SCIENTIFIC REPORTS

OPEN

Diversified secondary metabolite biosynthesis gene repertoire revealed in symbiotic dinoflagellates

Received: 11 September 2018
Accepted: 13 December 2018
Published online: 04 February 2019

Girish Beedessee[1], Kanako Hisata[1], Michael C. Roy[2], Frances M. Van Dolah[3], Noriyuki Satoh[1] & Eiichi Shoguchi[1]

Symbiodiniaceae dinoflagellates possess smaller nuclear genomes than other dinoflagellates and produce structurally specialized, biologically active, secondary metabolites. Till date, little is known about the evolution of secondary metabolism in dinoflagellates as comparative genomic approaches have been hampered by their large genome sizes. Here, we overcome this challenge by combining genomic and metabolomics approaches to investigate how chemical diversity arises in three decoded Symbiodiniaceae genomes (clades A3, B1 and C). Our analyses identify extensive diversification of polyketide synthase and non-ribosomal peptide synthetase genes from two newly decoded genomes of *Symbiodinium tridacnidorum* (A3) and *Cladocopium* sp. (C). Phylogenetic analyses indicate that almost all the gene families are derived from lineage-specific gene duplications in all three clades, suggesting divergence for environmental adaptation. Few metabolic pathways are conserved among the three clades and we detect metabolic similarity only in the recently diverged clades, B1 and C. We establish that secondary metabolism protein architecture guides substrate specificity and that gene duplication and domain shuffling have resulted in diversification of secondary metabolism genes.

Dinoflagellates of the family Symbiodiniaceae[1] (previously known as the genus *Symbiodinium*) exist freely in symbiotic associations with many invertebrates, such as corals, clams, and anemones. This invertebrate-Symbiodiniaceae mutualism seems to provide a competitive advantage[2], resulting in the production and exchange of metabolites by both organisms[3]. Members of this family are sources of unusual, large, polyhydroxyl and polyether compounds or so-called "super-carbon-chain compounds (SCCs)," composed of long-chain backbones functionalized by oxygen[4]. The Symbiodiniaceae SCCs are polyketide metabolites, that are biosynthesized via an assembly line mechanism by two important classes of modular enzymes, polyketide synthase (PKS) and non-ribosomal peptide synthase (NRPS)[5]. PKSs comprise three core domains: an acyl-transferase (AT) domain, an acyl-carrier protein (ACP), and a ketosynthase (KS) domain that work with optional domains[6]. Polyketide synthases are also closely related to fatty acid synthases (FASs) and share the same core of enzymatic activities, implying a common evolutionary history[7]. Based on protein organization, PKSs are further categorized into three types (Type I, II and III), and FASs into two (Type I and II)[8].

On the other hand, NRPSs are modular multi-enzyme complexes that synthesize a diverse array of biological active peptides or lipopeptides[9]. Biosynthesis of non-ribosomal peptides occurs via the action of catalytic modules within NRPS that are composed of three compulsory domains: adenylation (A-domain), thiolation (T-domain), and condensation (C-domain), supported by other domains[10]. PKS and NRPS pathways often cross-talk such that a polyketide product is elongated by NRPS or *vice versa* to produce hybrid natural products, thereby increasing structural diversity[11]. Pathways involved in secondary metabolite biosynthesis are among the most rapidly evolving genetic elements[12]. Mutations, domain rearrangements, and module duplications within

[1]Marine Genomics Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa, 904-0495, Japan. [2]Instrumental Analysis Section, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa, 904-0495, Japan. [3]College of Charleston, School of Sciences and Mathematics, 66 George St., Charleston, South Carolina, 29424, USA. Correspondence and requests for materials should be addressed to G.B. (email: girish.beedessee@oist.jp) or E.S. (email: eiichi@oist.jp)

PKS and *NRPS* genes account for generation of novel, diverse small-molecules[12]. Thus, there exist several entry points where combinatorial potential can arise.

Several of these SCCs such as zooxanthellatoxins (ZTs) and zooxanthellamides (ZADs) have been isolated from several Symbiodiniaceae clades and a clade-to-metabolite relationship has been proposed and experimentally supported, in which strains of specific Symbiodiniaceae clades produce specific metabolites[13]. Nakamura *et al*. (ref.[14]) suggested the existence of shared biogenetic processes, such as the polyketide pathway with glycine as the starting substrate, yielding products with structural similarities to palytoxins and ZTs. Over the years, other secondary metabolites have been isolated from these clades, but their ecological roles and biosynthetic pathways have yet to be identified[15]. A preliminary genomic survey reported the presence and organization of secondary metabolite genes in *Breviolum minutum* (B1), overcoming limitations of previous transcriptomic surveys[16]. Availability of new Symbiodiniaceae genomes now allows us to survey and compare genes associated with metabolite biosynthesis[17–20]. However, how chemical diversity arises within Symbiodiniaceae is still unknown. Evolution of novel chemistry depends on diversity-generating metabolism, which comprises broad-substrate enzymes[21]. Metabolic pathways accept many different substrates, generating diverse chemical products and this provides organisms with unique chemistry to face environmental challenges[22].

To investigate existence of shared biosynthetic pathways, we cultured three species of the family Symbiodiniaceae namely *Symbiodinium tridacnidorum* (a.k.a clade A3), *Breviolum minutum* (a.k.a clade B1), and *Cladocopium* sp. (a.k.a clade C) that produce different metabolites, and surveyed their genomes[17,20] for genes involved in polyketide and non-ribosomal peptide biosynthesis. Additionally, we examined how these genomes are equipped to expand their gene repertoire for biosynthesis of complex secondary metabolites and suggest possible diversification mechanisms that may contribute to such chemical variability and modularity.

Results

Phylogenetic and syntenic analyses of ketosynthase and acyltransferase domains. The tree shows that majority KS domains clustered according to their domain organization types under a reliable node (BI posterior probability: 0.79 & maximum likelihood probability: 99) (Fig. 1, Supplemental Information, Figs S1 and S2). Recently, Kohli *et al*. (ref.[23]) described contigs encoding multiple PKS domains in the dinoflagellates, *Gambierdiscus excentricus* and *Gambierdiscus polynesiensis*. Those sequences clustered into three dinoflagellate groups (Dinoflagellate PKS I, II and III) (blue highlighted inset of Fig. 1). We confirmed the presence of 25 KS sequences each from clades A3 and C. Our analysis showed only one gene model (B1030341.t1) associated with Type II fatty acid synthesis (FabF-KASII) and one gene model (B1027279.t1) in the FabB-KASI group. The result mirrored the clear demarcation between Type II FAS and Type I PKS & FAS[24]. Our analysis additionally revealed the expanded nature of *KS* genes into nine PKS groups (Dinoflagellate PKS I–III and Symbiodiniaceae PKS I–VI) associated with either mono- or multifunctional domains (Fig. 1). One group (Dinoflagellate PKS-I) was closely related to cyanobacterial KS sequences. Scanning the GC profile of PKS-I group scaffolds of clade C showed some regions of higher GC content (45–46.5%), compared to the average genomic GC content of 43.0%, indicative of gene transfer (Supplemental Information, Fig. S3). ~3% (3/83) of the sequences contain the cTP (chloroplast transit peptide) signal while 12% (10/83) contained mitochondrial targeting peptide (mTP) or secretory signal each (Fig. 1).

A striking feature among the three genomes is the high number (26) of *trans*-AT genes in contrast to *cis*-AT (4) (Fig. 2). A phylogenetic tree of the AT domain consisted of two main nodes, *cis*-AT and *trans*-AT (BI posterior probability: 1.00 & maximum likelihood probability: 81) (Fig. 2, Supplemental Information, Figs S4 and S5), that deviated from the classical substrate-based clustering[25]. 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 (Fig. 2). Protein structure and function prediction by I-TASSER showed that most Symbiodiniaceae AT sequences pertain to the family of malonyl-CoA ACP transferase, based on the motif GAFH (highlighted blue in Fig. 2). Downstream of the active site serine, a motif (YASH or HAFH) is involved in the choice of either methylmalonyl-CoA or malonyl-CoA, respectively[26]. ~9% (3/33) of AT gene models contained the cTP or mTP signals (Fig. 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 (Supplemental Information, Fig. S6a). 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) (Supplemental Information, Fig. S6b–d) while only four *PKS*-containing gene clusters were found to be shared among all three clades (green boxes in Supplemental Information, Fig. S6b–d). The observed rearrangements within the syntenic scaffolds included mainly deletions. Transposons were found on scaffolds carrying *PKS*- and *NRPS*-coding 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 (Supplemental Information, Tables S6 and S7). Taken together, these results indicate that *PKS* genes have diversified in each Symbiodiniaceae species by several evolutionary processes.

Phylogenetic analysis of adenylation and condensation domain subtypes ($^{L}C_{L}$, $^{D}C_{L}$, Cyc and dual) in NRPS proteins. To understand if freestanding A-domains identified in Symbiodiniaceae genomes follow the same non-ribosomal code of traditional NRPS systems[27], we performed a phylogenetic comparison involving 117 adenylation sequences from different taxa. In addition, the amino acid substrate of adenylation domain was predicted by latent semantic indexing method[28]. One major observation was that freestanding A-domains appear in three major nodes (BI posterior probability: 1.00 & maximum likelihood probability: 72–100) that utilize tryptophan, glycine, and phenylalanine as substrates (three highlighted groups in Fig. 3a, Supplemental Information, Figs S7 and S8). In contrast, other proteins with di- or multi-domains displayed affinity for various substrates (Fig. 3a).

Figure 1. Phylogenetic tree of ketosynthase (KS) domains of prokaryotic and eukaryotic polyketide and fatty acid synthases. Analysis of ketosynthase, FabB-KASI, and FabF-KASII domains shows diversification of these domains into nine groups, comprised of mono-and multifunctional domains. Dots and squares indicate posterior probabilities of 0.70–0.89 and 0.9–1.0, respectively, generated by Bayesian inference. Inserts provide details of sub-groups as well as gene model architecture. C, M, and S denote chloroplasts, mitochondria, and secretory signal peptides, respectively.

Phylogenetic analysis of condensation domains was dominated by functional categories of C-domains rather than species phylogeny or substrate specificity alone (Fig. 3b). Classification of C-domains by NaPDoS software indicated that Symbiodiniaceae genomes are rich in $^{L}C_{L}$ subtypes (BI posterior probability: 0.73), that catalyze the condensation of two L-amino acids (Fig. 3b, Supplemental Information, Figs S9, S10), in contrast to a $^{D}C_{L}$ that links an L-amino acid to a D-amino acid. Our survey revealed the presence of six condensation domains with the consensus motif (HHxxxDG) being maintained, except for G being substituted with L and N in B1036245.t1 and Cs535_g6.t1, respectively (Fig. 3b). The phylogeny also supports the close relationship between $^{L}C_{L}$ and starter C-domains and dual and $^{D}C_{L}$ domains, as previously reported in bacterial genomes, confirming the reliability of our analysis[29]. These results demonstrate the specificity of NRPS genes for specific amino acids, thus introducing a degree of chemical diversity in non-ribosomal peptide biosynthesis.

Identification of metabolites and biosynthetic gene clusters from Symbiodiniaceae genomes. Polyols were identified based on their high-resolution mass data, as summarized in Beedessee *et al.* (ref.[16]). Doubly charged ions (negative ions) were searched for larger polyols (>2600 Da) in the MS spectra. Sample A3 showed the presence of zooxanthellatoxin-B (ZT-B), albeit in small amount, with an m/z of 1414.74 for the $[M\text{-}2H]^{2-}$ (Supplemental Information, Fig. 11a). Only zooxanthellamide D (ZAD-D) could be identified from sample B1 with extracted ions at *m/z* 1050.57 for the $[M+H]^{+}$ (Supplemental Information, Fig. 11b). No SCCs could be identified from sample C despite presence of many polyols (Supplemental Information, Figs 11c–12a).

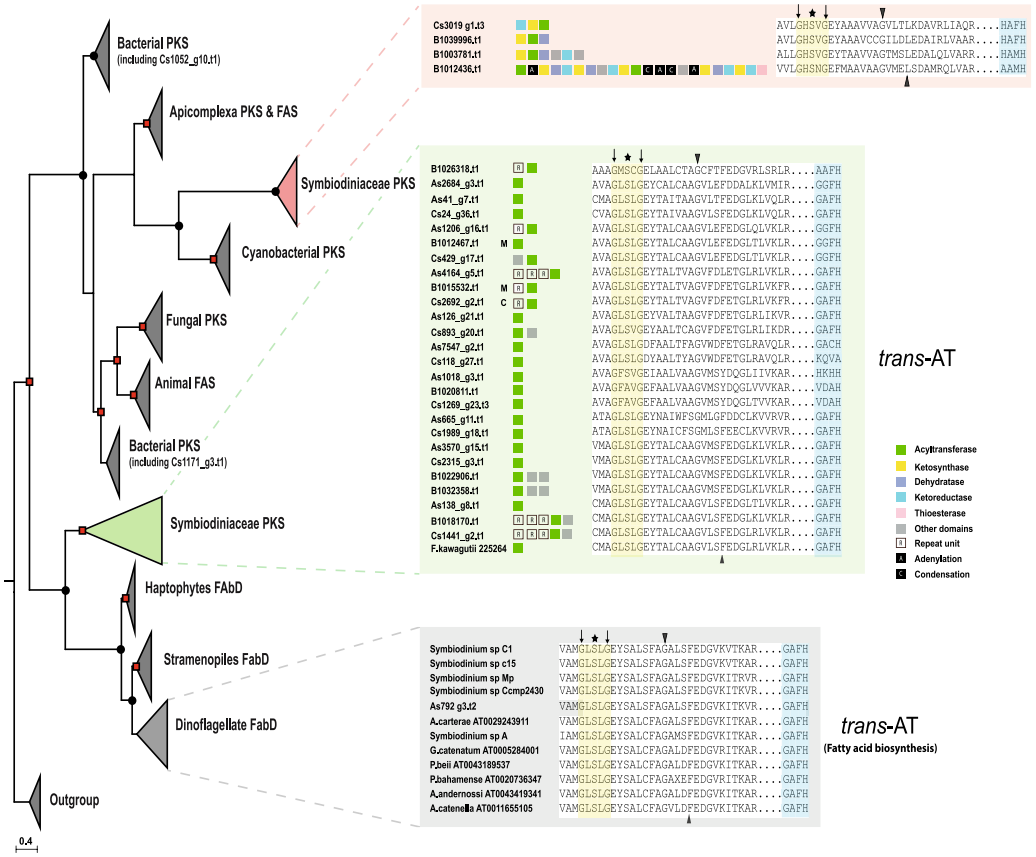

Figure 2. Phylogenetic tree of acyltransferase (AT) domain of prokaryotic and eukaryotic polyketide and fatty acid synthases. Analysis of acyltransferase domain show clear demarcation between *cis*- and *trans*-AT. Dots and square indicate posterior probability 0.70–0.89 and 0.9–1.0, respectively, generated by Bayesian inference. Details of sequences are provided in box inserts. Asterisk indicates active site residue, black triangles indicate conserved residues characteristic for specific substrate groups, and black arrows indicate overall conserved residues used by HMM[25]. The motif, GAFH, present in most Symbiodiniaceae sequences reflects the prediction of I-TASSER. C, M and S depicts chloroplast, mitochondria, and secretory signal peptide, respectively.

Samples B1 and C also showed similar LC-MS profiles and contained some identical unknown SCCs in the molecular weight range of 2,600–2,850 Da (Supplemental Information, Fig. 12b). It should be noted that other polyhydroxy SCCs were also detected in the crude methanol extracts of all samples and none of them corresponds to known zooxanthella polyhydroxy molecules[16].

Analysis using antiSMASH on the three Symbiodiniaceae matched four PKS/NRPS-containing clusters to known biosynthetic gene clusters, with similarities between 25–46% (Fig. 4a). Clade A3 harbors a gene cluster with similarity to ajudazol and phenalamide biosynthetic genes from *Streptomyces* species and *Chondromyces crocatus* while clade B1 shares similarity with a phenalamide biosynthetic cluster from *Chondromyces crocatus*. High sequence similarity was noted in clade A3, offering an example of module duplication between modules of gene models in one scaffold, as well as between modules of different scaffolds (Fig. 4b). To examine the localization of KS protein, antibodies against the KS domain were used. Immunolocalization showed that KS proteins were mainly associated with reticulate chloroplasts in clade C (Supplemental Information, Fig. 13), although the possibility remains that KS proteins are localized in other organelles. Similar observations on the location of KS proteins in chloroplasts have been reported in *Karenia brevis*[30].

Discussion

Evolution of modularity within Symbiodiniaceae genomes. The genomic analysis presented here reveals expanded genetic diversity of metabolite-producing capacity in Symbiodiniaceae dinoflagellates. The polyketide biosynthesis machinery gains its functional and genetic modularity by changes through combinatorial events assisted by gene duplication, horizontal gene transfer (HGT), and recombination[31]. Our analysis shows that domain as well as module duplications established an important evolutionary mechanism toward modularity (Fig. 4b). Dinoflagellate genomes are scattered with large numbers of repeats, with frequent recombination events, and possess genes with high copy numbers due to duplication[17–19]. These features might have led to decomposition of Type I multifunctional PKS clusters, a phenomenon involving shuffling of domains and modules previously observed in bacteria[7]. However, there is increasing evidence of multifunctional PKS domains in several dinoflagellates, indicating that multifunctionality coevolves with monofunctional domains[16,23,32]. Our data show that monofunctional PKSs are related to multifunctional PKS (Fig. 1) but it remains unclear whether

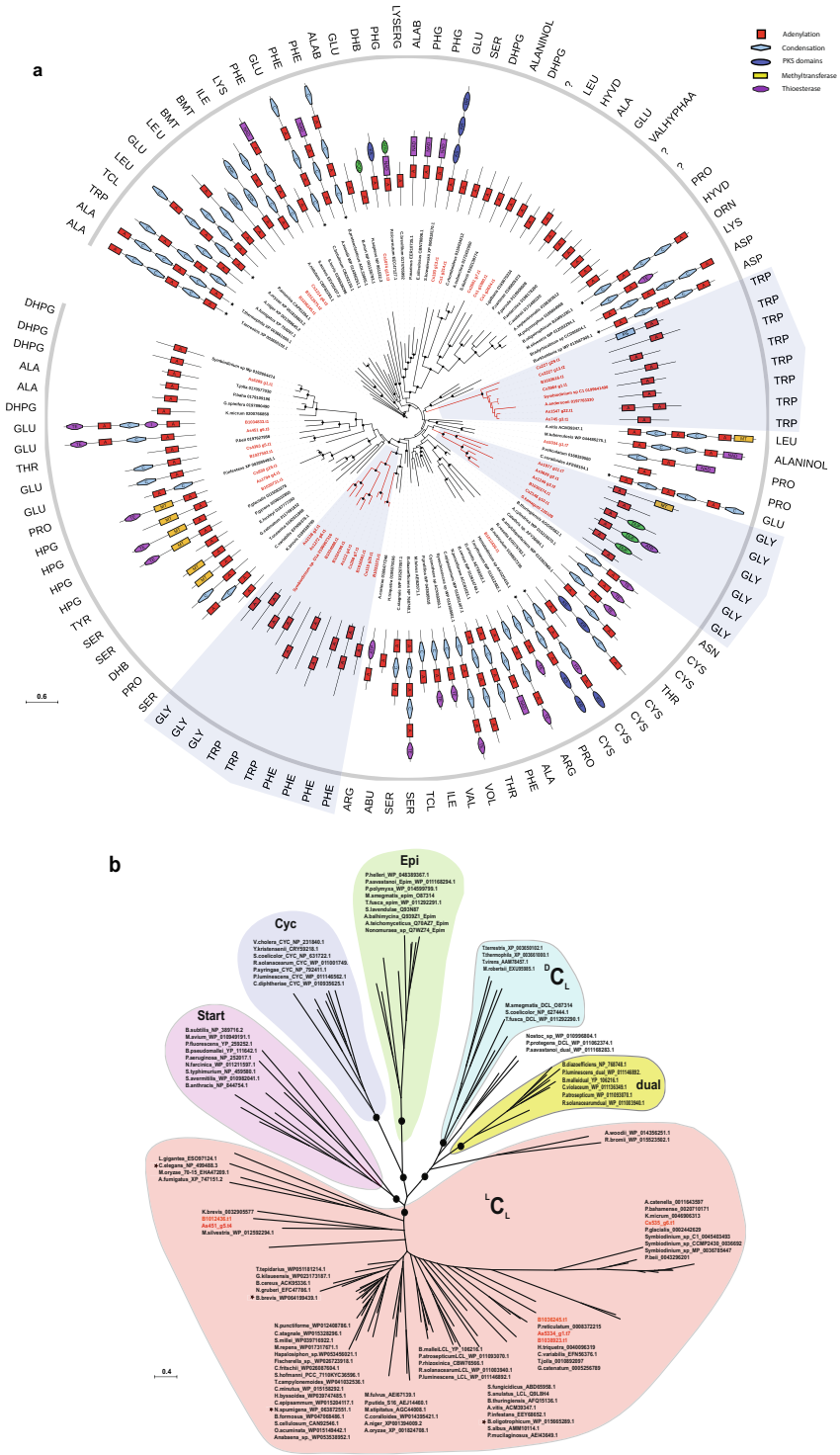

Figure 3. Phylogenetic analysis of adenylation (A) and condensation (C) domains of prokaryotic and eukaryotic NRPS. Dots indicate posterior probability ≥0.70 generated by Bayesian inference. (**a**) Analysis of adenylation domains shows specificity of monofunctional domains from the family Symbiodiniaceae toward three amino acids (glycine, tryptophan, and phenylalanine) as indicated by shaded regions. The specificity of the A-domain was determined using the Latent Semantic Indexing of the LSI-based A-domain predictor[28]. Colored blocked display domain organization and asterisks indicate multifunctional proteins that are too long to display. Details of protein sequences are provided in Supplemental Information, Table S4. (**b**) Condensation domains from Symbiodiniaceae belong to the $^{L}C_{L}$ type (shown in red). Asterisks indicate sequences with different specificities beside group subtype specificity. (Epi = epimerization domain, dual = dual/epimerization domain, Cyc = cyclization domain).

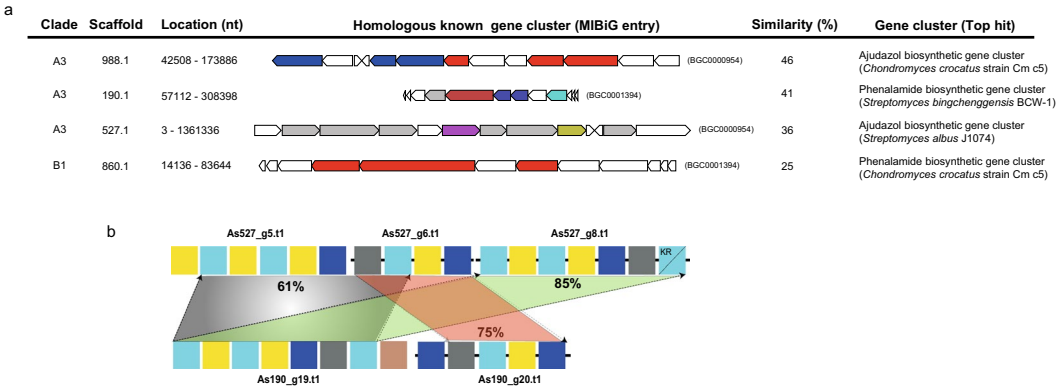

Figure 4. Multifunctional *PKS* genes in Symbiodiniaceae. (**a**) Table showing homologous gene clusters and similarities of different scaffolds from three clades obtained using antiSMASH version 4.1.0. Details of each gene cluster can be obtained using the MIBiG (Minimum Information about a Biosynthetic Gene cluster) entry number and is accessible at https://mibig.secondarymetabolites.org/repository.html. (**b**) Homology comparison of two scaffolds (527.1 and 190.1 of clade A3) shows an example of module duplication. Numbers indicate the percentage of identity shared between sequences. Details of modules are depicted in Fig. 1.

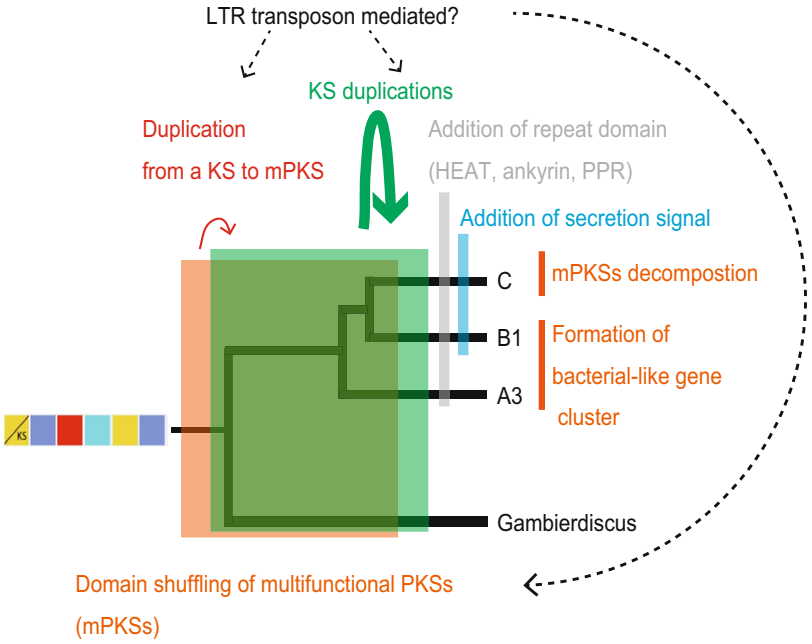

Figure 5. Evolution of *KS* gene in dinoflagellates. Several mechanisms may have contributed to biosynthetic diversification in Symbiodiniaceae. Bacterial-like gene clusters can be conserved and retained in several species. Decomposition of such multifunctional polyketide synthases and extensive duplication may have been mediated by LTR transposons, resulting in addition of secretory signals and repeat domains in the three clades (A3, B1 and C). On the other hand, duplication of a hybrid PKS-NRPS was not detected.

fusion of monofunctional PKS domains led to multifunctionality or *vice versa*. Retrotransposons may have been important contributors in the expansion of PKS and NRPS, since 34% and 47% of the scaffolds, respectively, are predicted to contain LTR signatures (Fig. 5, Supplemental Information, Table S6). Retrogenes account for >20% of all genes in Symbiodiniaceae clades[33]. The Ty1/copia LTR retrotransposon has been proposed as a likely candidate driver for retroposition in *Oxyrrhis marina*[34].

HGT has been suggested as a significant event contributing to gene innovation with recent evidence linking HGT to various biological processes[35]. HGT is thought to contribute to genome innovation in *Fugacium kawagutii*, with 41 out of 56 potential HGT genes being of marine bacterial origin[18]. Gene transfer of *PKS* genes has been suggested in *Karenia brevis*[36]. Multiple rounds of intra- and intergenic gene duplication have been associated with the expansion of the light-harvesting complex (LHC) gene family in *Breviolum minutum* B1, suggesting gene conversion and/or genome rearrangement as an impetus for diversification[37]. Interestingly, monofunctional, probably *trans*-acting domains of either PKS or NRPS, are often fused with repeats units like HEAT (*h*untingtin,

*e*longation factor 3, *A* subunit of protein phosphatase 2A and *T*OR1), ankyrin and pentatricopeptide (PPR) repeats (Fig. 5). HEAT repeats have been found in transport-related proteins while the ankyrin repeat family is the second largest dinoflagellate protein family in *Breviolum minutum* and is known to facilitate protein-protein interactions involved in several intracellular biological processes[17,38–40]. On the other hand, PPR proteins are nuclear-encoded, but are targeted to plastids and mitochondria, where they are involved in RNA processing and editing[41–43].

Evolution of polyketide biosynthesis. Fatty acid synthesis is probably carried out by type II FAS in dinoflagellates, based on a clear distinction between genes involved in fatty acid and polyketide biosynthesis[24]. Our data show that PKS domains have undergone extensive diversification in all three Symbiodiniaceae genomes. A plausible explanation for this expansion might be their involvement in novel functions, supported by the fact that ~15% of KS and ~9% of AT proteins possess targeting signal peptide, directed towards different organelles (Fig. 5). An FAS-like multi-domain polyketide synthase has been identified in *Durinskia baltica*[44], associated with fatty acid biosynthesis. A recent transcriptomic survey of the dinoflagellate *Hematodinium* sp. revealed only Type I FAS[45], while another study on *Gambierdiscus* spp. revealed a distinct Type II FAS system along with single KS domains[23], suggesting a uniqueness of these pathways to specific dinoflagellates. Both Type I and Type II FAS systems can exist, as in *Toxoplasma*[46]. Some taxa possess only cytosolic type I, as in *Cryptosporidinium parvum*, while others have only plastid Type II, as in *Plasmodium falciparum*[47]. Clearly, apicomplexan and dinoflagellate ancestors possessed both systems.

AT domains of *cis*-AT display specificity towards various extender units (e.g. methylmalonyl-CoA, hydroxymalonyl-ACP, methoxymalonyl-ACP, etc) while *trans*-ATs are specific for malonyl-CoA. Stand-alone AT proteins have been reported in several PKSs with modules lacking AT domains and these proteins provide malonyl building blocks for the ACP domains of PKS[48,49]. Our analysis shows that these stand-alone *trans*-AT proteins are dominant in Symbiodiniaceae genomes, forming a major group that may undergo independent evolution compared to canonical *cis*-AT domains. The existence of such *cis*- and *trans*-AT has been reported in bacteria and interpreted as proof of independent evolution[50]. Bacterial *cis*-AT PKS have evolved mainly via module duplication and horizontal/vertical acquisition of entire assembly lines[7], while *trans*-AT tends to recombine and to form novel gene clusters in a mosaic-like fashion[51], as seen in the global pattern of AT in Symbiodiniaceae genomes (Fig. 2). Shelest *et al*. (ref.[8]) found that noniterative PKSs in algae depend mainly on *trans*-AT and are features of multimodular PKS. Interestingly, we observe fragments of genes have been retained even between dinoflagellate genera (e.g. Dinoflagellate PKS-III in Fig. 1), attesting how several evolutionary events such as gene duplication and domain shuffling, with help of repeat domains and LTR retrotransposition have promoted diversification of *PKS* genes (Fig. 5).

Evolution of non-ribosomal peptide biosynthesis. Few studies have reported NRPS in dinoflagellate transcriptomes[52,53]; however, detailed analyses of NRPS remain limited. To our knowledge, this is the first study that look at the role and affinities of adenylation and condensation domains in dinoflagellates. Compared to Type I PKS, NRPSs were reduced in number. NRPSs are known to be less abundant in eukaryotic microalgae[8]. A sequence of amino acids within the A-domain catalytic pocket appears to govern recognition and activation of an amino acid substrate. Thus, any point mutations within this segment can drastically change the specificity of the A-domain. A mono-modular adenylation domain favors incorporation of polar and non-polar amino acids during peptide synthesis (Fig. 3a). A conserved domain organization in mono/bi-modular NRPSs exists in fungal species, implying that this architecture is critical for its function[54]. Single A- or A-T domains can interact with other NRPS components to achieve biosynthesis by successful activation and transfer of the substrate to the C domain in either the same or different NRPS[55]. NRPSs are mainly modular enzymes with several domains; however, there are reports of nonmodular enzymes among fungal subfamilies[54,56].

Conserved secondary metabolic pathways in the family Symbiodiniaceae. Symbiodiniaceae lineages diversified from the ancestral clade A ~160 MYA, at the beginning of the Eocene[1,57] and have adapted to different environments, performing critical functions in reef ecosystems, as well as serving as photosynthetic endosymbionts of different phyla[15]. New genomes now allow us to compare biosynthetic pathways, shedding light on the organization and role of pathways and their contribution to ecological success. Several biosynthetic gene clusters are conserved between *Symbiodinium tridacnidorum* (clade A3), *Breviolum minutum* (clade B1), and *Cladocopium sp*. (clade C) (Supplemental Information, Fig. S6b–d), despite the divergence time[1]. Rosic *et al*. (ref.[58]) reported the importance of conserved phosphatidylinositol signaling pathways in four Symbiodiniaceae clades and their contribution to symbiotic interactions. We found that clades A3 and B1 produce unique polyketides, supporting the clade-metabolite hypothesis[13]. Metabolite profiles of different Symbiodiniaceae species are influenced by different temperatures and light regimes[59]. On the other hand, metabolomic similarity was detected only between clades B1 and C. At this stage, it is difficult to link specific metabolites to specific pathways, but these results suggest that novel pathways must have evolved in the common ancestor of clades B1 and C to provide a common set of metabolites, irrespective of their hosts and environments. Biological systems regulate biochemical and cellular processes when subjected to environmental changes[60]. This study shows that Symbiodiniaceae genomes encode PKS and NRPS enzymes with broad substrate tolerance as a cost-effective way of generating chemical diversity.

The Screening hypothesis suggest that organisms that produce many chemicals, have more chances of enhanced fitness because greater chemical diversity increases the chance of producing metabolites with unique traits, as illustrated by zooxanthellatoxins and zooxanthellamides[61]. So why are only a few major pathways conserved among these species? It might be beneficial for organisms to elongate existing pathways to generate new chemical diversity, instead of originating entirely new pathways[62]. Dinoflagellates are known to form harmful

algal blooms, that negatively affect ecosystems via the accumulation of toxins through food webs that can cause classical seafood poisoning. Thus, insights into their biosynthesis can provide new ways for detection of toxin in environmental samples[63]. From a biotechnological perspective, such novel polyketide biochemistries can provide valuable tools for the combinatorial biosynthesis of future medicines[64].

In conclusion, we surveyed three genomes for genes associated with secondary metabolism. We showed that PKS genes are more diversified than NRPS genes and that several evolutionary processes have contributed to this diversification. Furthermore, these genes displayed a degree of substrate specificity and flexibility that has been maintained evolutionarily, irrespective of host system. These results demonstrate that Symbiodiniaceae genomes are well equipped to generate chemical diversity when it comes to secondary metabolite biosynthesis. This comparative genomic study provides preliminary insights into how dinoflagellate genomes adapt to hosts' environment and addresses the functional roles of secondary metabolites in such symbiotic relationships.

Methods

Symbiodiniaceae cultures. *Breviolum minutum* (Clade B1, strain Mf1.05b) was isolated from the stony coral, *Montastraea (Orbicella) faveolata* by Dr. Mary Alice Coffroth (University of New York, Buffalo, USA) and *Symbiodinium tridacnidorum* (clade A3, strain Y106) and *Cladocopium* sp. (clade C, strain Y103) were isolated from the clam *Tridacna crocea* and bivalve *Fragum* sp., respectively, by late Dr. Terufumi Yamasu (University of the Ryukyus, Okinawa, Japan). Cultures were maintained in autoclaved, artificial seawater containing 1X Guillard's (F/2) marine-water enrichment solution (Sigma-Aldrich: G0154), supplemented with antibiotics (ampicillin (100 μg/mL), kanamycin (50 μg/mL), and streptomycin (50 μg/mL)). Culturing and sampling were performed according to the protocol of Shoguchi *et al.* (ref.[17]).

Data retrieval. Throughout this manuscript, we adopted the revised terminology[1] but retain the previous familiar clade terminology and tag gene models from the three Symbiodiniaceae genomes (A3, B1 and C) with the letters A, B, and C to improve the readability and interpretation. To understand diversification and molecular evolution of PKS and FAS, we performed an extensive search for *PKS* (*KS* & *AT*) and *FAS* (*FabB-KASI*, *FabF-KASII* & *FabD*) genes within three Symbiodiniaceae genomes, as these domains are conserved[65]. The genome browser MarinegenomicsDB (http://marinegenomics.oist.jp/genomes/gallery/) and the *Fugacium kawagutii* browser (http://web.malab.cn/symka_new/genome.jsp) were accessed in order to retrieve PKS (KS & AT), FAS (FabB-KASI, FabF-KASII & FabD) and NRPS (A & C) sequences from clades A3, B1, and C and *Fugacium kawagutii*, respectively[18,66]. In addition, transcriptome datasets for several dinoflagellates, apicomplexans, stramenopiles, and haptophytes were downloaded from the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (http://datacommons.cyverse.org/browse/iplant/home/shared/imicrobe/camera) and surveyed for comparative analysis[67]. Amino acid sequences of several other prokaryotes, fungal, animal, and chlorophyte PKS and NRPS domains were obtained from NCBI Genbank with additional sequences from dinoflagellates[23,68]. Further NRPS sequences from *Proteobacteria*, *Firmicutes*, and *Cyanobacteria* were obtained from Wang *et al.* (ref.[5]). Functional prediction and conserved active site residues in sequences were identified using Pfam[69]. Only PKS, FAS, and NRPS sequences with full domains and conserved active sites were used in the analysis. Details of gene IDs and their transcriptome support are provided in Supplemental Information, Tables S1–S4.

Phylogenetic analysis. Type I and II PKS/FAS and condensation (C) & adenylation (A) domain sequences representing different taxa were used for Bayesian inference and maximum likelihood analysis. Four amino acid (aa) domain sequence datasets comprising of 233 KS sequences (226 aa), 96 AT sequences (208 aa), 117 A-sequences (400 aa), and 110 C-sequences (260 aa) were aligned using the MUSCLE algorithm[70]. Sites within alignments where homology was ambiguous (e.g. large insertions and deletions) were removed prior to phylogenetic analyses. Maximum likelihood phylogenetic analysis was performed using RaxML with 1000 bootstraps using the GAMMA and Le-Gasquel amino acid replacement matrix[71]. Bayesian inference was conducted with MrBayes v.3.2[72] using the same replacement model and run to maximum of six million generations and four chains or until the posterior probability approached 0.01. Statistics and trees were summarized using a burn-in of 25% of the data. Using two methods provided a convenient way to verify different phylogenetic estimates, since each method has its intrinsic strengths and assumptions about the evolutionary process. Trees were edited using Figtree (http://tree.bio.ed.ac.uk/software/figtree/).

***In silico* analysis of PKS and NRPS genes and genomic locations.** Monomer prediction based on specificity of the A-domain was determined using the Latent Semantic Indexing of the LSI-based A-domain predictor[28]. NaPDos was used to determine C-domain types[73]. For AT domains, sequences were compared to the Hidden Markov Model-based ensemble (HMM) generated by Khayatt *et al.* (ref.[25]). Additional information about substrate specificity was detected using I-TASSER[74]. AntiSMASH (Antibiotics & Secondary Metabolite Analysis SHell) version 4.1.0 was used with default settings to identify *NRPS* and *PKS* gene clusters within scaffold regions using nucleotides sequences as queries[75]. Subcellular localization of PKS gene products toward organelles (e.g. chloroplast and mitochondria) or the presence of signal peptide or membrane anchor was determined using ChloroP 1.1 and TargetP 1.1 using a cut-off score of $\geq$0.50 each and the subcellular localization predictor, DeepLoc[76–78]. NUCmer operation of SyMap v4.2 (Synteny Mapping and Analysis Program) was used to align and visualize syntenic relationships between the three clades[79]. Scaffold information and descriptions of these genomes were imported into SyMap as GFFs (General Feature Files). To determine orthologs, we performed an all-against-all BLAST search of *PKS*-coding scaffolds of one genome against itself at a BLAST bit score cutoff of $\geq$100 and e-value $\leq e^{-20}$. Outputs were parsed and processed, and orthologous pair detection was conducted using custom perl scripts. Possible segmental duplications were visualized using Circos[80]. GC content variations

in *PKS*-coding scaffolds were analyzed using GC-profile with a halting parameter of 100[81]. LTR Finder 1.05 was used with defaults parameters to search for long terminal repeat (LTR) retrotransposon-specific features[82].

Polyol extraction from Symbiodiniaceae cultures. All cultured biomass samples were treated as previously described[8]. Cultured cells were collected by centrifugation (9,000 × g and 14,000 g, 10 min, 10 °C). After discarding the supernatant, a cell pellet was extracted with methanol (three times) at room temperature. Methanol (400 μL) was added to the biomass followed by vortexing (1 min), sonication (10 min), and centrifugation (14,000 × g, 10 min, 10 °C) to give a methanol extract. The resulting clear solution was transferred into a new tube. By adding methanol (400 μL) to the residue, a 2nd methanol extraction was conducted in the same fashion. The 2nd clear methanol extract was again collected and stored at −30 °C. Additional methanol (400 μL) was added to the residue, vortexed (1 min), and kept overnight at room temperature. After centrifugation, the 3rd methanol extract was pooled with the previous extracts (total 1,200 μL), and marked as crude extract. To remove lipophilic materials from the crude extract, an aliquot (50 μL) of the crude extract was suspended in 50 μL water-methanol (90:10) containing 0.5% formic acid. The suspension was vortexed (30 sec) and centrifuged (14,000 × g, 10 min, 10 °C) to give a clean solution. The clean solution was transferred into a new tube (the stock solution) and the insoluble part was discarded. The stock solution was kept at −30 °C before NanoLC-MS analysis or immediately analyzed after dilution. All crude extracts were lyophilized and stored at −30 °C.

NanoLC-MS analysis of Symbiodiniaceae methanol 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* 400–2,000 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 (stock solution) was diluted (1:50) by adding water-methanol (50:50) containing 0.25% formic acid and separated on a capillary ODS column (50 × 0.18 mm, 3 μm, C_{18}, Supelco). A 20-min gradient (10% B for 0–2 min, 10–100% B for 2–10 min, hold 100% B for 10–15 min, equilibration 10% B for 15.1–20.0 min, where solvent A was water: acetonitrile 98:2 and solvent B was water: acetonitrile 2:98, both containing 0.1% formic acid; flow rate 2.0 μL/min, injection, 2.0 μL) was used for polyol separation.

KS protein localization. KS protein localizations were visualized using a modified version of the protocol of Berdieva *et al.* (ref.[83]). Briefly, cells were prefixed in methanol: F/2 medium (1:1) at RT for 15 min. Samples were then fixed in methanol at −20 °C overnight. Cells were washed in PBS, followed by permeabilization with 1% Triton X-100 for 15 min (5 min for clade B1), further washed with PBS and blocked with 5% normal goat serum-PBST for 1 h. Subsequently cells were incubated overnight at 4 °C with primary anti-KS antibodies at a 1:100 dilution in blocking solution. Primary antibody solution was then removed with 3 × 5-min PBS washes and cells were incubated with Alexa Fluor 488 (Abcam Cat #ab150077) secondary antibody for 1 h at RT (1:100 in blocking solution), ending with several PBS washes. Coverslips were mounted in DAPI-containing Vectashield on glass slides and visualized using a Zeiss Axio-Observer Z1 LSM780 confocal microscope under a Plan-APOCHROMAT 63X/1.4 oil DIC objective lens. Fluorescence excitation/emission wavelengths were 410/482 nm for DAPI, 499/614 nm for Alexa Fluor 488, and 649/740 nm for chlorophyll autofluorescence. Data were acquired using Zeiss ZEN version 14.0.8.201 software. For negative controls, primary antibodies were omitted. Z-stacks profiles were analyzed using ImageJ[84]. DIC imaging was performed using a Zeiss Image-Z1 under 40X.

Data Availability

The datasets supporting the conclusions of this article are available are accessible in the DDBJ/EMBL/NCBI database with BioProject IDs PRJDB3242 (clade A3), PRJDB732 (clade B1), and PRJDB3243 (clade C), respectively. Raw data for metabolite profiling is accessible at the genome browser site (http://marinegenomics.oist.jp/gallery/).

References

1. LaJeunesse, T. C. *et al.* Systematic revision of Symbiodiniaceae highlights the antiquity and diversity of coral endosymbionts. *Curr. Biol.* **28**, 2570–2580.e6, https://doi.org/10.1016/j.cub.2018.07.008 (2018).
2. Trench, R. K. The Cell Biology of Plant-Animal Symbiosis. *Annu Rev Plant Physiol Plant Mol Biol.* **30**, 485–531, https://doi.org/10.1146/annurev.pp.30.060179.002413 (1979).
3. Lewis, D. H. & Smith, D. C. The autotrophic nutrition of symbiotic marine coelenterates with special reference to hermatypic Corals. I. Movement of photosynthetic products between the symbionts. *Proc R Soc Lond B Bio Sci.* **178**, 111–129, https://doi.org/10.1098/rspb.1971.0055 (1971).
4. Uemura, D. Bioactive *polyethers*. In: Scheuer, P. J. (ed.) *Bioorganic Marine Chemistry*, Vol 4, 1–31 (Springer-Verlag,1991).
5. 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, https://doi.org/10.1073/pnas.1401734111 (2014).
6. Hertweck, C. The Biosynthetic Logic of Polyketide Diversity. *Angew Chem Int Ed* **48**, 4688–4716, https://doi.org/10.1002/anie.200806121 (2009).
7. Jenke-Kodama, H., Sandmann, A., Müller, R. & Dittmann, E. Evolutionary implications of bacterial polyketide synthases. *Mol Biol Evol.* **22**, 2027–2039, https://doi.org/10.1093/molbev/msi193 (2005).
8. 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, https://doi.org/10.1186/s12864-015-2222-9 (2015).
9. Schwarzer, D., Finking, R. & Marahiel, M. A. Nonribosomal peptides: from genes to products. *Nat Prod Rep.* **20**, 275–287, https://doi.org/10.1039/B111145K (2003).
10. Marahiel, M. A., Stachelhaus, T. & Mootz, H. D. Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chem Rev.* **97**, 2651–2674, https://doi.org/10.1021/cr960029e (1997).
11. Du, L., Sánchez, C. & Shen, B. Hybrid peptide-polyketide natural product: biosynthesis and prospects toward engineering novel molecules. *Metab Eng.* **3**, 78–95, https://doi.org/10.1006/mben.2000.0171 (2001).

12. Fischbach, M. A., Walsh, C. T. & Clardy, J. The evolution of gene collectives: How natural selection drives chemical innovation. *Proc Natl Acad Sci USA* **105**, 4601–8, https://doi.org/10.1073/pnas.0709132105 (2008).
13. Fukatsu, T. *et al*. Zooxanthellamide D, a polyhydroxy polyene amide from a marine dinoflagellate, and chemotaxonomic perspective of the *Symbiodinium* Polyols. *J Nat Prod.* **70**, 407–411, https://doi.org/10.1021/np060596p (2007).
14. Nakamura, H., Kawase, Y., Maruyama, K. & Muria, A. 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, https://doi.org/10.1246/bcsj.71.781 (1998).
15. Gordon, B. R. & Leggat, W. Symbiodinium-Invertebrate symbioses and the role of metabolomics. *Mar Drugs* **8**, 2546–2568, https://doi.org/10.3390/md8102546 (2010).
16. Beedessee, G., Hisata, K., Roy, M. C., Satoh, N. & Shoguchi, E. Multifunctional polyketide synthase genes identified by genomic survey of the symbiotic dinoflagellate, *Symbiodinium minutum*. *BMC Genomics* **16**, 941, https://doi.org/10.1186/s12864-015-2195-8 (2015).
17. Shoguchi, E. *et al*. Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. *Curr Biol* **23**, 1399–1408, https://doi.org/10.1016/j.cub.2013.05.062. (2013).
18. Lin, S. *et al*. The *Symbiodinium kawagutii* genome illuminates dinoflagellate gene expression and coral symbiosis. *Science* **350**, 691–694, https://doi.org/10.1126/science.aad0408 (2015).
19. Aranda, M. *et al*. Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle. *Sci Rep.* **6**, 39734, https://doi.org/10.1038/srep39734 (2016).
20. Shoguchi, E. *et al*. Two divergent *Symbiodinium* genomes reveal conservation of a gene cluster for sunscreen biosynthesis and recently lost genes. *BMC Genomics* **19**, 458, https://doi.org/10.1186/s12864-018-4857-9 (2018).
21. Williams, D. H., Stone, M. J., Hauck, P. R. & Rahman, S. K. Why are secondary metabolites (natural products) biosynthesized? *J Nat Prod.* **52**, 1189–1208, https://doi.org/10.1021/np50066a001 (1989).
22. Murray, S. A. *et al*. Unravelling the functional genetics of dinoflagellates: A review of approaches and opportunities. *Perspect Phycol.* **3**, 37–52, https://doi.org/10.1127/pip/2016/0039 (2016).
23. Kohli, G. S. *et al*. Role of Modular Polyketide Synthases in the Production of Polyether Ladder Compounds in Ciguatoxin-Producing Gambierdiscus polynesiensis and G. excentricus (Dinophyceae). *J Eukaryot Microbiol.* **64**, 691–706, https://doi.org/10.1111/jeu.12405 (2017).
24. Kohli, G. S., John, U., Van Dolah, F. M. & Murray, S. A. Evolutionary distinctiveness of fatty acid and polyketide synthesis in eukaryotes. *ISME J.* **10**, 1877–1890, https://doi.org/10.1038/ismej.2015.263. (2016).
25. Khayatt, B. I., Overmars, L., Siezen, R. J. & Francke, C. 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, https://doi.org/10.1371/journal.pone.0062136 (2013).
26. Tang, Y., Kim, C.-Y., Mathews, I. I., Cane, D. E. & Khosla, C. 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, https://doi.org/10.1073/pnas.0601924103 (2006).
27. Stachelhaus, T., Mootz, H. D. & Marahiel, M. A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem Biol.* **6**, 493–505, https://doi.org/10.1016/S1074-5521(99)80082-9 (1999).
28. Baranašić, D. *et al*. 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, https://doi.org/10.1007/s10295-013-1322-2 (2014).
29. Rausch, C., Hoof, I., Weber, T., Wohlleben, W. & Huson, D. H. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol Biol.* **7**, 78, https://doi.org/10.1186/1471-2148-7-78 (2007).
30. Monroe, E. A., Johnson, J. G., Wang, Z., Pierce, R. K. & Van Dolah, F. M. Characterization and expression of nuclear-encoded polyketide synthases in the brevetoxin-producing dinoflagellate *Karenia brevis*. *J Phycol.* **46**, 541–552, https://doi.org/10.1111/j.1529-8817.2010.00837.x (2010).
31. Thattai, M., Burak, Y. & Shraiman, B. The origins of specificity in polyketide synthase protein interactions. *PLoS Comput Biol.* **3**, e186, https://doi.org/10.1371/journal.pcbi.0030186 (2007).
32. Van Dolah, F. M., Kohli, G. S., Morey, J. S. & Murray, S. A. Both modular and single-domain Type I polyketide synthases are expressed in the brevetoxin-producing dinoflagellate, Karenia brevis (Dinophyceae). *J Phycol.* **53**, 1325–1339, https://doi.org/10.1111/jpy.12586 (2017).
33. Song, B. *et al*. Comparative genomics reveals two major bouts of gene retroposition coinciding with crucial periods of Symbiodinium evolution. *Genome Biol Evol.* **9**, 2037–2047, https://doi.org/10.1093/gbe/evx144 (2017).
34. Lee, R. *et al*. Analysis of EST data of the marine protist *Oxyrrhis marina*, an emerging model for alveolate biology and evolution. *BMC Genomics* **15**, 122, https://doi.org/10.1186/1471-2164-15-122 (2014).
35. Wisecaver, J. H., Brosnahan, M. L. & Hackett, J. D. Horizontal gene transfer is a significant driver of gene innovation in dinoflagellates. *Genome Biol Evol.* **5**, 2368–2381, https://doi.org/10.1093/gbe/evt179 (2013).
36. Lopez-Legentil, S., Song, B., DeTure, M. & Baden, D. G. 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, https://doi.org/10.1007/s10126-009-9197-y (2010).
37. Maruyama, S., Shoguchi, E., Satoh, N. & Minagawa, J. Diversification of the light-harvesting complex gene family via intra-and intergenic duplications in the coral symbiotic alga Symbiodinium. *PLoS ONE* **10**, e0119406, https://doi.org/10.1371/journal.pone.0119406 (2015).
38. Cook, A., Bono, F., Jinek, M. & Conti, E. Structural biology of nucleocytoplasmic transport. *Annu Rev Biochem.* **76**, 647–671, https://doi.org/10.1146/annurev.biochem.76.052705.161529 (2007).
39. Bennett, V. & Baines, A. J. Spectrin and ankyrin-based pathways: metazoan inventions for integrating cells into tissues. *Physiol Rev.* **81**, 1353–1392, https://doi.org/10.1152/physrev.2001.81.3.1353 (2001).
40. Mosavi, L. K., Cammett, T. J., Desrosiers, D. C. & Peng, Z. Y. The ankyrin repeat as molecular architechture for protein recognition. *Protein Sci.* **13**, 1435–1448, https://doi.org/10.1110/ps.03554604 (2004).
41. Colcombet, J. *et al*. Systematic study of subcellular localization of *Arabidopsis* PPR proteins confirms a massive targeting to organelles. *RNA Biol.* **10**, 1557–1575, https://doi.org/10.4161/rna.26128 (2013).
42. Fujii, S. & Small, I. The evolution of RNA editing and pentatricopeptide repeat genes. *New Phytologist* **191**, 37–47, https://doi.org/10.1111/j.1469-8137.2011.03746.x (2011).
43. Nakamura, T., Yagi, Y. & Kobayashi, K. Mechanistic Insight into Pentatricopeptide Repeat Proteins as Sequence-Specific RNA-Binding Proteins for Organellar RNAs in Plants. *Plant Cell Physiol.* **53**, 1171–1179, https://doi.org/10.1093/pcp/pcs06 (2012).
44. Hehenberger, E., Burki, F., Kolisko, M. & Keeling, P. J. Functional Relationship between a Dinoflagellate Host and Its Diatom Endosymbiont. *Mol Biol Evol.* **33**, 2376–2390, https://doi.org/10.1093/molbev/msw109 (2016).
45. Gornik, S. G. *et al*. Endosymbiosis undone by stepwise elimination of the plastid in a parasitic dinoflagellate. *Proc Natl Acad Sci USA* **112**, 5767–5772, https://doi.org/10.1073/pnas.1423400112 (2015).
46. Seeber, F. & Soldati-Favre, D. Metabolic pathways in the apicoplast of apicomplexa. *Int Rev Cell Mol Biol.* **281**, 161–228, https://doi.org/10.1016/S1937-6448(10)81005-6. (2010).
47. Zhu, G. Current Progress in the Fatty Acid Metabolism in *Cryptosporidium parvu*m. *J Eukaryot Microbiol.* **51**, 381–388, https://doi.org/10.1111/j.1550-7408.2004.tb00384.x (2004).

48. Piel, J. A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of Paederus beetles. *Proc Natl Acad Sci USA* **99**, 14002–14007, https://doi.org/10.1073/pnas.222481399 (2002).
49. Cheng, Y. Q., Tang, G. L. & Shen, B. Type I polyketide synthase requiring a discrete acyltransferase for polyketide biosynthesis. *Proc Natl Acad Sci USA* **100**, 3149–3154, https://doi.org/10.1073/pnas.0537286100 (2003).
50. Piel, J., Hui, D., Fusetani, N. & Matsunaga, S. Targeting modular polyketide synthases with iteratively acting acyltransferases from metagenomes of uncultured bacterial consortia. *Environ Microbiol.* **6**, 921–927, https://doi.org/10.1111/j.1462-2920.2004.00531.x (2004).
51. Nguyen, T. *et al*. Exploiting the mosaic structure of trans-acyltransferase polyketide synthases for natural product discovery and pathway dissection. *Nat Biotechnol.* **26**, 225–233, https://doi.org/10.1038/nbt1379 (2008).
52. Salcedo, T., Upadhyay, R. J., Nagasaki, K. & Bhattacharya, D. Dozens of toxin-related genes are expressed in a nontoxic strain of the dinoflagellate *Heterocapsa circularisquama*. *Mol Biol Evol.* **29**, 1503–1506, https://doi.org/10.1093/molbev/mss007 (2012).
53. Cooper, J. T., Sinclair, G. A. & Wawrik, B. Transcriptome analysis *of Scrippsiella trochoidea* CCMP 3099 reveals physiological changes related to nitrate depletion. *Front Microbiol.* **7**, 639, https://doi.org/10.3389/fmicb.2016.00639 (2016).
54. Bushley, K. E. & Turgeon, B. G. Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. *BMC Evol Biol.* **10**, 26, https://doi.org/10.1186/1471-2148-10-26 (2010).
55. Mootz, H. D., Schwarzer, D. & Marahiel, M. A. Ways of assembling complex natural products on modular nonribosomal peptide synthetases. *ChemBioChem* **3**, 490–504, https://doi.org/10.1111/j.1529-8817.2010.00837.x (2002).
56. Donadío, S., Monciardini, P. & Sosio, M. Polyketide synthases and nonribosomal peptide synthetases: The emerging view from bacterial genomics. *Nat Prod Rep.* **24**, 1073–1109, https://doi.org/10.1039/B514050C (2007).
57. Pochon, X., Montoya-Burgos, J., Stadelmann, B. & Pawlowski, J. Molecular phylogeny, evolutionary rates, and divergence timing of the symbiotic dinoflagellate genus Symbiodinium. *Mol Phylogenet Evol.* **38**, 20–30, https://doi.org/10.1016/j.ympev.2005.04.028 (2006).
58. Rosic, N. *et al*. Unfolding the secrets of coral-algal symbiosis. *ISME J.* **9**, 844–856, https://doi.org/10.1038/ismej.2014.182 (2015).
59. Klueter, A., Crandall, J., Archer, F., Teece, M. & Coffroth, M. Taxonomic and environmental variation of metabolite profiles in marine dinoflagellates of the genus Symbiodinium. *Metabolites* **5**, 74–99, https://doi.org/10.3390/metabo5010074 (2015).
60. Hannah, M. A. *et al*. 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, https://doi.org/10.1104/pp.109.147306 (2010).
61. Jones, C. G. & Firn, R. D. On the evolution of plant secondary metabolite chemical diversity. *Phil Trans R Soc Lond B* **333**, 273–280, https://doi.org/10.1098/rstb.1991.0077 (1991).
62. Firn, R. D. & Jones, C. G. Natural products-a simple model to explain chemical diversity. *Nat Prod Rep.* **20**, 382–391, https://doi.org/10.1039/B208815K (2003).
63. Kellmann, R., Stüken, A., Orr, R. J., Svendsen, H. M. & Jakobsen, K. S. Biosynthesis and molecular genetics of polyketides in marine dinoflagellates. *Mar Drugs.* **8**, 1011–48, https://doi.org/10.3390/md8041011 (2010).
64. Rein, K. S. & Snyder, R. V. The biosynthesis of polyketide metabolites by dinoflagellates. *Adv Appl Microbiol.* **59**, 93–125, https://doi.org/10.1016/S0065-2164(06)59004-5 (2006).
65. Kroken, S., Glass, N. L., Taylor, J. W., Yoder, O. C. & Turgeon, B. G. Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. *Proc Natl Acad Sci USA* **100**, 15670–15675, https://doi.org/10.1073/pnas.2532165100 (2003).
66. Koyanagi, R. *et al*. MarinegenomicsDB: An integrated genome viewer for community-based annotation of genomes. *Zool Sci.* **30**, 797–800, https://doi.org/10.2108/zsj.30.797 (2013).
67. Keeling, P. J. *et al*. 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, https://doi.org/10.1371/journal.pbio.1001889 (2014).
68. Eichholz, K., Beszteri, B. & John, U. Putative monofunctional type I polyketide synthase units: A dinoflagellate-specific feature? *PLoS One* **7**, e48624, https://doi.org/10.1371/journal.pone.0048624 (*2012*).
69. Punta, M. *et al*. The Pfam protein families database. *Nucl Acids Res.* **40**, D290–D301, https://doi.org/10.1093/nar/gkr1065 (2012).
70. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl Acids Res.* **32**, 1792–1797, https://doi.org/10.1093/nar/gkh340 (2004).
71. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313, https://doi.org/10.1093/bioinformatics/btu033 (2014).
72. Ronquist, F. *et al*. MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Syst Biol.* **61**, 539–542, https://doi.org/10.1093/sysbio/sys029 (2012).
73. Ziemert, N. *et al*. The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity. *PLoS ONE* **7**, e34064, https://doi.org/10.1371/journal.pone.0034064 (2012).
74. Zhang, Y. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**, 40, https://doi.org/10.1186/1471-2105-9-40 (2008).
75. Blin, K. *et al*. antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucl Acids Res.* **45**, W36–W41, https://doi.org/10.1093/nar/gkx319 (2017).
76. Emanuelsson, O., Nielsen, H. & von Heijne, G. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.* **8**, 978–984, https://doi.org/10.1110/ps.8.5.978 (1999).
77. Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc.* **2**, 953–971, https://doi.org/10.1038/nprot.2007.131 (2007).
78. Armenteros, J. J. A., Sønderby, C. K., Sønderby, S. K., Nielsen, H. & Winther, O. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* **33**, 3387–3395, https://doi.org/10.1093/bioinformatics/btx431 (2017).
79. Soderlund, C., Bomhoff, M. & Nelson, W. SyMAPv3.4: a turnkey synteny system with application to plant genomes. *Nucl Acids Res.* **39**, e68, https://doi.org/10.1093/nar/gkr123 (2011).
80. Krzywinski, M. *et al*. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645, https://doi.org/10.1101/gr.092759.109 (2009).
81. Gao, F. & Zhang, C. T. GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. *Nucl Acids Res.* **34**, W686–W691, https://doi.org/10.1093/nar/gkl040 (2006).
82. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucl Acids Res.* **35**, W265–268, https://doi.org/10.1093/nar/gkm286 (2007).
83. Berdieva, M., Pozdnyakov, I., Matantseva, O., Knyazev, N. & Skarlato, S. Actin as a cytoskeletal basis for cell architecture and a protein essential for ecdysis in *Prorocentrum minimum* (Dinophyceae, Prorocentrales). *Phycol Res.* **66**, 127–136, https://doi.org/10.1111/pre.12214 (2018).
84. Schindelin, J. *et al*. Fiji: an open-source platform for biological-image analysis. *Nature methods* **9**, 676–682, https://doi.org/10.1038/nmeth.2019 (2012).

Acknowledgements

GB is supported by a Japanese Society for the Promotion of Science (JSPS) Research Fellowship for Young Scientists (17J00597). This work was supported partly by JSPS (no. K07454 to E.S) and by generous funding by Okinawa Institute of Science and Technology Graduate University to the Marine Genomics Unit. We thank Steven D. Aird for editing the manuscript. The authors are grateful to Dr. Mary Alice Coffroth and Dr. Michio Hidaka for providing the samples. The authors are thankful to Dr. Chuya Shinzato (The University of Tokyo, Japan) for helpful comments on genomic analysis and to the OIST sequencing and imaging sections for their support.

Author Contributions

G.B., E.S., and N.S. conceived and designed the research. G.B., K.H., and E.S. analyzed the genomic data. M.R. performed the mass sample preparation, data acquisition, and data interpretation. G.B., M.R., F.V.D., and E.S. wrote the paper. All authors read and approved the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-37792-0.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.