Article

Symbiodinium genomes reveal adaptive evolution of functions related to symbiosis

Huanle Liu[1], Timothy G. Stephens[1], Raúl A. González-Pech[1], Victor H. Beltran[2], Bruno Lapeyre[3,4], Pim Bongaerts[5], Ira Cooke[3], David G. Bourne[2,6], Sylvain Forêt[7,*], David J. Miller[3], Madeleine J. H. van Oppen[2,8], Christian R. Voolstra[9], Mark A. Ragan[1] and Cheong Xin Chan[1,10,†]

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

[2]Australian Institute of Marine Science, Townsville, QLD 4810, Australia

[3]ARC Centre of Excellence for Coral Reef Studies and Department of Molecular and Cell Biology, James Cook University, Townsville, QLD 4811, Australia

[4]Laboratoire d'excellence CORAIL, Centre de Recherches Insulaires et Observatoire de l'Environnement, Moorea 98729, French Polynesia

[5]Global Change Institute, The University of Queensland, Brisbane, QLD 4072, Australia

[6]College of Science and Engineering, James Cook University, Townsville, QLD 4811, Australia

[7]Research School of Biology, Australian National University, Canberra, ACT 2601, Australia

[8]School of BioSciences, The University of Melbourne, VIC 3010, Australia

[9]Red Sea Research Center, Division of Biological and Environmental Science and Engineering, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia

[10]School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD 4072, Australia

[*]Deceased 17 December 2016

[†]Correspondence and requests for materials should be addressed to Cheong Xin Chan (c.chan1@uq.edu.au)

Abstract

Symbiosis between dinoflagellates of the genus *Symbiodinium* and reef-building corals forms the trophic foundation of the world's coral reef ecosystems. Here we present the first draft genome of *Symbiodinium goreaui* (Clade C, type C1: 1.03 Gbp), one of the most ubiquitous endosymbionts associated with corals, and an improved draft genome of *Symbiodinium kawagutii* (Clade F, strain CS-156: 1.05 Gbp), previously sequenced as strain CCMP2468, to further elucidate genomic signatures of this symbiosis. Comparative analysis of four available *Symbiodinium* genomes against other dinoflagellate genomes led to the identification of 2460 nuclear gene families that show evidence of positive selection, including genes involved in photosynthesis, transmembrane ion transport, synthesis and modification of amino acids and glycoproteins, and stress response. Further, we identified extensive sets of genes for meiosis and response to light stress. These draft genomes provide a foundational resource for advancing our understanding *Symbiodinium* biology and the coral-algal symbiosis.

Introduction

Coral reefs provide habitats for one-quarter to one-third of all marine species[1]. Although typically surrounded by nutrient-poor waters, coral reefs show high rates of primary productivity, with the fixed carbon supporting not only the biomass of reef organisms but also commercial and recreational fisheries and aquaculture. Reef-building corals rely on the symbiosis between the coral animal *per se* and photosynthetic dinoflagellates of the genus *Symbiodinium*. This symbiosis is based on mutual nutrient exploitation, with corals providing shelter and inorganic nutrients to their algal partners, while *Symbiodinium* supply their coral hosts with photosynthates that can meet up to 95% of the corals' energy requirements[2,3].

The relationship between *Symbiodinium* and their host determines not only the rate of coral-reef growth (calcium carbonate deposition), but also how the system responds to environmental stress[3]. Many studies have shown that coral-*Symbiodinium* mutualism is susceptible to environmental factors including temperature, light and salinity. Exposure to ultraviolet radiation, thermal stress or a combination of both can initiate photoinhibition, decoupling of carbon flow between symbiont and host, oxidative damage and breakdown of the symbiosis, a phenomenon known as coral bleaching[4]. Unless the symbiosis is soon re-established the coral host is at risk of starvation, disease and eventual death[5]. In recent decades, coral bleaching has led to large-scale mortality on coral reefs around the world, with the most recent global coral bleaching event (2014-2016) now confirmed as the longest and most severe on record[6,7].

Despite the critical importance of this coral-dinoflagellate symbiosis, little is known about the underlying molecular mechanisms (apart from photosynthesis and carbon exchange), largely due to the lack of comprehensive understanding of what molecules, pathways and functions *Symbiodinium* can contribute. Genomes of *Symbiodinium* (and of

66 dinoflagellates more broadly) are known for their idiosyncratic features including non-
67 canonical splice sites, extensive methylation[8] and large sizes, up to 250 Gbp[9]. Their plastid
68 genomes occur as plasmid-like minicircles[10-12]; their mitochondrial genomes harbor only
69 three protein-coding genes and lack stop codons, and both mitochondrial[13-15] and nuclear[16]
70 transcripts are extensively edited.

71 *Symbiodinium* are classified into nine clades[17-19], with members of Clades A, B, C and
72 D responsible for the vast majority of associations with scleractinian corals[20]. Draft genomes
73 have been published for representatives of Clades A, B and F[17-19], with sequence comparisons
74 demonstrating them to be highly divergent[18]. Genome sequences are still lacking for Clade C,
75 the most ubiquitous and diverse clade associated with tropical reef corals[21], at least some sub-
76 clades ("types") of which are ecologically partitioned[22].

77 Here we report draft genomes of two *Symbiodinium* from the Pacific Ocean: *S. goreaui*
78 (type C1; isolated from the acroporid coral *Acropora tenuis*) from the Great Barrier Reef, and
79 *S. kawagutii* CS-156 (=CCMP2468, Clade F) from Hawaii. *Symbiodinium* type C1 is one of
80 two "living ancestors" (along with type C3) of Clade C[21], and one of the most dominant types
81 associated with reef corals in both Indo-Pacific and Caribbean waters[20]. *S. goreaui* has been
82 reported from >150 coral species on Australia's Great Barrier Reef, representing >80% of the
83 studied coral genera in this region[23] across environments from reef flats to lower mesophotic
84 depths[23-25]. In contrast, *S. kawagutii* CS-156 (=CCMP2468) was isolated during attempts to
85 culture the symbiont from *Montipora verrucosa* (Todd LaJeunesse, *personal*
86 *communication*). This isolate has yet to be verified to occur in mutualistic symbiosis with any
87 coral, and appears incapable of establishing experimental symbiosis with cnidarian hosts[26].
88 Instead *S. kawagutii* may be exclusively a symbiont of foraminifera, or occur free-living at
89 low environmental densities but proliferate opportunistically in culture. As some genome data

have been published for *S. kawagutii* CCMP2468[18], we used these in combination with new data from the present study to generate a refined genome assembly. The genomes of *S. goreaui* and *S. kawagutii* offer a platform for comparative genomic analyses between two of the most-recently diverged *Symbiodinium* lineages Clades C and F, and published genome sequences in the more-basal Clades A and B.

Adopting a comparative approach using both genome and transcriptome data, we systematically investigated genes and functions that are specific to *Symbiodinium* vis-à-vis other dinoflagellates, and their association with the establishment and maintenance of symbiosis. We also computationally identify genes and functions for which there is evidence of adaptive selection in *Symbiodinium*. This is the most-comprehensive comparative analysis so far of *Symbiodinium* genomes, and the first to include a prominent endosymbiont of corals of Indo-Pacific and Caribbean reefs.

Results

Genomes of S. goreaui and S. kawagutii

We sequenced and generated two draft *Symbiodinium* genome assemblies *de novo*, for *S. goreaui* (Clade C, 1.03 Gbp) and for *S. kawagutii* (Clade F, 1.05 Gbp). Details of data generation and assembly statistics are shown in Supplementary Tables S1 and S2 respectively. Our *S. goreaui* assembly consists of 41,289 scaffolds (N50 length 98,034 bp). For *S. kawagutii*, we first verified that our data (from isolate CS-156) and the published data (from the synonym isolate CCMP2468) are indeed from the same culture of origin (see Supplementary Methods and Supplementary Figure S1). Compared to the published assembly (Lin et al.[18]), independent mapping of their ten fosmid sequences[18] onto our preliminary CS-156 assembly yielded up to 43-fold and 37-fold fewer gaps and mismatches respectively (Supplementary Figure S2). We later combined both datasets in a single *de novo* assembly,

yielding 16,959 scaffolds (N50 length 268,823 bp). Genome-size estimates based on *k*-mer coverage are 1.19 Gbp for *S. goreaui* and 1.07 Gbp for *S. kawagutii* (Supplementary Table S3), comparable to those for other sequenced *Symbiodinium* genomes. We also recovered sequences putatively derived from their plastid genomes (Supplementary Tables S4, S5 and S6) including their distinct core conserved regions (Supplementary Table S7), and from their mitochondrial genomes; see Supplementary Note for details.

The repeat content of the assembled genomes ranged from 16.0% (*S. kawagutii*) to 27.9% (*S. microadriaticum*); a large peak in transposable element (TE) abundance observed at high divergence (Kimura distance[27] 15-25) in all genomes (Supplementary Figure S3) suggests that most extant transposable elements are remnants of an ancient burst of TE activity. TE activity has been broadly linked to genome size in plants[28-30], so reduced TE activity may be connected with the relative compactness of *Symbiodinium* genomes in comparison to those of other dinoflagellates. However, as these genomes are still in draft, the impact of assembly completeness on the patterns of repeat divergence cannot be dismissed.

Using a stringent threshold to remove genome scaffolds of potential bacterial or viral origin (Methods), we predict 35,913 and 26,609 high-quality gene models respectively for *S. goreaui* and *S. kawagutii* (Supplementary Table S8). Usage profiles of codons and amino acids are shown in Supplementary Figures S4 and S5 respectively, and non-canonical splice sites in Supplementary Table S9 and Supplementary Figure S6. Although we report fewer genes, the majority (67.0% and 64.4% respectively for *S. goreaui* and *S. kawagutii*) have transcriptome support, and we generally recovered more of the 458 conserved core eukaryote genes (Supplementary Figure S7), 371 of which are common to all four *Symbiodinium* based on the predicted gene models (Figure 1a; Supplementary Table S10); similar results are observed for the corresponding genome sequences (Supplementary Figure S7). About 94% of

the predicted genes have introns, similar to *S. microadriaticum* (98.2%) and *S. minutum* (95.3%); the earlier *S. kawagutii* genome assembly[18] may have underestimated the proportion of intron-containing genes (Supplementary Table S8) due to a less-stringent approach to gene prediction. All coding sequences have higher G+C content (56.7% in *S. goreaui* and 55.0% in *S. kawagutii*) than does the genome overall, comparable to coding sequences from other *Symbiodinium* (57.7% in *S. microadriaticum* and 52.7% in *S. minutum*).

Sequence divergence and synteny

Despite the seemingly high number of protein-coding genes, an earlier analysis of syntenic blocks[17] found only several hundred blocks conserved in a pairwise manner among three published *Symbiodinium* genomes. Here we included our two new genome sequences in this analysis, and focused further on syntenic collinear blocks, requiring each block to contain the same genes in the same order and orientation with no gene losses (Methods). The *S. goreaui* and *S. kawagutii* genomes share the most collinear blocks, 889 blocks implicating 8621 genes; 62 of these blocks are of size >15, with the largest containing 76 genes (Supplementary Table S11). Thus substantial proportions of genes in these two genomes occur in clusters: for cluster size ≥ 5 genes, 32.4% and 24.0% of *S. kawagutii* and *S. goreaui* genes respectively. These are likely to be underestimates, as the assemblies remain fragmentary. At the other end of the spectrum, the genomes of *S. microadriaticum* and *S. goreaui* share only 86 collinear blocks of size ≥ 5, with maximum size 12 and implicating 588 genes in total (Supplementary Table S11). These results suggest that (a) although Clades C and F are divergent, they nonetheless are the most-closely related among the four analysed *Symbiodinium* genomes (in line with their phylogenetic relationship); and (b) C and F are more divergent from Clade A than from Clade B (in line with their phylogenetic relationship). Thus whole-genome sequences support and extend earlier conclusions, based

on common marker sequences, about phylogenetic relationships among *Symbiodinium* clades[31-33] and the remarkable divergence among *Symbiodinium* lineages[17,18].

Gene and protein functions

All annotated genes from *S. goreaui* and *S. kawagutii* genomes are listed in Supplementary Tables S12 and S13 respectively. Of the 35,913 proteins predicted in *S. goreaui*, 31,646 (88.1%) show similarity (BLASTP, $E \leq 10^{-5}$) to sequences in UniProt; among these, 29,198 (81.3% of 35,913) and 19,718 (54.9%) are annotated with Gene Ontology (GO) terms or Pfam domains (Supplementary Tables S12 and S14). For *S. kawagutii*, 21,947 of 26,609 proteins (82.5%) find a match in UniProt (Supplementary Tables S13 and S14). *Protein kinase* (Pfam PF00069), *reverse transcriptase* (PF07727), *ion transport protein* (PF00520) and *ankyrin repeats* (PF12796) are among the most-abundant domains in *Symbiodinium*, appearing among the ten most-abundant for each of the four genomes (Supplementary Table S15). Ankyrin repeat motifs are important in the recognition of surface proteins, and more generally in protein-protein interactions[34]. Thus proteins potentially involved in host-symbiont interaction (with phosphorylation, ion transport and protein recognition/interaction domains) are well represented in the predicted *Symbiodinium* proteomes.

We compared functions of proteins predicted from these four *Symbiodinium* genomes to a set of 27 more-narrowly scoped eukaryotes: 17 alveolates (ten other dinoflagellates, four ciliates, two apicomplexans and *Perkinsus marinus*), stramenopiles (two diatoms) and Archaeplastida (four rhodophytes, three chlorophytes and *Arabidopsis*). This 31-taxon set (1,136,347 proteins; Supplementary Tables S16 and S17) represents lineages in which one or more endosymbioses are implicated in plastid origin[35-37]; these proteins were clustered (based on sequence similarity) into 56,530 groups of size two or greater (Supplementary Table S17; see Methods). Using this 31-taxon dataset as background, we assessed the over- or under-

representation of protein domains within our various groups of *Symbiodinium* proteins. We found 270 domains (Supplementary Table S18) to be significantly overrepresented (adjusted $p \leq 0.05$) in *Symbiodinium*. Interestingly, many domains e.g. *C-5 cytosine-specific DNA methylase* (PF00145), *planctomycete cytochrome c* (PF07635) and RNA polymerase *sigma-70 region 2* (PF04542) are also overrepresented in the four *Symbiodinium* genomes in a similar comparison against 880,909 proteins in a 15-taxon set that includes ten other dinoflagellates and the immediate outgroup *Perkinsus marinus* (Supplementary Table S19). Thus compared to related eukaryotes and to other dinoflagellates, *Symbiodinium* is enriched in functions involved in methylation of cytosine, (photosynthetic) energy production and RNA polymerisation. Hydroxymethylation of uracil is common (12-70%) in dinoflagellate genomes[38]; while cytosine methylation has been described in *Symbiodinium*[39,40], our findings suggest that cytosine methylation is more prominent in *Symbiodinium* than in these other dinoflagellates.

Activation of some retrotransposons is part of the stress-response mechanism in diatoms[41], plants[42,43] and other eukaryotes[44,45]. The *reverse transcriptase* domain (PF07727) is enriched in *Symbiodinium* in both the 31-taxon and 15-taxon sets, suggesting that retrotransposition could be a prominent mechanism of stress response in *Symbiodinium* and dinoflagellates. Although we set a stringent threshold for removing putative bacterial or viral sequences (see Methods), 40 (~0.1%) of the final 41,289 genome scaffolds of *S. goreaui* have significant hits (BLASTN $E \leq 10^{-20}$) to the virus genomes[46] isolated from the same *S. goreaui* (type C1) strain, with 16 identical regions (76-609 bp) distributed in nine scaffolds of lengths ranging from 1092 to 7,338,656 bp. Whether this indicates introgression of viral sequences remains to be determined.

Positive selection of Symbiodinium genes

Using a branch-site model based on the ratio of substitution rates in non-synonymous (dN) to synonymous (dS) sites[47] (Methods and Supplementary Figure S8), we identified *Symbiodinium* genes under positive selection in comparison to ten other dinoflagellates, with *Perkinsus marinus* as the outgroup (15 taxa: Supplementary Tables S16 and S17). The reference species tree (Figure 2a) was computed following Price and Bhattacharya[48]. We then based our analysis of adaptive evolution on all orthologous sets for which the protein tree is topologically congruent with our reference tree (Methods).

Of our 44,282 homologous sets, 2460 containing 7987 *Symbiodinium* proteins show evidence of positive selection in one or more *Symbiodinium* lineages; 1069 of these sets are annotated with GO terms (Supplementary Table S20). Figure 2b shows the terms (level 3) in the three GO hierarchies that are shared by ≥ 5% of these 1069 sets. In the Biological Process hierarchy, metabolic processes are highly represented (*primary metabolic process* [292] and *macromolecule compound metabolic process* [243] are among the four most-frequent terms), followed by *oxidation reduction* [96] and transport (*establishment of localization* [90], *transport* [90], and *transmembrane transport* [80]). Highly represented terms in the Molecular Function hierarchy point to binding of diverse molecules and ions e.g. *protein binding* [173] and metabolism (*hydrolase* [390], *transferase* [344]). In Cellular Component, *cell part* [113], *membrane* [86] and *membrane part* [59] are the most frequent. Thus in *Symbiodinium* as represented by these four assemblies, broad aspects of metabolism, and transport including across membranes, show evidence of positive selection, in line with their recognised importance in cnidarian-dinoflagellate symbioses[17].

We further assessed enrichment of GO terms against all annotated terms specifically in the four *Symbiodinium* genomes (Supplementary Table S21) compared with the other

dinoflagellates in this study. Here we consider themes among Biological Process terms. The first theme is that functions associated with photosynthetic light reactions are enriched among the positively selected *Symbiodinium* genes; *photosynthesis, light reaction* and *Photosystem II assembly* are significantly over-represented (adjusted $p \leq 0.05$), as are Cellular Component terms related to plastid functions e.g. *thylakoid*, *photosynthetic membrane*, *intracellular membrane-bounded organelle* (Supplementary Table S21). Coral bleaching has been associated with the loss of light-harvesting proteins and the subsequent inactivation of photosystem II (PSII) in *Symbiodinium* under combined light and temperature stress[49-51]. These authors reported that coral bleaching associated with algal photobleaching can be ameliorated, at least in part, by thermal acclimation of *Symbiodinium* to improve the thermal tolerance of PSII.

The second emerging theme involves the transport of ions and metabolites across membranes. *Intracellular transport*, *cytosolic transport*, *transition metal ion transport* and *copper ion transport* as well as terms related to transmembrane transport of amino acids, organic acids and carboxylic acids are significantly enriched (Supplementary Table S21); these functions underpin multiple physiological processes, including but not limited to pH regulation, calcification and photosynthetic carbon fixation[52]. Investigated *Symbiodinium* are enriched in membrane transporters compared with other dinoflagellates[17]. These biological processes are especially relevant to the maintenance and regulation of coral-dinoflagellate symbiosis[52,53], possibly including its sensitivity and/or response to environmental stress.

The third theme is the enrichment of functions related to the biosynthesis and modification of amino acids and glycoproteins (Supplementary Table S21) e.g. *protein phosphorylation*, *peptide biosynthesis process*, *protein ADP-ribosylation*, *protein glycosylation*, *D-amino acid metabolic process* and *glycoprotein biosynthetic process*. Corals

lack the capacity to synthesise a number of amino acids (e.g. cysteine in *Acropora digitifera*[54]), thus selection acting on the synthesis of amino acids may indicate the critical role of *Symbiodinium* in supplying amino acids both for self-preservation and for the coral host. Glycoprotein molecules are often surface-localised and in microbes form the basis of microbe-associated molecular patterns (MAMPs) which, in conjunction with a host-associated pattern recognition receptor, mediate recognition by a host[55]. Both in culture and *in hospite*, *Symbiodinium* exude glycoconjugates[56-59]. Where investigated, cell-surface glycan profiles are stable over time within a *Symbiodinium* culture but can differ between clades within a host[60]. *N*-acetyl and mannosyl residues are prominent constituents of *Symbiodinium* cell-surface glycans, making them candidates for MAMPs that could participate in the establishment of symbiosis. Lin et al.[18] reported a *S. kawagutii* glycan biosynthesis pathway distinct from that of *S. minutum*, again pointing to a possible role of glycans in specificity of host recognition[60,61]. Neubauer et al.[62] demonstrated that the thrombospondin type 1 repeat (TSR) from the sea anemone *Aiptasia pallida* contains binding sites for glycosaminoglycan, and that blocking TSR led to decreased colonisation by *S. minutum*. Our results offer the first evidence of positive selection of functions underlying the biosynthesis and modification of amino acids and glycoproteins, suggesting that these functions are critical in the establishment of cnidarian-dinoflagellate symbioses.

Our fourth emerging theme is stress response. Enriched terms annotated for the positively selected genes include *cell redox homeostasis*, *translation initiation* and 22 terms describing the negative regulation of gene expression, transcription, RNA biosynthesis and cellular biosynthetic and metabolic processes (Supplementary Table S21). Environmental stressors can disrupt the cellular redox homeostasis and break down the coral-dinoflagellate symbiosis. Negative regulation of transcription may represent a stress response that safeguards the genome from accumulating DNA damage[63]; a similar stress response has been

observed in coral[64]. Other enriched functions that may be related to stress response include *mitotic nuclear division*, *translation*, and various processes of nucleotide biosynthesis and modification e.g. *RNA methylation, rRNA methylation*, *DNA replication*, *RNA processing,* and *deoxyribonucleotide biosynthetic process*. Our results thus provide evidence that stress response is under positive selection in *Symbiodinium*, presumably (given their lifestyle) in relation to the establishment and/or maintenance of symbiosis.

Do Symbiodinium have sex?

Symbiodinium have been hypothesised to reproduce sexually and to have a diploid life stage[65] but definitive evidence for sex, e.g. karyogamy and meiosis, has yet to be observed[31,66-68]. The ability to reproduce sexually offers increased efficiency of selection and adaptation[69]. So far, the strongest evidence supporting meiotic potential in *Symbiodinium* comes from patterns of population-genetic variation revealed in allozymes, randomly amplified polymorphic DNA and other molecular markers[22,31,70-73]. Indeed, for some markers a higher degree of genetic variation has been observed in certain *Symbiodinium* clades than in dinoflagellates known to reproduce sexually[70]. More recently, differential gene expression analysis[74] using a heterologous culture from which our sequenced *S. goreaui* was derived revealed an enrichment of gene functions related to meiosis under thermal stress, suggesting a switch from mitosis to meiosis under stress conditions.

Schurko and Logsdon[75] described a "meiosis detection toolkit", a set of 51 genes specific or related to meiosis[76,77] that collectively point to a capacity for meiosis even in divergent or specialised eukaryotic genomes[78]. Incomplete genome coverage or assembly, sequence divergence, paralogy, patterns of overlapping function and evolutionary specialisation mean that not all 51 need to be present or detectable for a lineage to be assessed as probably sexual, or only recently asexual[75,77]. Thirty-one of these genes were

earlier identified in *Symbiodinium* Clades A and B[76]. Here, BLASTP search ($E \leq 10^{-5}$) of predicted proteins in these four *Symbiodinium* genomes recovered matches corresponding to 48 of the of 51 toolkit genes in *S. microadriaticum*, 47 in *S. minutum* and *S. goreaui*, and 46 in *S. kawagutii* (Figure 3a and Supplementary Table S22). Eight of the 11 meiosis-specific proteins were detected in all four *Symbiodinium*. REC114, SAD1 and XRS2 found weaker matches ($E \geq 10^{-14}$) in one to two genomes, although confirmatory UniProt domains were usually present (Supplementary Table S22). RAD17 is the *Schizosaccharomyces pombe* homolog of *S. cerevisiae* RAD24[79], for which we find highly significant matches ($E \leq 10^{-127}$) in all four *Symbiodinium*. Moreover, 15 of the 51 genes show evidence of positive selection in *Symbiodinium* against other dinoflagellates (Supplementary Table S22). Our data imply that these four *Symbiodinium* are, or until recently have been, capable of meiosis.

Response to light stress

Mycosporine-like amino acids (MAAs) are UV-protective compounds that, in corals and other marine organisms, also act as antioxidants scavenging reactive oxygen species (ROS). Up to five MAAs have been reported in *Symbiodinium* (Clades A, B and C) isolated from cnidarian hosts[80,81]. The MAA biosynthetic pathway involves dehydroquinate synthase (DHQS), *O*-methyltransferase (O-MT), an ATP-grasp and non-ribosomal peptide synthetase (NRPS)[82,83]. In cyanobacteria, a short-chain dehydrogenase may play a role in converting shinorine to palythine-serine[84]. Genes encoding these four MAA-biosynthetic enzymes were reported absent from the *S. kawagutii* genome[18]. Here, using known proteins in bacteria, fungi and cnidarians as queries, we recovered all five enzymes including the short-chain dehydrogenase from the *S. microadriaticum*, *S. goreaui* and *S. kawagutii* genomes (Supplementary Table S23); ATP-grasp was not found in *S. minutum*. These enzymes were earlier reported absent from *S. kawagutii*, and it was proposed that their absence can be compensated via coral-*Symbiodinium* co-evolution[18]; this hypothesis remains to be

331 investigated, but we note that this *S. kawagutii* isolate has not been observed in association
332 with an animal host[26].

333 Scytonemin is a UV-blocker first reported in terrestrial cyanobacteria[83,85], and in
334 contrast to MAAs was thought to be synthesised exclusively by cyanobacteria[86]. The genome
335 of *Nostoc punctiforme* contains an 18-gene operon that specifies proteins of scytonemin
336 biosynthesis and regulation, including the synthesis of aromatic amino-acid precursors[87,88]. Its
337 expression is up-regulated by UV radiation[89]. Homologs of six of these 18 genes have been
338 described in the coral *Acropora ditigifera*, and were considered putative instances of lateral
339 genetic transfer[90]. We find 12 of these 18 genes in the genomes of *S. goreaui* and in *S.*
340 *kawagutii*, 11 in *S. microadriaticum* and ten in *S. minutum* (Figure 3b, Supplementary Table
341 S24).

342 Genes responsible for biosynthesis of tryptophan (*trpA*, *trpB*, *trpC*, *trpD* and *trpE*) and
343 the two key enzymes of chorismate (and aromatic amino acid) biosynthesis, *aroG* and *aroB*
344 (dehydroquinate synthase, also important for MAA biosynthesis), are found in all
345 *Symbiodinium* genomes, albeit so far in different scaffolds; these genes are also present in
346 *Arabidopsis thaliana* although not in corals or *Hydra* which, like most other animals, are
347 unable to synthesise tryptophan. The recovery of more of these 18 genes in *Symbiodinium*
348 than in corals or other animals (Figure 3b) could reflect the impact of endosymbiotic
349 association of ancestral cyanobacteria during the course of plastid evolution in all
350 photosynthetic eukaryotes including dinoflagellates[35-37]. The presence of multiple gene copies
351 (Supplementary Table S24) also implicates genetic duplication. Our findings suggest that
352 *Symbiodinium* has the capacity to produce scytonemin, either by itself or jointly within the
353 holobiont.

Discussion

Symbiodinium can form associations with a wide range of cnidarian hosts (as well as some other marine invertebrates and protists) across broad geographic and time scales[91]. The symbiosis between corals and *Symbiodinium* relies on compatible host-symbiont recognition and sustainable nutrient exchange, both of which are vulnerable to external environmental factors including temperature and light. A sustainable coral-*Symbiodinium* association requires an adaptive capacity in the face of environmental extremes.

In this study we generated the first draft genome of S. goreaui (Clade C), a much-improved draft genome of S. kawagutii (Clade F) and high-quality gene models for both. Comparative analysis revealed remarkable divergence among the genomes of Symbiodinium from four clades, consistent with previous single-gene phylogenetic relationships. We identified 2460 Symbiodinium gene families that are under positive selection, many of which encode functions directly relevant to the establishment and/or maintenance of symbiosis with the coral host. We also identified complete, or near-complete, sets of genes indicative of the presence of meiosis and several mechanisms of stress tolerance, functions that have until now been considered absent from S. kawagutii. Our results demonstrate the remarkable genomic capacity of Symbiodinium to synthesize key metabolites that are essential to the establishment of symbiosis with coral hosts, and to respond to environmental stress.

S. goreaui (type C1) belongs to one of the most globally dominant clades (Clade C) on coral reefs, and analysis of its draft genome has provided important new insights into coral-algal symbiosis. This genomic resource is already demonstrating utility in the identification of symbiont fractions in *de novo* sequencing of coral tissues[92,93], and will provide a foundation for targeted studies into the molecular biology, physiology and of this crucial symbiosis and its adaptation to a changing environment.

Methods

Biological materials and sequencing data

Symbiodinium goreaui (Clade C, type C1; AIMS-aten-C1-MI-cfu-B2, now AIMS culture collection SCF055-01) is a single-cell monoclonal culture first isolated from the coral *Acropora tenuis* at Magnetic Island (Queensland, Australia) at 3 m depth[94]. *Symbiodinium kawagutii* CCMP2468 (Clade F) was maintained as a monoclonal culture. Genomic DNA was extracted from these isolates using the Qiagen DNeasy Plant Mini Kit. We generated a total of 116.0 Gb of sequence data (2 paired-end libraries of 230- and 500-bp inserts, plus 3 mate-pair libraries of 3-, 6- and 9-Kbp inserts) for *S. goreaui*, and a total of 92.2 Gb of sequence data (1 paired-end library of 230-bp inserts, plus 3 mate-pair libraries of 4-, 6- and 9-Kbp inserts) for *S. kawagutii*, in each case using the Illumina HiSeq 2500 Rapid Chemistry platform. See Supplementary Methods and Supplementary Table S1 for details.

Genome assembly and annotation

We adopted a combined approach to *de novo* genome assembly (Supplementary Methods) in which multiple assembly programs (CLC Genomics Workbench (Qiagen), SPAdes[95] and ALLPATHS-LG[96]) were first used independently. The quality of each assembly was assessed based on full-length recovery of phylogenetic markers and known coding sequences. Once the best assembly (the master assembly) was identified, other assemblies were used to refine it via merging scaffolds and filling gaps. We adopted a comprehensive *ab initio* approach for gene prediction using all available dinoflagellate proteins, as well as all *Symbiodinium* genes and transcriptomes, as guiding evidence. Our approach combines evidence-based methods i.e. PASA (with TransDecoder)[97], AUGUSTUS[98] and MAKER[99], and unsupervised machine-learning (GeneMark-ES[100] and SNAP[101]; alternative splice sites were specified in these methods by modifying the relevant scripts. We then used EvidenceModeler[102] to combine

multiple sets of predicted genes, allocating a heavier weight for evidence-based predictions than for those produced by unsupervised approaches; for details see Supplementary Methods. Final genome assemblies, predicted gene models and proteins are available at https://cloudstor.aarnet.edu.au/plus/index.php/s/6yziMf2ygWjGu0L.

We adopted multiple approaches to assess genome completeness. Established methods including CEGMA[103] and BUSCO[104] are based on conserved genes in a limited number of eukaryote model organisms that are distantly related to dinoflagellates. The use of these methods resulted in relatively low recovery of conserved eukaryote genes in *Symbiodinium* (e.g. 33-42% of BUSCO genes; Supplementary Figure S7) when run at default setting. We further assessed completeness using BLAST based on predicted proteins from the gene models and the assembled genome scaffolds. For each genome, we followed Baumgarten et al.[105] and searched (BLASTP, $E \leq 10^{-5}$) against the predicted proteins using the 458 CEGMA proteins[103]. We also searched against the CEGMA proteins using the genome scaffolds (BLASTX $E \leq 10^{-5}$), against genome scaffolds using the 458 CEGMA proteins (TBLASTN, $E \leq 10^{-5}$), and against genome scaffolds using the 458 CEGMA transcripts (TBLASTX, $E \leq 10^{-5}$) (Supplementary Table S10, Supplementary Figure S7).

Identification and removal of bacterial and viral sequences

Bacterial and viral sequences were identified and removed following Aranda et al.[17]. Briefly, we used our genome scaffolds (BLASTN, $E \leq 10^{-20}$) to query the complete and draft bacterial genomes in NCBI, and the viral genomes in NCBI and PhAnToMe (http://phantome.org). We applied more-stringent criteria than did Aranda et al.[17] to identify putative contaminating sequences, removing from the final assembly any scaffold sequence that contains regions that match a bacterial or viral genome (BLASTN bit score > 1000, $E \leq 10^{-20}$) where such regions

425 constitute >10% of the overall scaffold length (Supplementary Methods, Supplementary
426 Figure S9).

427 ***Analysis of genome synteny and collinearity***

428 Using all predicted genes and their associated genomic positions, we assessed the number of
429 syntenic collinear blocks (i.e. regions with the same genes coded in the same order, free from
430 rearrangement or loss) shared pairwise among genomes of *S. microadriaticum* (Clade A)[17], *S.*
431 *minutum* (B)[19], *S. goreaui* (C) and *S. kawagutii* (F). We used BLASTP ($E \leq 10^{-5}$) to search
432 for similar proteins within each genome, and among all of them. Next we used MCScanX[106]
433 with parameter -s 5 to sort the BLASTP matches (alignments) based on genomic positions; to
434 minimise the number of collinear gene pairs arising from tandem repeats, we report only
435 collinear blocks that consist of five or more genes.

436 ***Functional annotation of gene models***

437 We adopted a similar approach to Aranda et al.[17] to annotate gene models. Protein sequences
438 predicted using the standard genetic code were used as query (BLASTP, $E \leq 10^{-5}$) first
439 against Swiss-Prot, and those with no Swiss-Prot hits subsequently against TrEMBL (both
440 databases from UniProt release 2016_10). Gene Ontology (http://geneontology.org/) terms
441 associated with Swiss-Prot and TrEMBL hits were obtained using the UniProt-GOA mapping
442 (release 2016_10).

443 ***Identification of homologous protein groups and gene families***

444 Our workflow for delineation of sets of putatively homologous proteins, multiple sequence
445 alignment, generation of protein-family and reference trees, and analysis of selection is
446 shown in Supplementary Figure S8. Using predicted proteins from 31 phyletically diverse
447 organisms including *Symbiodinium* for which genome and/or transcriptome data are available
448 (Supplementary Table S25), we identified sets of putatively homologous proteins using

449 OrthoFinder[107] and consider the corresponding gene sets (families) to be homologs.
450 Following Harlow et al.[108] and Beiko et al.[109], we consider sequences within single-copy sets
451 (i.e. those in which each genome is represented no more than once) to be orthologs. We refer
452 to sets that contain proteins only from *Symbiodinium*, plus the *Symbiodinium* singletons, as
453 *Symbiodinium*-specific. For enrichment analysis of annotated features (GO terms or Pfam
454 domains), we compared the features within the *Symbiodinium*-specific set against those in
455 each background set (i.e. the 31-taxon set and, separately, the 15-taxon set) using a
456 hypergeometric test; features with an adjusted[110] p-value < 0.05 were considered significant.

457 ***Analysis of positive selection in Symbiodinium genes***

458 For the 15-taxon set we sorted the 311,651 protein sets into 1,654 single-copy (ortholog) and
459 16,836 multi-copy sets. Multiple sequence alignments were carried out using MAFFT v7.245
460 at -linsi mode[111]; questionably aligned columns and rows were removed from these
461 alignments using trimAl[112] with the -automated1 option.

462 Branch-site models (BSMs; see below) require a reference topology. We follow Price
463 and Bhattacharya[48] to generate the reference species tree. The trimmed single-copy protein
464 alignments were concatenated prior to maximum-likelihood (ML) inference of the species
465 phylogeny using IQTREE[113]; each alignment represents a partition for which the best
466 evolutionary model was determined independently. Support for each node was assessed using
467 2000 rapid bootstraps. The species tree so generated (Figure 2a) is similar to that of Price and
468 Bhattacharya[48], with very strong support (≥ 96% bootstrap replicates) for all internal nodes;
469 the *Symbiodinium* and Suessiales (*Symbiodinium* + *Polarella glacialis*) clades received 100%
470 bootstrap support.

471 Of all trimmed protein alignments, those with ≥ 60 aligned positions and ≥ 4 sequences
472 were used in subsequent analysis. For multi-copy protein sets, we imposed further filtering

criteria. We first inferred individual ML trees for the multi-copy sets using IQ-TREE, and each resulting protein tree was compared with the reference species tree. Those congruent with the reference species tree at genus level, and in which all *Symbiodinium* are resolved as an exclusive monophyletic clade, were judged paralog-free and used in subsequent BSM analysis (Supplementary Figure S8). Among the 16,836 multi-copy sets of the 15-taxon analysis, 1788 (10.6%) resolve all *Symbiodinium* sequences into an exclusive monophyletic clade and are topologically congruent at genus level with the reference species tree (i.e. contain co-orthologs but not paralogs) and were retained, while the remaining 15,048 failed one or both of these filtering criteria (i.e. contain presumed paralogs) and were not analysed further (Supplementary Figure S8). The percentages of missing data and parsimoniously informative sites in all filtered protein alignments for the 15-taxon set are detailed in Supplementary Table S26. For each filtered alignment, we used the corresponding coding-sequence alignment (codon alignment) generated using PAL2NAL[114] in the BSM analysis. Some predicted protein sequences in MMETSP[115] do not match their corresponding CDS, sometimes due to problematic translation and other times due to a frameshift. For these, we used MACSE[116] to derive the codon alignments.

We applied the branch-site model (BSM) implemented in the *codeml* program in PAML 4.9[117] to detect positive selection signal unique to the *Symbiodinium* lineage. BSMs allow the dN/dS ratio ω to vary among both sites and branches, making it possible to infer selection at both. We computed two models: a null model with fixed ω = 1, and an alternative model that estimates ω in our defined foreground branches (here, the node that leads to all *Symbiodinium* lineages). We then compared the likelihoods of these two models to determine the better fit. To reduce false positives we applied *q*-value estimation for false discovery rate control, as implemented in R package *qvalue* to adjust *p*-values. Instances with an adjusted *p*

497 ≤ 0.05 were considered significant. See Supplementary Note and Supplementary Figure S10

498 for analysis of gene gain and gene loss in *Symbiodinium*.

499 ## References

500 1. Plaisance, L., Caley, M.J., Brainard, R.E. & Knowlton, N. The diversity of coral reefs:
501 what are we missing? *PLoS ONE* **6**, e25026 (2011).

502 2. Falkowski, P.G., Dubinsky, Z., Muscatine, L. & Porter, J.W. Light and the
503 bioenergetics of a symbiotic coral. *Bioscience* **34**, 705-709 (1984).

504 3. Muscatine, L. & Porter, J.W. Reef corals: mutualistic symbioses adapted to nutrient-
505 poor environments. *Bioscience* **27**, 454-460 (1977).

506 4. Glynn, P.W. Coral reef bleaching: facts, hypotheses and implications. *Global Change*
507 *Biol.* **2**, 495-509 (1996).

508 5. Hoegh-Guldberg, O. Climate change, coral bleaching and the future of the world's coral
509 reefs. *Mar. Freshw. Res.* **50**, 839-866 (1999).

510 6. Albright, R. et al. Ocean acidification: linking science to management solutions using
511 the Great Barrier Reef as a case study. *J. Environ. Manage.* **182**, 641-650 (2016).

512 7. Hughes, T.P. et al. Global warming and recurrent mass bleaching of corals. *Nature* **543**,
513 373-377 (2017).

514 8. Lin, S. Genomic understanding of dinoflagellates. *Res. Microbiol.* **162**, 551-569 (2011).

515 9. LaJeunesse, T.C., Lambert, G., Andersen, R.A., Coffroth, M.A. & Galbraith, D.W.
516 *Symbiodinium* (Pyrrhophyta) genome sizes (DNA content) are smallest among
517 dinoflagellates. *J. Phycol.* **41**, 880-886 (2005).

518 10. Barbrook, A.C., Voolstra, C.R. & Howe, C.J. The chloroplast genome of a
519 *Symbiodinium* sp. Clade C3 isolate. *Protist* **165**, 1-13 (2014).

520 11. Howe, C.J., Nisbet, R.E.R. & Barbrook, A.C. The remarkable chloroplast genome of
521 dinoflagellates. *J. Exp. Bot.* **59**, 1035-1045 (2008).

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

524 13. Nash, E.A., Nisbet, R.E.R., Barbrook, A.C. & Howe, C.J. Dinoflagellates: a
525 mitochondrial genome all at sea. *Trends Genet.* **24**, 328-335 (2008).

526 14. Waller, R.F. & Jackson, C.J. Dinoflagellate mitochondrial genomes: stretching the rules
527 of molecular biology. *BioEssays* **31**, 237-245 (2009).

528 15. Jackson, C.J. et al. Broad genomic and transcriptional analysis reveals a highly derived
529 genome in dinoflagellate mitochondria. *BMC Biol.* **5**, 41 (2007).

16. Liew, Y.J., Li, Y., Baumgarten, S., Voolstra, C.R. & Aranda, M. Condition-specific RNA editing in the coral symbiont *Symbiodinium microadriaticum*. *PLoS Genet.* **13**, e1006619 (2017).

17. Aranda, M. et al. Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle. *Sci. Rep.* **6**, 39734 (2016).

18. Lin, S. et al. The *Symbiodinium kawagutii* genome illuminates dinoflagellate gene expression and coral symbiosis. *Science* **350**, 691-694 (2015).

19. Shoguchi, E. et al. Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. *Curr. Biol.* **23**, 1399-1408 (2013).

20. LaJeunesse, T.C. et al. Low symbiont diversity in southern Great Barrier Reef corals, relative to those of the Caribbean. *Limnol. Oceanogr.* **48**, 2046-2054 (2003).

21. LaJeunesse, T.C. "Species" radiations of symbiotic dinoflagellates in the Atlantic and Indo-Pacific since the Miocene-Pliocene transition. *Mol. Biol. Evol.* **22**, 570-581 (2005).

22. Thornhill, D.J., Lewis, A.M., Wham, D.C. & LaJeunesse, T.C. Host-specialist lineages dominate the adaptive radiation of reef coral endosymbionts. *Evolution* **68**, 352-367 (2014).

23. Tonk, L., Bongaerts, P., Sampayo, E.M. & Hoegh-Guldberg, O. *SymbioGBR*: a web-based database of *Symbiodinium* associated with cnidarian hosts on the Great Barrier Reef. *BMC Ecol.* **13**, 7 (2013).

24. Bongaerts, P. et al. Prevalent endosymbiont zonation shapes the depth distributions of scleractinian coral species. *R. Soc. Open Sci.* **2**, 140297 (2015).

25. Chan, Y.L. et al. Generalist dinoflagellate endosymbionts and host genotype diversity detected from mesophotic (67-100 m depths) coral *Leptoseris*. *BMC Ecol.* **9**, 21 (2009).

26. Yuyama, I., Higuchi, T. & Mezaki, T. *Symbiodinium kawagutii* (clade F) coats the surface of *Aropora solitaryensis*, resulting in the formation of a sheet-like crust. *Proc. 13th Intl Coral Reef Symp.*, 49-56 (2016).

27. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111-120 (1980).

28. Vitte, C., Panaud, O. & Quesneville, H. LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. *BMC Genomics* **8**, 218 (2007).

29. Tenaillon, M.I., Hufford, M.B., Gaut, B.S. & Ross-Ibarra, J. Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. *Genome Biol. Evol.* **3**, 219-229 (2011).

30. Lee, S.I. & Kim, N.S. Transposable elements and genome size variations in plants. *Genomics Inform* **12**, 87-97 (2014).

566 31. LaJeunesse, T.C. Investigating the biodiversity, ecology, and phylogeny of
567 endosymbiotic dinoflagellates in the genus *Symbiodinium* using the ITS region: in
568 search of a "species" level marker. *J. Phycol.* **37**, 866-880 (2001).

569 32. Pochon, X. & Gates, R.D. A new *Symbiodinium* clade (Dinophyceae) from soritid
570 foraminifera in Hawai'i. *Mol. Phylogenet. Evol.* **56**, 492-497 (2010).

571 33. van Oppen, M.J.H., Mieog, J.C., Sánchez, C.A. & Fabricius, K.E. Diversity of algal
572 endosymbionts (zooxanthellae) in octocorals: the roles of geography and host
573 relationships. *Mol. Ecol.* **14**, 2403-2417 (2005).

574 34. Mosavi, L.K., Cammett, T.J., Desrosiers, D.C. & Peng, Z.Y. The ankyrin repeat as
575 molecular architecture for protein recognition. *Protein Sci.* **13**, 1435-1448 (2004).

576 35. Chan, C.X., Gross, J., Yoon, H.S. & Bhattacharya, D. Plastid origin and evolution: new
577 models provide insights into old problems. *Plant Physiol.* **155**, 1552-1560 (2011).

578 36. Howe, C.J., Barbrook, A.C., Nisbet, R.E.R., Lockhart, P.J. & Larkum, A.W.D. The
579 origin of plastids. *Phil. Trans. R. Soc. B* **363**, 2675-2685 (2008).

580 37. Yoon, H.S. et al. Tertiary endosymbiosis driven genome evolution in dinoflagellate
581 algae. *Mol. Biol. Evol.* **22**, 1299-1308 (2005).

582 38. Rae, P.M. Hydroxymethyluracil in eukaryote DNA: a natural feature of the Pyrrophyta
583 (dinoflagellates). *Science* **194**, 1062-1064 (1976).

584 39. Blank, R.J., Huss, V.R. & Kersten, W. Base composition of DNA from symbiotic
585 dinoflagellates: a tool for phylogenetic classification. *Arch. Microbiol.* **149**, 515-520
586 (1988).

587 40. ten Lohuis, M.R. & Miller, D.J. Hypermethylation at CpG-motifs in the dinoflagellates
588 *Amphidinium carterae* (Dinophyceae) and *Symbiodinium microadriaticum*
589 (Dinophyceae): evidence from restriction analyses, 5-azacytidine and ethionine
590 treatment. *J. Phycol.* **34**, 152-159 (1998).

591 41. Maumus, F. et al. Potential impact of stress activated retrotransposons on genome
592 evolution in a marine diatom. *BMC Genomics* **10**, 624 (2009).

593 42. Ito, H. et al. A stress-activated transposon in *Arabidopsis* induces transgenerational
594 abscisic acid insensitivity. *Sci. Rep.* **6**, 23181 (2016).

595 43. Ramallo, E., Kalendar, R., Schulman, A.H. & Martínez-Izquierdo, J.A. *Reme1*, a *Copia*
596 retrotransposon in melon, is transcriptionally induced by UV light. *Plant Mol. Biol.* **66**,
597 137-150 (2008).

598 44. Faure, E., Best-Belpomme, M. & Champion, S. X-irradiation activates the *Drosophila*
599 1731 retrotransposon LTR and stimulates secretion of an extracellular factor that
600 induces the 1731-LTR transcription in nonirradiated cells. *J Biochem* **120**, 313-319
601 (1996).

602 45. Hunter, R.G., Gagnidze, K., McEwen, B.S. & Pfaff, D.W. Stress and the dynamic
603 genome: steroids, epigenetics, and the transposome. *Proc. Natl. Acad. Sci. U. S. A.* **112**,
604 6828-6833 (2015).

605 46. Weynberg, K.D. et al. Prevalent and persistent viral infection in cultures of the coral
606 algal endosymbiont *Symbiodinium*. *Coral Reefs* **36**, 773-784 (2017).

607 47. Kimura, M. Preponderance of synonymous changes as evidence for the neutral theory
608 of molecular evolution. *Nature* **267**, 275-276 (1977).

609 48. Price, D.C. & Bhattacharya, D. Robust Dinoflagellata phylogeny inferred from public
610 transcriptome databases. *J. Phycol.* **53**, 725-729 (2017).

611 49. Takahashi, S., Yoshioka-Nishimura, M., Nanba, D. & Badger, M.R. Thermal
612 acclimation of the symbiotic alga *Symbiodinium* spp. alleviates photobleaching under
613 heat stress. *Plant Physiol.* **161**, 477-485 (2013).

614 50. Lesser, M.P. Elevated temperatures and ultraviolet radiation cause oxidative stress and
615 inhibit photosynthesis in symbiotic dinoflagellates. *Limnol. Oceanogr.* **41**, 271-283
616 (1996).

617 51. Lesser, M.P. Oxidative stress causes coral bleaching during exposure to elevated
618 temperatures. *Coral Reefs* **16**, 187-192 (1997).

619 52. Barott, K.L., Venn, A.A., Perez, S.O., Tambutte, S. & Tresguerres, M. Coral host cells
620 acidify symbiotic algal microenvironment to promote photosynthesis. *Proc. Natl. Acad.*
621 *Sci. U. S. A.* **112**, 607-612 (2015).

622 53. Zoccola, D. et al. Bicarbonate transporters in corals point towards a key step in the
623 evolution of cnidarian calcification. *Sci. Rep.* **5**, 9983 (2015).

624 54. Shinzato, C. et al. Using the *Acropora digitifera* genome to understand coral responses
625 to environmental change. *Nature* **476**, 320-323 (2011).

626 55. Davy, S.K., Allemand, D. & Weis, V.M. Cell biology of cnidarian-dinoflagellate
627 symbiosis. *Microbiol. Mol. Biol. Rev.* **76**, 229-261 (2012).

628 56. Lin, K.L., Wang, J.T. & Fang, L.S. Participation of glycoproteins of zooxanthellal cell
629 walls in the establishment of a symbiotic relationship with the sea anemone, *Aiptasia*
630 *pulchella*. *Zool. Stud.* **39**, 172-178 (2000).

631 57. Markell, D.A., Trench, R.K. & Iglesias-Prieto, R. Macromolecules associated with the
632 cell walls of symbiotic dinoflagellates. *Symbiosis* **12**, 19-31 (1992).

633 58. Markell, D.A. & Trench, R.K. Macromolecules exuded by symbiotic dinoflagellates in
634 culture: amino acid and sugar composition. *J. Phycol.* **29**, 64-68 (1993).

635 59. Markell, D.A. & Wood-Charlson, E.M. Immunocytochemical evidence that symbiotic
636 algae secrete potential recognition signal molecules *in hospite*. *Mar. Biol.* **157**, 1105-
637 1111 (2010).

60. Logan, D.D.K., LaFlamme, A.C., Weis, V.M. & Davy, S.K. Flow-cytometric characterization of the cell-surface glycans of symbiotic dinoflagellates (*Symbiodinium* spp.). *J. Phycol.* **46**, 525-533 (2010).

61. Wood-Charlson, E.M., Hollingsworth, L.L., Krupp, D.A. & Weis, V.M. Lectin/glycan interactions play a role in recognition in a coral/dinoflagellate symbiosis. *Cell. Microbiol.* **8**, 1985-1993 (2006).

62. Neubauer, E.F. et al. A diverse host thrombospondin-type-1 repeat protein repertoire promotes symbiont colonization during establishment of cnidarian-dinoflagellate symbiosis. *eLife* **6**, e24494 (2017).

63. Ljungman, M. The transcription stress response. *Cell Cycle* **6**, 2252-2257 (2007).

64. Mohamed, A.R. et al. The transcriptomic response of the coral *Acropora digitifera* to a competent *Symbiodinium* strain: the symbiosome as an arrested early phagosome. *Mol. Ecol.* **25**, 3127-3141 (2016).

65. Fitt, W.K. & Trench, R.K. Endocytosis of the symbiotic dinoflagellate *Symbiodinium microadriaticum* Freudenthal by endodermal cells of the scyphistomae of *Cassiopeia xamachana* and resistance of the algae to host digestion. *J. Cell Sci.* **64**, 195-212 (1983).

66. Freudenthal, H.D. *Symbiodinium* gen. nov. and *Symbiodinium microadriaticum* sp. nov., a zooxanthella: taxonomy, life cycle, and morphology. *J. Protozool.* **9**, 45-52 (1962).

67. Taylor, F.J.R. Implications and extensions of the serial endosymbiosis theory of the origin of eukaryotes. *Taxon* **23**, 229-258 (1974).

68. Trench, R.K. Diversity of symbiotic dinoflagellates and the evolution of microalgal-invertebrate symbioses. *Proc. 8th Intl Coral Reef Symp.* **2**, 1275-1286 (1997).

69. de Visser, J.A. & Elena, S.F. The evolution of sex: empirical insights into the roles of epistasis and drift. *Nat. Rev. Genet.* **8**, 139-149 (2007).

70. Baillie, B.K., Belda-Baillie, C.A. & Maruyama, T. Conspecificity and Indo-Pacific distribution of *Symbiodinium* genotypes (Dinophyceae) from giant clams. *J. Phycol.* **36**, 1153-1161 (2000).

71. Pettay, D.T., Wham, D.C., Pinzón, J.H. & LaJeunesse, T.C. Genotypic diversity and spatial-temporal distribution of *Symbiodinium* clones in an abundant reef coral. *Mol. Ecol.* **20**, 5197-5212 (2011).

72. Reichman, J.R., Wilcox, T.P. & Vize, P.D. PCP gene family in *Symbiodinium* from *Hippopus hippopus*: low levels of concerted evolution, isoform diversity, and spectral tuning of chromophores. *Mol. Biol. Evol.* **20**, 2143-2154 (2003).

73. Santos, S.R. & Coffroth, M.A. Molecular genetic evidence that dinoflagellates belonging to the genus *Symbiodinium* Freudenthal are haploid. *Biol. Bull.* **204**, 10-20 (2003).

676 74. Levin, R.A. et al. Sex, scavengers, and chaperones: transcriptome secrets of divergent
677 *Symbiodinium* thermal tolerances. *Mol. Biol. Evol.* **33**, 2201-2215 (2016).

678 75. Schurko, A.M. & Logsdon, J.M., Jr. Using a meiosis detection toolkit to investigate
679 ancient asexual "scandals" and the evolution of sex. *BioEssays* **30**, 579-589 (2008).

680 76. Chi, J., Parrow, M.W. & Dunthorn, M. Cryptic sex in *Symbiodinium* (Alveolata,
681 Dinoflagellata) is supported by an inventory of meiotic genes. *J. Eukaryot. Microbiol.*
682 **61**, 322-327 (2014).

683 77. Malik, S.B., Pightling, A.W., Stefaniak, L.M., Schurko, A.M. & Logsdon, J.M., Jr. An
684 expanded inventory of conserved meiotic genes provides evidence for sex in
685 *Trichomonas vaginalis*. *PLoS ONE* **3**, e2879 (2008).

686 78. Speijer, D., Lukeš, J. & Eliáš, M. Sex is a ubiquitous, ancient, and inherent attribute of
687 eukaryotic life. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 8827-8834 (2015).

688 79. Griffiths, D.J., Barbet, N.C., McCready, S., Lehmann, A.R. & Carr, A.M. Fission yeast
689 *rad17*: a homologue of budding yeast *RAD24* that shares regions of sequence similarity
690 with DNA polymerase accessory proteins. *EMBO J.* **14**, 5812-5823 (1995).

691 80. Banaszak, A.T., LaJeunesse, T.C. & Trench, R.K. The synthesis of mycosporine-like
692 amino acids (MAAs) by cultured, symbiotic dinoflagellates. *J. Exp. Mar. Biol. Ecol.*
693 **249**, 219-233 (2000).

694 81. Rosic, N.N. & Dove, S. Mycosporine-like amino acids from coral dinoflagellates. *Appl.*
695 *Environ. Microbiol.* **77**, 8478-8486 (2011).

696 82. Balskus, E.P. & Walsh, C.T. The genetic and molecular basis for sunscreen
697 biosynthesis in cyanobacteria. *Science* **329**, 1653-1656 (2010).

698 83. Gao, Q. & Garcia-Pichel, F. Microbial ultraviolet sunscreens. *Nat. Rev. Microbiol.* **9**,
699 791-802 (2011).

700 84. D'Agostino, P.M. et al. Comparative profiling and discovery of novel glycosylated
701 mycosporine-like amino acids in two strains of the cyanobacterium *Scytonema* cf.
702 *crispum*. *Appl. Environ. Microbiol.* **82**, 5951-5959 (2016).

703 85. Rastogi, R.P., Sonani, R.R. & Madamwar, D. Cyanobacterial sunscreen scytonemin:
704 role in photoprotection and biomedical research. *Appl. Biochem. Biotechnol.* **176**, 1551-
705 1563 (2015).

706 86. Garcia-Pichel, F. & Castenholz, R.W. Characterization and biological implications of
707 scytonemin, a cyanobacterial sheath pigment. *J. Phycol.* **27**, 395-409 (1991).

708 87. Balskus, E.P. & Walsh, C.T. Investigating the initial steps in the biosynthesis of
709 cyanobacterial sunscreen scytonemin. *J. Am. Chem. Soc.* **130**, 15260-15261 (2008).

710 88. Soule, T., Stout, V., Swingley, W.D., Meeks, J.C. & Garcia-Pichel, F. Molecular
711 genetics and genomic analysis of scytonemin biosynthesis in *Nostoc punctiforme*
712 ATCC 29133. *J. Bacteriol.* **189**, 4465-4472 (2007).

89. Soule, T., Garcia-Pichel, F. & Stout, V. Gene expression patterns associated with the biosynthesis of the sunscreen scytonemin in *Nostoc punctiforme* ATCC 29133 in response to UVA radiation. *J. Bacteriol.* **191**, 4639-4646 (2009).

90. Shinzato, C., Mungpakdee, S., Satoh, N. & Shoguchi, E. A genomic approach to coral-dinoflagellate symbiosis: studies of *Acropora digitifera* and *Symbiodinium minutum*. *Front. Microbiol.* **5**, 336 (2014).

91. Frankowiak, K. et al. Photosymbiosis and the expansion of shallow-water corals. *Sci. Adv.* **2**, e1601122 (2016).

92. Bongaerts, P. et al. Deep reefs are not universal refuges: reseeding potential varies among coral species. *Sci. Adv.* **3**, e1602373 (2017).

93. Thomas, L., Kennington, W.J., Evans, R.D., Kendrick, G.A. & Stat, M. Restricted gene flow and local adaptation highlight the vulnerability of high-latitude reefs to rapid environmental change. *Global Change Biol.* **23**, 2197-2205 (2017).

94. Howells, E.J. et al. Coral thermal tolerance shaped by local adaptation of photosymbionts. *Nat. Clim. Change* **2**, 116-120 (2012).

95. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455-477 (2012).

96. Gnerre, S. et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 1513-1518 (2011).

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

98. Stanke, M. et al. AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435-439 (2006).

99. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).

100. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494-6506 (2005).

101. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).

102. Haas, B.J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).

103. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067 (2007).

748 104. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M.
749 BUSCO: assessing genome assembly and annotation completeness with single-copy
750 orthologs. *Bioinformatics* **31**, 3210-3212 (2015).

751 105. Baumgarten, S. et al. The genome of *Aiptasia*, a sea anemone model for coral
752 symbiosis. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11893-11898 (2015).

753 106. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene
754 synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).

755 107. Emms, D.M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome
756 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**,
757 157 (2015).

758 108. Harlow, T.J., Gogarten, J.P. & Ragan, M.A. A hybrid clustering approach to
759 recognition of protein families in 114 microbial genomes. *BMC Bioinformatics* **5**, 45
760 (2004).

761 109. Beiko, R.G., Harlow, T.J. & Ragan, M.A. Highways of gene sharing in prokaryotes.
762 *Proc. Natl. Acad. Sci. U. S. A.* **102**, 14332-14337 (2005).

763 110. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing
764 under dependency. *The Annals of Statistics* **29**, 1165-1188 (2001).

765 111. Katoh, K., Kuma, K.-I., Toh, H. & Miyata, T. MAFFT version 5: improvement in
766 accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511-518 (2005).

767 112. Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. trimAl: a tool for automated
768 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973
769 (2009).

770 113. Nguyen, L.T., Schmidt, H.A., von Haeseler, A. & Minh, B.Q. IQ-TREE: a fast and
771 effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol.*
772 *Biol. Evol.* **32**, 268-274 (2015).

773 114. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence
774 alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609-612
775 (2006).

776 115. Keeling, P.J. et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project
777 (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans
778 through transcriptome sequencing. *PLoS Biol.* **12**, e1001889 (2014).

779 116. Ranwez, V., Harispe, S., Delsuc, F. & Douzery, E.J. MACSE: Multiple Alignment of
780 Coding SEquences accounting for frameshifts and stop codons. *PLoS ONE* **6**, e22594
781 (2011).

782 117. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**,
783 1586-1591 (2007).

784

Acknowledgements

We thank Manuel Aranda for advance access to the *S. microadriaticum* genome data, and Todd LaJeunesse for information on the original isolation of *S. kawagutii*. This project was supported by the Reef Future Genomics (ReFuGe) 2020 International Consortium and the Great Barrier Reef Foundation. The data reported in this work were supported by funding from Bioplatforms Australia through the Australian Government National Collaborative Research Infrastructure Strategy (NCRIS). In memory of SF, our friend and colleague who is sorely missed.

Author contributions

HL, MAR and CXC conceived the study and designed the experiments. HL, TGS, RGP and CXC conducted all computational analyses. VHB and BL established the algal cultures and extracted the DNA. HL, TGS, RGP, IC, MAR and CXC analysed and interpreted the results. HL and CXC prepared all figures, tables, and the first draft of this manuscript. SF and CRV provided analytical tools and scripts. HL, TGS, MAR and CXC wrote the manuscript. PB, IC, DGB, DJM, MJHvO and CRV assisted in experimental design and writing of the manuscript. All authors reviewed, commented on and approved the final manuscript.

Data availability

All sequence data generated from this study will be available at the NCBI Short Read Archive (accessions XXXX and XXXX). Assembled genomes, predicted gene models and proteins are available at https://cloudstor.aarnet.edu.au/plus/index.php/s/6yziMf2ygWjGu0L.

Code availability

The customised scripts for AUGUSTUS and PASA used in this study were previously published in Aranda et al.[17]; they are available at http://smic.reefgenomics.org/download/.

Competing financial interests

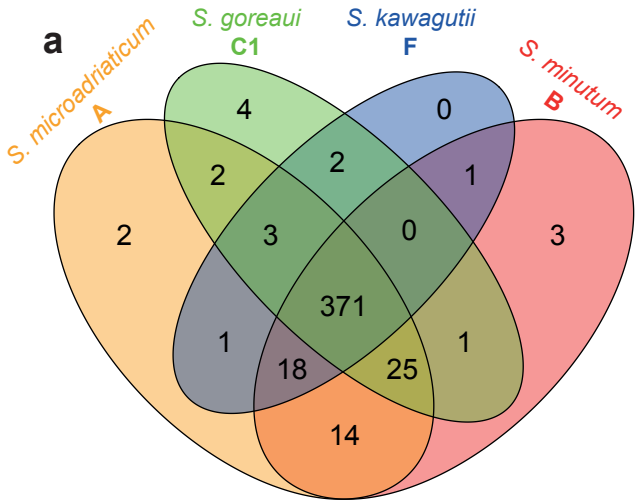
The authors declare no competing financial interests.

Figure legends

Figure 1. Comparison of *Symbiodinium* genomes. (**a**) Number of recovered core eukaryote genes in each genome based on CEGMA. (**b**) Number of identified syntenic collinear blocks (and the corresponding number of implicated genes) between each genome pair.

Figure 2. Testing for positive selection acting on *Symbiodinium* genomes. (**a**) The reference 15-species supertree of *Symbiodinium*, dinoflagellates and *Perkinsus marinus* (as outgroup) based on single-copy orthologous genes. Support based on 2000 rapid bootstraps is shown on each internal node, and the branch length is the number of substitutions per site. (**b**) Percentage of positively selected genes with annotated GO (level 3) terms in *Symbiodinium*, shown for principal hierarchies Biological Function, Molecular Function and Cellular Component.

Figure 3. Recovery of genes in *Symbiodinium*. (**a**) Meiosis-related genes recovered in the genomes of *S. microadriaticum* (Clade A), *S. minutum* (Clade B), *S. goreaui* (Clade C) and *S. kawagutii* (Clade F). (**b**) Scytonemin biosynthesis genes in *Symbiodinium* genomes relative to the coral *Acropora digitifera*, sea anemone *Nematostella vectensis*, hydra (*Hydra magnipapillata*) and the green plant *Arabidopsis thaliana*. The order of the 18-gene cluster in the cyanobacteria *Nostoc punctiforme* is used as a reference, with presence (+) and absence (-) of a gene in each species are shown. Figure modified from Shinzato et al.[90].



b

	S. microadriaticum clade **A**	*S. minutum* clade **B**	*S. goreaui* clade **C**
S. minutum clade **B**	**370** (2816)		
S. goreaui clade **C**	**86** (588)	**155** (1125)	
S. kawagutii clade **F**	**121** (893)	**173** (1323)	**889** (8621)

Figure 1. Comparison of *Symbiodinium* genomes. (**a**) Number of recovered core eukaryote genes in each genome based on CEGMA. (**b**) Number of identified syntenic collinear blocks (and the corresponding number of implicated genes) between each genome pair.

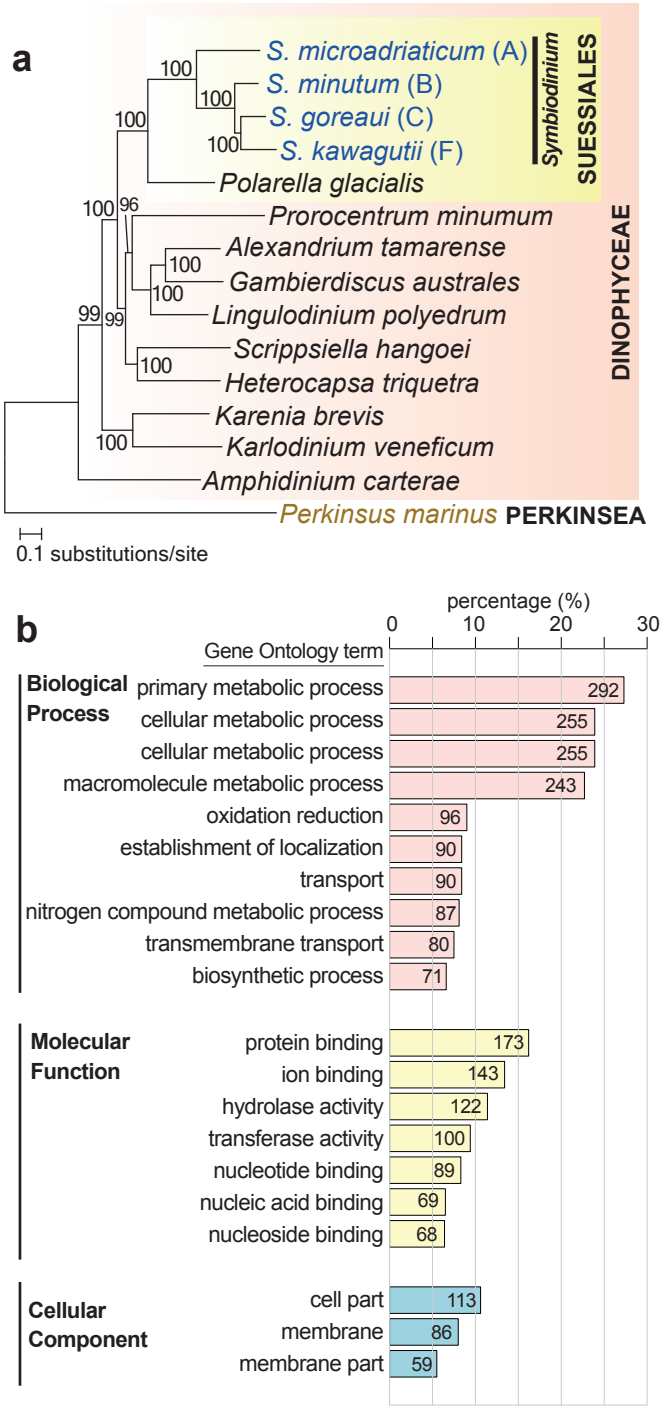


Figure 2. Testing for positive selection acting on *Symbiodinium* genomes. (**a**) The reference 15-species supertree of *Symbiodinium*, dinoflagellates and *Perkinsus marinus* (as outgroup) based on single-copy orthologous genes. Support based on 2000 rapid bootstraps is shown on each internal node, and the branch length is the number of substitutions per site. (**b**) Percentage of positively selected genes with annotated GO (level 3) terms in *Symbiodinium*, shown for principal hierarchies Biological Process, Molecular Function and Cellular Component.

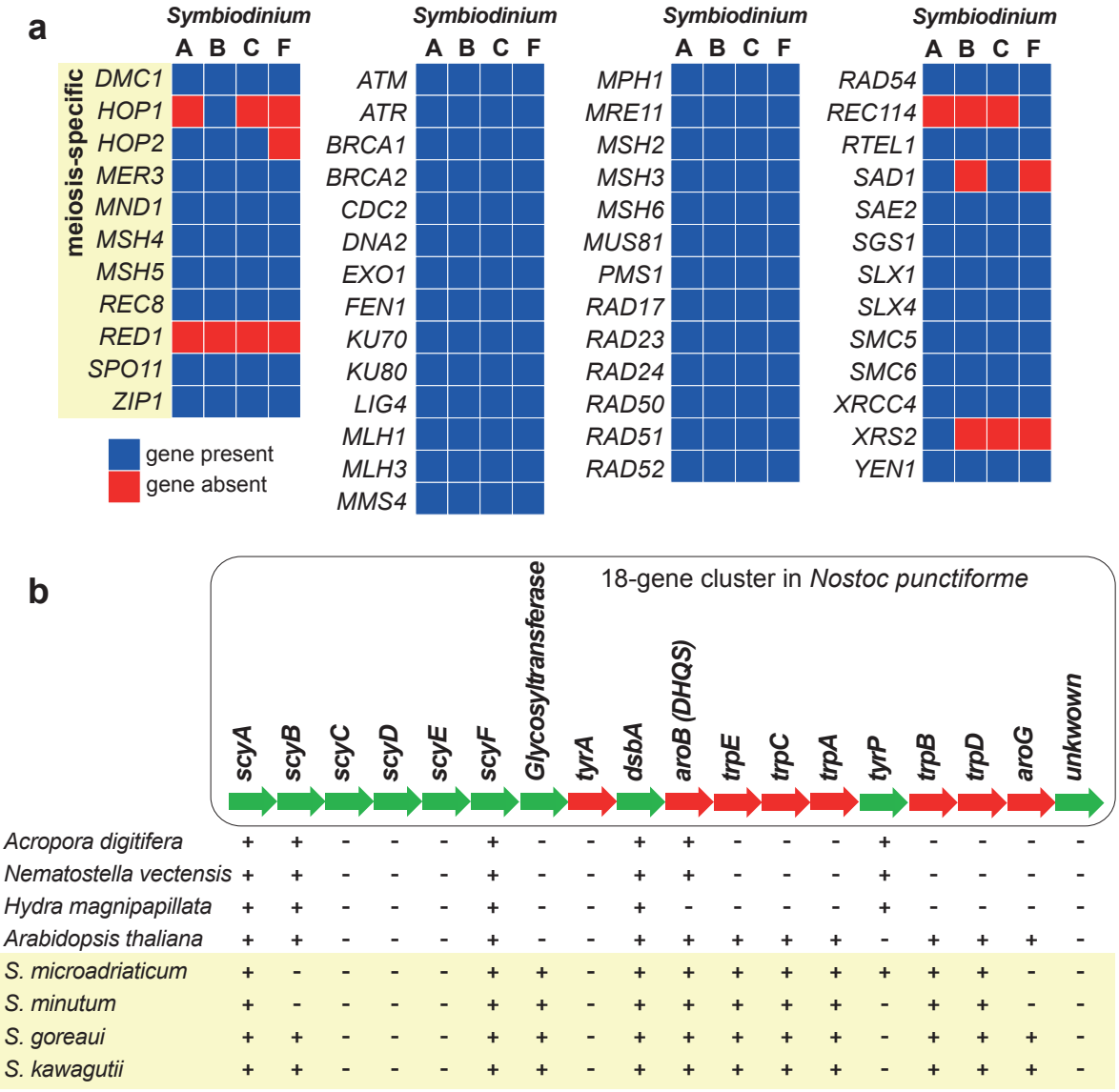


Figure 3. Recovery of genes in *Symbiodinium*. (**a**) Meiosis-related genes recovered in the genomes of *S. microadriaticum* (Clade A), *S. minutum* (Clade B), *S. goreaui* (Clade C) and *S. kawagutii* (Clade F). (**b**) Scytonemin biosynthesis genes in *Symbiodinium* genomes relative to the coral *Acropora digitifera*, sea anemone *Nematostella vectensis*, hydra (*Hydra magnipapillata*) and the green plant *Arabidopsis thaliana*. The order of the 18-gene cluster in the cyanobacteria *Nostoc punctiforme* is used as a reference, with presence (+) and absence (-) of a gene in each species are shown. Figure modified from Shinzato et al.[90].