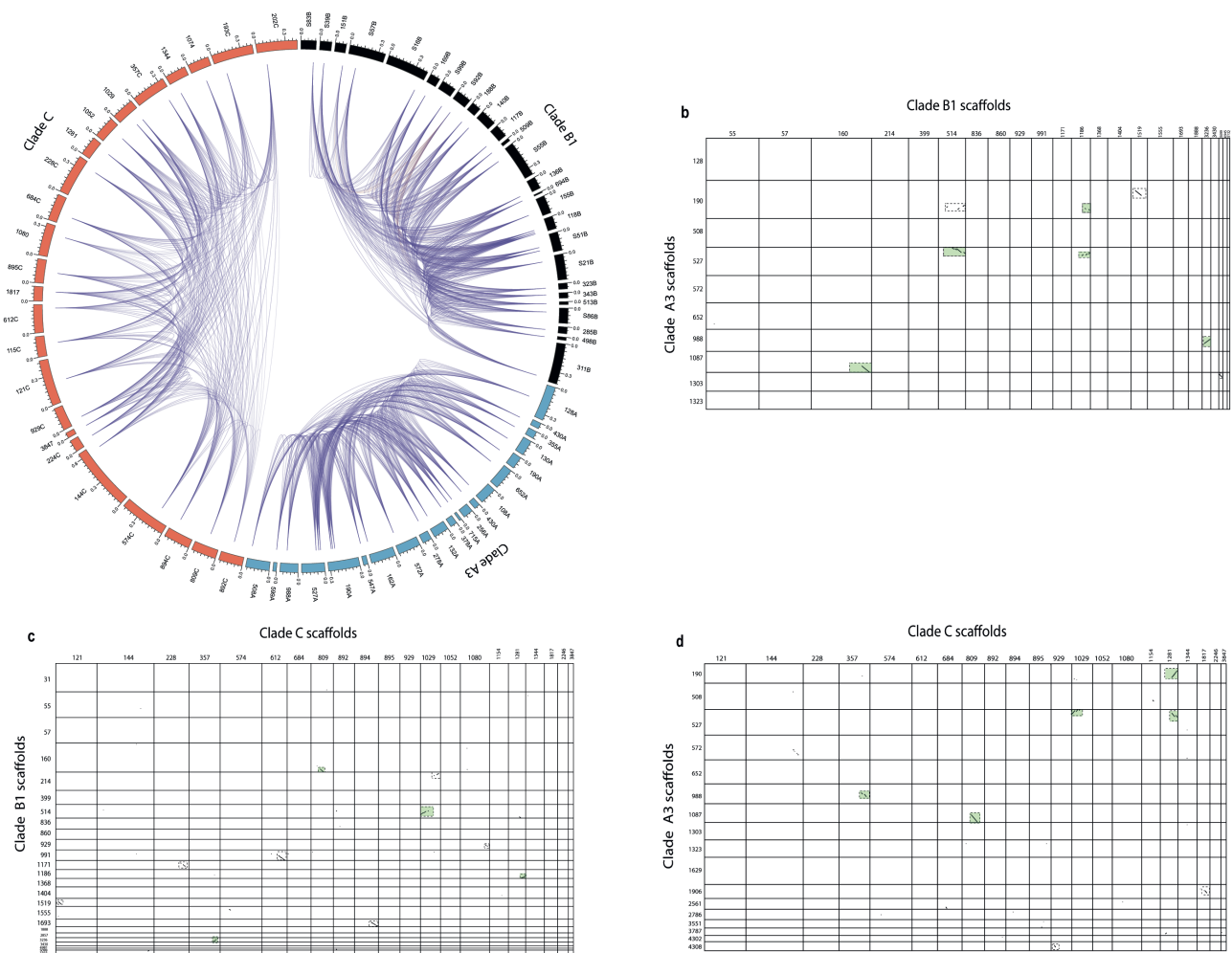


Figure 2.3 | Pathway duplication and conservation within and across *Symbiodiniaceae*. (a) Plot showing duplicate gene distribution within *PKS*-containing scaffolds of three *Symbiodiniaceae* genomes. Colored sections (black = clade B1, orange = clade C, blue = clade A3) represent scaffolds studied in Fig. 1. A link represents a possible duplication event between two domains. (b) Synteny plot of clade A3 and B1 *PKS*-containing scaffolds. (c) Synteny plot of clade B1 and C *PKS*-containing scaffolds. (d) Synteny plot of clade A3 and C *PKS*-containing scaffolds. Dotted boxes highlight regions of significant homology between genomes. Green colored dotted boxes show common regions shared among the three genomes.

2.3.2 Phylogenetic analysis of adenylation and condensation domain subtypes ($^L\text{C}_L$, $^D\text{C}_L$, Cyc and dual E/C) in NRPS proteins

To get a better understanding of freestanding A-domains identified in Symbiodiniaceae genomes, as to whether they obey the same non-ribosomal code of traditional NRPS systems (Stachelhaus *et al.*, 1999), a phylogenetic analysis involving 117 adenylation sequences from several taxa was performed. One significant result was that a freestanding A-domain from Symbiodiniaceae falls into three major groups that utilize tryptophan, glycine, and phenylalanine as substrates, respectively (three highlighted clades in Figure 2.4a). On the contrary, other proteins with di- or multi-domains demonstrated affinity for various substrates. Phylogenetic analysis of condensation domains was directed by functional categories of C-domains instead of species phylogeny or substrate specificity alone. Four specific functional categories were clearly supported, namely (1) ordinary C-domains, that are composed of $^L\text{C}_L$ and $^D\text{C}_L$, (2) heterocyclization (Cyc) domains, (3) dual E/C domains and (4) starter domains, which are found on initiation modules (Figure 2.4b). NaPDOS classification showed that Symbiodiniaceae are rich in $^L\text{C}_L$ subtypes, which catalyze the condensation of two L-amino acids. Both catalysts possess a conserved His-motif in their active sites with a consensus sequence of HHxxxDG, where x can be any residue. This survey revealed the existence of six condensation domains with the consensus motif being maintained, except for G being substituted with L and N in B1036245.t1 and Cs535_g6.t1, respectively. This analysis also confirms the close relationship between $^L\text{C}_L$ and starter C domains and between dual E/C and $^D\text{C}_L$ domains, as previously reported in bacterial genomes, adding reliability of this analysis (Rausch *et al.*, 2007). These results show that NRPS genes are specific for certain amino acids, thus contributing to a degree of chemical diversity in non-ribosomal peptide biosynthesis.

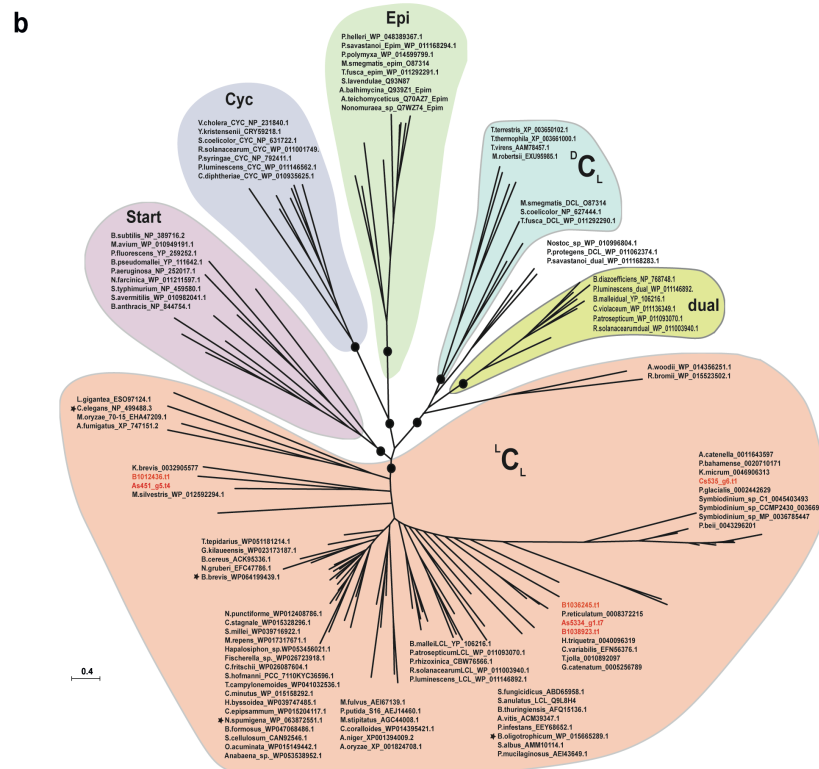
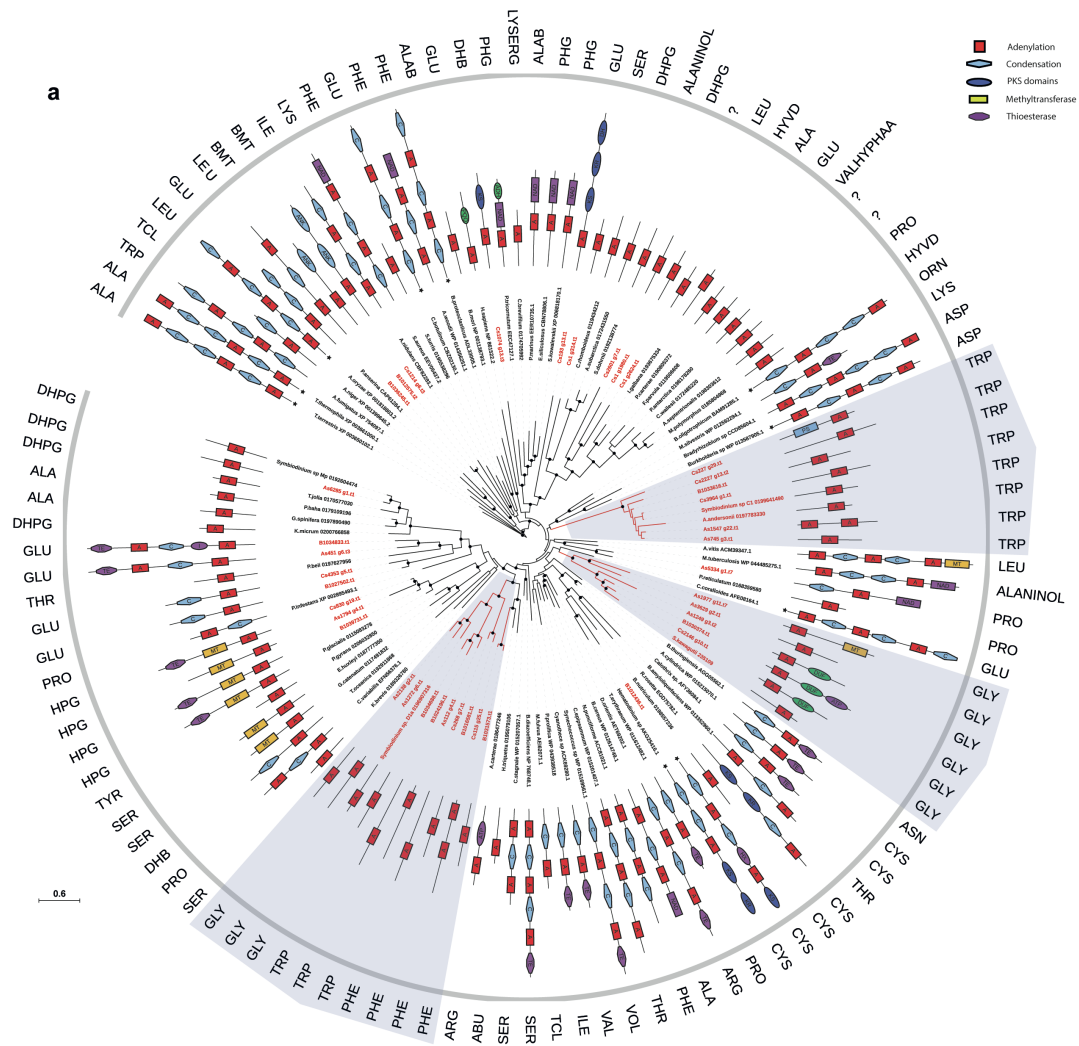


Figure 2.4 | Phylogenetic comparison of adenylation (A) and condensation (C) domains of prokaryotic and eukaryotic NRPS. A posterior probability ≥ 0.70 generated by Bayesian inference is indicated by dots. **(a)** Analysis of adenylation domains shows specificity of monofunctional domains from Symbiodiniaceae toward glycine, tryptophan, and phenylalanine (boxed by blue). **(b)** Condensation domains from Symbiodiniaceae belong to the $^L C_L$ type (shown in red).

2.3.3 Identification of metabolites and biosynthetic gene clusters from *Symbiodinium* genomes

Based on high-resolution mass data as summarized in Beedessee *et al.* (2015), polyols were identified. From MS spectra, doubly charged ions (negative ions) were searched for the larger polyols (>2600 Da). The presence of zooxanthellatoxin-B (ZT-B) with an m/z of 1414.74 for the $[M-2H]^{2-}$ was detected in Sample A3 showed (Appendix Figure C). Only zooxanthellamide D (ZAD-D) was identified from sample B1 with extracted ions at m/z 1050.57 for the $[M+H]^+$ (Appendix Figure D). Similar LC-MS profiles were noticeable for sample B1 and C, with identical unknown SCCs within the range of 2,600-2,850 Da (Appendix Figure D). The antiSMASH analysis on Symbiodiniaceae genomes matched four PKS-NRPS clusters to reported biosynthetic gene clusters, with similarities ranging between 25-46% (Figure 2.5a). A biosynthetic gene cluster with similarity to ajudazol and phenalamide biosynthesis was identified in clade A3 while a second phenalamide biosynthetic cluster was detected in clade B1. An example of module duplication in one scaffold, as well as between modules of different scaffolds can be seen in Figure 2.5b. Immunolocalization indicated that KS proteins were detected in only reticulate chloroplasts of clade C (Appendix Figure E), although KS proteins can be localized to other organelles as have been reported in *Karenia brevis* (Monroe *et al.*, 2010).

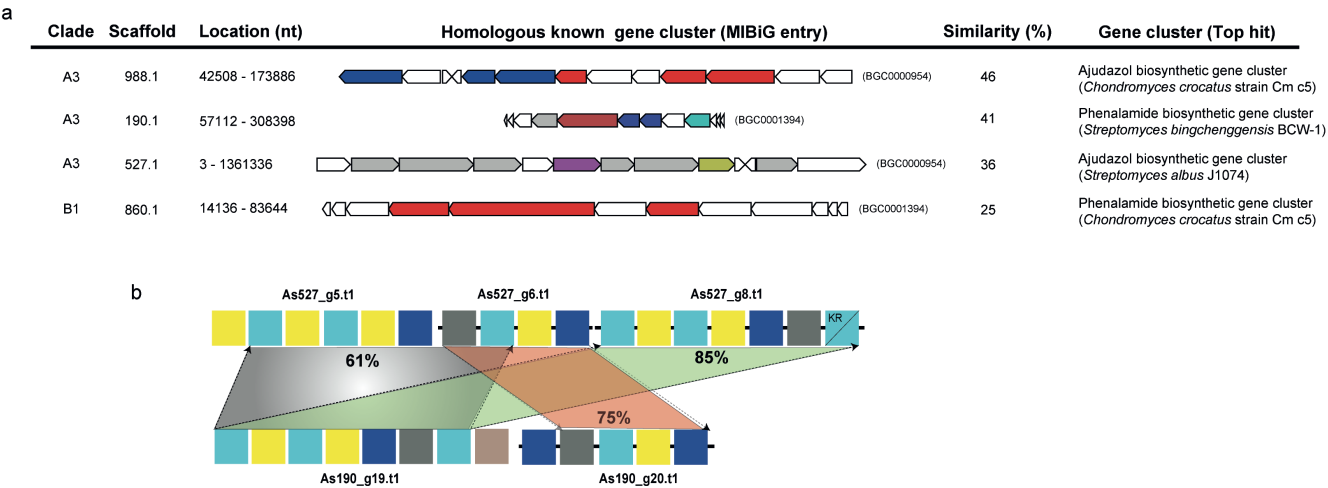


Figure 2.5 | Multifunctional *PKS* genes in Symbiodiniaceae. (a) Table showing gene clusters and similarities of different scaffolds from Symbiodiniaceae obtained using antiSMASH. Details of each gene cluster can be obtained using the MIBiG (Minimum Information about a Biosynthetic Gene cluster) entry number and is accessible at <https://mibig.secondarymetabolites.org/repository.html> (b) An example of module duplication between two scaffolds (527.1 and 190.1 of clade A3). Numbers signify the percentage of identity shared between sequences.

2.4 Discussion

2.4.1 Evolution of modularity within three Symbiodiniaceae genomes

The genomic analysis reveals the expanded genetic diversity of metabolite-producing capacity in Symbiodiniaceae dinoflagellates. The polyketide biosynthesis machinery increases its functional and genetic modularity by modifications through combinatorial events assisted by gene duplication, horizontal gene transfer (HGT), and recombination (Thattai *et al.*, 2007). The presence of many monofunctional KS or AT domains within these genomes raises questions about the evolution of modularity. The present analysis shows that module as well as domain duplications prove to be an important evolutionary mechanism toward modularity

(Figure 2.5b). Large numbers of repeats are scattered within dinoflagellate genomes, with frequent recombination events, and expansion of genes due to duplication (Shoguchi *et al.*, 2013; Lin *et al.*, 2015; Aranda *et al.*, 2016). These characteristics might have contributed to decomposition of Type I multifunctional PKS clusters, an event involving shuffling of domains and modules previously observed (Jenke-Kodama *et al.*, 2005). There is now an increasing number of reports on multifunctional PKS domains in several dinoflagellates, demonstrating that multifunctionality coevolves with monofunctional domains (Beedessee *et al.*, 2015; Kohli *et al.*, 2017; Van Dolah *et al.*, 2017). The data show that monofunctional PKSs are closely linked to multifunctional PKS (Figure 2.1), but it is unclear whether fusion of monofunctional PKS domains directed multifunctionality or *vice versa*. Another important contributor in the expansion of PKS and NRPS may have been retrotransposons because 34-47% of the scaffolds are predicted to contain LTR signatures (Appendix B). Retrogenes have been known to account for >20% of all genes in *Symbiodinium* clades (Song *et al.*, 2017). For retroposition events in *Oxyrrhis marina*, Ty1/copia LTR retrotransposon has been proposed as a likely candidate (Lee *et al.*, 2014).

Another significant event contributing to gene innovation is HGT, with recent evidence for association of HGT with several biological processes including metabolism (Wisecaver *et al.*, 2013). HGT is assumed to contribute to genome innovation in *Symbiodinium kawagutii* (Lin *et al.*, 2015). PKS gene transfer has been proposed in *Karenia brevis* (Lopez-Legentil *et al.*, 2010). On the other hand, gene duplication has contributed to the expansion of the light-harvesting complex (LHC) gene family in *Symbiodinium minutum* B1 (Maruyama *et al.*, 2015). Interestingly, monofunctional domains of either PKS and NRPS, are often merged with repeat units like HEAT (huntingtin, elongation factor 3, α subunit of protein phosphatase 2A and TOR1), ankyrin and pentatricopeptide (PPR) repeats. Ankyrin repeat family is a major protein family in the dinoflagellate *Breviolum minutum*, facilitating protein-protein interactions while

HEAT repeats play a role in protein transport (Bennett *et al.*, 2001; Mosavi *et al.*, 2004; Cook *et al.*, 2007). PPR proteins, on the other hand, are nuclear-encoded, but target plastids and mitochondria, where they participate in RNA processing and editing (Colcombet *et al.*, 2013; Fujii *et al.*, 2011; Nakamura *et al.*, 2012).

2.4.2 Evolution of polyketide biosynthesis

It was suggested that fatty acid synthesis could be carried out by Type II FAS in dinoflagellates (Kohli *et al.* 2016), based on a strong distinction between genes involved in fatty acid and polyketide biosynthesis. Only a single orthologue, B1030341.t1, was found to be associated with Type II fatty acid synthesis (*FabF-KASII*). The data show that PKS domains have undergone widespread diversification in all the three Symbiodiniaceae genomes. A conceivable explanation for this expansion might be their participation in novel functions, as suggested by the fact that ~ 15% of KS and ~9% of AT proteins have a target signal peptide, directed towards different organelles. In *Durinskia baltica*, a FAS-like multi-domain polyketide synthase has been found to associate with fatty acid biosynthesis (Hehenberger *et al.*, 2016). Recent transcriptomic assessment of the dinoflagellate *Hematodinium* sp. showed only Type I FAS (Gornik *et al.*, 2015), while another study on *Gambierdiscus* spp. revealed a distinct Type II FAS system together with single KS domains (Kohli *et al.*, 2017), signifying possible distinctiveness of these pathways to specific dinoflagellates. Although transcriptome data is not exclusive, both Type I and Type II FAS systems can co-exist, as in *Toxoplasma* (Seeber *et al.*, 2010). In some taxa only, cytosolic Type I are present, as in *Cryptosporidium parvum*, while in others only the plastid Type II, as in *Plasmodium falciparum* (Zhu *et al.*, 2004). Clearly, apicomplexan and dinoflagellate ancestors possessed both systems.

AT domains of *trans*-ATs are specific for malonyl-CoA while *cis*-AT are specific display towards various extender units (e.g. hydroxymalonyl-ACP, methylmalonyl-CoA,

methoxymalonyl-ACP, etc). Stand-alone AT proteins have been described in several PKSs with modules devoid of AT domains and these proteins provide malonyl as building blocks for the ACP domains of PKS (Piel, 2002; Cheng *et al.*, 2003). The present analysis shows that these stand-alone *trans*-AT proteins are the main AT types in Symbiodiniaceae genomes, establishing a major group that may undergo independent evolution in contrast to canonical *cis*-AT domains. The presence of such *cis*- and *trans*-AT clades has been described in bacteria and has been taken as a proof of independent evolution (Piel *et al.*, 2004). *cis*-AT PKS of bacterial origin have evolved mainly via horizontal/vertical acquisition and module duplication of entire assembly lines (Jenke-Kodama *et al.*, 2005) while *trans*-AT appears to recombine and lead to novel gene clusters in a mosaic-like fashion (Nguyen *et al.*, 2008), as observed globally for AT in Symbiodiniaceae genomes (Figure 2.2). Noniterative PKSs in algae depend largely on *trans*-AT and are features of multimodular PKS (Shelest *et al.*, 2015).

2.4.3 Evolution of non-ribosomal peptide biosynthesis

There are a few studies reporting NRPS from dinoflagellate transcriptomes (Salcedo *et al.*, 2012; Cooper *et al.*, 2016). The present study is the first study that aimed at looking at the affinities and role of adenylation and condensation domains in dinoflagellates. In contrast to Type I PKS, NRPSs were fewer in number within the three Symbiodiniaceae genomes. NRPS genes are known to be rare in eukaryotic microalgae (Shelest *et al.*, 2015). A stretch of amino acids within the A domain catalytic pocket governs recognition and activation of an amino acid substrate. Therefore, any point mutations within this segment can significantly change the specificity of the A domain. Incorporation of non-polar and polar amino acids during peptide synthesis is favored by a mono-modular adenylation domain (Figure 2.4a). Mono/bi-modular NRPSs present in fungal species contain a conserved domain organization that is important for its function (Bushley *et al.*, 2010). Solitary A or A-T domains can interact with other NRPS

proteins to accomplish biosynthesis by successful activation and transfer of the substrate to the condensation domain in the same or different NRPS (Mootz *et al.*, 2002). NRPSs are primarily modular enzymes with multiple domains, although, nonmodular enzymes have been reported in fungal subfamilies (Bushley *et al.*, 2010). Freestanding A, C, or PCP proteins act *in trans* to form NRPS modules and may be involved in natural product biosynthesis, devoid of the peptide moiety (Donadío *et al.*, 2007).

2.4.4 Secondary metabolic pathways are conserved in the family Symbiodiniaceae

Symbiodiniaceae lineages diversified from the ancestral clade A ~160 MYA, at the beginning of the Eocene (LaJeunesse *et al.*, 2018) and adjusted to different niches, playing critical functions in reef ecosystems as well as serving as endosymbionts of different phyla (Gordon & Leggat, 2010). Symbiodiniaceae genomes allow us to compare biosynthetic pathways, providing insights on the organization and contribution of pathways to ecological success. Several gene clusters are conserved between *Symbiodinium tridacnidorum* (clade A3), *Breviolum minutum* (clade B1), and *Cladocopium* sp. (clade C) (Figure 2.3b-d), despite their different divergence time (LaJeunesse *et al.*, 2018). The importance of conserved phosphatidylinositol signaling pathways in four Symbiodiniaceae towards symbiotic interactions have been reported (Rosic *et al.*, 2015). Mass spectrometry analysis showed that *Symbiodinium tridacnidorum* (clade A3) and *Breviolum minutum* (clade B1) produce unique polyketides, supporting the clade-metabolite hypothesis (Fukatsu *et al.*, 2007). Different temperatures and light regimes can influence the metabolite profiles of different Symbiodiniaceae species (Klueter *et al.*, 2015). Interestingly, metabolomic similarity was detected only between *Breviolum minutum* and *Cladocopium* sp. It is difficult to link specific metabolites to specific pathways, but this result suggest that new pathways must have evolved in the common ancestor of *Breviolum minutum* and *Cladocopium* sp. to generate a joint set of

metabolites, irrespective of their environment and hosts. Biological systems control their biochemical and cellular activities when subjected to environmental changes (Hannah *et al.*, 2010).

Taken together, these results show how Symbiodiniaceae genomes encode the necessary enzymes (PKSs and NRPSs) with broad substrate tolerance as an effective way of producing chemical diversity. The “Screening hypothesis” proposes that organisms that synthesize many chemicals, have more chances of improved fitness because greater chemical diversity, more the chance of producing metabolites with unique traits, as shown by zooxanthellatoxins and zooxanthellamides (Jones & Firn, 1991). But this does not answer as to why only a few major pathways are conserved among the Symbiodiniaceae. It might be favorable for organisms to extend existing pathways to create chemical diversity than to originate entirely novel pathways (Firn & Jones, 2003).

3 Genome analysis of *Amphidinium gibbosum*

3.1 Introduction

Dinoflagellate biology defies many genetic and cellular features commonly attributed to eukaryotes lifestyle. Presence of unusual upstream promoters, non-canonical splice site, 5-hydroxymethyluracil in nuclear genome DNA and higher rate of translational regulation are clear deviations from most other eukaryotes (Shoguchi *et al.*, 2013; Aranda *et al.*, 2016). Dinoflagellate genomics undoubtedly will enrich our basic understanding of the functionality and evolution of eukaryotes genomes. In recently years, there has been a growing number of dinoflagellates transcriptomes (Erdner *et al.*, 2006; Moustafa *et al.*, 2010; Bayer *et al.*, 2012; Keeling *et al.*, 2014); however, only a reference genome would provide insights into how gene numbers, and their organization and position, which are important for developing any transgenic approach (Shoguchi *et al.*, 2013; Aranda *et al.*, 2016). The unusually large genomes of dinoflagellates have been the major limitation and it is only in the past five years that draft genomes of the smallest dinoflagellates of the family Symbiodiniaceae has been achieved (Lin *et al.*, 2015; Aranda *et al.*, 2016; Shoguchi *et al.*, 2018; Liu *et al.*, 2018). To increase our understanding of dinoflagellate genomics, new genomes are needed and keeping in mind the special feature of dinoflagellates, *Amphidinium* genus proves to be an ideal candidate with a reasonable genome size of ~ 5.9 Gb (LaJeunesse, 2005).

Amphidinium species are most abundant in benthic ecosystems, and *Amphidinium carterae* is easily grown and accessible from culture collections, making an ideal dinoflagellate model (Murray & Patterson, 2002; Lee, 2003). *Amphidinium* spp. were used in the understanding of the peridinin-chlorophyll-protein light-harvesting antenna complex (Hofmann, 1996), unique mitochondrial genome (Nash *et al.*, 2007), and the first reported genetic transformation of a dinoflagellate (Ten Lohuis & Miller, 1998). Some species of

Amphidinium have been reported to be associated with pelagic harmful algal blooms (HABs) (Lee, 2003; Baig *et al.*, 2006; Gárate-Lizárraga, 2012). Dense bloom of *Amphidinium carterae* (Genotype 2) was reported in a shallow coastal lagoon in south-eastern Australia (Murray *et al.*, 2015). Symbiosis between *Amphidinium klebsii* (Kofoid & Swezy, 1921) and the acoel *Amphiscolops langerhansii* is essential for the survival of the flatworm (Taylor, 1971). Similar symbiotic relationship was reported in *Amphidinium klebsii*-like algae and *Amphiscolops* sp. (Lopes, 1994).

The genus *Amphidinium* Claparède et Lachmann 1859 is among the largest and most diverse marine dinoflagellates, with approximately 120 species (Murray & Patterson, 2002). This genus has recently gained interest because some species produce ichthyotoxic substances. *Amphidinium* is a member athecate dinoflagellate similar to Gymnodiniaceae, as species are devoid of cellulosic thecal plates. However, molecular data did not support the monophyly of the Gymnodiniaceae (Daugbjerg, 2000) nor a close relationship between *Amphidinium* and other genera of Gymnodiniaceae (Murray *et al.*, 2004; Jørgensen *et al.*, 2004; Murray *et al.*, 2005; Zhang *et al.*, 2007a). There is a high level of genetic diversity (37%) in the sequences of D1-D6 regions of the large subunit ribosomal rRNA (LSU) within taxa of this genus, indicating either a high evolutionary rate in the rRNA genes of members of this genus when compared to other dinoflagellates or they represent a diverse ancient group (Murray *et al.*, 2012). rRNA studies show that *Amphidinium* may be a relatively early evolving lineage of dinoflagellates (Murray *et al.*, 2004; Jørgensen *et al.*, 2004; Murray *et al.*, 2005; Zhang *et al.*, 2007a). This genus was redefined using stricter morphological criteria and now includes approximately 20 known species (Murray, 2003; Karafas *et al.*, 2017).

The genus *Amphidinium* possesses an intricate secondary metabolism that generate several macrolides and polyketides, unique in structure and cytotoxicity activity (Kobayashi & Kubota, 2007). This group of cytotoxic macrolides, amphidinolides, have unusual odd-

numbered lactone rings, a feature observed from more than half of the isolated compounds. Several studies have attempted to isolate PKS genes from *Amphidinium* species to understand amphidinolide biosynthesis but has remained inconclusive due to absence of complete KS sequences (Kubota *et al.*, 2006; Murray *et al.*, 2012). One such macrolide, Amphidinolide H induces multinucleated cells by disrupting actin organization in cells and can be exploited as potential anticancer drug lead (Chakraborty & Das, 2001). However, the mode of action of several other amphidinolides remains to be explored.

To understand how dinoflagellates evolve innovation in secondary metabolism, I sequenced the genome of the basal *Amphidinium gibossum* and survey genes that generate such structural and biological uniqueness. I further examine the mechanisms that can result in biosynthesis of small and large complex metabolites in the family Symbiodiniaceae and *A. gibossum*.

3.2 Materials and Methods

3.2.1 Biological sample and genome size estimation

Amphidinium gibossum was originally isolated by Dr. Takaaki Kubota (Showa Pharmaceutical University, Tokyo, Japan) from the inner cell of acoel flatworms, *Amphiscolops* species found near Ishigaki Island. The culture was maintained in artificial seawater containing 1X Guillard's (F/2) marine-water enrichment solution and antibiotic-antimycotic mix in a 25°C incubator under a 12:12 light and dark cycle. Subculture was performed ~ every 4 weeks with fresh medium and handled strictly aseptically.

3.2.2 Genome size estimation

For an estimation of the *A. gibossum* genome size, nuclear DNA was measured using fluorescence-activated cell sorting (FACS) and the frog *Xenopus laevis* as an internal control of known genome size. Nuclei extraction and staining for *A. gibossum* and the internal control were performed using the Partec CyStainPI absolute T kit (Partec #05-5023) following the manufacturer's protocol and the fluorescence signals were measured with a BD Accuri C6 cell analyzer (BD Bioscience). The reported measurement for *A. gibossum* reflects the 1C genome content as *Amphidinium* is reported to be haploid in culture. K-mer analysis was performed using Jellyfish (v2.1.3) (Marcais and Kingsford, 2011) using K values ranging from 75-85 and resulting histograms were visualized using GenomeScope (Vurture *et al.*, 2017) to survey the genome size and repeat content.

3.2.3 DNA sample preparation and sequencing

Cells were centrifuged at 3000 g for 10 minutes and washed using TEN buffer (100 mM Tris-Cl pH 8, 100 mM EDTA pH 8, 1.5 M NaCl, 0.5 mg/mL proteinase K and 7% SDS) for 2 hours at 65°C so as to lyse bacteria. DNA was extracted using a modified protocol (Doyle, 1987). DNA was further cleaned using ethanol precipitation. DNA was fragmented and paired-end libraries with insert size of 620-820 bp were prepared. Libraries were quantified by qPCR and Bioanalyzer and sequenced using an Illumina Miseq using manufacturer's protocols. This generated ~ 10 Gb of 2 x 300 bp long paired-end data. The same library was further sequenced using Hiseq 2500 and generated ~586 Gb of 2 x 125 bp data. Reads were merged and then trimmed using Trimmomatic (v0.35) (Bolger, 2014) and quality-checked using FastQC (v0.11.4) (Andrews, 2010). Additionally, mate pair libraries were constructed using Nextera technology with 3-18 kb inserts selected using the Bluepippin and SageELF system. Mate pair libraries were sequenced using Hiseq 4000 generating ~200 Gb data. Raw mate paired reads

were filtered using NextClip (v1.31) (Legget *et al.*, 2014). Genome assembly was conducted using Platanus (v2.1.4) (Kajitani *et al.*, 2014). The assembled genome was then subjected to two rounds of scaffolding using the quality-controlled mate-pair reads using SSPACE (V3.0) (Boetzer *et al.*, 2011). Gaps in the scaffolds were filled using GapCloser (v1.12) (Luo, 2012). Scaffolds < 500 nucleotides were removed from the assembly.

3.2.4 Evaluation of genome assembly completeness and removal of bacterial and viral sequences

The scaffolded *Amphidinium* genome was checked for genome assembly completeness using BUSCO (Simão *et al.*, 2015), where the presence of 303 highly conserved eukaryotic genes (CEGs) was determined. Additionally, blast suite was used to recover the 458 CEGs from CEGMA (Parra *et al.*, 2007) against the *Amphidinium* genome so as to identify potential homologs at a cutoff value of $1e^{-5}$. BLASTN searches against several databases was conducted: draft and complete bacterial genomes (<ftp://ftp.ncbi.nih.gov/genomes/Bacteria/all.fna.tar.gz>, ftp://ftp.ncbi.nih.gov/genomes/Bacteria_DRAFT/) and viral genomes from NCBI and PhanToME (<ftp://ftp.ncbi.nih.gov/genomes/Viruses/all.fna.tar.gz>, <http://phantome.org>). A combination of cutoffs (total bit score >1000, e-value $\leq 10^{-20}$) was used to identify scaffolds with similarities to bacterial and viral sequences.

3.2.5 Transcriptome assembly for generating gene models

Cells were subjected to standard condition (12:12 light and dark cycle) after which RNA was extracted and cDNA library constructed using Truseq stranded mRNA sample preparation kit (Illumina). Libraries were quantified and validated by qPCR and a 2100 Agilent Bioanalyzer. The validated library was subsequently sequenced using two lanes of HiSeq 2500 (Illumina). Reads were trimmed using Trimmomatic (v0.35) (Bolger, 2014), and quality-checked using

FastQC (v0.11.4) and assembled *de novo* using Trinity v2.3.2 (Haas *et al.*, 2003). In order to confirm splice sites, the assembled transcriptome was mapped to the genome using GMAP (Wu and Watanabe, 2005). BLAT (Kent, 2002) was also used to confirm such splice sites and found to be less accurate than GMAP. The assembled transcriptome was found to have a BUSCO of 80.9 % and not significantly different from the stress transcriptome (Chapter 4).

3.2.6 cDNA construction, Iso-Seq sequencing and data processing

RNA was extracted using from several culture treatments and pooled using RNA PureLink reagent. High-quality RNAs were (RIN > 7.0) were used for cDNA synthesis using the Clontech SMARTer PCR cDNA kit. Size fractionation (0.7-2.5, 2.5-7 and > 7kb) was conducted using SageELF system (Sage Science, Beverly, MA, USA). Libraries were sequenced on the Pacific Biosciences RS II platform with the P6-P4 chemistry with 360 min movie lengths. A total of 16 SMRT cells were sequenced. Raw sequencing data were processed using the RS_Iso-Seq protocol. HQ and LQ reads were error corrected using proovread v2.14 using Illumina RNA-seq data obtained from 2 lanes. Reads were then merged and ‘cd-hit-est’ from CD-HIT v4.6 (Li & Godzik, 2006) was used to remove redundancy with parameters: -c 0.99 -G 0 -aL 0.00 -aS 0.99 -AS 30 -M 0 -d 0 -p 1 -T 24. Non-redundant transcripts were further processed with Cogent (<https://github.com/Magdoll/Cogent>).

3.2.7 Annotation of repetitive elements and gene models generation

In order to annotate TEs, *de novo* repeats within the genome were identified using an l-mer size of 17 bp by employing RepeatScout (Price *et al.*, 2005). A combined library was made that consisting of *de novo* repeats and known eukaryotic TEs from RepBase (January 2017 edition); this library was then used to locate and annotate repetitive elements in the assembled genome using RepeatMasker (Smit *et al.*, 2013). RNA-seq reads were mapped to a soft-masked

genome using STAR aligner (Dobin *et al.*, 2013) and used in the BRAKER2 pipeline (version 2) (Hoff *et al.*, 2016) using GeneMark-ES v3.32 (Lomsadze *et al.*, 2005) and Augustus v3.2.3 (Stanke *et al.*, 2003). Parameters generated were used and UTR prediction was performed again using Augustus v3.2.3 (Stanke *et al.*, 2003). To improve gene prediction accuracy, hints (intron and exon) were generated as additional evidence of gene structure and location by mapping both Illumina and Isoseq transcripts to the genome using GMAP (Wu & Watanable, 2005) and STAR (Dobin *et al.*, 2013). These hints were then used to perform a final gene prediction using a soft-masked genome using a modified version of Augustus v3.2.2, where the source code was changed to take into account the non-canonical exon-intron boundary (GA-AG). The final set of predicted proteins was annotated against UniProt (SwissProt and TrEMBL) (Magrane *et al.*, 2011) and PFAM (Punta *et al.*, 2012). Briefly, BLASTP searched for all protein models were undertaken against SwissProt and TrEMBL databases (October 2018 release).

3.2.8 Pfam and KEGG pathway analysis

Amino acid sequences of selected organisms were applied to Pfam (Punta *et al.*, 2012) domain search using HMMER v3.1b2 (Finn *et al.*, 2011) and hits larger than $1e-5$ were discarded. For KEGG pathway analysis, the online service on KEGG Automatic Server (KAAS) was used to assign each predicted gene to KEGG ortholog (bi-directional best hit method) and mapped orthologs to KEGG pathways.

3.2.9 Phylogenetic analysis of PKS and NRPS proteins

The dataset used in Beedessee *et al.* (2019) was repopulated with proteins sequences of ketosynthase, acyltransferase, adenylation and condensation from *A. gibossum* genome. In brief, active site residues of sequences were confirmed by Pfam (Punta *et al.*, 2012) and aligned using MUSCLE algorithm. They consisted of 244 KS sequences (225 aa), 104 AT sequences

(208 aa), 121 A-sequences (272 aa), and 111 C-sequences (253 aa), respectively (Edgar *et al.*, 2004). Bayesian inference and maximum likelihood analysis were performed as described in Beedessee *et al* (2019). Substrate specificity of *A. gibossum* AT sequences was generated using I-TASSER (Zhang, 2008). In order to determine the A-domain specificity and C-domain types, LSI-based A-domain predictor and NaPDos were used, respectively (Baranašić *et al.*, 2014; Ziemert *et al.*, 2012). PKS proteins subcellular localization was detected using ChloroP 1.1 and TargetP 1.1, and further confirmed with DeepLoc (Emanuelsson *et al.*, 1997; Emanuelsson *et al.*, 2007; Armenteros *et al.*, 2017).

3.2.10 PKS proteins immunolocalization

Cells were first fixed in 2% paraformaldehyde in seawater, washed three times with PBS and incubated in 50% methanol: PBS (5 mins). Cells were then deposited on poly-L-lysine coated coverslips, blocked with 5% normal goat serum for 1 hr, and subsequently incubated with primary anti-PKS antibodies (KS and KR) (provided by Dr. Frances Van Dolah, College of Charleston, USA) at 1:100 dilution overnight at 4°C. Cells were then incubated with Alexa fluor-488-conjugated secondary antibodies for 1 hr at room temperature. Coverslips were then mounted with Vectashield on glass slides and observed under a Zeiss Axio-Observer Z1 LSM 780 microscope. Data were collected using the ZEN software (version 14.0.8.201). For negative controls, cells were treated with PBS instead of the primary antibodies. Stacks were analysed using ImageJ (Schindelin *et al.*, 2012).

3.3 Results

3.3.1 Genomic features of *A. gibossum*

The draft genome of *A. gibossum* was assembled into a 7.0 Gb assembly; the *k*-mer analysis and FACS (Fluorescence-activated cell sorting) estimated a genome size of 6.3 Gb and 6.5 Gb, respectively (Figure 3.1b; Table 3.1). The scaffold N50 of the assembled genome is 166.5 kb, encoding 85139 genes, of which ~ 48% have matches in available databases (Appendix H). The completeness of the genome was assessed by using the 303 conserved BUSCO genes. This resulted in a low completeness BUSCO score of 27.4 % (83/303) (Appendix G). Such low score has been reported for dinoflagellates (Aranda *et al.*, 2016) since dinoflagellates evolutionary origin dates back to ~1.5-1.9 billion years (Nei *et al.*, 2001; Parfrey *et al.*, 2011). However, using the 458 conserved CEGMA genes, at least 73% of the homologs were recovered using several BLAST approaches (Appendix I).

Non-canonical splice site analysis confirmed the utilization of GC and GA (5' donor splice site) (Figure 3.1c). Gene orientation analysis showed that *A. gibossum* has a similar clustering of unidirectionally genes as other dinoflagellates of the Symbiodiniaceae family (Shoguchi *et al.*, 2013; Aranda *et al.*, 2015) (Figure 3.1d). 16 scaffolds with a combined length of ~11 Mbp were obtained as hits for bacterial and viral contaminants, which were removed from the assembly.

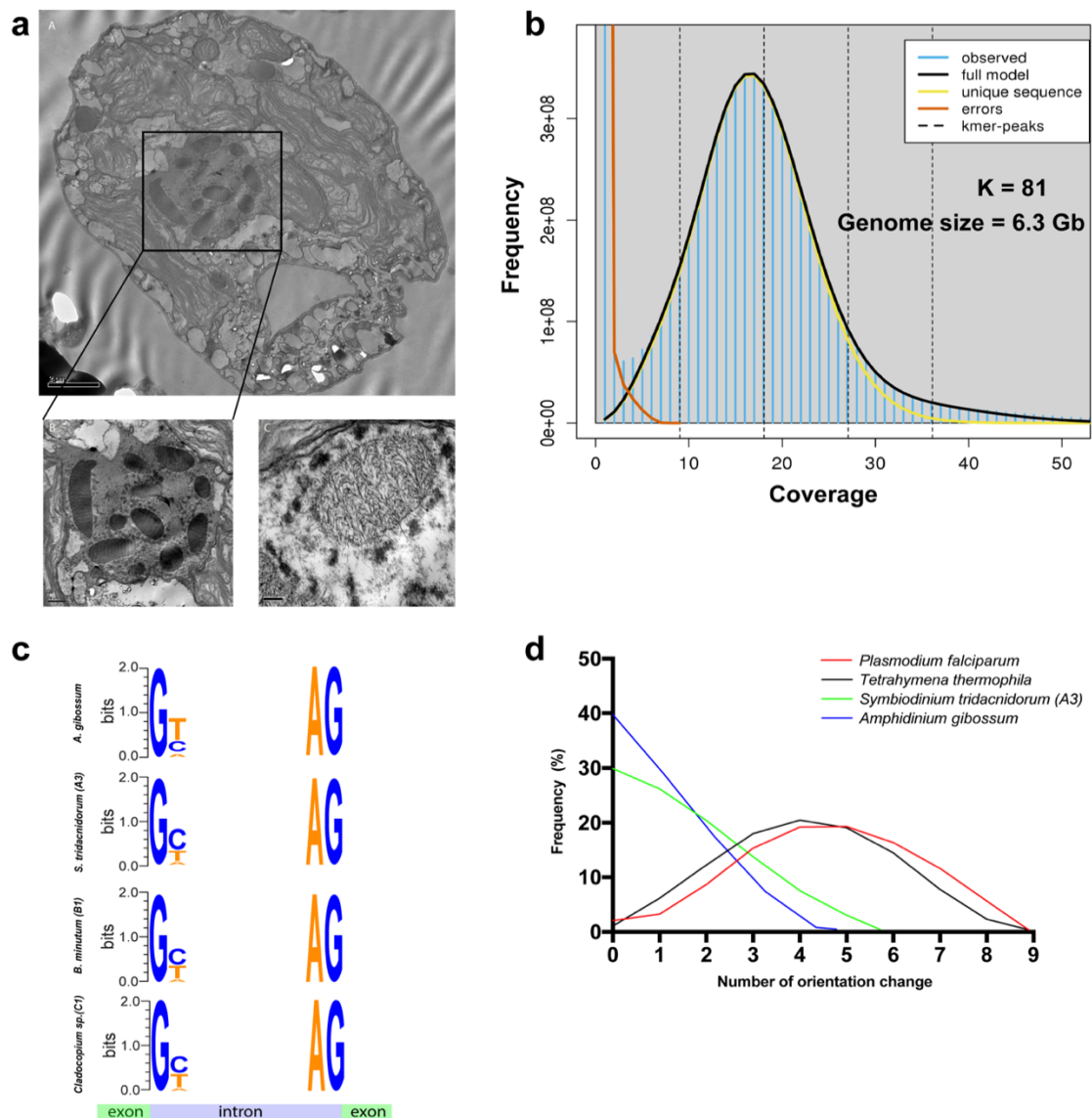


Figure 3.1 | General features of *A. gibossum*. (a) Transmission electron microscopy of *A. gibossum* with lower insert showing detailed region of the condensed chromosomes. (b) Genome estimation using Genomescope at $K = 81$. (c) Non-canonical splice sites in *A. gibossum* with comparison to other *Symbiodiniaceae* genomes. (d) Gene orientation changes in *A. gibossum* using a 9-gene sliding window and 9-gene step

Table 3.1: Genome statistics of *Amphidinium gibbosum* and other Symbiodinaceae

		<i>Amphidinium gibbosum</i>	<i>Breviolum minutum</i> (B1)*	<i>Symbiodinium tridacnidorum</i> (A3)#	<i>Cladocopium</i> sp. (C)#
	Total assembly length (bp)	7,034,147,423	615,520,517	766,659,703	704,779,698
	Scaffold (N50)	166.4K	126.2k	133.4k	248.9k
	G+C content (%)	47.1	43.6	49.9	43
Genes	No. of genes	85,139	41,925	69,018	65,832
	Average length of genes (bp)	26,201	11,959	8,834	8,192
	Average length of transcripts (nt)	1,423	2,067	1,423	1,479
	Gene models supported by EST (%)	85.2	77.2	67.5	62.5
Exons	No. of exons per gene	8.1	19.6	13.38	11.27
	Average length (bp)	185	99.8	105	130
	Total length (Mb)	126	82.1	98.2	97.3
Introns	No. of genes with introns (%)	93.7	95.3	83.4	80.3
	Average length (bp)	3468	499	561	622
	First two nucleotides at 5' splice sites	GT/GC/GA	GT/GC/GA	GT/GC/GA	GT/GC/GA
	Total length (Mb)	1715	331.5	481.8	421.2

*From Shoguchi *et al.* (2013); #From Shoguchi *et al.* (2018)**Table 3.2: Top 30 abundant domains in *A. gibbosum***

Pfam domain	Pfam ID	Function	Agb	Str	Bmi	Cla	Fka	Pfa	Ehu	Tth	Tbr	Gth
DUF4116	PF13475.5	Domain of unknown function (DUF4116)	13747	856	525	672	233	0	17	0	0	66
LRR_8	PF13855.5	Leucine rich repeat	5987	664	1963	2342	677	40	1003	296	408	1077
LRR_4	PF12799.6	Leucine Rich repeats (2 copies)	3437	1062	1427	2720	930	98	802	1702	686	1546
Ank_3	PF13606.5	Ankyrin repeat	2063	10804	4185	7439	2496	68	1444	381	153	1717
Ank	PF00023.29	Ankyrin repeat	2023	0	0	7044	2383	52	1341	322	132	1580
Ank_5	PF13857.5	Ankyrin repeats (many copies)	1951	9823	3716	6663	2123	60	1323	302	110	1464
Ank_4	PF13637.5	Ankyrin repeats (many copies)	1851	8877	3605	6115	2095	52	1246	352	112	1444
RVT_1	PF00078.26	Reverse transcriptase (RNA-dependent DNA polymerase)	1741	1143	369	1016	472	0	7	7	1	52
TPR_1	PF00515.27	Tetratricopeptide repeat	1459	1399	1224	2351	933	0	731	2724	0	1405
Ank_2	PF12796.6	Ankyrin repeats (3 copies)	1451	6962	2802	4663	1567	45	974	253	86	1041
PPR	PF01535.19	PPR repeat	1311	10508	5093	6033	820	8	417	104	84	198
TPR_2	PF07719.16	Tetratricopeptide repeat	1262	1752	1428	2485	1050	112	915	2943	329	1524
PPR_2	PF13041.5	PPR	1311	10508	5113	6025	774	16	373	22	93	269
TPR_11	PF13414.5	TPR repeat	1150	531	414	530	356	17	285	1609	152	856
Ephrin_rec_like	PF07699.12	Putative ephrin-receptor like	1092	1069	753	1154	355	53	69	360	0	1441
EF-hand_1	PF00036.31	EF hand	1089	3441	2726	2744	929	102	647	798	126	847
TPR_17	PF13431.5	Tetratricopeptide repeat	1078	488	451	1029	503	206	413	1890	110	831
EF-hand_6	PF13405.5	EF-hand domain	1057	3454	2718	2783	942	105	645	823	138	809
TPR_14	PF13428.5	Tetratricopeptide repeat	1039	843	876	1789	639	31	723	1666	264	1128
Pkinase	PF00069.24	Protein kinase domain	993	1475	989	1088	510	206	740	1894	232	603
TPR_8	PF13181.5	Tetratricopeptide repeat	971	977	735	1823	761	54	548	2683	194	1193
LRR_1	PF00560.32	Leucine Rich Repeat	933	360	573	2448	670	0	655	37	49	817
PPR_1	PF12854.6	PPR repeat	884	6095	3077	3616	470	8	334	17	43	196
PPR_3	PF13812.5	Pentatricopeptide repeat domain	830	6564	3445	4014	490	5	241	17	64	210
WD40	PF00400.31	WD domain, G-beta repeat	770	1728	980	1450	539	325	980	1793	573	1865
Pkinase_Tyr	PF07714.16	Protein tyrosine kinase	753	1287	852	915	424	105	657	1353	228	575
LRR_6	PF13516.5	Leucine Rich repeat	716	1539	1149	1680	776	22	1761	10655	290	1057
EF-hand_7	PF13499.5	EF-hand domain pair	710	2413	1889	1851	593	72	397	581	105	574
RCC1_2	PF13540.5	Regulator of chromosome condensation (RCC1) repeat	705	8647	2221	6193	496	57	406	284	57	519
TPR_12	PF13424.5	Tetratricopeptide repeat	680	1482	1196	2448	855	33	662	2084	174	1005

Agb, *Amphidinium gibbosum*; Str, *Symbiodinium tridacnidorum* (A3); Bmi, *Breviolum minutum* (B1); Cla, *Cladocopium* sp. (C); Fka, *Fugacium kawagutii* (F); Pfa, *Plasmodium falciparum*; Ehu, *Emiliania huxleyi*; Tth, *Tetrahymena thermophila*; Tbr, *Trypanosoma brucei*; Gth, *Guillardia theta*

3.3.2 Evidence of multifunctional PKS transcripts in *A. gibossum*

Till date, short reads RNA sequencing has been the major approach to understand the transcriptome; however, such an approach cannot generate full-length sequences for each RNA. Several studies focusing on recovery of *PKS* genes from short read transcriptome assemblies mainly reported single domain *PKS* genes (Pawlowicz *et al.*, 2014; Meyer *et al.*, 2015; Kohli *et al.*, 2015). I employed Pacbio Isoform sequencing (Iso-Seq) so as to gather direct evidence for *PKS* transcript production and its isoforms. After error-correction with Illumina short reads, 10 *PKS* transcripts were recovered, which consisted of several isoforms as well as complete multifunctional *PKS* genes (Figure 3.2). Interestingly, none of the transcripts contained the acyltransferase (AT) domain; this domain has been reported to be expressed mainly as a *trans*-acting in the *Symbiodiniaceae* dinoflagellates (Beedessee *et al.*, 2019) and is also observed from gene models from *A. gibossum*. To my knowledge, this is the first direct long read evidence of *PKS* transcripts from any dinoflagellate.

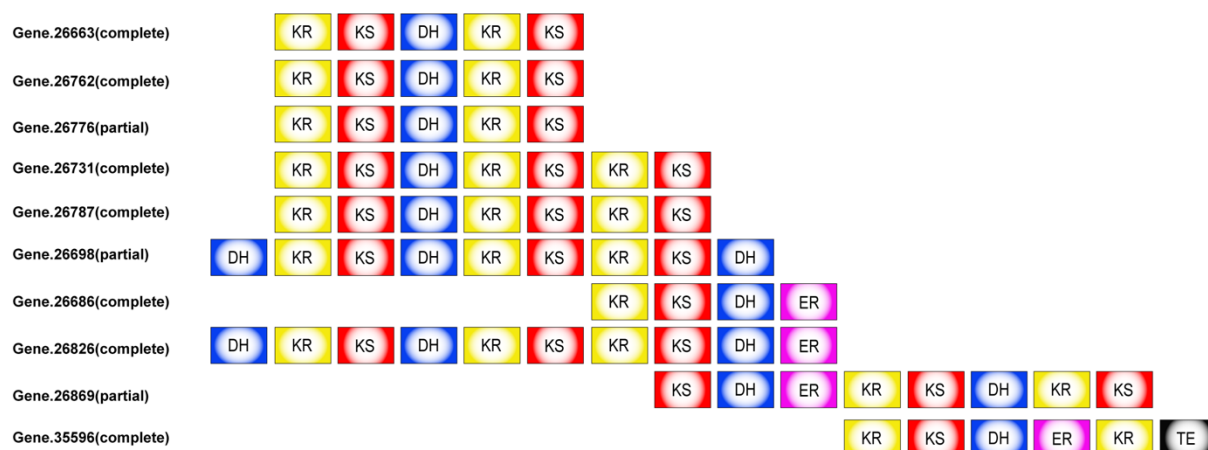


Figure 3.2 | PKS transcripts recovered from Iso-Seq (KS, ketosynthase; KR, ketoreductase, DH; dehydratase, ER, enoylreductase; TE, thioesterase).

3.3.3 Features of abundant domains, pathway and repetitive elements analysis

Pfam analysis showed that Leucine rich repeat (LRR), ankyrin, tetratricopeptide (TPR) and pentatricopeptide repeat (PPR) domains are the most abundant domains in *A. gibossum* (Table 3.2). Similar domain enrichment has been reported in several *Symbiodiniaceae* (Shoguchi *et al.*, 2013; González-Pech *et al.*, 2017; Shoguchi *et al.*, 2018). LRR, ankyrin and TPR repeat domains are known to play important roles in protein-protein interactions (Blatch & Lässle, 1999; Kobe & Kajava, 2001; Mosavi *et al.*, 2004) while PPR proteins are involved in RNA editing (Fujii *et al.*, 2011). Shoguchi *et al.* (2013) reported < 10% of the *Breviolum minutum* (B1) genome consist of transposons and tandem repeats. The *A. gibossum* genome consists of 29% of repetitive elements composed of simple repeats (1.97%), low complexity repeats (0.39%), satellite repeats (0.02%), LINEs (0.02%), LTR elements (0.03%), DNA elements (0.1%) and unclassified repeats (27.4%) (Appendix J). This repeat content is certainly an underestimation of the real repeat content considering the fragmented nature of the genome. Additionally, the genome assembly has a gap rate of 25%.

In order to understand how *A. gibossum* gene models are conserved at pathway level, predicted genes were mapped to KEGG reference pathways while comparing with other dinoflagellates and eukaryotes. This analysis resulted in the recovery of 388 KEGG pathways indicating that *A. gibossum* retains most of the pathways present in other eukaryotes and is comparable to other dinoflagellates (Figure 3.3).

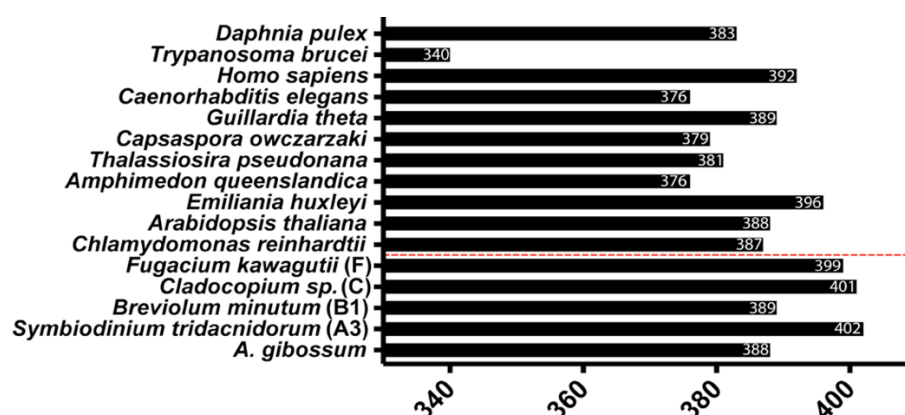


Figure 3.3 | KEGG pathway analysis of *A. gibossum*

3.3.4 Analyses of ketosynthase, acyltransferase, adenylation and condensation domains

In order to understand the evolution of *PKS* (*KS* & *AT*) and *NRPS* (*A* & *C*) genes in the *A. gibossum* genome, previous dataset used in Beedessee *et al.* (2019) was repopulated with *A. gibossum* gene models and subjected to phylogenetic analysis. Figure 3.4a shows that the *A. gibossum* *KS* domains clustered in a dinoflagellate-specific group under a reliable node (Bayesian Inference: 0.97). Additionally, this analysis supports the expanded nature of *A. gibossum* *KS* genes with cTP (chloroplast transit peptide) signal detected in 1 out of 14 of the sequences while 2 out of 14 sequences contained mitochondrial targeting peptide (mTP) or secretory signal each (Figure 3.4a). *A. gibossum* genome data contain fewer *KS* and *AT* genes compared to the Symbiodiniaceae dinoflagellates. Only eight *AT* genes were recovered from *A. gibossum* genome, which are *trans*-acting in nature (Bayesian Inference: 0.70) (Figure 3.4b). I-TASSER prediction suggests that these *AT* sequences belong mainly to the family of malonyl-CoA ACP transferase and thus brings malonyl-CoA for chain elongation. This result mirrors previous observation in Symbiodiniaceae, where malonyl-CoA is the preferred building block and attributes a degree of conservation for secondary metabolism in dinoflagellates.

A phylogenetic analysis involving 121 adenylation sequences was performed to understand the nature of freestanding A-domains identified in the *A. gibossum* genome. One major observation is that a freestanding A-domain falls into three major groups that utilize cysteine, valine, and phenylalanine as substrates, respectively (three highlighted yellow in Figure 3.5a). This is in contrast to glycine, tryptophan and phenylalanine as main substrates in Symbiodiniaceae (Beedessee *et al.*, 2019). The condensation enzyme (g6187.t1) in *A. gibossum* belong to ^LC_L subtype, same as that of Symbiodiniaceae and thus involved in condensation of two L-amino acids. This may be a common feature in dinoflagellates (Figure 3.5b). To understand cellular localization of PKS protein, antibodies against the KS and KR domains were employed. Immunolocalization indicated that KS and KR proteins were detected near membrane vesicles (Figure 3.6), although such proteins can be localized to other organelles (Monroe *et al.*, 2010).

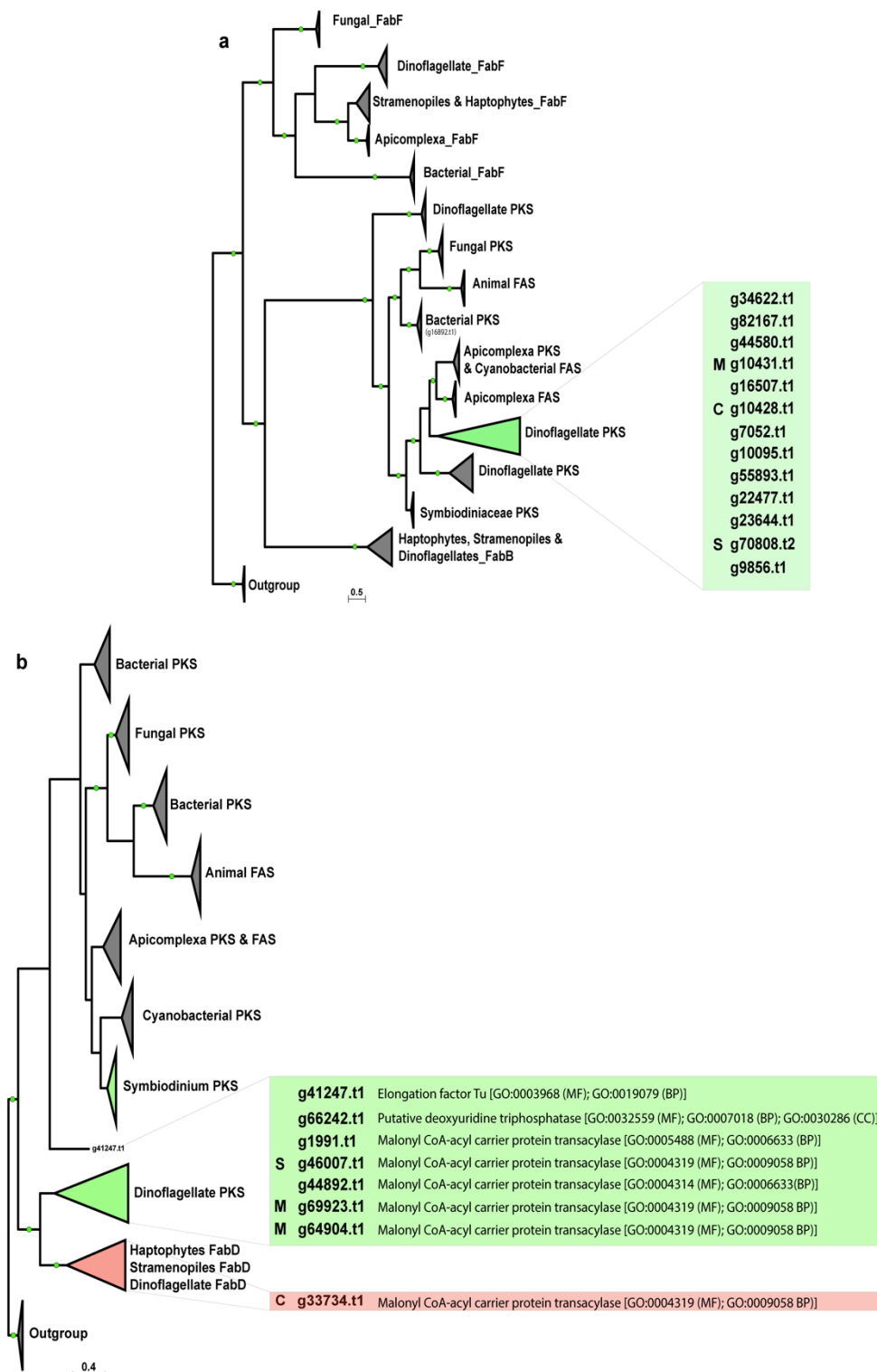


Figure 3.4 | Phylogenetic analysis of (a) ketosynthase (KS) and (b) acyltransferase (AT) domains using Bayesian inference. Green circles indicate a probability ≥ 0.75 .

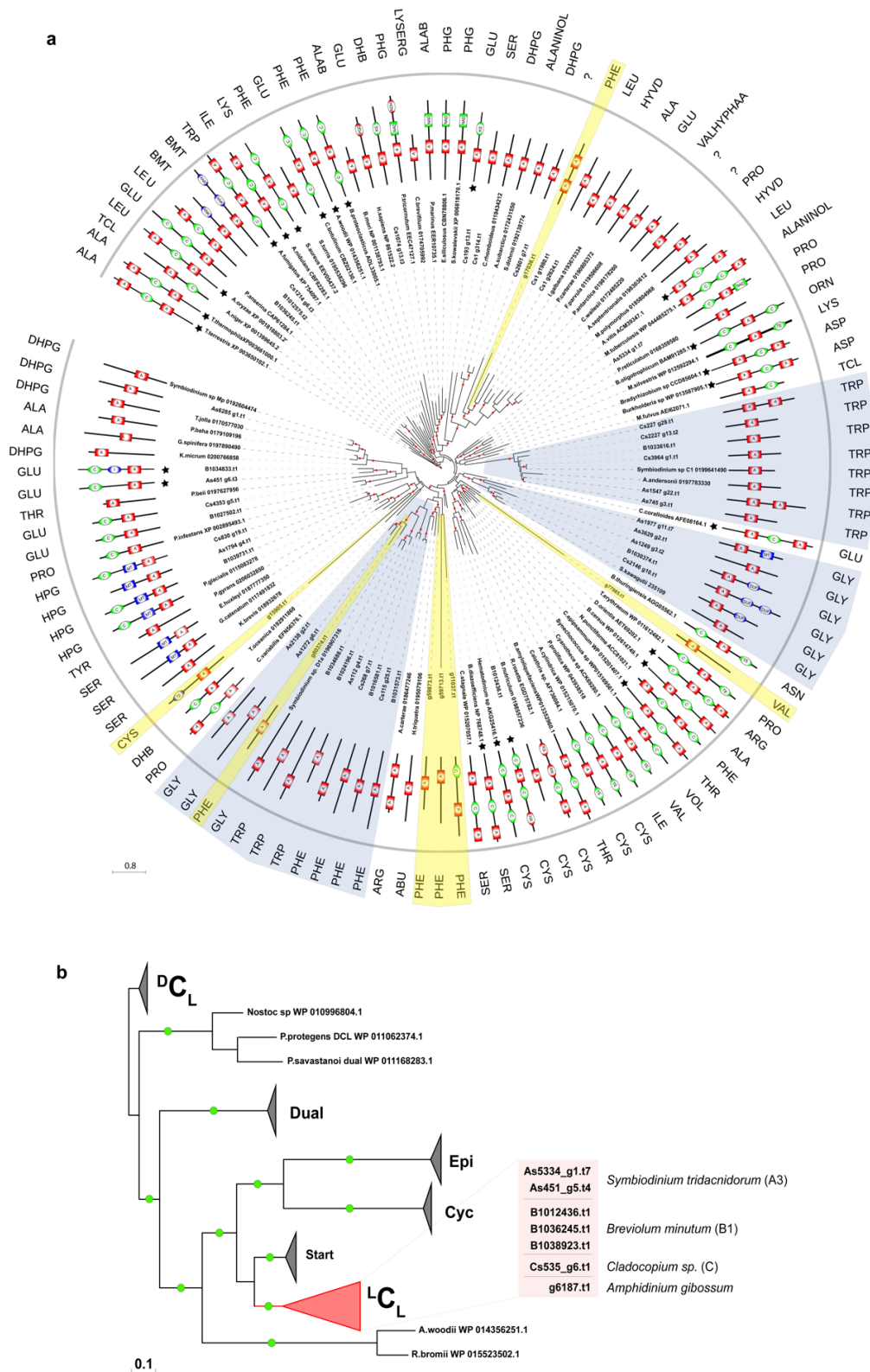


Figure 3.5 | Phylogenetic analysis of (a) adenylation and (b) condensation domains using Bayesian inference. The blue and yellow areas are Symbiodiniaceae and *A. gibossum* adenylation sequences, respectively. Green circles indicate a probability ≥ 0.75 .

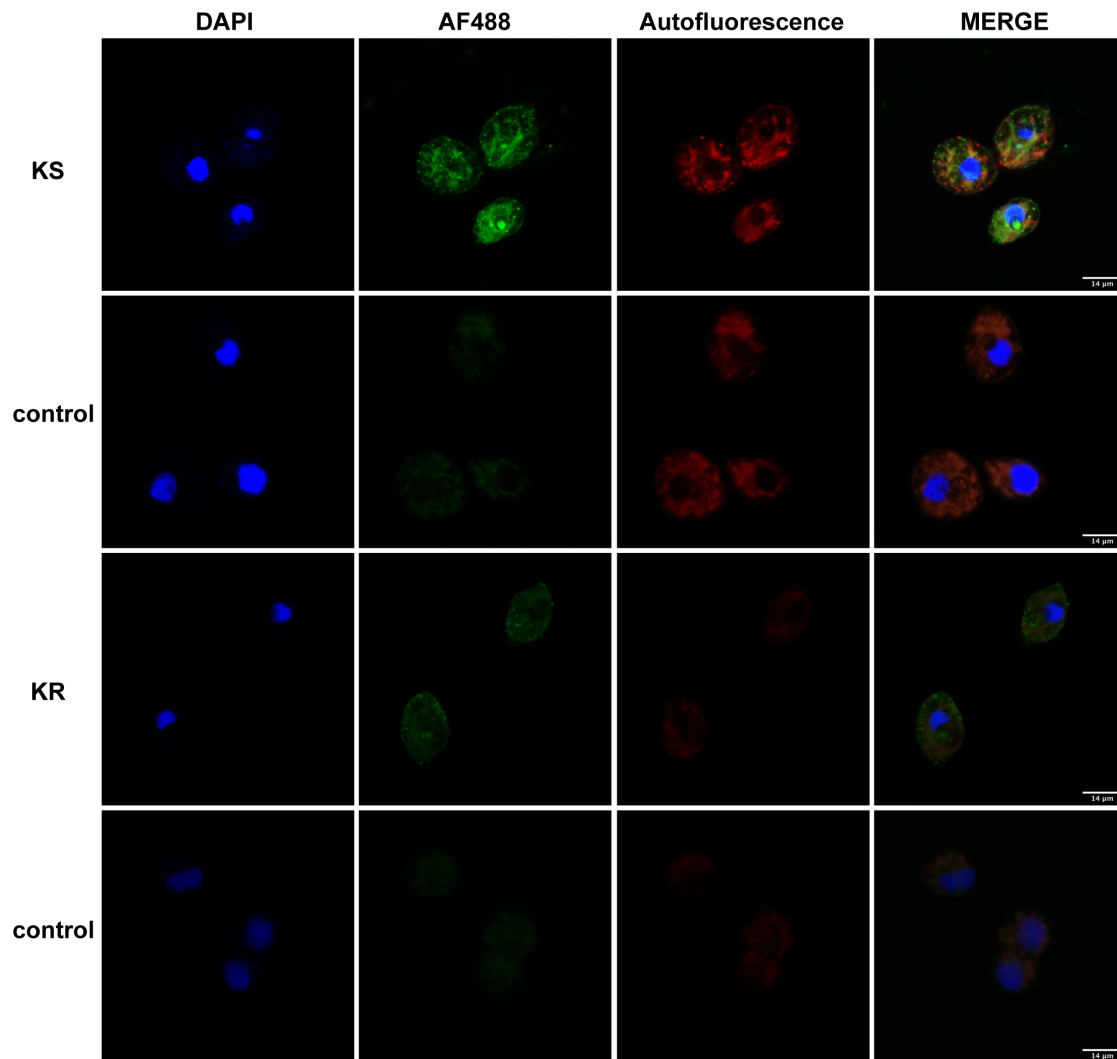


Figure 3.6 | Immunofluorescent staining of *Amphidinium* cells with anti-KS and anti-KR antibodies. Confocal images of antibodies show localization of KS proteins. Nuclei are stained blue with DAPI. KS proteins are in green, and merged images of nuclei and KS/KR protein staining, respectively. Scale bars are 10 µm in the panels.

3.4 Discussion

3.4.1 The advances of genomic findings of *A. gibossum*

The most comprehensive study till date provided preliminary insights into *Amphidinium carterae* genome employed a PCR approach to obtain genomic (120 kb) and cDNA (98 kb) sequences for 47 genes (Bachvaroff & Place, 2008). Sixteen genes were found to be in tandem arrays, and comparison of cDNAs to genomic copies revealed a polyadenylation sequence motif corresponding to AAAAG/C in the genome at the exact polyadenylation site. Interestingly, only 4/47 genes showed a more typical eukaryotic intron density with > 5 introns, namely polyketide synthase (18 introns), translation initiation factor 3 (8 introns), small nuclear ribonuclear protein (6 introns) and *psbO* (9 introns), respectively. The results obtained need to be interpreted in light of using a PCR approach despite similar patterns being observed in *Lingulodinium polyedrum* (Le, 1997) and *Symbiodinium* (Reichman *et al.*, 2003). The present study reveals that intron length is ~ 7 times longer than from reported genomes and occupied a significant portion of genes.

3.4.2 Biochemistry of secondary metabolism in dinoflagellates

Dinoflagellates focus exclusively on the biosynthesis of polyketides that are usually polyol in nature with occurrence of multiple ether rings, either fused as spirocyclic or in a ladder frame structure or both (Wagoner *et al.*, 2014). Polyketide biosynthesis is similar to that of fatty acids; the chain starts with initiation by acetyl CoA, extension by a series of Claisen ester condensation reactions with malonyl CoA, and termination when the required length and functionality is reached (Kellman *et al.*, 2010). These polyketides rarely contain nitrogen and the carbon skeleton is commonly assembled from acetate, with the addition of an amino acid (mainly glycine) in the assembly process to form hybrid polyketides (Walsh *et al.*, 2013). This incorporation is more common in other organisms than dinoflagellates, where only utilization

of glycine has been reported (Jones *et al.*, 2010; Wenzel & Müller, 2009). Glycine is one of the predominant substrates of the adenylation domain of Symbiodiniaceae (Beedessee *et al.*, 2019) and is involved in the generation of unique hybrid molecules such as zooxanthellatoxin B (ZT-B) and zooxanthellamide D (ZAD-D) (Figure 3.7 a, b).

Zooxanthellatoxin B (ZT-B), produced by *Symbiodinium tridacnidorum* (A3), is an example of how polyethers can arise via PKS/NRPS pathway. In ZT-B, glycine is incorporated as an extender unit. ZT-B is a 62-membered lactone, which is highly oxygenated and is made up of a spiroketal moiety and cyclic ethers along with the amide linkage from glycine (Nakamura *et al.*, 1995). ZAD-D, produced by *Breviolum minutum* (B1), is a linear polyhydroxylated polyene-type metabolite consisting of a C₂₇ amine and a C₂₁ acid substructures connected by a glycine (Fukatsu *et al.*, 2007). All the core enzymes needed for such biosynthesis can be recovered from both *Symbiodinium tridacnidorum* (A3) and *Breviolum minutum* (B1) genomes, proving that they play crucial roles in secondary metabolism. The presence of fused six-membered ring is evidence of the role of specialized enzymes such as epoxidases (Figure 3.7 a, b).

3.4.3 Secondary metabolism machinery is conserved in dinoflagellates

The genus *Amphidinium* is a rich source of polyketides, with amphidinolides being an expanding group of cytotoxic macrolides (Kobayashi & Tsuda, 2004) (Appendix I). The present study showed that the basic gene repertoire and substrate pool for secondary metabolism are same in the family Symbiodiniaceae and *A. gibossum*. Bachvaroff and Place (2008) suggested that *PKS* gene is present as a low-copy-numbered tandem array gene with low intron density in *Amphidinium carterae*. The genomic data and survey of *PKS* genes within *A. gibossum* is in agreement with the observation of Bachvaroff and Place (2008).

While amphidinolides structures are unique, some similarities can be found among them. C1-C8 and C7-C29 of amphidinolide U correspond to C1-C8 and C12-C34 of amphidinolides A and C, respectively. These similarities would suggest the existence of a common biogenic origin of amphidinolides. Biosynthetic tracer studies revealed that all the carbons of amphidinolides are derived from acetate (Rein & Snyder, 2009).

An attempt to isolate *PKS* genes from *Amphidinium* sp. Y-42 recovered a fragment that contained six regions corresponding to ketosynthase (KS), acyltransferase (AT), dehydratase (DH), ketoreductase (KR) and thioesterase (TE) domains, respectively with several frame-shifts present within and between catalytic domains (Kubota *et al.*, 2006). Approximately 15% of this 36.4-kb insert consisted of protein-coding sequence and would theoretically encode catalytic functions for only 1 cycle of a 26-membered polyketide. Extrapolation of these data suggests that all the genes needed for complete synthesis of an amphidinolide can occupy upto 500 kb of genomic DNA (Kellmann *et al.*, 2010). A survey of the genomic data of *A. gibossum* confirmed that such long cluster of *PKS* genes are not present; additional support for such possibility is the recovery of only 4000-5000 amino acids long *PKS* gene cluster from Iso-Seq data. Since each ketosynthase enzyme adds 2 carbon to a growing polyketide chain, a 26-membered polyketide would require at least 12 rounds of carbon addition, implying that such a long cluster may not be present in *A. gibossum*. Similar survey in *Breviolum minutum* recovered the longest *PKS* protein, made up to 10,601 amino acids (Beedessee *et al.*, 2015).

Thus, amphidinolide biosynthesis can happen by 2 ways, (1) monofunctional *PKS* proteins form an enzyme complex, and iteratively catalyze addition of substrate or (2) multifunctional small *PKS* protein utilize substrate in many cycles, to finally yield a product. Monofunctional *PKS* domains are key features in both Symbiodiniaceae and *A. gibossum* genomes (Beedessee *et al.*, 2015; Beedessee *et al.*, 2019).

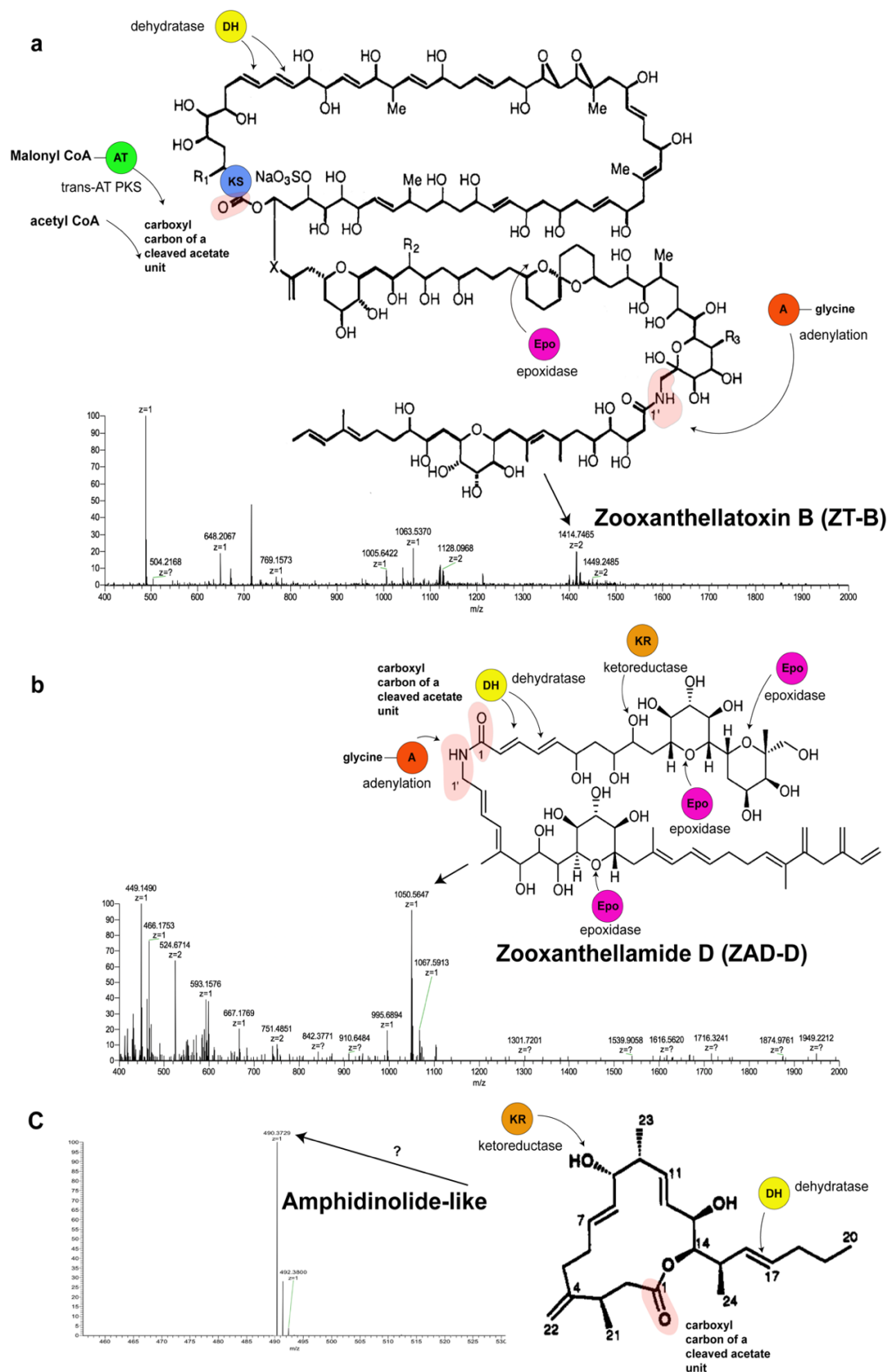


Figure 3.7 | Biosynthesis of specialized metabolites from Symbiodiniaceae and *A. gibosum* dinoflagellates showing (a) zoosxanthellatoxin B; (b) zoosxanthellamide D; (c) Amphidinolide-like molecule.

4 Transcriptome analysis of *Amphidinium gibbosum*

4.1 Introduction

Several studies demonstrated that parameters such as light, nutrients, temperature, and salinity affect photosynthesis and growth of dinoflagellates, but it is still debatable whether environmental factors influence toxicity. Toxin production in cultures of *Alexandrium tamarense* has been reported to be influenced by nutrient supplementation (Wang *et al.*, 2002). Saxitoxin-producing species of *Alexandrium* is found to have lower toxin content under nitrate starvation and higher toxin content under phosphate starvation (Erdner & Anderson, 2006). Despite different life styles, all organisms including dinoflagellate species require carbon (C), phosphorus (P) and nitrogen (N).

Phosphorus is an important nutrient for phytoplankton growth as it is involved for cellular structures (membranes, DNA, RNA), metabolism (nucleotides), energy storage (ATP), cell signaling (IP₃, cAMP) and biochemical regulation (protein phosphorylation) (Karl., 2014; Lin *et al.*, 2016). Phosphate limits phytoplankton growth when excess nitrogen is present in eutrophic coastal waters (Lin *et al.*, 2016). Phytoplankton undergo several changes under phosphate limitation namely change in P transporters (Perry, 1976), cell membrane remodeling (Shemi *et al.*, 2016), bypassing processes consuming phosphate in glycolysis (Wurch *et al.*, 2011). Nitrogen is essential for the synthesis of amino acids, chlorophylls, nucleic acids, and toxins and any change in concentration of nitrogenous compounds can significantly affect metabolism (Dagenais-Bellefeuille & Morse, 2013).

Dinoflagellates, in general, exhibit limited transcriptional regulation and recent reports of microRNAs in dinoflagellates suggest that microRNAs can be involved in post-transcriptional regulation to control gene expression (Baumgarten *et al.*, 2013; Gao *et al.*, 2013; Lin *et al.*, 2015; Geng *et al.*, 2015; Dagenais-Bellefeuille *et al.*, 2017). However, molecular

mechanisms associated to nutrient starvation remain unclear and no investigation of the role of microRNAs and their expression profiles have been conducted in *Amphidinium*. In this chapter, I present a transcriptomic and microRNAomic analysis of *Amphidinium gibossum* to examine the effects of nitrogen and phosphate limitation on gene expression and to explore if any post-translational regulation via the role of microRNA is involved during such nutrient stress.

4.2 Materials and Methods

4.2.1 Biological sample

Amphidinium gibossum was originally isolated by Dr. Takaaki Kubota (Showa Pharmaceutical University, Tokyo, Japan) from the inner cells of acoel flatworms, *Amphiscolops* species found near Ishigaki Island. The culture was maintained in artificial seawater containing 1X Guillard's (F/2) marine-water enrichment solution and antibiotic-antimycotic mix in a 25°C incubator under a 12:12 light-dark cycle. Subcultures were performed ~ every 4 weeks with fresh medium and handled strictly aseptically.

4.2.2 Culture and nutrient treatment

For the nitrate-depletion experiment, culture medium was prepared by supplementing artificial seawater (ASW) with F/2 medium containing a reduced nitrate concentration (150 μ M). For the phosphate-depletion experiment, the phosphate level was 22 μ M. A phosphate and nitrate-replete treatment was set up as the control, where phosphate and nitrate concentration were 36 μ M and 880 μ M, respectively. Both deplete, and replete treatment were carried in triplicate. First measurements were started after 24 hours of stabilisation and this was counted as Day 1. Nitrate and phosphate levels were monitored every two days using Greiss and phosphomolybdenum blue spectrophotometric methods (Miranda *et al.*, 2001; Parsons, 1984) until their concentration were undetectable. Other physiological parameters such as cell

concentration, chlorophyll *a* and photochemical efficiency (F_v/F_m ratio) were monitored at the same time. Cell counts were obtained by fixing cells in formalin and visualisation using a haemocytometer. A 1-ml sample was centrifuged, and the cell pellet was immersed in N, N-dimethylformamide (DMF) and kept at -20 °C for at least 12 hours in order to extract chlorophyll *a*, which was then measured using a Turner Trilogy (Turner Designs fluorometer, USA) and averaged to per cell content. Photochemical efficiency was monitored with a Xe-PAM (Walz, Germany) (Appendix M).

4.2.3 Transcriptome analysis, annotation and differential gene expression

For nutrient stress experiment, cells were subjected to standard growth condition (12:12 light and dark cycle) without any antibiotics. On the day the level of dissolved nitrate and phosphate were undetectable, $\sim 10^7$ cells were collected and snap frozen in liquid nitrogen and ground using a cryopress. RNA was extracted from 3 control, 3 nitrate-depleted and 3 phosphate-depleted samples using PureLink reagent. 4 µg of RNA was used for cDNA library construction using Truseq stranded RNA Sample preparation kit (Illumina). Nine libraries were quantified and validated by qPCR and a 2100 Agilent Bioanalyzer respectively and sequenced in two lanes of Hiseq 4000 (Illumina). Reads were trimmed using Trimmomatic (v0.35) (Bolger, 2014) and quality-checked using FastQC (v0.11.4) and assembled *de novo* using Trinity (v2.3.2) (Haas *et al.*, 2003). The assembly was processed with CD-HIT-EST (v4.6.7) using a clustering threshold of 0.95 (Li and Godzik., 2006). BLASTN searches against several databases was conducted: draft and complete bacterial genomes (<ftp://ftp.ncbi.nih.gov/genomes/Bacteria/all.fna.tar.gz>, ftp://ftp.ncbi.nih.gov/genomes/Bacteria_DRAFT/) from NCBI. A combination of cutoffs (total bit score >1000, e-value $\leq 10^{-20}$) was used to identify scaffolds with similarities to bacterial sequences.

Functional annotation of non-redundant contigs was performed using the Trinotate pipeline (<https://trinotate.github.io/>) against several databases namely Uniprot (Swiss-Prot), GeneBank non-redundant (nr), Kyoto Encyclopedia of Genes and Genomes (KEGG), EggNog using Blast with an E-value cut-off of 10^{-5} (Altschul *et al.*, 1990; Kanehisa *et al.*, 2012). The transcriptome gene completeness was evaluated using BUSCO v3.0.2 (Simão *et al.*, 2015). For identification of differentially expressed transcripts, expression abundance was quantified using RSEM (Li & Dewey, 2011). The R package, EdgeR was used to identify differentially expressed genes using adjusted *p*-values (q-value) determined by Benjamini, Krieger and Yekutieli correction of the PRISM package. Gene ontology terms for functional enrichment was performed using Fishers Exact test in topGo using the parent-child analysis to categorize whether differential expressed genes were enriched in molecular function, cellular components and biological processes (Alexa & Rahnenfuhrer, 2010). KEGG pathway enrichment was performed using DAVID using the Fisher's Exact test (Huang *et al.*, 2009).

4.2.4 Bioinformatic analysis of small RNA

Small RNAs were isolated from 2 µg of RNA obtained from the same pellets used for total RNA extraction (including nutrient treatment), using the NEXTflex™ Small RNA-seq kit V3 (Bioo Scientific). All isolations were quantified and quality-checked using a nanodrop (ThermoScientific) and a 2100 Bioanalyser (Agilent, Santa Cruz, USA), respectively. Single-end reads (1 x 50 bp) were generated from nine libraries on a Hiseq 2500 platform. Reads were cleaned by removing adapter and polyA/N sequences using Cutadapt-1.4.1 (Martin, 2011), and only reads within the range of 17-25 were kept. Reads were further collapsed using the collapse_reads.pl script of MiRDeep2 package (Friedlander *et al.*, 2012). Sequences having hits to various non-coding RNAs (rRNAs, tRNAs, snRNAs, snoRNAs and scRNAs) of the RNACentral database (The RNACentral Consortium, 2015) were discarded. Bowtie v1.1.12

(Langmead *et al.*, 2009) was used to map clean small RNA reads to the *Amphidinium gibossum* genome with no mismatch and 1 alignment. Mapped reads were further queried against known miRNAs in miRBase 22.0 (<http://www.mirbase.org>). miRNAs were annotated using the miRdeep2 package. miRNA criteria used by Baumgarten *et al.*, (2013) were applied to the list of annotated miRNAs. Expression level of miRNAs was conducted and normalized using the quantifier.pl script of miRdeep2 package where processed reads were mapped to identified miRNA precursors. EdgeR was then used to identify differentially expressed miRNAs at FDR < 0.05 (adjusted *p*-value), determined by Benjamini, Krieger and Yekutieli of the PRISM package and $|\log_2(\text{FC})| > 1$. Only miRNAs present in at least 2 replicates were considered further. For predicting the mRNA targets of the miRNAs, the 3'UTR sequences of unigenes by employing miRanda with strict criteria (Enright *et al.*, 2003). GO and KEGG pathway enrichment were performed for the predicted target unigenes of differentially expressed miRNAs using topGO and DAVID, respectively (Alexa & Rahnenfuhrer, 2010; Huang *et al.*, 2009).

4.2.5 Identification of key proteins in microRNA biogenesis pathways

In order to confirm the presence of a miRNA biogenesis pathway, sequences of three core protein families involved in RNA interference (i.e. Argonaute, Dicer and HEN1) were retrieved for model organisms (*H. sapiens*, *C. elegans*, *S. pombe*, *D. melanogaster*, *A. thaliana*) from UniProtKB (Magrane, 2011). The sequences were then queried against predicted proteins from *A. gibossum* transcriptome using BLASTP at e-values < $1e^{-10}$. The hits were then searched for specific domains (a PAZ domain and a pair of RNase III domains for Dicer, a Piwi and Dicer domains for Argonaute, and a methyltransferase domain for HEN1) needed for functional activity using InterProScan (Jones *et al.*, 2014). An alignment of the homologs against retrieved

RNAi proteins from model organisms was conducted using Clustal Omega (Sievers *et al.*, 2011) and visualize using Jalview (Clamp *et al.*, 2004).

4.2.6 Mass spectrometry

Culture medium as well as cell pellet was saved for mass spectrometry analysis. Cell pellet was extracted with methanol: ether (3:1), three times at room temperature. Methanol (400 µl) was added to the biomass followed by vortex (1 min), sonication (10 min), and centrifugation (14,000 g, 10 min, 10°C) to give the first extract. The resulting clear solution was transferred into a new tube. By adding methanol: ether (400 µl) to the residue, the 2nd extraction was carried out in the same fashion. The clear extract was again collected in the 1st extract and stored at -30°C. Additional methanol: ether (400 µl) was added to the residue and vortexed (1 min). After centrifugation, the 3rd extract was pooled with the previous extracts (total 1,200 µl) and marked as crude extract and lyophilized. For analysis, extract was suspended with ether: water (1:1), vortexed and centrifuged (14,000 g, 10 min, 10°C), giving the organic and aqueous layer. This step was repeated. The organic layer was dried and reconstituted with 200 µl methanol. A 20 µl aliquot was dissolved in 100 µl 50% water-methanol containing 0.25% formic acid. The suspension was centrifuged, and clean solution was transferred to a new tube and further diluted with methanol-water (1:1) solution and analyzed immediately.

4.2.7 NanoLC-MS analysis of the *Amphidinium* extract

A Thermo Scientific hybrid (LTQ Orbitrap) mass spectrometer was used for MS data collection. The mass spectrometer was equipped with a HPLC (Paradigm MS4, Michrom Bioresources Inc.), an auto-sampler (HTC PAL, CTC Analytics) and a nanoelectrospray ion source (NSI). The high-resolution MS spectrum was acquired at 60,000 resolution in FTMS

mode (Orbitrap), full mass range m/z 200-1,500 Da with capillary temperature (200 °C), spray voltage (1.9 kV), and both positive and negative ion modes were used. The lipid-depleted crude extract (the stock solution) was separated on a capillary ODS column (50 × 0.18 mm, 3 µm, C₁₈, Supelco). A 30-min gradient (30% B for 0-2 min, 30-80% B for 2-12 min, hold 80% B for 12-16 min, hold 100% B in 20-25 min, equilibration 30% B in 20-30 min; where solvent A is aqueous-acetonitrile-formic acid 98:2:0.1% and solvent B is acetonitrile-water-formic acid 98:2:0.1% ; flow rate 2.0 µl/min, injection 2.0 µl loop) was used for separation.

4.3 Results

4.3.1 Transcriptome assembly and functional annotation

De novo assembly of the RNA-seq reads yielded 322,846 unigenes with an average length of 762 bp and N50 of 1237 bp. After clustering using CD-HIT-EST, a total of 186,803 were obtained with an average length of 739 bp and N50 of 1278 bp. GO terms were assigned to 78,037 (42%) unigenes. Cellular and functional processes were the most highly represented groups (29%) (Figure 4.1). KEGG pathway classification revealed 422 pathways to be present (Appendix N). Of these, metabolic and biosynthesis of secondary metabolites pathways were those with the highest number of enzymes (1187). BUSCO analysis using the eukaryotic database identified 81.2% of the 303 BUSCOs (80.2% complete; 4% fragmented) suggesting a large part of the transcriptome was represented.

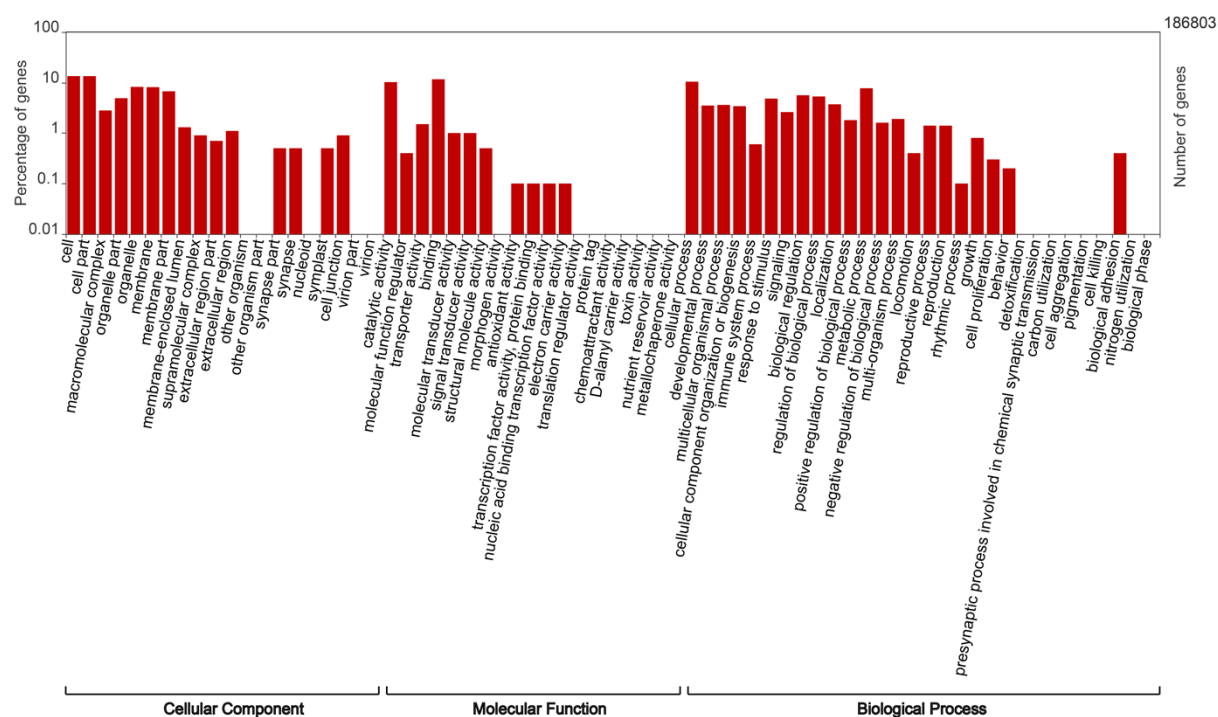


Figure 4.1 | Gene annotation of *Amphidinium gibosum* unigenes using gene ontology (GO).

4.3.2 Differential expression analysis under nitrogen starvation

A total of 624 genes were differentially expressed (Figure 4.2a), among which 250 and 374 were up- and downregulated respectively, under N-depleted condition relative to control ($|\log_2(\text{FC})| > 1$ and $q\text{-value} < 0.001$). The inclusion of biological replicates allowed better correlation of the observed gene expression. Only 16 *PKS* and *NRPS* unigenes were found to be differentially expressed at $|\log_2(\text{FC})| > 2$ and $p < 0.001$ (Figure 4.2b). Differentially expressed genes were analyzed using topGO enrichment based on Biological process and Molecular Function, on a ranking of ($|\log_2(\text{FC})| > 1$ and $q\text{-value} < 0.001$). This analysis showed that N starvation has significant effects on nitrogen transport and metabolism (upregulated) and anion uptake (downregulated) (Figure 4.2c). KEGG pathway enrichment analysis using DAVID confirmed this observation with nitrogen metabolism, being the most enriched pathway among upregulated genes (Table 4.1), while three pathways (Bile secretion,

Proteoglycans in cancer and Pancreatic secretion) related to bicarbonate uptake were most enriched among downregulated genes (q-value < 0.001) (Table 4.2).

4.3.3 Differential expression analysis under phosphate starvation

A total of 16,494 were differentially expressed, among which 16,449 and 45 unigenes were up- and downregulated, respectively, under P-depleted condition relative to control ($|\log_2(\text{FC})| > 2$ and q-value < 0.001) (Figure 4.3a). Only 528 *PKS* and *NRPS* unigenes were found to be differentially expressed at $|\log_2(\text{FC})| > 2$ and $p < 0.05$ (Figure 4.3b). Differentially expressed genes were analyzed using topGO enrichment based on Biological process and Molecular Function, according to a ranking of ($|\log_2(\text{FC})| > 2$ and q-value < 0.001). This analysis showed that P starvation has significant effects on small molecule biosynthesis (upregulated) and anion uptake (downregulated) (Figure 4.3c). KEGG pathway enrichment analysis using DAVID confirmed this result with Ribosome, metabolic and biosynthesis of secondary metabolites pathways as being the most enriched pathway among upregulated genes (q-value < 0.01) (Table 4.3).

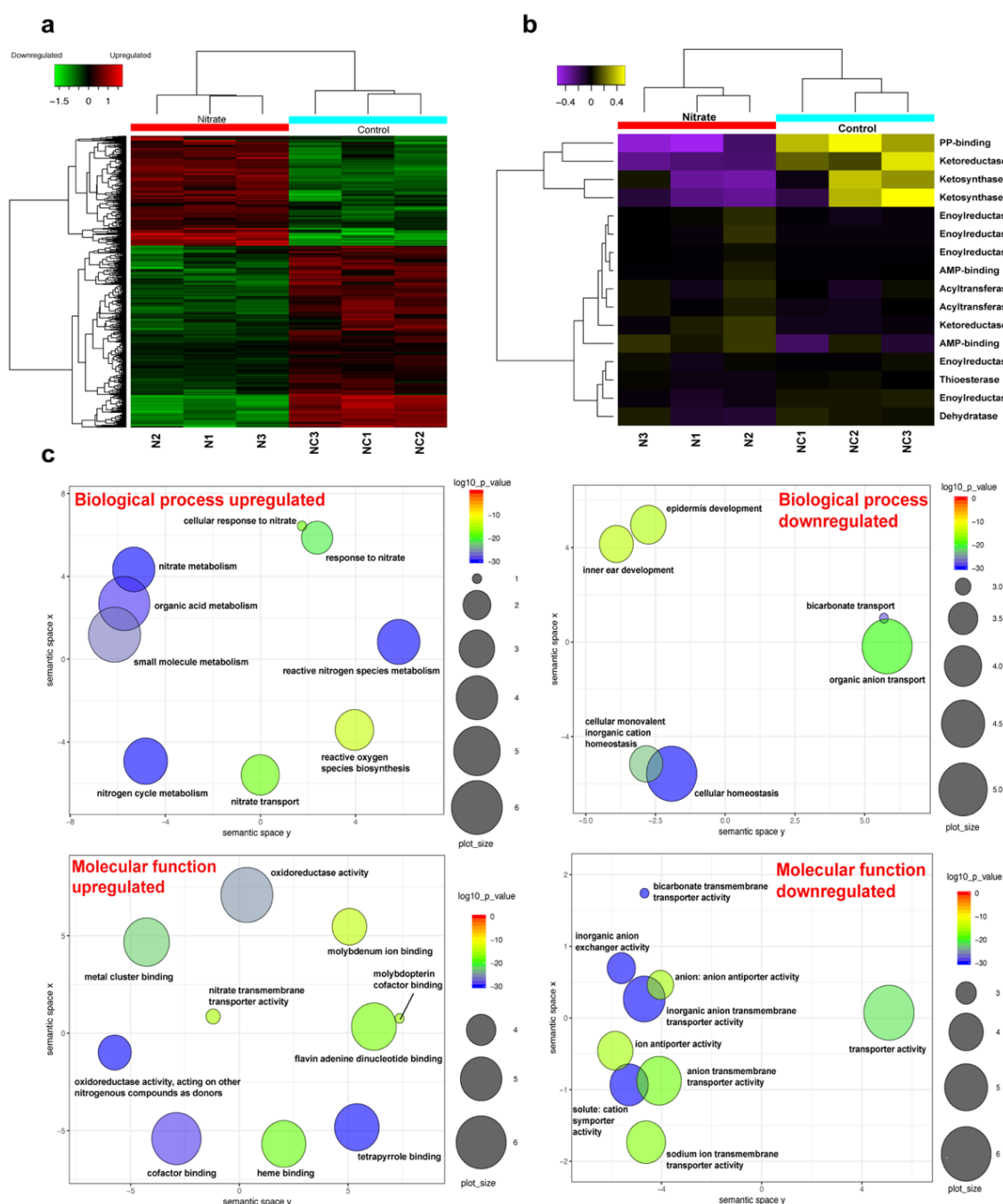


Figure 4.2 | (a) Global expression profile of differentially expressed genes under nitrogen starvation ($q\text{-value} < 0.001$ and $|\log_2(\text{FC})| > 1$). (b) Expression profile of *PKS* and *NRPS* genes ($p < 0.05$ and $|\log_2(\text{FC})| > 2$). (c) Functional enrichment of GO-terms analyzed with topGO and summarized with REVIGO showing top 10 hits in nitrogen starvation. Scatterplots summarize the GO terms based on semantic similarities. Similar GO terms remain in close proximity together. Bubble color represent the adjusted $p < 0.001$ while circle size shows the frequency of the GO term.

Table 4.1: Significantly enriched KEGG pathways upregulated under N starvation

Pathway	Gene count	Fold enrichment	p-value (Fisher's Exact Test)
Nitrogen metabolism	6	234	1.70E-05

Table 4.2: Significantly enriched KEGG pathways downregulated under N starvation

Pathway	Gene count	Fold enrichment	p-value (Fisher's Exact Test)
Bile secretion	5	118.3	1.10E-06
Proteoglycans in cancer	2	49.3	6.70E-04
Pancreatic secretion	4	78.9	2.40E-04

Table 4.3: Significantly enriched KEGG pathways upregulated under P starvation

Pathway	Gene count	Fold enrichment	p-value (Fisher's Exact Test)
Ribosome	25	2.5	3.70E-07
Metabolic pathways	78	1.5	6.80E-06
Biosynthesis of secondary mebolites	38	1.5	2.00E-03
Inositol phosphate metabolism	10	2.4	1.70E-03
Purine metabolism	13	2.1	3.00E-03
mRNA surveillance pathway	12	2.0	6.40E-03
Carbon metabolism	12	2.0	6.40E-03
Pyrimidine metabolism	10	2.3	3.40E-03
Glyoxylate and dicarboxylate metabolism	6	3.1	2.30E-03
Biosynthesis of antibiotics	11	2.1	5.00E-03
Endocytosis	16	1.7	1.10E-03
Insulin signaling pathways	7	2.6	6.10E-03
Proteoglycans in cancer	6	2.7	6.90E-03
Glutamatergic synapse	8	2.2	1.00E-02
Neuroactive ligand-receptor interaction	8	2.2	1.00E-02
Biosynthesis of amino acids	12	1.7	3.10E-02
Ascorbate and aldarate metabolism	5	2.6	1.90E-02
Ubiquitin mediated proteolysis	5	2.6	1.90E-02
Phosphatidylinositol	8	1.9	3.00E-02

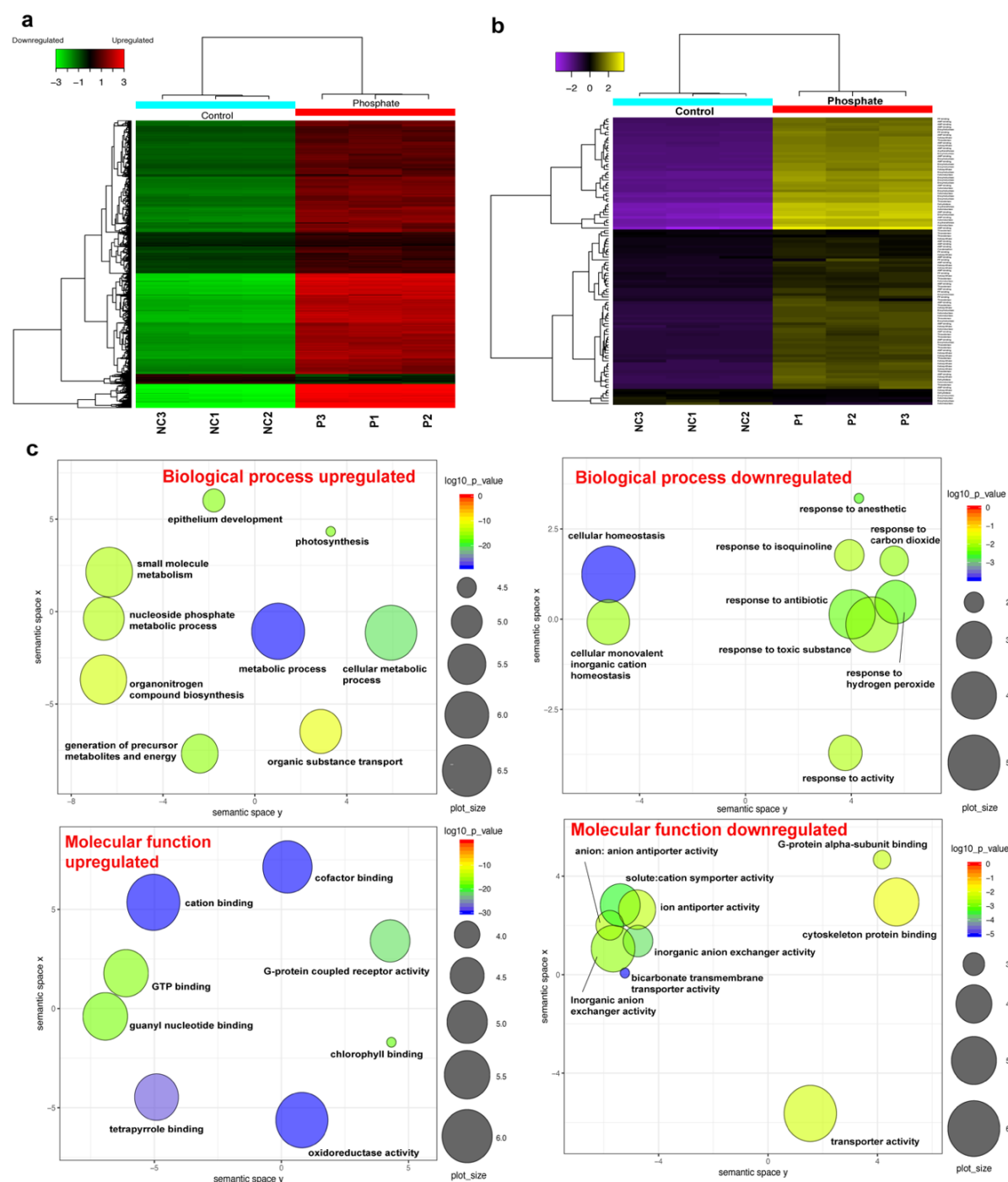


Figure 4.3 | (a) Global expression profile of differentially expressed genes under phosphate starvation ($q\text{-value} < 0.001$ and $|\log_2(\text{FC})| > 2$). (b) Expression profile of *PKS* and *NRPS* genes ($p < 0.05$ and $|\log_2(\text{FC})| > 2$). (c) Functional enrichment of GO-terms analyzed with topGO and summarized with REVIGO for phosphate starvation showing top 10 hits. Scatterplots summarize the GO terms based on semantic similarities; Similar GO terms remain in close proximity together. Bubble color represents the adjusted $p < 0.001$ while circle size shows the frequency of the GO term.

4.3.4 Identification of miRNAs, differential expression and target prediction

Using the *Amphidinium gibossum* genome, 107 miRNAs could be predicted using stringent criteria (no mismatch allowed during alignment). Of these, 84 mature miRNAs and 23 novel miRNAs were identified. miRNA lengths ranged between 17 to 25 nucleotides, with a peak at 18 nucleotides (Appendix O). This result contrasts to typical miRNAs in animals and plants with length 22 and 21 nt, respectively and is in agreement with observation from diatoms and haptophytes (Lopez-Gomollon *et al.*, 2014). To identify miRNAs involved in N and P starvation, normalized expression of miRNAs was compared. Under nitrogen starvation, only 1 miRNA (bdi-miR7721-5p) was found to be differentially expressed ($q\text{-value} < 0.05$, $\log_2(\text{FC}) > 2$) while under phosphate starvation, 3 miRNAs (has-miR-6874-5p, bdi-miR7721-3p and a novel miRNA from *A. gibossum*) were found to be significantly differentially expressed ($q\text{-value} < 0.05$). The fold change of these three miRNAs were > 18 compared to control, suggesting that they may have important effects during nitrogen and phosphate starvation.

To understand the role of these three differentially expressed miRNAs, potential targets were predicted using miRanda 3.3a (Enright *et al.*, 2003). The miRNA, bdi-miR7721-5p, under N starvation was found to have 303 potential target genes. KEGG pathway enrichment analysis showed that pyruvate metabolism, dilated cardiomyopathy, GABAergic synapse and hypertonic cardiomyopathy were the most enriched pathways (fold enrichment > 24 , $q\text{-value} < 0.001$, Fisher Exact test). The GO molecular function enrichment suggests that this miRNA target genes are involved mainly in myosin light chain kinase activity and extracellular matrix protein binding (Appendix P). The upregulated miRNA under P starvation was found to have 2711 potential target genes. KEGG pathway enrichment analysis showed that fructose-mannose metabolism, proteoglycans in cancer and N-glycan biosynthesis pathways (fold-enrichment > 4 , $q\text{-value} < 0.01$, Fisher Exact test). The GO molecular function suggest that these miRNA target genes participate in a series of activities including hydrolase activity,

nucleic acid and carbohydrate binding, transporter binding and cation binding as the top 5 most enriched GO terms (Appendix Q).

4.3.5 Metabolomics analysis

To get better insights into what metabolites are synthesized under N and P starvation, samples were collected at Day 1, 7 and 14, and crude extracts were prepared, and analyzed using NanoLC-MS, with focus on isolation of polar compounds. One major observation was that these metabolites appeared to be novel in nature and didn't not matched any reported metabolites from *Amphidinium* species till date (Appendix K). By Day 7, all the samples appeared to favor the production of small molecules within the range of m/z 450-550. However, by Day 14, larger molecules could be detected. Interestingly, by Day 14, several molecules were found to be common between the treated and control samples (Figure 4.4). This confirms that under nutrient starvation, cells continue to synthesize some common polyketide in nature as supported by NMR data (Appendix L).

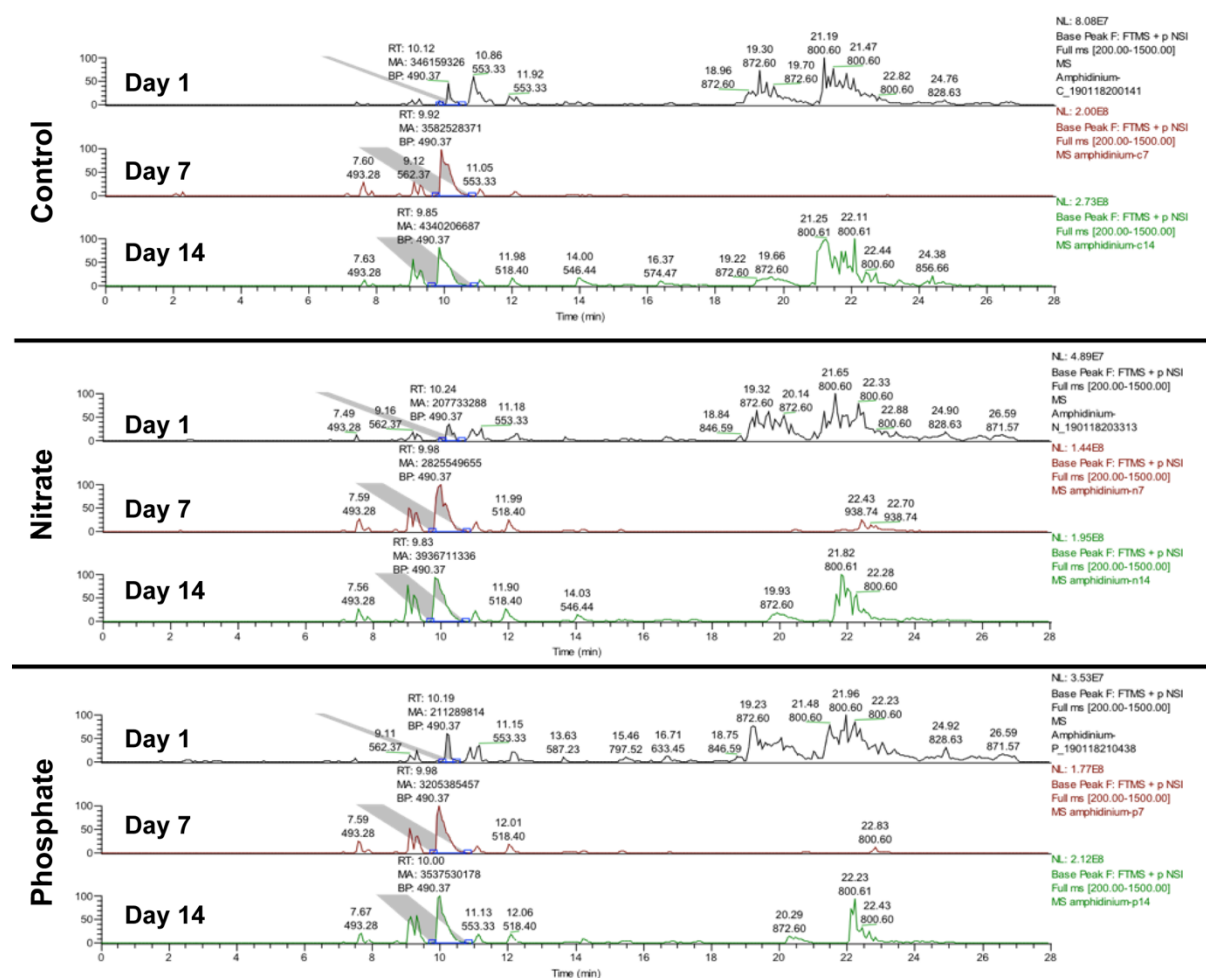
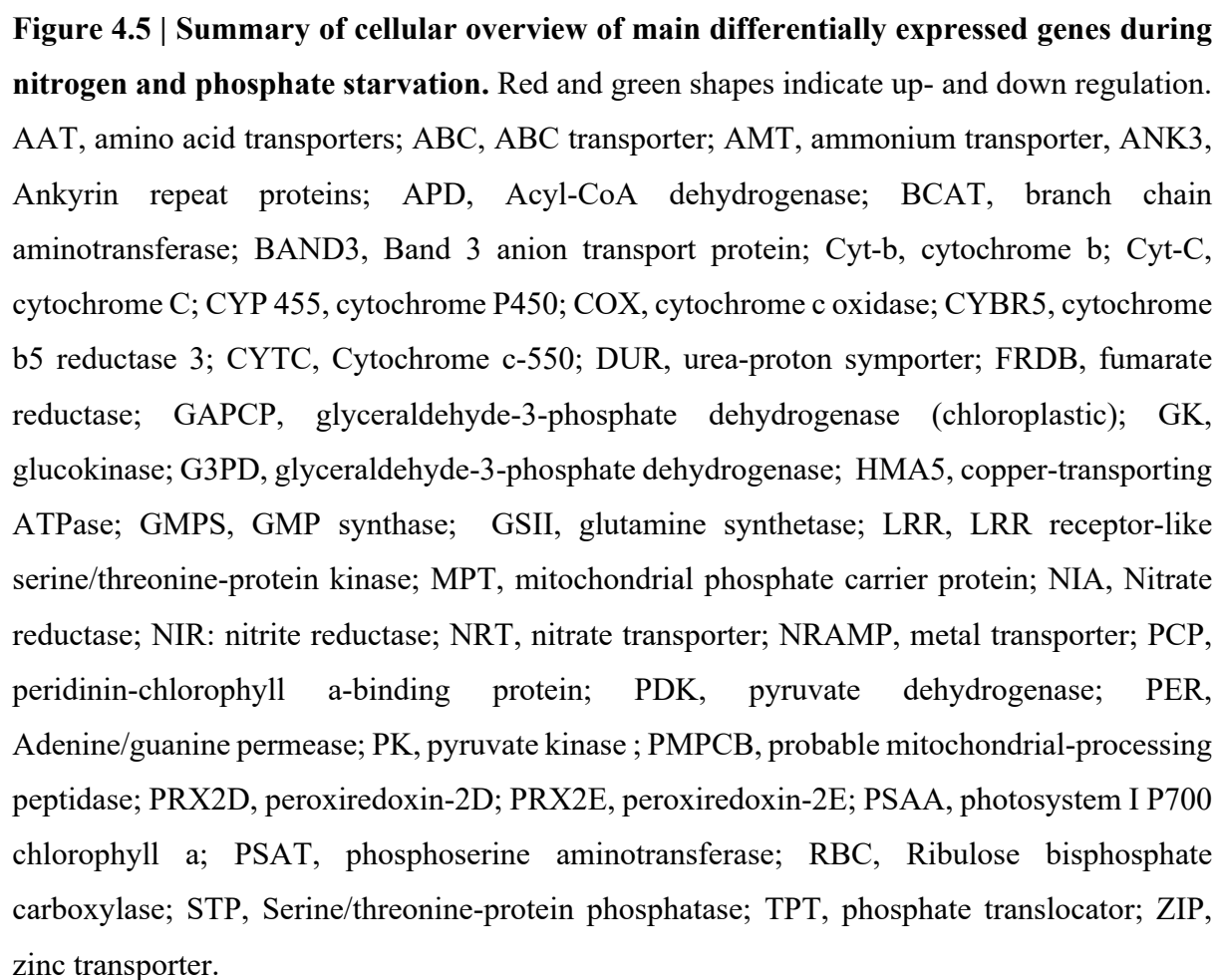


Figure 4.4 | NanoLC-MS profile of the methanol extract of *Amphidinium gibosum* at three time points (Day 1, Day 7 and Day 14) under the control, nitrate and phosphate stress.



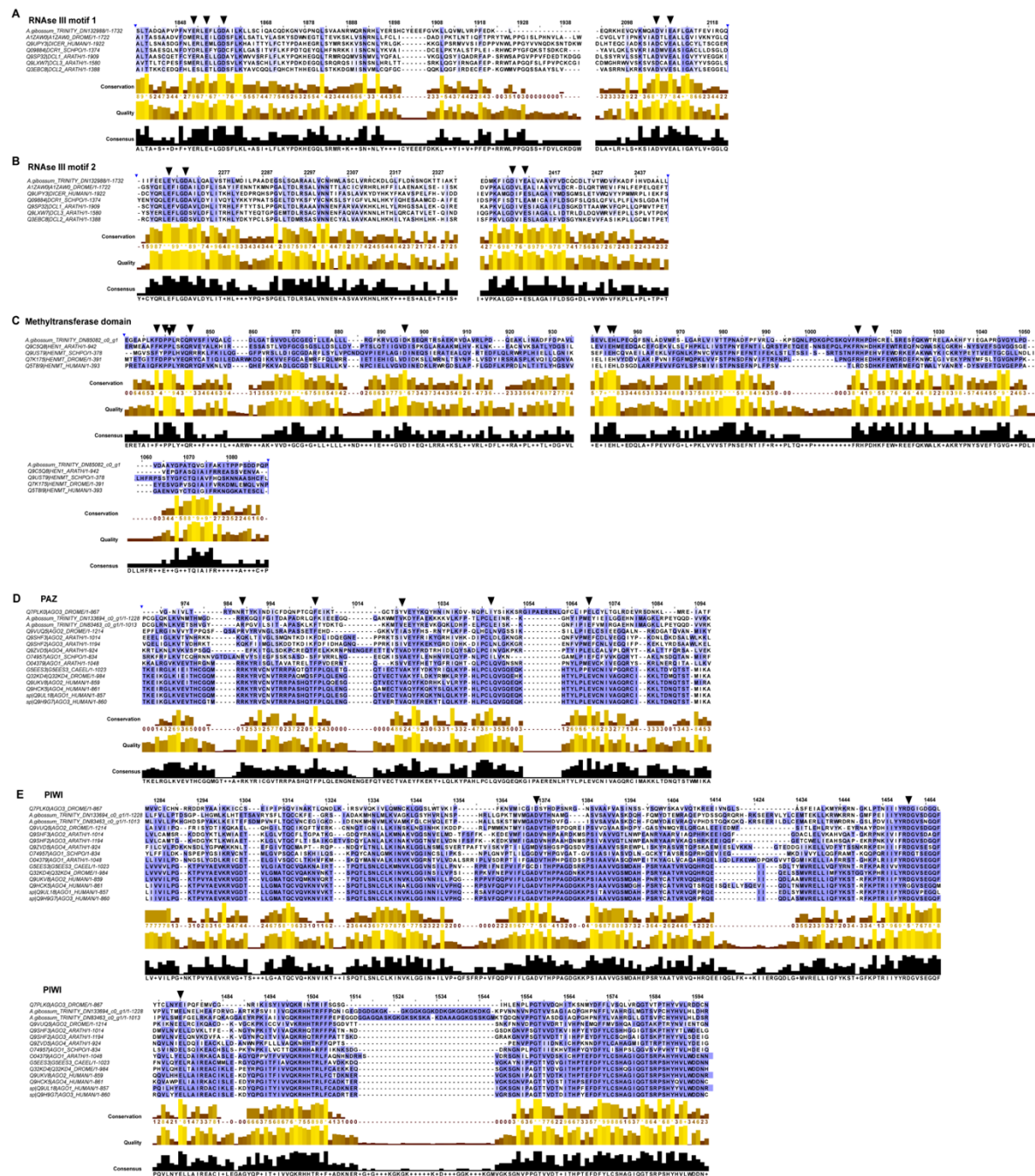


Figure 4.6 | Alignment of functional domains of the *A. gibosum* homolog. Alignments depicts Dicer RNase III (A) motif 1 and (B) motif 2, (C) small RNA 2'-O-methyltransferase (HEN1), argonaute protein, D (PAZ domain) and (E) PIWI domain with homologs from model organisms. Key functional residues are depicted with black triangle.

4.4 Discussion

4.4.1 Nitrogen metabolism

Nitrogen starvation can influence a cascade of physiological and transcriptomic modifications to ensure survival. *A. gibossum* experienced an up-regulation of genes involved in nitrate assimilation, nitrate reduction and nitrite reduction, indicative of nitrogen metabolism. This is an expected observation as reported in other dinoflagellates and diatoms when exposed to nitrogen starvation (Morey *et al.*, 2011; Bender *et al.*, 2014). However, two nitrogen starvation studies in the dinoflagellates, *Scrippsiella trochoidea* and *Amphidinium carterae* failed to find nitrate and nitrite transporters along with nitrate and nitrite reductases to be differentially expressed using RNAseq analysis (Copper *et al.*, 2016; Lauritano *et al.*, 2017). This contrasted with results obtained from the diatom *Phaeodactylum tricornutum* where these reductases and transporters genes were upregulated (Maheshwari *et al.*, 2010). This analysis in *A. gibossum* has better sequencing resolution and samples were collected only when nitrate concentration in medium was actually undetectable, ensuring nutrient depletion. This is critical because phytoplankton are known to store excess nutrients in vacuoles (Lin *et al.*, 2016). Furthermore, *A. gibossum* appeared to tune its C level and N intake during the starvation stage by downregulating bicarbonate export system (Figure 4.5). Overall, these data indicate that the dinoflagellate *A. gibossum* is capable of incorporating and utilizing several forms of dissolved organic and inorganic nitrogen sources to satisfy its N requirement.

4.4.2 Phosphate metabolism

Phosphate starvation produced a significant transcriptional response with several biological processes being upregulated to deal with the lack of phosphate. The transcriptome data identified the key upregulation of membrane transporters involved the uptake of amino acids, ammonium, dissolved organic phosphate (DOP), metal ions and nitrate. Dissolved

inorganic deficiency can be overcome by the utilization of DOP, which are hydrolyzed to release phosphate (Lin *et al.*, 2016) (Figure 4.5). This suggests that *A. gibossum* is able to utilize various sources of phosphate while downregulating genes involved in bicarbonate export as observed in N starvation.

Key components of the ATP-consuming glycolysis pathway were significantly upregulated, and this is consistent with previous reports in green algae where low inorganic phosphate level activates this pathway (Botha and Turpin, 1990). Additionally, several ribosomal proteins were upregulated since they are involved in ATP-driven protein synthesis to meet the cell demand for metabolism and phosphate uptake. Photosynthetic processes of *A. gibossum* did not suffer from P limitation as shown in Figure 4.5. On the contrary, the carbon-fixing potential increased with several plastid key components being upregulated including phosphate transporters. This increase may be necessary to refuel the increased cellular processes as observed in the alga, *Prymnesium parvum* (Liu *et al.*, 2015).

4.4.3 Secondary metabolism during nutrient starvation

Toxin content in dinoflagellates is known to alter when subjected to different environmental parameters such as light, temperature, salinity, N or P variations (Han *et al.*, 2016). The present analysis shows that genes involved in secondary metabolite biosynthesis are upregulated in *A. gibossum* only when subjected to P starvation (Figure 4.3c). The end product of glycolysis, pyruvate is converted to acetyl-CoA, which is one of the starter groups for polyketide biosynthesis (Hopwood *et al.*, 2004). The present analysis revealed that PKS genes expression is higher under P starvation in contrast to N starvation, and this is consistent with other reports that phosphate limited cells have higher toxicity than replete cells (Frangopulos *et al.*, 2004; Liu *et al.*, 2015; Han *et al.*, 2016; Hii *et al.*, 2016). On the other hand, a previous study found that N-starvation results in an increase in brevetoxin in *Karenia brevis* (Hardison

et al., 2012). No significantly increased expression in *PKS* genes under N-limitation was reported in *Amphidinium carterae* (Lauritano *et al.*, 2017). The observed *PKS* gene expression under P starvation can be explained by the evolutionary theory, which predicts that microalgal growth slows under nutrient limitation as cells tend to divert greater carbon resources to defense mechanisms (Ianora *et al.*, 2006). The increased photosynthetic activity observed during P starvation in *A. gibossum* would be a coordinated physiological response to provide energy necessary for secondary metabolite biosynthesis.

4.4.4 *Amphidinium gibossum* RNAi pathway and its role in nutrient starvation

RNAi pathway and miRNAs were reported in at least five eukaryotic lineages (Cock *et al.*, 2017) including some dinoflagellate species (Baumgarten *et al.*, 2013; Gao *et al.*, 2013; Lin *et al.*, 2015; Geng *et al.*, 2015; Dagenais-Bellefeuille *et al.*, 2017). I identified 1 Dicer and 2 Argonaute homologs from the transcriptome data; the two RNase III domains needed in cleavage of guide-passenger duplex along with the PAZ domain in Dicer (Zhang *et al.*, 2004). For Argonaute protein, a Piwi and PAZ domains were conserved with other organisms. Additionally, one homolog of the small RNA 2'-O-methyltransferase (HEN1) was identified, which is required for final maturation of a subclass of small RNAs (Huang *et al.*, 2009). The presence and conservation of the core proteins in Dicer, Argonaute and HEN1 suggest the presence of a functional RNAi machinery in *A. gibossum* (Figure 4.6).

GO, KEGG, and PFAM enrichment analysis of the potential targets of differentially expressed miRNAs was conducted to gain better insights into possible post-transcriptional effect during nutrient stress. During N starvation, KEGG enrichment of the target of the miRNA, bdi-miR7721-5p, was the pyruvate metabolism (38.4-fold enrichment, $p < 0.001$, Fisher Exact test). This would directly affect the production of acetyl-CoA, thereby secondary metabolism. Involvement of miRNAs affecting secondary metabolites biosynthesis were

reported previously (Biswas *et al.*, 2016). On the other hand, during phosphate starvation, PFAM enrichment of target genes showed leucine rich repeat, cyclic nucleotide-binding domain, ion transport protein, protein kinase domain, and PPR repeat among the top five protein families (> 2 -fold enrichment, $p < 0.001$, Fisher Exact test). The leucine rich repeat family was the most enriched (4-fold) and this protein family is known to be involved in post-transcriptional regulation (Shivaprasad *et al.*, 2012). Contrasting results have shown that RNA recognition motif 2 (RRM-2) is the most enriched motif in miRNA target genes (8.93-fold) in *Prorocentrum donghaiense* under phosphate limitation (Shi *et al.* 2017).

Overall, these results suggest that miRNAs are regulating the expression of certain genes involved in secondary metabolism under N stress; however, metabolite profiling didn't reveal any drastic reduction in metabolite production. It is possible that these genes are involved in the biosynthesis of a smaller set of metabolites, that are derived in minor fractions compared to the most predominant molecules.

5 Conclusion

5.1 Symbiodiniaceae genomes generate chemical diversity by expanding its secondary metabolism genes

I surveyed three Symbiodiniaceae genomes for genes involved in secondary metabolism. *PKS* genes were more expanded than *NRPS* genes and multiple evolutionary processes contributed to this expansion. Additionally, these genes exhibit a degree of substrate specificity and flexibility that are evolutionarily preserved, irrespective of host. These results demonstrate that these genomes are equipped to generate chemical diversity for secondary metabolite biosynthesis. The present comparative genomic approach provides insights into a secondary metabolism code in the late-diverging Symbiodiniaceae dinoflagellates that may reflect the different adaptations to their environment. However, such a secondary metabolic code needs to be tested by integrating new dinoflagellates genomes in future.

5.2 *A. gibossum* genome illuminates conserved secondary metabolism in dinoflagellates

In order to investigate the degree of conservation of secondary metabolism among dinoflagellates, the genome of the basal dinoflagellate *A. gibossum* was decoded. Till date, only Symbiodiniaceae have been sequenced and *A. gibossum* genome provides new insights into dinoflagellate biology. The assembly was 7.0 Gb in size and predicted to have 85,139 genes with intron length seven times longer than those reported in previously sequenced Symbiodiniaceae genomes. Despite being symbiotic and mixotrophic, two very different modes of nutrition, our data suggested that secondary metabolism machinery is conserved between Symbiodiniaceae and *A. gibossum*. However, the metabolites synthesized are unique, indicating their potential roles in maintaining their symbiotic lifestyle.

5.3 Transcriptome approaches to understand *A. gibossum* secondary metabolism

The present study reveals that the metabolic responses to nitrogen and phosphate starvation in the dinoflagellate *A. gibossum* are regulated both at transcriptional and post-transcriptional stages. The integrated omics approach is powerful to provide new insights into how nutrient stress affect metabolic and cellular processes. Such stress involves the upregulation of a group of ion transporters while downregulation of the release of important ions like bicarbonate. The role of miRNAs in regulating the production of secondary metabolites under nitrogen stress is first demonstrated here in dinoflagellates. The present study improved on previous reports in providing replicates, increased sequencing depth and integrated the genome, transcriptome, microRNAome, and metabolome. Incorporating proteome data will provide a clearer picture into the regulation and biosynthesis of secondary metabolites in dinoflagellates.

5.4 Concluding remarks

Decoding new dinoflagellate genomes and transcriptomes enhanced resources to understand dinoflagellate genomics, which is still in its infancy. Insights from dinoflagellate genomes can reveal key milestones in eukaryote evolution. The genomic data on dinoflagellate secondary metabolites and the biosynthetic pathways, which are highly complex can possibly pave the way for specific markers for algal blooms.

References

- Alexa A, Rahnenfuhrer J (2010) topGO: Enrichment analysis for Gene Ontology. R package version 2.22.0.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215, 403-410.
- Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- Aranda, M., Li, Y., Liew, Y. J., Baumgarten, S., Simakov, O., Wilson, M. C., *et al.* (2016). Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle. *Sci Rep*, 6, 39734.
- Armenteros, J.J.A., Sønderby, C.K., Sønderby, S.K., Nielsen, H., Winther, O. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* 33, 3387-3395.
- Bachvaroff, T. R., & Place, A. R. (2008). From Stop to Start: Tandem Gene Arrangement, Copy Number and Trans-Splicing Sites in the Dinoflagellate *Amphidinium carterae*. *PLoS One*, 3,e2929.
- Baig, H.S., Saifullah, S.M. & Dar, A. (2006). Occurrence and toxicity of *Amphidinium carterae* Hulburt in the North Arabian Sea. *Harmful Algae*, 5, 133-140.
- Baranašić, D. *et al.* (2014). Predicting substrate specificity of adenylation domains of nonribosomal peptide synthetases and other protein properties by latent semantic indexing. *J Ind Microbiol Biotechnol*. 41, 461-467.
- Baumgarten, S., Bayer, T., Aranda, M., Liew, Y. J., Carr, A., Micklem, G., & Voolstra, C. R. (2013). Integrating microRNA and mRNA expression profiling in *Symbiodinium microadriaticum*, a dinoflagellate symbiont of reef-building corals. *BMC Genomics*, 14, 704.

- Bayer, T., Aranda, M., Sunagawa, S., Yum, L. K., DeSalvo, M. K., Lindquist, E., *et al.* (2012). Symbiodinium Transcriptomes: Genome Insights into the Dinoflagellate Symbionts of Reef-Building Corals. *PLoS One*, 7,e35269.
- Beedessee, G., Hisata, K., Roy, M. C., Satoh, N., & Shoguchi, E. (2015). Multifunctional polyketide synthase genes identified by genomic survey of the symbiotic dinoflagellate, *Symbiodinium minutum*. *BMC Genomics*, 16, 941.
- Beedessee, G., Hisata, K., Roy, M. C., Van Dolah, F. M., Satoh, N., & Shoguchi, E. (2019). Diversified secondary metabolite biosynthesis gene repertoire revealed in symbiotic dinoflagellates. *Scientific Reports*, 9:1204
- Bender, S. J., Durkin, C. A., Berthiaume, C. T., Morales, R. L., & Armbrust, E. V. (2014). Transcriptional responses of three model diatoms to nitrate limitation of growth. *Frontiers in Marine Science*, 1,3
- Bennett, V., Baines, A.J. (2001). Spectrin and ankyrin-based pathways: metazoan inventions for integrating cells into tissues. *Physiol Rev.* 81,1353-1392.
- Bhattacharya, D., Yoon, H. S., Hedges, S. B., & Hackett, J. D. (2007). Eukaryotes (Eukaryota). In S. B. Hedges (Ed.), *In The Timetree of Life*: Oxford Univ. Press. Pp 116-120.
- Berdieva, M., Pozdnyakov, I., Matantseva, O., Knyazev, N., Skarlato, S. (2018). Actin as a cytoskeletal basis for cell architecture and a protein essential for ecdysis in *Prorocentrum minimum* (Dinophyceae, Prorocentrales). *Phycol Res.* 66, 127-136.
- Biswas, S., Hazra, S., & Chattopadhyay, S. (2016). Identification of conserved miRNAs and their putative target genes in *Podophyllum hexandrum* (Himalayan Mayapple). *Plant Gene*, 6, 82-89.
- Blatch, G. L., & Lassel, M. (1999). The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. *Bioessays*, 21, 932-939.

- Blin, K. *et al.* (2017). antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucl Acids Res.* 45, W36-W41.
- Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., & Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, 27, 578-579.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, 30, 2114-2120.
- Botha, F.C., & Turpin, D.H. (1990). Molecular, Kinetic, and Immunological Properties of the 6-Phosphofructokinase from the Green Alga *Selenastrum minutum*. *Plant Physiol*, 93, 871-879.
- Bouligand, Y., & Norris, V. (2001). Chromosome separation and segregation in dinoflagellates and bacteria may depend on liquid crystalline states. *Biochimie*, 83, 187-192.
- Bushley, K.E. & Turgeon, B.G. (2010). Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. *BMC Evol Biol.* 10, 26.
- Caboche, S., Pupin, M., Leclère, V., Fontaine, A., Jacques, P., Kuchero, G. (2008). NORINE: a database of nonribosomal peptides. *Nucl Acids Res.* 36, D326-D331.
- Caffrey, P. (2003). Conserved amino acid residues correlating with ketoreductase stereospecificity in modular polyketide synthases. *ChemBioChem* 4, 654-657.
- Chakraborty, T. & Das, S. (2001). Chemistry of Potent Anti-Cancer Compounds, Amphidinolides. *Current Medicinal Chemistry-Anti-Cancer Agents*, 1, 131-149.
- Cheng, Y.Q., Tang, G.L., Shen, B. (2003). Type I polyketide synthase requiring a discrete acyltransferase for polyketide biosynthesis. *Proc Natl Acad Sci USA* 100, 3149-3154.
- Chow, M. H., Yan, K. T., Bennett, M. J., & Wong, J. T. (2010). Birefringence and DNA condensation of liquid crystalline chromosomes. *Eukaryot Cell*, 9, 1577-1587.
- Clamp, M., Cuff, J., Searle, S. M., & Barton, G. J. (2004). The Jalview Java alignment editor. *Bioinformatics*, 20, 426-427.

- Cock, J. M., Liu, F., Duan, D., Bourdareau, S., Lipinska, A. P., Coelho, S. M., & Tarver, J. E. (2017). Rapid Evolution of microRNA Loci in the Brown Algae. *Genome Biol Evol*, 9, 740-749.
- Coffroth, M. A., & Santos, S. R. (2005). Genetic Diversity of Symbiotic Dinoflagellates in the Genus Symbiodinium. *Protist*, 156, 19-34.
- Colcombet, J. *et al.* (2013). Systematic study of subcellular localization of *Arabidopsis* PPR proteins confirms a massive targeting to organelles. *RNA Biol*. 10, 1557-1575.
- Cook, A., Bono, F., Jinek, M., Conti, E. (2007). Structural biology of nucleocytoplasmic transport. *Annu Rev Biochem*. 76, 647-671.
- Cooper, J. T., Sinclair, G. A., & Wawrik, B. (2016). Transcriptome Analysis of *Scrippsiella trochoidea* CCMP 3099 Reveals Physiological Changes Related to Nitrate Depletion. *Front Microbiol*, 7, 639.
- Dagenais-Bellefeuille, S., Beauchemin, M., & Morse, D. (2017). miRNAs Do Not Regulate Circadian Protein Synthesis in the Dinoflagellate *Lingulodinium polyedrum*. *PLoS One*, 12, e0168817.
- Dagenais-Bellefeuille, S., & Morse, D. (2013). Putting the N in dinoflagellates. *Front Microbiol*, 4:369
- Daugbjerg, N., Hansen, G., Larsen, J., & Moestrup, O. (2000). Phylogeny of some of the major genera of dinoflagellates based on ultrastructure and partial LSU rDNA sequence data, including the erection of three new genera of unarmoured dinoflagellates. *Phycologia*, 39, 302-317.
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., *et al.* (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29, 15-21.
- Donadio, S., Monciardini, P., Sosio, M. (2007). Polyketide synthases and nonribosomal peptide synthetases: The emerging view from bacterial genomics. *Nat Prod Rep*. 24, 1073-1109.

- Doyle, J.J., & Doyle, J.L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, 19, 11-15.
- Du, L., Sánchez, C., Shen, B. Hybrid peptide-polyketide natural product: biosynthesis and prospects toward engineering novel molecules. *Metab Eng.* 3, 78-95.
- Dutta, S., Whicher, J.R., Hansen, D.A., Hale, W.A., Chemler, J.A., Congdon, G.R. *et al.* (2014). Structure of a modular polyketide synthase. *Nature* 510 (7506), 512-517.
- Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl Acids Res.* 32, 1792-1797.
- Eichholz, K., Beszteri, B., & John, U. (2012). Putative Monofunctional Type I Polyketide Synthase Units: A Dinoflagellate-Specific Feature? *PLoS One*, 7, e48624.
- Emanuelsson, O., Nielsen, H., von Heijne, G. (1999). ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.* 8, 978-984.
- Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc.* 2, 953-971.
- Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., & Marks, D. S. (2003). MicroRNA targets in *Drosophila*. *Genome Biol*, 5(R1).
- Erdner, D.L., & Anderson, D. M. (2006). Global transcriptional profiling of the toxic dinoflagellate *Alexandrium fundyense* using Massively Parallel Signature Sequencing. *BMC Genomics*, 7, 88.
- Finn, R.D., Clements, J., & Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(suppl), W29-W37.
- Finn, R.D. & Jones, C.G. (2003). Natural products-a simple model to explain chemical diversity. *Nat Prod Rep.* 20, 382-391.
- Fischbach, M.A., Walsh, C.T., Clardy, J. (2008). The evolution of gene collectives: How natural selection drives chemical innovation. *Proc Natl Acad Sci USA* 105, 4601-8.

- Frangópulos, M., Guisande, C., deBlas, E., & Maneiro, I. (2004). Toxin production and competitive abilities under phosphorus limitation of *Alexandrium* species. *Harmful Algae*, 3, 131-139.
- Friedländer, M.R., Mackowiak, S.D., Li, N., Chen, W., & Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucl Acids Res.* 40, 37-52.
- Fujii, S. & Small, I. (2011). The evolution of RNA editing and pentatricopeptide repeat genes. *New Phytologist* 191, 37-47
- Fukatsu, T., Onodera, K., Ohta, Y., Oba, Y., Nakamura, H., Shintani, T., *et al.* (2007). Zooxanthellamide D, a Polyhydroxy Polyene Amide from a Marine Dinoflagellate, and Chemotaxonomic Perspective of the Symbiodinium Polyols. *J Nat Prod*, 70,407-11.
- Galluzzi, L., Bertozzini, E., Penna, A., Perini, F., Garcés, E., & Magnani, M. (2009). Analysis of rRNA gene content in the Mediterranean dinoflagellate *Alexandrium catenella* and *Alexandrium taylori*: implications for the quantitative real-time PCR-based monitoring methods. *Journal of Applied Phycology*, 22, 1-9.
- Gao, D., Qiu, L., Hou, Z., Zhang, Q., Wu, J., Gao, Q., & Song, L. (2013). Computational Identification of MicroRNAs from the Expressed Sequence Tags of Toxic Dinoflagellate *Alexandrium Tamarense*. *Evolutionary Bioinformatics*, 9,479-485.
- Gao, F. & Zhang, C.T. (2006). GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. *Nucl Acids Res.* 34, W686-W691,
- Gárate-Lizárraga, I. (2012). Proliferation of *Amphidinium carterae* (GYMNODINIALES: GYMNODINIACEAE) in Bahia De La Paz, Gulf of California. *CICIMAR Oceanides*, 27, 37-49.

- Geng, H., Sui, Z., Zhang, S., Du, Q., Ren, Y., Liu, Y., *et al.* (2015). Identification of microRNAs in the Toxigenic Dinoflagellate *Alexandrium catenella* by High-Throughput Illumina Sequencing and Bioinformatic Analysis. *PLoS One*, 10, e0138709.
- Gonzalez-Pech, R. A., Ragan, M. A., & Chan, C. X. (2017). Signatures of adaptation and symbiosis in genomes and transcriptomes of Symbiodinium. *Sci Rep*, 7, 15021.
- Gordon, B.R. & Leggat, W. (2010). *Symbiodinium*-Invertebrate symbioses and the role of metabolomics. *Mar Drugs* 8, 2546-2568.
- Gornik, S.G. *et al.* (2015). Endosymbiosis undone by stepwise elimination of the plastid in a parasitic dinoflagellate. *Proc Natl Acad Sci USA* 112, 5767-5772.
- Guillebault, D., Sasorith, S., Derelle, E., Wurtz, J. M., Lozano, J. C., Bingham, S., *et al.* (2002). A new class of transcription initiation factors, intermediate between TATA box-binding proteins (TBPs) and TBP-like factors (TLFs), is present in the marine unicellular organism, the dinoflagellate *Cryptothecodinium cohnii*. *J Biol Chem*, 277, 40881-40886.
- Haapala, O.K & Soyer, M-O. (1873). Structure of Dinoflagellate Chromosomes. *Nature New Biol.*, 244, 195-197.
- Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., *et al.* (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc.* 8, 1494-512.
- Hackett, J.D., Anderson, D.M., Erdner, D.L., & Bhattacharya, D. (2004). Dinoflagellates: A remarkable evolutionary experiment. *American Journal of Botany*, 91, 1523-1534.
- Hackett, J.D., Yoon, H.S., Butterfield, N.J., Sanderson, M.J., & Bhattacharya, D. (2007). Plastid endosymbiosis: sources and timing of the major events. In A. K. Falkowski (Ed.), *Evolution of Primary Producers in the Sea*. pp 109-132

- Han, K., Lee, H., Anderson, D.M., & Kim, B. (2016). Paralytic shellfish toxin production by the dinoflagellate *Alexandrium pacificum* (Chinhae Bay, Korea) in axenic, nutrient-limited chemostat cultures and nutrient-enriched batch cultures. *Mar Pollut Bull*, *104*, 34-43.
- Hannah, M.A. *et al.* (2010). Combined transcript and metabolite profiling of *Arabidopsis* grown under widely variant growth conditions facilitates the identification of novel metabolite-mediated regulation of gene expression. *Plant Physiol.* *152*, 2120-2129.
- Hardison, R., Sunda, W.G., Wayne, L.R., Shea, D., & Tester, P.A. (2012). Nitrogen Limitation increase brevetoxins in *Karenia brevis* (Dinophyceae): Implications for bloom toxicity. *J. Phycol.* *48*, 844-58.
- Harlow, L. D., Koutoulis, A., & Hallengraeff, G. M. (2007). S-adenosylmethionine synthetase genes from eleven marine dinoflagellates. *Phycologia*, *46*, 46-53.
- He, J., Yang, Y., Xu, H., Zhang, X., & Li, X. M. (2005). Olanzapine attenuates the okadaic acid-induced spatial memory impairment and hippocampal cell death in rats. *Neuropsychopharmacology*, *30*, 1511-1520.
- Hehenberger, E., Burki, F., Kolisko, M., Keeling, P.J. (2016). Functional Relationship between a Dinoflagellate Host and Its Diatom Endosymbiont. *Mol Biol Evol.* *33*, 2376-2390,
- Hertweck, C. (2009). The Biosynthetic Logic of Polyketide Diversity. *Angew Chem Int Ed* *48*, 4688-4716.
- Hii, K. S., Lim, P. T., Kon, N. F., Takata, Y., Usup, G., & Leaw, C. P. (2016). Physiological and transcriptional responses to inorganic nutrition in a tropical Pacific strain of *Alexandrium minutum*: Implications for the saxitoxin genes and toxin production. *Harmful Algae*, *56*, 9-21.
- Hoff, K.J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, *32*, 767-769.

- Hofmann, E., Wrench, P. M., Sharples, F. P., Hiler, R. G., Welte, W., & Diederichs, K. (1996). Structural Basis of Light Harvesting by Carotenoids: Peridinin-Chlorophyll-Protein from *Amphidinium carterae*. *Science*, 272, 1788-1791.
- Hoppenrath, M., & Leander, B.S. (2010). Dinoflagellate phylogeny as inferred from heat shock protein 90 and ribosomal gene sequences. *PLoS One*, 5, e13220.
- Hopwood, D. A. (2004). Cracking the Polyketide Code. *PLoS Biology*, 2, e35.
- Hou, Y. & Lin, S. (2009). Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. *PLoS One*, 4, e6978.
- Howe, C.J., Nisbet, R.E., & Barbrook, A.C. (2008). The remarkable chloroplast genome of dinoflagellates. *J Exp Bot*, 59, 1035-1045.
- Huang, D., Sherman, B.T., Tan, Q., Collins, J.R., Alvord, W.G., Roayaei, J., et al. (2007). The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology*, 8, R183.
- Huang, Y., Ji, L., Huang, Q., Vassilyev, D. G., Chen, X., & Ma, J. B. (2009). Structural insights into mechanisms of the small RNA methyltransferase HEN1. *Nature*, 461, 823-827.
- Ianora, A., Boersma, M., Cassoti, R., Fontana, A., Harder, J., Hoffmann, F., et al. (2006). New trends in marine chemical ecology. *Estuaries and Coasts*, 29, 531-551.
- Jackson, C.J., Norman, J.E., Schnare, M.N., Gray, M.W., Keeling, P.J., & Waller, R. F. (2007). Broad genomic and transcriptional analysis reveals a highly derived genome in dinoflagellate mitochondria. *BMC Biol*, 5, 41.
- Janouškovec, J., Gavelis, G. S., Burki, F., Dinh, D., Bachvaroff, T. R., Gornik, S. G., et al. (2017). Major transitions in dinoflagellate evolution unveiled by phylotranscriptomics. *Proc Natl Acad Sci USA*, 114, E171-E180.
- Jenke-Kodama, H., Sandmann, A., Müller, R., Dittmann, E. (2005). Evolutionary implications of bacterial polyketide synthases. *Mol Biol Evol*. 22, 2027-2039.

- Johnson, J.G., Morey, J.S., Neely, M.G., Ryan, J.C., & Van Dolah, F.M. (2012). Transcriptome remodeling associated with chronological aging in the dinoflagellate, *Karenia brevis*. *Mar Genomics*, 5, 15-25.
- Jones, A.C., Monroe, E.A., Eisman, E.B., Gerwick, L., Sherman, D.H., & Gerwick, W. H. (2010). The unique mechanistic transformations involved in the biosynthesis of modular natural products from marine cyanobacteria. *Natural Product Reports*, 27, 1048-65.
- Jones, C.G. & Firn, R.D. (1991). On the evolution of plant secondary metabolite chemical diversity. *Phil Trans R Soc Lond B* 333, 273-280.
- Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., *et al.* (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30, 1236-1240.
- Jørgensen, M. F., Murray, S., & Daugbjerg, N. (2004). Amphidinium Revisited. I. Redefinition of Amphidinium (Dinophyceae) Based on Cladistic and Molecular Phylogenetic Analyses1. *Journal of Phycology*, 40, 351-365.
- Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., *et al.* (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res*, 24, 1384-1395.
- Kamikawa, R., Nishimura, H. & Sako, Y. (2009). Analysis of the mitochondrial genome, transcripts, and electron transport activity in the dinoflagellate *Alexandrium catenella* (Gonyaulacales, Dinophyceae). *Phycological Research*, 57, 1-11.
- Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res*, 40(Database issue), D109-114.
- Karafas, S., Teng, S. T., Leaw, C. P., & Alves-de-Souza, C. (2017). An evaluation of the genus Amphidinium (Dinophyceae) combining evidence from morphology, phylogenetics, and toxin production, with the introduction of six novel species. *Harmful Algae*, 68, 128-151.

- Karl, D. M. (2014). Microbially mediated transformations of phosphorus in the sea: new views of an old cycle. *Ann Rev Mar Sci*, 6, 279-337.
- Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., *et al.* (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol*, 12, e1001889.
- Kellmann, R., Stüken, A., Orr, R. J. S., Svendsen, H. M., & Jakobsen, K. S. (2010). Biosynthesis and Molecular Genetics of Polyketides in Marine Dinoflagellates. *Marine Drugs*, 8, 1011-1048.
- Kent, W.J. (2002) BLAT- the BLAST-like alignment tool. *Genome Res*. 12,656-64.
- Khayatt, B.I., Overmars, L., Siezen, R.J., Francke, C. (2013). Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using Ensembles of substrate specific Hidden Markov Models. *PLoS ONE* 8, e62136
- Kita, M., Ohishi, N., Washida, K., Kondo, M., Koyama, T., Yamada, K., & Uemura, D. (2005). Symbioimine and neosymbioimine, amphoteric iminium metabolites from the symbiotic marine dinoflagellate Symbiodinium sp. *Bioorganic & Medicinal Chemistry*, 13, 5253-5258.
- Klueter, A., Crandall, J., Archer, F., Teece, M., Coffroth, M. (2015). Taxonomic and environmental variation of metabolite profiles in marine dinoflagellates of the genus *Symbiodinium*. *Metabolites* 5, 74-99.
- Kobayashi, J., & Kubota, T. (2007). Bioactive macrolides and polyketides from marine dinoflagellates of the genus Amphidinium. *J Nat Prod*, 70, 451-460.
- Kobayashi, J., & Tsuda, M. (2004). Amphidinolides, bioactive macrolides from symbiotic marine dinoflagellates. *Nat Prod Rep*, 21, 77-93.
- Kobe, B., & Kajaba, A. V. (2001). The leucine-rich repeat as a protein recognition motif. *Current Opinion in Structural Biology*, 11, 725-732.

- Kohli, G.S., John, U., Figueroa, R. I., Rhodes, L.L., Harwood, D.T., Groth, M., *et al.* (2015). Polyketide synthesis genes associated with toxin production in two species of *Gambierdiscus* (Dinophyceae). *BMC Genomics*, 16, 410.
- Kohli, G.S., Campbell, K., John, U., Smith, K.F., Fraga, S., Rhodes, L.L *et al.* (2017). Role of Modular Polyketide Synthases in the Production of Polyether Ladder Compounds in Ciguatoxin-Producing *Gambierdiscus polynesiensis* and *G. excentricus* (Dinophyceae). *J Eukaryot Microbiol.* 64, 691-706.
- Kohli, G.S., John, U., Van Dolah, F.M., Murray, S.A. (2016). Evolutionary distinctiveness of fatty acid and polyketide synthesis in eukaryotes. *ISME J.* 10, 1877-1890.
- Koyanagi, R. *et al.* (2013). MarinegenomicsDB: An integrated genome viewer for community-based annotation of genomes. *Zool Sci.* 30, 797-800.
- Kroken, S., Glass, N.L., Taylor, J.W., Yoder, O.C., Turgeon, B.G. (2003). Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. *Proc Natl Acad Sci USA* 100, 15670-15675
- Krzywinski, M. *et al.* (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19,1639-1645
- LaJeunesse, T. C., Lambert, G., Andersen, R. A., Coffroth, M. A., & Galbraith, D. W. (2005). Symbiodinium (Pyrrophyta) genome sizes (DNA content) are smallest among dinoflagellates. *Journal of Phycology*, 41, 880-886.
- LaJeunesse, T. C., Parkinson, J. E., Gabrielson, P. W., Jeong, H. J., Reimer, J. D., Voolstra, C. R., & Santos, S. R. (2018). Systematic Revision of Symbiodiniaceae Highlights the Antiquity and Diversity of Coral Endosymbionts. *Current Biology*, 28, 2570-2580.e2576.
- Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10, R25.

- Lauritano, C., De Luca, D., Ferrarini, A., Avanzato, C., Minio, A., Esposito, F., & Ianora, A. (2017). De novo transcriptome of the cosmopolitan dinoflagellate *Amphidinium carterae* to identify enzymes with biotechnological potential. *Sci Rep*, 7, 11701.
- Le, Q.H., Markovic, P., Hastings, J.W., Jovine, R.V., Morse, D. (1997). Structure and organization of the peridinin-chlorophyll a-binding protein gene in *Gonyaulax polyedra*. *Mol Gen Genet*, 255, 595-604.
- Lee, J.J., Olea, R., Cevalco, M., Pochon, X., Correia, M., Shpigel, M., & Pawlowski, J. (2003). A Marine Dinoflagellate, *Amphidinium eilatiensis* n. sp., from the Benthos of a Mariculture Sedimentation Pond in Eilat, Israel. *J Eukaryot Microbiol*, 50, 439-448.
- Lee, R. *et al.* (2014). Analysis of EST data of the marine protist *Oxyrrhis marina*, an emerging model for alveolate biology and evolution. *BMC Genomics* 15, 122.
- Leggett, R. M., Clavijo, B. J., Clissold, L., Clark, M. D., & Caccamo, M. (2014). NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics*, 30, 566-568.
- Lewis, D.H., & Smith, D.C. (1971). The autotrophic nutrition of symbiotic marine coelenterates with special reference to hermatypic corals. I. Movement of photosynthetic products between the symbionts. *Proceedings of the Royal Society B*, 178, 111-129.
- Li, B., & Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, 323.
- Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22, 1658-1659.
- Lidie, K.B. & van Dolah, F.M. (2007). Spliced leader RNA-mediated trans-splicing in a dinoflagellate, *Karenia brevis*. *J Eukaryot Microbiol*, 54, 427-35.
- Lin, S. (2011). Genomic understanding of dinoflagellates. *Research in Microbiology*, 162, 551-569.

- Lin, S., Litaker, R.W., Sunda, W.G., & Wood, M. (2016). Phosphorus physiological ecology and molecular mechanisms in marine phytoplankton. *Journal of Phycology*, 52, 10-36.
- Lin, S., Zhang, H., & Gray, M.W. (2008). RNA editing in dinoflagellates and its implications for the evolutionary history of the editing machinery. In H. Smith (Ed.), *RNA and DNA Editing: Molecular Mechanisms and Their Integration into Biological Systems*, pp 280-309
- Lin, S.J., Cheng, S. F., Song, B., Zhong, X., Lin, X., Li, W. J., *et al.* (2015). The *Symbiodinium kawagutii* genome illuminates dinoflagellate gene expression and coral symbiosis. *Science*, 350, 691-694.
- Liu, H., Stephens, T. G., Gonzalez-Pech, R. A., Beltran, V. H., Lapeyre, B., Bongaerts, P., *et al.* (2018). Symbiodinium genomes reveal adaptive evolution of functions related to coral-dinoflagellate symbiosis. *Commun Biol*, 1, 95.
- Liu, Z., Koid, A. E., Terrado, R., Campbell, V., Caron, D. A., & Heidelberg, K. B. (2015). Changes in gene expression of *Prymnesium parvum* induced by nitrogen and phosphorus limitation. *Front in Microbiol*, 6, 631.
- Lomsadze, A., Burns, P. D., & Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research*, 42, e119-e119.
- Lopes, R.M., Silveira, M. (1994). Symbiosis between a pelagic flatworm and a dinoflagellate from a tropical area: structural observations. *Hydrobiologia*, 287, 277-284.
- Lopez-Gomollon, S., Beckers, M., Rathjen, T., Moxon, S., Maumus, F., Mohorianu, I., *et al.* (2014). Global discovery and characterization of small non-coding RNAs in marine microalgae. *BMC Genomics*, 15,697.
- Lopez-Legentil, S., Song, B., DeTure, M., Baden, D.G. (2010). Characterization and localization of a hybrid non-ribosomal peptide synthetase and polyketide synthase gene from the toxic dinoflagellate *Karenia brevis*. *Marine Biotechnology* 12, 32-41.

- Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., *et al.* (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1,18.
- Magrane, M., & UniProt, C. (2011). UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)*, 2011, bar009.
- Maheswari, U., Jabbari, K., Petit, J.-L., Porcel, B. M., Allen, E. A., Cadoret, J. P., *et al.* (2010). Digital expression profiling of novel diatom transcripts provides insight into their biological functions. *Genome Biol*, 11(R85).
- Marahiel, M. A., Stachelhaus, T., & Mootz, D. H. (1997). Modular Peptide Synthetases Involved in Nonribosomal Peptide Synthesis. *Chem Rev*, 97, 2651-2673.
- Marcais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27 764-770.
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17, 10-12.
- Maruyama, S., Shoguchi, E., Satoh, N., Minagawa, J. (2015). Diversification of the light-harvesting complex gene family via intra-and intergenic duplications in the coral symbiotic alga *Symbiodinium*. *PLoS ONE* 10, e0119406,
- Meyer, J.M., Rödelberger, C., Eichholz, K., Tillmann, U., Cembella, A., McGaughan, A., & John, U. (2015). Transcriptomic characterisation and genomic glimps into the toxigenic dinoflagellate *Azadinium spinosum*, with emphasis on polyketide synthase genes. *BMC Genomics*, 16,27.
- Miranda, K.M., Espey, M.G., & Wink, D.A (2001). A rapid, simple spectrophotometric method for simultaneous detection of nitrate and nitrite. *Nitric Oxide*, 5,62-71
- Monroe, E. A., Johnson, J. G., Wang, Z., Pierce, R. K., & Van Dolah, F. M. (2010). Characterization and Expression of Nuclear-Encoded Polyketide Synthases in the Brevetoxin-Producing Dinoflagellate *Karenia Brevis*. *Journal of Phycology*, 46, 541-552.

- Monroe, E. A., & Van Dolah, F. M. (2008). The toxic dinoflagellate *Karenia brevis* encodes novel type I-like polyketide synthases containing discrete catalytic domains. *Protist*, 159, 471-482.
- Mootz, H.D., Schwarzer, D., Marahiel, M.A. (2002). Ways of assembling complex natural products on modular nonribosomal peptide synthetases. *ChemBioChem* 3, 490-504
- Morey, J.S., Monroe, E.A., Kinney, A.L., Beal, M., Johnson, J.G., Hitchcock, G. L., & Dolah, F. M. V. (2011). Transcriptomic response of the red tide dinoflagellate, *Karenia brevis*, to nitrogen and phosphorus depletion and addition. *BMC Genomics*, 12,346.
- Mosavi, L.K., Cammett, T.J., Desrosiers, D.C., Peng, Z.Y. (2004). The ankyrin repeat as molecular architecture for protein recognition. *Protein Sci.* 13,1435-1448.
- Mount, S.M., Burks, C., Hertz, G., Stormo, G.D., White, O., & C,F. (1992). Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Res*, 20, 4255-4262.
- Moustafa, A., Evans, A.N., Kulis, D.M., Hackett, J.D., Erdner, D.L., Anderson, D.M., & Bhattacharya, D. (2010). Transcriptome profiling of a toxic dinoflagellate reveals a gene-rich protist and a potential impact on gene expression due to bacterial presence. *PLoS One*, 5, e9688.
- Murray, S., & Patterson, D. J. (2002). The benthic dinoflagellate genus *Amphidinium* in south eastern Australian waters including three new species. *European Journal of Phycology*, 37, 279-298.
- Murray, S. (2003). Diversity and phylogenetics of sand-dwelling dinoflagellates. Ph.D. Thesis, University of Sydney.
- Murray, S., Flø Jørgensen, M., Daugbjerg, N., & Rhodes, L. (2004). *Amphidinium* Revisited. II. Resolving Species Boundaries in the *Amphidinium Operculatum* Species Complex (Dinophyceae), Including the Descriptions of *Amphidinium Trulla* Sp. Nov. And *Amphidinium Gibbosum*. Comb. *Journal of Phycology*, 40, 366-382.

- Murray, S., Flø Jørgensen, M., Ho, S.Y., Patterson, D.J., & Jermini, L.S. (2005). Improving the analysis of dinoflagellate phylogeny based on rDNA. *Protist*, 156, 269-86.
- Murray, S. A., Garby, T., Hoppenrath, M., & Neilan, B. A. (2012). Genetic Diversity, Morphological Uniformity and Polyketide Production in Dinoflagellates (Amphidinium, Dinoflagellata). *PLoS One*, 7, e38253.
- Murray, S. A., Suggett, D. J., Doblin, M. A., Kohli, G. S., Seymour, J. R., Fabris, M., & Ralph, P. J. (2016). Unravelling the functional genetics of dinoflagellates: a review of approaches and opportunities. *Perspectives in Phycology*, 3, 37-52.
- Murray, S.A. *et al.* (2016). Unravelling the functional genetics of dinoflagellates: A review of approaches and opportunities. *Perspect Phycol.* 3, 37-52.
- Nagai, H., Torigue, K., Satake, M., Murata, M., Yasumoto, T., Hirota, H. (1992). Gambieric acids: unprecedented potent antifungal substances isolated from cultures of a marine dinoflagellate *Gambierdiscus toxicus*. *J. Am. Chem. Soc.*, 114, 1102-1103.
- Nagai, H., Mikami, Y., Yazawa, K., & Gono, T. (1993). Biological activities of novel polyether antifungals, gambieric acids A and B from a marine dinoflagellate *Gambierdiscus toxicus*. *Journal of Antibiotics*, 46, 520-522.
- Nakamura, H., Asari, T., Fujimaki, K., Maruyama, K., & Murai, A. (1995a). Zooxanthellatoxin-B, Vasoconstrictive Congener of Zooxanthellatoxin A from a Symbiotic Dinoflagellate *Symbiodinium sp.* *Tetrahedron Letters*, 36, 7255-7258.
- Nakamura, H., Asari, T., Murai, A. (1995b). Zooxanthellatoxin-A, a Potent Vasoconstrictive 62- Membered Lactone from a Symbiotic Dinoflagellate. *J. Am. Chem. Soc.*, 117, 550-551.
- Nakamura, H., Kawase, Y., Maruyama, K., Murai, A. (1998). Studies on Polyketide Metabolites of a Symbiotic Dinoflagellate, *Symbiodinium sp.* A New C30 Marine Alkaloid, Zooxanthellamine, a Plausible Precursor for Zoanthid Alkaloids. *Bull Chem Soc Jpn* 71, 781-787.

- Nakamura, T., Yagi, Y., Kobayashi, K. (2012). Mechanistic Insight into Pentatricopeptide Repeat Proteins as Sequence-Specific RNA-Binding Proteins for Organellar RNAs in Plants. *Plant Cell Physiol.* 53, 1171-1179.
- Nash, E.A., Barbrook, A.C., Edwards-Stuart, R.K., Bernhardt, K., Howe, C.J., & Nisbet, R.E. (2007). Organization of the mitochondrial genome in the dinoflagellate *Amphidinium carterae*. *Mol Biol Evol*, 24, 1528-1536.
- Nei, M., Xu, P., & Glazko, G. (2001). Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proc Natl Acad Sci USA*, 98, 2497-2502.
- Nguyen, T. *et al.* (2008) Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. *Nat Biotechnol.* 26, 225-233.
- Onodera, K., Nakamura, H., Oba, Y., & Ojika, M. (2004). Zooxanthellamide B, a novel large polyhydroxy metabolite from a marine dinoflagellate of *Symbiodinium* sp. *Biosci Biotechnol Biochem*, 68, 955-958.
- Parfrey, L. W., Lahr, D. J. G., Knoll, A. H., & Katz, L. A. (2011). Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc Natl Acad Sci USA*, 108, 13624-13629.
- Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23, 1061-1067.
- Parsons, T. R. (1984). *A Manual of Chemical & Biological Methods for Seawater Analysis*. New York, NY: Pergamon Press.
- Pawlowicz, R., Morey, J. S., Darius, H. T., Chinain, M., & Van Dolah, F. M. (2014). Transcriptome sequencing reveals single domain Type I-like polyketide synthases in the toxic dinoflagellate *Gambierdiscus polynesiensis*. *Harmful Algae*, 36, 29-37.

- Perry, M. J. (1976). Phosphate utilization by an oceanic diatom in phosphorus-limited chemostat culture and in the oligotrophic waters of the central North Pacific. *Limnology and Oceanography*, 21, 88-107.
- Piel, J. (2002). A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of *Paederus* beetles. *Proc Natl Acad Sci USA* 99, 14002-14007
- Piel, J., Hui, D., Fusetani, N., Matsunaga, S. (2004). Targeting modular polyketide synthases with iteratively acting acyltransferases from metagenomes of uncultured bacterial consortia. *Environ Microbiol.* 6, 921-927.
- Pochon, X., Gates, R.D. (2010). A new *Symbiodinium* clade (Dinophyceae) from soritid foraminifera in Hawai'i. *Molecular Phylogenetics and Evolution* 56,492-497.
- Price, A.L., Jones, N.C., & Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics*, 21 Suppl 1, i351-358.
- Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., *et al.* (2012). The Pfam protein families database. *Nucleic Acids Res*, 40(Database issue), D290-301.
- Rae, P. M. (1976). 5-Hydroxymethyluracil in the DNA of a Dinoflagellate. *Science*, 194, 1062-1064.
- Rae, P.M.M., & Steele, R.E. (1978). Modified bases in the DNAs of unicellular eukaryotes: An examination of distributions and possible roles, with emphasis on hydroxymethyluracil in dinoflagellates. *BioSystems*, 10, 37-53.
- Rausch, C., Hoof, I., Weber, T., Wohlleben, W., Huson, D.H. (2007). Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol Biol.* 7, 78
- Reichman, J.R., Wilcox, T.P., & Vize, P.D. (2003). PCP gene family in *Symbiodinium* from *Hippopus hippopus*: low levels of concerted evolution, isoform diversity, and spectral tuning of chromophores. *Mol Biol Evol*, 20, 2143-2154.

- Rein, K.S., & J.B.(1999). Polyketides from dinoflagellates: origins, pharmacology and biosynthesis. *Comparative Biochemistry and Physiology Part B*, 124, 117-131.
- Rein, K.S., & Snyder, R.V. (2006). The Biosynthesis of Polyketide Metabolites by Dinoflagellates. *Advances in Applied Microbiology Volume 59* (pp. 93-125).
- Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S. *et al.* (2012). MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Syst Biol.* 61, 539-542.
- Rix, U., Fischer, C., Remsing, L.L., Rohr, J. (2002). Modification of post-PKS tailoring steps through combinatorial biosynthesis. *Natural Product Reports* 19,542-580.
- Rizzo, P.J., & Nooden, L.D. (1972). Chromosomal Proteins in the Dinoflagellate Alga *Gyrodinium cohnii*. *Science*, 176, 796-797.
- Rizzo, P.J. (2003). Those amazing dinoflagellate chromosomes. *Cell Research*, 13, 215-217.
- Rosic, N. *et al.* Unfolding the secrets of coral-algal symbiosis. (2015). *ISME J.* 9, 844-856
- Salcedo, T., Upadhyay, R. J., Nagasaki, K., & Bhattacharya, D. (2012). Dozens of toxin-related genes are expressed in a nontoxic strain of the dinoflagellate *Heterocapsa circularisquama*. *Mol Biol Evol*, 29, 1503-1506.
- Salois, P. & Morse, D. (1997). Characterisation and molecular phylogeny of a protein kinase cDNA from the dinoflagellate *Gonyaulax* (Dinophyceae). *J Phycol*, 33, 1063-1072.
- Schindelin, J. *et al.* (2012) Fiji: an open-source platform for biological-image analysis. *Nature methods* 9, 676-682.
- Schwarzer, D., Finking, R., & Marahiel, M. A. (2003). Nonribosomal peptides: from genes to products. *Natural Product Reports*, 20, 275-287.
- Seeber, F.& Soldati-Favre, D. (2010). Metabolic pathways in the apicoplast of apicomplexa. *Int Rev Cell Mol Biol.* 281,161-228.

- Shelest, E., Heimerl, N., Fichtner, M., Sasso, S. Multimodular type I polyketide synthases in algae evolve by module duplications and displacement of AT domains *in trans*. *BMC Genomics* 16, 1015.
- Shemi, A., Schatz, D., Fredricks, H. F., Van Mooy, B. A., Porat, Z., & Vardi, A. (2016). Phosphorus starvation induces membrane remodeling and recycling in *Emiliania huxleyi*. *New Phytol*, 211, 886-898.
- Shen, B. (2003). Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Current Opinion in Chemical Biology*, 7, 285-295.
- Shi, X., Lin, X., Li, L., Li, M., Palenik, B., & Lin, S.(2017). Transcriptomic and microRNAomic profiling reveals multi-faceted mechanisms to cope with phosphate stress in a dinoflagellate. *The ISME Journal*, 11(10), 2209-2218.
- Shivaprasad, P.V., Chen, H.M., Patel, K., Bond, D.M., Santos, B.A., & Baulcombe, D.C. (2012). A microRNA superfamily regulates nucleotide binding site-leucine-rich repeats and other mRNAs. *Plant Cell*, 24, 859-874.
- Shoguchi, E., Beedessee, G., Tada, I., Hisata, K., Kawashima, T., Takeuchi, T., *et al.* (2018). Two divergent *Symbiodinium* genomes reveal conservation of a gene cluster for sunscreen biosynthesis and recently lost genes. *BMC Genomics*, 19, 458.
- Shoguchi, E., Shinzato, C., Kawashima, T., Gyoja, F., Mungpakdee, S., Koyanagi, R., *et al.* (2013). Draft Assembly of the *Symbiodinium minutum* Nuclear Genome Reveals Dinoflagellate Gene Structure. *Current Biology*, 23, 1399-1408.
- Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., *et al.* (2014). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7, 539-539.

- Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., & Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31, 3210-3212.
- Slamovits, C.H., Saldarriaga, J.F., Larocque, A., & Keeling, P.J. (2007). The highly reduced and fragmented mitochondrial genome of the early-branching dinoflagellate *Oxyrrhis marina* shares characteristics with both apicomplexan and dinoflagellate mitochondrial genomes. *J Mol Biol*, 372, 356-368.
- Smit A.F.A., Hubley, R., & Green P. RepeatMasker Open-3.0. 1996-2010.
(<http://w.w.w.repeatmasker.org>)
- Snyder, R.V., Gibbs, P.D., Palacios, A., Abiy, L., Dickey, R., Lopez, J.V., & Rein, K.S. (2003). Polyketide synthase genes from marine dinoflagellates. *Mar Biotechnol (NY)*, 5, 1-12.
- Soderlund, C., Bomhoff, M., Nelson, W. (2011). SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucl Acids Res.* 39, e68,
- Song, B. *et al.* (2017). Comparative genomics reveals two major bouts of gene retroposition coinciding with crucial periods of *Symbiodinium* evolution. *Genome Biol Evol.* 9, 2037-2047.
- Stachelhaus, T., Mootz, H.D., Marahiel, M.A. (1999). The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem Biol.* 6, 493-505.
- Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312-1313.
- Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, 24, 637-644.
- Strieker, M., Tanovic, A., & Marahiel, M.A. (2010). Nonribosomal peptide synthetases: structures and dynamics. *Curr Opin Struct Biol* 20, 234-240.

- Tang, Y., Kim, C-Y., Mathews, I.I., Cane, D.E., Khosla, C. (2006). The 2.7-Å crystal structure of a 194-kDa homodimeric fragment of the 6-deoxyerythronolide B synthase. *Proc Natl Acad Sci USA* 103, 11124-11129.
- Taroncher-Oldenburg, G & Anderson, DM. (2000). Identification and Characterization of Three Differentially Expressed Genes, Encoding Adenosylhomocysteine Hydrolase, Methionine Aminopeptidase, and a Histone-Like Protein, in the Toxic Dinoflagellate *Alexandrium fundyense*. *App. Environ. Microbiol.* 66, 2105-2112.
- Taylor, D.L. (1971). On the symbiosis between *Amphidinium klebsii* [Dinophyceae] and *Amphiscolops langerhansi* [Turbellaria: Acoela]. *J. Mar. Biol. Assoc. U.K.*, 301-313.
- ten Lohuis, M.R., & Miller, D.J. (1998). Genetic transformation of dinoflagellates (Amphidinium and Symbiodinium)-expression of GUS in microalgae using heterologous promoter constructs. *The Plant Journal*, 13, 427-435.
- Thattai, M., Burak, Y., Shraiman, B. (2007). The origins of specificity in polyketide synthase protein interactions. *PLoS Comput Biol.* 3, e186.
- Trench, R.K. The Cell Biology of Plant-Animal Symbiosis. *Annu Rev Plant Physiol Plant Mol Biol.* 30, 485-531.
- Tunez, I., Munoz, M.D.C., Feijoo, M., Munoz-Castaneda, R., Bujalance, I., Valdelvira, M.E., & Lopez, M. (2003). Protective melatonin effect on oxidative stress induced by okadaic acid into rat brain. *J Pineal Res*, 34, 265-268.
- Uemura, D. Bioactive polyethers. In: Scheuer PJ (ed.) *Bioorganic Marine Chemistry*, Vol 4, 1-31 (Springer-Verlag, 1991).
- Van Dolah, F.M., Kohli, G.S., Morey, J.S., Murray, S.A., & Lin, S. (2017). Both modular and single-domain Type I polyketide synthases are expressed in the brevetoxin-producing dinoflagellate, *Karenia brevis* (Dinophyceae). *J Phycol*, 53, 1325-1339.

- Van Wagoner, R.M., Satake, M. & Wright, J.L. (2014). Polyketide biosynthesis in dinoflagellates: what makes it different? *Nat Prod Rep.* 31,1101-37.
- Veldhuis, M.J.W., Cucci, T.L., Sieracki. (1997). Cellular DNA content of marine phytoplanktons using two new fluorochromes: taxonomic and ecological implications. *J Phycol*, 33, 527-541.
- Vurture, G.W., Sedlazeck, F.J., Nattestad, M., Underwood, C.J., Fang, H., Gurtowski, J., & Schatz, M.C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, 33, 2202-2204.
- Waller, R.F., & Jackson, C.J. (2009). Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology. *Bioessays*, 31, 237-245.
- Walsh, C.T., O'Brien, R.V., & Khosla, C. (2013). Nonproteinogenic Amino Acid Building Blocks for Nonribosomal Peptide and Hybrid Polyketide Scaffolds. *Angewandte Chemie International Edition*, 52, 7098-7124.
- Wang, D., Ho, A.Y.T., & Hsieh, D.P.H. (2002). Production of C2 toxin by *Alexandrium tamarense* CI01 using different culture methods. *Journal of Applied Phycology*, 14, 461-468.
- Wang, D.Z. (2008). Neurotoxins from marine dinoflagellates: a brief review. *Marine Drugs*, 6, 349-371.
- Wang, H., Fewer, D.P., Holm, L., Rouhiainen, L., Sivonen, K. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. *Proc Natl Acad Sci USA* 111, 9259-9264.
- Wenzel, S.C., & Muller, R. (2007). Myxobacterial natural product assembly lines: fascinating examples of curious biochemistry. *Nat Prod Rep*, 24, 1211-1224.
- Williams, D.H., Stone, M.J., Hauck, P.R., Rahman, S.K. (1989). Why are secondary metabolites (natural products) biosynthesized? *J Nat Prod.* 52,1189-1208.

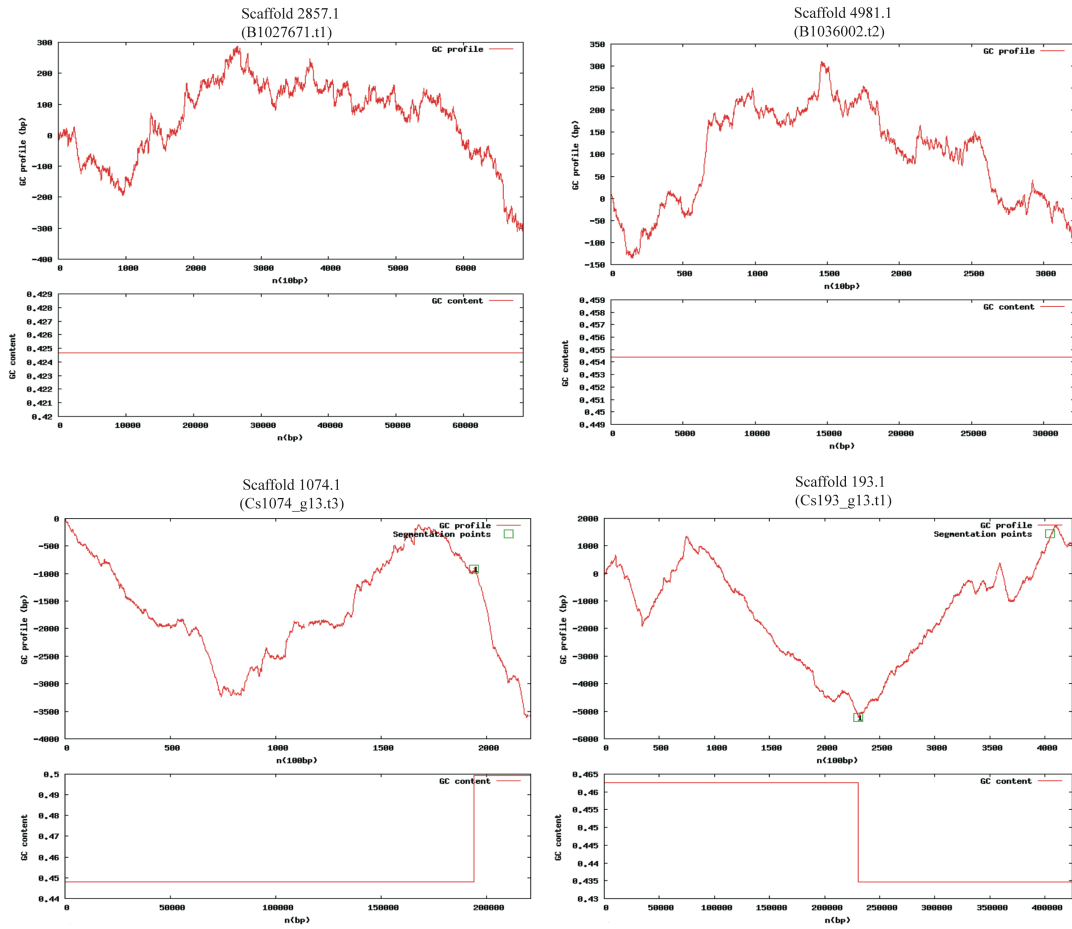
- Wisecaver, J.H., & Hackett, J.D. (2011). Dinoflagellate genome evolution. *Annu Rev Microbiol*, 65, 369-387.
- Wisecaver, J.H., Brosnahan, M.L., Hackett, J.D. (2013). Horizontal gene transfer is a significant driver of gene innovation in dinoflagellates. *Genome Biol Evol.* 5, 2368-2381
- Wu, T.D., & Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21, 1859-1875.
- Wurch, L.L., Bertrand, E.M., Saito, M.A., Van Mooy, B.A., & Dyhrman, S.T. (2011). Proteome changes driven by phosphorus deficiency and recovery in the brown tide-forming alga *Aureococcus anophagefferens*. *PLoS One*, 6, e28949.
- Xu, Z. & Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucl Acids Res.* 35, W265-268,
- Ziemert, N. *et al.* (2012). The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity. *PLoS ONE* 7, e34064.
- Zhang, H., Kolb, F.A, Jaskiewicz, L., Westhof, E., Filipowicz W. (2004). Single processing center models for human Dicer and bacterial RNase III. *Cell* 118, 57-68.
- Zhang, H., Bhattacharya, D., & Lin, S. (2007). A Three-Gene Dinoflagellate Phylogeny Suggests Monophyly of Prorocentrales and a Basal Position for Amphidinium and Heterocapsa. *Journal of Molecular Evolution*, 65, 463-474.
- Zhang, H., Hou, Y., Miranda, L., Campbell, D.A., Sturm, N.R., Gaasterland, T., & Lin, S. (2007). Spliced leader RNA trans-splicing in dinoflagellates. *Proc Natl Acad Sci U S A*, 104, 4618-4623.
- Zhang, M. Q. (1998). Statistical features of human exons and their flanking regions. *Human Molecular Genetics*, 7, 919-932.
- Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 9, 40.

Zhang, Z., Green, B. R., & Cavalier-Smith, T. (1999). Single gene circles in dinoflagellate chloroplast genomes. *Nature*, 400, 155-159.

Zhu, G. (2004). Current Progress in the Fatty Acid Metabolism in *Cryptosporidium parvum*. *J Eukaryot Microbiol.* 51,381-388.

APPENDIX A

Figure showing GC plots of 4 scaffolds associated with dinoflagellate PKS-I (Figure 2.1). Gaps less than 1% of the input scaffold fasta sequences were filtered. Plots were generated using a halting parameter of 100. The GC profile is shown in red while segmentation points are depicted by the numbered green boxes.



Appendix B Features of LTR-retrotransposons identified with PKS-associated scaffolds studied.

Clade	Scaffold #	Location	Length	Strand	Score ^a	5'-LTR ^b	3'-LTR ^c	PBS (Primer Binding Sites) ^d	PPT (Polynurine tract) ^e
B1	1171.1	95013 - 98077	3065	–	6	95013 - 95130 Len: 118	97960 - 98077 Len: 118	[14/17] 97872 - 97888 (-AlaCGC)	[11/15] 95168 - 95182
	2011.1	23 - 6079	6057	–	6	23 - 173 Len: 151	5929 - 6079 Len: 151	[14/17] 5908 - 5924 (-GlnCTG)	[13/15] 174 - 188
	5132.1	18094 - 25320	7227	+	7	18094 - 19213 Len: 1120	24159 - 25320 Len: 1162	[15/18] 19282 - 19299 (MetCAT)	[11/15] 24111 - 24125
	55.1	181361 - 187363	6003	–	7	181361 - 181472 Len: 112	187250 - 187363 Len: 114	[15/21] 187188 - 187208 (-IleTAT)	[11/15] 181516 - 181530
	57.1	96596 - 100499	3904	–	6	96596 - 97893 Len: 1298	99189 - 100499 Len: 1311	[15/25] 99098 - 99122 (-MetCAT)	[12/15] 97949 - 97963
	991.1	10007 - 16876	6870	–	6	10007 - 10738 Len: 732	16156 - 16876 Len: 721	[15/18] 16068 - 16085 (-GlnCTG)	[11/15] 10816 - 10830
		13136 - 16374	3239	+	6	13136 - 13346 Len: 211	16179 - 16374 Len: 196	[14/19] 13417 - 13435 (LeuTAG)	[12/15] 16164 - 16178
		13392 - 16876	3485	–	6	13392 - 13882 Len: 491	16391 - 16876 Len: 486	[16/24] 16359 - 16382 (-SerCGA)	[11/15] 13960 - 13974
		110819 - 118941	8123	+	7	110819 - 110926 Len: 108	118834 - 118941 Len: 108	[14/19] 110973 - 110991 (GlnCTG)	[12/15] 118750 - 118764
	214.1	10996 - 23070	12075	+	6	10996 - 11125 Len: 130	22959 - 23070 Len: 112	[15/19] 11203 - 11221 (AlaTGC)	[12/15] 22944 - 22958
		233839 - 242959	9121	–	7	233839 - 233994 Len: 156	242804 - 242959 Len: 156	[15/19] 242763 - 242781 (-LeuTAG)	[11/15] 233995 - 234009
	5099.1	19008 - 25190	6183	–	6	19008 - 19160 Len: 153	25024 - 25190 Len: 167	[15/20] 24990 - 25009 (-SerCGA)	[13/15] 19235 - 19249
	214.1	10996 - 23070	12075	+	6	10996 - 11125 Len: 130	22959 - 23070 Len: 112	[15/19] 11203 - 11221 (AlaTGC)	[12/15] 22944 - 22958
		233839 - 242959	9121	–	8	233839 - 233994 Len: 156	242804 - 242959 Len: 156	[15/19] 242763 - 242781 (-LeuTAG)	[11/15] 233995 - 234009
	2052.1	52858 - 62137	9280	–	6	52858 - 52960 Len: 103	62036 - 62137 Len: 102	[15/20] 61936 - 61955 (-AspGTC)	[11/15] 52966 - 52980
	31.1	384910 - 392457	7548	–	6	384910 - 385045 Len: 136	392322 - 392457 Len: 136	[14/21] 392224 - 392244 (-SerAGA)	[12/15] 385118 - 385132
	3551.1	22815 - 26576	3762	–	6	22815 - 22937 Len: 123	26463 - 26576 Len: 114	ND	[13/15] 22997 - 23011
		25224 - 39923	14700	–	6	25224 - 25331 Len: 108	39818 - 39923 Len: 106	ND	[11/15] 25347 - 25361
		25824 - 44954	19131	–	6.5	25824 - 26076 Len: 253	44706 - 44954 Len: 249	[14/19] 44659 - 44677 (-SerCGA)	[11/15] 26080 - 26094
		25826 - 45437	19612	–	6	25826 - 26560 Len: 735	44708 - 45437 Len: 730	[14/19] 44659 - 44677 (-SerCGA)	[12/15] 26578 - 26592
		27416 - 46553	19138	–	6	27416 - 28184 Len: 769	45798 - 46553 Len: 756	ND	[14/15] 28202 - 28216
		38907 - 44166	5260	+	6	38907 - 39910 Len: 1004	43214 - 44166 Len: 953	[14/18] 39930 - 39947 (LeuTAG)	[12/15] 43138 - 43152
		39072 - 43376	4305	+	6	39072 - 39185 Len: 114	43264 - 43376 Len: 113		[11/15] 43191 - 43205
		39698 - 49594	9897	+	6	39698 - 39910 Len: 213	49382 - 49594 Len: 213	[14/18] 39930 - 39947 (LeuTAG)	[12/15] 49300 - 49314
		40452 - 44835	4384	–	6	40452 - 40576 Len: 125	44711 - 44835 Len: 125	[14/19] 44659 - 44677 (-SerCGA)	[11/15] 40610 - 40624
		43228 - 45094	1867	–	6	43228 - 43559 Len: 332	44751 - 45094 Len: 344	[14/19] 44659 - 44677 (-SerCGA)	[12/15] 43593 - 43607
		44734 - 50473	5740	+	6	44734 - 44938 Len: 205	50269 - 50473 Len: 205	ND	[12/15] 50240 - 50254
	2561.1	45561 - 57214	11654	–	6	45561 - 46004 Len: 444	56768 - 57214 Len: 447	ND	[11/15] 46067 - 46081
	572.1	40146 - 44099	3954	+	6	40146 - 40852 Len: 707	43402 - 44099 Len: 698	[14/21] 40904 - 40924 (SerAGA)	[14/15] 43366 - 43380
		42198 - 46411	4214	+	6	42198 - 43173 Len: 976	45414 - 46411 Len: 998	[14/18] 43177 - 43194 (LeuTAG)	[11/15] 45399 - 45413
		42652 - 46411	3760	+	7	42652 - 43173 Len: 522	45907 - 46411 Len: 505	[14/18] 43177 - 43194 (LeuTAG)	[13/15] 45881 - 45895
		229591 - 231955	2365	–	6	229591 - 229821 Len: 231	231739 - 231955 Len: 217	[14/19] 231658 - 231676 (-MetCAT)	ND
		229591 - 231955	2365	–	6	229591 - 230049 Len: 459	231505 - 231955 Len: 451	[14/19] 231424 - 231442 (-MetCAT)	ND
		229591 - 231955	2365	–	6	229591 - 229935 Len: 345	231625 - 231955 Len: 331	[14/19] 231538 - 231556 (-MetCAT)	ND

A3		229591 - 231955	2366	–	7	229591 - 230106 Len: 516	231448 - 231955 Len: 508	[14/20] 231376 - 231395 (-MetCAT)	ND
	1629.1	70450 - 79049	8600	–	6	70450 - 70720 Len: 271	78779 - 79049 Len: 271	ND	[13/15] 70768 - 70782
		241976 - 251263	9288	+	6	241976 - 242420 Len: 445	250808 - 251263 Len: 456	ND	[12/15] 250728 - 250742
	190.1	144382 - 146079	1698	+	6	144382 - 144603 Len: 222	145857 - 146079 Len: 223	[14/20] 144647 - 144666 (GlnCTG)	ND
		213011 - 215482	2472	+	6	213011 - 213143 Len: 133	215343 - 215482 Len: 140	[14/20] 213222 - 213241 (LeuAAG)	[11/15] 215266 - 215280
		213011 - 215482	2472	+	6	213011 - 213143 Len: 133	215346 - 215482 Len: 137	[14/20] 213222 - 213241 (LeuAAG)	[11/15] 215266 - 215280
		213059 - 215446	2388	+	6	213059 - 213166 Len: 108	215330 - 215446 Len: 117	[14/20] 213222 - 213241 (LeuAAG)	[11/15] 215266 - 215280
		213059 - 215446	2388	+	6	213059 - 213166 Len: 108	215343 - 215446 Len: 104	[14/20] 213222 - 213241 (LeuAAG)	[11/15] 215266 - 215280
		213059 - 215452	2394	+	6	213059 - 213166 Len: 108	215343 - 215452 Len: 110	[14/20] 213222 - 213241 (LeuAAG)	[11/15] 215266 - 215280
		276622 - 277917	1296	+	7	276622 - 276762 Len: 141	277783 - 277917 Len: 135	[15/21] 276805 - 276825 (LeuAAG)	[12/15] 277733 - 277747
	41.1	96440 - 112431	15992	+	6	96440 - 96548 Len: 109	112325 - 112431 Len: 107	[14/18] 96623 - 96640 (ProTGG)	[11/15] 112293 - 112307
		276004 - 284536	8533	–	6.5	276004 - 276211 Len: 208	284327 - 284536 Len: 210	[17/26] 284291 - 284316 (-ThrCGT)	[12/15] 276223 - 276237
		330216 - 339229	9014	–	6	330216 - 330329 Len: 114	339115 - 339229 Len: 115	[14/19] 339047 - 339065 (-LeuTAG)	[11/15] 330374 - 330388
		399839 - 410874	11036	–	6	399839 - 399942 Len: 104	410771 - 410874 Len: 104	[14/16] 410694 - 410709 (-IleAAT)	[11/15] 399955 - 399969
	1206.1	56613 - 58066	1454	–	6	56613 - 56721 Len: 109	57958 - 58066 Len: 109	[15/19] 57931 - 57949 (-GlnCTG)	[12/15] 56768 - 56782
		56728 - 58457	1730	–	6	56728 - 56840 Len: 113	58345 - 58457 Len: 113	[14/19] 58278 - 58296 (-LeuTAA)	[12/15] 56856 - 56870
	4164.1	59417 - 61467	2051	+	6	59417 - 59730 Len: 314	61142 - 61467 Len: 326	[14/18] 59775 - 59792 (ThrTGT)	[12/15] 61075 - 61089
		61027 - 62264	1238	+	6	61027 - 61137 Len: 111	62154 - 62264 Len: 111	[14/16] 61164 - 61179 (SerTGA)	[12/15] 62082 - 62096
	126.1	72758 - 79309	6552	+	7	72758 - 72931 Len: 174	79134 - 79309 Len: 176	[16/21] 72971 - 72991 (MetCAT)	[14/15] 79091 - 79105
		124617 - 126040	1424	+	6	124617 - 124803 Len: 187	125885 - 126040 Len: 156	[14/19] 124817 - 124835 (TyrGTA)	[12/15] 125844 - 125858
		124694 - 126040	1347	+	6	124694 - 124803 Len: 110	125933 - 126040 Len: 108	[14/19] 124817 - 124835 (TyrGTA)	[12/15] 125844 - 125858
	1018.1	132419 - 152551	20133	–	6	132419 - 132526 Len: 108	152444 - 152551 Len: 108	[14/22] 152344 - 152365 (-SerCGA)	[11/15] 132536 - 132550
		153216 - 158216	5001	–	6	153216 - 153342 Len: 127	158106 - 158216 Len: 111	ND	[11/15] 153345 - 153359
		156331 - 167768	11438	–	6	156331 - 156457 Len: 127	167643 - 167768 Len: 126	[14/20] 167546 - 167565 (-AlaCGC)	ND
		160683 - 177505	16823	–	6	160683 - 160790 Len: 108	177398 - 177505 Len: 108	[14/17] 177316 - 177332 (-GlnCTG)	[12/15] 160854 - 160868
	665.1	7795 - 24932	17138	–	6	7795 - 7898 Len: 104	24796 - 24932 Len: 137	[14/18] 24706 - 24723 (-MetCAT)	[11/15] 7945 - 7959
	138.1	159246 - 164831	5586	+	6	159246 - 159740 Len: 495	164337 - 164831 Len: 495	[14/18] 159802 - 159819 (LeuTAA)	[12/15] 164242 - 164256
	792.1	31964 - 34502	2539	+	6	31964 - 32077 Len: 114	34391 - 34502 Len: 112	[15/20] 32114 - 32133 (LeuTAG)	ND
		98119 - 110925	12807	–	6	98119 - 98439 Len: 321	110597 - 110925 Len: 329	[14/19] 110569 - 110587 (-GluTTC)	[12/15] 98440 - 98454
		98119 - 111134	13016	–	6	98119 - 98665 Len: 547	110597 - 111134 Len: 538	[14/19] 110569 - 110587 (-GluTTC)	[11/15] 98666 - 98680
	527.1	225586 - 227135	1550	+	6	225586 - 225718 Len: 133	227003 - 227135 Len: 133	ND	[12/15] 226949 - 226963
	508.1	80141 - 88969	8829	+	7	80141 - 80378 Len: 238	88732 - 88969 Len: 238	[14/18] 80408 - 80425 (SerAGA)	[12/15] 88678 - 88692
		120408 - 128933	8526	–	6	120408 - 120508 Len: 101	128833 - 128933 Len: 101	[16/20] 128779 - 128798 (-PheGAA)	[11/15] 120509 - 120523
		120408 - 129053	8646	+	6	120408 - 120508 Len: 101	128953 - 129053 Len: 101	[14/17] 120544 - 120560 (AspGTC)	[11/15] 128886 - 128900
		128938 - 137665	8728	–	7	128938 - 129053 Len: 116	137545 - 137665 Len: 121	[16/20] 137511 - 137530 (-PheGAA)	[11/15] 129054 - 129068
		129137 - 137810	8674	+	6	129137 - 129251 Len: 115	137696 - 137810 Len: 115	[14/17] 129276 - 129292 (AspGTC)	[11/15] 137618 - 137632
		183849 - 193303	9455	+	6	183849 - 184136 Len: 288	193020 - 193303 Len: 284	[14/18] 184195 - 184212 (SerTGA)	ND
	892.1	6622 - 26025	19404	–	6	6622 - 6881 Len: 260	25689 - 26025 Len: 337	[14/18] 25603 - 25620 (-IleTAT)	[13/15] 6966 - 6980
		7371 - 16651	9281	+	6	7371 - 8903 Len: 1533	15110 - 16651 Len: 1542	ND	[12/15] 15016 - 15030

	32386 - 41575	9190	–	6	32386 - 32552 Len: 167	41413 - 41575 Len: 163	[15/20] 41325 - 41344 (-SerCGA)	[12/15] 32628 - 32642
	76453 - 84886	8434	–	6	76453 - 76627 Len: 175	84689 - 84886 Len: 198	[14/18] 84650 - 84667 (-GlnCTG)	ND
894.1	60142 - 63238	3097	–	6	60142 - 60655 Len: 514	62741 - 63238 Len: 498	ND	[13/15] 60685 - 60699
	60142 - 63304	3163	–	6	60142 - 60655 Len: 514	62807 - 63304 Len: 498	ND	[13/15] 60685 - 60699
	84093 - 107701	23609	–	6	84093 - 86248 Len: 2156	105526 - 107701 Len: 2176	ND	[14/15] 86282 - 86296
	86137 - 97800	11664	–	6	86137 - 86248 Len: 112	97689 - 97800 Len: 112	ND	[14/15] 86282 - 86296
	103043 - 104868	1826	–	6	103043 - 103196 Len: 154	104716 - 104868 Len: 153	[14/17] 104649 - 104665 (-GlnTTG)	ND
	254187 - 255899	1713	+	6	254187 - 254467 Len: 281	255618 - 255899 Len: 282	ND	[13/15] 255584 - 255598
	254187 - 257103	2917	+	6	254187 - 254370 Len: 184	256919 - 257103 Len: 185	ND	[13/15] 256885 - 256899
	254191 - 256389	2199	+	6	254191 - 254643 Len: 453	255910 - 256389 Len: 480	ND	[13/15] 255872 - 255886
	254191 - 256389	2199	+	6	254191 - 254741 Len: 551	255816 - 256389 Len: 574	ND	[14/15] 255775 - 255789
574.1	197963 - 199919	1957	+	6	197963 - 198067 Len: 105	199815 - 199919 Len: 105	[14/19] 198130 - 198148 (SerCGA)	[12/15] 199748 - 199762
	229375 - 232257	2883	+	6.5	229375 - 229583 Len: 209	232040 - 232257 Len: 218	[16/20] 231962 - 231981 (-GlnTTG)	[11/15] 229604 - 229618
	237751 - 239203	1453	–	6	237751 - 237899 Len: 149	239055 - 239203 Len: 149	[14/20] 239003 - 239022 (-GlnCTC)	[13/15] 237939 - 237953
144.1*	71526 - 79598	8073	–	6	71526 - 71719 Len: 194	79410 - 79598 Len: 189	[15/21] 79321 - 79341 (-LeuTAG)	[13/15] 71735 - 71749
	209207 - 211111	1905	+	6	209207 - 209330 Len: 124	210989 - 211111 Len: 123	[15/21] 209408 - 209428 (AlaTGC)	[11/15] 210959 - 210973
	209250 - 211111	1862	+	6	209250 - 209393 Len: 144	210968 - 211111 Len: 144	[15/21] 209408 - 209428 (AlaTGC)	[11/15] 210938 - 210952
	209250 - 211111	1862	+	6	209250 - 209372 Len: 123	210968 - 211111 Len: 144	[15/21] 209408 - 209428 (AlaTGC)	[11/15] 210938 - 210952
	464327 - 466467	2141	+	6	464327 - 464439 Len: 113	466355 - 466467 Len: 113	[16/23] 464483 - 464505 (SerCGA)	[11/15] 466340 - 466354
	464472 - 465948	1477	–	6	464472 - 464629 Len: 158	465789 - 465948 Len: 160	[14/21] 465707 - 465727 (-SerTGA)	[11/15] 464646 - 464660
	527399 - 530149	2751	+	6	527399 - 527896 Len: 498	529669 - 530149 Len: 481	[15/22] 527975 - 527996 (GlnCTG)	[13/15] 529580 - 529594
	527399 - 530158	2760	+	6	527399 - 527905 Len: 507	529669 - 530158 Len: 490	[15/22] 527975 - 527996 (GlnCTG)	[13/15] 529580 - 529594
2246.1	113 - 2512	2400	+	6	113 - 385 Len: 273	2219 - 2512 Len: 294	[14/18] 441 - 458 (TyrGTA)	[11/15] 2176 - 2190
	527 - 2646	2120	–	6	527 - 635 Len: 109	2538 - 2646 Len: 109	[14/20] 2479 - 2498 (-LeuTAA)	[11/15] 673 - 687
929.1	121850 - 123407	1558	+	6	121850 - 121967 Len: 118	123283 - 123407 Len: 125	[14/23] 122024 - 122046 (GlnCTG)	ND
	233694 - 235328	1635	+	7	233694 - 233845 Len: 152	235177 - 235328 Len: 152	[14/19] 233906 - 233924 (AsnGTT)	[11/15] 235134 - 235148
121.1	91055 - 104339	13285	+	6	91055 - 91247 Len: 193	104147 - 104339 Len: 193	[16/22] 91319 - 91340 (SerCGA)	[13/15] 104096 - 104110
	115223 - 125158	9936	–	6	115223 - 115379 Len: 157	124998 - 125158 Len: 161	[14/18] 124898 - 124915 (-SerTGA)	[13/15] 115388 - 115402
	116875 - 132473	15599	+	7	116875 - 117139 Len: 265	132206 - 132473 Len: 268	[15/20] 117219 - 117238 (GlnCTG)	[12/15] 132144 - 132158
	117902 - 126724	8823	–	6	117902 - 118042 Len: 141	126577 - 126724 Len: 148	[14/20] 126506 - 126525 (-LeuCAG)	[11/15] 118082 - 118096
	121024 - 122569	1546	+	6	121024 - 121272 Len: 249	122321 - 122569 Len: 249	[14/19] 121335 - 121353 (MetCAT)	ND
	127000 - 133770	6771	–	6	127000 - 127177 Len: 178	133593 - 133770 Len: 178	[14/18] 133563 - 133580 (-TyrGTA)	[12/15] 127215 - 127229
	198329 - 210153	11825	–	6	198329 - 198436 Len: 108	210044 - 210153 Len: 110	[14/16] 210030 - 210045 (-ValAAC)(ValTAC)	[11/15] 198468 - 198482
	384659 - 403461	18803	–	6	384659 - 384758 Len: 100	403361 - 403461 Len: 101	[14/17] 403287 - 403303 (-GlyTCC)	[13/15] 384769 - 384783
1154.1	69587 - 87034	17448	–	6	69587 - 69692 Len: 106	86923 - 87034 Len: 112	[14/16] 86855 - 86870 (-GluTTC)	[12/15] 69733 - 69747
	197004 - 211504	14501	+	7	197004 - 197360 Len: 357	211156 - 211504 Len: 349	[16/17] 197410 - 197426 (LeuCAA)	[12/15] 211080 - 211094
612.1	279163 - 287299	8137	–	6	279163 - 279320 Len: 158	287149 - 287299 Len: 151	[15/21] 287078 - 287098 (-GlnCTG)	[12/15] 279366 - 279380
1817.1	132492 - 141135	8644	+	7	132492 - 132644 Len: 153	140982 - 141135 Len: 154	[15/21] 132713 - 132733 (LeuTAA)	[11/15] 140940 - 140954
895.1	139108 - 146591	7484	–	6	139108 - 139287 Len: 180	146371 - 146591 Len: 221	[14/18] 146332 - 146349 (-PheGAA)	ND

C1

1080.1*	9 - 5995	5987	+	6	9 - 128 Len: 120	5879 - 5995 Len: 117	ND	[11/15] 5803 - 5817
	9 - 6856	6848	+	6	9 - 136 Len: 128	6725 - 6856 Len: 132	ND	[11/15] 6651 - 6665
	105049 - 106815	1767	-	6	105049 - 105166 Len: 118	106698 - 106815 Len: 118	[14/21] 106638 - 106658 (-SerCGA)	ND
	105049 - 109600	4552	+	6	105049 - 105166 Len: 118	109483 - 109600 Len: 118	ND	[11/15] 109391 - 109405
	105049 - 109887	4839	+	6	105049 - 105166 Len: 118	109770 - 109887 Len: 118	ND	[11/15] 109723 - 109737
	106135 - 109652	3518	+	6	106135 - 106327 Len: 193	109483 - 109652 Len: 170	ND	[11/15] 109391 - 109405
	106322 - 108000	1679	-	6	106322 - 106429 Len: 108	107889 - 108000 Len: 112	ND	[12/15] 106508 - 106522
	106363 - 108300	1938	-	6	106363 - 106523 Len: 161	108149 - 108300 Len: 152	ND	[11/15] 106600 - 106614
	106363 - 108455	2093	-	6	106363 - 106639 Len: 277	108149 - 108455 Len: 307	ND	[11/15] 106706 - 106720
684.1	67758 - 75155	7398	+	6	67758 - 68154 Len: 397	74767 - 75155 Len: 389	[16/19] 68219 - 68237 (IleTAT)	[11/15] 74721 - 74735
	69889 - 76205	6317	-	6	69889 - 70124 Len: 236	75999 - 76205 Len: 207	[14/21] 75923 - 75943 (-GluCTC)	[12/15] 70205 - 70219
	69889 - 76435	6547	+	6	69889 - 70350 Len: 462	75999 - 76435 Len: 437	[16/27] 70388 - 70414 (ProTGG)	[11/15] 75936 - 75950
	69889 - 76599	6711	-	6	69889 - 70469 Len: 581	75999 - 76599 Len: 601	[14/21] 75923 - 75943 (-GluCTC)	[12/15] 70550 - 70564
	70134 - 76435	6302	+	6	70134 - 70350 Len: 217	76215 - 76435 Len: 221	[16/27] 70388 - 70414 (ProTGG)	[11/15] 76140 - 76154
	70134 - 76599	6466	-	6	70134 - 70469 Len: 336	76215 - 76599 Len: 385	[14/18] 76121 - 76138 (-GluCTG)	[12/15] 70550 - 70564
	159518 - 160782	1265	-	6.5	159518 - 159635 Len: 118	160666 - 160782 Len: 117	[15/21] 160643 - 160663 (-LeuAAG)	[11/15] 159701 - 159715
	184999 - 187139	2141	-	6	184999 - 185274 Len: 276	186891 - 187139 Len: 249	[15/20] 186792 - 186811 (-ThrCGT)	[11/15] 185299 - 185313
	185026 - 187021	1996	-	6	185026 - 185466 Len: 441	186619 - 187021 Len: 403	ND	[14/15] 185481 - 185495
	241203 - 249130	7928	+	6	241203 - 241415 Len: 213	248918 - 249130 Len: 213	[15/19] 241430 - 241448 (PheGAA)	[13/15] 248898 - 248912
228.1	275218 - 276629	1412	+	6	275218 - 275420 Len: 203	276458 - 276629 Len: 172	ND	[12/15] 276438 - 276452
	275429 - 277241	1813	-	6	275429 - 275615 Len: 187	277109 - 277241 Len: 133	ND	[11/15] 275661 - 275675
357.1	62674 - 72364	9691	+	6	62674 - 63129 Len: 456	71911 - 72364 Len: 454	[14/19] 63207 - 63225 (ThrCGT)	[14/15] 71896 - 71910
	149759 - 161813	12055	+	7	149759 - 149917 Len: 159	161655 - 161813 Len: 159	[15/21] 149921 - 149941 (ValCAC)	[12/15] 161640 - 161654
	180238 - 182969	2732	+	6	180238 - 180551 Len: 314	182656 - 182969 Len: 314	ND	[11/15] 182587 - 182601
	210322 - 212492	2171	+	6	210322 - 210504 Len: 183	212265 - 212492 Len: 228	[15/20] 210555 - 210574 (MetCAT)	[12/15] 212213 - 212227
	210596 - 212731	2136	+	7	210596 - 210879 Len: 284	212473 - 212731 Len: 259	[14/20] 210882 - 210901 (SerTGA)	[12/15] 212383 - 212397
	210769 - 212731	1963	+	7	210769 - 210879 Len: 111	212620 - 212731 Len: 112	[14/20] 210882 - 210901 (SerTGA)	[12/15] 212581 - 212595
	210769 - 212908	2140	+	6	210769 - 211104 Len: 336	212620 - 212908 Len: 289	ND	[12/15] 212581 - 212595
1344.1	111723 - 113537	1815	-	6	111723 - 111982 Len: 260	113295 - 113537 Len: 243	[14/18] 113207 - 113224 (-GlnCTG)	[14/15] 111996 - 112010
1074.1	81236 - 82665	1430	-	6	81236 - 81459 Len: 224	82476 - 82665 Len: 190	[15/20] 82453 - 82472 (-LeuTAA)	[13/15] 81491 - 81505
	159010 - 160425	1416	-	6.5	159010 - 159205 Len: 196	160249 - 160425 Len: 177	[14/17] 160157 - 160173 (-IleAAT)	[11/15] 159221 - 159235
	159063 - 160349	1287	-	7	159063 - 159181 Len: 119	160249 - 160349 Len: 101	[14/17] 160157 - 160173 (-IleAAT)	[11/15] 159221 - 159235
	214814 - 217867	3054	-	6	214814 - 214921 Len: 108	217760 - 217867 Len: 108	ND	[11/15] 214922 - 214936
193.1	143198 - 147027	3830	+	6	143198 - 143975 Len: 778	146229 - 147027 Len: 799	[14/19] 143979 - 143997 (GlnCTG)	[13/15] 146152 - 146166
	143278 - 146193	2916	+	6	143278 - 143505 Len: 228	145949 - 146193 Len: 245	[14/18] 143524 - 143541 (IleTAT)	[12/15] 145853 - 145867
	143676 - 147484	3809	+	6	143676 - 143807 Len: 132	147328 - 147484 Len: 157	[15/19] 143835 - 143853 (MetCAT)	[12/15] 147312 - 147326
	143813 - 147027	3215	+	6	143813 - 143975 Len: 163	146866 - 147027 Len: 162	[14/19] 143979 - 143997 (GlnCTG)	[11/15] 146778 - 146792
	189084 - 191040	1957	-	7	189084 - 189277 Len: 194	190843 - 191040 Len: 198	[16/21] 190754 - 190774 (-IleAAT)	[11/15] 189285 - 189299
24.1	242805 - 244398	1594	+	6	242805 - 243036 Len: 232	244152 - 244398 Len: 247	ND	[12/15] 244107 - 244121

	287257 - 289315	2059	+	6	287257 - 287531 Len: 275	289026 - 289315 Len: 290	ND	[11/15] 288963 - 288977
	287298 - 289180	1883	–	6	287298 - 287420 Len: 123	289064 - 289180 Len: 117	ND	[12/15] 287447 - 287461
	460580 - 468901	8322	+	6	460580 - 460755 Len: 176	468725 - 468901 Len: 177	ND	[13/15] 468630 - 468644
429.1	89475 - 95162	5688	–	6	89475 - 89620 Len: 146	95018 - 95162 Len: 145	ND	[13/15] 89688 - 89702
2692.1	62963 - 64872	1910	+	6	62963 - 63180 Len: 218	64634 - 64872 Len: 239	ND	[12/15] 64594 - 64608
893.1	2955 - 10919	7965	–	6.5	2955 - 3075 Len: 121	10800 - 10919 Len: 120	[14/20] 10725 - 10744 (-SerTGA)	[12/15] 3089 - 3103
	75145 - 83093	7949	+	6	75145 - 75279 Len: 135	82960 - 83093 Len: 134	[15/21] 75290 - 75310 (GlnTTG)	[13/15] 82877 - 82891
	85020 - 86568	1549	+	6	85020 - 85166 Len: 147	86423 - 86568 Len: 146	[14/20] 85192 - 85211 (PheGAA)	[13/15] 86324 - 86338
	85020 - 87877	2858	+	6	85020 - 85204 Len: 185	87692 - 87877 Len: 186	[14/18] 85277 - 85294 (SerCGA)	[14/15] 87609 - 87623
	85132 - 88028	2897	+	6	85132 - 85268 Len: 137	87891 - 88028 Len: 138	[14/18] 85277 - 85294 (SerCGA)	[11/15] 87848 - 87862
	85192 - 89954	4763	–	6	85192 - 85938 Len: 747	89254 - 89954 Len: 701	[15/19] 89156 - 89174 (-AlaTGC)	[11/15] 85957 - 85971
	85323 - 89954	4632	+	6	85323 - 85938 Len: 616	89370 - 89954 Len: 585	[14/19] 85987 - 86005 (ProTGG)	[11/15] 89355 - 89369
	85323 - 89980	4658	+	6	85323 - 85964 Len: 642	89370 - 89980 Len: 611	[14/19] 85987 - 86005 (ProTGG)	[11/15] 89355 - 89369
	85796 - 89980	4185	+	6	85796 - 85964 Len: 169	89853 - 89980 Len: 128	[14/19] 85987 - 86005 (ProTGG)	[11/15] 89838 - 89852
	85903 - 87446	1544	+	6	85903 - 86100 Len: 198	87249 - 87446 Len: 198	[15/22] 86155 - 86176 (GlnCTG)	[12/15] 87177 - 87191
	87951 - 89391	1441	–	6	87951 - 88088 Len: 138	89254 - 89391 Len: 138	[15/19] 89156 - 89174 (-AlaTGC)	[12/15] 88111 - 88125
118.1	14575 - 16472	1898	–	6	14575 - 14758 Len: 184	16286 - 16472 Len: 187	[14/18] 16262 - 16279 (-ArgTCT)	[11/15] 14811 - 14825
	14575 - 16685	2111	–	6	14575 - 14925 Len: 351	16286 - 16685 Len: 400	[14/18] 16262 - 16279 (-ArgTCT)	[11/15] 14959 - 14973
	18699 - 21393	2695	–	6	18699 - 19035 Len: 337	21041 - 21393 Len: 353	ND	[11/15] 19036 - 19050
	69834 - 71178	1345	+	6	69834 - 69996 Len: 163	71021 - 71178 Len: 158	ND	[11/15] 70972 - 70986
	76336 - 77823	1488	+	7	76336 - 76493 Len: 158	77668 - 77823 Len: 156	[14/19] 76575 - 76593 (MetCAT)	[12/15] 77594 - 77608
	76336 - 78568	2233	+	7	76336 - 76493 Len: 158	78414 - 78568 Len: 155	[14/19] 76575 - 76593 (MetCAT)	[12/15] 78390 - 78404
	216947 - 236907	19961	+	6	216947 - 217171 Len: 225	236693 - 236907 Len: 215	[14/19] 217179 - 217197 (IleTAT)	[12/15] 236660 - 236674
	402912 - 404459	1548	+	6	402912 - 403012 Len: 101	404359 - 404459 Len: 101	ND	[12/15] 404306 - 404320
	533681 - 535037	1357	–	6	533681 - 533828 Len: 148	534889 - 535037 Len: 149	ND	[14/15] 533888 - 533902
	533681 - 535050	1370	–	6	533681 - 533866 Len: 186	534889 - 535050 Len: 162	ND	[14/15] 533888 - 533902
1269.1	63723 - 72250	8528	–	7	63723 - 63869 Len: 147	72104 - 72250 Len: 147	[15/21] 72059 - 72079 (-SerAGA)	[13/15] 63882 - 63896
1989.1	138962 - 140373	1412	+	6	138962 - 139106 Len: 145	140238 - 140373 Len: 136	[15/19] 139110 - 139128 (SerTGA)	[12/15] 140204 - 140218
202.1	330201 - 338192	7992	–	6	330201 - 330435 Len: 235	337955 - 338192 Len: 238	[14/20] 337863 - 337882 (-ThrCGT)	[12/15] 330483 - 330497
	331030 - 333855	2826	–	6.5	331030 - 331345 Len: 316	333535 - 333855 Len: 321	[15/21] 333490 - 333510 (-GlnCTG)	[12/15] 331413 - 331427
	335519 - 337527	2009	+	6	335519 - 335941 Len: 423	337092 - 337527 Len: 436	[14/17] 336008 - 336024 (LeuCAA)	[13/15] 337077 - 337091
	418240 - 420681	2442	–	6	418240 - 418807 Len: 568	420102 - 420681 Len: 580	[14/20] 420009 - 420028 (-GluCTC)	[13/15] 418833 - 418847

a: Score is an integer varying from 0 to 11.

b & c: Location of the LTRs. 5'LTR and 3'LTR are two similar regions. A typical LTR retrotransposon has a structure called TG.CA box, with TG at the 5' extremity of 5'LTR and CA at the 3' extremity of 3'LTR.

d: The first number in square brackets is number of matched bases and the second is total alignment length. Following is signal positions. String in parentheses is the tRNA type and anti-codon.

e: The first is the number of purines and length of putative PPT.

ND: Not detected.

*: Only first 9 outputs are shown out of 24

Features of LTR-retrotransposons identified with NRPS-associated scaffolds studied.

Clade	Scaffold #	Location	Length	Strand	Score ^a	5'-LTR ^b	3'-LTR ^c	PBS (Primer Binding Sites) ^d	PPT (Polypurine tract) ^e
A3	451.1	124432 - 127126	2695	+	6	124432 - 124613 Len: 182	126946 - 127126 Len: 181	[15/20] 124630 - 124649 (GlnCTG)	[13/15] 126920 - 126934
		124768 - 129338	4571	+	6	124768 - 124899 Len: 132	129207 - 129338 Len: 132	[14/20] 124958 - 124977 (AlaTGC)	[12/15] 129110 - 129124
	112.1	130245 - 133532	3288	+	6	130245 - 130356 Len: 112	133421 - 133532 Len: 112	[14/20] 130365 - 130384 (PheGAA)	[12/15] 133404 - 133418
		130357 - 133527	3171	-	6	130357 - 130471 Len: 115	133408 - 133527 Len: 120	ND	[11/15] 130549 - 130563
	1977.1	36433 - 43341	6909	-	6.5	36433 - 36594 Len: 162	43179 - 43341 Len: 163	[16/24] 43125 - 43148 (-AsnGTT)	[14/15] 36633 - 36647
		122602 - 124406	1805	+	6	122602 - 122819 Len: 218	124212 - 124406 Len: 195	[14/19] 122867 - 122885 (LeuTAG)	[13/15] 124117 - 124131
	2138.1	6877 - 22893	16017	+	6	6877 - 6982 Len: 106	22788 - 22893 Len: 106	[15/19] 7025 - 7043 (LeuTAG)	[12/15] 22698 - 22712
		22900 - 27800	4901	+	6.5	22900 - 23005 Len: 106	27699 - 27800 Len: 102	[14/17] 23040 - 23056 (IleTAT)	[11/15] 27659 - 27673
B1	745.1	22201 - 23780	1580	-	6	22201 - 22314 Len: 114	23668 - 23780 Len: 113	[15/22] 23584 - 23605 (-IleAAT)	[11/15] 22396 - 22410
	3629.1	16587 - 24405	7819	-	7	16587 - 16755 Len: 169	24237 - 24405 Len: 169	ND	[12/15] 16756 - 16770
	4570.1	19443 - 27781	8339	-	7	19443 - 19722 Len: 280	27502 - 27781 Len: 280	ND	[12/15] 19768 - 19782
C1	535.1	288896 - 297474	8579	+	6.5	288896 - 288996 Len: 101	297372 - 297474 Len: 103	[14/19] 289012 - 289030 (MetCAT)	[12/15] 297357 - 297371
	1214.1	142212 - 157440	15229	-	6	142212 - 142355 Len: 144	157290 - 157440 Len: 151	[15/22] 157211 - 157232 (-AlaCGC)	[13/15] 142356 - 142370
	193.1	143198 - 147027	3830	+	6	143198 - 143975 Len: 778	146229 - 147027 Len: 799	[14/19] 143979 - 143997 (GlnCTG)	[13/15] 146152 - 146166
		143278 - 146193	2916	+	6	143278 - 143505 Len: 228	145949 - 146193 Len: 245	[14/18] 143524 - 143541 (IleTAT)	[12/15] 145853 - 145867
		143676 - 147484	3809	+	6	143676 - 143807 Len: 132	147328 - 147484 Len: 157	[15/19] 143835 - 143853 (MetCAT)	[12/15] 147312 - 147326
		143813 - 147027	3215	+	6	143813 - 143975 Len: 163	146866 - 147027 Len: 162	[14/19] 143979 - 143997 (GlnCTG)	[11/15] 146778 - 146792
		189084 - 191040	1957	-	7	189084 - 189277 Len: 194	190843 - 191040 Len: 198	[16/21] 190754 - 190774 (-IleAAT)	[11/15] 189285 - 189299
	2146.1	31479 - 36824	5346	-	7	31479 - 31731 Len: 253	36571 - 36824 Len: 254	[14/19] 36547 - 36565 (-ProTGG)	[12/15] 31761 - 31775
		32468 - 37643	5176	-	6	32468 - 32971 Len: 504	37158 - 37643 Len: 486	[14/22] 37077 - 37098 (-AlaAGC)	[12/15] 33010 - 33024
		35948 - 40886	4939	-	6	35948 - 36131 Len: 184	40707 - 40886 Len: 180	[14/20] 40631 - 40650 (-LeuTAA)	ND
		36018 - 41388	5371	+	7	36018 - 36539 Len: 522	40877 - 41388 Len: 512	[15/21] 36608 - 36628 (SerTGA)	[13/15] 40794 - 40808
	227.1*	5759 - 37699	1941	+	6	35759 - 36172 Len: 414	37285 - 37699 Len: 415	[15/19] 36253 - 36271 (LeuCAA)	[11/15] 37226 - 37240
		85652 - 91173	5522	-	6	85652 - 86139 Len: 488	90728 - 91173 Len: 446	[14/19] 90701 - 90719 (-AlaTGC)	[11/15] 86156 - 86170
		85652 - 91219	5568	-	6	85652 - 86139 Len: 488	90728 - 91219 Len: 492	[14/19] 90701 - 90719 (-AlaTGC)	[11/15] 86156 - 86170
		85652 - 91265	5614	-	6	85652 - 86093 Len: 442	90772 - 91265 Len: 494	[14/19] 90701 - 90719 (-AlaTGC)	[11/15] 86156 - 86170
		85652 - 91276	5625	+	6	85652 - 86075 Len: 424	90864 - 91276 Len: 413	[14/19] 86093 - 86111 (SerCGA)	[11/15] 90781 - 90795
		85652 - 91401	5750	+	6	85652 - 85970 Len: 319	91102 - 91401 Len: 300	ND	[11/15] 91019 - 91033
		85699 - 91081	5383	-	6	85699 - 86139 Len: 441	90682 - 91081 Len: 400	[15/23] 90651 - 90673 (-IleAAT)	[11/15] 86156 - 86170
		85745 - 91127	5383	-	6	85745 - 86139 Len: 395	90682 - 91127 Len: 446	[15/23] 90651 - 90673 (-IleAAT)	[11/15] 86156 - 86170

	85791 - 91035	5245	—	6	85791 - 86139 Len: 349	90682 - 91035 Len: 354	[15/23] 90651 - 90673 (-IleAAT)	[11/15] 86156 - 86170
268.1	334018 - 342657	8640	+	8	334018 - 334147 Len: 130	342526 - 342657 Len: 132	[16/23] 334160 - 334182 (GlnCTG)	[11/15] 342478 - 342492
830.1	42518 - 44223	1706	—	6	42518 - 42693 Len: 176	44044 - 44223 Len: 180	[14/18] 43968 - 43985 (-IleAAT)	[11/15] 42695 - 42709
	120620 - 122987	2368	+	6	120620 - 120849 Len: 230	122781 - 122987 Len: 207	[14/19] 120862 - 120880 (IleAAT)	[11/15] 122722 - 122736
	120652 - 123784	3133	+	6	120652 - 120868 Len: 217	123569 - 123784 Len: 216	[14/18] 120938 - 120955 (LeuTAG)	[11/15] 123527 - 123541
	120899 - 123577	2679	+	7	120899 - 121301 Len: 403	123213 - 123577 Len: 365	[14/18] 121338 - 121355 (LeuTAG)	[11/15] 123176 - 123190
	120899 - 123577	2679	+	7	120899 - 121166 Len: 268	123349 - 123577 Len: 229	[14/18] 121210 - 121227 (LeuTAG)	[11/15] 123255 - 123269
	120899 - 123577	2679	+	7	120899 - 121429 Len: 531	123077 - 123577 Len: 501	[14/18] 121471 - 121488 (LeuTAG)	[11/15] 123040 - 123054
	121125 - 123024	1900	+	6	121125 - 121515 Len: 391	122653 - 123024 Len: 372	ND	[11/15] 122585 - 122599

a: Score is an integer varying from 0 to 11.

b & c: Location of the LTRs. 5'LTR and 3'LTR are two similar regions. A typical LTR retrotransposon has a structure called TG.CA box, with TG at the 5' extremity of 5'LTR and CA at the 3' extremity of 3'LTR.

d: The first number in square brackets is number of matched bases and the second is total alignment length. Following is signal positions. String in parentheses is the tRNA type and anti-co

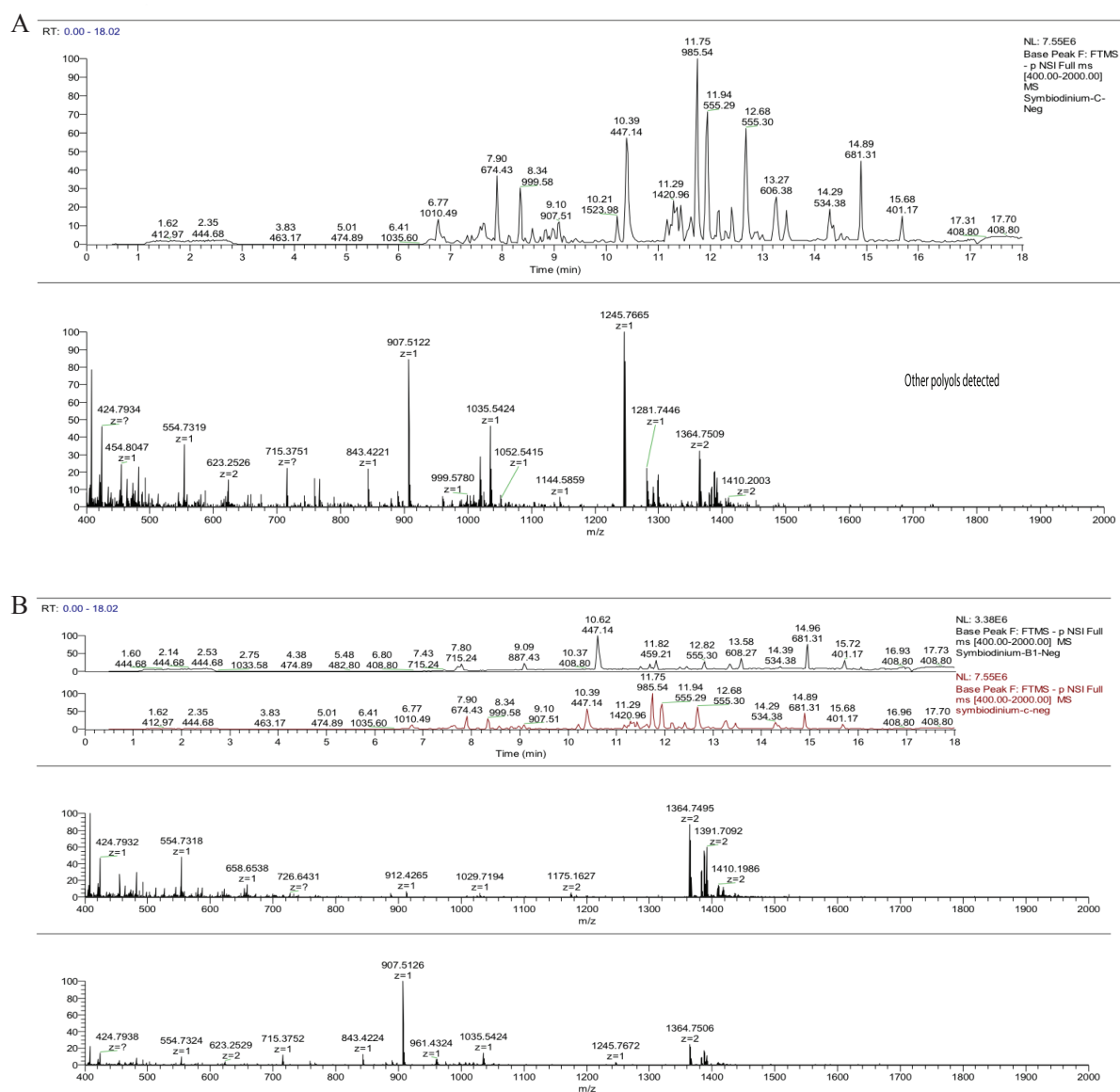
e: The first is the number of purines and length of putative PPT.

ND: Not detected.

*: Only first 9 outputs are shown out of 24

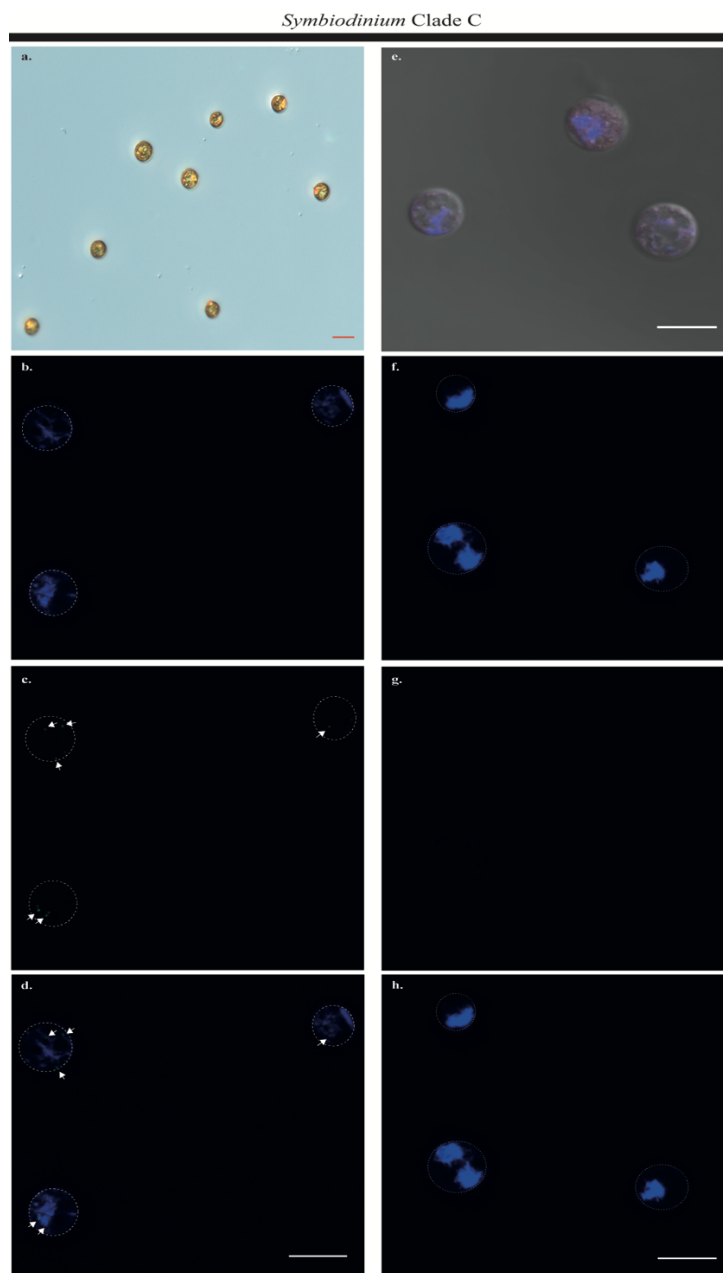
APPENDIX D

Figure showing nanoLC-MS (negative ion) profile and mass spectrum (expanded) of the methanol extract of Clade C **(A)**. **(B)**. Similarity of nanoLC-MS (negative ion) profile and mass spectrum (expanded) of the methanol extract of Clade C and B1.



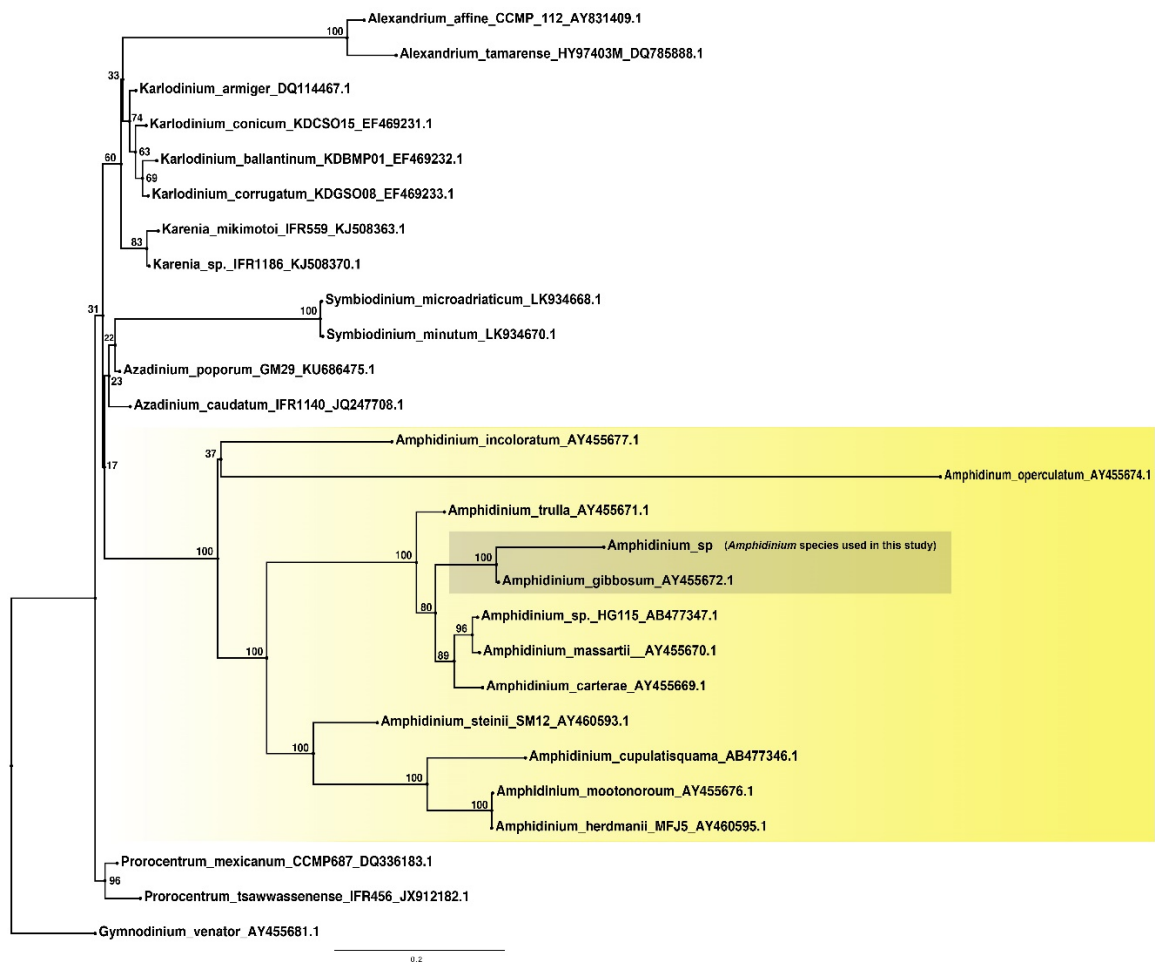
APPENDIX E

Figure showing immunofluorescent staining of *Symbiodinium* cells with anti-KS antibody. **a.** Differential interference contrast (DIC) imaging (40x). **b-d.** Confocal images (63x) of clade C stained with KS antibody showing localization of KS proteins (arrows). Nuclei are stained blue with DAPI (**b**), KS proteins are in green (**c**) and merged image of nuclei and KS protein staining (**d**). **e.** DIC imaging of cells at 63x showing detailed peripheral localization of chloroplasts (red autofluorescence) and nuclei (blue). **f-h.** Confocal images (63x) of control cells stained with only secondary antibody. Nuclei are stained with DAPI (**f**), but no KS protein were stained (**g**). Merged image of nuclei and no-KS staining (**h**). White dotted lines show the cell outlines. Scale bars are 10 μm in the panels.



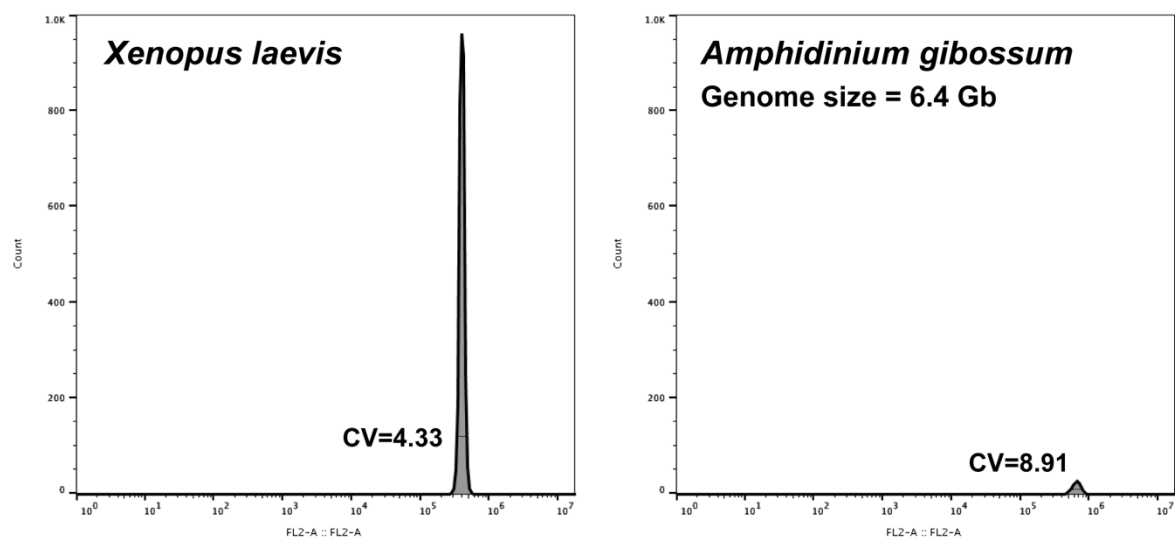
APPENDIX F

Figure showing phylogenetic analysis of alignment of *Amphidinium* partial LSU rDNA sequences using maximum likelihood. Values at nodes represent bootstrap support.



APPENDIX G

Figure showing comparison with the 3.1Gb *Xenopus laevis* genome suggesting that *A. gibossum* haploid genome is approximately 6.4 Gb in size, which matches Genomescope estimation at K= 81. CV is the Coefficient of Variation.



APPENDIX H

Table showing (a) details of genome assembly based on statistics of scaffolds of size ≥ 500 bp. (b) Annotation statistics for gene models.

A

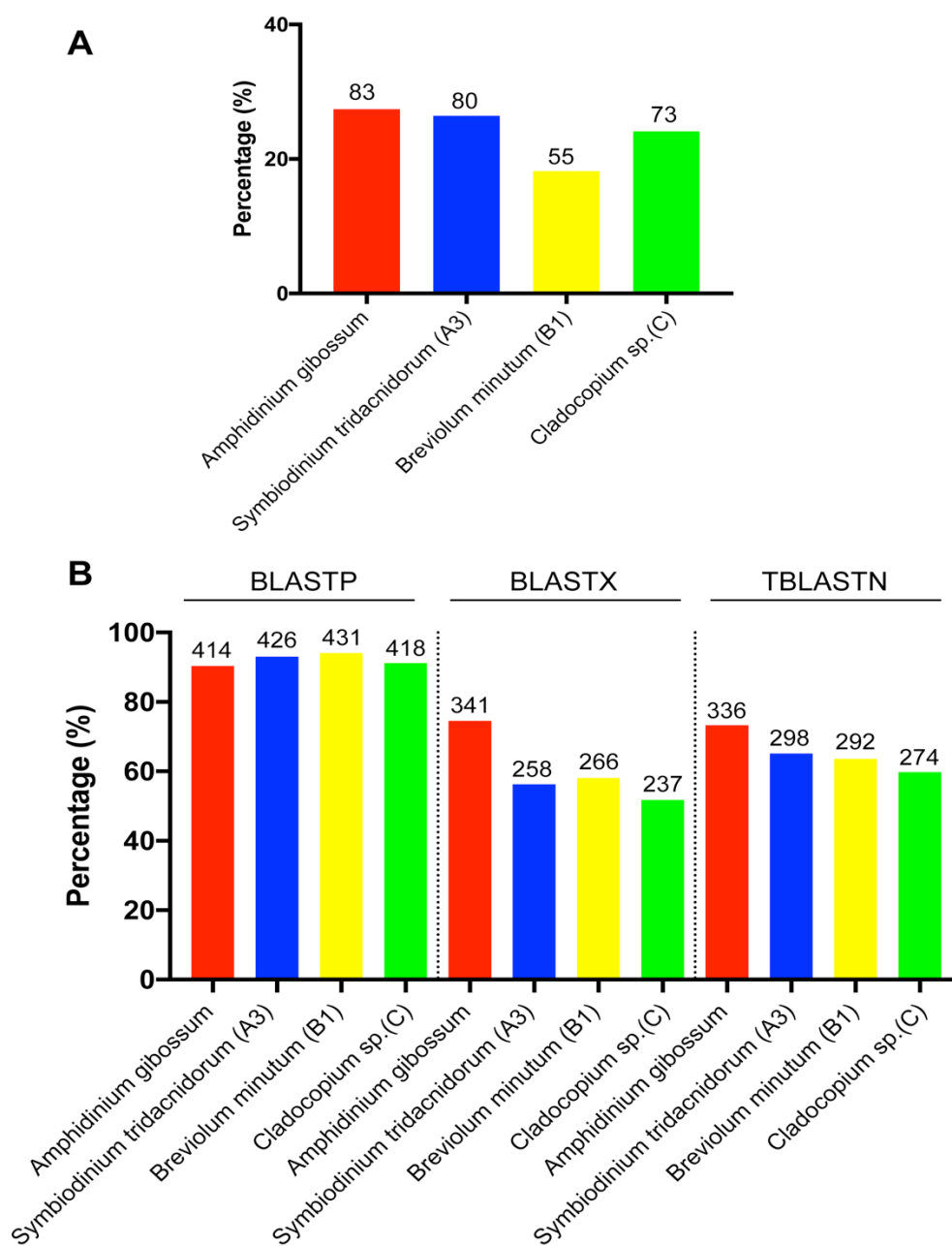
# scaffolds (≥ 1000 bp)	174,692
# scaffolds (≥ 5000 bp)	108,913
# scaffolds (≥ 10000 bp)	76,033
# scaffolds (≥ 25000 bp)	46,743
# scaffolds (≥ 50000 bp)	34,084
Total length (≥ 0 bp)	7,564,364,748
Total length (≥ 1000 bp)	6,949,758,089
Total length (≥ 5000 bp)	6,825,375,768
Total length (≥ 10000 bp)	6,574,923,633
Total length (≥ 25000 bp)	6,102,117,029
# scaffolds	298,063
Total length (≥ 50000 bp)	5,649,811,202
Largest scaffold	3,442,467
Total length	7,034,147,423
GC (%)	47.09
N50	166,499
N75	71,058
L50	12,063
L75	27,848

B

Annotation	Number	Percentage (%)
Total number of gene models	85139	100
Number of gene models with Swiss-Prot top hits	18015	21.1
Number of gene models with TrEMBL top hits	41117	48.3
Number of gene models with association with KEGG orthologs	3827	4.5
Number of gene models with annotated Pfam domains	5442	6.4

APPENDIX I

Figure showing (A) recovery of 303 BUSCO genes in *Amphidinium gibossum*, and (B) recovery of 458 CEGMA genes based on several BLAST analyses using BLASTP for predicted proteins and genome scaffolds for BLASTX and TBLASTN, respectively.



APPENDIX J

Table showing *A. gibossum* repeat content.

Category	Element	Number of occurrences	Number of bp covered
DNA transposon	Academ-1	121	4703
	Academ-2	3	127
	Academ-H	2	83
	CMC-Chapaev	1050	43084
	CMC-Chapaev-3	115	4641
	CMC-EnSpm	19149	1767122
	CMC-Mirage	1	40
	CMC-Transib	2682	144587
	Crypton	361	13437
	Crypton-A	73	2792
	Crypton-C	2	124
	Crypton-F	13	606
	Crypton-H	542	24172
	Crypton-S	20	1144
	Crypton-V	1070	37362
	Crypton-X	2	115
	Dada	707	29827
	Ginger	3188	249028
	IS3EU	596	22225
	Kolobok	27	1104
	Kolobok-E	7	403
	Kolobok-Hydra	1561	101722
	Kolobok-T2	1905	99366
	MULE-F	7	389
	MULE-MuDR	19897	1437967
	MULE-NOF	85	3786
	Maverick	2912	157819
	Merlin	673	24815
	Novosib	1409	152653
	P	1882	82424
	P-Fungi	6	207
	PIF-HarbS	8	307
	PIF-Harbinger	3838	174415
	PIF-ISL2EU	157	6661
	PIF-Spy	560	27903
	PiggyBac	211	9005
	PiggyBac-A	1	42
	PiggyBac-X	44	3446
	Sola-1	951	67602
	Sola-2	147	6910
	Sola-3	630	69590
	TcMar	2675	146449
	TcMar-Ant1	8	391
	TcMar-Cweed	3	111
	TcMar-Fot1	1052	62995
	TcMar-Gizmo	14	2065

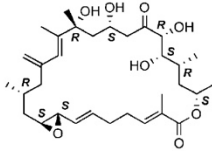
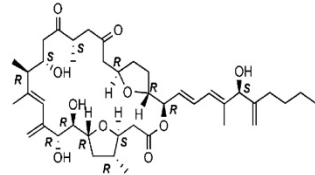
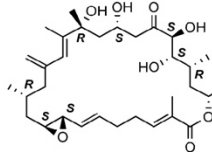
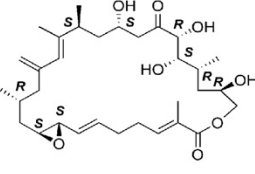
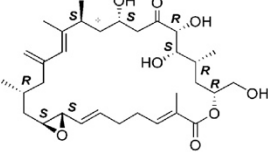
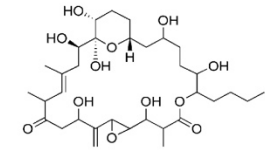
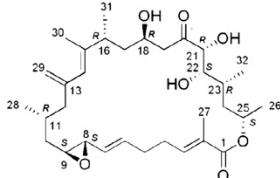
	TcMar-ISRm11	242	8716
	TcMar-Mariner	293	15801
	TcMar-Pogo	15	738
	TcMar-Sagan	3	127
	TcMar-Stowaway	105	4630
	TcMar-Tc1	3364	211876
	TcMar-Tc2	49	4872
	TcMar-Tc4	33	1252
	TcMar-Tigger	18	776
	TcMar-m44	8	255
	Zator	8	378
	Zisupton	4866	303673
	hAT	1004	49962
	hAT-Ac	5088	329748
	hAT-Blackjack	291	12307
	hAT-Charlie	2465	157368
	hAT-Pegasus	95	6023
	hAT-Restless	3	87
	hAT-Tag1	910	40818
	hAT-Tip100	1582	63810
	hAT-hAT1	12	763
	hAT-hAT19	13	454
	hAT-hAT5	16	570
	hAT-hAT6	2	61
	hAT-hATm	340	12001
	hAT-hATw	74	3221
	hAT-hATx	3	120
	hAT-hobo	11	648
LINE	CR1	352	51472
	CR1-Zenon	20	774
	CRE	107	5717
	CRE-Ambal	53	3285
	CRE-Odin	3	201
	Deceiver	1	17
	Dong-R4	17	566
	Dualen	7	380
	Genie	2	177
	I	138	6550
	I-Jockey	2077	164909
	L1	3393	225473
	L1-DRE	24	1311
	L1-Tx1	1913	112143
	L1-Zorro	3	51
	L2	4005	269015
	Penelope	3456	187633
	Proto1	17	1088
	Proto2	9	436

	R1	1492	134471
	R1-LOA	16	706
	R2	168	11286
	R2-Hero	66	3347
	R2-NeSL	332	36457
	RTE-BovB	188	8569
	RTE-RTE	21	553
	RTE-X	195	12313
	Rex-Babar	94	4096
	Tad1	52	2487
LTR	Bhikhari	11	473
	Caulimovirus	35	1365
	Copia	3691	179769
	DIRS	1735	79064
	ERV-Foamy	3	86
	ERV-Lenti	1	69
	ERV1	3823	201937
	ERV4	67	2668
	ERVK	1564	78030
	ERVL	169	7493
	ERVL-MaLR	7	418
	Gypsy	16541	1240334
	Ngaro	483	21957
	Pao	762	36725
	Viper	5	234
Other	DNA_virus	9	393
RC	Helitron	4753	275930
	Helitron-2	60	2844
Retroposon	SVA	9	604
SINE	5S	1	64
	5S-Deu-L2	1	43
	5S-RTE	5	146
	7SL	2	62
	B2	1	28
	B4	55	2381
	ID	23	730
	MIR	3	73
	RTE	2	96
	RTE-BovB	2	130
	U	1	47
	tRNA	740	39558
	tRNA-5S	1	19
	tRNA-7SL	5	194

	tRNA-CR1	2	49
	tRNA-Ceph-RTE	4	260
	tRNA-Core	97	3206
	tRNA-Core-RTE	3	170
	tRNA-Deu	5	207
	tRNA-Deu-L2	2	138
	tRNA-I	6	201
	tRNA-L1	2	52
	tRNA-L2	7	272
	tRNA-Mermaid	2	151
	tRNA-Meta	20	724
	tRNA-RTE	5	269
	tRNA-Sauria	1	41
	tRNA-V-CR1	2	41
	TATE	1	55
	Y-chromosome centromeric	1 1	64 13
Unspecified		9789040	2068177725
Total interspersed		9942170	2078587436
Low_complexity RNA		331120 1	29212862 85
Satellite	5S	43	2694
	acro	39	3443
	macro	27	786
	telo	3	53
Simple_repeat		2230179	148606652
rRNA		469	24880
snRNA		19	566
tRNA		221	6139
Total		12514653	2257532404

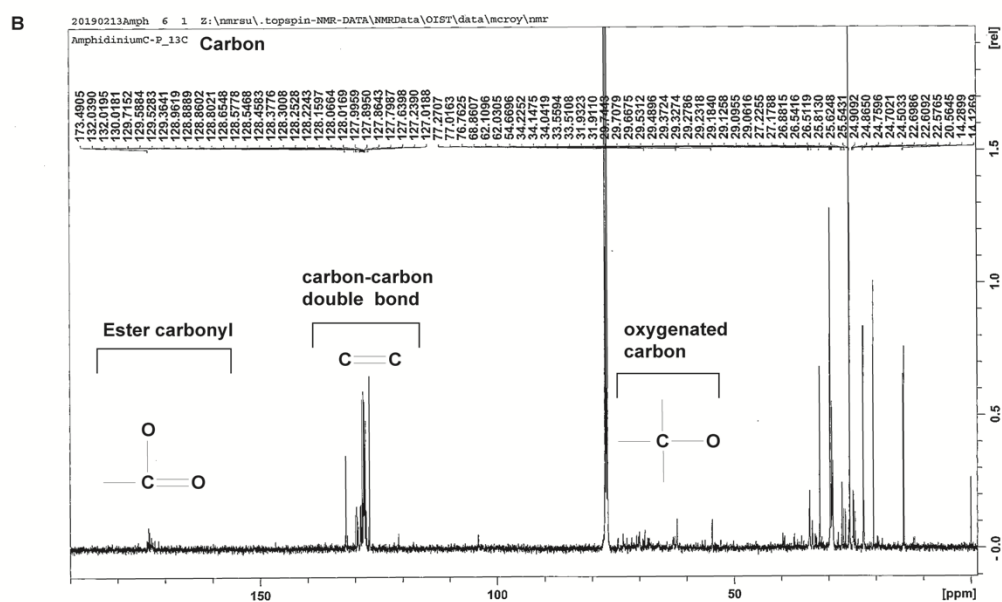
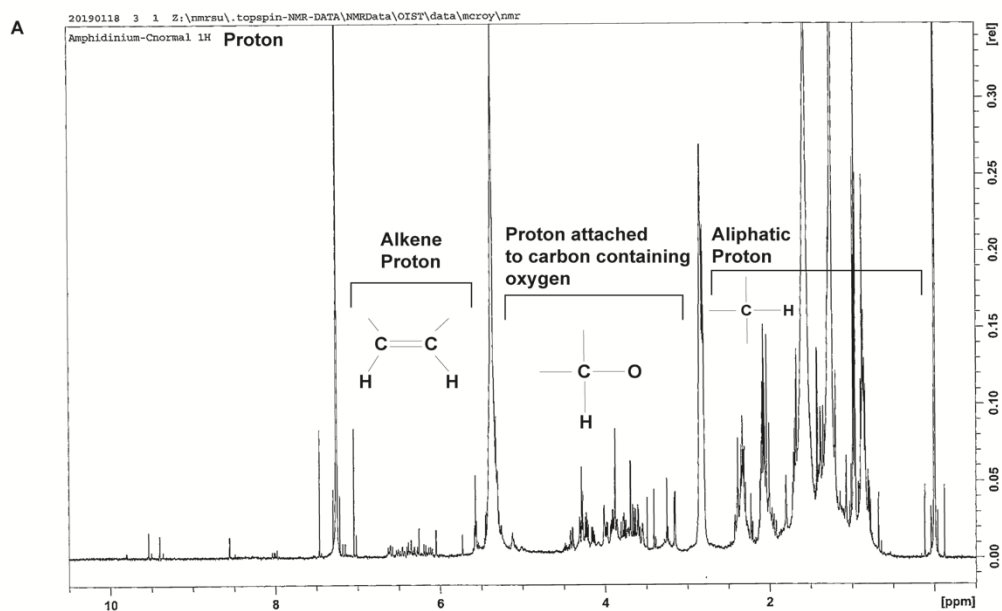
APPENDIX K

Table showing the lactone size, cytotoxicities (IC_{50} , $\mu\text{g/mL}$) against murine lymphoma (L1210), and human epidermoid carcinoma cells (KB) of some potent amphidinolides

	Lactone size	L1210	KB	
Amphidinolide B	26	0.00014	0.0042	
Amphidinolide C	25	0.0058	0.0046	
Amphidinolide D	26	0.019	0.08	
Amphidinolide G	26	0.0054	0.0059	
Amphidinolide H	26	0.00048	0.00052	
Amphidinolide N	26	0.00005	0.00006	
Amphidinolide B5	26	0.00012	0.004	

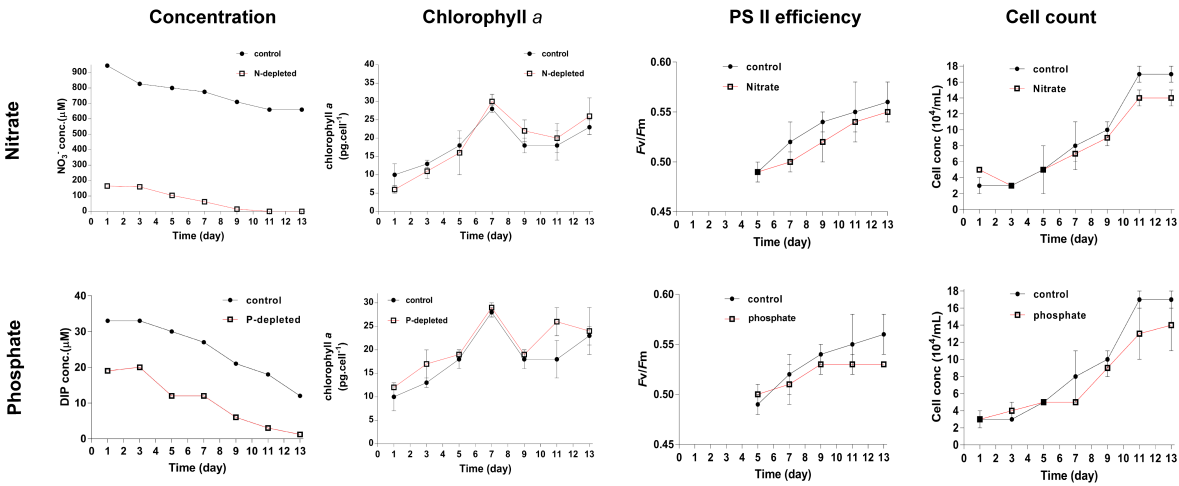
APPENDIX L

Figure showing NMR profile of methanol extract with distinct amphidinolide-like features
(A) proton NMR (B) carbon NMR.



APPENDIX M

Figure showing physiological parameters of *Amphidinium gibosum* under N-deplete and P-deplete conditions.



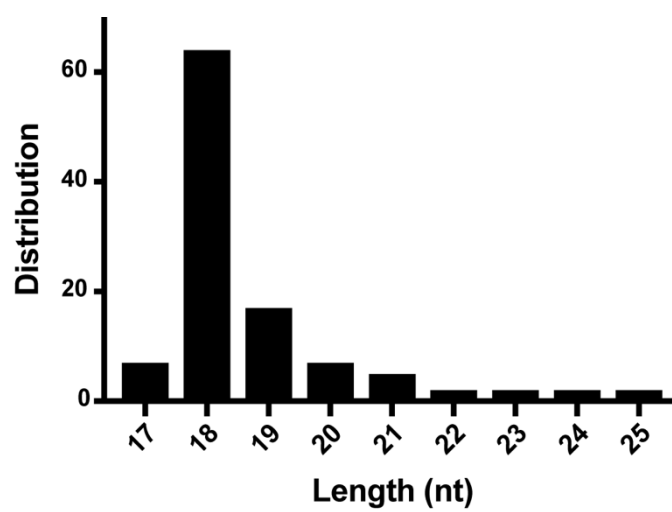
APPENDIX N

Table showing top 10 KEGG pathways in *A. gibosum* transcriptome. Numbers of pathways recovered are indicated in parenthesis.

Top 10 represented KEGG pathways	
ko01100	Metabolic pathways (841)
ko01110	Biosynthesis of secondary metabolites (346)
ko01130	Biosynthesis of antibiotics (233)
ko01120	Microbial metabolism in diverse environments (193)
ko01230	Biosynthesis of amino acids (109)
ko03010	Ribosome (105)
ko03040	Spliceosome (94)
ko01200	Carbon metabolism (91)
ko00230	Purine metabolism (77)
ko04141	Protein processing in endoplasmic reticulum (75)

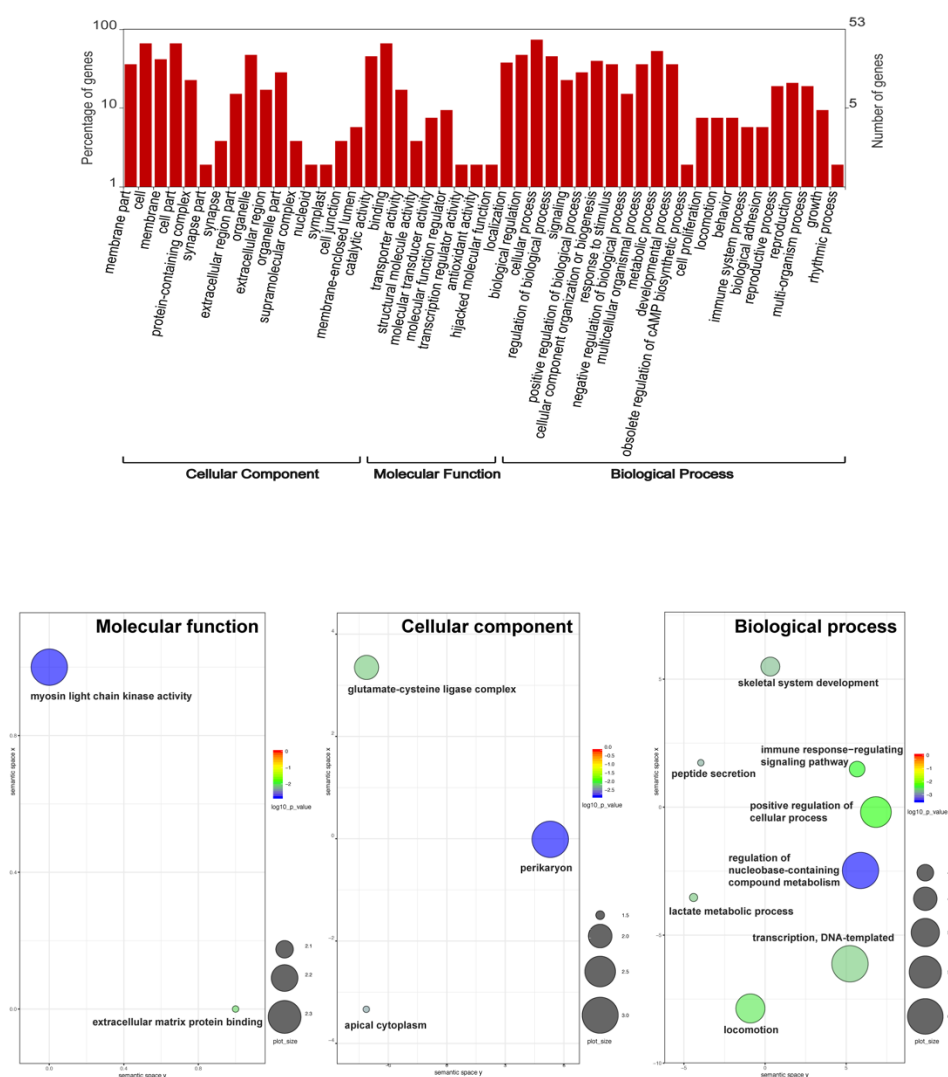
APPENDIX O

Figure showing length and distribution of microRNAs detected from *A. gibbosum*



APPENDIX P

Upper figure showing gene ontology classification for 303 predicted target unigenes of one differentially expressed miRNA under nitrate stress in three GO categories. Only 53 genes have GO annotations. Lower figure shows the functional enrichment of GO-terms analyzed with topGO and summarized with REVIGO. Bubble color represent the adjusted $p < 0.01$ while circle size shows the frequency of the GO term.



APPENDIX Q

Upper figure showing gene ontology classification for 2711 predicted target unigenes of three differentially expressed miRNA under phosphate stress in three GO categories. Only 580 genes have GO annotations. Lower figure shows the functional enrichment of GO-terms analyzed with topGO and summarized with REVIGO. Bubble color represent the adjusted $p < 0.01$ while circle size shows the frequency of the GO term.

