Supplementary Information

Whole-genome duplication in an algal symbiont serendipitously confers thermal tolerance to corals

Katherine E. Dougan[1,2]*, Anthony J. Bellantuono[3], Tim Kahlke[4], Raffaela M. Abbriano[4], Yibi Chen[1], Sarah Shah[1], Camila Granados-Cifuentes[2], Madeleine J. H. van Oppen[5,6], Debashish Bhattacharya[7], David J. Suggett[4], Cheong Xin Chan[1]*, Mauricio Rodriguez-Lanetty[2]*

Affiliations:

[1]The University of Queensland, School of Chemistry and Molecular Biosciences, Australian Centre for Ecogenomics, Brisbane, QLD 4072, Australia

[2]Florida International University, Department of Biological Sciences, Miami, FL 33099, U.S.A.

[3]Florida International University, Department of Biological Sciences, Biomolecular Science Institute, Miami, FL 33099, U.S.A.

[4]University of Technology Sydney, Climate Change Cluster, NSW 2007, Australia

[5]School of Biosciences, The University of Melbourne, Parkville, VIC 3010, Australia

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

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

*Correspondence: k.dougan@uq.edu.au (K. E. Dougan), c.chan1@uq.edu.au (C. X. Chan), rodmauri@fiu.edu (M. Rodriguez-Lanetty)

Materials and Methods

Symbiodiniaceae cultures and cell preparation for DNA extraction

The *Durusdinium trenchii* strain CCMP2556 was originally isolated by Dr Mary Alice Coffroth (Buffalo University, New York, USA) from an *Orbicella faveolata* coral sampled at 10m depth on Tennessee Reef, Florida Keys (24.7335° N, 20.7669° W). The CCMP2556 culture was first acquired from the National Center for Marine Algae and Microbiota (Bigelow Laboratory for Ocean Sciences, East Boothbay, Maine), and maintained for one year in PROV-50 medium (NCMA) with an antimicrobial cocktail (kanamycin [50 μg/mL], ampicillin [100 μg/mL], streptomycin [50 μg/mL], amphotericin B [2.5 μg/mL]) to minimize bacterial and fungal contamination. During this time, the culture was passaged in the antimicrobial cocktail and inspected via light microscopy to confirm culture quality and the absence of microbial contamination on a monthly basis. PCR amplification and sequencing of the cp23S marker sequence was also regularly conducted for comparison to the known *D. trenchii* sequence to verify the genetic identity of the cultures. No other *Durusdinium* spp. were maintained in the laboratory prior to or during this time. To prepare the CCMP2556 cells for extraction of genomic DNA, the culture was cleaned using methods adapted from Su et al. (2007) to further minimize bacterial contamination. Briefly, 5×10^7 log-phase cells were pelleted via centrifugation (500 *g*, 3 min) and resuspended in 40 mL sterile PROV-50 medium using the antimicrobial cocktail above. Cells were then pelleted and resuspended in PROV-50 medium with 0.005% Tween-80 and EDTA (0.1 M, pH 8.0). Resuspended cells were incubated at 25°C for 1 hour on a platform

rocker followed by the addition of lysozyme (0.5 mg/mL) and SDS (0.25%) and another incubation at room temperature (RT) with rocking for 10 minutes. Finally, cells were washed three times with PROV-50 medium containing the antimicrobial cocktail, and snap-frozen in liquid nitrogen prior to DNA extraction.

The *D. trenchii* strain SCF082 (previously designated UTSD amur-D-MI) was originally isolated from an *Acropora muricata* coral colony on the Great Barrier Reef (Magnetic Island) and maintained at the Australian Institute for Marine Science (AIMS) culture collection, prior to shipment to University of Technology Sydney. SCF082 was grown in xenic cultures in 200mL IMK medium (Daigo's IMK powder for microalgae) for 2 weeks at 24ºC under 25 μmol/s/m^2 light and pelleted in 50mL Falcon tubes at 1,000 RPM (ELMI CM-6MT centrifuge ~184 *g*). To reduce bacterial contamination, cell pellets were washed three times by re-suspending them in sterile PBS buffer (50 mL) and pelleting again at 1,000 RPM (ELMI CM-6MT centrifuge ~184 *g*). Bacterial contamination was assessed before and after wash based on plating on Marine Agar, and cell count by flow cytometry: cells were stained with SYBR Green for 15 minutes in the dark prior to detection by a Cytoflex S flow cytometer (Ex: 488 nm, Bandpass filter: 525/40). Flow cytometer results showed a 90% reduction in bacterial contamination. Final pellets were combined and pelleted once again at 4,000 *g* (5 min), and snap-frozen in liquid nitrogen before DNA extraction.

Extraction of genomic DNA and sequencing of 10X Genomics linked reads

To extract genomic DNA from CCMP2556 cells, frozen cells (above) were ground together with acid-washed 425-600 μm glass beads (Sigma) in a liquid nitrogen-chilled mortar and pestle. The ground cells (~50 mg) were used for DNA extraction using the MagAttract HMW DNA Kit

(Qiagen) following the manufacturer's protocol. To minimize DNA degradation, wide-bore pipette tips were used for all steps involving the handling of lysate or DNA, and mixing was performed by gentle inversion instead of the manufacture-indicated pipetting. The final DNA was eluted in Buffer AE (Qiagen; 10 mM Tris-Cl 0.5 mM EDTA, pH 9.0) and stored at -20°C. The DNA sample was analyzed using pulse-field gel electrophoresis which indicated the presence of an intact sample (mean fragment size ~100 kb), confirming DNA integrity. The Chromium library was constructed using 1 ng of genomic DNA as input, and sent for sequencing at HudsonAlpha Institute for Biotechnology (Huntsville, AL) using the Illumina HiSeq X platform (two lanes, 2×150 bases paired-end; table S1).

To extract genomic DNA from SCF082, cell lysis was performed on snap-frozen cells using a custom-made French press with two passes at 1,000 psi. Lysate was collected and subjected to standard extractions using ice-cold phenol:chloroform:isoamyl alcohol (25:24:1, v/v/v), and chloroform:isoamyl alchohol (24:1, v/v). DNA was precipitated using sodium acetate (pH 5.2; final concentration 0.3M), and an equal volume of ice-cold isopropanol. Precipitated DNA was washed with 70% ethanol and resuspended in TE buffer (overnight, room temperature) to allow for high-molecular weight fragments to redissolve. DNA was size selected using PluePippin (SageScience) pulse-field gel electrophoresis prior to 10X Chromium library preparation. The library was sent to Ramaciotti Center for Genomics, University of New South Wales (Sydney, Australia) for sequencing using Illumina NovaSeq 6000 (S1 flow cell, 2×150 bases paired-end).

De novo genome assembly

For each isolate, a preliminary draft genome was assembled *de novo* using 10X Genomics Supernova v2.1.1. For CCMP2556, the estimated genome coverage (~100×) exceeded the

optimal range (38–56×) of the Supernova assembler; we subsampled the 1.6B reads to 600M reads (~60× coverage). For SCF082, coverage estimates were observed to be impacted due to the presence of contaminant DNA from microbial sources in the sequencing reads; the *de novo* assembly was generated using all 1.4B reads with the flag *–accept_extreme_coverage*.

Presence of putative contaminant scaffolds in the supernova assemblies was investigated using a comprehensive approach adapted from Iha *et al.* (*33*) informed by read coverage, G+C content, taxonomic designation, and *de novo* transcriptome mapping (fig. S16). Taxon-annotated G+C-coverage plots (fig. S20) were generated using the Blobtools suite (v1.1) (*34*) to identify scaffolds in each assembly that deviated by read coverage, taxonomic sequence similarity, and/or G+C content (fig. S26). Read coverage was obtained via BWA (v0.7.17) mapping of Longranger (v2.2.2) (*35*) quality-trimmed reads (default settings) to the genome assembly. The taxonomic identity of scaffolds was assigned based on BLASTN search ($E \leq 10^{-20}$) against genome sequences from bacteria, archaea, viruses, and alveolates in the NCBI nt database (release 2021-05-10). *De novo* transcriptome assemblies were mapped to the genome assemblies using Minimap2 (v2.18) (*36*) within which we have modified the codes to account for non-canonical splice sites of dinoflagellates. Scaffolds that were designated as non-dinoflagellate were removed from the assemblies if they lacked mapped transcripts from the corresponding *de novo* transcriptome assembly, when <10% of mapped transcripts indicate evidence of introns in the genomes. We considered a scaffold as a putative contaminant if (a) its sequence coverage or G+C content is out of the 1.5*interquartile range, and (b) it lacks any transcript support defined above. Upon removal of these putative contaminant sequences from the CCMP2556 assembly,

105 the filtered assembly was incorporated in the database as the *D. trenchii* reference for assessing
106 the assembled scaffolds of SCF082.

107 The filtered genome assemblies were then further refined using two separate scaffolding
108 programs, first with L_RNA_Scaffolder (*37*) followed by ARBitR (*38*) to yield the final
109 assemblies. Briefly, *de novo* assembled transcripts were mapped onto the filtered genome
110 assemblies using minimap2 modified to recognize non-canonical splice sites of dinoflagellates,
111 and the mapping results were used to further scaffold the assembled genome sequences using
112 L_RNA_Scaffolder at default parameters. Longranger basic quality-trimmed linked genome
113 reads were mapped to the scaffolded genome assemblies with BWA for further scaffolding and
114 collapsing with ARBitR, which incorporates the distance information from linked read
115 sequencing when merging and scaffolding assemblies.

116 Genome size was estimated following the approach by Lin et al. (2015), where it is inferred from
117 *k*-mers enumerated from filtered genome-sequence data (independently at k = 19, 21, 23, 25, 27,
118 29 and 31), and the final estimate represents the average of these values. Sequence similarity of
119 the CCMP2556 and SCF082 genomes was assessed using the nucmer function (--mum) from
120 MUMmer v4.0.0 (*39*), requiring alignments of 100bp, 1kb, and 10kb with a minimum sequence
121 identity at 90%.

122 **Ab initio prediction of protein-coding genes**

123 The prediction of protein-coding genes from assembled genome sequences was conducted using
124 a comprehensive dinoflagellate-specific workflow (*40*). This workflow incorporates multiple
125 evidence types including transcriptome and protein sequence data, integrating the results from

multiple predictors to yield high-quality gene models (table S2). RNAseq data was used to inform gene prediction and assembled separately for each of the four treatments in both isolates, via both *de novo* and genome-guided modes, yielding a total of 2,014,780 transcripts for CCMP2556 and 1,007,955 transcripts for SCF082 across the eight assembled transcriptomes in each isolate.

Gene models were functionally annotated using BLASTP ($E \leq 10^{-5}$) searches against the UniProtKB (release 2021_03; Swiss-Prot and TrEMBL) database. Pfam domains within predicted proteins were annotated using PfamScan ($E \leq 0.001$) in conjunction with the Pfam-A database. Completeness of the genomes and the predicted proteins was performed using BUSCO5 (*41*) and the alveolata_odb10 and eukaryota_odb10 datasets under both genome and protein modes. A PCA plot of metrics of the predicted gene models compared to the rest of Suessiales was carried out using the prcomp package in R. Orthofinder v2.3.10 (*42*) was used to infer homologous gene sets among Suessiales species using NCBI-BLAST.

Inference of collinear gene blocks and putative ohnologs

Collinear gene blocks were identified with the program MCScanX (*43*) using both intra-species mode (-b 1) to identify putative duplicate gene blocks within each genome (i.e. segmental and/or whole-genome duplication) and inter-species mode (-b 2) to identify syntenic gene blocks between the two genomes. For each comparison, all-vs-all BLASTP search results were restricted to the top five hits for each protein with 50% sequence alignment of either query or subject. Predicted genes from each genome were classified according to their duplication history using MCScanX's *duplicate_gene_classifier* as singleton, dispersed duplicates (i.e. duplicates separated by more than 20 genes), proximal duplicates (i.e. duplicates separated by less than 20

148 genes), tandem duplicates, and WGD/segmental duplicates (i.e. ohnologs). The rates of
149 asynonymous (K_a) and of synonymous (K_s) substitutions were estimated for ohnolog gene pairs
150 using the *add_ka_to_ks_to_synteny.pl* perl script following Clustal Omega v1.2.4 alignment.

151 The putative WGD event was further investigated for CCMP2556 using the wgd pipeline (*44*).
152 Briefly, homologous protein clusters were inferred using a Markov Clustering algorithm (*45*)
153 from the previous all-vs-all BLASTP search used for MCScanX and aligned with MAFFT (*46*).
154 Phylogenetic trees of homologous protein clusters inferred with FastTree2 (*47*) were then used to
155 estimate K_s values for each cluster using codeml from the PAML package (*48*). A Gaussian-
156 mixture model (GMM) was applied to the K_s distribution, using a four-component model that
157 provided the best fit for the data according to Akaike information criterion (AIC), yielding a final
158 node-averaged histogram of K_s distribution. The *D. trenchii* NIES-2907 assembly from Shoguchi
159 *et al.* (*49*) was not included in this and further analyses considering it was assembled from short-
160 read Illumina sequencing and is the smallest Symbiodiniaceae genome assembly with the lowest
161 number of predicted genes.

162 **Ohnolog gene expression divergence**

163 To assess gene expression, all RNA-Seq reads were first quality-trimmed using fastp (mean
164 Phred quality $\geq$ 30 across a 4bp window; minimum read length of 50bp). To assess divergence of
165 expression between each ohnolog in an ohnolog-pair, trimmed RNA-Seq reads were mapped to
166 the corresponding genome using Hisat2 v2.2.1 (--concordant-only) with a Hierarchical Graph
167 FM index that is informed by annotated exon and splice sites. Counts of uniquely mapped
168 paired-end (PE) reads overlapping with coding sequence (CDS) regions were then enumerated

using featureCounts (-p --countReadPairs --B -C) from Subread v0.2.3 (*50*) and fragments per kilobase of transcript per million mapped reads (FPKM) calculated from the counts.

To investigate the gene expression profiles of ohnolog pairs, an ohnolog pair with one or both non-functionalised (i.e. non-expressed) gene(s) were first identified based on mapping of RNA-Seq reads to these ohnolog copies. To consider an ohnolog as "expressed", a mean FPKM > 1 was required in at least one of the four treatments. Ohnolog pairs for which one copy was not expressed (FPKM < 1 in all conditions for one ohnolog; FPKM ≥ 1 in one or more conditions for the other) represent non-F in one ohnolog copy. In cases for which both ohnologs were expressed (FPKM > 1), the $\log_2$(FPKM + 1) expression values were used to classify ohnolog pairs according to their expression profile into four groups: High Correlation and Similar Expression (HCSE), High Correlation and Different Expression (HCDE), Low Correlation and Similar Expression (LCSE), and Low Correlation and Different Expression (LCDE). A Spearman correlation test was used to assess similarity in expression profiles: ohnolog pairs with coefficients > 0.7 or < -0.7 designated as High Correlation (HC), whereas those with coefficients between -0.7 and 0.7 as Low Correlation (LC). A paired *t*-test of across all treatments and replicates was used to identify ohnolog pairs with Similar Expression (SE, $p < 0.05$) or Different Expression (DE, $p > 0.05$). Ohnolog pairs that switched in dominance across treatments and are potential cases of sub-functionalization or neo-functionalization, were identified using unique mapping results (i.e. no mapping from a read pair to more than one gene) using a paired *t*-test on the $\log_2$(FPKM+1) values for each treatment separately, requiring each ohnolog to be dominant to the other in at least one treatment ($p < 0.05$) and were input to topGO for Gene Ontology enrichment analysis ($p < 0.01$).

A weighted gene co-expression network analysis (WGCNA) was then performed on all genes in R using the WGCNA package. The raw counts from htseq-count were first filtered to remove lowly (<10 reads) expressed genes and variance normalized counts were acquired using the standard DESeq2 workflow followed by its *varianceStabilizingTransformation*. As symbiosis is a strong driver of expression in Symbiodiniaceae, using the inferred soft-thresholding power for reducing noise and setting a required threshold for gene correlations would have yielded a mean connectivity of over 4,000 at the inferred power of 6. Therefore a weighted, unidirectional co-expression network was inferred using a power of 18, the recommended value for signed networks with less than 20 samples. Co-expression similarity and adjacency was then assessed and converted into a topological overlap matrix (TOM). A gene dendrogram was then produced through hierarchical clustering (method=complete) of the corresponding TOM dissimilarity and branches, representing highly co-expressed genes, were cut using dynamicTreeCut (deepSplit=2, pamRespectsDendro=FALSE, minModuleSize=30). Module eigengenes were then clustered (method=average) followed by the merging of those with similar expression correlations (eigengene correlation >= 75). Significant relationships between modules and the factors symbiosis and temperature was then assessed and a *p*-value and correlation assigned to the module.

The distribution of ohnolog gene pairs throughout the WGCNA modules was then examined and pairs classified into different categories using a decision tree (fig. S8) that represent the level of conservation and/or divergence of gene expression in relation to symbiosis and temperature. Due to the stronger effect of symbiosis compared to temperature on gene expression, nearly all ohnolog pairs for which both genes were expressed (9,684 ohnolog pairs) exhibited a correlation

to symbiosis in at least one of the two ohnologs in a pair (9,499 ohnolog pairs) while far fewer pairs exhibited any correlation to temperature (2,294 ohnolog pairs). To account for these disparate effects, the expression patterns of ohnolog pairs for each of the two variables were represented as proportions, in context of the number of ohnolog pairs exhibiting a correlation to that specific variable. The expression patterns of ohnolog pairs were first broken down into those with "Similar Profiles", or when both ohnologs in a pair exhibited the same expression pattern of either a Positive or Negative correlation to both symbiosis and temperature. Other ohnolog pairs were designated as having "Dissimilar Profiles" as the two ohnologs exhibited conflicting patterns to either symbiosis and/or temperature. Of those with "Dissimilar Profiles", if they had diverged in expression pattern (i.e. one positive and one negatively correlated) to one of the two factors they were considered as "Divergent" for that factor or "Conserved" if they exhibited the same pattern for that factor. Those pairs where only one of the two ohnologs exhibited a significant correlation to a factor were considered as a "Gain/Loss" scenario, where one of the two genes either gained or lost a relationship to that factor.

Detection of metabolic pathways preferentially retained in duplicate

Metabolic enzyme detection was conducted using PRIAM v2 (January 2018 release). Following the formula used in Aury *et al.* (*19*), we inferred metabolic pathways that were preferentially retained in duplicate following WGD. Briefly, we identified metabolic enzymes that had been uniquely retained as ohnologs or singletons. We then compared the proportion of enzymes uniquely retained as ohnologs to singletons, to the background proportion of the number of ohnologs and singletons annotated in the genome. This tests whether the number of uniquely retained metabolic enzymes for a particular pathway exceeds the background levels that would

be expected to occur by random. We additionally required (a) five or more distinct enzymatic proteins to be identified as uniquely retained in either duplicate or singleton, and (b) pathways to be significantly overrepresented in both isolates.

Ohnolog splicing divergence

To examine the level of conservation of splice junctions in ohnolog pairs, all *de novo* assembled transcripts were first aligned to the genome using a modified minimap2 v2.20 and then input to PASA for identification and annotation of splice sites. Splice sites categorized as alternative acceptor, alternative donor, alternative exon, retained exon, and skipped exon were retained for subsequent analysis. Each identified splice event was assigned two unique identifiers to represent the upstream and downstream positions of the splice event, along with its gene identifier and genomic location. The upstream and downstream 300bp-region for each splice event were then extracted using the bedtools v2.30 *flank* and *getfasta* functions. An all-*versus*-all BLASTN search of the extracted splice junction sequences was used to identify sequence similarity ($E \leq 10^{-5}$) between the sequences. Custom Python scripts were then used to filter the BLASTN results to identify conserved splice junctions, defined as splice junctions where both the upstream and downstream regions for a splice event in two ohnologs were significantly similar ($E \leq 10^{-5}$). The splice junction profile for each ohnolog pair was then converted to a binary representation, where the presence of a splice junction in an ohnolog was represented as 1 and the absence of a splice junction represented as 0 (i.e., conserved splice junctions represented as 1 in both ohnologs compared to 1 and 0 for those that were not conserved. A Kendall's rank correlation was then conducted in R to identify ohnolog pairs that exhibited high level of conservation in splice

junctions. An exact binomial test was also performed to identify ohnolog pairs that had diverged in terms of total splice junctions ($p < 0.05$).

For inferring differential exon usage (DEU) within genes among the treatment conditions, gene models were first broken up into exon "counting bins" using the Python script *dexseq_prepare_annotation.py* from DEXSeq. The relative usage of each exon bin, i.e. the number of transcripts mapping to the bin or to the gene, was then calculated from the hisat2 BAM file using *dexseq_count.py*. The DEXSeq R package was then used to infer differential exon usage within genes using a generalized linear model, correcting for significance at the gene level using the Benjamini-Hochberg method. For singletons and ohnologs, we extracted and plotted the $\log_2$(fold-change) of highly significant DEU (adjusted $p < 0.01$) to view the differentially expressed exons.

RNA editing

Detection of mRNA editing events was inferred using JACUSA2 with hisat2-mapped RNA-seq reads and bowtie2-mapped longranger-trimmed 10X genomic reads using the *--very-sensitive* mode. PCR duplicates were removed from the BAM files using the *MarkDuplicates* function from the Picard Toolkit v2.26.2 (https://broadinstitute.github.io/picard/). Putative sites of mRNA editing identified by JACUSA2 were first filtered to meet the criteria of the D, Y, H feature filters. A second round of filtering was then performed using the R package JACUSA2helper, further requiring a score ≥ 1.15, >10 read coverage in DNA, >5 read coverage in each RNA replicate, observed bases ≤ 2, and the inference of RNA editing across all four treatments.

Supplementary Text

Genome assembly of *Durusdinium trenchii*

Coral reef ecosystem health is critically regulated by the functional performance of the microalgae (Family: Symbiodiniaceae) that thrive in endosymbiosis with reef-forming corals. Optimum metabolically-derived nutrient exchange between these microalgae and coral (*2, 3*) fuels coral growth and calcification essential to reef productivity and accretion. However, amplifying anthropogenic stressors are increasingly driving sub-optimum environmental conditions for symbiosis that result in an accumulation of toxic metabolic biproducts (e.g. Reactive Oxygen Species; (*51*)), and in turn endosymbiont rejection from the host ("coral bleaching"). Episodes of mass – region-wide – coral bleaching have become increasingly common place changing the entire nature of coral reef systems (*52, 53*). Importantly, Symbiodiniaceae is an extremely taxonomic (*1*) and functionally (e.g. Suggett et al. 2015) diverse family, and more "stress-compatible" (or rather those suited to more suboptimal reef environment conditions) taxa can proliferate in coral tissues enabling persistence under conditions where corals would otherwise bleach (Berkelmans & van Oppen 2006, Claar et al. 2020), including present day "natural extremes", such as super-hot tidal pools and lagoons (Oliver & Palumbi 2010, Camp et al. 2019).

Resolving how these "stress compatible" symbionts have evolved and function has remained a major goal in efforts to better manage reef systems into the future, including through more aggressive active interventions to evolve more heat-tolerant symbionts or manipulating host-symbiont partnerships (van Oppen et al. 2015, Voolstra et al. 2021). Indeed, this evidence for WGD in *D. trenchii* has been linked to its thermotolerance, which is widely supported by eco-physiological observations whereby *D. trenchii* often replaces the pre-dominant coral symbiont

299 taxa under thermal stress, thereby conferring increased thermal tolerance to the host (*54-56*).
300 Generally greater thermotolerance of *D. trenchii* and the relatively recent spread of this species
301 in the Caribbean (*22*) demonstrate the potential for *D. trenchii* to provide corals with more
302 resilience to ocean warming.

303 We generated high-quality *de novo* genome assemblies from two isolates of *D. trenchii*:
304 CCMP2556 (total length of assembly = 1.70 Gb; N50 = 750 Kb; 29, 137 scaffolds) and SCF082
305 (total length of assembly = 1.64 Gb; N50 = 398.5 Kb; 44,682 scaffolds) using 10X Genomics
306 linked reads (table S1). The CCMP2556 assembly is the most contiguous reported in
307 Symbiodiniaceae aside from the recent chromosome-level assemblies for *Symbiodinium*
308 *microadriaticum* (*57*) and *Breviolum minutum* (*58*). The estimated genome size based on *k*-mers
309 is 1.05 Gb for CCMP2556 and 1.65 Gb for SCF082 (table sS1 and S2; see Methods); however,
310 these sizes may be underestimated when a high amount of repetitive or duplicated DNA exist in
311 a genome. The identity of the two isolates was confirmed by the comparison of ribosomal large
312 subunit sequences that were identical between the two isolates and near identity of ITS2
313 sequences with one mismatch in the 639bp D1/D2 region shared between the two genomes (fig.
314 S1).

315 We used existing transcriptome data (*24, 59*) to guide prediction of protein-coding genes from
316 the assembled genome sequences using a customized approach tailored for dinoflagellates (see
317 Methods), yielding 55,799 protein-coding genes for CCMP2556 and 53,519 for SCF082 (table
318 S3). In both cases, over 75% of the predicted genes are supported by transcriptome evidence
319 (table S3). Genome data completeness, assessed using predicted proteins against the
320 Benchmarking Universal Single-Copy Ortholog (BUSCO) (*41*) for Alveolata (alveolata_odb10),

was 74.8% (CCMP2556) and 82.4% (SCF082) (table S5; see Methods). These numbers are higher than for an earlier published draft genome of *D. trenchii* (*49*) and those of most published dinoflagellate genomes analyzed using the BUSCO assessment of completeness.

Within the Order Suessiales, which contains both the family Symbiodiniaceae and the genus *Polarella*, *D. trenchii* genes revealed distinct patterns relative to introns and repetitive elements. A principal component analysis of various attributes of the predicted genes identifies longer introns and more introns per gene in the *D. trenchii* genome (fig. S11), although this trend was not specific to any category of gene duplication (fig. S12). Compared to other Symbiodiniaceae taxa (*8*), the repeat landscapes of both *D. trenchii* genomes (fig. S3) show a distinct distribution of predominantly LTR repeats, of which many are highly conserved (those constituting ~2% of the genome sequences have Kimura distances centered between 0 and 5). Long-interspersed nuclear elements (LINEs) and some LTRs appear to be highly divergent (Kimura distances >25), lending support to the notion of ancient LINE expansion in Symbiodiniaceae and *Polarella*, pre-dating the diversification of Suessiales (*8, 60*).

Detection of mRNA editing

Dinoflagellate genomes are known to use mRNA editing (*61*), a form of post-transcriptional regulation that can diversify an organism's response beyond what is coded in the genome, which has yet to be examined in context of WGD. Combining the transcriptome and genome data in a comparative analysis for CCMP2556, we identified 3,754 mRNA edited sites in gene regions: 1,802 implicated in synonymous substitutions, 1,909 in missense (non-synonymous) mutations, and 43 in nonsense mutations (e.g. premature stop codons or truncation of protein sequences). Of these, most edited sites were identified in dispersed duplicates (1,351 = sites) followed by

343 ohnologs (1,255 sites), singletons (500 sites), tandem duplicates (92 sites) and proximal
344 duplicates (33 sites; table S13). Among genes expressed in cells grown in the four distinct
345 conditions, we identified 3,754 (*ex-hospite* at 28°C), 2,643 (*in-hospite* at 28°C), 2,687 (*ex-*
346 *hospite* at 34°C), and 2,654 (*in-hospite* at 34°C) mRNA-edited sites. Of the total sites, more than
347 one-half (1,112 [54.8%]) were identified in all four treatments, but all the edited sites were
348 identified in genes expressed *ex-hospite* at 28°C (fig. S19); none were exclusively found *in-*
349 *hospite* or in *ex-hospite* cells grown at 34°C. These results suggest that the majority of mRNA
350 editing in *D. trenchii* occurs *ex-hospite* or in the absence of temperature stress *in-hospite*, and
351 that there is a decrease in mRNA editing following the transition to a symbiotic lifestyle and/or
352 exposure to thermal stress.

353 We then examined ohnolog divergence by searching for evidence of differential mRNA editing
354 in relation to symbiotic lifestyle (i.e. edited site exclusive to either *ex-hospite* or *in-hospite*)
355 and/or temperature stress (i.e. edited site exclusive to 28°C or 34°C). Of the 16 genes identified
356 with differential mRNA editing (table S14), the majority of identified genes (15 genes; 93.8%)
357 were associated with the symbiotic lifestyle, including a light-harvesting complex stress-related
358 protein 3.1 (LHCSR3.1), sporozoite surface antigen MB2, two peridinin-chlorophyll *a*-binding
359 proteins, and an insulin-degrading enzyme. These symbiosis-driven edited sites were
360 predominantly located in CDS regions, with the edited sites of only two genes, hypoxia-
361 inducible factor prolyl hydroxylase 2 (HIF-PH2) and G-protein-signaling modulator 1 (Gpsm1),
362 restricted to intron regions. The latter may represent editing in non-coding RNA and/or
363 regulatory regions within introns. In contrast, differential mRNA editing between 28°C and 34°C
364 was only found in two genes, Gpsm1 and a reticulocyte-binding protein 2 homolog a, with both

365 located in intron regions. Overall, these results reveal that the symbiotic lifestyle had a greater
366 influence on the retention and divergence of ohnologs, suggesting *D. trenchii* may have adapted
367 to life in different symbiotic modes.

368 **Global patterns in ohnolog divergence**

369 To gain deeper insights into the evolution of ohnologs post-WGD, we investigated their gene
370 structure. We examined exon-intron structure by mapping *de novo* and genome-guided
371 transcriptome assemblies to the genome to identify putative locations of alternative splicing
372 (table S15). We compared the average number of splice junctions per gene among different
373 categories of gene duplication (i.e. singleton, proximal duplicate, dispersed duplicate, tandem
374 duplicate, and ohnolog) to determine the number of splice junctions, or genomic locations
375 implicated in alternative splicing events among the ohnologs. On average, ohnologs contained
376 significantly more splice junctions per gene (800) compared to any other category (each <400;
377 fig. S22 and S23), which may indicate novel functions and/or a method of nonsense mRNA-
378 mediated decay.

379 We then compared the alternative splicing profiles within ohnolog pairs to determine whether
380 alternative splicing might be implicated in the divergence of gene duplicates following WGD.
381 For each ohnolog pair, we determined the number of intron/exon splice junctions, then assessed
382 the conservation of each splice junction by comparing upstream and downstream regions. We
383 found that the number and conservation of splice junctions are both diverged in 569 ohnolog
384 pairs, these numbers are similar, but the junctions are not conserved in 1587 pairs, whereas the
385 numbers are dissimilar (but most junctions that are present are conserved) in 1786 pairs (table

386 S16). Taken together, the increased numbers of splice junctions and DEU in ohnologs implicate
387 divergence in alternative splicing in the evolution of ohnologs following WGD.

388 We assessed the variability in exon usage due to alternative splicing by combining gene isoforms
389 into non-overlapping exon "counting bins" and compared the number of mapped transcripts per
390 exon across the gene expression treatments. We refer to this index as differential exon usage
391 (DEU). We compared a reduced model that controlled for the other as a covariate (i.e. reduced
392 model for temperature included life mode as a covariate) to identify significant DEU across
393 treatments (Benjamini-Hochberg adjusted $p < 0.025$) and then inferred DEU at the gene level
394 (FDR < 0.025) by accounting for the multiple tests of DEU for each exon in a gene. We found
395 that 10,857 genes exhibit significant DEU: most of these are in relation to life mode (10,341
396 genes) followed by temperature (1,444 genes) with only a few significant in both life mode and
397 temperature (7 genes) (fig. S24 and table S17).

398 In contrast to small-scale duplication events, WGD enables the divergence of entire metabolic
399 pathways. We assessed metabolic pathways in *D. trenchii* that were retained in duplicate, and if
400 they exhibit divergence in expression of the implicated genes. Using the predicted proteins of
401 both CCMP2556 and SCF082, we used the method of Aury *et al.* (*19*) to assess biased retention
402 of ohnolog pairs coding for enzymes in metabolic pathways, by mapping the associated Enzyme
403 Commission numbers to the KEGG database (*62*); see Methods. The proportion of enzymes
404 coded by genes that were uniquely retained as ohnologs or singletons, compared to their overall
405 proportions in the genome, was used to determine which KEGG pathways were preferentially
406 retained in duplicate following WGD. In total, we identified 98 metabolic pathways that were
407 preferentially retained in duplicate in both isolates (table S10).

Figs. S1 to S20

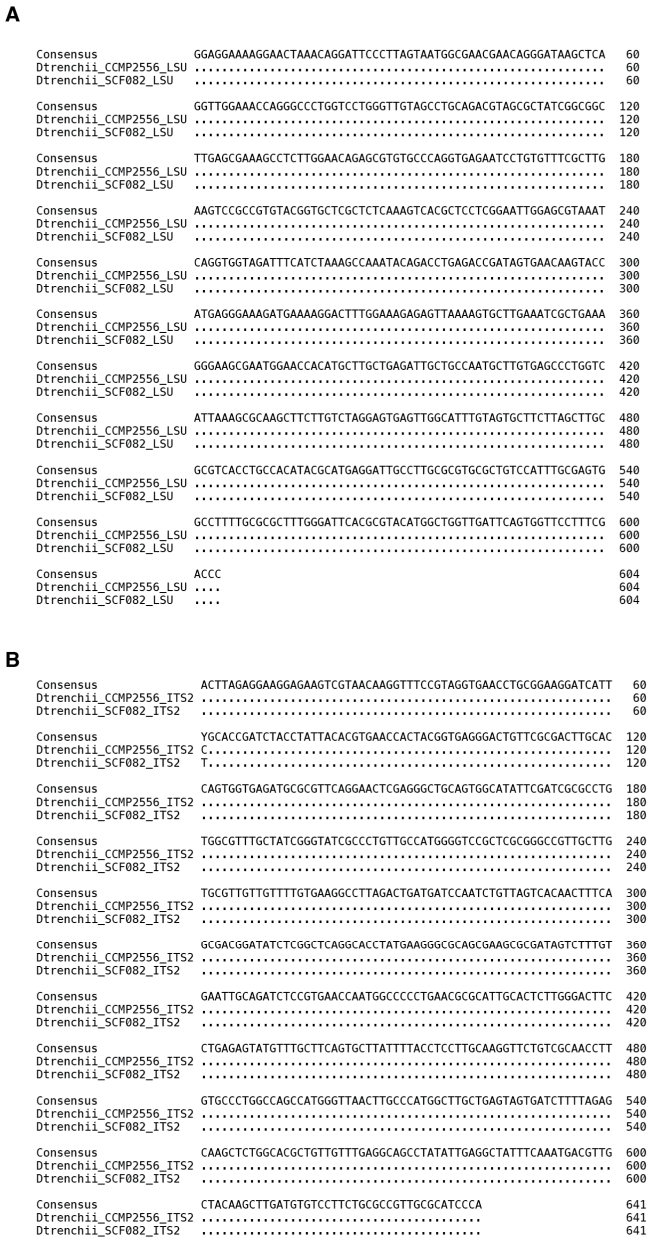
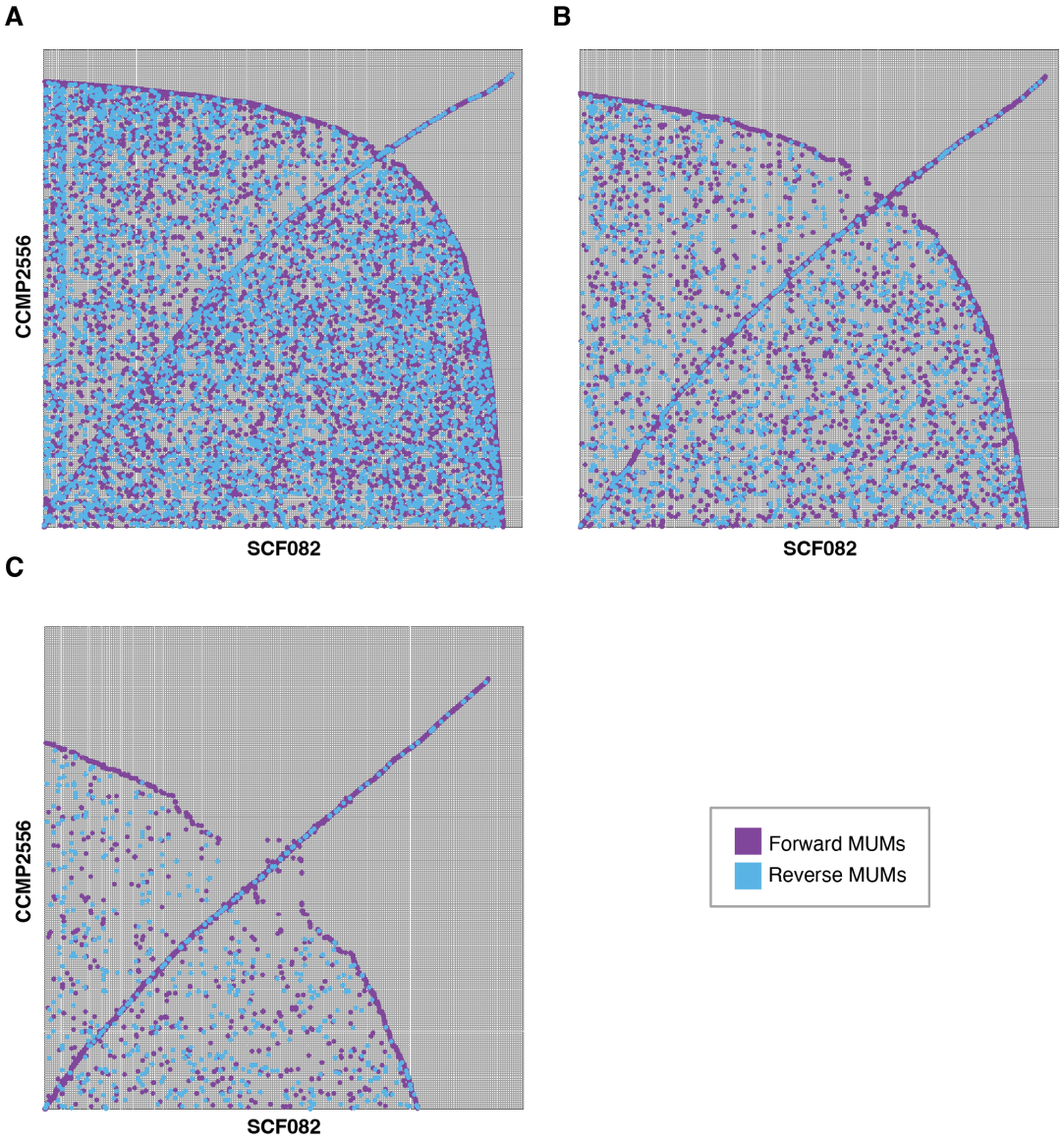


Fig. S1. Marker gene alignment.

Multiple sequence alignments for the (A) ITS2 and (B) LSU marker gene sequences for *D. trenchii* isolates CCMP2556 and SCF082.



413

414 **Fig. S2. Dotplot showing sequence similarity between the *D. trenchii* CCMP2556 and**

415 **SCF082 genome assemblies.**

416 Dotplots indicate the sequence identity across the genomes requiring an alignment identity of

417 90% and alignment lengths of (A) 100 bp (B) 1 kb and (C) 10 kb. Maximal unique matches

418 (MUMs) in the forward direction are indicated with purple and in the reverse direction blue.

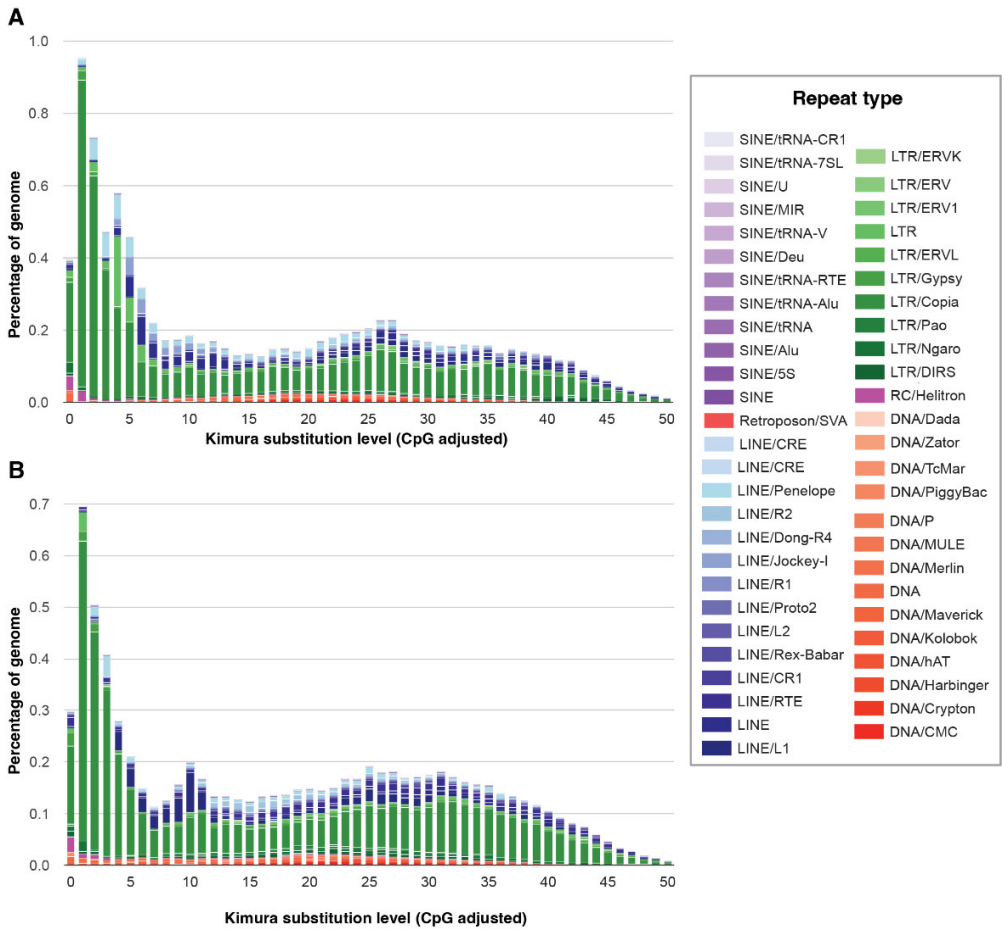


Fig. S3. Repeat landscapes for *D. trenchii* (A) CCMP2556 and (B) SCF082.

Distribution of different repeat familiess and types identified by RepeatModeler and RepeatMasker. Kimura substitution level is represented on the x-axis and the percentage of the genome length covered by each repeat type on the y-axis. Lower Kimura values represent more recently diverging sequences and higher values more anciently diverging sequences.

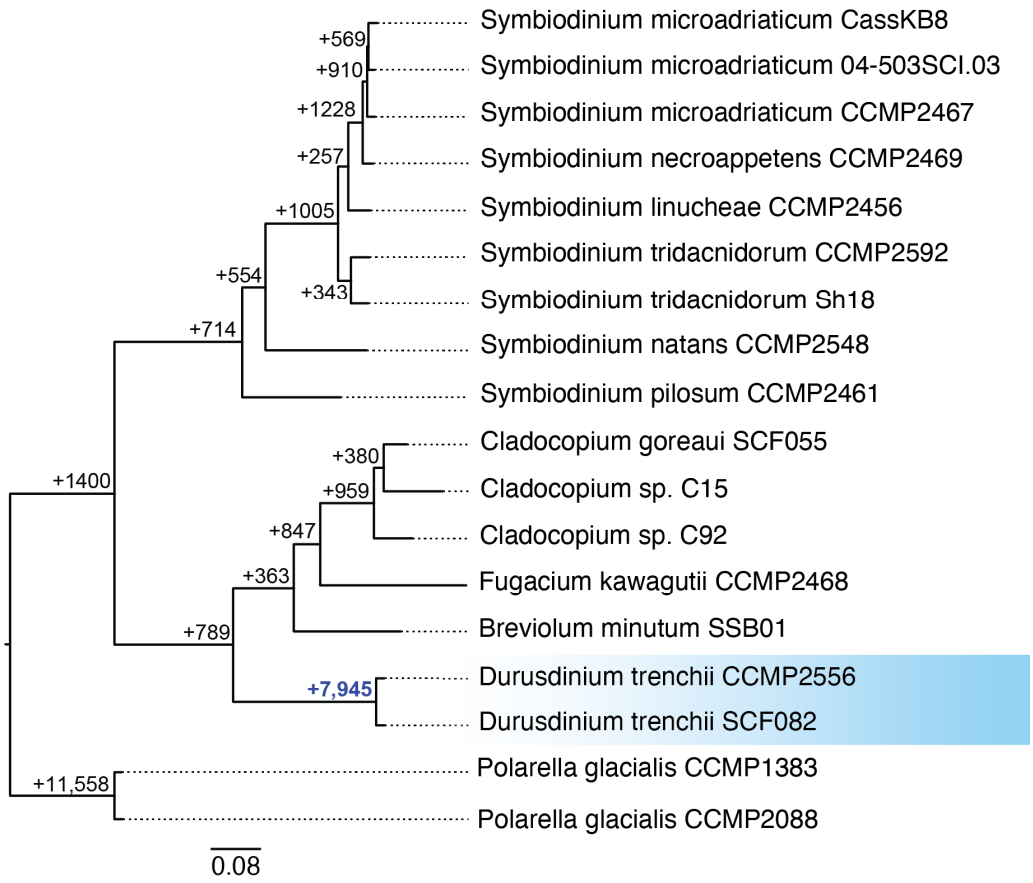
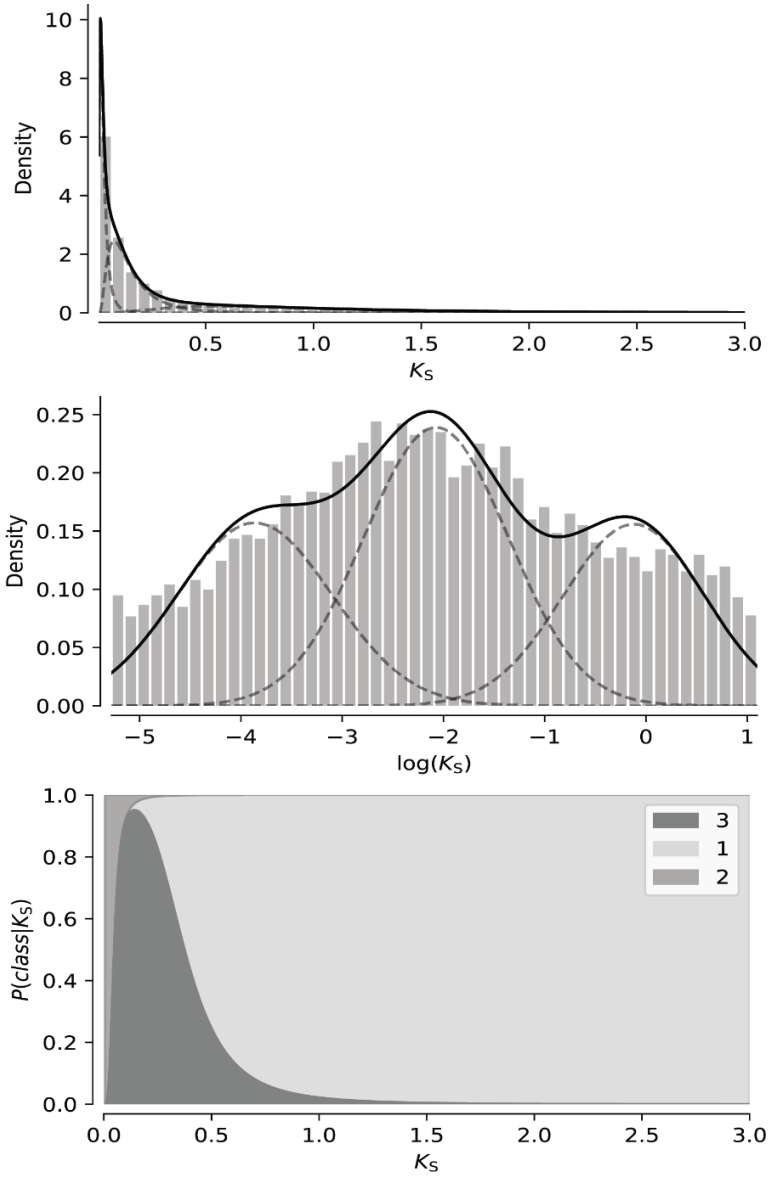


Fig. S4. Phylogenetic inference of gene duplication events in protein families for different lineages.

The number of protein families experiencing a gene duplication event is represented at each node of the representative species phylogeny as inferred by OrthoFinder2. Only clade-specific gene family duplication events across Suessiales that have at least 50% support for occurring at that node within the tree are shown.



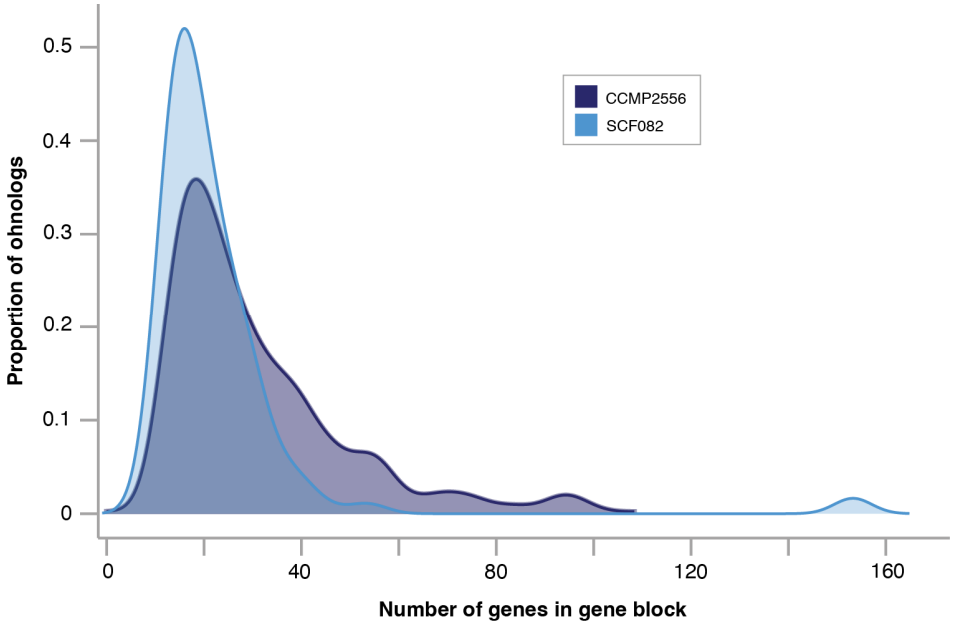
433

434 **Fig. S5. Ks plot of D. *trenchii* CCMP2556 isolate**

435 (A) Standard Ks plot for CCMP2556. (B) Log-normalized Ks plot for CCMP2556. (C)

436 Probability of genes along Ks plot falling into different categories according to the model.

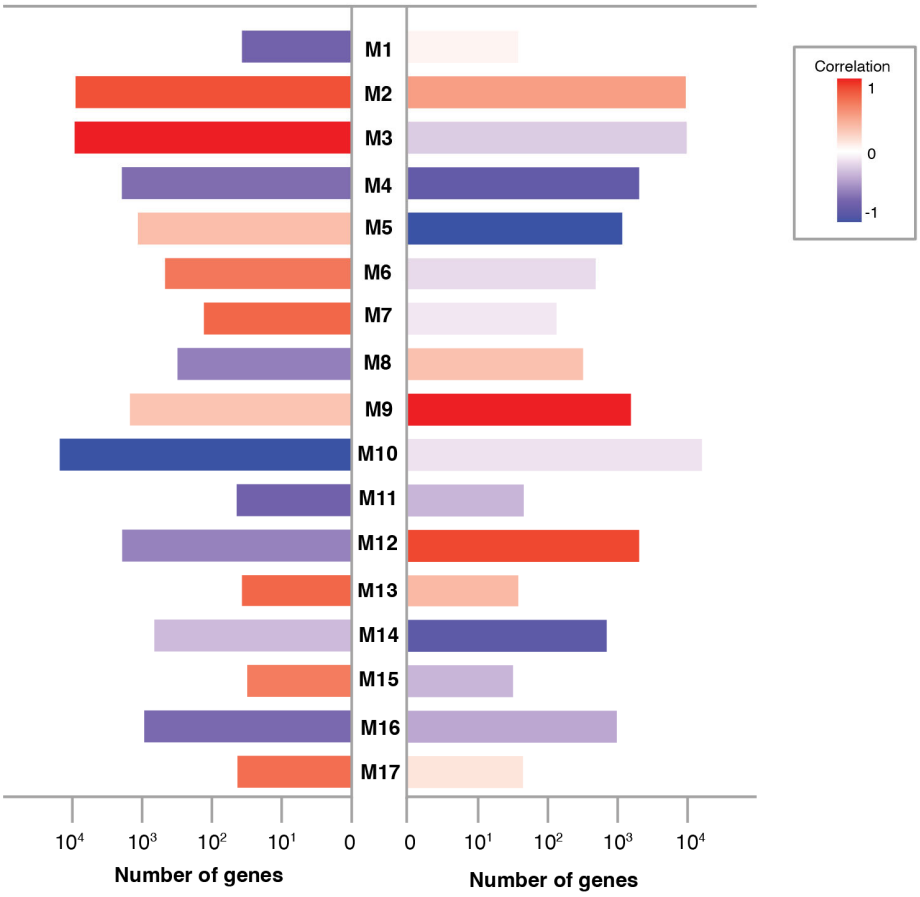
437



438

439 **Fig. S6. Number of genes in duplicated collinear gene blocks for CCMP2556 and SCF082**

440 Density plot of *D. trenchii* ohnologs relative to the size of the duplicated collinear gene block



441

442 **Fig. S7. Barplot of significant modules in the WGCNA analysis**

443 Bar heights represent the number of genes in each module. Filled in bars were statistically

444 significant ($p < 0.05$: filled bar; $p > 0.05$ outlined bar) and colors represent their Pearson

445 correlation coefficient to that variable.

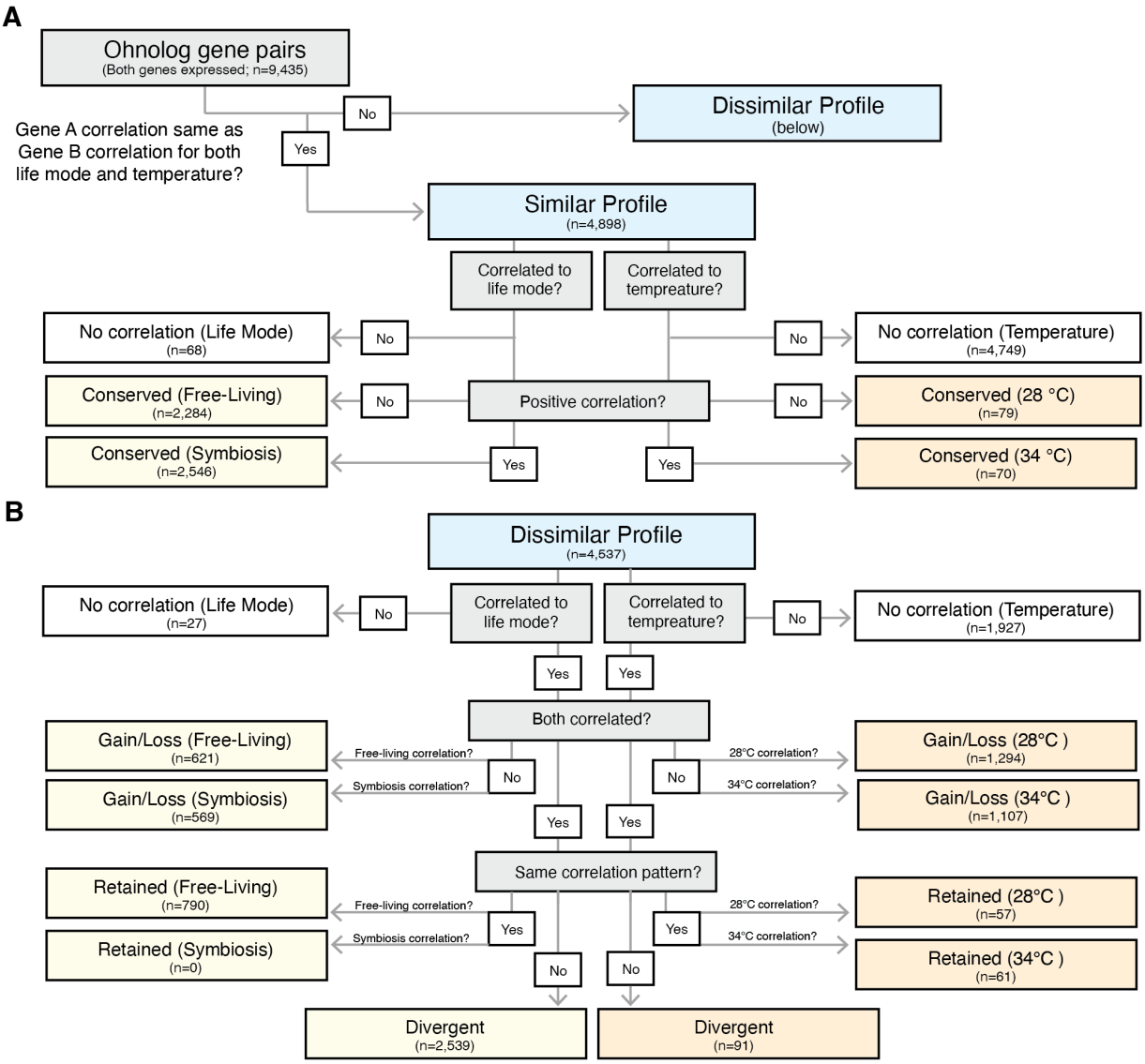
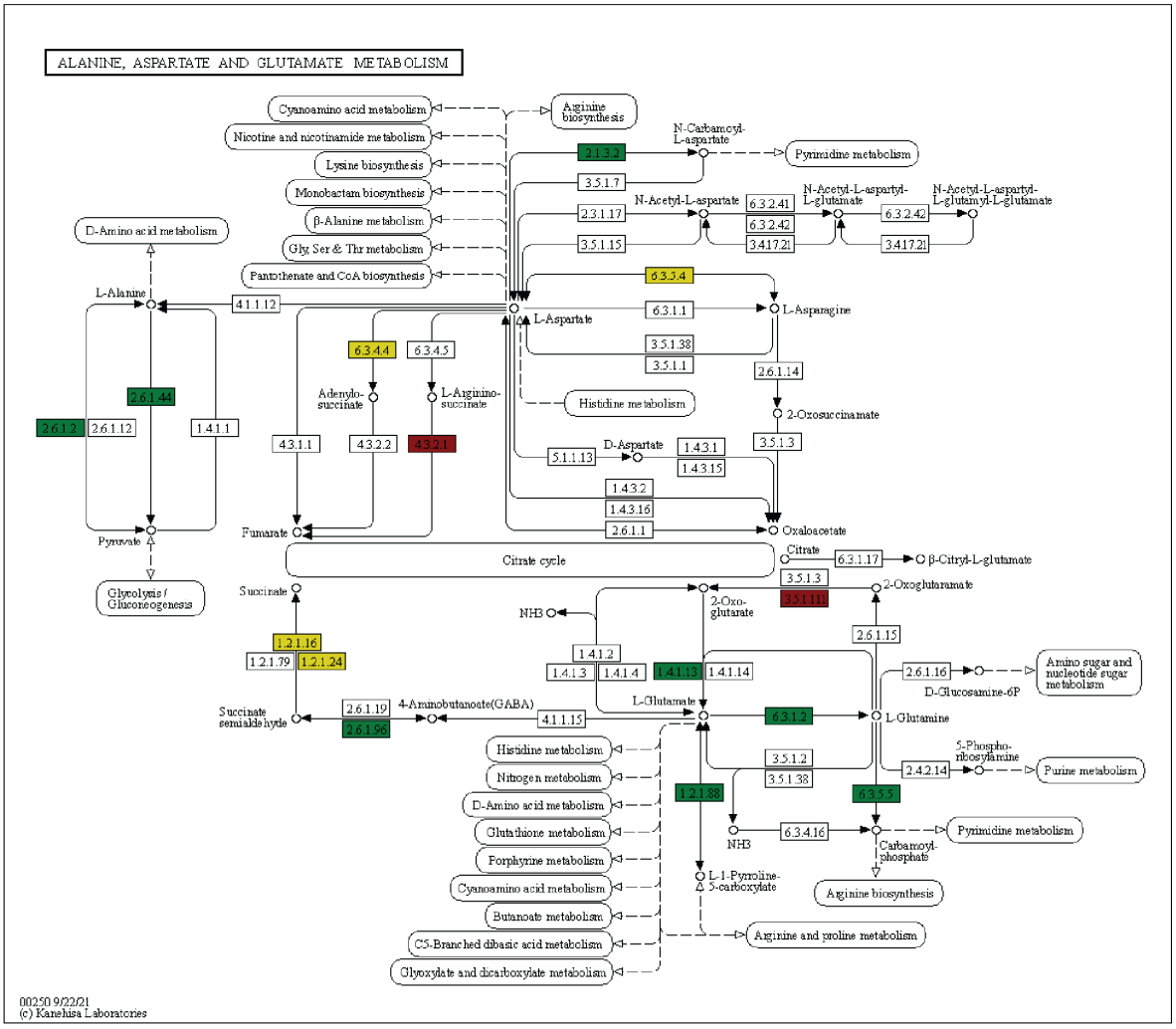


Fig. S8. Decision tree for classifying ohnolog pairs into expression categories according to their corresponding WGCNA modules.

(A) Ohnolog-pairs that maintained the same expression correlation for both life mode and temperature were designated as having "Similar Profiles", and were further classified by whether this conserved expression of the ohnolog-pair was correlated to the free-living or symbiotic life

452 modes or to 28°C or 34°C temperature. (B) Those ohnolog-pairs exhibiting some change in
453 expression profile were designated as “Dissimilar Profiles”. Those with dissimilar profiles were
454 first broken down by whether both ohnologs were correlated to life mode or temperature.
455 Ohnolog-pairs where both ohnologs were correlated one of the two factors were then further
456 classified by whether their correlations were both in favor of one either symbiosis (i.e.
457 “Retained: Symbiosis) or free-living lifestyles (i.e. “Retained: Free-Living) for life mode, or
458 34°C (i.e. “Retained: 34°C) and 28°C (i.e. “Retained: 28°C) for temperature. If they exhibitied
459 contrasting expression correlations, they were labelled as “Divergent”. If only one ohnolog
460 exhibited a correlation, they were labelled as “Gain/Loss” and with the directionality of the
461 ohnolog correlated to that factor.



Symbiosis: Conserved and Retained

Symbiosis: Conserved and Retained and Divergent

Free-Living: Conserved and Retained

Free-Living: Conserved and Retained and Divergent

Divergent

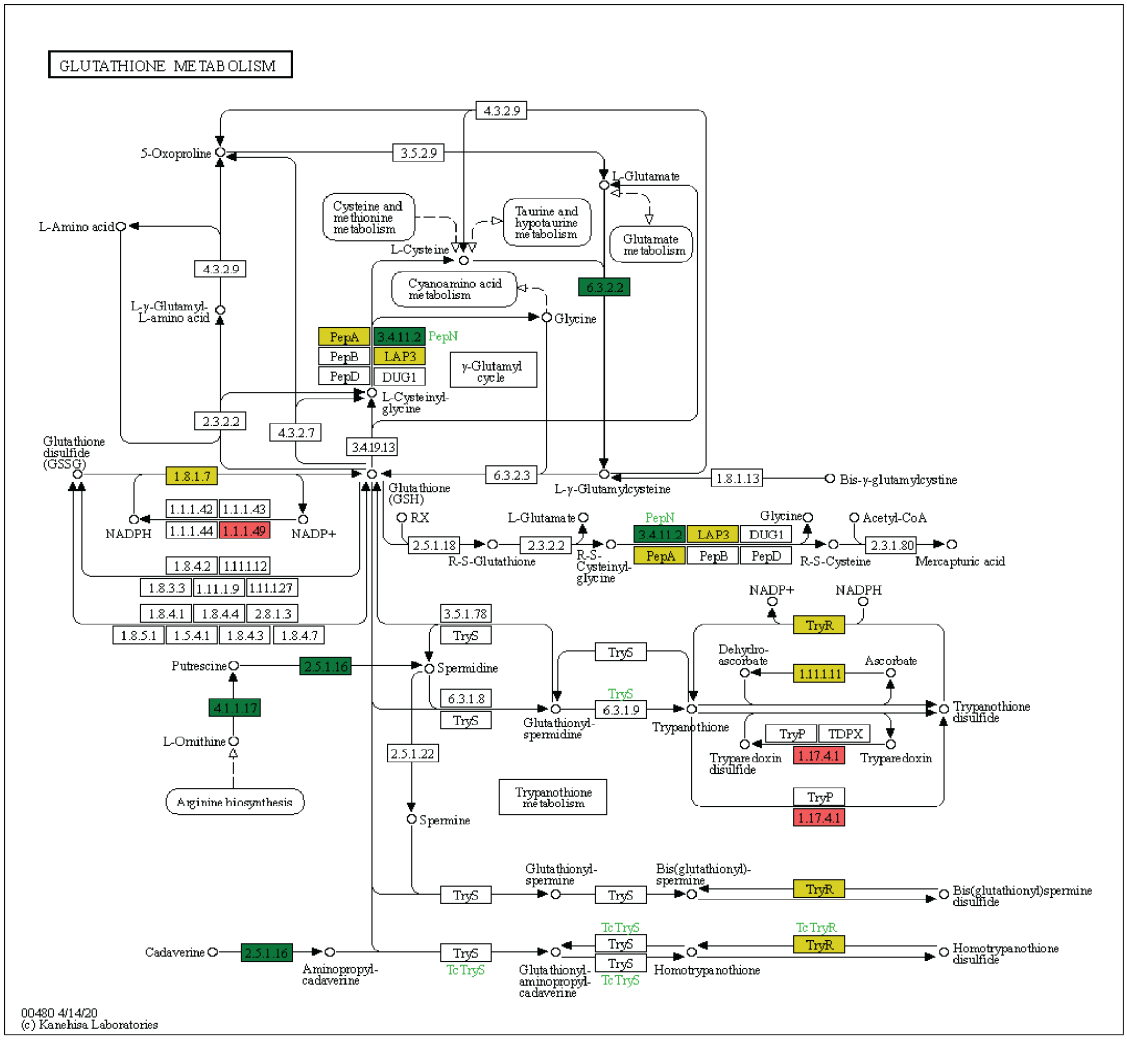
Symbiosis: Conserved and Retained and Free-Living: Conserved and Retained

462

463 **Fig. S9. KEGG pathway for Alanine, Aspartate, and Glutamate Metabolism, with colored**

464 **enzymes indicating KEGG IDs of ohnologs retained in duplicate annotated by PRIAM**

465 **from the Conserved, Retained, and Divergent WGCNA ohnolog-pair categories.**



Symbiosis: Conserved and Retained

Symbiosis: Conserved and Retained and Divergent

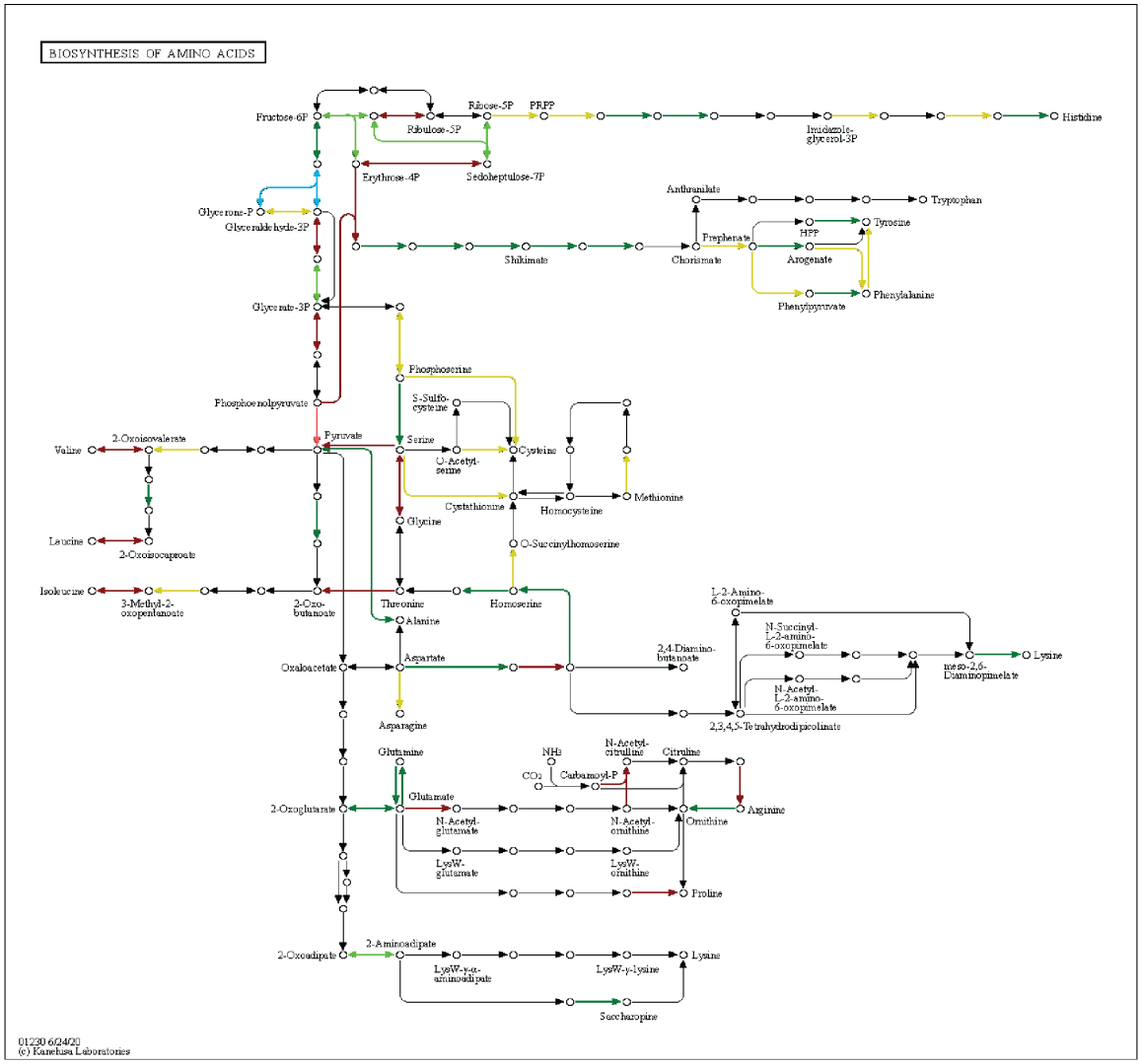
Free-Living: Conserved and Retained

Free-Living: Conserved and Retained and Divergent

Divergent

Symbiosis: Conserved and Retained and Free-Living: Conserved and Retained

Fig. S10. KEGG pathway for Glutathione Metabolism, with colored enzymes indicating KEGG IDs of ohnologs retained in duplicate annotated by PRIAM from the Conserved, Retained, and Divergent WGCNA ohnolog-pair categories.



Symbiosis: Conserved and Retained

Symbiosis: Conserved and Retained and Divergent

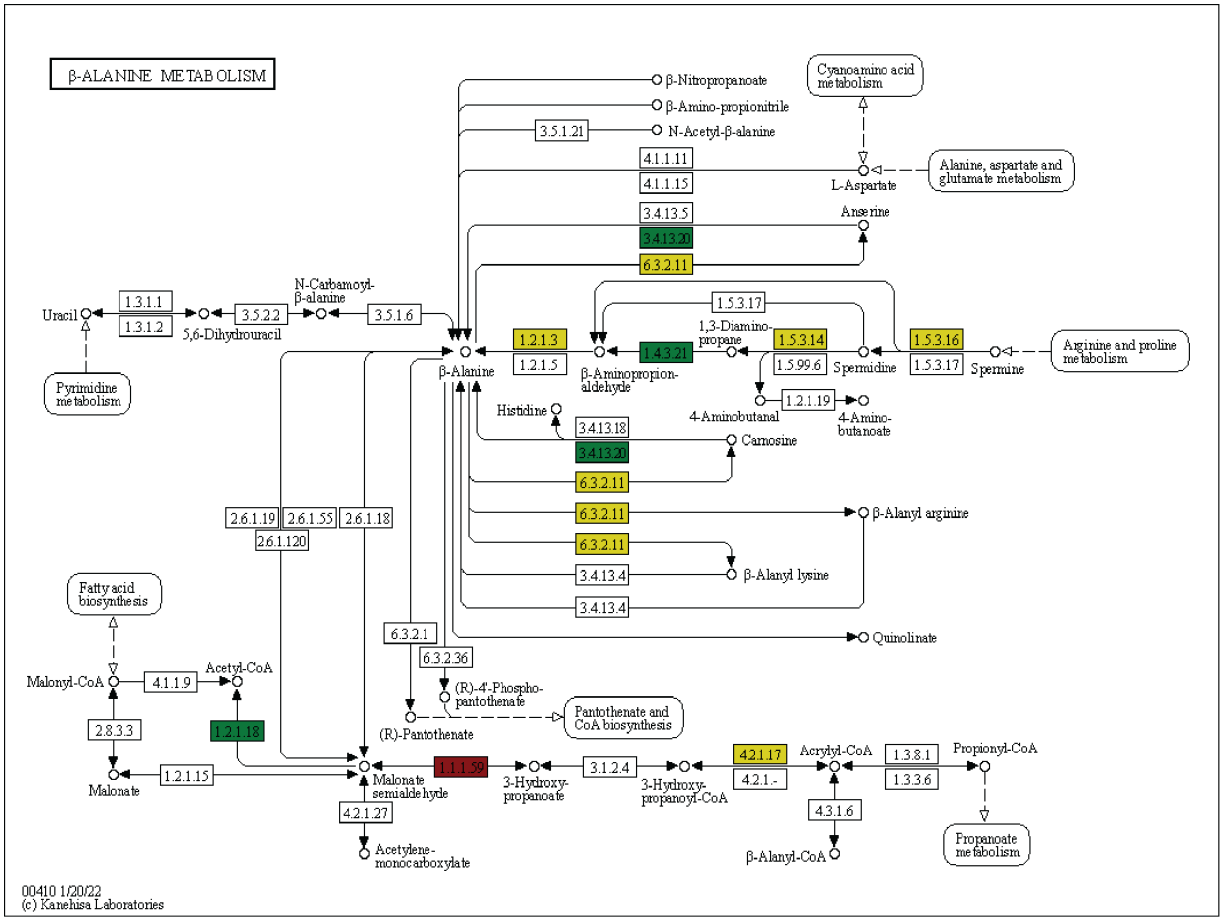
Free-Living: Conserved and Retained

Free-Living: Conserved and Retained and Divergent

Divergent

Symbiosis: Conserved and Retained and Free-Living: Conserved and Retained

Fig. S11. KEGG pathway for Biosynthesis of Amino Acids, with colored enzymes indicating KEGG IDs of ohnologs retained in duplicate annotated by PRIAM from the Conserved, Retained, and Divergent WGCNA ohnolog-pair categories.



Symbiosis: Conserved and Retained

Symbiosis: Conserved and Retained and Divergent

Free-Living: Conserved and Retained

Free-Living: Conserved and Retained and Divergent

Divergent

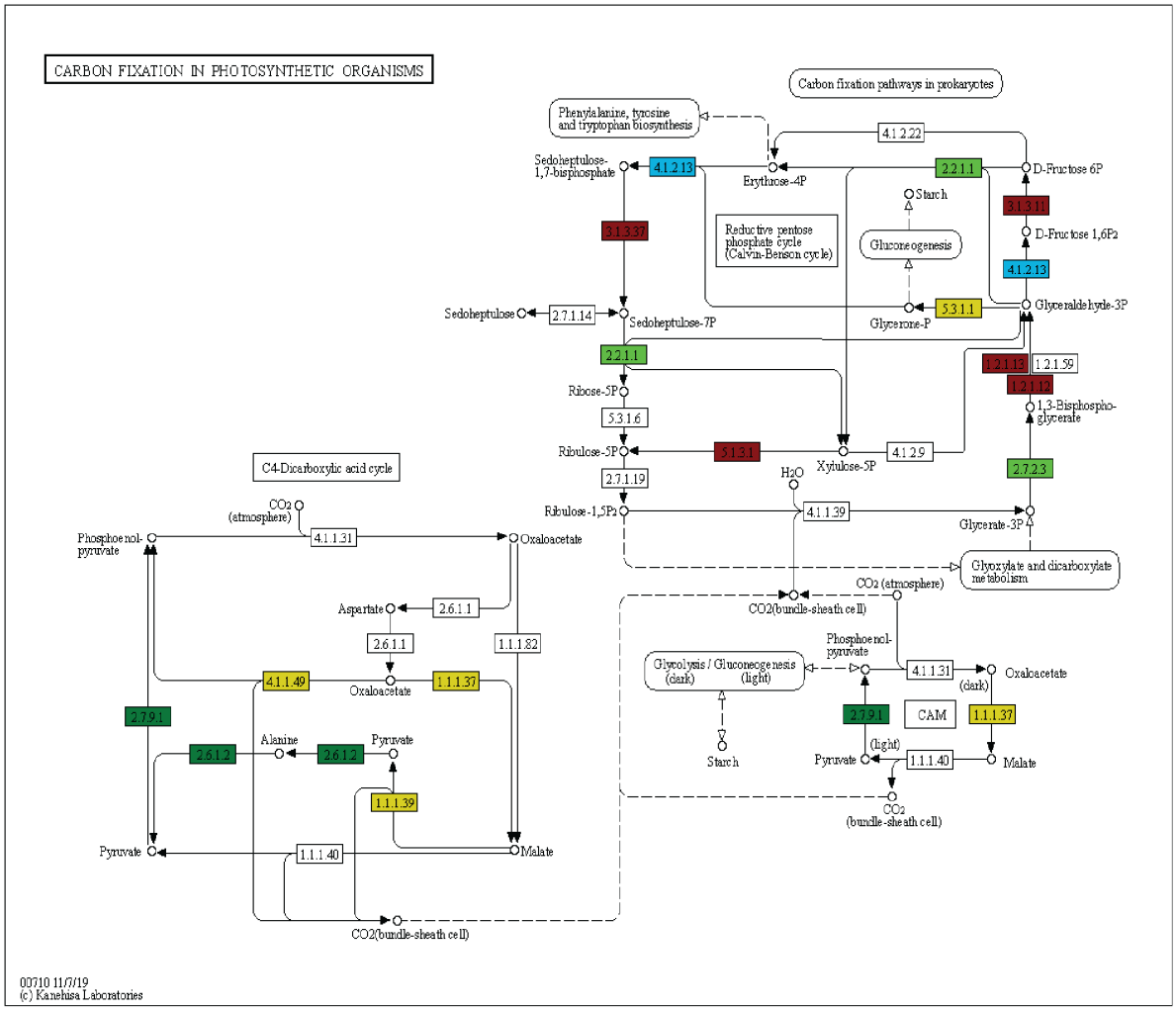
Symbiosis: Conserved and Retained and Free-Living: Conserved and Retained

475

476 **Fig. S12. KEGG pathway for β-Alanine Metabolism, with colored enzymes indicating**

477 **KEGG IDs of ohnologs retained in duplicate annotated by PRIAM from the Conserved,**

478 **Retained, and Divergent WGCNA ohnolog-pair categories.**



Symbiosis: Conserved and Retained

Symbiosis: Conserved and Retained and Divergent

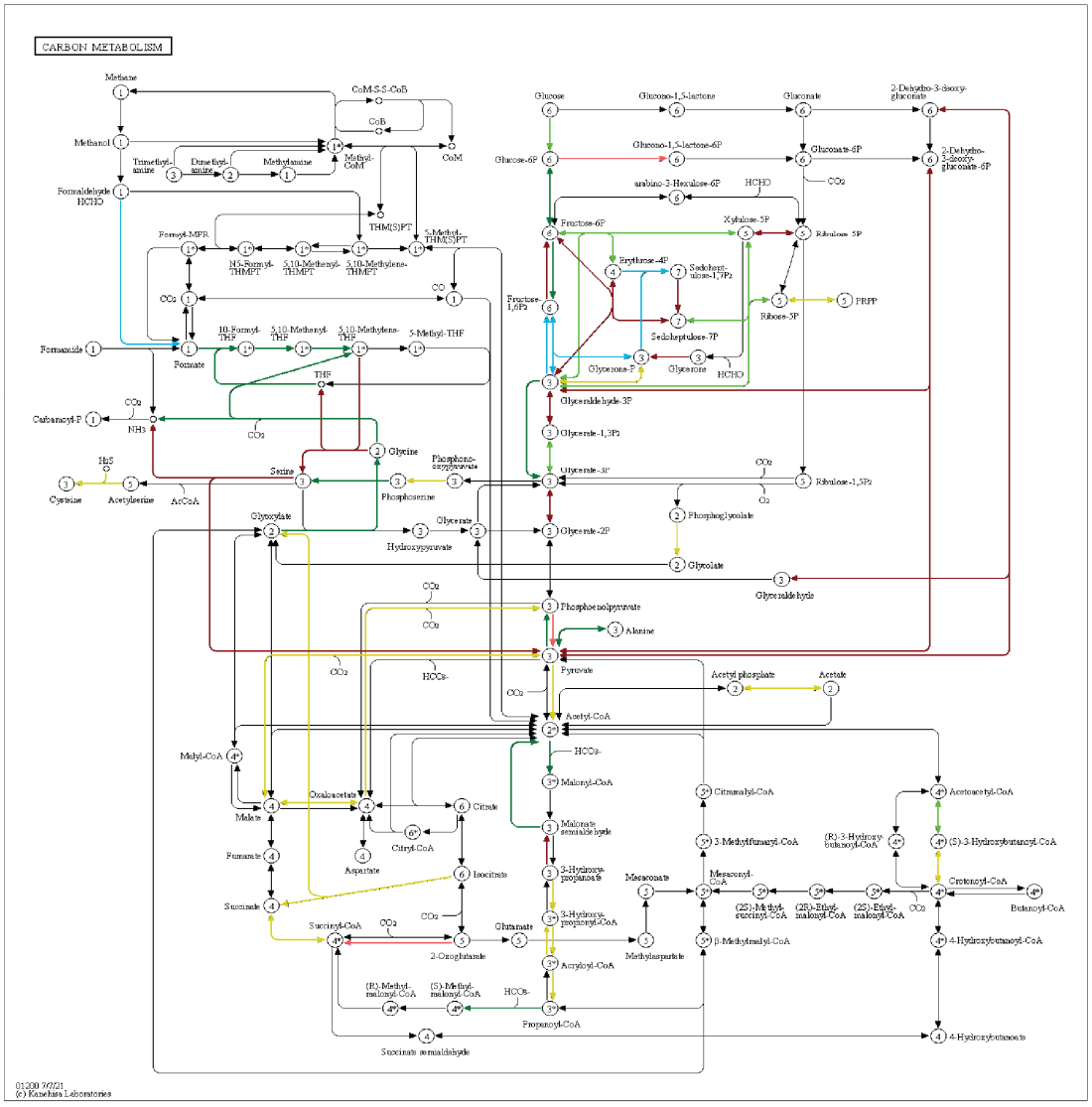
Free-Living: Conserved and Retained

Free-Living: Conserved and Retained and Divergent

Divergent

Symbiosis: Conserved and Retained and Free-Living: Conserved and Retained

Fig. S13. KEGG pathway for Carbon Fixation in Photosynthetic Organisms, with colored enzymes indicating KEGG IDs of ohnologs retained in duplicate annotated by PRIAM from the Conserved, Retained, and Divergent WGCNA ohnolog-pair categories.



Symbiosis: Conserved and Retained

Symbiosis: Conserved and Retained and Divergent

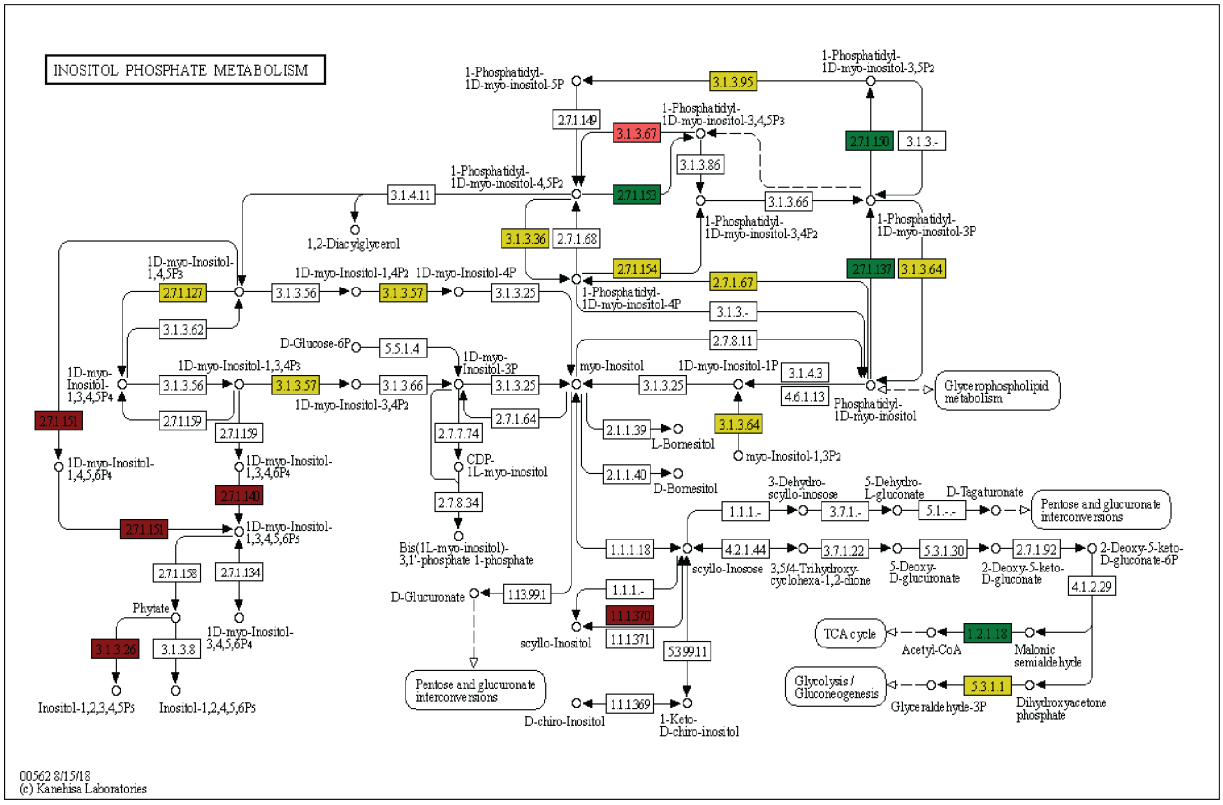
Free-Living: Conserved and Retained

Free-Living: Conserved and Retained and Divergent

Divergent

Symbiosis: Conserved and Retained and Free-Living: Conserved and Retained

Fig. S14. KEGG pathway for Carbon Metabolism, with colored enzymes indicating KEGG IDs of ohnologs retained in duplicate annotated by PRIAM from the Conserved, Retained, and Divergent WGCNA ohnolog-pair categories.



Symbiosis: Conserved and Retained

Symbiosis: Conserved and Retained and Divergent

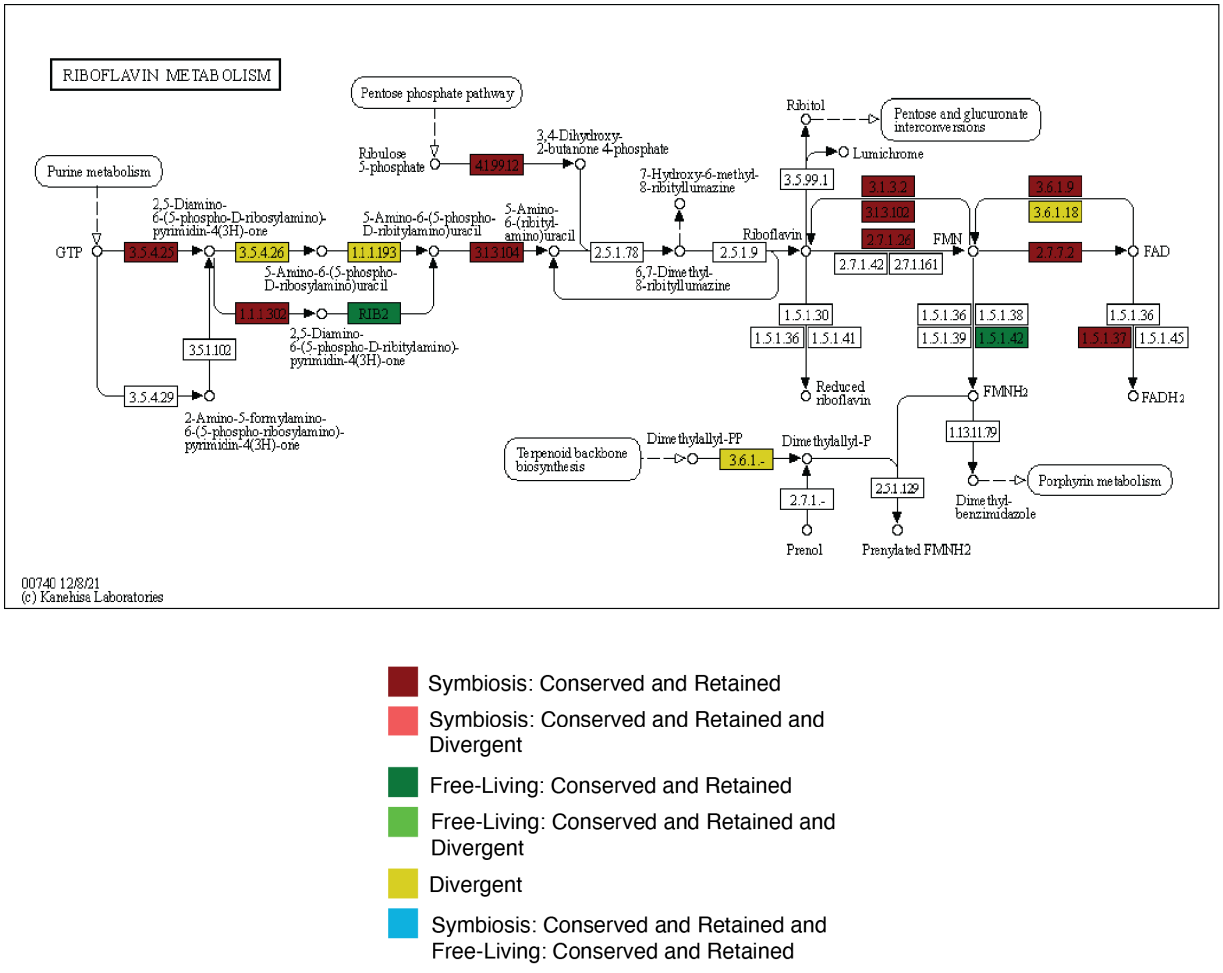
Free-Living: Conserved and Retained

Free-Living: Conserved and Retained and Divergent

Divergent

Symbiosis: Conserved and Retained and Free-Living: Conserved and Retained

Fig. S15. KEGG pathway for Inositol Phosphate Metabolism, with colored enzymes indicating KEGG IDs of ohnologs retained in duplicate annotated by PRIAM from the Conserved, Retained, and Divergent WGCNA ohnolog-pair categories.



Symbiosis: Conserved and Retained

Symbiosis: Conserved and Retained and Divergent

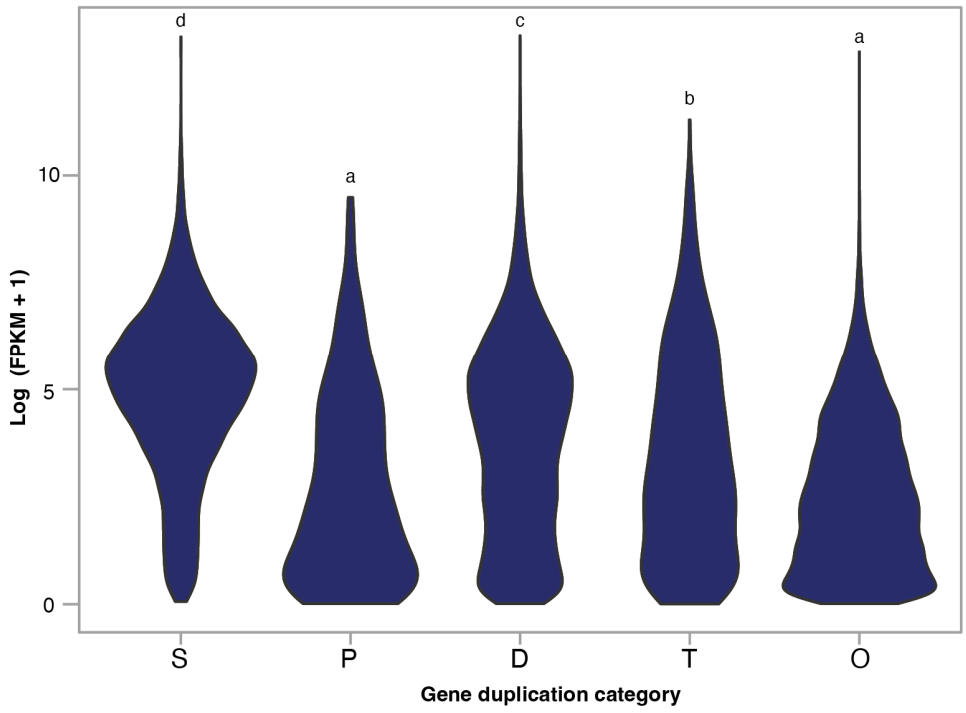
Free-Living: Conserved and Retained

Free-Living: Conserved and Retained and Divergent

Divergent

Symbiosis: Conserved and Retained and Free-Living: Conserved and Retained

Fig. S16. KEGG pathway for Riboflavin Metabolism, with colored enzymes indicating KEGG IDs of ohnologs retained in duplicate annotated by PRIAM from the Conserved, Retained, and Divergent WGCNA ohnolog-pair categories.



495

496 **Fig. S17. Gene expression violin plots across gene duplication categories.**

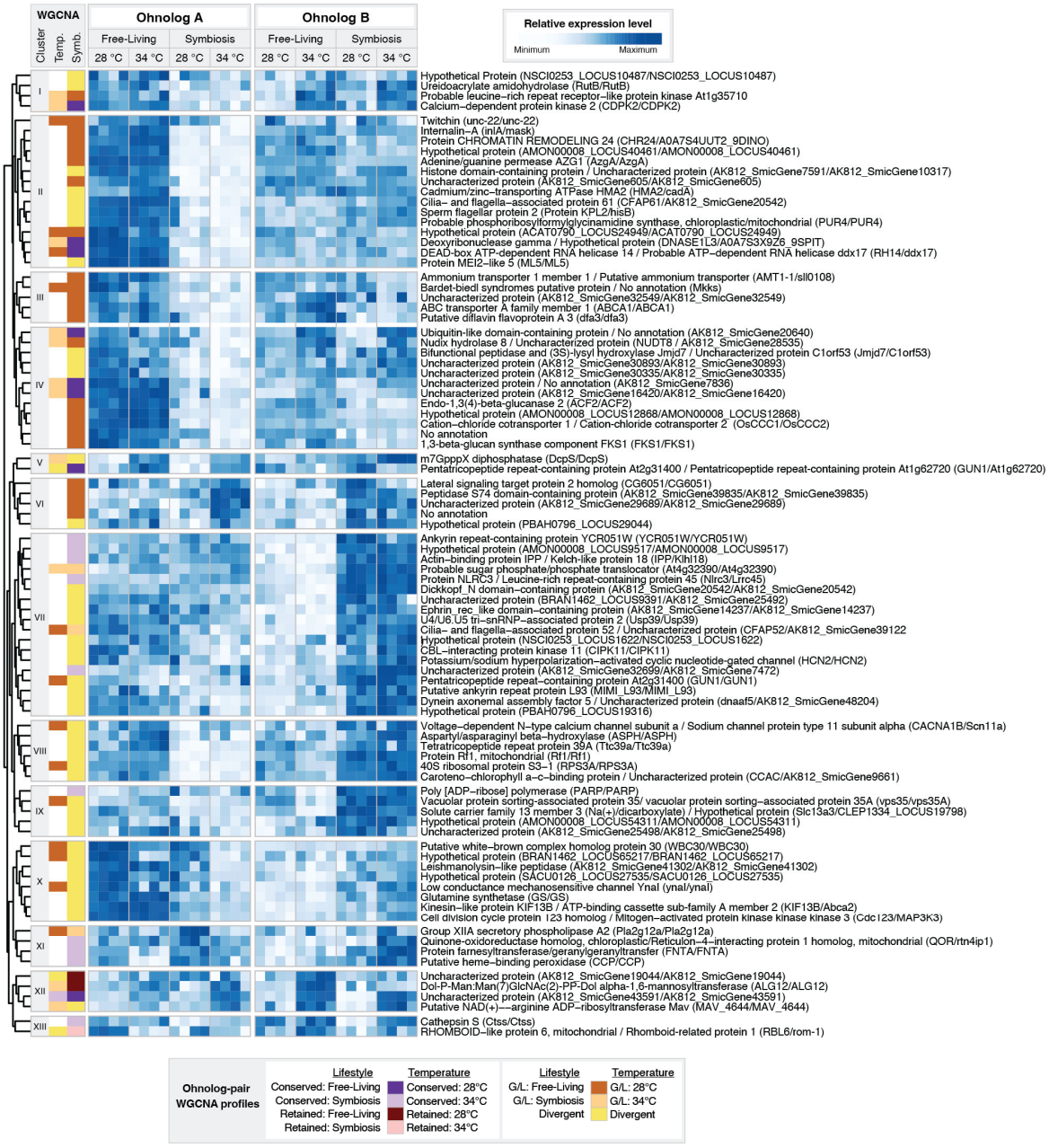
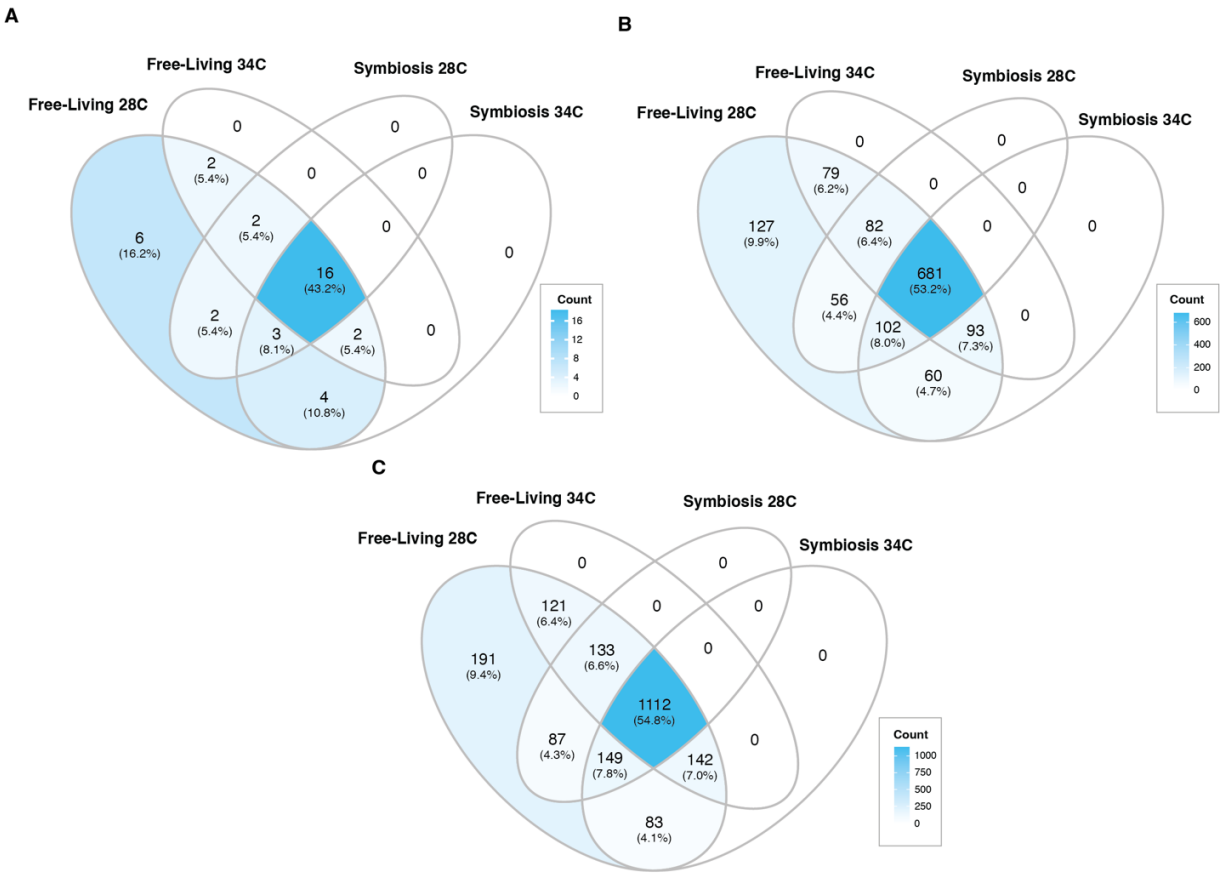


Fig. S18. Gene expression heatmap of the 90 ohnolog-pairs exhibiting expression dominance in different growth conditions.

Ohnolog-pairs are clustered according to their expression patterns using Euclidean distancesand expresson levels scaled within each row from the minimum to maximum value to allow comparison of expression level between the two ohnologs in a pair.



503

504 **Fig. S19. Summary of consensus mRNA-edited sites in CCMP2556 across the four RNAseq**

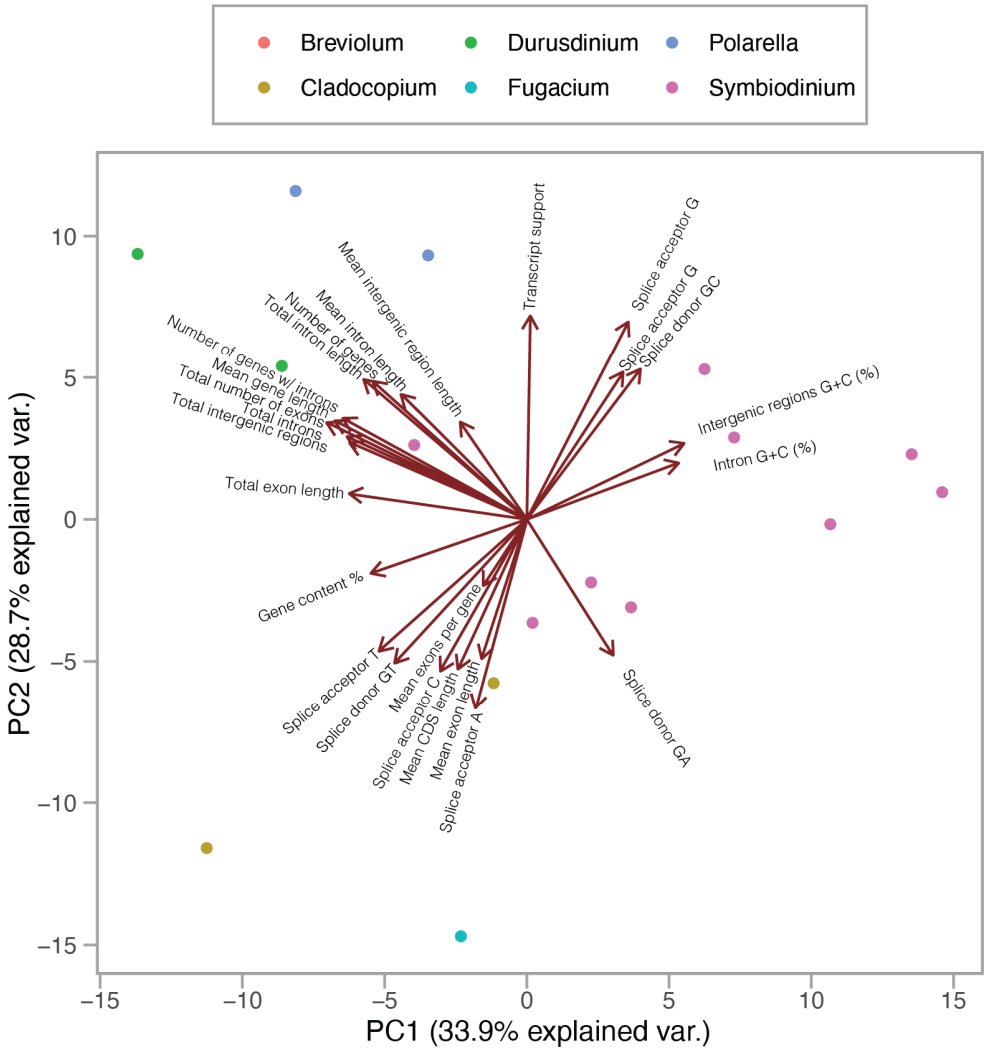
505 **treatments.**

506 The numbers represent the number of distinct edited sites for those that have a (A) high impact

507 (ie. stop gained, stop lost, start lost) on gene function (B) moderate (i.e. missense edit) impact on

508 gene function, and (C) all consensus edited sites including synonymous changes.

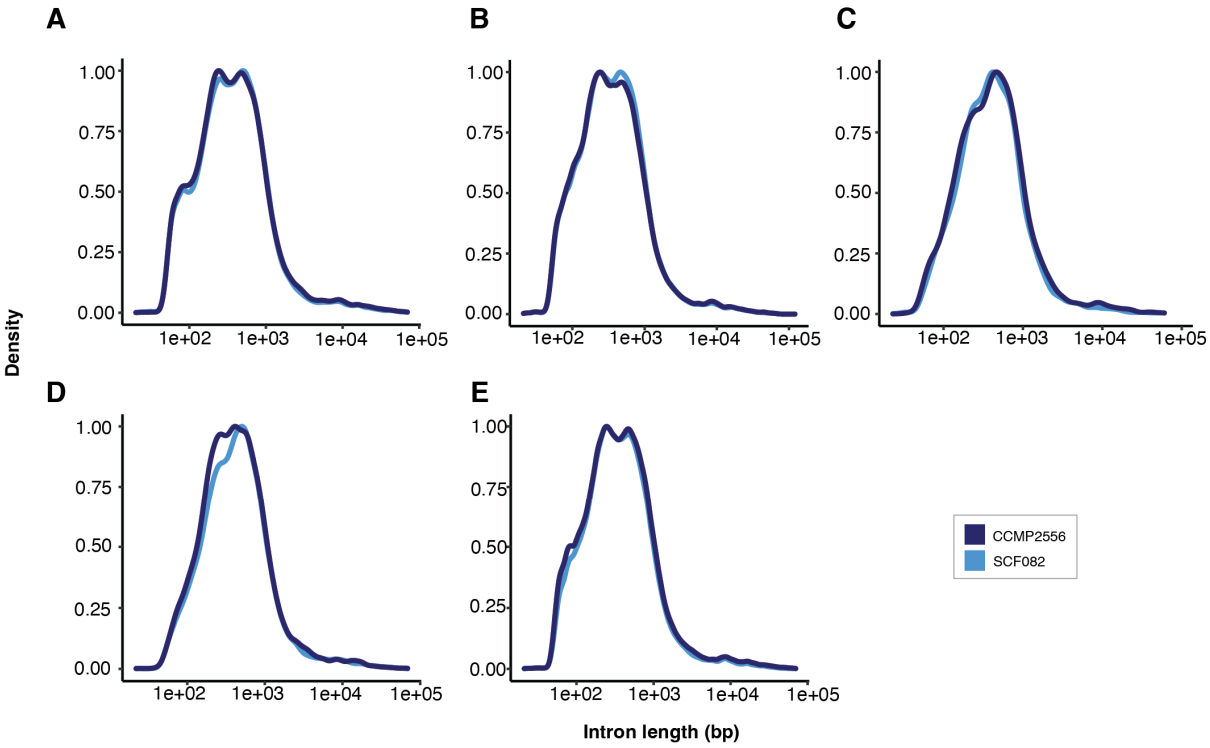
509



510

511 **Fig. S20. Principal component analysis of metrics of the gene model predictions within**

512 **Suessiales.**



513

514 **Fig. S21. Intron length distribution across gene duplication categories.**

515 The intron length distribution across (A) singletons (B) proximal duplicates (C) dispersed

516 duplicates (D) tandem duplicates and (E) ohnologs.

517

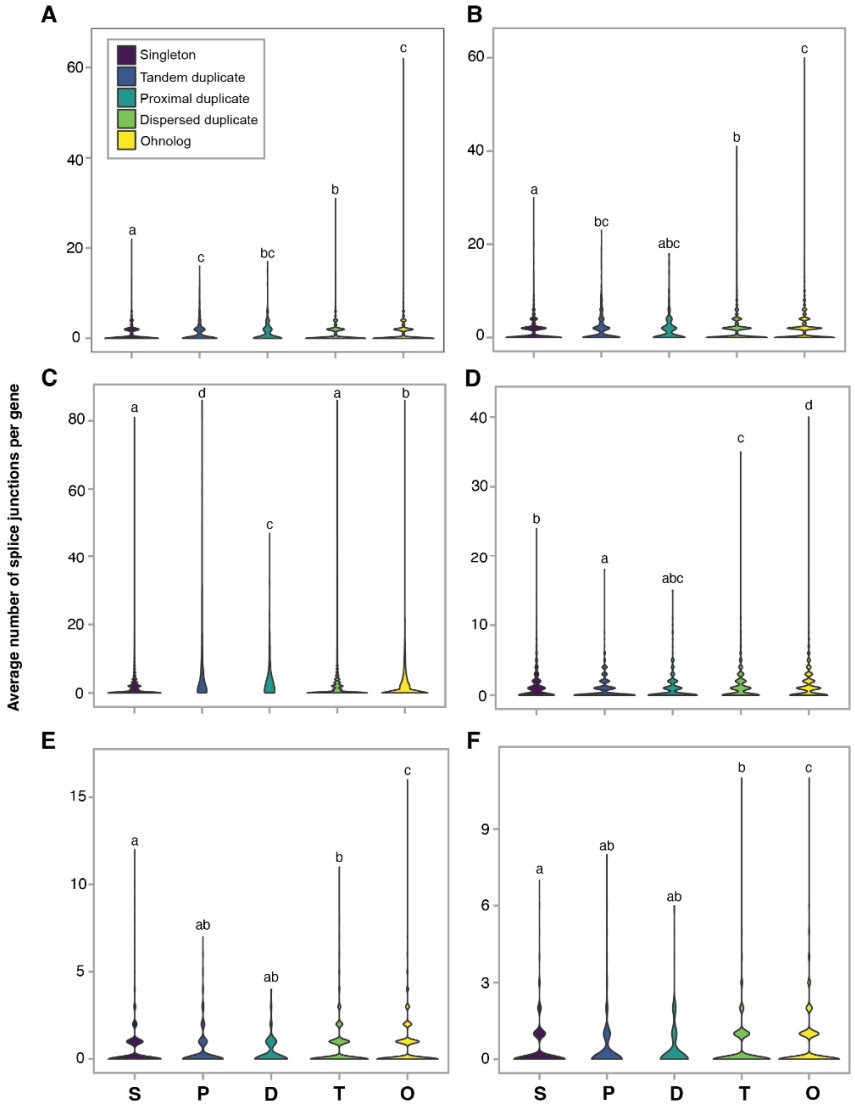
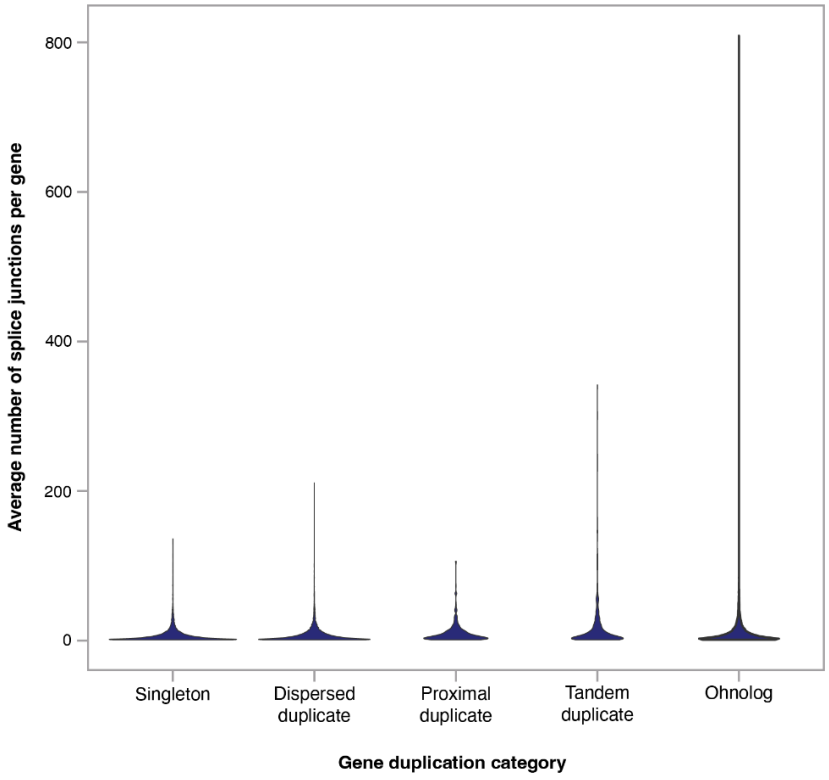


Fig. S22. Violin plots of the average number of splice junctions per gene across the gene duplication categories.

The average number of splice junctions per gene is shown for the splice junction types (A) alternate acceptor (B) alternate donor (C) alternate exon (D) retained intron (E) intron start and (F) intron end. Statistically different groups identified through a Kruskal-Willis post-hoc test are identified with letters above the plot.



525

526 **Fig. S23. Splice site numbers per gene across duplication categories**

527

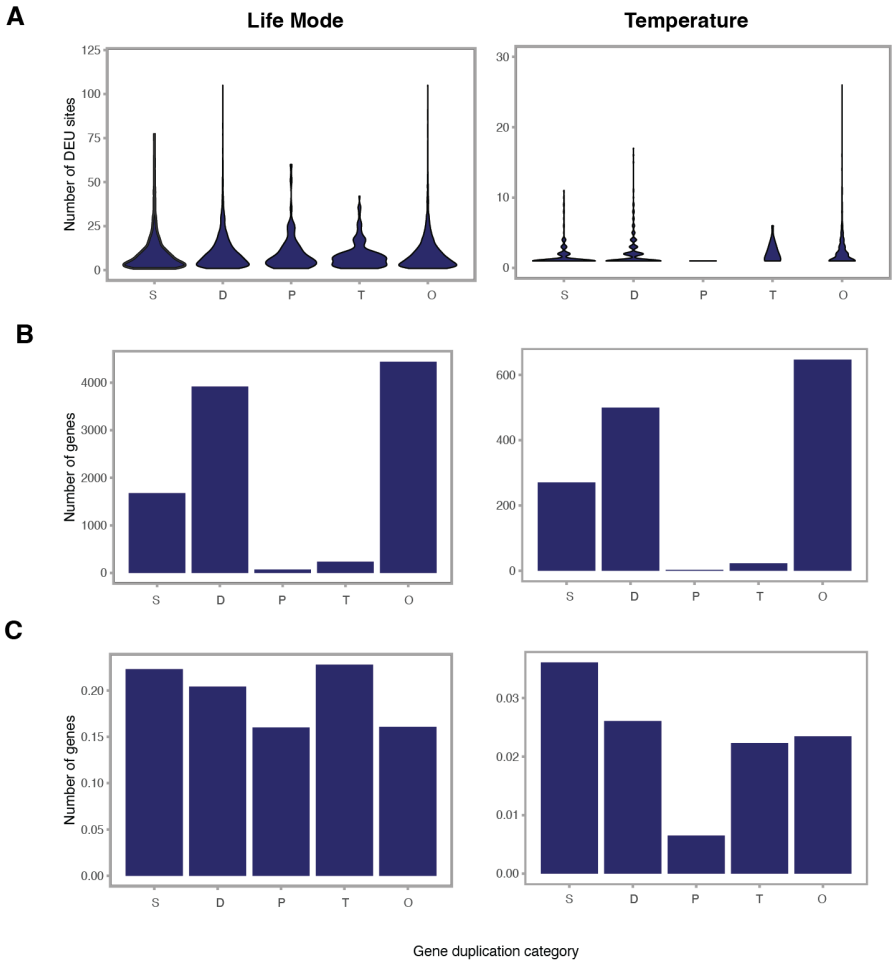
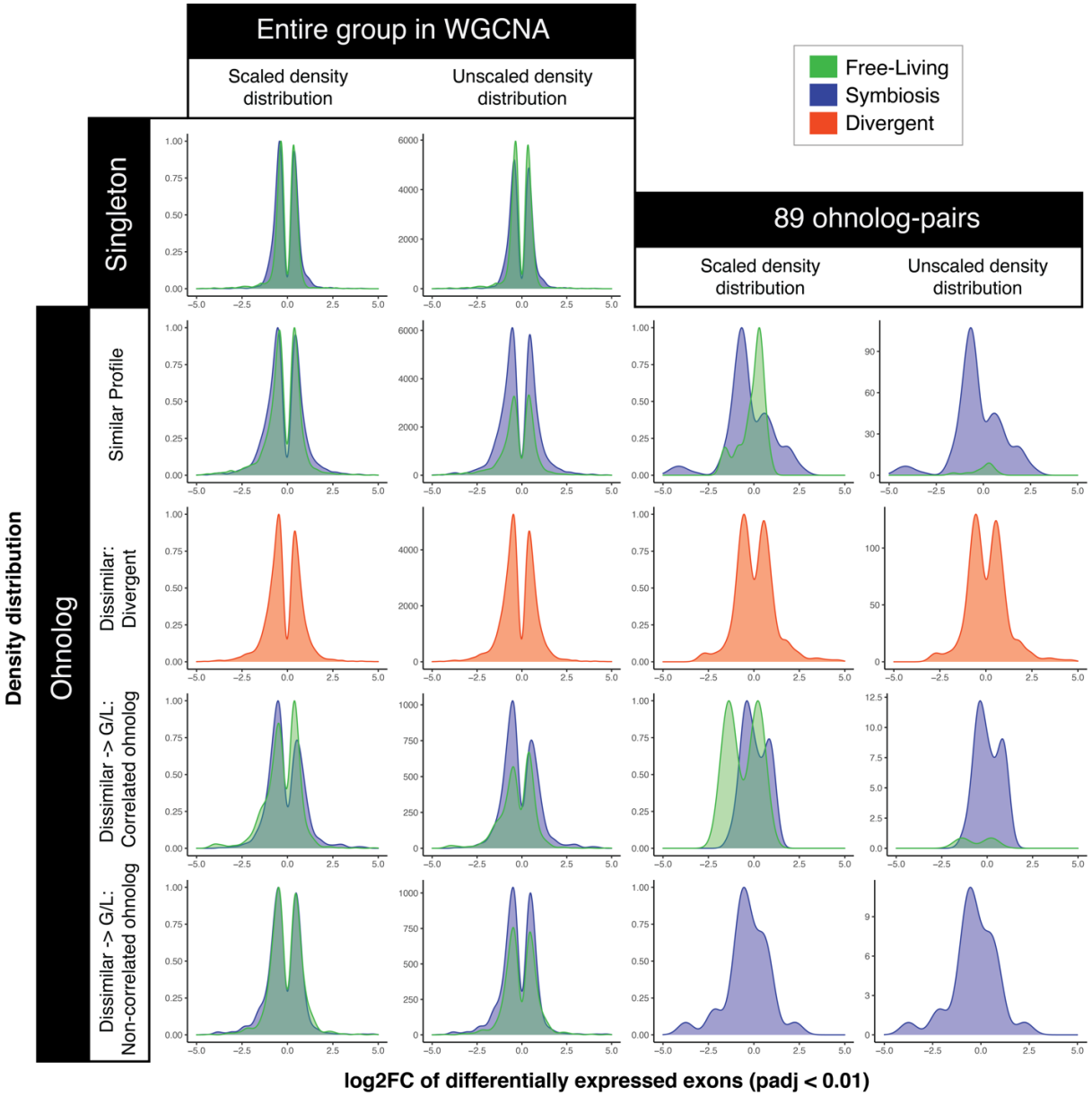


Fig. S24. CCMP2556 Differential Exon Usage (DEU) summary

Summary of the DEU results for both Life Mode and Temperature across the gene duplication categories: Singleton, Dispersed duplicate, Proximal duplicate, Tandem duplicate, and Ohnolog. representing the number of genes (FDR < 0.05) from each duplication category of Singleton, Dispersed duplicate, Proximal duplicate, Tandem duplicate, and Ohnolog exhibiting at least one site with DEU (adjsted p < 0.05) in terms of (A) total genes and (B) proportion of genes. (C) Violin plot of the number of significant DEU sites per gene.



536

537 **Fig. S25. Density distribution of differentially used exons across different gene duplication**

538 **categories**

539 Scaled and unscaled density plots of the log2FC of differentially used exons ($p < 0.001$) for (i)

540 singletons, (ii) all "Conserved" and "Retained" in context of lifestyle, (iii) the ohnolog from

541 "Gain/Loss" pairs exhibiting a correlation to lifestyle, and (iv) all ohnologs exhibiting a

correlation to a particular lifestyle from the 90 divergent ohnolog-pairs. Directionality of the log2FC change is indicated along the x-axis, with different colors indicating their gene-level correlation to either the free-living (green) or symbiotic (purple) lifestyles.

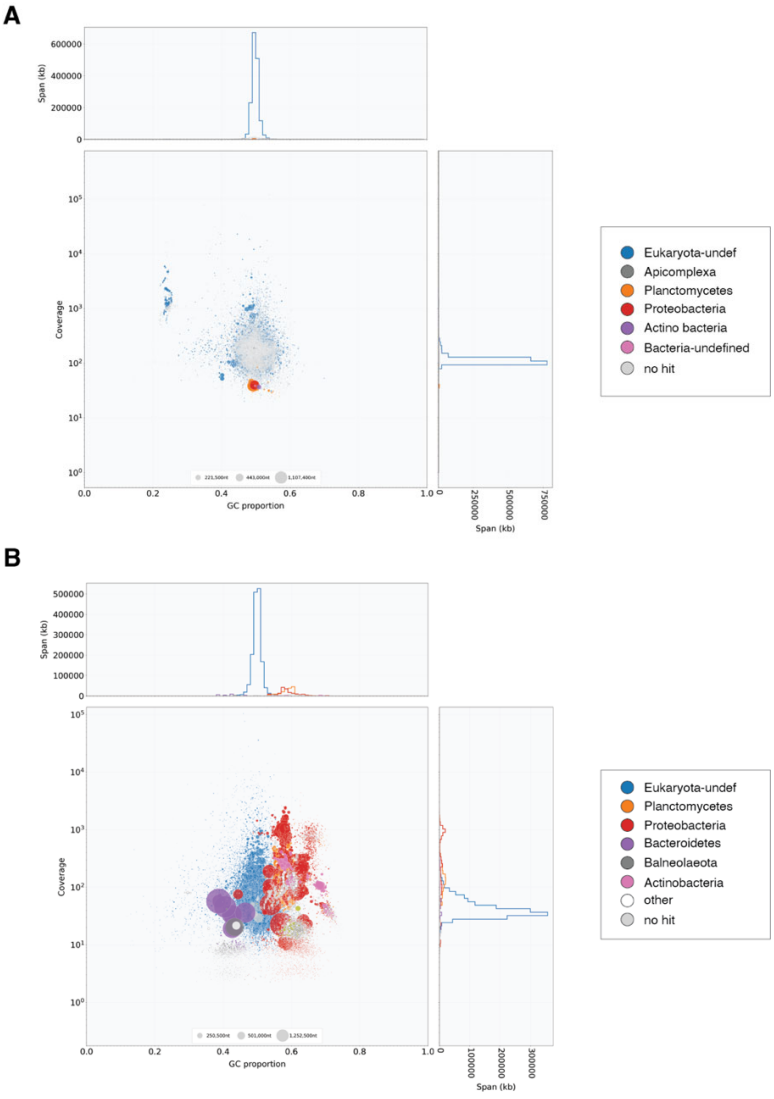


Fig. S26. Blobtools taxon-annotated GC-coverage plots for *D. trenchii* (A) CCMP2556 and (B) SCF082 genome assemblies.

Each dot represents a scaffold with read coverage on the y-axis and percent GC along the x-axis. The size of each dot corresponds to the length of the scaffold and the color to the scaffold's taxa annotation from BLAST.

References:

33. C. Iha *et al.*, Genomic adaptations to an endolithic lifestyle in the coral-associated alga *Ostreobium*. *Current Biology* **31**, 1393-1402. e1395 (2021).

34. D. R. Laetsch, M. L. Blaxter, BlobTools: Interrogation of genome assemblies. *F1000Research* **6**, 1287-1287 (2017).

35. P. Marks *et al.*, Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Research* **29**, 635-645 (2019).

36. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).

37. W. Xue *et al.*, L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC Genomics* **14**, 1-14 (2013).

38. M. Hiltunen, M. Ryberg, H. Johannesson, ARBitR: an overlap-aware genome assembly scaffolder for linked reads. *Bioinformatics* **37**, 2203-2205 (2021).

39. A. L. Delcher, S. L. Salzberg, A. M. Phillippy, Using MUMmer to identify similar regions in large sequence sets. *Current Protocols in Bioinformatics*, 10.13. 11-10.13. 18 (2003).

40. Y. Chen, R. A. González-Pech, T. G. Stephens, D. Bhattacharya, C. X. Chan, Evidence that inconsistent gene prediction can mislead analysis of dinoflagellate genomes. *Journal of Phycology* **56**, 6-10 (2020).

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

42. D. M. Emms, S. Kelly, OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**, 157-157 (2015).

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

44. A. Zwaenepoel, Y. Van de Peer, wgd—simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* **35**, 2153-2155 (2019).

45. A. J. Enright, S. Van Dongen, C. A. Ouzounis, Graph clustering by flow simulation. *Nucleic Acids Research* **30**, 1575-1584 (2000).

46. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772-780 (2013).

47. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

48. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586-1591 (2007).

49. E. Shoguchi *et al.*, A new dinoflagellate genome illuminates a conserved gene cluster involved in sunscreen biosynthesis. *Genome Biology and Evolution* **13**, evaa235 (2021).

50. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).

51. D. J. Suggett, D. J. Smith, Coral bleaching patterns are the outcome of complex biological and environmental networking. *Global Change Biology* **26**, 68-79 (2020).

52. T. P. Hughes *et al.*, Emergent properties in the responses of tropical corals to recurrent climate extremes. *Current Biology* **31**, 5393-5399. e5393 (2021).

53. T. P. Hughes *et al.*, Spatial and temporal patterns of mass bleaching of corals in the Anthropocene. *Science* **359**, 80-83 (2018).

54. R. N. Silverstein, R. Cunning, A. C. Baker, Tenacious D: *Symbiodinium* in clade D remain in reef corals at both high and low temperature extremes despite impairment. *Journal of Experimental Biology* **220**, 1192-1196 (2017).

55. R. Cunning, P. Gillette, T. Capo, K. Galvez, A. C. Baker, Growth tradeoffs associated with thermotolerant symbionts in the coral *Pocillopora damicornis* are lost in warmer oceans. *Coral Reefs* **34**, 155-160 (2015).

56. R. N. Silverstein, R. Cunning, A. C. Baker, Change in algal symbiont communities after bleaching, not prior heat exposure, increases heat tolerance of reef corals. *Global Change Biology* **21**, 236-249 (2015).

57. A. Nand *et al.*, Genetic and spatial organization of the unusual chromosomes of the dinoflagellate *Symbiodinium microadriaticum*. *Nature Genetics* **53**, 618-629 (2021).

58. G. K. Marinov *et al.*, Transcription-dependent domain-scale three-dimensional genome organization in the dinoflagellate *Breviolum minutum*. *Nature Genetics* **53**, 613-617 (2021).

59. E. F. Camp *et al.*, Proteome metabolome and transcriptome data for three Symbiodiniaceae under ambient and heat stress conditions. *Scientific Data* **9**, 153 (2022).

60. T. G. Stephens *et al.*, Genomes of the dinoflagellate *Polarella glacialis* encode tandemly repeated single-exon genes with adaptive functions. *BMC Biology* **18**, 1-21 (2020).

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

62. M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, The KEGG resource for deciphering the genome. *Nucleic Acids Research* **32**, D277-D280 (2004).