Appendix S1

Two-phase approach for identification of putative contaminant sequences

For each genome assembly, scaffolds <1Kb were removed, before putative contaminants were identified using a robust two-phase strategy (Figure S1). First, following Liu et al. (2018), scaffolds were compared using BLASTn against a database of archaeal, bacterial and viral genome sequences retrieved from the NCBI RefSeq (release 88) database. Scaffolds with ≥ 5% of their length covered by significant BLASTn hits ($E \leqslant 10^{-20}$, bit score > 1000) were deemed putative contaminants and removed from the assembly. Second, the lengths of the remaining scaffolds were plotted against their G+C percentage, from which scaffolds (>10 Kb) with irregular G+C content (i.e., scaffolds that deviate from the expected normal distribution; Figure S2) were identified and removed. Specifically, long scaffolds (>100Kb) in the genome assemblies of *Cladocopium goreaui* and *Fugacium kawagutii* were removed if their corresponding G+C content is > 52% and > 53%, respectively (Figure S2). This process yielded the final revised genome assemblies that were used for subsequent analysis.

The customized gene-prediction workflow for dinoflagellate genomes

Our customized workflow for predicting genes from dinoflagellate genomes is available at https://github.com/TimothyStephens/Dinoflagellate_Annotation_Workflow; an overview of this workflow is shown in Fig. S1. For each genome assembly, protein-coding genes were predicted using an approach similar to Liu et al. (2018). For each assembly, a *de novo* repeat library was generated using RepeatModeler v1.0.11 (http://www.repeatmasker.org/RepeatModeler/) and combined with known repeats in the RepeatMasker database (release 20171107) to create a customized repeat library. Repeats were masked in each assembly using RepeatMasker v4.0.7 (http://www.repeatmasker.org/) and the associated customized repeat library.

For each published transcriptome assembly, vector sequences were removed using SeqClean (Chen et al. 2007) based on the UniVec (build 10) database. The PASA pipeline v2.3.3 (Haas et al. 2003) (modified to recognize non-canonical GA donor splice sites; see https://github.com/chancx/dinoflag-alt-splice) and TransDecoder v5.2.0 (Haas et al. 2013) were used to predict protein-coding genes using the corresponding vector-trimmed transcriptome assemblies (hereinafter transcript-based genes). The protein sequences from multi-exon

transcript-based genes with complete 5′ and 3′-ends were searched (BLASTp, $E \leqslant 10^{-20}$) against a combined sequence database of RefSeq proteins (release 88) and available Symbiodiniaceae proteins. Only genes with significant BLASTp hits (> 80% query coverage) were retained and checked for transposable elements using HHblits (Remmert et al. 2011), searching against the JAMg transposon database (https://github.com/genomecuration/JAMg), and Transposon-PSI (http://transposonpsi.sourceforge.net/). Proteins putatively identified as transposable elements were removed and those remaining were clustered using CD-HITS v4.6.8 (ID=75%) (Li & Godzik 2006); the representative sequence (and associated transcript-based gene) from each cluster was retained. The remaining genes were further processed by the *Prepare_golden_genes_for_predictors.pl* script from the JAMg package (modified to recognize non-canonical GA donor splice sites; https://github.com/genomecuration/JAMg).

This produced a set of high-quality "golden" genes that was used as the training set for gene prediction using SNAP (version 2006-07-28) (Korf 2004) and AUGUSTUS v3.3.1 (Stanke et al. 2006). Genes were also predicted using GeneMark-ES v4.38 (Lomsadze et al. 2018) and MAKER protein2genome v2.31.10 (Holt & Yandell 2011) using available Symbiodiniaceae and SwissProt proteins (downloaded 27/06/2018). AUGUSTUS (see https://github.com/chancx/dinoflag-alt-splice) and MAKER were modified to recognize the non-canonical GA donor splice sites, while SNAP and GeneMark-ES only recognized GT and GC splice sites.

Genes predicted by GeneMark-ES, MAKER, PASA (transcript-based genes), SNAP and AUGUSTUS were integrated into a single combined set using EVidenceModeler v1.1.1 (Haas et al. 2008). At this stage, Liu et al. (2018) incorporated the repeat information (i.e., repetitive regions were excluded from subsequent gene prediction). Here, however, we specifically ignored repeat information (i.e., repetitive regions were included in subsequent gene prediction); this was to specifically allowed for prediction of genes within the repetitive regions. The weightings used for the integration of predicted genes with EVidenceModeler were as follows: GeneMark-ES 2, SNAP 2, AUGUSTUS 6, MAKER 8, PASA 10. Genes produced by EVidenceModeler were retained if they were constructed using evidence from PASA or at least two other prediction methods. This integrated approach minimizes our dependency on a single method, as the algorithm implemented by distinct prediction tools can yield different results given the same

genome features. For instance, AUGUSTUS has been shown to predict genes with an unrealistically large number of introns when trained using a low-quality dataset (Hoff & Stanke 2019).

References

Chen, Y. A., Lin, C. C., Wang, C. D., Wu, H. B. & Hwang, P. I. 2007. An optimized procedure greatly improves EST vector contamination removal. *BMC Genomics* **8**:416.

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Jr., Hannick, L. I., Maiti, R., Ronning, C. M., Rusch, D. B., Town, C. D., Salzberg, S. L. & White, O. 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**:5654-66.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel, R., LeDuc, R. D., Friedman, N. & Regev, A. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**:1494-512.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R. & Wortman, J. R. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**:R7.

Hoff, K. J. & Stanke, M. 2019. Predicting genes in single genomes with AUGUSTUS. *Curr Protoc Bioinformatics* **65**:e57.

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

Korf, I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* **5**:59.

Li, W. & Godzik, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**:1658-9.

Liu, H., Stephens, T. G., González-Pech, R. A., Beltran, V. H., Lapeyre, B., Bongaerts, P., Cooke, I., Aranda, M., Bourne, D. G., Foret, S., Miller, D. J., van Oppen, M. J. H., Voolstra, C. R., Ragan, M. A. & Chan, C. X. 2018. *Symbiodinium* genomes reveal adaptive evolution of functions related to coral-dinoflagellate symbiosis. *Commun Biol* **1**:95.

Lomsadze, A., Gemayel, K., Tang, S. & Borodovsky, M. 2018. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. *Genome Res* **28**:1079-89.

Remmert, M., Biegert, A., Hauser, A. & Soding, J. 2011. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **9**:173-5.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. & Morgenstern, B. 2006. AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res* **34**:W435-9.