Supplementary Information

**Symbiodinium genomes reveal positive selection of symbiosis-related genes**

Huanle Liu¹, Timothy G. Stephens¹, Raúl A. González-Pech¹, Victor H. Beltran², Bruno Lapeyre³, Pim Bongaerts⁴, Ira Cooke³, David G. Bourne²,⁶, Sylvain Forêt¹,*, David J. Miller³, Madeleine J. H. van Oppen²,⁸, Christian R. Voolstra⁹, Mark A. Ragan¹ and Cheong Xin Chan¹,¹⁰,†

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

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

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

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

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

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

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

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

⁹Red 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

¹⁰School 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)
Supplementary Methods

Cultures and DNA extractions

_Symbiodinium goreaui_ (Clade C, type C1) originally isolated from the tissue of _Acropora tenuis_ from Magnetic Island, Australia was established in single-cell monoclonal culture first as AIMS-aten-C1-MI-cfu-B2 (later renamed SCF055-01) and maintained at the Australian Institute of Marine Science, Townsville, Australia. Genomic DNA was extracted using the DNeasy Plant Mini Kit (Qiagen) following the manufacturer’s protocol. The culture of _Symbiodinium kawagutii_ CS-156 (also known as CCMP2468) was first acquired from the Australian National Algae Culture Collection (ANACC). Unique cells were first selected under the microscope and grown in 24-well plates, from which unique cells were transferred onto agar plates. Their growth was monitored under the microscope to ensure colony formation before a colony was selected for further culturing in liquid medium. Throughout the experiment, the cells were cultured in f/2 medium containing ampicillin (100 μg/mL), kanamycin (50 μg/mL) and streptomycin (50 μg/mL). PCR amplification using generic bacterial primers¹ was performed regularly to identify potential bacterial contamination. High molecular-weight genomic DNA was extracted following Shoguchi et al.².

Generation and processing of sequencing data

For each isolate, sequence data (2 × 150 bp reads) were generated using multiple paired-end and mate-pair libraries on the Illumina HiSeq2500 platform at the Australian Genome Research Facility, Melbourne. Details of insert length for each paired-end and mate-pair libraries are shown in Supplementary Table S1. Specifically, one of the paired-end libraries (of insert length 250 bp) was designed such that the read-pairs of 2 × 150 bp would overlap. In total, we generated 116.0 Gb (614.6 million reads) and 92.2 Gb (774.1 million reads) of sequence data for _S. goreaui_ (type C1) and _S. kawagutii_ (Clade F) respectively. Compared to
S. goreau, we generated fewer sequence data for S. kawagutii because some genome data of the same isolate are publicly available (see next section).

Adapter sequences were removed from the raw sequence data using Trimmomatic, and erroneous bases were corrected using Quake. For reads generated from the paired-end libraries, pairs with overlapping reads were merged into longer, single-end reads using BBMerge (http://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/); we treated the other reads as bona fide paired-end reads. All reads generated from the mate-pair libraries were processed and classified using NXtrim based on read-orientation information (as observed based on the presence of adapter sequences in Nextera mate-pair libraries) into (a) paired-end, (b) single-end and (c) bona fide mate-pair reads. Due to the high standard deviation of estimated insert-lengths among the reads in (a), we treat both (a) and (b) as single-end reads. Details of all processed reads are shown in Supplementary Table S1.

**Comparative analysis of S. kawagutii genome sequence data**

To ensure that the sequence data we generated for S. kawagutii CS-156 (=CCMP2468) are indeed from the same source as the published data, we compared the sequence reads between the two data sources by mapping our reads onto the assembled genome in Lin et al., and conversely the reads in Lin et al. against our SPAdes genome assembly (see De novo genome assembly below), using CLC Genomics Workbench. As shown in Figure 1a, about 89% of our reads mapped at high quality (MAPQ score ≥ 30) to the published genome assembly from Lin et al. In comparison, 96.0% and 87.2% of the reads respectively from our dataset and from Lin et al. mapped (MAPQ score ≥ 30) to our SPAdes genome assembly (Supplementary Figure S1B). We recovered identical sequences of the phylogenetic marker genes (18S ribosomal RNA and ITS2) from both genome datasets. To further assess assembly quality, we aligned the contigs from our preliminary genome assembly, and from the
published assembly of *S. kawagutii*, against each of the ten fosmid sequences from Lin et al.³. Our SPAdes assembly has orders of magnitude fewer gaps and mismatches than the published assembly (Supplementary Figure S2). In subsequent genome assembly (below), we combined both published sequence reads from Lin et al.³ and our processed reads as a single dataset.

**De novo genome assembly**

For each isolate we adopted a novel, integrative approach using multiple methods to assemble the genome *de novo*. First, to minimise assembly errors we systematically assessed the distances between read-pairs in all sequencing libraries. To do this, we first assembled all processed (single, paired-end and mate-pair) reads using CLC Genomics Workbench v7.5.1 (Qiagen, Aarhus, Denmark) to generate an initial assembly; at this step the insert-length information for each sequencing library was based on the estimate given by the sequencing provider. We then mapped all reads to the assembled contigs and derived a more-accurate estimate of read-pair distances (*i.e.* via size of insert-length) for each sequencing library using *CollectInsertSizeMetrics* tool in Picard (https://broadinstitute.github.io/picard/).

Second, we assembled all processed reads using the more-accurate estimate of read-pair distances above, independently using (a) CLC Genomics Workbench v7.5.1, (b) SPAdes⁷ and (c) ALLPATHS-LG⁸. For CLC and SPAdes, the contigs were further joined into longer scaffolds using mate-pair reads with SSPACE⁹; ALLPATHS-LG yielded genome scaffolds directly. Gaps within scaffolds were further filled using GapFiller¹⁰ at the default setting, thereby yielding three preliminary assemblies: the (a) CLC, (b) SPAdes and (c) ALLPATHS-LG assemblies. In addition to assembly statistics, we further assessed the quality of each assembly based on (a) full-length recovery of phylogenetic markers (18S ribosomal RNA and internal transcribed spacer region ITS2), (b) full-length recovery of coding sequences of
known organellar genes, and (c) genome completeness based on conserved core eukaryote genes using CEGMA\textsuperscript{11} and BUSCO\textsuperscript{12} (below and Supplementary Table S10). As reference we used all publicly available \textit{Symbiodinium} internal transcribed spacer (ITS) regions (both ITS1 and ITS2), mitochondrial-encoded genes and chloroplast-encoded genes in NCBI. Compared to the other assemblies, the CLC assembly yielded the highest number of complete ITS sequences and BUSCO genes both for \textit{S. goreau}\textit{i} and for \textit{S. kawagutii}; we therefore used the CLC assembly as the master assembly for each genome.

Third, we refined these master assemblies using MUMmers in GMCloser\textsuperscript{13} by filling the gaps and merging scaffolds using contigs from the SPAdes and ALLPATHS-LG assemblies, followed by another step of gap-filling using GapFiller\textsuperscript{10}. This gave us the refined master assemblies.

Finally, putative bacterial and viral sequences were removed based on BLASTN searches using the refined master assemblies as query against the NCBI RefSeq nucleotide database (release 76) based on two stringent criteria: that the sequence aligns with a bacterial or viral sequence with bit score $\geq 1000$, and the alignment has $E \leq 10^{-20}$. Following these criteria, 322 and 90 scaffolds were removed respectively from the \textit{S. goreau}\textit{i} and \textit{S. kawagutii} assemblies.

\textbf{Identification of bacterial and viral sequences in the genome scaffolds}

To identify putative bacterial and viral sequences in the genome scaffolds of \textit{S. goreau}\textit{i} and \textit{S. kawagutii}, we follow the approach of Aranda et al.\textsuperscript{14} with some modifications. In brief, we first searched the scaffolds (BLASTN) against a database of bacterial and viral genomes (see Methods), and identified those with hits at bit score $> 1000$ and $E \leq 10^{-20}$; we consider these as significant hits. We then examined the sequence cover of these regions in each scaffold, and identified the percentage (in length) contributed by these regions relative to the scaffold’s
full length. Aranda et al.\textsuperscript{14} used an arbitrary threshold of 50\% sequence cover as indication of putative bacterial or viral contaminant, and thus removed scaffolds containing >50\% of putative bacterial or viral regions. Here, we systematically assessed the number of implicated genome scaffolds across the different thresholds of percentage sequence cover of putative bacterial or viral regions, and the corresponding gene models in these scaffolds (Supplementary Figure S9). At the most-stringent threshold (0\% sequence cover) any scaffold with any significant bacterial or viral hits is considered a contaminant, here 333 and 90 scaffolds respectively in \textit{S. goreau}i and \textit{S. kawagutii} (Supplementary Figures S9A and S9B); these represent <1\% of the total assembled scaffolds in each genome. In contrast, at the lenient threshold of 90\% sequence cover, only 32 and 2 scaffolds from \textit{S. goreau}i and \textit{S. kawagutii} are considered contaminants. In both genomes, the number of scaffolds shows a sharp decrease from thresholds at 0\% to 10\% sequence cover, followed by a gradual decrease as the subsequent thresholds become less stringent. A similar trend is observed with the implicated gene models on these scaffolds (Supplementary Figures S9C and S9D). The 0\% threshold may be too strict in these cases, since bacterial-like genes are known to be present in dinoflagellates. Here we chose 10\% as the deciding threshold, i.e. any scaffold with significant bacterial or viral hits covering >10\% of its length is considered a contaminant. In this way 129 and 33 scaffolds (and the gene models implicated within) were removed respectively from \textit{S. goreau}i and \textit{S. kawagutii}.

\textbf{Genome annotation and gene prediction}

For each genome assembly, a \textit{de novo} repeat library was first derived using RepeatModeler (http://www.repeatmasker.org/RepeatModeler/). All repeats (including known repeats in RepeatMasker database release 20150807) were masked using RepeatMasker (http://www.repeatmasker.org/).
We used transcriptome data to guide functional annotation assembled genomes. For *S. goreau*ii, we used the published transcriptome data (NCBI accession GSE72763) from Levin et al.\textsuperscript{15}. For *S. kawagutii*, we used the transcriptome data of CCMP2468 (MMETSP0132; RNA-Seq reads after filtering for adapters and low-quality reads) available from MMETSP\textsuperscript{16}, and the published transcripts (generated using the 454 platform) from Lin et al.\textsuperscript{3}. For RNA-Seq data, we assembled the reads using Trinity independently in “de novo” mode and “genome-guided” modes\textsuperscript{17}, after which vector sequences were trimmed using SeqClean (https://sourceforge.net/projects/seqclean/) based on UniVec database (ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/; build v9.0).

Following Aranda et al.\textsuperscript{14}, we used a customised PASA\textsuperscript{18} script (available at http://smic.reefgenomics.org/download/) that recognises an additional donor splice site (GA), and used the program alongside TransDecoder\textsuperscript{18} to predict coding sequences (CDS) in each genome. These CDS were searched (BLASTX, \( E \leq 10^{-20} \)) against a customised protein database that consists of RefSeq proteins release 76 and other annotated or predicted *Symbiodinium* proteins (total of 49,732,862 sequences; Supplementary Table S26). Only near full-length CDS were included in the subsequent analysis; here we determine this based on near full-length alignment of a CDS to a protein in the database (>70%) CDS were isolated, using a script provided with Trinity.

The near full-length gene models were checked for transposable elements using HHBLITS (probability=80\% and E-value=10^{-5})\textsuperscript{19} searching against the JAMg transposon database (https://sourceforge.net/projects/jamg/files/databases/), as well as with Transposon-PSI (http://transposonpsi.sourceforge.net/). Genes models containing transposable elements were removed from the gene set, and redundancy reduction was conducted using CD-HIT (ID=75\%)\textsuperscript{20}. The remaining gene models were processed using the
Prepare\_golden\_genes\_for\_predictors.pl (http://jamg.sourceforge.net/) script from the JAMg pipeline (altered to recognise GA donor splice sites). This script produces a set of “golden genes” which were used as a training set for the gene-prediction packages AUGUSTUS\textsuperscript{21} and SNAP\textsuperscript{22}. Following Aranda et al.\textsuperscript{14}, we used a customised code of AUGUSTUS (available at http://smic.reefgenomics.org/download/) so it recognises GA donor splice sites, and trained it to predict both coding sequences and untranslated regions; SNAP was trained for both GT and GC donor splice sites. Soft-masked genomes were passed to GeneMark-ES\textsuperscript{23} for training and gene prediction.

UniProt-SwissProt (release 2016\_01) proteins, MMETSP Suessiales proteins and the predicted \textit{Symbiodinium} proteins (described above) were clustered using CD-HIT (ID=100\%). The clustered proteins were used to produce a set of gene predictions using MAKER\textsuperscript{24} with protein2genome; the custom repeat library was used by RepeatMasker as part of MAKER prediction. A primary set of predicted genes was produced using the EvidenceModeler software package\textsuperscript{25} which had been altered to recognise GA donor splice sites. This package combines the gene predictions from PASA, SNAP, AUGUSTUS, GeneMark-ES and MAKER protein2genome, as well as the masked repeats (using custom repeat library), into a single set of evidence-based predictions. The weightings used for the package were: PASA 10, Maker protein 8, AUGUSTUS 6, SNAP 2 and GeneMark-ES 2.

The assembled genome scaffolds, predicted gene models and proteins are temporarily available at https://cloudstor.aarnet.edu.au/plus/index.php/s/6yziMf2ygWjGu0L.

\textbf{Analysis of plastid genomes}

As plastid genomes of dinoflagellates occur as minicircles, here we focused on our ALLPATHS-LG genome assemblies. To identify putative plastid genome fragments in our genome data, we used plastid gene sequences identified in \textit{Symbiodinium} type C3\textsuperscript{26},
*Symbiodinium minutum*\(^2^7\) and *Heterocapsa triquetra*\(^2^8\) as queries in BLASTN searches against our genome assemblies. To identify the conserved core regions in the putative plastid genome sequences, we set a high mismatch penalty (match score = 1, mismatch scores = -4, gap opening cost = 5, and gap extension cost = 2, \(E \leq 10^{-10}\)) in reciprocal BLASTN searches. The identified core region was then used to identify other genome scaffolds that were not previously identified by alignment with known plastid-encoded genes. These scaffolds were searched against the NCBI’s non-redundant nucleotide database (BLASTN at default parameters) to assess if they align to any known genes. All scaffolds identified as being of plastid origin, both those encoding known plastid genes and those encoding only core regions, were checked for circularisation using pairwise BLASTN (\(E \leq 10^{-10}\)). Artemis\(^2^9\) and Artemis Comparison Tool (ACT)\(^3^0\) were used to annotate the isolated scaffolds. The putative plastid genome sequences and their annotation are temporarily available at [https://cloudstor.aarnet.edu.au/plus/index.php/s/6yziMf2ygWjGu0L](https://cloudstor.aarnet.edu.au/plus/index.php/s/6yziMf2ygWjGu0L).

**Analysis of mitochondrial genomes**

Mitochondrial genes from the dinoflagellates *Alexandrium catenella* and *Karlodinium micrum* were used as queries to identify putative mitochondrial genome fragments within our ALLPATHS-LG assemblies using BLASTN (\(E \leq 10^{-10}\)). Nucleotide sequences of the *cox1*, *cox3* (cytochrome oxidase subunits 1 and 3 of complex IV) and *cob* (cytochrome *b* of complex III) genes and fragments of the large subunit rRNA (LSU rRNA) and the small subunit rRNA (SSU rRNA) were retrieved from the NCBI non-redundant nucleotide database. Scaffolds with *cox1*, *cox3* and *cob* hits were considered putative mitochondrial genome fragments and were assessed for evidence of circularisation using pairwise BLASTN. The putative mitochondrial genome sequences and their annotation are temporarily available at [https://cloudstor.aarnet.edu.au/plus/index.php/s/6yziMf2ygWjGu0L](https://cloudstor.aarnet.edu.au/plus/index.php/s/6yziMf2ygWjGu0L).
Identification of sets of homologous proteins and gene families

Supplementary Figure S8 depicts our workflow for delineation of sets of putatively homologous proteins, multiple sequence alignment, generation of protein-family and reference trees, and analysis of selection. Protein sequences were generated computationally, using the standard genetic code, from genome and/or transcriptome sequences of 31 organisms including *Symbiodinium* (Supplementary Table S17; 31-taxon set). Similarly, a 15-taxon set (14 dinoflagellates and the outgroup *Perkinsus marinus*) was established. Sequences of length < 30 amino acids were removed, and sets of putatively homologous proteins were generated using OrthoFinder\textsuperscript{31}. Sets that contain \( \geq 4 \) proteins, including at least one from a *Symbiodinium*, were taken forward. We assume that all proteins within each set (and thus the corresponding coding genes) share a common ancestor. Sequences in which no genome is represented more than once (single-copy sets) are assumed to be orthologs, while those in multi-copy sets may include co-orthologs and/or paralogs. The workflow includes steps designed to recover sets with co-orthologs (but no paralogs).

Evolution of gene families

We performed gain-and-loss analysis on the gene sets corresponding to the protein sets generated using OrthoFinder\textsuperscript{31} under a Dollo parsimony model\textsuperscript{32}, using *dollop* as implemented in PHYLIP 3.69 (http://evolution.genetics.washington.edu/phylip/). Here we focused on the *Symbiodinium* subtree (i.e. lineages for which genome data are available) with the immediate outgroup of *Polarella glacialis*. To assess the impact of Markov clustering granularity in OrthoFinder on our results, we analysed gene gain and gene loss using homologous protein sets that were generated independently using the inflation parameter \( I \) at 1.0, 1.5 and 2.0.
Supplementary Note

Nuclear genomes of S. goreau and S. kawagutii

The G+C contents of our assembled genomes of S. goreau (43.8%) and S. kawagutii (45.6%) are comparable to the earlier S. kawagutii assembly (44.0%)\(^3\) and to S. minutum (43.6%)\(^2\), and lower than for S. microadriaticum (50.5%)\(^1\). Figure S1A shows the extent of mapped sequence reads from S. goreau and from S. kawagutii to each assembled Symbiodinium genome. Most reads (>78%) mapped to the corresponding assembly. Only about 16.5% of reads from S. goreau mapped to the two S. kawagutii genomes, and conversely 17.6% of S. kawagutii reads to S. goreau. This is in contrast to 4.1% of S. goreau reads and 10.3% of S. kawagutii reads that mapped to S. microadriaticum. These results indicate a high extent of dissimilarity among Symbiodinium genomes, greatest between representatives of Clades C and A, and least (although not by far) between C and F.

We adopted a comprehensive ab initio approach for predicting genes, combining both evidence-based and unsupervised methods (Supplementary Methods). Average gene lengths (7671 bp in S. goreau, 6646 bp in S. kawagutii) are intermediate between the shortest (3788 bp in the earlier S. kawagutii genome)\(^3\) and longest average (12,898 bp in S. microadriaticum)\(^1\). In addition, we observed similar codon usage (Supplementary Figure S4) and amino acid profiles (Supplementary Figure S5) among the genes of S. goreau, S. kawagutii and S. microadriaticum; the latter shows a slight bias towards high-G+C codons. S. minutum\(^2\) shows a distinctive codon usage profile vis-à-vis the others, with substantially higher contents of arginine, serine and tryptophan in predicted protein sequences (Supplementary Figure S5).

Proportions of canonical (GC) and non-canonical 5′-donor splice sites (GA, GT) in Symbiodinium genomes are shown in Supplementary Table S9. These splice sites occur in
similar proportion in the genomes of *S. goreau*i and *S. kawagutii*, with GA << GC < GT (e.g. 19.5% GA, 36.1% GC and 44.4% GT in *S. kawagutii*); a similar pattern was observed in *S. minutum*², whereas in *S. microadriaticum*¹⁴ the canonical GC is more prominent (21.9% GA, 52.1% GC and 26.0% GT). Non-canonical 5′-donor sites were not explicitly considered in the gene-prediction process for the earlier *S. kawagutii* genome, although a dominance of GT sites (65.6%) was observed. In all *Symbiodinium* genomes, a non-canonical G usually immediately follows the acceptor splice site (Supplementary Table S9 and Supplementary Figure S6). Thus *Symbiodinium* of Clades C and F use donor splice sites similarly to Clade B, but different than Clade A; this is likely related to the higher G+C content in the *S. microadriaticum* genome (50.51%) than in the others (43.46-45.59%).

**Plastid genomes of *S. goreau*i and *S. kawagutii***

Plastid genomes of dinoflagellates occur as minicircles each with one or more protein-coding genes, plus a non-coding region that contains a highly conserved core²⁶,²⁸,³³. We identified putative minicircle sequences with plastid-encoded genes in each of our two genomes: nine sequences encoding 14 genes in *S. goreau*i (Supplementary Table S4), and 13 sequences encoding 13 genes in *S. kawagutii* (Supplementary Table S5). A highly conserved core region was identified, of 79 bp and 41 bp for *S. goreau*i and *S. kawagutii* respectively (Supplementary Table S7). We also identified putative “empty” minicircles (two in *S. goreau*i, one in *S. kawagutii*) that do not encode any gene. Empty minicircles have been described in other dinoflagellates, but were not seen in *Symbiodinium* type C3²⁶ or *S. minutum*²⁷. Plastid-encoded *psbl*, previously reported in *S. minutum* but not in *Symbiodinium* type C3, was found in minicircle-like sequences in both *S. goreau*i and *S. kawagutii*; we did not find direct evidence of circularisation in these sequences, but a core region is present in the *psbl*-encoding sequence from *S. goreau*i. Our results demonstrate that “empty”
minicircles and plastid-encoded psbl occur in Symbiodinium; full-length minicircle sequences in S. goreau and S. kawagutii remain to be validated.

Specifically from the S. goreau ALLPATHS-LG assembly we recovered eight scaffolds containing all 13 known plastid-encoded genes (Supplementary Table S4) described previously for Symbiodinium type C3, as well as a scaffold containing the psbl gene described previously in S. minutum. Some of the genes were recovered only in part, and none of the scaffolds shows clear evidence of circularisation. Three of the gene-coding scaffolds (SC1_Plastid_1, SC1_Plastid_2 and SC1_Plastid_3) displayed characteristics that have not been described before in dinoflagellate minicircles. These scaffolds are longer than minicircles observed in other dinoflagellates, and contain multiple core regions. They also encode multiple genes, some of which are fragmentary. Although minicircles with two genes have been observed in other dinoflagellates, studies in Symbiodinium type C3 and S. minutum have not recovered any such structures26,27.

All plastid genes except psaA, 16S rRNA and psaB were recovered in S. goreau. The 16S rRNA is fragmented in Symbiodinium; we recovered three partial fragments encoded on SC1_Plastid_3. Only 48% (973/2022 bp) of the psaA gene, representing the end of the sequence, was recovered in a single fragment on SC1_Plastid_1. One full-length and one partial (12%; 128/1029 bp) copy of the psbA gene are present in SC1_Plastid_2. The psaB gene is encoded over two fragments (representing regions of 50-1319 nt and 1502-2082 nt) that together comprise 89% (1851/2082 bp) of the length of the psaB described in type C3. The two fragments of the psaB gene are encoded sequentially, separated on the scaffold by 3 bp; the length of the missing sequence between the two fragments is 182 bp. Studies in other dinoflagellates suggest that genes such as psaA have undergone internal deletions as a way of reducing size to fit on a minicircle26,33, and such a process could also have impacted the psaB
gene described in type C3. A 79-bp core (Supplementary Table S7) was identified in *S. goreau*i and recovered in all scaffolds found to encode plastid genes. SC1_Plastid_8 contains a partial core region that is encoded at the beginning of the sequence and SC1_Plastid_9 contained a full length core region with two mismatches. The core region was used to isolate SC1_Plastid_10 that had not been previously identified as being of plastid origin. It is not circular, and a comparison with the NCBI NR database found no similarity with any known genes. The identity of this scaffold as a minicircle, and whether it is an “empty” minicircle, remain to be validated.

From the *S. kawagutii* data we recovered 13 scaffolds that contain 13 of the known plastid-encoded genes (Supplementary Table S5), with only *psbE* not recovered. Five of the scaffolds (SF_Plastid_1, SF_Plastid_2, SF_Plastid_3, SF_Plastid_4 and SF_Plastid_5) are outside the 2-4 Kbp size range observed in other dinoflagellates. The 16S rRNA was found as two partial fragments encoded on two scaffolds, consistent with our current understanding of the 16S rRNA gene structure in *Symbiodinium*. The *psaB*, *atpB*, *psaA*, *psbC* and *atpA* genes were recovered as multi-copy fragments encoded on the same scaffold. For example, the *atpA* gene in SF_Plastid_5 is present in two nearly complete copies, whereas in SF_Plastid_1 six partial fragments of the *psaB* gene were recovered. Gene structures of this nature have not been observed in any other dinoflagellate.

Three gene-encoding scaffolds (SF_Plastid_3, SF_Plastid_5 and SF_Plastid_12) were found to also be circular in *S. kawagutii*. SF_Plastid_3 has a 139-bp overlap (100% identity) at each end of the sequence, SF_Plastid_5 has a 1874-bp overlap (98% identity) and SF_Plastid_12 a 45-bp overlap (100% identity). A 41-bp core was identified in *S. kawagutii* and was recovered in only six of the plastid gene-encoding scaffolds. Two previously unidentified scaffolds (SF_Plastid_14 and SF_Plastid_15) were found to contain the *S.
kawagutii core region, with SF_Plastid_14 also found to be circular. Both scaffolds were checked against the NCBI nr database and show no similarity with any known encoded genes. Circularisation of SF_Plastid_14 occurs between the two encoded core regions that are positioned exactly at the start and end of the scaffold. The positioning of the core regions along the scaffold may be an artefact of assembly and so the identity of the scaffold, potentially as an “empty” minicircle, remains to be investigated.

In *S. goreaui* and *S. kawagutii*, some scaffolds identified as being of plastid origin show traits that have not been observed in other dinoflagellates. In both *S. goreaui* and *S. kawagutii* there are scaffolds which exceed the size expected of a minicircle (i.e. 1.3-3.0 Kbp\textsuperscript{26,27}), and encode gene structures that vary significantly from what has been observed for plastids of other dinoflagellates. Duplications within the scaffolds (encompassing genes, core regions and non-coding regions) and scaffolds composed of multiple minicircles could for instance be assembly artifacts. The data generation and assembly strategy adopted in this study was designed to recover nuclear rather than organellar genomes. These data serve as the first analysis platform for the organellar genomes of *S. goreaui* and *S. kawagutii*, and can be extended using a PCR sequencing strategy based on targeted primer design.

The G+C contents of the coding and non-coding regions for the plastid sequences in *S. goreaui* and *S. kawagutii* are summarised in Supplementary Table S6. In the plastid minicircles of *Symbiodinium* type C3, G+C content within the coding regions (36.03%) is lower than that within the non-coding regions (48.22%)\textsuperscript{26}. The gene-coding regions identified in our *S. goreaui* data have G+C content (36.80%) close to that published for type C3 (36.03%). The G+C content of the non-coding regions (45.64%), however, is three percentage points lower than the published data (48.22%), possibly due to the increased length of identified non-coding sequences (18,051 bp in *S. goreaui* and 11,475 bp in type
C3). Our *S. kawagutii* data have G+C content in their coding (35.46%) and non-coding regions (42.56%) below that observed in type C3 and in our *S. goreaui* data. For *S. kawagutii* both the length of the recovered coding and non-coding sequence are much higher than that of type C3. Many of the scaffolds encode duplicate regions of genes, increasing the total length of the coding sequence recovered beyond what is observed for type C3.

**Mitochondrial genomes of *S. goreaui* and *S. kawagutii**

We also identified sequences of mitochondrial genomes from *S. goreaui* (55,144 bp) and *S. kawagutii* (62,663 bp), each encoding the well-characterised three mitochondrial genes (*cox3*, *cox1* and *cob*) and fragments of large-subunit rRNAs.

We identified one scaffold from *S. goreaui*, and two from *S. kawagutii*, that contain genes known to be encoded in the mitochondrial genome of dinoflagellates. For *S. goreaui*, SC1_Mitochondria_1 (length 55,144 bp) was found to contain the *cox3*, *cox1* and *cob* genes, with a fragment of the LSU rRNA between the *cox3* and *cox1* genes. For *S. kawagutii* SF_Mitochondria_1 (length 62,663 bp) the *cox3*, *cox1* and *cob* genes were all recovered, and the LSU rRNA was found between the *cox3* and *cox1* genes. SF_Mitochondria_2 (length 35,965 bp) from *S. kawagutii* contains only the *cox1* gene. This configuration of *cox3*-LSU rRNA-*cox1*-cob is well-characterised in the apicomplexa *Plasmodium falciparum* and has been observed in dinoflagellates. The size of the mitochondrial genome in dinoflagellates has not yet been determined, but previous studies recovered a genome sequence of up to 42 Kbp. SC1_Mitochondria_1 from *S. goreaui* and SF_Mitochondria_1 from *S. kawagutii* both display characteristics of dinoflagellate mitochondrial genomes but are 55,144 and 62,663 bp in length respectively. Our findings suggest that the mitochondrial genome of *Symbiodinium* is considerably larger than what has been described previously at 42 Kbp in dinoflagellates.
**DMSP and DMS**

Dimethylsulphoniopropionate (DMSP) serves as an osmolyte and antioxidant for both the alga and the coral, as a nutrient for associated bacteria\(^3^7\) and as a signal in maintaining the complex interactions among holobiont partners\(^3^8\). In coral-associated *Symbiodinium*\(^3^9\) and other biological systems, DMSP lysase converts DMSP into dimethylsulphide (DMS). DMS, in turn, is the major input of biogenic sulphur into the marine boundary layer\(^4^0\) and can contribute to sulphate aerosols that nucleate cloud condensation, thereby lowering global temperatures\(^4^1\). DMSP concentrations are positively correlated with the thermal tolerance of *Symbiodinium*, and with bleaching tolerance of the coral *Acropora millepora*\(^4^2\).

Investigating the salinity-induced production of DMSP, Lyon et al.\(^4^3\) identified candidate enzymes of DMSP biosynthesis in the sea-ice diatom *Fragilariopsis cylindrus*. DMS production correlates with expression of *Alma1* (encoding DMSP lyase) in the bloom-forming cryptophyte *Emiliania huxleyi*\(^4^4,4^5\). Here we used BLASTP to search these sequences against our *Symbiodinium* gene models. We found three putative DMSP lyase genes in *S. goreau*, and four in *S. kawagutii* (Supplementary Tables S13 and S14); the predicted proteins show significant sequence similarity (\(E < 10^{-100}\) and nearly full-length alignment) to *Alma* genes in *E. huxleyi* and the Alma1 protein described in *Symbiodinium* Clade D, indicative of capacity to degrade DMSP into DMS. Since Raina et al.\(^4^6\) showed that corals can also produce DMSP, we further examined whether coral could degrade DMSP by searching for ALMA orthologs. We find that *A. digitifera* has the molecular machinery potentially to encode Alma1 and produce DMS. Since both coral and symbiotic alga have the genetic capacity to produce DMSP and DMS, we encourage future studies to confirm whether DMS is produced by corals and explore the mechanisms that regulate the dynamics of DMS(P) production and its influence on the complex interplay among corals, *Symbiodinium* spp. and other symbiotic microbes, especially those using DMSP as a source of sulphur.
**Evolution of gene families**

The results of gene gain and gene loss along *Symbiodinium* lineages with respect to the outgroup *P. glacialis* are shown in Supplementary Figure S10. Overall we observed a higher number of gene losses in the lineage leading to *S. kawagutii* (Clade F) than in that leading to *S. goreau* (Clade C), i.e. 3122 versus 1375 genes respectively among protein sets generated at $I = 1.0$ (Supplementary Figure S10A). More gene losses are inferred when the clustering is more granular, e.g. 6314 versus 3482 genes at $I = 2.0$ (Supplementary Figure S10C). In comparison, the lineages leading to *S. minutum* (Clade B) and *S. microadriaticum* (Clade A) appear to have gained more genes (186 and 218 genes respectively) than those leading to *S. kawagutii* and *S. goreau* (97 and 69 respectively) ($I = 2.0$; Supplementary S10C). The greater inferred number of gene loss in the lineages leading to Clades C and F, and of gene gain in lineages leading to Clades A and B, while interesting, remain to be validated with more-complete genome data from *Symbiodinium*.

**Supplementary References**


