Computational classification of SOD2 protein in coral and symbiotic algae from public genomic and transcriptomic data

Benjie Chen
benjie@alum.mit.edu

ABSTRACT

SOD2, or Manganese/Iron Superoxide Dismutase, plays a crucial role in maintaining cellular health across the tree of life. Existing literature is vague in identifying types of SOD2 proteins expressed in both corals and their symbiotic algae; most agree on the presence of a) mitochondrial targeting MnSOD in corals, and b) likely Mn dependent SOD2 in algae. This paper presents evidence for four classes of SOD2 proteins between coral hosts and their symbiotic algae. Public genomic and transcriptomic data reveal mitochondrial targeting and secreted versions of SOD2 among a wide range of cnidaria species, and chloroplast targeting (with a bipartite leader architecture) and secreted versions of SOD2 among several Symbiodinium species. The mature/catalytic region of SOD2 is well conserved within each class, but highly differentiated between classes. This paper presents the mature region HMM profiles for each class, with threshold scores high enough such that they are predictive of leader type.

Introduction

SOD2 proteins protect cells against oxidative damage by converting toxic superoxide radicals into hydrogen peroxide and diatomic oxygen. SOD2 proteins can bind either manganese (MnSOD) or iron (FeSOD) as co-factors; their structure consists of conserved residues His26, His74, Asp159, and His163 (human sequence coordinates) forming the anchors of a bipyramidal metal pocket and variations in the outer shell, based on bound metal, to ensure superoxide dismutase activity [1]. SOD2 is present throughout the tree of life, with bacterial SOD2 often preferring Fe as cofactor, mammalian SOD2 almost always requires Mn as a cofactor to function in mitochondria, and metal preference is not always phylogenetically constrained [2].

Existing literature is vague on the types of SOD2 proteins in coral hosts and their symbiotic algae. While there is consensus on the presence of mitochondrial MnSOD in corals, other forms of SOD2 in coral and symbiotic algae are rarely discussed/analyzed separately. Saragosti et al. [3] report experimental evidence for secreted superoxide dismutase activity in *Stylophora pistillata* without identifying the source of such activity. Dash et al. [4] experimentally identify two forms of MnSOD in *Hydra vulgaris*, a non-coral species in the Cnidaria phylum. Krueger et al. [5] computationally compare various annotated MnSOD and FeSOD sequences from various Symbiodinium species, with the grouping of sequences, based on putative annotation, inconsistent with findings of this paper. Indeed, most transcriptomic analyses in the field depend on blasting against UniProt or NCBI, or matching against the HMM profile K04564, a pan-metagenomic KEGG ortholog [6] for SOD2. Neither approach provides resolution beyond putative superoxide dismutase activity.

This paper presents results from a computational search and grouping of SOD2 protein sequences, using NCBI deposited genome sequences from 35 coral and 10 Symbiodinium species, as well as RNAseq datasets from 5 papers. The goal of this work is to present specific and unambiguous sequence-level profiles for SOD2 proteins in *Anthozoa* and *Dinophyceae*. Although these protein groups have features that are suggestive of biological and biochemical mechanism, the scope of this paper is limited to computational; there is no

Table 1. RNAseq sequences used in this paper

Species	Source of Raw Reads	Source of Transcriptome	Source of Protein Sequences
Acropora tenuis	[14]	This paper, Trinity	This paper, TransDecoder
Orbicella faveolata	[15]	This paper, Trinity	This paper, TransDecoder
Pseudodiploria clivosa	[15]	This paper, Trinity	This paper, TransDecoder
Siderastrea radians	[15]	This paper, Trinity	This paper, TransDecoder
Durusdinium trenchii	[16]	[16]	This paper, TransDecoder
Cladocopium goreaui	[14]	This paper, Trinity	This paper, TransDecoder
Breviolum B5	[15]	This paper, Trinity	This paper, TransDecoder
Breviolum faviinorum	[15]	This paper, Trinity	This paper, TransDecoder
Symbiodinium A3	[15]	This paper, Trinity	This paper, TransDecoder
Durusdinium trenchii	[17]	This paper, Trinity	This paper, TransDecoder
Cladocopium goreaui	[18]	[18]	This paper, TransDecoder

experimental work, beyond referenced RNAseq experiments, to functionally validate the different classes of proteins, determine their metal preference, or likely cellular location.

Specifically, a protein sequence database was constructed from a) protein sequences submitted to and associated with NCBI genome accessions, b) protein sequences exhaustively detected using 25k+ KO HMM profiles [6] on 6-frame translations of genomic sequences submitted to and associated with NCBI genome accessions, and c) protein sequences generated from transcriptomes from 5 published RNAseq studies. 210 seed sequences meeting the K04564 HMM profile minimum threshold [7] were selected from this sequence database.

Seed sequences were further filtered to those containing PF00081 and PF02777 Pfam FeSOD N and C-terminal domains, using Pfam's gathering score [8], and having four key metal-binding pocket residues [1]. A *mature protein sequence* between (inclusive) the 3rd AA of PF00081 and 100th AA of PF02777 was extracted for each remaining seed sequence.

A leader discovery algorithm was employed to look for, in predicted exons if protein sequence was obtained through a gene model (above (a)) or in upstream 3-frame translations (above (b)) or in the protein sequence (above (c)), possible start codons upstream of each mature protein sequence. This algorithm produced zero (if no suitable start codon was found) or more full-length SOD2 sequences for each mature sequence. This leader discovery step typically found predicted start sites for coral hosts, but was instrumental in finding complete SOD2 sequences in symbiotic algae genomic and transcriptomic data; tandem transcription [9] and alternative splicing sites [10] make gene prediction in dinoflagellates difficult otherwise.

DeepLoc 2.1 [11] was used to determine possible target cellular compartment and leader type of each full-length SOD2 sequence. MUSCLE multi-sequence alignments [12] were used to group mature protein sequences.

Five RNAseq studies are included in this work. Table 1 lists the papers and the dataset each contributes. RNAseq data provide further evidence for expression, and works around the complex splicing machinery in Symbiodinium. In some cases, raw reads from the original paper were obtained from NCBI and assembled using Trinity [13] into a transcriptome, and TransDecoder [13] was used to generate protein sequences.

Results

Manual inspection of DeepLoc 2.1 results over full-length SOD2 sequences suggests the presence of three patterns: a) a mitochondrial transit peptide (mTP) between 15 to 35 AAs upstream of PF00081 with sequence location predicted to be mitochondria, b) a signal peptide (SP) between 15 to 40 AAs upstream of PF00081 with sequence location predicted to be extracellular, and c) a signal peptide between 50 and 70 AAs upstream

Table 2. Breakdown of full-length SOD2 sequences

Group	Taxonomy	DeepLoc Predictions		Start codon upstream of PF00081	Unique full length seqs
		Leader	Location		
mT_cnidaria	Coral	mTP	Mitochondria	15 to 35 AAs	65
SP_cnidaria	Coral	SP	Extracellular	15 to 40 AAs	35
SP_cTP_algae	Symbiodinium	SP	Golgi apparatus	50 to 70 AAs	17
SP_algae	Symbiodinium	SP	Extracellular	15 to 35 AAs	6

of PF00081 with sequence location predicted to be the Golgi apparatus. Specifically, the first and last predictions were almost exclusively for corals and Symbiodinium, respectively, while the middle prediction includes sequences for both hosts and symbionts.

Additional inspection using DeepLoc, on the SP and Golgi apparatus sequences, reveals that once the SP is cleaved, DeepLoc predicts the remaining sequence as plastid bound. This observation is consistent with general consensus of bipartite leader architecture for chloroplast-targeting proteins in cells that have gone through secondary endosymbiosis [19]; proteins are first shuttled into the secretory pathway, through the Golgi apparatus, before reaching the plastid.

Based on these observations, leader discovery algorithm was adjusted to specifically look for start sites in these ranges, and the full-length SOD2 sequences were separated into four groups: mTP in coral, SP in coral, SP+cTP in Symbiodinium, and SP in Symbiodinium. Multi-sequence alignments of mature protein sequences within each bucket showed high degree of conservation. Overall, 147 full-length SOD2 sequences, 123 unique, exhibited these DeepLoc called features. Table 2 shows the breakdown by group. Figure 1 shows breakdown by taxonomy – the first columns on the left are transcriptomic sequences, then the rest of the columns are grouped by taxonomy of the genomic sequences.

A few genomes/species have high HMM scores but lack an expected DeepLoc prediction; these are likely false negatives due to DeepLoc being unable to predict a required leader type and/or compartment.

Four groups of SOD2 sequences, two in coral and two in Symbiodinium, are clearly distinguishable from each other, and cover >=90% of sequences with SOD2 signature

For each protein group, an HMM profile of the mature protein sequences was constructed using hmmer3 [20]. Figure 2 shows that the four HMM profiles are clearly distinguishable; when scoring each group against all four profiles, the correct profile, in orange, scores the highest, with no overlap with other scores. The lowest score from each set of orange marks is used as the "threshold score" for the corresponding HMM profile. The values are 303.5, 262.4, 441.8, and 421.3, respectively.

To test if there are other undiscovered SOD2 protein groups, 114 unique mature protein sequences were selected from all the genomic and RNAseq sequences that matched to the K04564 HMM profile, without checking for plausible start codons, types of leader sequences, and targeted cellular compartment. These 114 sequences were then scored against the mature protein HMM profiles, with no consideration for presence or type of leaders, from the four groups. 102 sequences, or 90%, scored above the threshold for an HMM profile. Out of the 12 negatives, 4 are just below the threshold for SP_cTP_algae and can be considered FNs. Thus 106 sequences, 93% of all K04564 matching sequences, can be classified by the four HMM profiles.

Interestingly, SP_cnidaria sequences are missing from all the genomes under the *Acroporidae* family. The *Acropora tenuis* RNAseq data from [14] was captured against its genomic sequences, which may also explain the lack of a SP_cnidaria sequence in that transcriptome. It is unclear if this omission in the genome is real, or due to some genome assembly artifact.

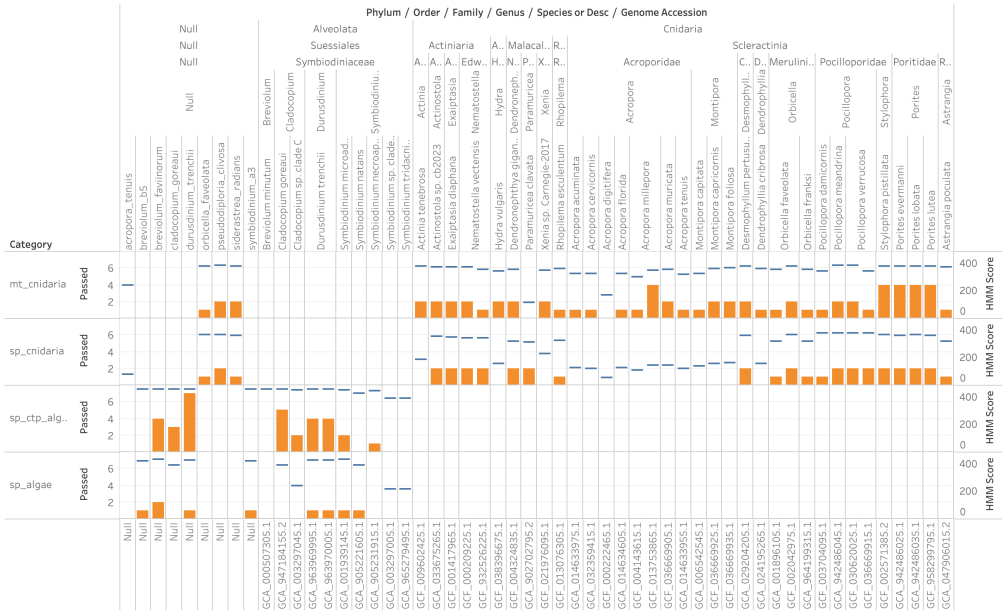

Figure 1. Orange vertical bars are the numbers of full-length SOD2 sequences per species/genome, for each of the four protein groups. Left columns are sequences from RNAseq studies. Right columns include sequences from genomes, grouped by taxonomy. Horizontal marks are the highest HMM scores from each set of sequences, scored against the HMM profile for that protein group. A high HMM score without accompanying orange vertical bar indicates a false negative – a protein sequence can be classified by HMM into a group, but DeepLoc is unable to predict the required leader type and/or compartment.

Figure 2. `hmmsearch` scores of protein groups (each in a pane), against all the HMM profiles (columns).

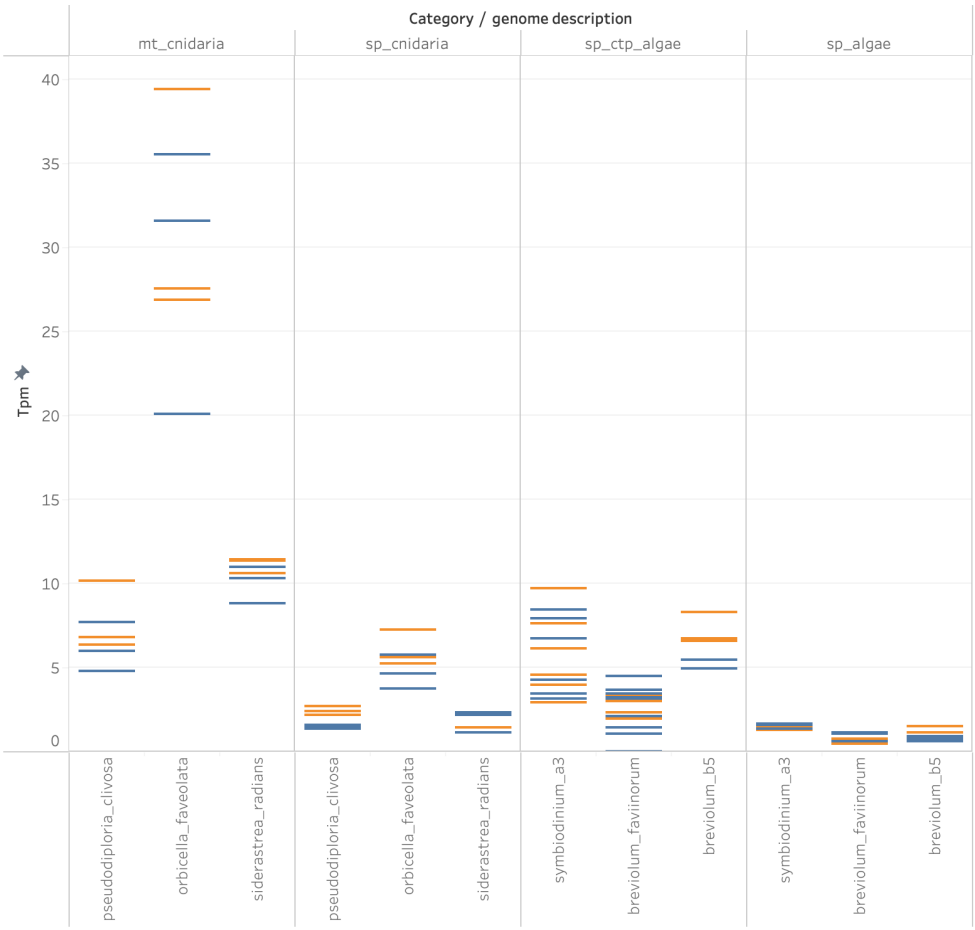

Figure 3. Transcript counts, in transcripts per million (TPM), for the four protein groups, from RNAseq data provided by [15]. Orange and blue lines are from treatment groups at elevated temperature and control groups at normal temperature, respectively. Differential expression analysis is out of scope for this paper.

RNAseq data provide expression evidence for all groups, with very weak expression of SP_algae

[15] provides RNA sequencing reads for three holobionts, (using paper's description) thermally sensitive *Pseudodiploria clivosa* and its symbiotic algae *Symbiodinium A3*, thermally robust *Orbicella faveolata* and its symbiotic algae *Breviolum faviinorum*, and thermally resistant *Siderastrea radians* and its symbiotic algae *Breviolum B5*.

Figure 3 shows that proteins from all four groups were present in the RNAseq data and registered positive read counts (transcripts per million, or TPM), although read counts for SP_algae are quite low. In depth comparison of these proteins between the treatment group at elevated temperature (orange lines) and the control group at normal temperature (blue lines) is beyond the scope of this paper. Transcript counts were generated using Salmon [21] on Trinity [13] assembled transcriptome.

The four groups of SOD2 proteins form distinct clades in an unrooted phylogenetic tree, suggesting independent evolution

160 mature SOD2 sequences consisting of the 123 from coral and Symbiodinium, plus 37 from annotated reference SOD2 proteins from across the tree of life, were clustered (98% sequence similarity) using MMSeqs [22], aligned using MUSCLE, and organized into an unrooted phylogenetic tree using IQTree [23].

Figure 4. Unrooted phylogenetic tree of the mature protein sequences from the four SOD2 protein groups and other reference sequences. UFBoot scores are shown next to the branches.

Figure 4 shows the tree as visualized using iTOL [24]. Each of the four protein groups forms an exclusive clade that includes every member of the group and shares a unique common ancestor. This tree structure suggests these four groups evolved independently, without intermixing with each other. Additionally, the mature protein sequences co-evolved, in lockstep, with the localization signals.

Discussion

Independent evolution of the four protein groups

Figure 4 strongly suggests that mT_cnidaria is the canonical mitochondrial MnSOD protein, clustered with reference primate and fungal mitochondrial-targeting MnSOD sequences with the enclosing fungal node supported by a branch bootstrap score of 100. Likely, mT_cnidaria evolved from the endosymbiosis event that gave rise to the eukaryotic mitochondria. In contrast, SP_cnidaria appears to have evolved distantly from mT_cnidaria, likely independent of mitochondrial endosymbiosis.

More sequences would boost confidence of any interpretation of the grouping of SP_algae and SP_cTP_algae in the tree. Currently, these two distinct groups appear to be paralogs that have come from gene duplication events.

Unverified metal specificity for three protein groups

mT_cnidaria is likely a canonical mitochondrial MnSOD. While there is no direct experimental evidence of manganese specificity of coral mitochondrial SOD2, this specificity has been demonstrated experimentally in other species [25].

Assigning metal specificity for the other three protein groups cannot be done computationally. Metal co-factor preference of a SOD2 protein does not always correlate with sequence; specificity depends on several non-conserved second sphere residues around the metal-binding pocket, different from the key structural residues this paper used to filter for functional SOD2 sequences. Additionally, while some proteins are Mn or Fe dependent, others are cambialistic and can work with either ion.

The metal annotations for the reference sequences in Figure 4 come from UniProt and are provided as a reference. Many of these annotations may not have been experimentally validated.

Unverified cellular compartment for three protein groups

The final destination of the SP_cnidaria, SP_cTP_algae, and SP_algae proteins has not been proven experimentally. While there is some evidence for secreted SOD2 activity in literature [3] [4], there is no link, yet, between that activity and sequence.

The leader type in these three groups comes from DeepLoc [11]. To sanity check DeepLoc predictions, protein sequences matching K09490, an ER chaperone slated for the secretory pathway, and K02639, ferredoxins that typically target the chloroplast, were passed to DeepLoc. For K09490 matching proteins, DeepLoc reports either a) SP as the leader and ER as the cellular compartment, or b) NLS as the leader and cytoplasm as the target compartment. Removing the SP, DeepLoc reports the ER or cytoplasm as the target compartment for the truncated sequences. For K02639 matching sequences, DeepLoc reports SP as the leader for some sequences, and Golgi apparatus as the target compartment. Removing the SP, DeepLoc reports plastid as the target compartment for the truncated protein sequences. These results indicate that DeepLoc predictions for the four SOD2 protein groups are unlikely coincidences.

Methods and Files

Seed Sequences

FASTA file for the 210 seed protein sequences is at `https://github.com/benjiec/paper-coral-sod2-hmm/tree/main/data/seed_sequences.faa`. A manifest that lists the genome accession is at `https://github.com/benjiec/paper-coral-sod2-hmm/tree/main/data/seed_sequences.tsv`.

Protein sequences from NCBI genomes

Genomic sequences and protein sequences (if available) from 35 coral genomes and 10 Symbiodinium genomes were downloaded from NCBI Genome database. *De novo* protein detection was also conducted on many of the genomes, by searching the full genomic 6-frame translations for hits against the entire KEGG KO HMM database [7], using `hmmscan` [20].

The *de novo* searching program is available at `https://github.com/benjiec/tangle-needle`.

Protein sequences from RNAseq data and transcript counts

Table 1 lists which transcriptomes and protein sequences were generated by this paper. Docker environment and instructions for using Trinity, TransDecoder [13] and Salmon [21] are available at `https://github.com/benjiec/tangle-pile`.

Transcript counts associated with protein sequences from each of the four protein groups are available at `https://github.com/benjiec/paper-coral-sod2-hmm/tree/main/data/transcript_counts.txt`.

Removal of non-functional and/or partial protein sequences

K04564 matching protein sequences were filtered to remove those that do not contain both PF00081 and PF02777, and those that do not possess key residues that form the metal-binding pocket. A rule-based filtering program was used. The code for the program is available at `https://github.com/benjiec/tangle-sieve`. The specific rule used for each protein group is available at `https://github.com/benjiec/tangle-coral/tree/main/coral/rules`.

Leader discovery

Source code for leader discovery is available at `https://github.com/benjiec/tangle-sieve`. Start codons upstream and at a specified distance from an anchor (e.g. where PF00081 aligns to a seed protein sequence) are used as the start of a full-length SOD2 sequence. If the seed protein sequence comes from RNAseq data, start codon searching only occurs in the TransDecoder generated protein sequence. If the seed protein sequence comes from an NCBI protein file, with an associated genome GFF file, start codon searching takes predicted exons from the GFF file into consideration. If the seed protein sequence comes from *de novo* HMM search (see above), start codon discovery begins in frame with the anchor and looks upstream until hitting a prior stop codon. Leader discovery in this last context is challenging and may not find the correct start codon if the leader transcript was spliced together.

This paper chose not to use the patch to AUGUSTUS, available at `http://smic.reefgenomics.org/download/`, because the patch assumes the sequence motif frequencies around each of the three DSS sites are independent of the DSS motif itself. Through inspection of likely DSS sites from several Symbiodinium genomes, that assumption does not appear to be valid.

Full-length and mature SOD2 sequences

147 full-length sequences are available at `https://github.com/benjiec/paper-coral-sod2-hmm/tree/main/data/full_length.faa`. The mature portions of the sequence, bounded by 3rd AA of PF00081 and

100th AA of PF02777, inclusively, are available at `https://github.com/benjiec/paper-coral-sod2-hmm/tree/main/data/mature.faa`. The reference mature sequences, included in the phylogenetic tree construction, are available at `https://github.com/benjiec/paper-coral-sod2-hmm/tree/main/data/reference_mature.faa`.

The phylogenetic tree constructed by IQTree [23] is available at `https://github.com/benjiec/paper-coral-sod2-hmm/tree/main/data/tree.contree`.

Alignment and HMM profile construction

Multi-sequence alignments of mature sequences, without the leader sequence and bounded by PF00081 and PF02777, were constructed using MUSCLE [12] after removing duplicate sequences. HMM profiles were constructed using `hmmbuild` [20] from the alignments.

The four HMM profiles as well as the FASTA files used to construct each profile are available at `https://github.com/benjiec/paper-coral-sod2-hmm/tree/main/profiles/`.

References

1. Perry, J., Shin, D., Getzoff, E. & Tainer, J. The structural biochemistry of the superoxide dismutases. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* (2010).
2. Sendra, K. *et al.* An ancient metalloenzyme evolves through metal preference modulation. *Nature Ecology and Evolution* (2023).
3. Saragosti, E., Tchernov, D., Katsir, A. & Shaked, Y. Extracellular production and degradation of superoxide in the coral *Stylophora pistillata and cultured Symbiodinium*. *PLOS One* (2010).
4. Dash, B., Metz, R., Huebner, H., Porter, W. & Phillips, T. Molecular characterization of two superoxide dismutases from *Hydra vulgaris*. *Gene* (2007).
5. Krueger, T. *et al.* Transcriptomic characterization of the enzymatic antioxidants FeSOD, MnSOD, APX and KatG in the dinoflagellate genus Symbiodinium. *BMC Evolutionary Biology* (2015).
6. Kanehisa, M. *et al.* From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research* (2006).
7. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* (2020).
8. Sonnhammer, E., Eddy, S., Birney, E., Bateman, A. & Durbin, R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Research* (1998).
9. Zhang, H., Hou, Y., Miranda, L. & Lin, S. Spliced leader RNA trans-splicing in dinoflagellates. *Microbiology* (2007).
10. Shoguchi, E. *et al.* Draft Assembly of the Symbiodinium minutum Nuclear Genome Reveals Dinoflagellate Gene Structure. *Current Biology* (2013).
11. Nielsen, H. Practical Applications of Language Models in Protein Sorting Prediction: SignalP 6.0, DeepLoc 2.1, and DeepLocPro 1.0. *Large Language Models (LLMs) in Protein Bioinformatics* (2025).
12. Edgar, R. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nature Communications* (2022).
13. Grabherr, M. *et al.* Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology* (2013).
14. Buerger, P. *et al.* Heat-evolved microalgal symbionts increase coral bleaching tolerance. *Science Advances* (2020).
15. Avila-Magaña, V. *et al.* Elucidating gene expression adaptation of phylogenetically divergent coral holobionts under heat stress. *Nature Communications* (2021).

16. Bellantuono, A., Dougan, K., Granados-Cifuentes, C. & Rodriguez-Lanetty, M. Free-living and symbiotic lifestyles of a thermotolerant coral endosymbiont display profoundly distinct transcriptomes under both stable and heat stress conditions. *Molecular Ecology* (2019).
17. Camp, E. *et al.* Proteome metabolome and transcriptome data for three Symbiodiniaceae under ambient and heat stress conditions. *Scientific Data* (2022).
18. Chen, Y. *et al.* Improved Cladocopium goreaui Genome Assembly Reveals Features of a Facultative Coral Symbiont and the Complex Evolutionary History of Dinoflagellate Genes. *Microorganisms* (2022).
19. McFadden, G. Endosymbiosis and evolution of the plant cell. *Current Opinion in Plant Biology* (1999).
20. Eddy, S. Accelerated Profile HMM Searches. *PLOS Computational Biology* (2011).
21. Patro, R., Duggal, G., Love, M., Irizarry, R. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* (2017).
22. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* (2017).
23. Wong, T. *et al.* IQ-TREE 3: phylogenomic inference software using complex evolutionary models. *Molecular Biology and Evolution* (2026).
24. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research* (2024).
25. Luk, E., Yang, M., Jensen, L., Bourbonnais, Y. & Culotta, V. Manganese activation of superoxide dismutase 2 in the mitochondria of Saccharomyces cerevisiae. *Journal of Biological Chemistry* (2015).