Supplementary Information

Structural rearrangements drive extensive genome divergence between symbiotic and free-living *Symbiodinium*

Raúl A. González-Pech[1], Timothy G. Stephens[1], Yibi Chen[1], Amin R. Mohamed[2], Yuanyuan Cheng[3,†], David W. Burt[3], Debashish Bhattacharya[4], Mark A. Ragan[1], Cheong Xin Chan[1,5]*

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

[2]Commonwealth Scientific and Industrial Research Organisation (CSIRO) Agriculture and Food, Queensland Bioscience Precinct, St Lucia, QLD 4072, Australia

[3]UQ Genomics Initiative, The University of Queensland, Brisbane, QLD 4072, Australia

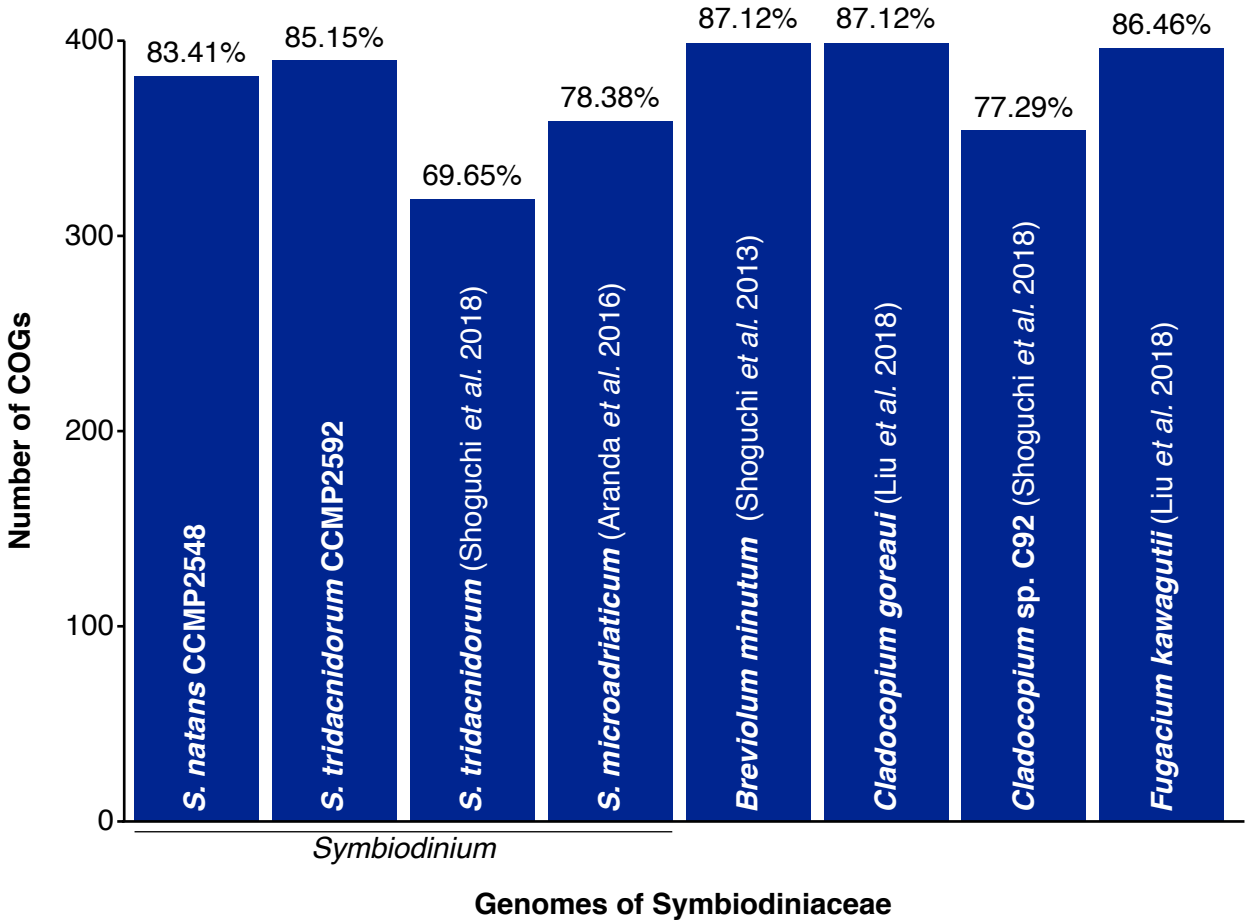
[4]Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ 08901, U.S.A.

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

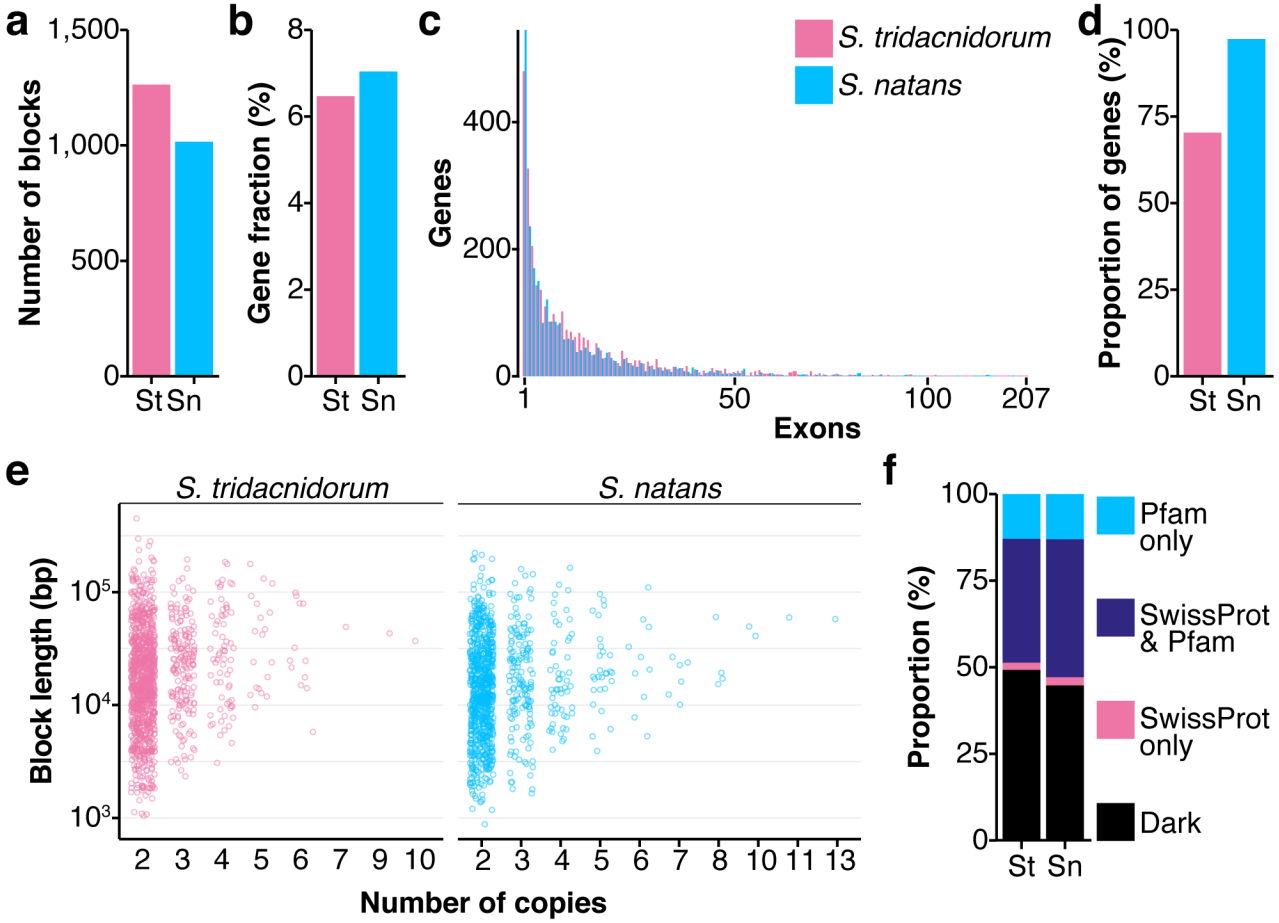
[†]Current address: School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW 2006, Australia

*Correspondence author (c.chan1@uq.edu.au)

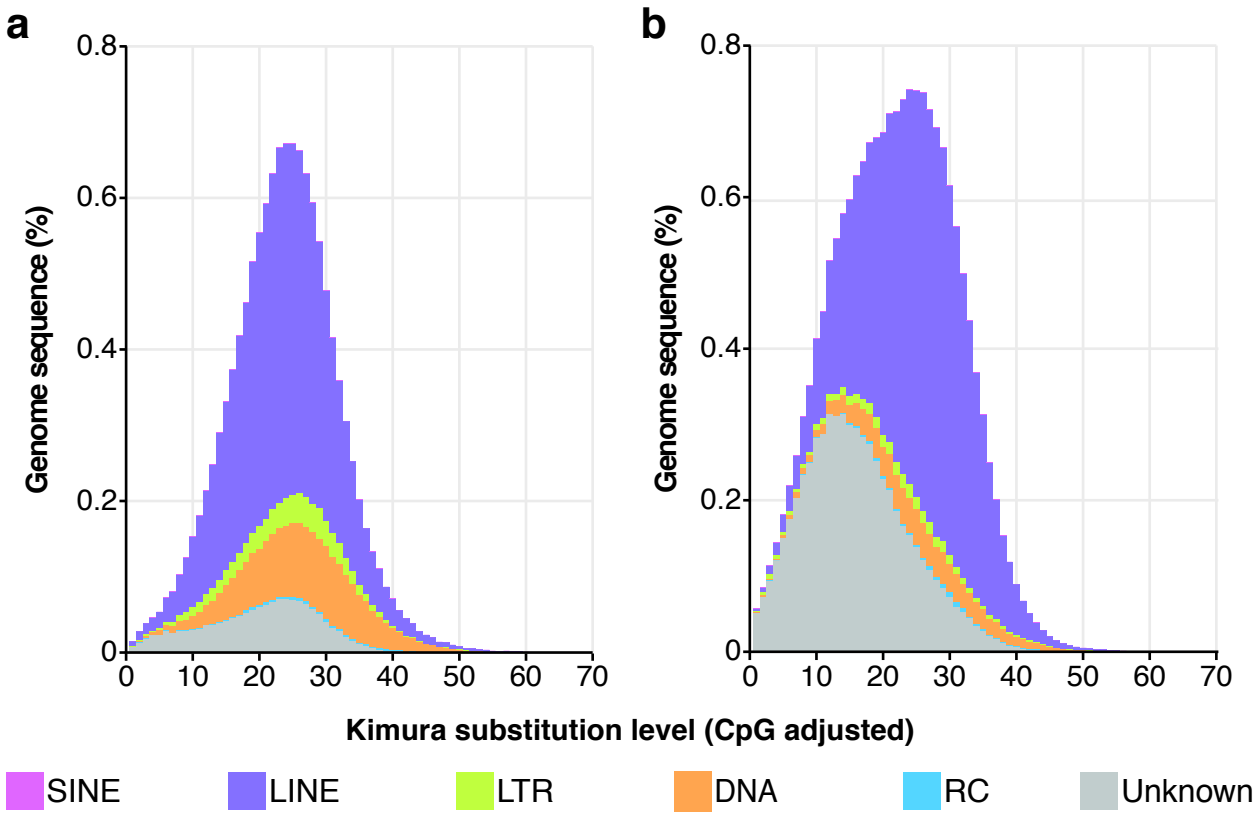
Supplementary Figures



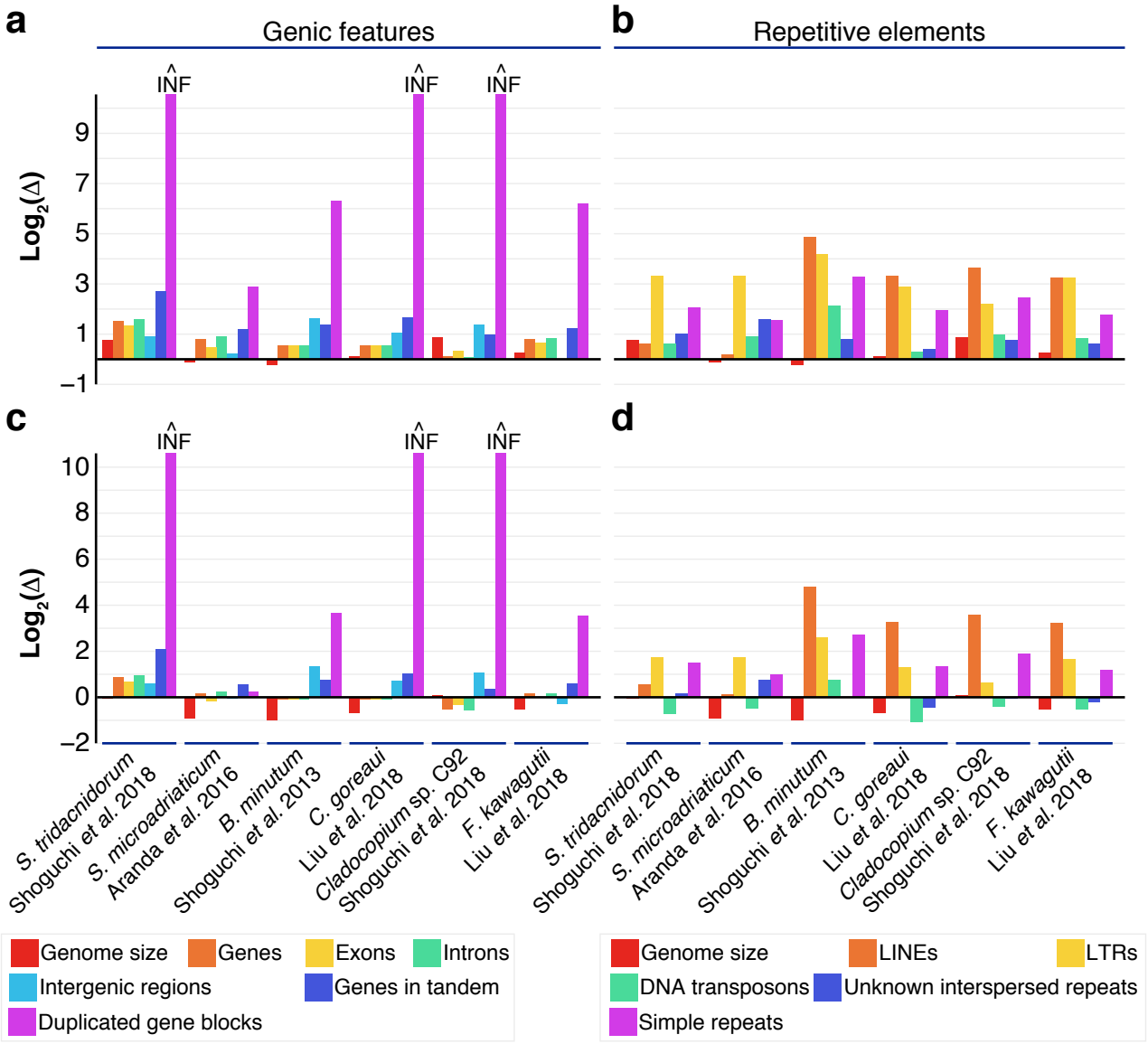
Supplementary Figure 1 Recovery of CEGMA Clusters of Orthologous Groups (COGs) by BLASTp similarity search (see Methods) in the predicted protein sequences of *S. natans* CCMP2548 and *S. tridacnidorum* CCMP2592, as well as in the most recently predicted protein sequences[1] of other available genomes of Symbiodiniaceae: *S. tridacnidorum*[2], *Symbiodinium microadriaticum*[3], *Breviolum minutum*[4], *Cladocopium goreaui*[5], *Cladocopium* sp. C92[2] and *F. kawagutii*[5]. The fraction of COGs for each genome, out of the 458 in CEGMA, is shown on top of its corresponding bar.



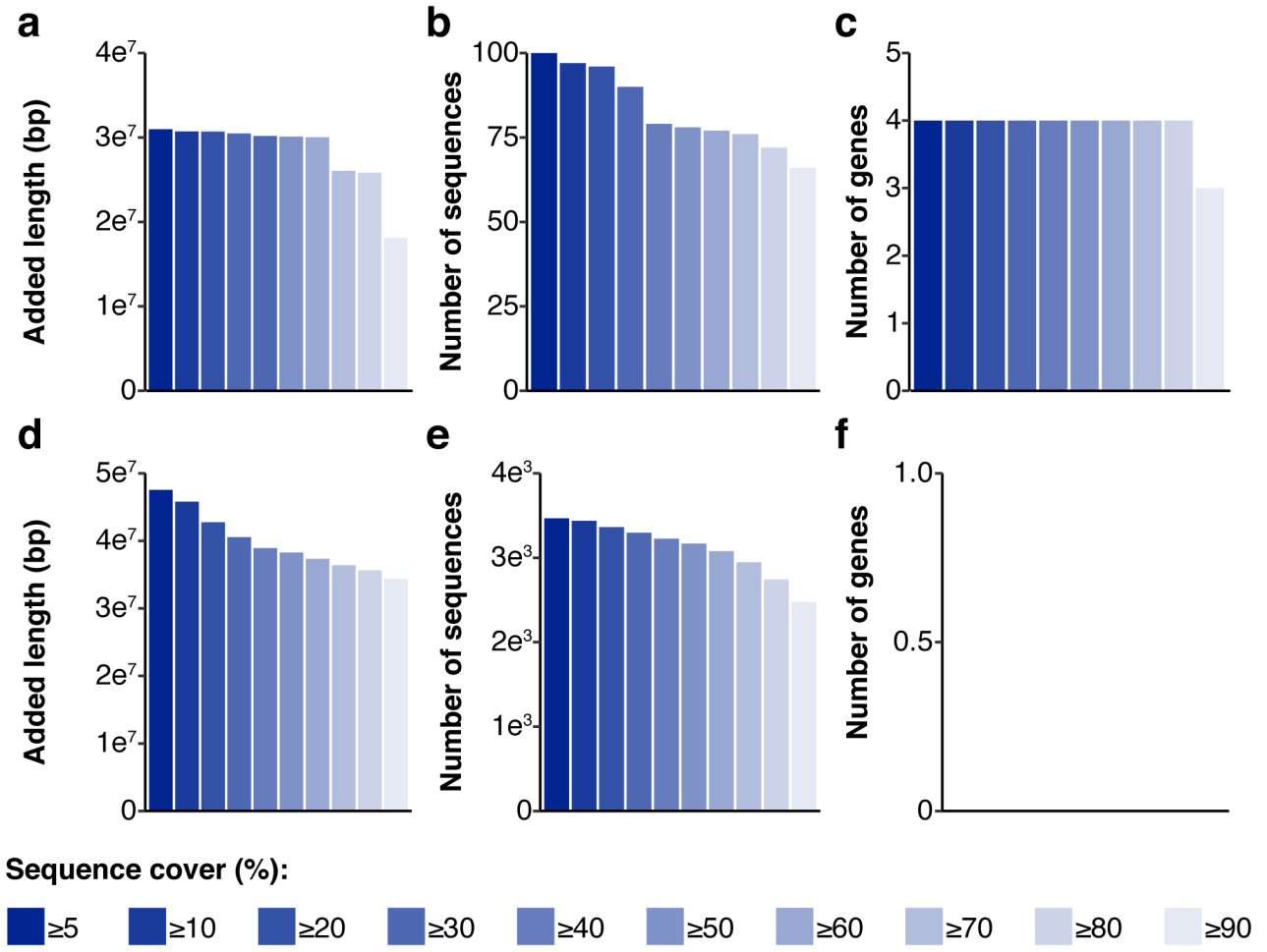
Supplementary Figure 2 (a) Number of blocks of genes arranged in tandem, and **(b)** fraction of genes (out of the total number of genes per genome) implicated in those blocks, found in genomes of *S. tridacnidorum* and *S. natans*. **(c)** Breakdown of the number of genes by the number of exons they contain. **(d)** Gene fraction (out of the total number of genes in tandem) with transcript support. **(e)** Block length distribution of genes in tandem as a function of the number of gene copies implicated in the blocks. **(f)** Fraction of genes with functional annotation based either on sequence similarity of their protein products with sequences in UniProt-SwissProt or identified Pfam protein domains, or both.



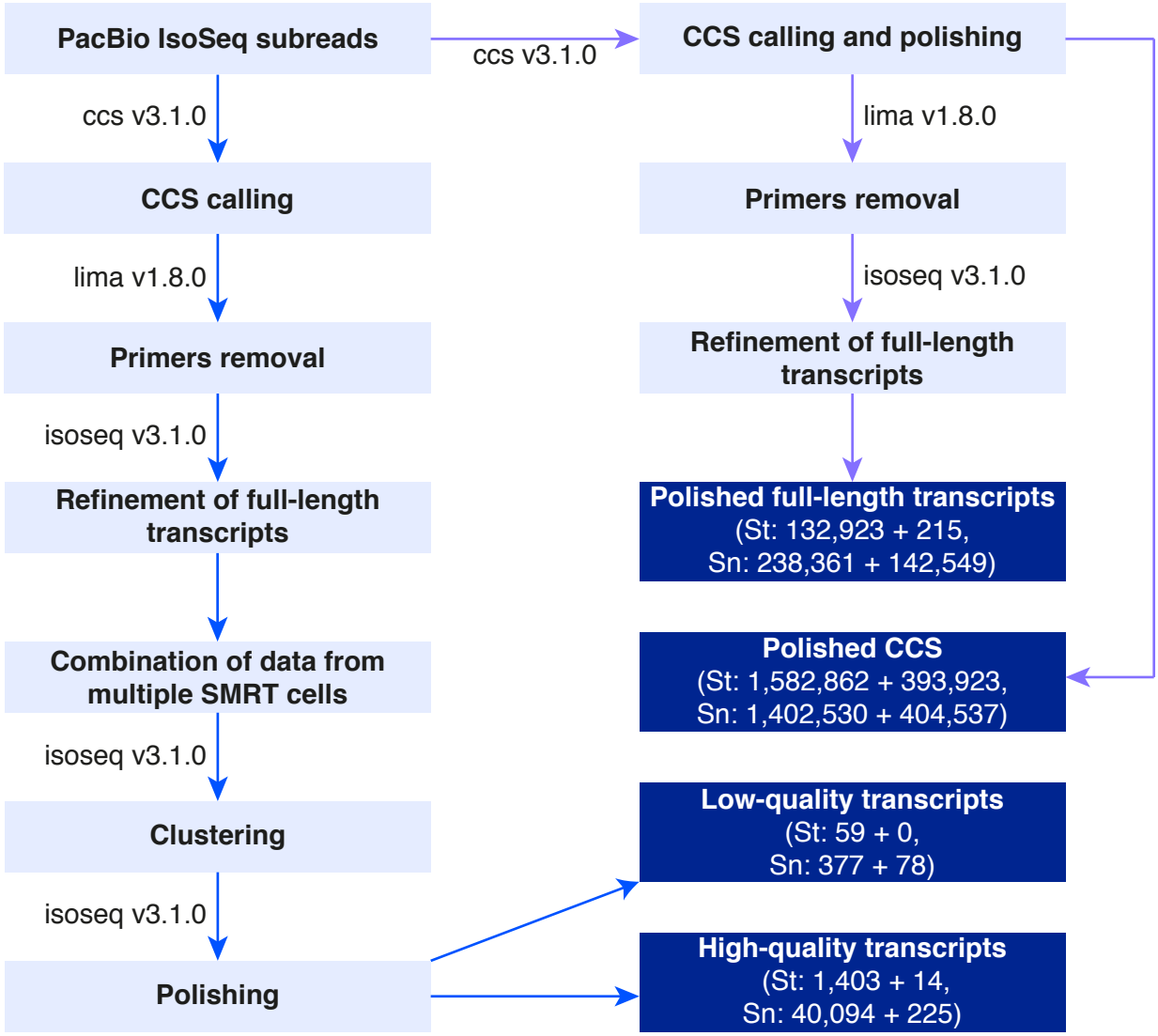
Supplementary Figure 3 Interspersed repeat landscapes from the preliminary genome assemblies (based on short-read data only) of *S. natans* **(a)** and *S. tridacnidorum* CCMP2592 **(b)**. The colour code of the different repeat classes (including multiple-copy genes) is shown at the bottom of the two charts.



Supplementary Figure 4 Ratio (Δ) of the sequence length covered by genic features **(a, c)** and repetitive elements **(b, d)** from other Symbiodiniaceae genome assemblies (based on short-sequence read data) relative to those from the hybrid assemblies of *S. tridacnidorum* **(a, b)** and *S. natans* **(c, d)**. The ratios between estimated genome sizes is given as reference, except for *S. tridacnidorum* and *Cladocopium* sp. C92 (Shoguchi *et al.* 2018)[2] that lack estimates, for which assembly total length was used instead.



Supplementary Figure 5 Added length **(a, d)** and count **(b, e)** of putatively contaminant sequences, and number of implicated genes **(c, f)**, in genomes assemblies from *S. tridacnidorum* **(a-c)** and *S. natans* **(d-f)**, based on different sequence cover thresholds (x-axis) of significant hits against bacterial, archaeal and viral genomes. The bars in each chart are coloured according to their corresponding query cover threshold following the colour code at the bottom.



Supplementary Figure 6 Diagram showing the detailed steps followed to process PacBio IsoSeq data to generate full-length transcript evidence for gene prediction. The traditional IsoSeq 3.1 workflow (blue arrows) was followed to obtain low- and high- quality transcripts. In an alternative approach (purple arrows), circular consensus sequences (CCS) were called and polished simultaneously. These polished CCS were further trimmed and refined into full-length transcripts skipping the clustering step. IsoSeq sequences from the DinoSL library were processed apart from the other libraries. Boxes in dark blue represent the transcript evidence subsequently used for gene prediction and the values in parentheses show the corresponding number of sequences from the standard (left) + the DinoSL (right) libraries for both *S. tridacnidorum* (St) and *S. natans* (Sn).

Supplementary References

1 Chen, Y., Stephens, T. G., Bhattacharya, D., González-Pech, R. A. & Chan, C. X. Evidence that inconsistent gene prediction can mislead analysis of algal genomes. *bioRxiv*, 690040, doi:10.1101/690040 (2019).

2 Shoguchi, E. *et al.* Two divergent *Symbiodinium* genomes reveal conservation of a gene cluster for sunscreen biosynthesis and recently lost genes. *BMC Genomics* **19**, 458, doi:10.1186/s12864-018-4857-9 (2018).

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

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

5 Liu, H. *et al. Symbiodinium* genomes reveal adaptive evolution of functions related to coral-dinoflagellate symbiosis. *Commun. Biol.* **1**, 95, doi:10.1038/s42003-018-0098-3 (2018).