

Structural rearrangements drive extensive genome divergence between symbiotic and free-living *Symbiodinium*

Raúl A. González-Pech¹, Timothy G. Stephens¹, Yibi Chen¹, Amin R. Mohamed², Yuanyuan Cheng^{3,†}, David W. Burt³, Debashish Bhattacharya⁴, Mark A. Ragan¹, Cheong Xin Chan^{1,5*}

¹Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia

²Commonwealth Scientific and Industrial Research Organisation (CSIRO) Agriculture and Food, Queensland Bioscience Precinct, St Lucia, QLD 4072, Australia

³UQ Genomics Initiative, The University of Queensland, Brisbane, QLD 4072, Australia

⁴Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ 08901, U.S.A.

⁵School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD 4072, Australia

[†]Current address: School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW 2006, Australia

*Corresponding author (c.chan1@uq.edu.au)

Abstract

Symbiodiniaceae are predominantly symbiotic dinoflagellates critical to corals and other reef organisms. *Symbiodinium* is a basal symbiodiniacean lineage and includes symbiotic and free-living taxa. However, the molecular mechanisms underpinning these distinct lifestyles remain little known. Here, we present high-quality *de novo* genome assemblies for the symbiotic *Symbiodinium tridacnidorum* CCMP2592 (genome size 1.3 Gbp) and the free-living *Symbiodinium natans* CCMP2548 (genome size 0.74 Gbp). These genomes display extensive sequence divergence, sharing only ~1.5% conserved regions ($\geq 90\%$ identity). We predicted 45,474 and 35,270 genes for *S. tridacnidorum* and *S. natans*, respectively; of the 58,541 homologous gene families, 28.5% are common to both genomes. We recovered a greater extent of gene duplication and higher abundance of repeats, transposable elements and pseudogenes in the genome of *S. tridacnidorum* than in that of *S. natans*. These findings demonstrate that genome structural rearrangements are pertinent to distinct lifestyles in *Symbiodinium*, and may contribute to the vast genetic diversity within the genus, and more broadly in Symbiodiniaceae. Moreover, the results from our whole-genome comparisons against a free-living outgroup support the notion that the symbiotic lifestyle is a derived trait in, and that the free-living lifestyle is ancestral to, *Symbiodinium*.

32 Introduction

33 Symbiodiniaceae are dinoflagellates (Order Suessiales) crucial for coral reefs because of their
 34 symbiotic relationship with corals and diverse marine organisms. Although these dinoflagellates do
 35 not display evident morphological diversity, their extensive genetic variation is well-recognised,
 36 prompting the recent systematic revision to family status^{1,2}. Sexual reproduction stages have not
 37 been directly observed in Symbiodiniaceae, but the presence of a complete meiotic gene repertoire
 38 suggests that they are able to reproduce sexually³⁻⁵. The potential sexual reproduction of
 39 Symbiodiniaceae has been used to explain their extensive genetic variation⁶⁻¹⁰.

40 The genetic diversity in Symbiodiniaceae is in line with their broad range of symbiotic associations
 41 with other organisms, covering a broad spectrum depending on host specificity, transmission mode
 42 and permanence in the host^{11,12}. Furthermore, some taxa are considered free-living because they
 43 have been found only in environmental samples, and in experiments fail to infect potential hosts<sup>13-
 44 15</sup>.

45 The basal lineage of Symbiodiniaceae (formerly clade A) consists of two monophyletic groups, one
 46 of which has been revised as *Symbiodinium sensu stricto*^{2,16}. *Symbiodinium* (as revised) includes a
 47 wide range of mutualistic, opportunistic and free-living forms. *Symbiodinium tridacnidorum*, for
 48 instance, encompasses isolates in ITS2-type A3 that are predominantly symbionts of giant clams in
 49 the Indo-Pacific Ocean². Although the nature of this symbiosis is extracellular, they can also
 50 establish intracellular symbiosis with cnidarian hosts both in experimental settings and in nature¹⁷.
 51 On the other hand, *Symbiodinium natans* (the type species of the genus) is free-living. *S. natans*
 52 occurs frequently in environmental samples, exhibits a widespread distribution and, thus far, has not
 53 been shown to colonise cnidarian hosts^{2,18}.

54 Symbiosis, or the lack thereof, has been predicted to impact genome evolution of
 55 Symbiodiniaceae¹². Most symbiotic Symbiodiniaceae are thought to be facultative to some extent,

56 with the potential to shift between a free-living motile stage (*i.e.* mastigote form) and a spherical
57 symbiotic stage (*i.e.* coccoid form). The genomes of facultative and recent intracellular symbionts
58 and parasites are usually very unstable, with extensive structural rearrangements, intensified activity
59 of transposable elements (TEs) and exacerbated gene duplication that leads to the accumulation of
60 pseudogenes^{19,20}. Symbiotic Symbiodiniaceae are thus expected to display similar genomic features.

61 In this study, we present draft *de novo* genome assemblies of *S. tridacnidorum* CCMP2592 and *S.*
62 *natans* CCMP2548. Using a comparative genomic approach, we found extensive genome-sequence
63 divergence and few shared families of predicted genes between the two species. A greater extent of
64 gene duplication, and the higher abundance of TEs and pseudogenes in *S. tridacnidorum* relative to
65 *S. natans* suggest that duplication and transposition underpin genome divergence between these
66 species.

67 Results

68 Genome sequences and predicted genes of *S. tridacnidorum* and *S. natans*

69 The genome sequences of *S. tridacnidorum* CCMP2592 and *S. natans* CCMP2548 were assembled
 70 *de novo* using both short- and long-read sequence data (**Error! Reference source not found.**,
 71 Supplementary Table 1). The estimated genome size is 1.29 Gbp for *S. tridacnidorum*, and 0.74
 72 Gbp for *S. natans* (Supplementary Table 2); the latter is the smallest reported for any
 73 Symbiodiniaceae genome to date. Using an integrative gene-prediction workflow tailored for
 74 dinoflagellate genomes (see Methods), we predicted 45,474 high-quality gene models in *S.*
 75 *tridacnidorum*, and 35,270 in *S. natans* (**Error! Reference source not found.**). The gene repertoire
 76 for each genome is more complete (85.15% and 83.41% recovery of core conserved eukaryote
 77 genes²¹ in *S. tridacnidorum* and *S. natans*, respectively) than other *Symbiodinium* genomes (<79%
 78 recovery; Supplementary Figure 1).

79 Genomes of *S. tridacnidorum* and *S. natans* are highly divergent

80 The genomes of *S. tridacnidorum* and *S. natans* are highly dissimilar from one another (Fig. 1).
 81 Only 14.70 Mbp (1.33%) of the genome sequence of *S. tridacnidorum* aligned to 11.84 Mbp
 82 (1.55%) of that of *S. natans* at 90% identity or greater. Most aligned genomic regions are short
 83 (<100 bp, Fig. 1a). About half of these regions represent repeats, and another ~40% represent genic
 84 regions that are common to both species (Fig. 1b). We observed a low mapping rate (<15%) of read
 85 pairs from one genome dataset against the genome assembly of the counterpart, and *vice versa* (Fig.
 86 1c). Using all predicted genes, we inferred 58,541 gene families (including 26,649 single-copy
 87 genes), many of which are exclusive to each species (Fig. 1d), *e.g.* 25,700 are specific to *S.*
 88 *tridacnidorum*. However, the predominant gene functions are conserved, as shown by the top ten
 89 most abundant protein domains encoded in the genes from both species (Fig. 1e). The composition
 90 of repetitive elements differs between the two genomes. Simple repeats and long interspersed
 91 nuclear elements (LINEs), for instance, are in smaller proportion in the genome of *S. tridacnidorum*

than they are in that of *S. natans* (Fig. 1F). Conversely, long terminal repeats (LTRs) and DNA transposons are more prominent in *S. tridacnidorum*.

Duplication events and transposable elements contribute to the divergence between *S. tridacnidorum* and *S. natans* genomes

We further assessed the distinct genome features in each species that may have contributed to the discrepancy in genome sizes. Specifically, we assessed, for each feature, the ratio (Δ) of the total length of the implicated sequence regions in *S. tridacnidorum* to the equivalent length in *S. natans* (Fig. 2). The genome size estimate for *S. tridacnidorum* is 1.74 times larger than that for *S. natans* (Supplementary Table 2); we use this ratio as a reference for comparison. Most of the examined genome features span a larger region in the genome of *S. tridacnidorum*, as expected. The Δ for each inspected genic feature (even for exons and introns separately), approximates 1.74. However, six features related to duplicated genes and repetitive elements have $\Delta > 1.74$. This observation suggests that gene duplication and repeats likely expanded in *S. tridacnidorum* (and/or contracted in *S. natans*), contributing to the genome-size discrepancy.

Tandem duplication of exons and genes is common in dinoflagellates, and may serve as an adaptive mechanism to enhance functions relevant for their biology^{22,23}. Whereas in some dinoflagellates genes in tandem arrays can have hundreds of copies, *e.g.* up to 5000 copies of the peridinin-chlorophyll a-binding protein (PCP) gene in *Lingulodinium polyedra*²⁴, these arrays are not as prominent in the genomes of *S. tridacnidorum* and *S. natans* (Supplementary Figure 2), with the largest array comprising 10 and 13 gene copies, respectively. The 13-gene array in *S. natans* encodes a full-length alpha amylase, whereas the remaining 12 copies are fragments of this gene and likely not functional. On the other hand, the 10-gene block in *S. tridacnidorum* contains genes encoding PCP; of these, seven contain duplets of PCP domains, lending support to the previous finding of the origin of a PCP form by duplication in Symbiodiniaceae²⁵; the remaining three copies contain 1, 6 and 14 PCP domains respectively. An additional gene, not part of the tandem array,

contains another PCP-duplet. The total 37 individual PCP domains (35 in a gene cluster and two in a separate duplet) supports the earlier size estimation (36 ± 12) of the PCP family in a genome of Symbiodiniaceae²⁶. In stark contrast, we only recovered a duplet of PCP domains among all predicted proteins of *S. natans*.

The length of duplicated gene blocks is drastically longer in *S. tridacnidorum* than in *S. natans* ($\Delta = 6.32$; Fig. 2). This observation, and the number of gene-block duplicates in each of the two species, suggests that segmental duplication has occurred more frequently during the course of genome evolution of *S. tridacnidorum*. We found 23 syntenic collinear blocks within the *S. tridacnidorum* genome (*i.e.* within-genome duplicated gene blocks) implicating 242 genes in total. Of these genes, 20 encode protein kinase functions (Supplementary Table 3) that are associated with distinct signalling pathways. In comparison, only five syntenic collinear blocks implicating 62 genes were found in the *S. natans* genome; these genes largely encode functions of cation transmembrane transport, relevant for the maintenance of pH homeostasis. Ankyrin and pentatricopeptide repeats are common in the predicted protein products of duplicated genes in both genomes.

Retroposition is another gene-duplication mechanism known to impact genome evolution of Symbiodiniaceae and other dinoflagellates^{22,27}. To survey retroposition in genomes of *S. tridacnidorum* and *S. natans*, we searched for relicts of the dinoflagellate spliced-leader (DinoSL) sequence in upstream regions of all predicted genes. Since the DinoSL is attached to transcribed genes by trans-splicing²⁸, genes containing these relicts represent the primary evidence of retroposition into the genome. We found 412 and 252 genes with conserved DinoSL relicts in *S. tridacnidorum* and *S. natans*, respectively. Genes with higher expression levels have been assumed to be more prone to be retroposed into the genome²⁹. The identified retroposed genes in the two species encode distinct functions based on the annotated Gene Ontology (GO) terms (Fig. 3a). This observation may be attributed to the preferential expression of functions that are (or were) relevant

141 to each species. For instance, peptide antigen binding (GO:0042605) might be important for host
 142 recognition in *S. tridacnidorum*³⁰.

143 Both retroposition and retrotransposition have been reported to contribute to gene-family expansion
 144 in Symbiodiniaceae³¹. Protein domains with functions related to retrotransposition were
 145 overrepresented in gene products of *S. tridacnidorum* relative to those of *S. natans* (Supplementary
 146 Table 4). However, the reverse transcriptase domains (PF00078 and PF07727) are abundant in both;
 147 they were found in 1313 predicted proteins in *S. tridacnidorum* and 591 in *S. natans*.

148 Retrotransposons can accelerate mutation rate³² and alter the architecture of genes in their flanking
 149 regions³³, and may explain the emergence of genes coding for reverse transcriptase domains (RT-
 150 genes) in these genomes. Other domains found in these proteins are involved in diverse cellular
 151 processes including ubiquitin-mediated proteolysis, DNA methylation, transmembrane transport
 152 and photosynthesis (Fig. 3b, Supplementary Table 5). The lack of overlap between functions
 153 enriched in genes containing DinoSL relicts and those in RT-genes indicates that retroposition and
 154 retrotransposition are independent processes. The abundance of repeats characteristic of TEs (such
 155 as LINEs and LTRs; Fig. 2) further supports the enhanced activity of retrotransposition in *S.*
 156 *tridacnidorum*. Although LINEs display high sequence divergence (Kimura distance³⁴ 20-30),
 157 potentially a remnant from an ancient burst of this type of element common to all Suessiales^{3,22},
 158 most LTRs and DNA transposons are largely conserved (Kimura distance < 5), suggesting that they
 159 may be active (Fig. 4). We note that these conserved LTRs and DNA transposons were recovered
 160 only in our hybrid genome assemblies incorporating both short- and long-read sequence data, and
 161 not in our preliminary genome assemblies based solely on short-read data (Supplementary Figure 3,
 162 Supplementary Table 6). This indicates that these conserved, repetitive regions can be resolved only
 163 using long-read sequence data (Supplementary Figure 4), highlighting the importance of long-read
 164 data in generating and assembling dinoflagellate genomes.

165 **High divergence among gene copies counteracts gene-family expansion in *S. tridacnidorum***

166 Duplicated genes can experience distinct fates^{35,36}. These fates can result in different scenarios
 167 depending on the divergence accumulated in the sequences. First, if the function remains the same
 168 or changes slightly (*e.g.* through subfunctionalisation), the duplicated gene sequences will remain
 169 similar, resulting in gene-family expansion. We assessed the difference in gene-family sizes
 170 between *S. tridacnidorum* and *S. natans* using Fisher's exact test (see Methods), and consider those
 171 with an adjusted $p \leq 0.05$ as significantly different (Fig. 5). Although events contributing to the
 172 increase of gene-copy numbers appear more prevalent in *S. tridacnidorum*, gene families are not
 173 drastically larger than those in *S. natans*; only 20 families are significantly larger. Of these 20
 174 families, one (OG0000004) putatively encodes protein kinases and glycosyltransferases that are
 175 necessary for the biosynthesis of glycoproteins, and another (OG0000013) encodes ankyrin and
 176 transport proteins (Supplementary Table 7). These functions are important for the recognition of
 177 and interaction with the host among symbiodiniacean symbionts³⁷⁻³⁹. In comparison, five gene
 178 families were significantly larger in *S. natans* than in *S. tridacnidorum*, of which one (OG0000003)
 179 encodes for a sodium-transporter and another (OG0000034) for a transmembrane protein. Many
 180 genes in the expanded families encode for retrotransposition functions in both genomes, lending
 181 support to the contributing role of retrotransposons in gene-family expansion in Symbiodiniaceae³¹.
 182 Although the functions of many other genes in these families could not be determined due to the
 183 lack of known similar sequences, they might be relevant for adaptation to specific ecological niches
 184 as previously proposed for dinoflagellates⁴⁰.

185 Second, if novel beneficial functions of the gene copies emerge (*i.e.* neofunctionalisation), the
 186 sequence divergence between gene copies may become too large to be recognised as the same
 187 family. This scenario could, at least partially, explain the higher number of single-copy genes
 188 exclusive to *S. tridacnidorum* (25,649) than those exclusive to *S. natans* (16,137). Whereas 13,320
 189 (82.54%) of the 16,137 single-copy genes of *S. natans* are supported by transcriptome evidence,
 190 only 13,189 (51.42%) of those 25,649 in *S. tridacnidorum* are. It remains unclear if these latter

191 represent functional genes. Moreover, the annotated functions of these single-copy genes exclusive
192 to each genome are similar in both species (Supplementary Table 8), suggesting the presence of
193 highly diverged homologs.

194 Finally, duplicated genes can undergo loss of function (*i.e.* nonfunctionalisation or
195 pseudogenisation). Pseudogene screening in both genomes (see Methods) identified 183,516
196 putative pseudogenes in *S. tridacnidorum* and 48,427 in *S. natans*. The nearly four-fold difference
197 in the number of pseudogenes between the two genomes further supports the notion that more-
198 frequent duplication events occur in *S. tridacnidorum*, and may explain the lower proportion of
199 genes with transcript support in this species (**Error! Reference source not found.**).

200 Our results suggest that the high sequence divergence of duplicated genes, potentially due to the
201 accumulation of mutations as a consequence of pseudogenisation, perhaps together with
202 neofunctionalisation, may hinder gene family expansion in the genome of *S. tridacnidorum*.

203 **Gene functions of *S. tridacnidorum* and *S. natans* are relevant to their lifestyle**

204 According to our analysis of enriched gene functions in *S. tridacnidorum* relative to *S. natans* based
205 on annotated GO terms, methylation and the biosynthesis of histidine and peptidoglycan were
206 among the most significant (Supplementary Table 9). The enrichment of methylation is not
207 surprising because retrotransposons of Symbiodiniaceae are known to have acquired
208 methyltransferase domains, likely contributing to the hypermethylated nuclear genomes of these
209 dinoflagellates⁴¹. The link between the extent of methylation in symbiodiniacean genomes and its
210 representation among predicted genes can be further assessed using methylation sequencing.

211 Although some corals can synthesise histidine *de novo*, metazoans generally lack this capacity⁴².
212 The enrichment of histidine biosynthesis in *S. tridacnidorum* may be a result of host-symbiont
213 coevolution or, alternatively, may explain why this species is a preferred symbiont over others (*e.g.*
214 *S. natans*). Biosynthesis of peptidoglycans is also important for symbiosis, because these molecules,

215 on the cell surface of Symbiodiniaceae, interact with host lectins as part of the symbiont recognition
216 process^{30,39}.

217 On the other hand, *S. natans* displays a wider range of enriched functions related to cellular
218 processes (Supplementary Table 9), as expected for free-living Symbiodiniaceae¹². One of the most
219 significantly overrepresented gene functions is the transmembrane transport of sodium. Whereas
220 this function is likely related to pH (osmotic) homeostasis with the extracellular environment, the
221 occurrence of a sodium:phosphate symporter (PF02690) in tandem, exclusive to *S. natans*, and the
222 abundance of a sodium:chloride symporter (PF00209) among the RT-genes (Supplementary Table
223 5) suggest that *S. natans* makes use of the Na⁺ differential gradient (caused by the higher Na⁺
224 concentration in seawater) for nutrient uptake in a similar fashion to the assimilation of inorganic
225 phosphate by the malaria parasite (*Plasmodium falciparum*) in the Na⁺-rich cytosol of the host's
226 erythrocytes⁴³.

227 **Are features underpinning genome divergence in Symbiodiniaceae ancestral or derived?**

228 To assess whether the genome features found in *S. tridacnidorum* were ancestral or derived relative
229 to *S. natans*, we compared the genome sequences from both species with those from the outgroup
230 *Polarella glacialis* CCMP1383²², a psychrophilic free-living species closely related to
231 Symbiodiniaceae (also in Order Suessiales).

232 A greater genome sequence proportion of *S. natans* (3.38%) than that of *S. tridacnidorum* (0.85%)
233 aligned to the *P. glacialis* genome assembly. Interestingly, the aligned regions in both cases
234 implicate only ~5 Mbp (~0.18%) of the *P. glacialis* genome sequence. This observation is likely
235 due to duplicated genome regions of *S. natans* that have remained highly conserved. Similarly, the
236 average percent identity of the best-matching sequences between any of the two *Symbiodinium*
237 genomes against *P. glacialis* is very similar (*i.e.* 92.13% and 92.56% for *S. tridacnidorum* and *S.*
238 *natans*, respectively). Nonetheless, regions occupied by duplicated genes are recovered in larger
239 proportions in *Symbiodinium* than in *P. glacialis* (Fig. 6). On the other hand, LTR retrotransposons

are evidently more prominent in *P. glacialis*. However, these LTRs are more diverged (Kimura distances 3-8)²² than those in the two *Symbiodinium* (Kimura distances < 5; Fig. 4), indicating an independent, more-ancient burst of these elements in *P. glacialis*.

Discussion

We report for the first time, based on whole-genome sequence data, evidence of structural rearrangements and TEs contributing to the extensive genomic divergence between the symbiotic *S. tridacnidorum* and the free-living *S. natans*, including the discrepancy in genome sizes. In comparison, structural rearrangements and TE activity are less prominent in the genomes of *S. natans* and the outgroup species *P. glacialis*.

Structural rearrangements, abundance of pseudogenes, and enhanced activity of TEs are common in facultative and recent intracellular symbionts and parasites^{19,20}, and are expected in symbiotic Symbiodiniaceae¹². Our results support this hypothesis. In this regard, our results agree with the notion that the symbiotic lifestyle is a derived trait in *Symbiodinium*, and that the free-living lifestyle is likely ancestral. Under this assumption, the genome proportion spanned by TEs and duplicated genes in *S. natans* is expected to be similar (if not smaller) than that in the outgroup *P. glacialis*. However, we found the proportion of duplicated genes to be larger in *S. natans* (Fig. 6), prompting two possible explanations. First, the pervasive simple repeats in the *P. glacialis* genome²², independently expanded along this lineage or possibly an ancestral trait in Suessiales, drastically diminishes the proportion of genic regions in the genome. Second, the free-living lifestyle of *S. natans* may be a derived trait in *Symbiodinium*, having passed through a symbiotic phase earlier in its evolutionary history. However, the robust placement of *S. natans* in the basal position alongside *Symbiodinium pilosum* (another free-living species) in the *Symbiodinium* phylogeny² contradicts this less-parsimonious explanation. Additional high-quality genome data from free-living and symbiotic taxa are thus required to gain a clearer understanding of the evolutionary transition(s) between free-living and symbiotic lifestyles in Symbiodiniaceae.

265 **Methods**

266 ***Symbiodinium* cultures**

267 Single-cell monoclonal cultures of two *Symbiodinium* (formerly Clade A) species were obtained
 268 from the Bigelow National Center for Marine Algae and Microbiota. *Symbiodinium natans* (strain
 269 CCMP2548) was originally collected from open ocean water in Hawaii, USA. *Symbiodinium*
 270 *tridacnidorum* (Clade A3, strain CCMP2592) was originally recovered from a stony coral
 271 (*Heliofungia actiniformis*) on the Great Barrier Reef, Australia. The cultures were maintained in
 272 multiple 100-mL batches (in 250-mL Erlenmeyer flasks) in f/2 (without silica) medium (0.2 mm
 273 filter-sterilized) under a 14:10 h light-dark cycle (90 $\mu\text{E}/\text{m}^2/\text{s}$) at 25 °C. The medium was
 274 supplemented with antibiotics (ampicillin [10 mg/mL], kanamycin [5 mg/mL] and streptomycin [10
 275 mg/mL]) to reduce bacterial growth.

276 **Nucleic acid extraction**

277 Genomic DNA was extracted following the 2×CTAB protocol with modifications. *Symbiodinium*
 278 cells were first harvested during exponential growth phase (before reaching 10^6 cells/mL) by
 279 centrifugation (3000 g, 15 min, room temperature (RT)). Upon removal of residual medium, the
 280 cells were snap-frozen in liquid nitrogen prior to DNA extraction, or stored at -80 °C. For DNA
 281 extraction, the cells were suspended in a lysis extraction buffer (400 μL ; 100 mM Tris-Cl pH 8, 20
 282 mM EDTA pH 8, 1.4 M NaCl), before silica beads were added. In a freeze-thaw cycle, the mixture
 283 was vortexed at high speed (2 min), and immediately snap-frozen in liquid nitrogen; the cycle was
 284 repeated 5 times. The final volume of the mixture was made up to 2% w/v CTAB (from 10% w/v
 285 CTAB stock; kept at 37 °C). The mixture was treated with RNase A (Invitrogen; final
 286 concentration 20 $\mu\text{g}/\text{mL}$) at 37 °C (30 min), and Proteinase K (final concentration 120 $\mu\text{g}/\text{mL}$) at 65
 287 °C (2 h). The lysate was then subjected to standard extractions using equal volumes of
 288 phenol:chloroform:isoamyl alcohol (25:24:1 v/v; centrifugation at 14,000 g, 5 min, RT), and
 289 chloroform:isoamyl alcohol (24:1 v/w; centrifugation at 14,000 g, 5 min, RT). DNA was

precipitated using pre-chilled isopropanol (gentle inversions of the tube, centrifugation at 18,000 g, 15 min, 4 °C). The resulting pellet was washed with pre-chilled ethanol (70% v/v), before stored in Tris-HCl (100 mM, pH 8) buffer. DNA concentration was determined with NanoDrop (Thermo Scientific), and DNA with $A_{230:260:280} \approx 1.0:2.0:1.0$ was considered appropriate for sequencing. Total RNA was isolated using the RNeasy Plant Mini Kit (Qiagen) following directions of the manufacturer. RNA quality and concentration were determined with an Agilent 2100 BioAnalyzer.

Genome sequence data generation and *de novo* assembly

In total, we generated 1021.63 Gbp (6.77 billion reads) of genome sequence data for *S. natans* and 259.57 Gbp (1.48 billion reads) for *S. tridacnidorum* (Supplementary Table 1). Short-read sequence data (2×150 bp reads) were generated using multiple paired-end (for both species) and mate-pair (for *S. natans* only) libraries on the Illumina HiSeq 2500 and 4000 platforms at the Australian Genome Research Facility (Melbourne) and the Translational Research Institute Australia (Brisbane). One of the paired-end libraries for *S. natans* (of insert length 250 bp) was designed such that the read-pairs of 2×150 bp would overlap. Genome size and sequence read coverage were estimated based on *k*-mer frequency analysis (Supplementary Table 2) as counted with Jellyfish v2.2.6, using only paired-end data.

Quality assessment of the raw paired-end data was done with FastQC v0.11.5, and subsequent processing with Trimmomatic v0.36⁴⁴. To ensure high-quality read data for downstream analyses, the paired-end mode of Trimmomatic was run with the settings:

ILLUMINACLIP:[AdapterFile]:2:30:10 LEADING:30 TRAILING:30 SLIDINGWINDOW:4:25 MINLEN:100 AVGQUAL:30; CROP and HEADCROP were run (prior to LEADING and TRAILING) when required to remove read ends with nucleotide biases. Overlapping read pairs from the library with insert size of 250 bp were merged with FLASH v1.2.11⁴⁵. Library adapters from the mate-pair data were removed with NxTrim v0.41⁴⁶. A preliminary *de novo* genome assembly per species was done for genome-guided transcriptome assembly (see below) with CLC

315 Genomics Workbench v7.5.1 (qiagenbioinformatics.com) using default parameters and the merged
316 pairs (for *S. natans*), the unmerged read pairs and the trim-surviving unpaired reads. The
317 preliminary assembly of *S. natans* was further scaffolded with SSPACE v3.0⁴⁷ and the mate-pair
318 filtered data.

319 Additionally, long-read sequence data were generated on a PacBio Sequel system at the Ramaciotti
320 Centre for Genomics (Sydney). These data and the paired-end libraries (adding up to a coverage of
321 152-fold for *S. natans* and 200-fold for *S. tridacnidorum*) were used for hybrid *de novo* genome
322 assembly (Supplementary Table 1) with MaSuRCA 3.3.0⁴⁸, following the procedure described in
323 the manual. Except for the PacBio sub-reads, filtered to a minimum length of 5 kbp, all sequence
324 data were input without being pre-processed, as recommended by the developer. The genome
325 assemblies were further scaffolded with transcriptome data generated in this study (see below)
326 using L_RNA_scaffolder⁴⁹.

327 **Removal of putative microbial contaminants**

328 To identify putative sequences from bacteria, archaea and viruses in the genome scaffolds we
329 followed the approach of Liu *et al.*³. In brief, we first searched the scaffolds (BLASTn) against a
330 database of bacterial, archaeal and viral genomes from RefSeq (release 88); hits with $E \leq 10^{-20}$ and
331 alignment bit score ≥ 1000 were considered as significant. We then calculated the proportion of
332 bases in each scaffold covered by significant hits. Next, we assessed the added length of implicated
333 genome scaffolds across different thresholds of these proportions, and the corresponding gene
334 models in these scaffolds as predicted from available transcripts using PASA v2.3.3⁵⁰ (see below),
335 with a modified script available at github.com/chancx/dinoflag-alt-splice) that recognises an
336 additional donor splice site (GA), and TransDecoder v5.2.0⁵⁰. This preliminary gene prediction was
337 done on the repeat-masked genome using clean transcripts, as described below. The most-stringent
338 sequence coverage ($\geq 5\%$) was selected as the threshold for all samples, *i.e.* any scaffold with

339 significant bacterial, archaeal or viral hits covering $\geq 5\%$ of its length was considered as
340 contaminant and removed from the assembly (Supplementary Figure 5).

341 **RNA sequence data generation and transcriptome assembly**

342 We generated transcriptome sequence data for both *S. tridacnidorum* and *S. natans* (Supplementary
343 Table 10). Short-read sequence data (2×150 bp reads) were generated using paired-end libraries on
344 the Illumina NovaSeq 6000 platform at the Australian Genome Research Facility (Melbourne).
345 Quality assessment of the raw paired-end data was done with FastQC v0.11.4, and subsequent
346 processing with Trimmomatic v0.35⁴⁴. To ensure high-quality read data for downstream analyses,
347 the paired-end mode of Trimmomatic was run with the settings: HEADCROP:10
348 ILLUMINACLIP:[AdapterFile]:2:30:10 CROP:125 SLIDINGWINDOW:4:13 MINLEN:50. The
349 surviving read pairs were further trimmed with QUADTrim v2.0.2
350 (bitbucket.org/arobinson/quadtrim) with the flags *-m 2* and *-g* to remove homopolymeric guanine
351 repeats at the end of the reads (a systematic error of Illumina NovaSeq 6000).

352 Transcriptome assembly was done with Trinity v2.1.1⁵¹ in two modes: *de novo* and genome-guided.
353 *De novo* transcriptome assembly was done using default parameters and the trimmed read pairs. For
354 genome-guided assembly, high-quality read pairs were aligned to the preliminary *de novo* genome
355 assembly using Bowtie v2.2.7⁵². Transcriptomes were then assembled with Trinity in the genome-
356 guided mode using the alignment information, and setting the maximum intron size to 100,000 bp.
357 Both *de novo* and genome-guided transcriptome assemblies from each sample were used for
358 scaffolding (see above) and gene prediction (see below).

359 **Full-length transcript evidence for gene prediction**

360 Full-length transcripts for *S. tridacnidorum* and *S. natans* were generated using the PacBio IsoSeq
361 technology. All sequencing was conducted using the PacBio Sequel platform at the Institute for
362 Molecular Bioscience (IMB) Sequencing Facility, The University of Queensland (Brisbane,
363 Australia; Supplementary Table 10). Full-length cDNA was first synthesised and amplified using

the TeloPrime Full-Length cDNA Amplification Kit (Lexogen) and TeloPrime PCR Add-on Kit (Lexogen) following the protocols provided in the product manuals. One synthesis reaction was performed for each sample using 821 ng from *S. tridacnidorum* and 1.09 µg from *S. natans* of total RNA as starting material. Next, 25 (*S. tridacnidorum*) and 23 (*S. natans*) PCR cycles were carried out for cDNA amplification. PCR products were divided into two fractions, which were purified using 0.5× (for *S. tridacnidorum*) and 1× (for *S. natans*) AMPure PB beads (Pacific Biosciences), and then pooled with equimolar quantities. The recovered 699 ng (*S. tridacnidorum*) and 761 ng (*S. natans*) of cDNA were used for sequencing library preparation with the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences). The cDNA from these libraries were sequenced in two SMRT cells.

To generate the dinoflagellate spliced-leader (DinoSL) specific transcript library, 12 PCR cycles were carried out for both samples using the conserved DinoSL fragment (5'-CCGTAGCCATTTTGGCTCAAG-3') as forward primer, the TeloPrime PCR 3'-primer as reverse primer, and the fraction of full-length cDNA purified with 0.5× (for *S. tridacnidorum*) and 1× (for *S. natans*) AMPure PB beads. The above-described PCR purification and sequencing library preparation methods were used for the DinoSL transcript libraries; cDNA from these libraries was sequenced in one SMRT cell per sample.

Due to the abundance of undesired 5'-5' and 3'-3' pairs, and to recover as much transcript evidence as possible for gene prediction, we followed two approaches (Supplementary Figure 6). First, the IsoSeq 3.1 workflow (github.com/PacificBiosciences/IsoSeq3/blob/master/README_v3.1.md) was followed. Briefly, circular consensus sequences (CCS) were generated from the subreads of each SMRT cell with ccs v3.1.0 without polishing, and setting the minimum number of subreads to generate CCS (--minPasses) to 1. Removal of primers was done with lima v1.8.0 in the IsoSeq mode, with a subsequent refinement step using isoseq v3.1.0. At this stage, the refined full-length transcripts of all SMRT cells (excluding those from the DinoSL library) were combined to be then

388 clustered by similarity and polished with isoseq v3.1.0. High- and low- quality transcripts resulting
389 from this approach were further used for gene prediction (see below).

390 For the second approach, we repeated the IsoSeq workflow with some modifications. We polished
391 the subreads with the Arrow algorithm and used at least three subreads per CCS with ccs v3.1.0 to
392 generate high-accuracy CCS. Primer removal and refinement were done as explained above. The
393 subsequent clustering and polishing steps were skipped. The resulting polished CCS and full-length
394 transcripts were also used for gene prediction. IsoSeq data from the DinoSL library were processed
395 separately following the same two approaches.

396 **Genome annotation and gene prediction**

397 We adopted the same comprehensive *ab initio* gene prediction approach reported in Chen *et al.*⁵³,
398 using available genes and transcriptomes of Symbiodiniaceae as guiding evidence. A *de novo* repeat
399 library was first derived for the genome assembly using RepeatModeler v1.0.11
400 (repeatmasker.org/RepeatModeler). All repeats (including known repeats in RepeatMasker database
401 release 20180625) were masked using RepeatMasker v4.0.7 (repeatmasker.org).

402 As direct transcript evidence, we used the *de novo* and genome-guided transcriptome assemblies
403 from Illumina short-read sequence data, as well as the PacBio IsoSeq full-length transcript data (see
404 above). We concatenated all the transcript datasets per sample and “cleaned” them with SeqClean
405 (sourceforge.net/projects/seqclean) and the UniVec database build 10.0. We used PASA v2.3.3⁵⁰,
406 customised to recognise dinoflagellate alternative splice donor sites (see above), and TransDecoder
407 v5.2.0⁵⁰ to predict coding sequences (CDS). These CDS were searched (BLASTp, $E \leq 10^{-20}$)
408 against a protein database that consists of RefSeq proteins (release 88) and a collection of available
409 and predicted (with TransDecoder v5.2.0⁵⁰) proteins of Symbiodiniaceae (total of 111,591,828
410 sequences; Supplementary Table 11). We used the *analyze_blastPlus_topHit_coverage.pl* script
411 from Trinity v2.6.6⁵¹ to retrieve only those CDS having a hit with >70% coverage of the database
412 protein sequence (*i.e.* nearly full-length) in the database for subsequent analyses.

413 The near full-length gene models were checked for TEs using HHblits v2.0.16 (probability = 80%
 414 and E -value = 10^{-5}), searching against the JAMg transposon database
 415 (sourceforge.net/projects/jamg/files/databases), and TransposonPSI (transposonpsi.sourceforge.net).
 416 Gene models containing TEs were removed from the gene set, and redundancy reduction was
 417 conducted using cd-hit v4.6^{54,55} (ID = 75%). The remaining gene models were processed using the
 418 *prepare_golden_genes_for_predictors.pl* script from the JAMg pipeline (altered to recognise GA
 419 donor splice sites; jamg.sourceforge.net). This script produces a set of “golden genes” that was used
 420 as training set for the *ab initio* gene-prediction tools AUGUSTUS v3.3.1⁵⁶ (customised to recognise
 421 the non-canonical splice sites of dinoflagellates, following the changes made to that available at
 422 smic.reefgenomics.org/download) and SNAP v2006-07-28⁵⁷. Independently, the soft-masked
 423 genome sequences were passed to GeneMark-ES v4.32⁵⁸ for unsupervised training and gene
 424 prediction. UniProt-SwissProt proteins (downloaded on 27 June 2018) and the predicted proteins of
 425 Symbiodiniaceae (Supplementary Table 11) were used to produce a set of gene predictions using
 426 MAKER v2.31.10⁵⁹ protein2genome; the custom repeat library was used by RepeatMasker as part
 427 of MAKER prediction. A primary set of predicted genes was produced using EvidenceModeler
 428 v1.1.1⁶⁰, modified to recognise GA donor splice sites. This package combined the gene predictions
 429 from PASA, SNAP, AUGUSTUS, GeneMark-ES and MAKER protein2genome into a single set of
 430 evidence-based predictions. The weightings used for the package were: PASA 10, Maker protein 8,
 431 AUGUSTUS 6, SNAP 2 and GeneMark-ES 2. Only gene models with transcript evidence (*i.e.*
 432 predicted by PASA) or supported by at least two *ab initio* prediction programs were kept. We
 433 assessed completeness by querying the predicted protein sequences in a BLASTp similarity search
 434 ($E \leq 10^{-5}$, $\geq 50\%$ query/target sequence cover) against the 458 core eukaryotic genes from
 435 CEGMA²¹. Transcript data support for the predicted genes was determined by BLASTn ($E \leq 10^{-5}$)
 436 similarity search, querying the transcript sequences against the predicted CDS from each genome.
 437 Genes for which the transcripts aligned to their CDS with at least 50% of sequence cover and 90%
 438 identity were considered as supported by transcript data.

439 **Gene-function annotation and enrichment analyses**

440 Annotation of the predicted genes was done based on sequence similarity searches against know
 441 proteins following the same approach as Liu *et al.*³, in which the predicted protein sequences were
 442 used as query (BLASTp, $E \leq 10^{-5}$, minimum query or target cover of 50%) against Swiss-Prot first,
 443 and those with no Swiss-Prot hits subsequently against TrEMBL (both databases from UniProt,
 444 downloaded on 27 June 2018). The best UniProt hit with associated Gene Ontology (GO,
 445 geneontology.org) terms was used to annotate the query protein with those GO terms using the
 446 UniProt-GOA mapping (downloaded on 03/06/2019). Pfam domains⁶¹ were searched in the
 447 predicted proteins of both *Symbiodinium* species using PfamScan⁶² ($E \leq 0.001$) and the Pfam-A
 448 database (release 30 August 2018)⁶¹.

449 Tests for enrichment of Pfam domains were done with one-tailed Fisher's exact tests, independently
 450 for over- and under-represented features; domains with Benjamini-Hochberg⁶³ adjusted $p \leq 0.05$
 451 were considered significant. Enrichment of GO terms was performed using the topGO
 452 Bioconductor package⁶⁴ implemented in R v3.5.1, applying Fisher's Exact test with the
 453 'elimination' method to correct for the dependence structure among GO terms. GO terms with a $p \leq$
 454 0.01 were considered significant.

455 **Comparative genomic analyses**

456 Whole-genome sequence alignment was carried out with nucmer v4.0.0⁶⁵ with the hybrid genome
 457 assembly of *S. natans* as reference and that of *S. tridacnidorum* as query, and using anchor matches
 458 that are unique in the sequences from both species (--mum). Sequences from both *Symbiodinium*
 459 genomes were queried in the same way against the genome sequence of *P. glacialis* CCMP1383²².
 460 Filtered read pairs (see above, Supplementary Table 1) from both species were aligned to their
 461 corresponding and counterpart genome sequences using bwa v0.7.13⁶⁶, and rates of mapping with
 462 different quality scores were calculated with SAMStat v1.5.1⁶⁷.

463 Groups of homologous sequences from the two *Symbiodinium* genomes were inferred with
 464 Orthofinder v2.3.1⁶⁸, and considered gene families. The significance of size differences of the gene
 465 families shared by *S. tridacnidorum* and *S. natans* was assessed with a two-tailed Fisher's exact test
 466 correcting p-values for multiple testing with the Benjamini-Hochberg method⁶³; difference in size
 467 was considered significant for gene families with adjusted $p \leq 0.05$.

468 We used the predicted genes and their associated genomic positions to identify potential segmental
 469 genome duplications in both *Symbiodinium* species, as well as in *P. glacialis*. First, we used
 470 BLASTp ($E \leq 10^{-5}$) to search for similar proteins within each genome; the hit pairs were filtered to
 471 include only those where the alignment covered at least half of either the query or the matched
 472 protein sequence. Next, we ran MCScanX⁶⁹ in intra-specific mode (*-b I*) to identify collinear
 473 syntenic blocks of at least five genes and genes arranged in tandem within each genome separately.

474 Identification of genes with DinoSL and pseudogenes was done in a similar way to Song *et al.*
 475 (2017)²⁷. We queried the original DinoSL sequence (DCCGUAGCCAUUUUGGCUCAAG)²⁸,
 476 excluding the first ambiguous position, against the upstream regions (up to 500 bp) of all genes in a
 477 BLASTn search, keeping the default values of all alignment parameters but with word size set to 9
 478 (*-word_size 9*). Pseudogene detection was done with tBLASTn, with the predicted protein for each
 479 genome as query against the genome sequence, with the regions covered by the predicted genes
 480 masked, as target. Matched regions with $\geq 75\%$ identity were considered part of pseudogenes and
 481 surrounding matching fragments were considered as part of the same pseudogene as long as they
 482 were at a maximum distance of 1 kbp from another pseudogene fragment and in the same
 483 orientation.

484 References

- 485 1 Rowan, R. & Powers, D. A. Ribosomal RNA sequences and the diversity of symbiotic
 486 dinoflagellates (zooxanthellae). *Proc. Natl. Acad. Sci. U. S. A.* **89**, 3639-3643 (1992).
- 487 2 LaJeunesse, T. C. *et al.* Systematic revision of Symbiodiniaceae highlights the antiquity and
 488 diversity of coral endosymbionts. *Curr. Biol.* **28**, 2570-2580, doi:10.1016/j.cub.2018.07.008
 489 (2018).

- 490 3 Liu, H. *et al.* *Symbiodinium* genomes reveal adaptive evolution of functions related to coral-
491 dinoflagellate symbiosis. *Commun. Biol.* **1**, 95, doi:10.1038/s42003-018-0098-3 (2018).
- 492 4 Chi, J., Parrow, M. W. & Dunthorn, M. Cryptic sex in *Symbiodinium* (Alveolata,
493 Dinoflagellata) is supported by an inventory of meiotic genes. *J. Eukaryot. Microbiol.* **61**,
494 322-327, doi:doi:10.1111/jeu.12110 (2014).
- 495 5 Morse, D. A transcriptome-based perspective of meiosis in dinoflagellates. *Protist*,
496 doi:10.1016/j.protis.2019.06.003 (2019).
- 497 6 Baillie, B., Monje, V., Silvestre, V., Sison, M. & Belda-Baillie, C. Allozyme electrophoresis
498 as a tool for distinguishing different zooxanthellae symbiotic with giant clams. *Proc. R. Soc.*
499 *Lond. B Biol. Sci.* **265**, 1949-1956 (1998).
- 500 7 Baillie, B. *et al.* Genetic variation in *Symbiodinium* isolates from giant clams based on
501 random-amplified-polymorphic DNA (RAPD) patterns. *Mar. Biol.* **136**, 829-836 (2000).
- 502 8 LaJeunesse, T. Diversity and community structure of symbiotic dinoflagellates from
503 Caribbean coral reefs. *Mar. Biol.* **141**, 387-400 (2002).
- 504 9 Pettay, D. T. & LaJeunesse, T. C. Long-range dispersal and high-latitude environments
505 influence the population structure of a “stress-tolerant” dinoflagellate endosymbiont. *PLoS*
506 *ONE* **8**, e79208, doi:10.1371/journal.pone.0079208 (2013).
- 507 10 Thornhill, D. J., Lewis, A. M., Wham, D. C. & LaJeunesse, T. C. Host-specialist lineages
508 dominate the adaptive radiation of reef coral endosymbionts. *Evolution* **68**, 352-367,
509 doi:doi:10.1111/evo.12270 (2014).
- 510 11 Baker, A. C. Flexibility and specificity in coral-algal symbiosis: diversity, ecology, and
511 biogeography of *Symbiodinium*. *Annu. Rev. Ecol. Evol. Syst.*, 661-689 (2003).
- 512 12 González-Pech, R. A., Bhattacharya, D., Ragan, M. A. & Chan, C. X. Genome evolution of
513 coral reef symbionts as intracellular residents. *Trends Ecol. Evol.*,
514 doi:10.1016/j.tree.2019.04.010 (2019).
- 515 13 Quigley, K., Bay, L. K. & Willis, B. Temperature and water quality-related patterns in
516 sediment-associated *Symbiodinium* communities impact symbiont uptake and fitness of
517 juveniles in the genus *Acropora*. *Front. Mar. Sci.* **4**, 401 (2017).
- 518 14 LaJeunesse, T. C. Investigating the biodiversity, ecology, and phylogeny of endosymbiotic
519 dinoflagellates in the genus *Symbiodinium* using the ITS region: in search of a “species” level
520 marker. *J. Phycol.* **37**, 866-880 (2002).
- 521 15 Nitschke, M. R., Davy, S. K., Cribb, T. H. & Ward, S. The effect of elevated temperature and
522 substrate on free-living *Symbiodinium* cultures. *Coral Reefs* **34**, 161-171,
523 doi:10.1007/s00338-014-1220-8 (2015).
- 524 16 Pochon, X., Montoya-Burgos, J. I., Stadelmann, B. & Pawlowski, J. Molecular phylogeny,
525 evolutionary rates, and divergence timing of the symbiotic dinoflagellate genus
526 *Symbiodinium*. *Mol. Phylogenet. Evol.* **38**, 20-30 (2006).
- 527 17 Lee, S. Y. *et al.* *Symbiodinium tridacnidorum* sp. nov., a dinoflagellate common to Indo-
528 Pacific giant clams, and a revised morphological description of *Symbiodinium*
529 *microadriaticum* Freudenthal, emended Trench & Blank. *Eur. J. Phycol.* **50**, 155-172,
530 doi:10.1080/09670262.2015.1018336 (2015).
- 531 18 Hansen, G. & Daugbjerg, N. *Symbiodinium natans* sp. nov.: A “free-living” dinoflagellate
532 from Tenerife (Northeast-Atlantic Ocean). *J. Phycol.* **45**, 251-263 (2009).
- 533 19 Moran, N. A. & Plague, G. R. Genomic changes following host restriction in bacteria. *Curr.*
534 *Opin. Genet. Dev.* **14**, 627-633, doi:10.1016/j.gde.2004.09.003 (2004).
- 535 20 McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nat.*
536 *Rev. Microbiol.* **10**, 13-16, doi:10.1038/nrmicro2670 (2011).
- 537 21 Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in
538 eukaryotic genomes. *Bioinformatics* **23**, 1061-1067, doi:10.1093/bioinformatics/btm071
539 (2007).

- 540 22 Stephens, T. G. *et al.* *Polarella glacialis* genomes encode tandem repeats of single-exon genes
541 with functions critical to adaptation of dinoflagellates. *bioRxiv*, 704437, doi:10.1101/704437
542 (2019).
- 543 23 Bachvaroff, T. R. & Place, A. R. From stop to start: tandem gene arrangement, copy number
544 and trans-splicing sites in the dinoflagellate *Amphidinium carterae*. *PLoS ONE* **3**, e2929,
545 doi:10.1371/journal.pone.0002929 (2008).
- 546 24 Le, Q. H., Markovic, P., Hastings, J. W., Jovine, R. V. M. & Morse, D. Structure and
547 organization of the peridinin-chlorophyll a-binding protein gene in *Gonyaulax polyedra*.
548 *Molecular and General Genetics MGG* **255**, 595-604, doi:10.1007/s004380050533 (1997).
- 549 25 Norris, B. J. & Miller, D. J. Nucleotide sequence of a cDNA clone encoding the precursor of
550 the peridinin-chlorophyll a-binding protein from the dinoflagellate *Symbiodinium* sp. *Plant*
551 *Mol. Biol.* **24**, 673-677, doi:10.1007/BF00023563 (1994).
- 552 26 Reichman, J. R., Wilcox, T. P. & Vize, P. D. PCP gene family in *Symbiodinium* from
553 *Hippopus hippopus*: low levels of concerted evolution, isoform diversity, and spectral tuning
554 of chromophores. *Mol. Biol. Evol.* **20**, 2143-2154, doi:10.1093/molbev/msg233 (2003).
- 555 27 Song, B. *et al.* Comparative genomics reveals two major bouts of gene retroposition
556 coinciding with crucial periods of *Symbiodinium* evolution. *Genome Biol. Evol.* **9**, 2037-2047,
557 doi:10.1093/gbe/evx144 (2017).
- 558 28 Zhang, H. *et al.* Spliced leader RNA trans-splicing in dinoflagellates. *Proc. Natl. Acad. Sci.*
559 *U. S. A.* **104**, 4618-4623 (2007).
- 560 29 Slamovits, C. H. & Keeling, P. J. Widespread recycling of processed cDNAs in
561 dinoflagellates. *Curr. Biol.* **18**, R550-R552, doi:10.1016/j.cub.2008.04.054 (2008).
- 562 30 Kirk, N. L. & Weis, V. M. in *The mechanistic benefits of microbial symbionts* (ed Christon
563 J. Hurst) 269-294 (Springer International Publishing, 2016).
- 564 31 Lin, S. *et al.* The *Symbiodinium kawagutii* genome illuminates dinoflagellate gene expression
565 and coral symbiosis. *Science* **350**, 691-694 (2015).
- 566 32 Quadrana, L. *et al.* Transposition favors the generation of large effect mutations that may
567 facilitate rapid adaption. *Nat. Commun.* **10**, 3421, doi:10.1038/s41467-019-11385-5 (2019).
- 568 33 Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution.
569 *Nat. Rev. Genet.* **10**, 691, doi:10.1038/nrg2640 (2009).
- 570 34 Kimura, M. A simple method for estimating evolutionary rates of base substitutions through
571 comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111-120 (1980).
- 572 35 Prince, V. E. & Pickett, F. B. Splitting pairs: the diverging fates of duplicated genes. *Nat. Rev.*
573 *Genet.* **3**, 827-837, doi:10.1038/nrg928 (2002).
- 574 36 Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science*
575 **290**, 1151, doi:10.1126/science.290.5494.1151 (2000).
- 576 37 Mohamed, A. R. *et al.* Transcriptomic insights into the establishment of coral-algal symbioses
577 from the symbiont perspective. *bioRxiv*, 652131, doi:10.1101/652131 (2019).
- 578 38 Davy, S. K., Allemand, D. & Weis, V. M. Cell biology of cnidarian-dinoflagellate symbiosis.
579 *Microbiol. Mol. Biol. Rev.* **76**, 229-261, doi:10.1128/mmbr.05014-11 (2012).
- 580 39 Weis, V. M. Cell biology of coral symbiosis: foundational study can inform solutions to the
581 coral reef crisis. *Integrative and Comparative Biology*, doi:10.1093/icb/icz067 (2019).
- 582 40 Stephens, T. G., Ragan, M. A., Bhattacharya, D. & Chan, C. X. Core genes in diverse
583 dinoflagellate lineages include a wealth of conserved dark genes with unknown functions. *Sci.*
584 *Rep.* **8**, 17175, doi:10.1038/s41598-018-35620-z (2018).
- 585 41 de Mendoza, A. *et al.* Recurrent acquisition of cytosine methyltransferases into eukaryotic
586 retrotransposons. *Nat. Commun.* **9**, 1341 (2018).
- 587 42 Ying, H. *et al.* Comparative genomics reveals the distinct evolutionary trajectories of the
588 robust and complex coral lineages. *Genome Biol.* **19**, 175, doi:10.1186/s13059-018-1552-8
589 (2018).
- 590 43 Saliba, K. J. *et al.* Sodium-dependent uptake of inorganic phosphate by the intracellular
591 malaria parasite. *Nature* **443**, 582-585, doi:10.1038/nature05149 (2006).

592 44 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
593 sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

594 45 Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome
595 assemblies. *Bioinformatics* **27**, 2957-2963, doi:10.1093/bioinformatics/btr507 (2011).

596 46 O'Connell, J. *et al.* NxTrim: optimized trimming of Illumina mate pair reads. *Bioinformatics*
597 **31**, 2035-2037, doi:10.1093/bioinformatics/btv057 (2015).

598 47 Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-
599 assembled contigs using SSPACE. *Bioinformatics* **27**, 578-579,
600 doi:10.1093/bioinformatics/btq683 (2011).

601 48 Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669-2677,
602 doi:10.1093/bioinformatics/btt476 (2013).

603 49 Xue, W. *et al.* L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC Genomics* **14**,
604 604, doi:10.1186/1471-2164-14-604 (2013).

605 50 Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript
606 alignment assemblies. *Nucleic Acids Res.* **31**, 5654-5666 (2003).

607 51 Grabherr, M. G. *et al.* Trinity: reconstructing a full-length transcriptome without a genome
608 from RNA-Seq data. *Nat. Biotechnol.* **29**, 644-652, doi:10.1038/nbt.1883 (2011).

609 52 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**,
610 357, doi:10.1038/nmeth.1923 (2012).

611 53 Chen, Y., Stephens, T. G., Bhattacharya, D., González-Pech, R. A. & Chan, C. X. Evidence
612 that inconsistent gene prediction can mislead analysis of algal genomes. *bioRxiv*, 690040,
613 doi:10.1101/690040 (2019).

614 54 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-
615 generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).

616 55 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein
617 or nucleotide sequences. *Bioinformatics* **22**, 1658-1659, doi:10.1093/bioinformatics/btl158
618 (2006).

619 56 Stanke, M. *et al.* AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids*
620 *Res.* **34**, W435-W439 (2006).

621 57 Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 1 (2004).

622 58 Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. Gene identification
623 in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494-6506,
624 doi:10.1093/nar/gki937 (2005).

625 59 Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management
626 tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491, doi:10.1186/1471-
627 2105-12-491 (2011).

628 60 Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVIDENCEModeler
629 and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, 1 (2008).

630 61 Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32**, D138-D141
631 (2004).

632 62 Li, W. *et al.* The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic*
633 *Acids Res.* **43**, W580-W584 (2015).

634 63 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful
635 approach to multiple testing. *Journal of the Royal Statistical Society. Series B*
636 *(Methodological)*, 289-300 (1995).

637 64 topGO: enrichment analysis for Gene Ontology v. 2 (2010).

638 65 Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput.*
639 *Biol.* **14**, e1005944, doi:10.1371/journal.pcbi.1005944 (2018).

640 66 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform.
641 *Bioinformatics* **26**, 589-595, doi:10.1093/bioinformatics/btp698 (2010).

642 67 Lassmann, T., Hayashizaki, Y. & Daub, C. O. SAMStat: monitoring biases in next generation
643 sequencing data. *Bioinformatics* **27**, 130-131 (2011).

Emms, D. M. & Kelly, S. OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences. *bioRxiv*, 466201, doi:10.1101/466201 (2018).

Wang, Y. *et al.* MCSscanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49-e49, doi:10.1093/nar/gkr1293 (2012).

Acknowledgements

R.A.G.P. is supported by an International Postgraduate Research Scholarship and a University of Queensland Centenary Scholarship. This work is supported by two Australian Research Council grants (DP150101875 awarded to M.A.R., C.X.C. and D.B., and DP190102474 awarded to C.X.C. and D.B.), and the computational resources of the National Computational Infrastructure (NCI) National Facility systems through the NCI Merit Allocation Scheme (Project d85) awarded to C.X.C. and M.A.R.

Author contributions

R.A.G.P., M.A.R. and C.X.C. conceived the study; R.A.G.P., T.G.S., A.R.M., D.W.B., D.B., M.A.R. and C.X.C. designed the analyses and interpreted the results; C.X.C. maintained the dinoflagellate cultures; C.X.C. and A.R.M. extracted biological materials for sequencing; Y. Cheng generated the long-read libraries for genome and full-length transcriptome sequencing; R.A.G.P. and Y. Chen conducted all computational analyses. R.A.G.P. prepared all figures and tables, and prepared the first draft of the manuscript; all authors wrote, reviewed, commented on and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Data availability

The assembled genomes, predicted gene models and proteins from *S. tridacnidorum* CCMP2592 and *S. natans* CCMP2548 are available at <https://cloudstor.aarnet.edu.au/plus/s/095Tqepmq2VBztd>.

669 Tables

670 **Table 1.** Statistics of *de novo* genome assemblies of *S. tridacnidorum* CCMP2592 and *S. natans*
671 CCMP2548.

Metric	<i>S. tridacnidorum</i>	<i>S. natans</i>
Overall G+C (%)	51.01	51.79
Number of scaffolds	6245	2855
Assembly length (bp)	1,103,301,044	761,619,964
N50 scaffold length (bp)	651,264	610,496
Max. scaffold length (Mbp)	4.01	3.40
Number of contigs (bp)	7913	4262
N50 contig length (bp)	356,695	358,021
Max. contig length (Mbp)	2.96	2.90
Gap (%)	0.02	0.02

672

673 **Table 2.** Statistics of predicted genes from genomes of *S. tridacnidorum* and *S. natans*.

Statistic	<i>S. tridacnidorum</i>	<i>S. natans</i>
Genes		
Number of genes	45,474	35,270
Mean gene (exons + introns) length (bp)	10647.95	8779.96
Mean CDS length (bp)	2033.50	1660.13
Gene content (total gene length/total assembly length, %)	43.87	40.66
CDS G+C (%)	57.32	58.16
Supported by transcript data (%)	61.73	82.99
Exons		
Average number per gene	16.15	15.66
Average length (bp)	125.89	106.00
Total length (bp)	92,471,373	58,552,877
Introns		
Number of genes with introns	40,282	30,171
Average length	568.48	485.61
Total length (bp)	391,733,376	251,116,222
G+C (%)	50.20	51.33
Intron-exon boundaries		
5'-donor splice sites (%)	GC (canonical)	56.38
	GT (non-canonical)	25.71
	GA (non-canonical)	17.91
		18.36
Nucleotide after the AG 3'-acceptor splice sites (%)	G	96.53
	A	1.98
	T	0.92
	C	0.57
Intergenic regions		
Average length (bp)	11,467.68	11,585.13
G+C (%)	50.20	51.50

674

Figure Legends

Fig. 1 Comparison of *S. tridacnidorum* and *S. natans* genomes

(a) Density polygon of the similarity between aligned genome sequences of *S. tridacnidorum* and *S. natans* as a function of the length of the aligned region in the query sequence. (b) Proportion of distinct genome features (by sequence length) among the aligned regions between the two genomes. (c) Overlap of the sequences with similarity between both genomes with predicted genes and repetitive elements. (d) Mapping rate of filtered read pairs generated for each species against the assembled genomes of itself and of the counterpart. 'St': *S. tridacnidorum*, 'Sn': *S. natans*. (e) Homologous gene families for the two genomes, showing the number of shared families and those that are exclusive to each genome. (f) Top ten most-abundant protein domains recovered, sorted in decreasing relative abundance (from bottom to top) among proteins of *S. tridacnidorum* (left) and those of *S. natans* (right). The abundance for each domain in both genomes is shown in each chart for comparison. Domains common among the top ten most abundant for both species are connected with a line between the charts. 'MORN': MORN repeat, 'RCC1': Regulator of chromosome condensation repeat, 'RVT': reverse transcriptase, 'DUF': domain of unknown function, 'PPR': pentatricopeptide repeat, 'EFH': EF-hand, 'IonTr': ion transporter, 'Pkin': protein kinase, 'Ank': ankyrin repeat, 'DNAmet': C-5 cytosine-specific DNA methylase. (g) Composition of sequence features for each of the two genomes, showing the percentage of sequences (by length) associated with distinct types of repetitive elements. 'St': *S. tridacnidorum*, 'Sn': *S. natans*.

Fig. 2 Contribution of genomic features to the distinct composition of *S. tridacnidorum* and *S. natans* genomes

Each genome feature was assessed based on the ratio (Δ) of the total length of the implicated sequence region in *S. tridacnidorum* to the equivalent length in *S. natans*, shown in log₂-scale. The ratio of the estimated genome sizes is shown as reference (marked with a dashed line). The untransformed Δ for each feature is shown in its corresponding bar. A genome feature with Δ

700 greater than the reference likely contributed to the discrepancy of genome sizes. Bars are coloured
701 based on the genome in which they are more abundant as shown in the legend.

702 **Fig. 3 Overrepresented functions in retroposed and RT-genes**

703 GO molecular functions enriched in genes with conserved DinoSL relicts in their upstream regions
704 (a) and genes coding for reverse transcriptase domains (RT-genes) (b).

705 **Fig. 4 Interspersed repeat landscapes of *S. tridacnidorum* and *S. natans***

706 Interspersed repeat landscapes of *S. natans* (a) and *S. tridacnidorum* (b). The colour code of the
707 different repeat classes is shown at the bottom of the charts.

708 **Fig. 5 Relative gene-family sizes in *S. tridacnidorum* and *S. natans***

709 Volcano plot comparing gene-family sizes against Fisher's exact test significance (p -value). The
710 colour of the circles indicates the species in which those gene families are larger according to the
711 top-right legend. The number of gene families with the same ratio and significance is represented
712 with the circle size following the bottom-right legend. Filled circles represent size differences that
713 are considered statistically significant (adjusted $p \leq 0.05$).

714 **Fig. 6 Genome proportion of distinct elements in genomes of *S. tridacnidorum*, *S. natans* and** 715 ***P. glacialis***

716 Proportion (in percentage of the sequence length) covered by different types of genome features in
717 the hybrid assemblies of *S. tridacnidorum*, *S. natans* and *P. glacialis*.

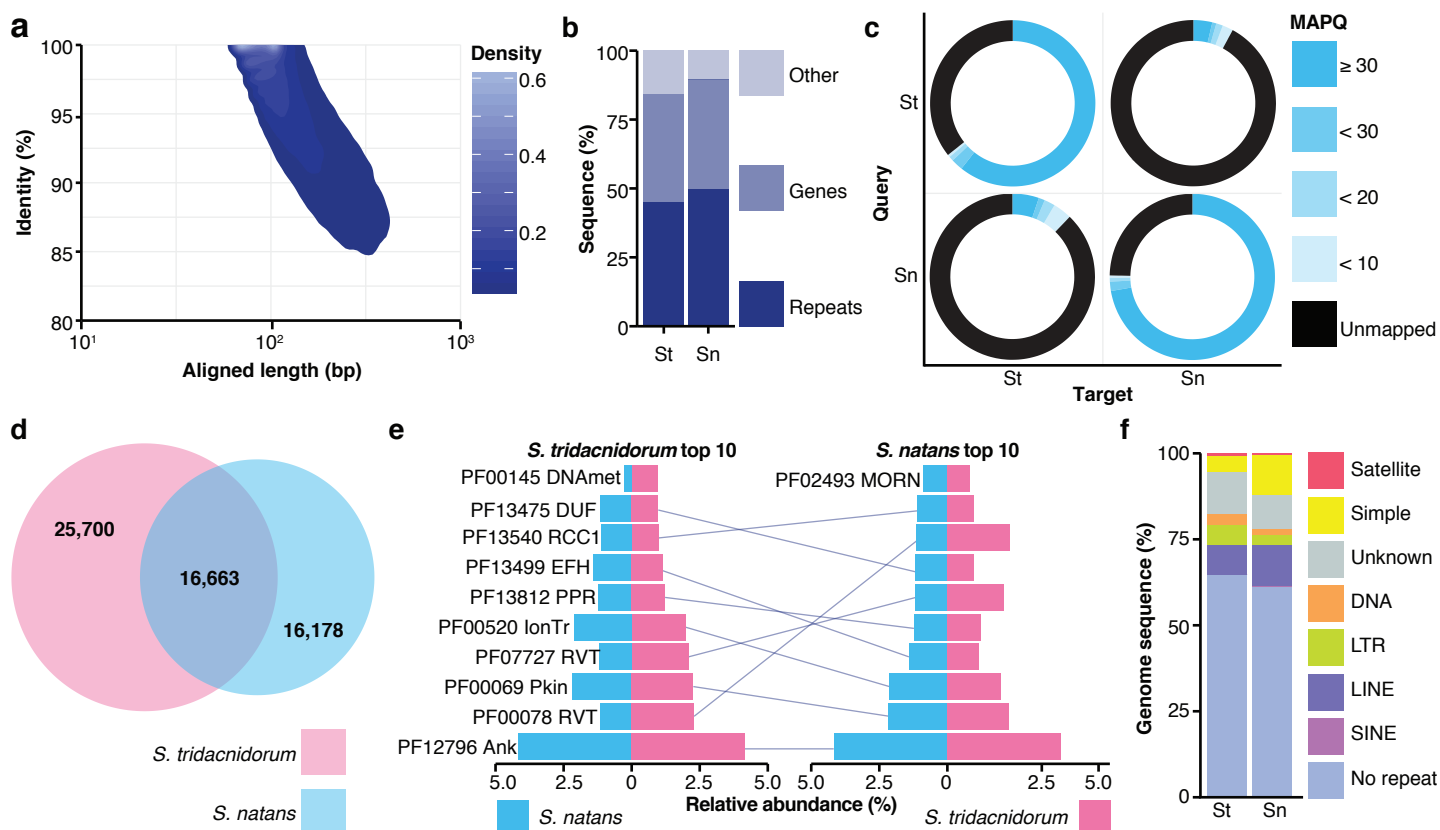


Fig. 1 Comparison of *S. tridacnidorum* and *S. natans* genomes

(a) Density polygon of the similarity between aligned genome sequences of *S. tridacnidorum* and *S. natans* as a function of the length of the aligned region in the query sequence. (b) Proportion of distinct genome features (by sequence length) among the aligned regions between the two genomes. Overlap of the sequences with similarity between both genomes with predicted genes and repetitive elements. (c) Mapping rate of filtered read pairs generated for each species against the assembled genomes of itself and of the counterpart. 'St': *S. tridacnidorum*, 'Sn': *S. natans*. (d) Homologous gene families for the two genomes, showing the number of shared families and those that are exclusive to each genome. (e) Top ten most-abundant protein domains recovered, sorted in decreasing relative abundance (from bottom to top) among proteins of *S. tridacnidorum* (left) and those of *S. natans* (right). The abundance for each domain in both genomes is shown in each chart for comparison. Domains common among the top ten most abundant for both species are connected with a line between the charts. 'MORN': MORN repeat, 'RCC1': Regulator of chromosome condensation repeat, 'RVT': reverse transcriptase, 'DUF': domain of unknown function, 'PPR': pentatricopeptide repeat, 'EFH': EF-hand, 'IonTr': ion transporter, 'Pkin': protein kinase, 'Ank': ankyrin repeat, 'DNAmet': C-5 cytosine-specific DNA methylase. (f) Composition of sequence features for each of the two genomes, showing the percentage of sequences (by length) associated with distinct types of repetitive elements. 'St': *S. tridacnidorum*, 'Sn': *S. natans*.

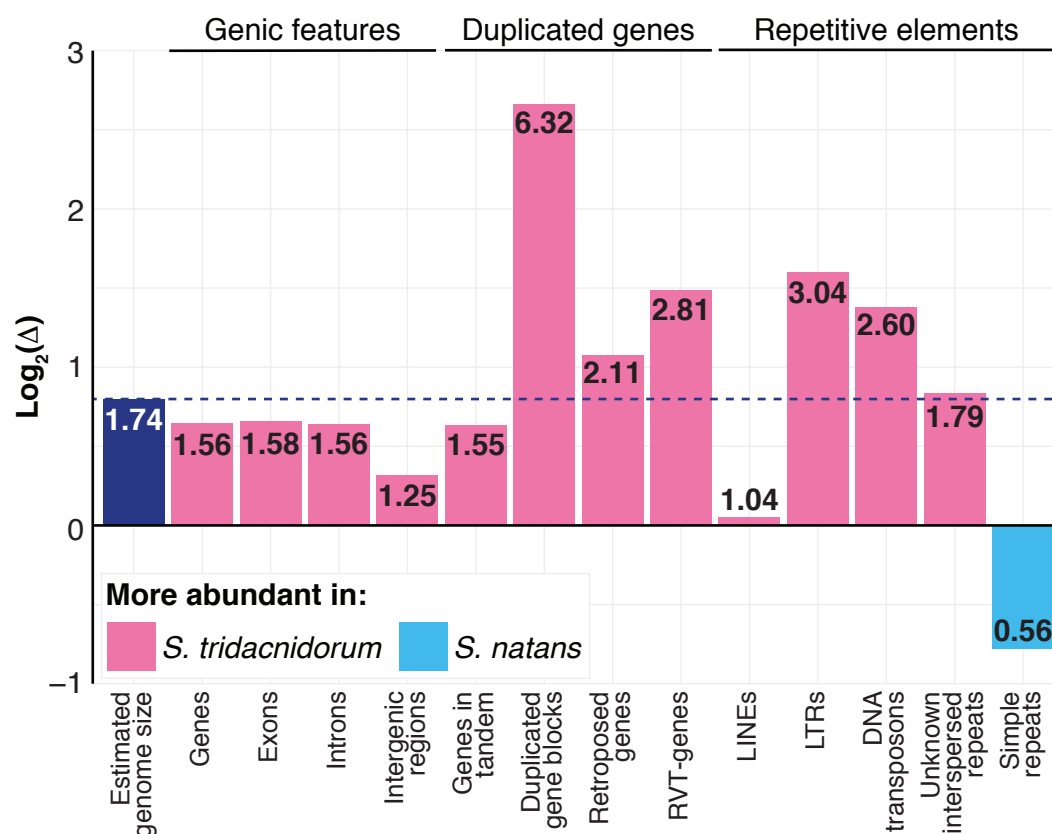


Fig. 2 Contribution of genomic features to the distinct composition of *S. tridacnidorum* and *S. natans* genomes

Each genome feature was assessed based on the ratio (Δ) of the total length of the implicated sequence region in *S. tridacnidorum* to the equivalent length in *S. natans*, shown in log_2 -scale. The ratio of the estimated genome sizes is shown as reference (marked with a dashed line). The untransformed Δ for each feature is shown in its corresponding bar. A genome feature with Δ greater than the reference likely contributed to the discrepancy of genome sizes. Bars are coloured based on the genome in which they are more abundant as shown in the legend.

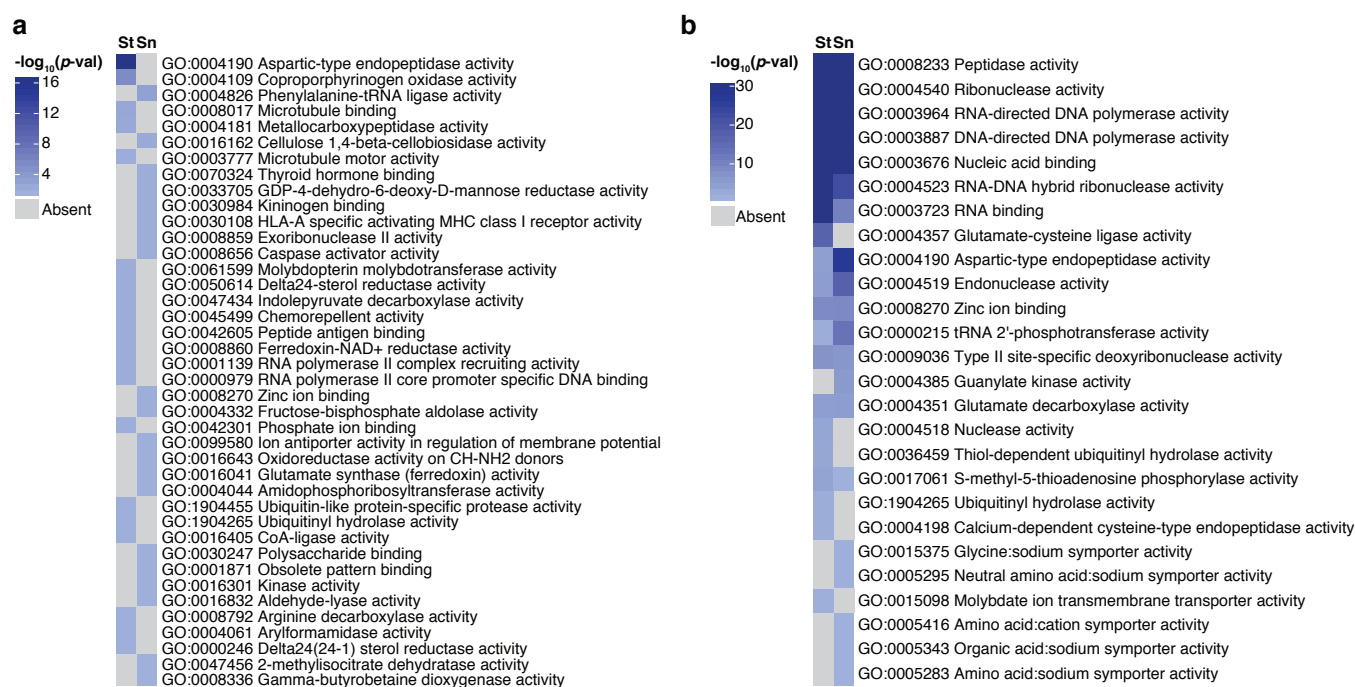


Fig. 3 Overrepresented functions in retroposed and RT-genes

GO molecular functions enriched in genes with conserved DinoSL relicts in their upstream regions **(a)** and genes coding for reverse transcriptase domains (RT-genes) **(b)**.

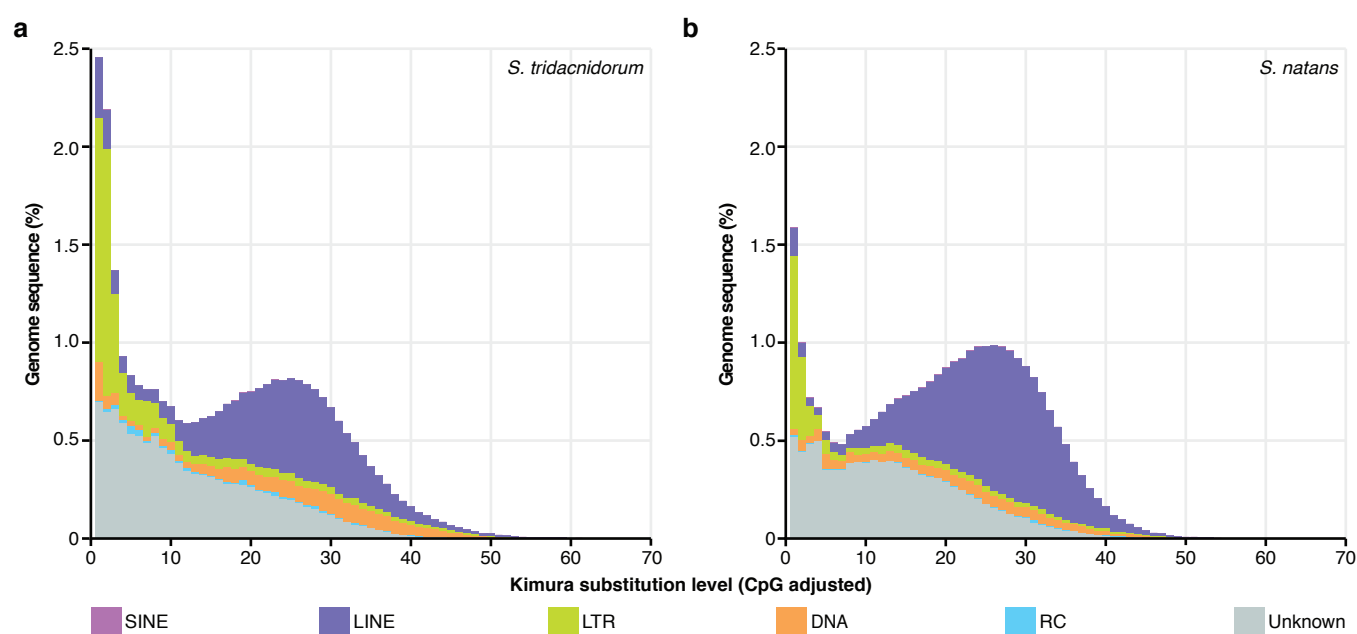


Fig. 4 Interspersed repeat landscapes of *S. tridacnidorum* and *S. natans*

Interspersed repeat landscapes of *S. natans* (**a**) and *S. tridacnidorum* (**b**). The colour code of the different repeat classes is shown at the bottom of the charts.

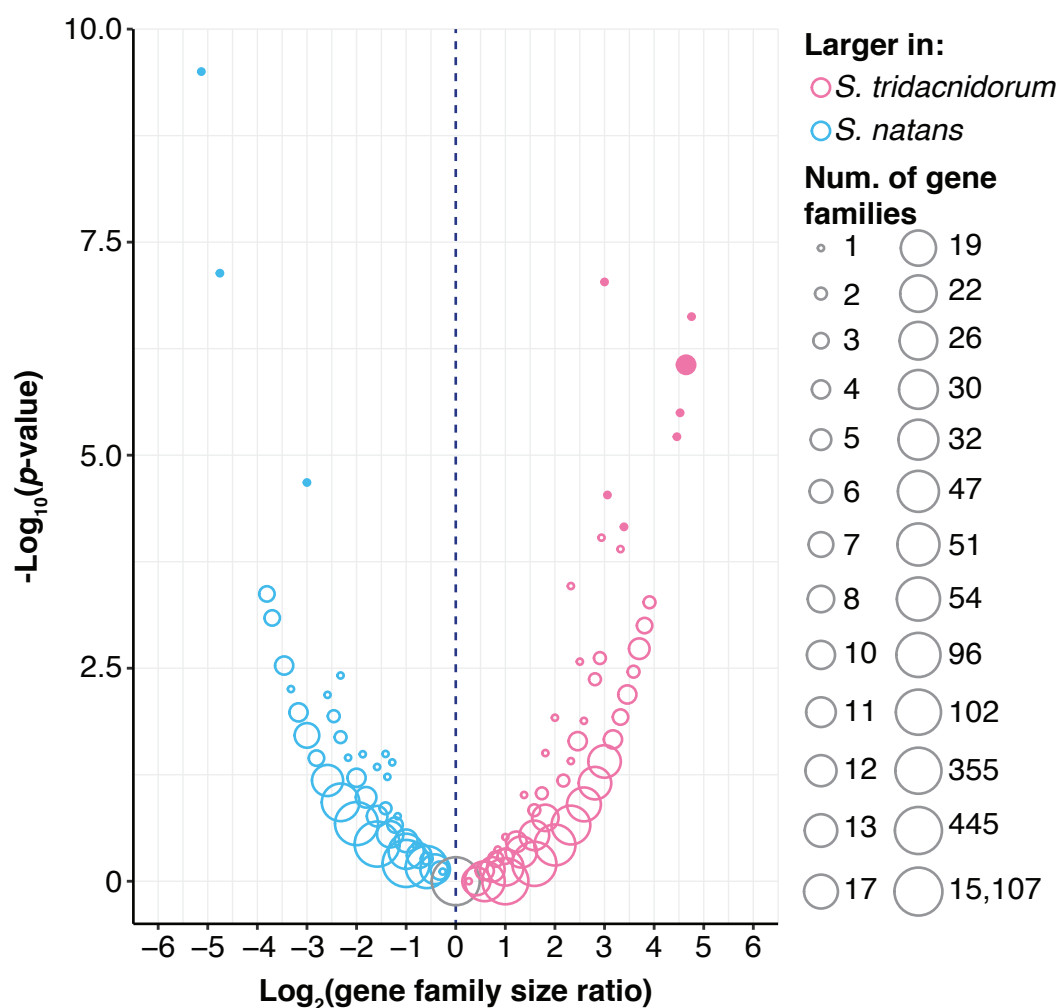


Fig. 5 Relative gene-family sizes in *S. tridacnidorum* and *S. natans*

Volcano plot comparing gene-family sizes against Fisher's exact test significance (p -value). The colour of the circles indicates the species in which those gene families are larger according to the top-right legend. The number of gene families with the same ratio and significance is represented with the circle size following the bottom-right legend. Filled circles represent size differences that are considered statistically significant (adjusted $p \leq 0.05$).

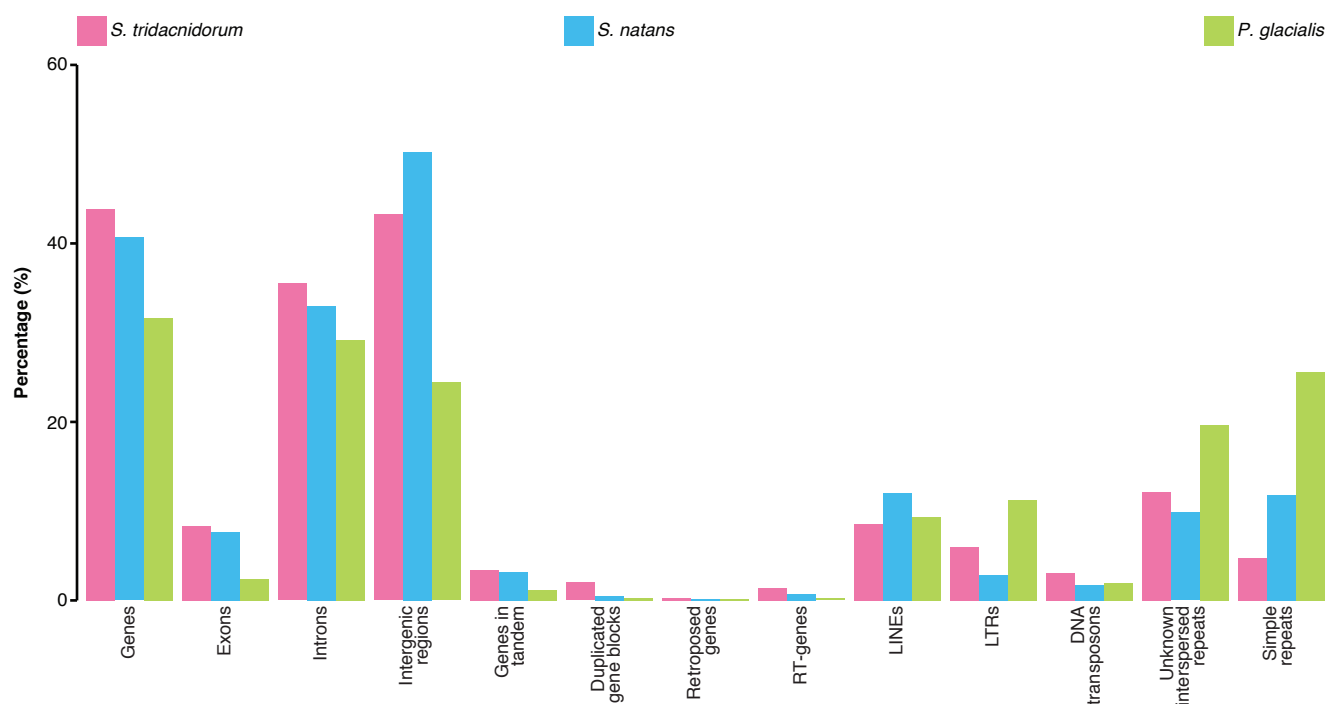


Fig. 6 Genome proportion of distinct elements in genomes of *S. tridacnidorum*, *S. natans* and *P. glacialis*

Proportion (in percentage of the sequence length) covered by different types of genome features in the hybrid assemblies of *S. tridacnidorum*, *S. natans* and *P. glacialis*.