Massive genome reduction occurred prior to the origin of coral algal symbionts

Sarah Shah[1], Katherine E. Dougan[1], Yibi Chen[1], Rosalyn Lo[1], Gemma Laird[1], Michael D. A. Fortuin[1], Subash K. Rai[2], Valentine Murigneux[2], Anthony J. Bellantuono[3], Mauricio Rodriguez-Lanetty[4], Debashish Bhattacharya[5,*], Cheong Xin Chan[1,*]

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

[2]The University of Queensland, Genome Innovation Hub, Brisbane, 4072 QLD, Australia.

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

[4]Florida International University, Department of Biological Sciences, Miami, FL 33099, USA.

[5]Rutgers University, Department of Biochemistry and Microbiology, New Brunswick, NJ, 08901, USA

*Corresponding authors: Debashish Bhattacharya (d.bhattacharya@rutgers.edu) and Cheong Xin Chan (c.chan1@uq.edu.au)

Abstract

Dinoflagellates in the Family Symbiodiniaceae (Order Suessiales) represent diverse, predominantly symbiotic lineages that associate with taxa such as corals and jellyfish. Their ancestor is believed to have been free-living, and the establishment of symbiosis (i.e., symbiogenesis) is hypothesised to have occurred multiple times during Symbiodiniaceae evolution. Among Symbiodiniaceae taxa, the genus *Effrenium* is an early diverging, free-living lineage that is phylogenetically positioned between two robustly supported groups of genera within which symbiotic taxa have emerged. The lack of symbiogenesis in *Effrenium* suggests that the ancestral features of Symbiodiniaceae may have been retained in this lineage. Here we present *de novo* assembled genomes and associated transcriptome data from three isolates of *Effrenium voratum*. We compared the *Effrenium* genomes (1.2-1.9 Gbp in size) and gene features with those of 16 Symbiodiniaceae taxa and other outgroup dinoflagellates. Surprisingly, we find that genome reduction predates the origin of Symbiodiniaceae and is not primarily explained by their symbiotic lifestyle. We postulate that adaptation to an extreme habitat (e.g., as in *Polarella glacialis*) or life in oligotrophic conditions resulted in the Suessiales ancestor having a haploid genome size $\leq$ 2Gbp, which is retained (or smaller) among all extant algae in this lineage. Our data reveal that the free-living lifestyle distinguishes *Effrenium* from symbiotic Symbiodiniaceae vis-à-vis longer introns, more-extensive mRNA editing, fewer (~30%) lineage-specific gene families, and lower (~10%) level of pseudogenisation. These results demonstrate how genome reduction and the adaptation to symbiotic versus free-living lifestyles intersect, and have driven the diversification and genome evolution of Symbiodiniaceae.

Introduction

Dinoflagellate algae in the Family Symbiodiniaceae comprise taxa that form symbioses with diverse marine organisms. Of particular importance to modern coral reefs, these dinoflagellates provide photosynthates *via* fixed carbon and essential nutrients to corals while resident in these cnidarians. The Symbiodiniaceae ancestor is believed to have been free-living (LaJeunesse et al. 2018) with members of this group forming symbiotic associations with corals as early as 230 million years ago (MYA) (Frankowiak et al. 2021). Symbiogenesis, or the establishment of a symbiotic relationship between two or more taxa (Guerrero et al. 2013), can drastically influence their evolution, adaptation, and speciation as observed in obligate parasites and diverse symbiotic taxa (Stephen et al. 2005; Morrison et al. 2007; Opperman et al. 2008; Kwan et al. 2012; McCutcheon and Moran 2012). This is termed the resident genome syndrome and was previously hypothesized to explain the observed patterns of Symbiodiniaceae genome evolution (González-Pech et al. 2019).

Based on current divergence time estimates for Symbiodiniaceae, the split between the basal genera (*Symbiodinium* and *Philozoon*) and the rest of the family occurred 166 MYA, whereas the more-recently branching symbiotic lineages diversified ~109 MYA (LaJeunesse et al. 2018). If the emergence of symbiogenesis coincides with the earliest fossil evidence from 230 MYA (Frankowiak et al. 2021), then different Symbiodiniaceae lineages would have arose and diversified during major global geological events. These events include the switch from aragonite to calcite seas (~190 MYA (Vulpius and Kiessling 2017)), the breakup of Pangea (150-230 MYA (Schettino and Turco 2009)), the diversification or extinction of potential hosts, e.g., the extinction of rudists 66 MYA (Stanley and van de Schootbrugge 2018; de Winter et al. 2020), and the overall change in coral morphology from the Triassic (201-252 MYA) to the Cretaceous (66-145 MYA) (Veron 2011; Pandolfi and Kiessling 2014). More recent examples include the rapid radiation of the genus *Cladocopium* 4-6 MYA due to geographic isolation (Thornhill et al. 2014) and the co-diversification of *Symbiodinium fitti* and their coral hosts (Turnham et al. 2021).

As described above, symbiogenesis is expected to impact the genome evolution of symbionts to varying extents within a broad spectrum of "facultativeness" that reflects the nature of the host association (i.e., with obligate free-living and obligate symbiont at opposing extremes) and underpins evolutionary processes such as genome streamlining,

genetic drift, expansion/contraction of mobile elements, pseudogenization, gene loss, and varying mutation rates (González-Pech et al. 2019). Previous studies investigating the effects of symbiogenesis on Symbiodiniaceae genomes have focused almost entirely on symbiotic genera, with the polar-dwelling, highly specialized *Polarella glacialis* (Stephens et al. 2020), a sister of Symbiodiniaceae and within the same Order Suessiales, providing the only free-living outgroup.

Haploid genome sizes of Symbiodiniaceae taxa and *P. glacialis* thus far have been estimated at < 2 Gbp based on available sequencing data (Li et al. 2020; Stephens et al. 2020; González-Pech et al. 2021; Chen et al. 2022), and < 5Gbp based on DNA staining and qPCR of marker sequences (LaJeunesse et al. 2005; Saad et al. 2020). Estimates based on DNA staining are generally larger than those based on sequencing data, likely due to the permanently condensed chromosomal structures of dinoflagellates that result in the overestimation of DNA content (Rizzo 1991, 2003). The diverse dinoflagellate taxa external to the Symbiodiniaceae are predominantly free-living and, in comparison, have massive genome sizes, e.g., 4.8 Gbp estimated from sequencing data for the bloom-forming *Prorocentrum cordatum* (Dougan et al. 2022b), and 200 Gbp based on DNA staining for *Alexandrium tamarense* (Hou and Lin 2009).

Among Symbiodiniaceae taxa, *Effrenium* is the early-diverging, exclusively free-living genus. The sole species, *E. voratum*, is globally distributed in temperate and subtropical waters (Jeong et al., 2014; LaJeunesse et al., 2018). Attempts to establish a symbiotic relationship between *E. voratum* and the anemone *Exaiptasia pallida* have been unsuccessful (Xiang et al. 2013; Gabay et al. 2018). Current understanding of Symbiodiniaceae evolutionary history suggests that *E. voratum* diverged 147 MYA from the basal, largely symbiotic genera of *Symbiodinium* and *Philozoon*, and prior to the other later-diverging symbiotic genera (LaJeunesse et al. 2018). Whereas genomes of other free-living species such as *Symbiodinium natans* (Hansen and Daugbjerg 2009) have been generated, these taxa belong to genera that also include symbiotic species and thus might have experienced a symbiotic lifestyle at some point in their history. We expect the genus *Effrenium* to have remained unaffected by the influence of symbiogenesis, and thus have retained the ancestral free-living lifestyle (and genome features) of Symbiodiniaceae.

102 In this study, we present *de novo* assembled genome and transcriptome data for three
103 isolates of *E. voratum*. Incorporating publicly available genome-scale data from 16
104 Symbiodiniaceae taxa plus four free-living taxa external to the Symbiodiniaceae in a
105 comparative genomic analysis, we examine genomic features in *E. voratum*, including mobile
106 elements, gene structures, gene-families, and pseudogenisation to gain insights into ancestral
107 features of Symbiodiniaceae, and more broadly, Suessiales genome evolution.

108 Results and Discussion

109 Genome-size reduction pre-dated divergence of Order Suessiales and Family 110 Symbiodiniaceae

111 Using a mix of short- and long-read sequencing data (Supplementary Table S1), we generated
112 *de novo* genome assemblies for three isolates of *E. voratum* (assembly sizes 1.1-1.3 Gb), with
113 estimated haploid genome sizes of 1.2-1.9 Gb (Supplementary Table S2, Supplementary
114 Figure S1), completeness (BUSCO recovery 67.2-77.2%), and number of predicted genes
115 (32,102-39,878) (Table 1, Supplementary Tables S3 and S4) comparable to the other
116 Symbiodiniaceae genomes (Shoguchi et al. 2013; Robbins et al. 2019; Stephens et al. 2020;
117 González-Pech et al. 2021; Nand et al. 2021; Chen et al. 2022; Dougan et al. 2022a). We
118 obtained all available genomic data from 23 dinoflagellate taxa: 19 from Symbiodiniaceae
119 (Order Suessiales), two sister taxa of *P. glacialis* (Order Suessiales), and two distantly related
120 free-living dinoflagellate taxa, *Prorocentrum cordatum* (Order Prorocentrales) and
121 *Amphidinium gibbosum* (Order Amphidiniales). The 21 Suessiales taxa were grouped into:
122 (a) the earlier-branching, largely symbiotic genus *Symbiodinium* (S1), (b) the three
123 exclusively free-living *E. voratum* isolates (Ev), (c) the later-branching symbiotic
124 Symbiodiniaceae lineages (S2), and (d) the free-living outgroup *P. glacialis* (Po) sister to
125 Family Symbiodiniaceae (Supplementary Table S3).

126 The most striking feature of genome evolution among Suessiales is the marked
127 reduction in genome size that occurred in the common ancestor of this lineage. Whereas free-
128 living dinoflagellates external to the Suessiales have genomes that range widely in size from
129 ca. 5-200 Gbp, and the parasitic *Amoebophrya ceratii* has a highly reduced genome at 0.1
130 Gbp, all Suessiales genomes have a much narrower size range from 0.7-2.0 Gbp, estimated
131 using sequencing data (Figure 1). This is accompanied by a ~40% loss in gene families prior

132 to the diversification of Suessiales, when compared to their closest dinoflagellate relatives
133 (node N1 in Figure 1). This pattern is reminiscent of the red algae (Rhodophyta), whose
134 common ancestor underwent massive genome reduction, precipitating the loss of canonical
135 eukaryotic features such as flagella-based motility, phytochromes, and autophagy
136 (Bhattacharya et al. 2018). These algae split into two monophyletic lineages, the
137 extremophilic Cyanidiophytina that specialized to life in hot spring environments, and the
138 species-rich mesophilic lineages (e.g., red seaweeds) that inhabit a variety of aquatic
139 environments (Bhattacharya et al. 2018). Most red algae have therefore smaller genomes
140 when compared to the green lineage and have adapted to diverse habitats through gene family
141 evolution and horizontal gene transfer. In an analogous fashion in dinoflagellates, it appears
142 that the Suessiales common ancestor underwent significant genome reduction, likely due to
143 life in extreme habitats (e.g., the psychrophilic *Polarella glacialis*). This streamlining of the
144 gene inventory may have played an important role in driving symbiotic associations with
145 cnidarians that offered nutrient-rich and protected habitats within the animal tissues. The
146 facultative lifestyle was likely retained in most Symbiodiniaceae because it offers the benefit
147 of sexual reproduction during the free-living stage. The most substantial recovery from
148 genome streamlining is offered by whole genome duplication, which has occurred in the
149 *Durusdinium* lineage (Dougan et al. 2022a).

150 **Genome features of *E. voratum* versus early- and later-diverging symbiotic lineages**

151 Genome sequences of the three *E. voratum* isolates share high similarity with 96% sequence
152 identity over 93% of bases (Figure 2A). Repetitive regions containing protein-coding genes
153 were highly conserved relative those of other Suessiales (the Order containing
154 Symbiodiniaceae and the earlier branching sister *P. glacialis*). We identified 98,344 core *k*-
155 mers ($k = 23$; all possible 23-base sequences) that are common in genomes of all Suessiales
156 taxa following Lo et al. (2022), and recovered 95% of core 23-mers in repetitive regions of *E.*
157 *voratum* (Supplementary Figure S2; see Methods). Among the three *E. voratum* genome
158 datasets, we recovered 5-32 putative mitochondrial genome sequences in each dataset that
159 encode the marker genes of *cob*, *cox1* and/or *cox3* (Supplementary Table S5), and 6-32
160 putative plastid genome sequences in each dataset that encode the 16S/23S rRNA or one of
161 the 11 plastid-encoded genes (Barbrook et al. 2014), including a putative empty minicircle
162 sequence (Supplementary Table S6).

163 Given the history of genome reduction, we investigated the traits that may
164 differentiate *Effrenium* from the symbiotic lineages of Symbiodiniaceae. We studied genome
165 size, intron evolution, gene family evolution, pseudogenisation, RNA editing, and
166 phylogenetic signals in Ev, S1 and S2, relative to the outgroup Po. The GC content of coding
167 regions varied among the Symbiodiniaceae lineages (Figure 2B) and were significantly lower
168 in S2 (mean 54.2%; $p < 0.05$) relative to Ev (61.0%) and S1 (57.7%); in intronic regions, the
169 mean GC is 44.6% (S2), 50.1% (Ev) and 50.4% (S1), whereas that of whole-genome
170 sequences is 45.8% (S2), 50.7% (Ev), and 50.6% (S1). GC-rich genome sequences confer
171 bendability to DNA helices (Vinogradov 2003) and may prevent cell freezing or desiccation
172 (Šmarda et al. 2014). Variation of GC content in dinoflagellate genomes does not appear to
173 correlate to lifestyle; among the free-living species external to Symbiodiniaceae, genomes of
174 *P. glacialis* and *A. gibbosum* displayed a mean GC content of 46.4%, similar to S2, whereas
175 genome of *Pr. cordatum* has the highest GC content described thus far for any dinoflagellate,
176 at 59.7% (Dougan et al. 2022b). Intracellular bacteria have a mutational bias towards low
177 genomic GC content, e.g. ~20% (Hershberg and Petrov 2010), but intracellular eukaryotes
178 display both low and high extreme GC content patterns, ranging as low as 24% in the malaria
179 parasite *Plasmodium falciparum* (Nikbakht et al. 2014) to 67% in the green algal symbiont
180 *Chlorella variabilis* in the ciliate *Paramecium* (Blanc et al. 2010). The lower GC content in
181 S2 genomes than the S1 counterparts underscores the dynamic nature of genomic GC content
182 in intracellular eukaryotes.

183 Mobile elements, particularly transposable elements (TEs), can influence genomic
184 architecture and base composition, and have been used to reconstruct the evolutionary history
185 of many species (Osanai-Futahashi et al. 2008; Rayan et al. 2016; Cai et al. 2021). Although
186 facultative Symbiodiniaceae symbionts (i.e., S1 and S2) are expected to contain a larger
187 proportion of mobile elements in their genomes compared to free-living (Ev) lineages
188 (González-Pech et al. 2019), no significant difference among the groups was observed in the
189 overall abundance (Supplementary Table S7, Supplementary Figure S3A) or conservation of
190 TEs (Supplementary Figure S3B). Although more-contiguous genome assemblies generally
191 contained more conserved repeats (Kimura substitution values centred around 3 for *D.*
192 *trenchii* CCMP2556, *C. goreaui* SCF055, the three *E. voratum* isolates, *S. natans*
193 CCMP2548, and *S. tridacnidorum* CCMP2592) compared to the others (i.e., values centred
194 around 20), we note divergent repeats (Kimura values centred around 25) in the chromosome-
195 scale assembly of *S. microadriaticum* CCMP2467 (Supplementary Figure S3B). This result

196 suggests a potential technical bias in the recovery of mobile elements. Despite this issue,
197 based on the proportions for distinct types of mobile elements in each group, we found
198 significantly ($p < 0.05$) more long interspersed nuclear elements (LINEs) in S1 (3.9%)
199 compared to Ev (1.8%) and to S2 (1.9%), with the proportion of the outgroup Po at 8%
200 (Figure 2C). This result suggests that S1 likely retained ancestral LINEs, lending support to
201 the notion of loss of LINEs after the diversification of basal Symbiodiniaceae genera
202 (González-Pech et al. 2021).

203 Introner elements (IEs) are a type of mobile element consisting of inverted and direct
204 repeat motifs found at 5′- and 3′-end of introns in diverse eukaryotes (Worden et al. 2009;
205 van der Burgt et al. 2012; Huff et al. 2016; Gozashti et al. 2022). Recent research on
206 dinoflagellates revealed that IEs are more abundant in free-living (implicating 10-12% of
207 genes) than in symbiotic/parasitic species (0.8-6.0% of genes) (Farhat et al. 2021; Dougan et
208 al. 2022b). Although not as high as in other free-living dinoflagellates, the Ev genomes
209 exhibit more IE-containing genes (5%) than do the S1 (4%) and S2 genomes (3%) (Figure
210 2D). IEs have been postulated to be non-autonomous and their mobility is dependent on
211 transposases encoded in dinoflagellate genomes (Farhat et al. 2021). We recovered
212 transposase protein sequences from most of the Symbiodiniaceae genomes in this study
213 (Supplementary Table S8), suggesting a capacity for IEs to be mobile.

214 Significantly ($p < 0.05$) longer genes were observed in Ev (mean 20 Kb) than S1 (8
215 Kb) and S2 (11 Kb) (Figure 2E), primarily driven by longer intron sizes (introns make up, on
216 average, 92% of a gene [Figure 2F]; sizes peak at 1 Kb [Figure 2G]) and higher intron density
217 per gene (mean of 18 for Ev, 14 for S1, 14 for S2) (Figure 2H). The driving mechanism for
218 this trend may reflect one of two evolutionary scenarios: (a) intron expansion in Ev, or (b)
219 intron contraction in S1/S2. In examples of the first scenario, TE-mediated insertions drive
220 intron expansion and are biased toward the 5′ ends of genes to prevent disrupting functional
221 elements (McCoy and Fire 2020), yielding larger intron sizes at 5′ ends; we did not observe
222 this in Ev (Figure 2I). In the second scenario, which has been observed in
223 endosymbiotic/parasitic organisms, the reduction of intron size and density occurs as a result
224 of genome reduction and/or streamlining induced by spatial confinement in the host organism
225 or cell (Wang et al. 2020a; Cai et al. 2021). Considering the evolutionary history of
226 Symbiodiniaceae, intron contraction in S1/S2 taxa due to their symbiotic lifestyle is more
227 plausible than invoking intron expansion in Ev. Large introns observed in Po (mean 1.4 Kb)

228 (Figure 2G), albeit at a lower intron density per gene (Figure 2H), lend further support to this
229 notion.

230 **Symbiogenesis shaped evolution of gene families and post-transcriptional processing in**
231 **Symbiodiniaceae**

232 To examine the effect of symbiotic lifestyle on protein family evolution in Symbiodiniaceae,
233 we first inferred 53,173 homologous protein families from all 811,611 protein sequences
234 predicted from the 21 Suessiales genomes (see Methods). Most protein families (47,353 of
235 53,173 [89%]) were shared among the four groups (S1+S2+Ev+Po); relative to functions
236 annotated in all families, these families were enriched in functions such as cellular motility,
237 biosynthetic processes for rRNA, antibiotics, and glycosides (Supplementary Table S9).
238 There were more lineage-specific protein families in S1 (6,389) and S2 (4,056) than in Ev
239 (1,734), and the two symbiotic groups (S1+S2) shared 3,357 protein families not found in the
240 other groups (Figure 3). The high number of protein families found only in symbiotic
241 lineages split by over 40 million years hints at convergent evolution from adapting to a
242 symbiotic lifestyle. These protein families were enriched for diverse functions including
243 signalling, apoptosis, protein splicing, photosynthesis, cell adhesion, and various transferase
244 activities (Figure 3). Incidentally, Po shared more protein families with symbiotic lineages
245 (1,290; S1+S2+Po) than with Ev (221; Ev+Po); these families were enriched in functions
246 such as autophagy and microtubule organization.

247 Facultative symbionts in Symbiodiniaceae are expected to display higher levels of
248 pseudogenisation, a major feature of the resident genome syndrome (González-Pech et al.
249 2019). To investigate this, we identified putative pseudogenes in the Symbiodiniaceae
250 genomes (Supplementary Table S10) following González-Pech et al. (2021). We defined the
251 level of pseudogenisation, Ψ, as a ratio of the number of putative pseudogenes to the number
252 of putative functional genes per homologous family; see Methods. We compared Ψ
253 independently for Ev (Ψ_{Ev}) against that for S1 (Ψ_{S1}), S2 (Ψ_{S2}), and the combined S1 and S2
254 (Ψ_{S1+S2}), then identified protein families that exhibited significant difference ($p < 0.05$) of this
255 value. More protein families displayed $\Psi_{S1} > \Psi_{Ev}$ (336) and $\Psi_{S2} > \Psi_{Ev}$ (273; Figure 4A),
256 compared to $\Psi_{S1} < \Psi_{Ev}$ (300) and $\Psi_{S2} < \Psi_{Ev}$ (126; Figure 4B). There were nine-fold more
257 protein families exhibiting $\Psi_{S1+S2} > \Psi_{Ev}$ (229; Figure 4A) than the opposite (25; Figure 4B);

258 these pseudogenes were associated with a wide range of functions, including cell cycle
259 processes and stimuli response (Figures 4C and 4D). The protein families that displayed
260 significantly higher Ψ in the symbiotic lineages were mostly mutually exclusive from the
261 3,357 families that putatively experienced convergent evolution (only 16–23 families were
262 represented in Ψ_{S1}, Ψ_{S2}, Ψ_{S1+S2}). We found negligible technical biases in the clustering of
263 homologous sequences that may affect our inference of pseudogenes, i.e., protein families
264 displaying Ψ remained stable at different clustering parameters (Supplementary Figure S4;
265 see Methods). These results suggest that in addition to convergent evolution in the symbiotic
266 lineages, these lineages of Symbiodiniaceae have experienced a greater extent of
267 pseudogenisation than did Ev.

268 Editing of mRNAs is a known mechanism in Symbiodiniaceae for increasing
269 variability in protein isoforms (Liew et al. 2017). To assess mRNA editing in *E. voratum*, we
270 focused on RCC1521 for which the genome assembly is the most contiguous among the three
271 we generated, and compared the rates and types of mRNA editing against a representative
272 from S1 (*S. microadriaticum* CCMP2467 (Liew et al. 2017)) and S2 (*D. trenchii* CCMP2556
273 (Dougan et al. 2022a)) for which the relevant data were available. We identified 45,009
274 unique mRNA edited sites in *E. voratum*, about 13-fold and 4-fold more than those in *S.*
275 *microadriaticum* and *D. trenchii*, respectively (Supplementary Table S11). A larger
276 proportion of genes in *E. voratum* (9,158, 28.5%) contained mRNA edits, compared to *S.*
277 *microadriaticum* (774, 1.6%) and *D. trenchii* (4,227, 7.6%), although the distributions of
278 substitution types are similar among the three species (Supplementary Figure S5A-D). Most
279 edits in *D. trenchii* and *E. voratum* (> 60%) were located within exons (Supplementary
280 Figure S5E), whereas in *S. microadriaticum* edits were evenly split between exons and
281 introns, although this may be due to the more-fragmented genome assembly of Liew et al.
282 (2017). As also observed in *S. microadriaticum* (Liew et al. 2017), there was a slight bias for
283 edits to occur toward 5′ ends of genes (within the first ~10% of gene length, $p < 0.05$;
284 Supplementary Figure S5F) and the edits tended to be located within 1 Kb of each other
285 (versus expected distribution, $p < 0.05$; Supplementary Figure S5G). The high level of
286 mRNA editing in Ev is consistent with the observation in another free-living dinoflagellate,
287 *Pr. cordatum* (42,969 edited sites; 32,067 within 10,169 [12%] genes) (Dougan et al. 2022b).
288 Together with our observation of fewer Ev-specific protein families than those specific to S1
289 or S2 (Figure 2), these results suggest a more pronounced role of mRNA editing in generating

290 functional diversity in free-living versus symbiotic dinoflagellate taxa, to potentially
291 compensate for fewer specialised gene functions in the former. The correlation and impact of
292 this RNA editing with the final protein products remains to be investigated by incorporating
293 experimental proteome data.

294 **Impact of symbiogenesis on phylogenetic signal of non-coding regions**

295 Our alignment-based phylogenetic trees inferred using multiple protein families (Figure 1),
296 the standard molecular marker of 18S rDNA (Figure 5A), and the ITS2 (Supplementary
297 Figure S6A) were consistent with previously established phylogenies (LaJeunesse et al. 2018;
298 Stephens et al. 2018; Shi et al. 2021). We then analysed the phylogenetic signal of whole-
299 genome sequence data using a *k*-mer-based alignment-free (AF) approach (see Methods),
300 focusing on distinct genome-sequence regions following Lo et al. (2022). Interestingly, the
301 inferred AF phylogenies of introns (Fig. 5B), repetitive regions (Supplementary Figure S6B),
302 repeat-masked whole-genome sequences (Fig. 5C), and entire whole-genome sequences
303 (Supplementary Figure S6C) placed Ev as the basal group, branching earlier than S1/S2. In
304 comparison, the AF phylogenies for coding regions (i.e., coding sequences [CDS; Fig. 5D]
305 and protein sequences [Supplementary Figure S6D]) were largely congruent with the 18S
306 rDNA phylogeny (Fig. 5A), placing S1 as earlier branching than Ev. The branching orders of
307 Ev vs S1/S2 in AF trees were supported by a robust jackknife support of ≥ 96% based on 100
308 subsampled replicates (Figure 5B-D; see Methods). This trend is consistent with visualisation
309 of the AF distances as a phylogenomic network (Supplementary Figure S6E), in which *E.*
310 *voratum* was more closely related to *P. glacialis* based on intron sequences, and to
311 *Symbiodinium* spp. based on CDS.

312 The incongruence between phylogenies of coding and non-coding regions with robust
313 support of the distinct clades clearly indicate differential selective pressure acting on these
314 two regions in Symbiodiniaceae genomes, as demonstrated in an earlier study (Lo et al.
315 2022), against the backdrop of incomplete lineage sorting, horizontal gene transfer,
316 hybridization, or convergent GC-biased gene conversion (Pessia et al. 2012; Jarvis Erich et
317 al. 2014; Chan et al. 2020). In addition, this result may also reflect by the retainment of
318 ancestral non-coding regions in the *E. voratum* genomes, and/or the loss of some non-coding
319 regions in symbiotic lineages due to genome streamlining.

320 **Symbiogenesis differentially affected early- versus late-branching symbiotic lineages of**
321 **Symbiodiniaceae**

322 Following a possible genome reduction in the free-living ancestor of the Suessiales, some
323 lineages may have developed symbiogenesis with a range of hosts, giving rise to the
324 Symbiodiniaceae Family. Common and distinct genomic features we observed between early
325 and late-branching symbiotic lineages of Symbiodiniaceae suggest an interplay between the
326 geological eras during which they arose, and the corresponding coral morphology and ocean
327 chemistry (Figure 6). Ancestral Symbiodiniaceae started inhabiting stony corals presumably
328 as early as 230 MYA in the late Triassic (Frankowiak et al. 2021) and may have driven the
329 Norian-Rhaetian reef bloom (Kiessling et al. 2009). These early Scleractinian corals (e.g.,
330 *Retiophyllia*) tended to be uniserial, i.e., possessing one corallite per branch, and phaceloid
331 with thick walls (Coates and Jeremy 1987; Stanley 2003), and thus were less efficient at
332 harvesting light (Enríquez et al. 2017). The ability of the early-diverging extant
333 *Symbiodinium* to thrive under high or variable light (LaJeunesse et al. 2018) may be a trait
334 inherited from their ancestor living in these ancient corals. Because these early
335 Symbiodiniaceae adapted to different hosts, they likely underwent genome streamlining
336 (González-Pech et al. 2019), experienced high pseudogenisation and a reduction in mRNA
337 editing and intron sizes (Figure 6), as observed here and in other studies (Liew et al. 2017;
338 González-Pech et al. 2021). Although these trends were also observed in the later-branching
339 symbionts, *Symbiodinium* uniquely retained ancestral LINE repeats (Figure 2C) which were
340 lost in later-branching Symbiodiniaceae including *E. voratum.*

341 *E. voratum* is estimated to have diverged ~147 MYA in the late Jurassic (LaJeunesse
342 et al. 2018). We did not find evidence for genome streamlining in *E. voratum*, but instead
343 found genomic hallmarks reminiscent of earlier-branching free-living lineages external to
344 Family Symbiodiniaceae, e.g., more IE-containing genes, larger intron size than in S1 and S2.
345 Given that *Effrenium* is also the sole genus within the Family that is exclusively free-living,
346 we postulate that this lineage has not been impacted by symbiogenesis, in contrast to S1 or
347 S2. Since the lineage diverged from S1, global events such as the breakup of Pangea
348 (Schettino and Turco 2009; Bulois et al. 2018) and the Cretaceous-Paleogene mass extinction
349 occurred, alongside changes in coral reef biomass (Figure 6). Without additional evidence
350 from fossil records or ancient DNA (oldest evidence of Suessiales is *P. glacialis* from just
351 9,000 years ago (De Schepper et al. 2019)), we cannot pinpoint a specific geological event or

352 biological explanation for why or how *Effrenium* may have eluded symbiosis with potential
353 hosts.

354 At the estimated time when S2 lineages diversified (~109 MYA), shallow-water
355 corals were described to be multiserial, flatter, and more efficient at harvesting light
356 (Enríquez et al. 2017), coinciding with a rise in ocean temperature (i.e., "greenhouse Earth",
357 when no continental glaciers exited), and a switch in ocean chemistry from an aragonite sea
358 to a calcite sea, which would have made it difficult for corals to build their aragonite
359 skeletons (Pandolfi and Kiessling 2014) (Figure 6). In contrast, bivalve rudists that could
360 build aragonite or calcite shells (Steuber and Löser 2000) radiated and flourished (Veron
361 2011; Stanley and van de Schootbrugge 2018). They likely harboured photosymbionts (de
362 Winter et al. 2020), presumably ancestral Symbiodiniaceae, given that extant
363 Symbiodiniaceae (e.g., *S. tridacnidorum*) can inhabit modern bivalves (Shoguchi et al. 2018).
364 Although genomes of S2 exhibit lower GC content than those of S1, our results clearly
365 indicate that both S1 and S2 underwent genome streamlining, which led to convergent
366 evolution of protein families associated to diverse functions essential to forming symbiotic
367 association with the host, such as cell signalling, apoptosis, and photosynthesis (Figure 3).
368 We posit that symbiogenesis drove genome evolution in Symbiodiniaceae and elicited
369 common features such as pseudogenisation, lowered mRNA editing and intron contraction,
370 but some features (such as LINE retention and GC content) were affected differently in
371 earlier vs later-branching symbiotic lineages.

372 **Concluding remarks**

373 Our results support a phase of genome reduction that occurred in the Suessiales ancestor.
374 Therefore, the Symbiodiniaceae have smaller genome sizes than most free-living
375 dinoflagellates. The mixture of obligate free-living to facultative lifestyles among
376 Symbiodiniaceae has resulted in divergent paths of genome evolution. Our results
377 demonstrate the retention of ancestral Symbiodiniaceae genome features in *E. voratum* (in
378 contrast to symbiotic lineages) despite multiple emergences of symbiogenesis over the past
379 200 million years, and how symbiogenesis could drive convergent evolution between the
380 earlier and later-branching symbiotic lineages. These observations support the notion that *E.*
381 *voratum* (and by extension the genus *Effrenium*) eluded symbiogenesis in its evolutionary
382 history. Therefore, *Effrenium* presents a useful free-living outgroup for studying the structural

383 and functional genome features of symbiotic Symbiodiniaceae, and the implications of these
384 features on ecology and evolution, including but not limited to host specificity and the
385 "facultativeness" of symbiotic association.

386 # Methods

387 ***E. voratum* cultures**

388 Cell cultures of *E. voratum* RCC1521, rt-383, and CCMP421 were provided by the
389 LaJeunesse Lab in Pennsylvania State University, U.S.A. They were incubated using Daigo's
390 IMK medium at 25°C under light-dark cycles of 14:10 hours. RCC1521 and rt-383 were
391 maintained at the University of Queensland, CCMP421 at Florida International University.

392 **Extraction of genomic DNA**

393 To extract genomic DNA (gDNA) for RCC1521 and rt-383, cells were pelleted by spinning
394 the culture at 300 *g* (5 min, room temperature [RT]). Then the cell pellet was resuspended in
395 100-500 µL (higher volumes for more cells) prewarmed (at 60°C) lysis buffer (100mM Tris-
396 HCl, 20mM EDTA, 4% CTAB (w/v), 1.4 NaCl, 1% 360,000 $g \cdot mol^{-1}$ PVP (w/v), 2% β-
397 mercaptoethanol). This mixture was transferred to liquid nitrogen pre-chilled mortar pestle
398 and ground to a fine powder and high molecular weight (HMW) gDNA was extracted as
399 described in Rai (2022). Briefly, Proteinase K (200 µL of stock at 20 mg/mL) was added,
400 multiple rounds of chloroform:isoamyl alcohol (24:1 v/v) extractions were performed, the
401 CTAB-DNA pellet was washed using ethanol, RNase A (4 µL of stock at 20 mg/mL) was
402 added, and HMW gDNA was captured by incubating with 7.5 M ammonium acetate. The
403 final HMW gDNA was resuspended in Tris-HCl (10 mM, pH 8) prewarmed at 50 °C, then
404 was stored at -20°C until sequencing.

405 *E. voratum* CCMP421 cells were pelleted and snap frozen in liquid nitrogen and
406 ground along with glass beads (diameter 425-600 µm), and the 2×CTAB method described in
407 Stephens et al. (2020) was followed. Briefly, the ground powder was transferred into a lysis
408 buffer (100 mM Tris-HCl pH 8, 20 mM EDTA pH 8, 1.4 M NaCl, 2% w/v CTAB), mixed
409 with RNAse A (final concentration 20 µg/mL) and incubated at 37°C (30 min), and then with
410 Proteinase K (final concentration 120 µg/mL) at 65°C (2h). The standard extractions using

411 phenol-chloroform-isoamyl alcohol (25:24:1 v/v; centrifuged at 14,000 *g*, 5 min, RT) and
412 chloroform-isoamyl alcohol (24:1 v/v; centrifuged at 14,000 *g*, 5 min, RT) were then
413 performed. The DNA was precipitated using chilled isopropanol (18,000 *g*, 5 min, 4°C),
414 washed using chilled 70% ethanol, then stored in Tris-HCl (10 mM, pH 8) until sequencing.

415 **Extraction of total RNA**

416 To extract total RNA from *E. voratum* RCC1521 for Iso-Seq sequencing, cell pellets were
417 lysed via five freeze-thaw cycles with 425-600 µm diameter glass beads added to cell pellet,
418 dipped in liquid nitrogen, vortexed, and thawed at 37°C), and then the QIAGEN RNeasy kit
419 was used following the plant tissue protocol. To increase transcriptome diversity, we
420 extracted more RNA using a different method. Total RNA was extracted for RNA-Seq
421 (RCC1521 and rt-383) and Iso-Seq (from rt-383) following Acosta-Maspons et al. (2019)
422 with slight modifications. Cells were first pelleted by centrifugation (300 *g*, 5 min). The cell
423 pellet was suspended in 100–500 µL lysis buffer (100mM Tris-HCl, 25 mM EDTA, 2%
424 CTAB w/v, 2M NaCl, 0.75 g/L spermidine trihydrochloride, 4% β-mercaptoethanol). This
425 mixture was transferred to a mortar and pestle that were prechilled in liquid nitrogen, and
426 ground to a fine powder. The powder was swirled with liquid nitrogen and transferred to a
427 Falcon tube (15 mL) standing in dry ice, then the liquid nitrogen was allowed to evaporate.
428 Lysis buffer (5 mL; prewarmed at 65°C) was added, mixed by vortexing, and the tube was
429 kept on ice for subsequent steps. An equal volume of chloroform:isoamyl alcohol (24:1 v/v)
430 was mixed in; the mixture was divided into aliquots (2mL) in Eppendorf tubes and
431 centrifuged (10,000 *g*, 10 min, 4°C). The clear supernatants were transferred to new tubes
432 (2mL each) before the extraction with chloroform:isoamyl alcohol (24:1 v/v) was repeated.
433 Nucleic acids were precipitated using LiCl (final concentration 2M) overnight at 4°C, and
434 centrifuged (17,000 *g*, 30 min, 4°C). The resulting pellet was resuspended in the SSTE buffer
435 (50 µL; 1 M NaCl, 0.5% SDS w/v, 10mM Tris-HCl pH 8, 1mM EDTA; prewarmed at 65°C);
436 the samples were not set on ice to avoid undesirable precipitates. Further chloroform:isoamyl
437 alcohol (24:1 v/v) extraction was performed (centrifuged 10,000 *g*, 10 min, 4°C), and the
438 RNA pellet was precipitated using an equal volume of 2-propanol (incubated 10 min at RT,
439 then centrifuged at 17,000 *g*, 10 min, 4°C). The pellet was washed with 80% ethanol (500
440 µL), dislodged by pulse vortexing (2 s), and the tube centrifuged (17,000 *g*, 5 min, 4°C); this
441 step was repeated. The final pellet was air-dried (10 min) in a fume hood, resuspended
442 nuclease-free water (25 µL), and stored at -80°C.

443 **Generation of genome data**

444 For short-read sequencing, the libraries for RCC1521 and rt-383 were prepared using the
445 Illumina TruSeq Nano kit with 350 bp targeted inserts following standard protocol, and then
446 sequenced on the NovaSeq 6000 platform at Australian Genome Research Facility
447 (Melbourne, Australia). For CCMP421, the genomic DNA library was prepared using the
448 Chromium Genome Reagent Kit v2 Chemistry following the manufacturer's protocol (Step 2
449 GEM generation and barcoding onwards) and then sequenced on the NovaSeq 6000 platform
450 at Florida International University.

451 For Nanopore long-read sequencing of RCC1521 and rt-383, the ligation kit SQK
452 LSK-109 was used following standard protocol. Each library was sequenced on a MinION
453 flow cell and rt-383 gDNA was further sequenced using a PromethION flow cell at the
454 Genome Innovation Hub (GIH), the University of Queensland (Brisbane, Australia). The
455 sequence reads were base-called using guppy v4.0.11 as part of the MinKNOW software
456 v20.06.18, with the minimum read quality filter of 7.

457 For PacBio long-read sequencing of RCC1521 and rt-383, the gDNA of RCC1521
458 (unsheared, for continuous long-read [CLR] library) and rt-383 (sheared in 15-20 Kb
459 fragments with Pippin Prep (Sage Science), for HiFi library) were used for library preparation
460 using the SMRTbell Express Template Prep Kit 2.0 following the manufacturer's protocol.
461 The RCC1521 and the rt-383 libraries were sequenced on the PacBio Sequel II platform,
462 respectively at the University of Washington PacBio Sequencing Services (Settle, WA, USA)
463 and at the University of Queensland Sequencing Facility (Brisbane, Australia). The CLR
464 reads for RCC1521 were acquired using the PacBio BAM2fastx toolkit, whereas the HiFi
465 CCS reads for rt-383 were obtained using the CCS module of the SMRT Link pipeline v8.0.

466 **Generation of transcriptome data**

467 Transcriptome data were generated for RCC1521 and rt-383. Illumina RNA-Seq libraries
468 were generated using polyA-selection with the Dynabeads mRNA purification Kit and the
469 Illumina Stranded mRNA Prep following standard protocols. Sequencing was performed on
470 the Illumina NovaSeq 6000 platform at the Australian Genome Research Facility
471 (Melbourne, Australia). PacBio Iso-Seq libraries were prepared using the NEBNext® Single

472 Cell/Low Input cDNA Synthesis and Amplification Module (New England BioLabs) and the
473 SMRTbell Express Template Prep Kit 2.0 following standard protocol and sequenced on the
474 PacBio Sequel II at the University of Queensland Sequencing Facility.

475 **Transcriptome assembly and processing**

476 For RNA-Seq data from RCC1521 and rt-383, adapters and unique molecular identifiers were
477 removed using Illumina's bcl2fastq v2.20.0.422, and the reads were trimmed (for polyG and
478 polyA) and filtered using fastp v0.20.0 (*-A -L 35 -g -x --cut_front --cut_window_size 4 --*
479 *cut_mean_quality 15*). The processed reads were assembled using Trinity v2.9.1 (Grabherr et
480 al. 2011) in both "de novo" (*--SS_lib_type RF --trimmomatic*) and "genome-guided" modes;
481 for the latter, RNA-Seq reads were first mapped to the assembled genome using HISAT2
482 (Kim et al. 2019) before Trinity was run using *--SS_lib_type RF --genome_guided_bam --*
483 *genome_guided_max_intron 70000*. We assessed the completeness of the each transcript set
484 using BUSCO v5.1.2 (Simão et al. 2015) against the alveolata_odb10 database
485 (Supplementary Table S1).

486 Iso-Seq transcripts do not require assembly. The raw Iso-Seq sequences of RCC1521
487 and rt-383 underwent CCS generation and demultiplexing using the standalone modules CCS
488 v4.2.0 and Lima v1.11.0. The rest of the IsoSeq processing steps (i.e., refining, clustering
489 isoforms, and polishing) were conducted using the IsoSeq pipeline v3.3.0, resulting in a final
490 set of high-quality transcripts.

491 **Estimation of genome size from sequencing data**

492 Illumina short reads were used for estimating genome size based on *k*-mers. First, the reads
493 from each genome dataset were processed to remove potential adapters (for 10X linked-reads
494 of CCMP421, specifically the first 23 bases) and polyG tails using fastp v0.20.0 (Chen et al.
495 2018). Jellyfish v2.3.0 (Marçais and Kingsford 2011) was used to obtain *k*-mers of sizes 17-
496 31. The *k*-mer distribution was then plotted for each *k,* and the genome size was estimated as
497 the sum of observed *k*-mers divided by the mean coverage (corresponding to the peak of the
498 curve), averaged from the different *k*-mer sizes (Supplementary Table S2). The ploidy of the
499 genome datasets was assessed using GenomeScope2 (Ranallo-Benavidez et al. 2020) based
500 on 21-mer distribution and a better fit was observed using a haploid (p=1) model in all three

501 cases. Genome size estimations of other dinoflagellates (Supplementary Table S12) were
502 taken from Rizzo et al. (1982), Hou and Lin (2009), Sano and Kato (2009), and Kohli et al.
503 (2015).
504
505 ***De novo* genome assembly**

506 *De novo* genome assemblies were generated for RCC1521 and rt-383 adopting a hybrid-data
507 approach, combining Illumina (short reads), PacBio (long reads), and Nanopore (long reads)
508 data using MaSuRCA (Zimin et al. 2013; Wang et al. 2020b) v4.0.1 for RCC1521, and v3.4.2
509 for rt-383, with the built-in CABOG as the final assembler; the key distinction between these
510 two versions of MaSuRCA is the six-fold decrease in run-time in v4.0.1 relative to that for
511 v3.4.2, with negligible impact on the yielded assemblies. For CCMP421, the *de novo* genome
512 assembly was generated from 10X Genomics linked-read sequencing data using Supernova
513 v2.1.1 (Weisenfeld et al. 2017).

514 RCC1521 and rt-383 assemblies were further scaffolded with L_RNA_scaffolder
515 (Xue et al. 2013), using IsoSeq transcripts and *de novo* assembled transcripts from RNA-Seq
516 (above). For CCMP421, linked-read distance information was first used to refine the
517 assembly with ARBitR v0.2 (*-m 27k -s 10k*) (Hiltunen et al. 2020) prior to scaffolding with
518 L_RNA_scaffolder (Xue et al. 2013). Due to the low quality of the publicly available
519 transcriptome data of CCMP421 (i.e., MMETSP1110 (Johnson et al. 2019) with only 54%
520 mapped to the corresponding assembled genome; Supplementary Table S13), we used the *de*
521 *novo* assembled transcripts from RCC1521 and rt-383 to scaffold the CCMP421 genome
522 assembly via L_RNA_scaffolder.

523 To ensure high quality of each genome assembly, we identified and removed potential
524 contaminant sequences of bacterial or archaeal sources following a decision tree based on
525 analysis using BlobTools v1.1 (Zimin et al. 2017) as described in earlier studies of algal
526 genomes (Iha et al. 2021; Chen et al. 2022); this step yielded the final assembly for each
527 isolate. For each assembly, we assessed data completeness using BUSCO v5.1.2 (Simão et al.
528 2015) against the alveolata_odb10 database. Genome-sequence similarity among the three *E.*
529 *voratum* isolates was assessed using nucmer implemented in the MUMmer package
530 v4.0.0beta2 (*--mum*) (Marçais et al. 2018).

531 **Identification of mitochondrial and plastid genome sequences**

532 We identified putative mitochondrial and plastid genome sequences from the three isolates of
533 *E. voratum*. To search for mitochondrial scaffolds, we followed Stephens et al. (2020) by
534 adopting BLASTn v2.10.0+ (Altschul et al. 1990) search against the assembled genomes,
535 using the protein-coding sequences of *B. minutum* mitochondrial genes as queries (GenBank
536 accessions LC002801.1 and LC002802.1; $E \leq 10^{-10}$). We used BEDtools merge (Quinlan and
537 Hall 2010) to merge overlapping BLAST hits of the same gene *(-s -o collapse -c 1,2,3,4,5,6*),
538 then used BEDtools intersect (*-wa -wb*) to check for overlaps with the predicted gene
539 annotations. Genome scaffolds with BLAST hits that did not have any other predicted genes
540 were considered putative mitochondrial genomes. We manually annotated the genes on the
541 putative genome scaffolds using Artemis (Carver et al. 2011) with translation table 4.

542 To identify plastid genome fragments that are known to be shorter and may not be
543 recovered in a hybrid genome assembly combing long- and short-read sequence data, we
544 performed an independent short-read only genome assembly for each isolate using CLC
545 Genomics Workbench v21.0.4 (Supplementary Table S14). We used BLASTn search ($E \leq$
546 10^{-10}) using protein-coding sequences of plastid-encoded genes for *Cladocopium* sp. C3
547 (GenBank accessions HG515015.1-HG515028.1) as query, and annotated the putative
548 genome scaffolds in Artemis (translation table 11). To look for empty minicircles, we first
549 determined the core region of all plastid genome scaffolds. We masked coding regions of the
550 scaffolds using BEDtools *maskfasta* and did pairwise BLASTn searches. The region that was
551 common to most scaffolds was considered the core region. We then used this as query in
552 BLASTn searches among unannotated *E. voratum* short-read genome scaffolds. A scaffold
553 was considered empty if they did not have any hits in the NCBI nr database (November
554 2020). Finally, we looked for evidence of circularisation in all of the recovered organellar
555 genome scaffolds using nucmer (*--mum -l 0*) and mummerplot (*--layout -- png*) from
556 MUMmer 4.0.0beta2 (Marçais et al. 2018) to self-align the genome sequences and visualise
557 the alignments.

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

559 To predict protein-coding genes, we used an integrated, multi-method workflow customised
560 for dinoflagellates (Chen et al. 2022) incorporating transcriptome and protein evidence

561 (pipeline available at
562 https://github.com/TimothyStephens/Dinoflagellate_Annotation_Workflow). We first
563 predicted repetitive elements from the genome assembly, followed by gene predictions based
564 on a) the genome only, b) protein sequence similarity, and c) mRNA transcripts.

565 *De novo* repeat families were predicted from the genome assembly using
566 RepeatModeler v2.0.1 (Flynn et al. 2020), and these repeats were added to the Dfam database
567 (dfam.org; downloaded June 2019) to guide RepeatMasker v4.1.0 (Smit et al. 2017) in
568 repeat-masking the genome assembly. Next, GeneMark-ES v4.65 (Lomsadze et al. 2005) was
569 used for *ab initio* gene prediction on the masked genome assembly. Protein-based gene
570 prediction was performed using MAKER v2.31.10 (*protein2genome*) (Cantarel et al. 2008)
571 modified to recognize dinoflagellate alternative splice sites. It integrated the custom repeat
572 library from the repeat analysis step, and BLASTn/x (Camacho et al. 2009) searches were
573 performed against the combined protein sequence database of SwissProt (released March
574 2020) and the Suessiales sequences, hereby "Suessiales_pep", listed in Table S7 of Chen et
575 al. (2022).

576 For transcript-based gene prediction, Iso-Seq transcripts where available (i.e., for
577 RCC1521 and rt-383) were mapped on the corresponding genome assembly using minimap2
578 v2.18 (Li 2018) for which the code was modified to recognize dinoflagellate alternative
579 splice sites, using options *--secondary=no -ax splice:hq -uf --splice-flank=no.* The assembled
580 transcripts from RNA-Seq (both *de novo* and genome-guided) for each isolate were mapped
581 to the corresponding genome assembly using BLAT (Kent 2002); for CCMP421, *de novo*
582 assembled transcripts from RCC1521 and rt-383, plus the assembly of these reads guided by
583 the CCMP421 genome, were used in this step (due to poor quality of the publicly available
584 CCMP421 transcriptome data; see Supplementary Tables S1 and S13). The resulting GFF3
585 files were input into PASA v2.4.1 (*--IMPORT_CUSTOM_ALIGNMENTS_GFF3 --*
586 *transcribed_is_aligned_orient -C -R --MAX_INTRON_LENGTH 70000*) (Haas et al. 2003)
587 which was modified to recognize dinoflagellate alternative splice sites.

588 The PASA-predicted genes were filtered in the following steps: they were searched
589 against the combined RefSeq (release 98) and the Suessiales_pep database using BLASTp
590 v2.2.26 (e-value < 10^{-20}, both query and subject coverage > 80%) (Camacho et al. 2009),
591 putative transposon sequences were removed via running HHBLITS v3.3.0 (Remmert et al.

592 2012) and TransposonPSI v1.0.0 (Haas 2010) against the UniRef30_3030_03 database from
593 Uniclust (uniclust.mmseqs.com) (Mirdita et al. 2017), redundant sequences were removed
594 using CD-HIT v4.8.1 (*-c 0.75 -n 5*) (Li and Godzik 2006), and the script
595 *Prepare_golden_genes_for_predictors.pl* from the JAMg pipeline
596 (https://github.com/genomecuration/JAMg) was used to produce a highly curated set of
597 "golden genes". These golden genes were used to guide the *ab initio* gene prediction tools
598 SNAP (Korf 2004) and AUGUSTUS v3.4.0 (Stanke et al. 2006) on the repeat-masked
599 genome assembly. Gene models from the five tools (MAKER, GeneMark-ES, PASA, SNAP,
600 AUGUSTUS) were integrated using EVidenceModeler v1.1.1 (Haas et al. 2008). Finally, the
601 resulting gene models were refined to correct exon boundaries, add UTRs, and incorporate
602 gene models from alternative splicing using the *Load_Current_Gene_Annotations.dbi* and
603 *Launch_PASA_pipeline.pl* from the PASA pipeline (Campbell et al. 2006; Haas 2008)
604 iteratively for three rounds to yield the final gene models.

605 We assessed the completeness of the predicted protein sequences using BUSCO
606 v5.1.2 (Simão et al. 2015) against the alveolata_odb10 database in "protein" mode. Genes
607 with transcript support were identified by aligning transcripts for each isolate to their
608 predicted coding sequences using BLASTn (e-value < 10^{-5}, percent identity ≥ 90 %, subject
609 cover ≥ 50 %).

610 **Functional annotation of predicted genes**

611 Functions of the predicted protein sequences were annotated based on BLASTp searches (e-
612 value < 10^{-5} and query/subject cover ≥ 50 %) against the SwissProt (2022_01) database.
613 Then, those that had no hits were searched against TrEMBL (2022_01) database; the function
614 of the top protein hit was assumed to the putative function of the query protein. The UniProt
615 IDs were converted to Gene Ontology (GO) terms via the UniProtKB ID mapping tool
616 (https://www.uniprot.org/id-mapping) in December 2022.

617 **Inferring phylogenetic tree based on multiple sequence alignment**

618 To infer phylogenies for the 18S rDNA and ITS2 marker sequences, we first recovered 18S
619 and ITS2 sequences from the three genome assemblies of *E. voratum* using BLASTn.
620 Reference sequences for the 18S rDNA were downloaded from

621 https://doi.org/10.5061/dryad.1717129 (79 sequences) (LaJeunesse et al. 2018) and ITS2
622 from SymPortal (https://symportal.org; "published post-MED sequences" of 8,409 sequences
623 downloaded 10 September 2021) (Hume et al. 2019), respectively. For each marker sequence
624 set, multiple sequence alignment was generated using MAFFT v7.471 in *mafft-linsi* mode
625 (Katoh and Standley 2013), trimmed using trimAl v1.4.rev15 *(-automated1*) (Capella-
626 Gutiérrez et al. 2009), from which a maximum-likelihood phylogenetic tree was inferred
627 using IQ-TREE v2.1.3 *(-nm 2000 -bb 2000 -m MFP*) (Nguyen et al. 2015).

628 To reconstruct a reference species tree of dinoflagellates based on strictly orthologous
629 protein sequences, we incorporated 1,603,073 predicted protein sequences from 33
630 dinoflagellate taxa, comprising 21 Suessiales taxa (including the three *E. voratum* isolates in
631 this study) and 12 other taxa external to Suessiales (Supplementary Table S12) (John et al.
632 2019; Johnson et al. 2019; Beedessee et al. 2020). These sequences were clustered into
633 homologous sets using OrthoFinder v2.5.4 (Emms and Kelly 2019), from which a species
634 tree was estimated from strictly orthologous sets.

635 **Alignment-free phylogenetic inference and core *k*-mers**

636 We used an alignment-free approach to infer phylogenetic relationships from (a) whole-
637 genome sequences (WGS) and from distinct genomic regions of (b) repeat-masked WGS, (c)
638 coding sequences (CDS), (d) introns, (e) annotated repeats, and (f) predicted protein
639 sequences. Each of these distinct regions were extracted from assembled genome sequences
640 using *gff3_file_to_feature_files.pl* implemented in PASA (Haas et al. 2003). We followed Lo
641 et al. (2022) to identify optimal *k*-mer length (*k*) for each of these datasets. Briefly, for each
642 dataset, *k*-mers at varied length *k* were enumerated using Jellyfish v2.3.0 (Marçais and
643 Kingsford 2011); for all datasets, odd-numbered *k* between 13 and 27 were used, except for
644 the repeats dataset (odd-numbered *k* values between 13 and 51 were used) and protein
645 sequences (odd-numbered *k* values between 3 and 9 were used). For each dataset except the
646 protein sequences, the optimal *k* was determined based on the cumulative proportion of
647 unique *k*-mers and the cumulative proportion of distinct *k*-mers, at the point when
648 distributions of both proportions reached a plateau (Supplementary Figure S7); *k* value
649 determined this way was found to yield the greatest distinguishing power for phylogenetic
650 analysis (Greenfield and Roehm 2013). For protein sequences, we inferred alignment-free
651 phylogenies (see below) from each *k* and chose the *k* with a topology that best matched the

652 18S rDNA alignment-based tree as implemented in Lo et al. (2022). The optimal *k* was
653 identified as 23 for WGS and repeat-masked WGS, 19 for CDS, 21 for introns, 51 for
654 annotated repeats, and 9 for protein sequences. Jellyfish v2.3.0 was used to extract *k*-mers at
655 the corresponding optimal *k* length for each dataset, with the option *-C* used to enforce
656 strand-specific directionality for the intron and CDS datasets.
657 To infer alignment-free (AF) phylogenies based on *k*-mers (i.e. using the optimal *k* for
658 each corresponding dataset identified above), we derived pairwise distance based on D_2^S
659 statistic (Reinert et al. 2009) following Chan et al. (Chan et al. 2014), using *d2ssect*
660 (https://github.com/bakeronit/d2ssect). These pairwise distances were then used to infer a
661 phylogenetic tree using *neighbor* implemented in PHYLIP v3.698 (Felsenstein 2005). For
662 each tree, we assessed node support based on jackknife analysis of 100 "pseudo-replicates"
663 following Bernard et al. (2016). In each pseudo-replicate, 40% of the data, in iteratively
664 subsampled 100-bp regions at random, were deleted using the Python script *jackknife.py* from
665 which an AF tree was inferred; the R script *Jackknife.r* was then used to calculate jackknife
666 support value in percentage, among the pseudo-replicate trees, for each node in the original
667 AF tree. These scripts are available at https://github.com/chanlab-genomics/alignment-free-
668 tools.
669 To identify core *k*-mers (Bernard et al. 2018) that are shared by all 21 Suessiales
670 genomes used in this study, we used the optimal $k = 23$ for the WGS dataset. Using the
671 extracted 23-mers from the entire WGS dataset as input, core 23-mers were identified using
672 the bash command *comm* (-12). BEDtools (Quinlan and Hall 2010) *intersect* was used to find
673 regions of overlap between the core *k*-mers and different genomic features.

674 **Analysis of gene family evolution**

675 To examine the gene family evolution between *E. voratum* and the earlier-/later-branching
676 symbiotic lineages of Symbiodiniaceae, we grouped 21 Suessiales protein sequence datasets
677 into three groups: Ev, S1, and S2 (Supplementary Table S3), with Po (*P. glacialis*
678 CCMP1383 and CCMP2088) as the outgroup. The 811,661 protein sequences from the 21
679 Suessiales taxa were clustered into homologous families using OrthoFinder v2.5.4 (Emms
680 and Kelly 2019). We then subset the shared/exclusive protein families among the different
681 groups (Ev, S1, S2, and Po). GO enrichment was performed using the topGO package in R
682 (algorithm = "elim", statistic = "fisher") (Alexa & Rahnenfuhrer 2021) for six comparisons:
683 (a) shared genes in Ev+Po (test set) versus all genes in Ev+Po (background), (b) shared genes

684 in S1+S2 versus all genes in S1 and S2, (c) shared genes in S1+S2+Po versus all genes in S1,
685 S2, and Po, (d) shared genes in S1+S2+Ev+Po versus all genes in the 21 taxa, (e) exclusive
686 genes to S1 versus all S1 genes, and (f) exclusive genes to Ev versus all Ev genes.

687 **Identification of pseudogenes**

688 Pseudogenes were identified following the method described in González-Pech et al. (2021)
689 based on tBLASTn search using the predicted protein sequences as query against the
690 corresponding genome sequences for which the predicted gene model sequences were
691 masked. Matched regions (≥ 75% identity) were considered fragments of pseudogenes, and
692 fragments at no more than 1 Kb apart and in the same orientation were considered
693 collectively as a pseudogene.

694 In this analysis, we focused on 752,954 protein sequences from the 19 Suessiales taxa,
695 specifically excluding *S. natans* and *S. pilosum* from S1 to avoid signatures of free-living
696 lifestyle in these taxa interfering with potential signatures of symbiogenesis. The protein
697 sequences were first clustered into homologous families using OrthoFinder v2.5.4 (Emms
698 and Kelly 2019). We define the extent of pseudogenization, Ψ, as the ratio of the number of
699 putative pseudogenes to the number of putative functional genes in a homologous family. We
700 determined this value independently for Ev (Ψ_{Ev}) against that for S1 (Ψ_{S1}), S2 (Ψ_{S2}), and the
701 combined S1 and S2 (Ψ_{S1+S2}); a protein family with $\Psi_{S1} > \Psi_{Ev}$ indicates a greater extent of
702 pseudogenisation in S1 than in Ev. We then ran a one-way ANOVA test and Tukey's test
703 using the R package *rstatix* (https://cran.r-project.org/package=rstatix) to assess correlation
704 among the Ψ values of Ev, S1, S2, and the combined S1+S2 groups.

705 To assess the robustness of our approach to the inflation parameter (*I*) in single-
706 linkage clustering within OrthoFinder that modulates granularity (i.e., sizes and numbers) of
707 the resulting protein families, we ran OrthoFinder independently using $I = 1.1$, 1.3, 1.7 and
708 2.0 in addition to the default value of 1.5, and performed the analysis of pseudogenisation per
709 above.

710 **Analysis of mRNA editing**

711 For the analysis of mRNA editing, we focused on *E. voratum* RCC1521, and the
712 representative genomes for S1 (*S. microadriaticum* CCMP2467) and S2 (*D. trenchii*
713 CCMP2556), for which high-quality genome and transcriptome data are available. Editing of
714 mRNAs was identified using JACUSA2 (Piechotta et al. 2017), based on observed nucleotide
715 variants in the mapping of transcripts onto the genome, relative to the mapping of genome
716 sequence reads onto the genome. First, gDNA reads were mapped on each genome using
717 BWA-mem v0.7.17-r1198 (Li and Durbin 2009) using default settings. Then, for each
718 genome, RNA-Seq reads were mapped using HISAT2 v2.2.0 *(--rna-strandness RF*) (Kim et
719 al. 2019) to the assembled genome sequences with a HGFM HISAT2 index (*hisat2-build --*
720 *exon --ss*) informed by the annotated splice sites. To generate the index, the gene annotation
721 file in GFF3 was converted to the GTF format using Gffread (Pertea and Pertea 2020), and
722 splice site and exon locations extracted with the *hisat2_extract_splice_sites.py* and
723 *hisat2_extract_exons.py* scripts. Iso-Seq reads, where available, were mapped using
724 minimap2 v2.18 (Li 2018) for which the code was modified to recognize alternative splice
725 sites of dinoflagellates, with options *--splice-flank=no –secondary=no -ax splice:hq -uf –*
726 *junc-bed*.

727 Duplicate mappings were removed from the gDNA BAM files using Picard
728 MarkDuplicates (*ASSUME_SORTED=true REMOVE_DUPLICATES=true*
729 *CREATE_INDEX=TRUE VALIDATION_STRINGENCY=LENIENT*). The MD field
730 documenting mismatched and deleted bases was added to the gDNA BAM files with
731 samtools calmd (*-b*), as required as input for JACUSA2. JACUSA2 analysis was performed
732 on the gDNA, RNA-Seq, and Iso-Seq BAM files using option *-a D,Y,H* to remove false
733 positives caused by read starts/ends, indels, splice sites, and homopolymers. Due to the
734 different strand directionality, *-P2 RF-FIRSTSTRAND* was specified for the runs
735 incorporating RNA-Seq data, and *-P2 FR-SECONDTRAND* for the runs using Iso-Seq data.
736 The results from JACUSA2 were overlayed using BEDtools intersect (Quinlan and Hall
737 2010) at *-s -wo* with predicted genes (including isoforms), and the edited sites were
738 visualized using JACUSA2helper (https://github.com/dieterich-lab/JACUSA2helper).

739 We followed Liew et al. (2017) to assess 5′ bias in the location of RNA editing and
740 the propensity for edits to occur together. We calculated the frequency of the locations of

741 each edit with respect to the features they were in (exon, intron, gene), normalized by the
742 length of each feature. For each edit, its distance (in bp) to the closest upstream edit, and that
743 to the closest downstream edit where all within the same gene, were determined. The average
744 of these two values was used as the per-edit observed distance to neighbouring edits.

745 **Analysis of introner elements**

746 To identify introner elements (IEs), we used Pattern Locator (Mrázek and Xie 2006) to search
747 for the patterns described in Farhat et al. (2021): inverted repeats of 8-20 nucleotides within
748 30 bases of the 5′ and 3′ ends of each intron, flanked by direct repeats of 3-5 nucleotides. We
749 first used Seqkit (Shen et al. 2016) to obtain the first and last 30 bases at each end of introns,
750 then used Pattern Locator to identify the IEs.

751 Data availability statement

752 The assembled genome, predicted gene models and proteins, and the identified organellar
753 genome sequences from this study are available at https://doi.org/10.48610/1f0377a.

754 Competing interests

755 Authors declare that they have no competing interests.

756 Author contributions

757 Conceptualization, SS, KED, DB and CXC; methodology, SS, KED, YC, SKR, AJB, VM
758 and CXC; formal analysis, SS, KED, YC, RL, GL, MDAF, and VM; investigation, SS, KED;
759 resources, SS, SKR, AJB, MRL, CXC, writing—original draft preparation, SS; writing—
760 review and editing, SS, KED, DB, and CXC; visualisation, SS; supervision, KED, DB, CXC;
761 funding acquisition, MRL, DB and CXC. All authors have read and agreed to the published
762 version of the manuscript.

763 Funding

764 This research was supported by the University of Queensland Research Training Program
765 scholarship (SS and YC), the Australian Research Council grant DP19012474 awarded to

CXC and DB, the University of Queensland Genome Innovation Hub Collaborative Research grant awarded to CXC, and the NSF-IOS CAREER (1453519) grant awarded to MRL. DB was also supported by NSF grant NSF-OCE 1756616 and a NIFA-USDA Hatch grant (NJ01180).

Acknowledgements

This project is supported by high-performance computing facilities at the National Computational Infrastructure (NCI) National Facility systems through the NCI Merit Allocation Scheme (Project d85) awarded to CXC, the University of Queensland Research Computing Centre, and computing facility at the Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences at the University of Queensland.

References

Acosta-Maspons A, González-Lemes I, Covarrubias AA. 2019. Improved protocol for isolation of high-quality total RNA from different organs of *Phaseolus vulgaris* L. BioTechniques. 66:96-98.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology. 215:403-410.

Barbrook AC, Voolstra CR, Howe CJ. 2014. The chloroplast genome of a *Symbiodinium* sp. clade C3 isolate. Protist. 165:1-13.

Beedessee G, Kubota T, Arimoto A, Nishitsuji K, Waller RF, Hisata K, Yamasaki S, Satoh N, Kobayashi Ji, Shoguchi E. 2020. Integrated omics unveil the secondary metabolic landscape of a basal dinoflagellate. BMC Biology. 18:139.

Bernard G, Chan CX, Ragan MA. 2016. Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. Scientific Reports. 6:28970.

Bernard G, Greenfiled P, Ragan MA, Chan CX, Claesson MJ. 2018. *k*-mer similarity, networks of microbial genomes, and taxonomic rank. mSystems. 3:e00257-18.

Bhattacharya D, Qiu H, Lee J, Su Yoon H, Weber APM, Price DC. 2018. When less is more: red algae as models for studying gene loss and genome evolution in eukaryotes. Critical Reviews in Plant Sciences. 37:81-99.

795 Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, Lindquist E, Lucas S,
796 Pangilinan J, Polle J, et al. 2010. The *Chlorella variabilis* NC64A genome reveals adaptation
797 to photosymbiosis, coevolution with viruses, and cryptic sex. The Plant Cell. 22:2943-2955.

798 Bulois C, Pubellier M, Chamot-Rooke N, Delescluse M. 2018. Successive rifting events in
799 marginal basins: the example of the Coral Sea region (Papua New Guinea). Tectonics. 37:3-
800 29.

801 Cai L, Arnold BJ, Xi Z, Khost DE, Patel N, Hartmann CB, Manickam S, Sasirat S, Nikolov
802 LA, Mathews S, et al. 2021. Deeply altered genome architecture in the endoparasitic
803 flowering plant *Sapria himalayana* Griff. (Rafflesiaceae). Current Biology. 31:1002-
804 1011.e1009.

805 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009.
806 BLAST+: architecture and applications. BMC Bioinformatics. 10:421.

807 Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR. 2006. Comprehensive analysis
808 of alternative splicing in rice and comparative analyses with *Arabidopsis*. BMC Genomics.
809 7:327.

810 Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sánchez Alvarado A,
811 Yandell M. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model
812 organism genomes. Genome Research. 18:188-196.

813 Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated
814 alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25:1972-1973.

815 Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. 2011. Artemis: an integrated
816 platform for visualization and analysis of high-throughput sequence-based experimental data.
817 Bioinformatics. 28:464-469.

818 Chan CX, Bernard G, Poirion O, Hogan JM, Ragan MA. 2014. Inferring phylogenies of
819 evolving sequences without multiple sequence alignment. Scientific Reports. 4:6504.

820 Chan KO, Hutter CR, Wood PL, Grismer LL, Brown RM. 2020. Larger, unfiltered datasets
821 are more effective at resolving phylogenetic conflict: introns, exons, and UCEs resolve
822 ambiguities in golden-backed frogs (Anura: Ranidae; genus *Hylarana*). Molecular
823 Phylogenetics and Evolution. 151:106899.

824 Chang FH. 1983. Winter phytoplankton and microzooplankton populations off the coast of
825 Westland, New Zealand, 1979. New Zealand Journal of Marine and Freshwater Research.
826 17:279-304.

827 Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor.
828 Bioinformatics. 34:i884-i890.

829 Chen Y, Shah S, Dougan KE, van Oppen MJH, Bhattacharya D, Chan CX. 2022. Improved
830 *Cladocopium goreaui* genome assembly reveals features of a facultative coral symbiont and
831 the complex evolutionary history of dinoflagellate genes. Microorganisms. 10:1662.

832 Coates AG, Jeremy BCJ. 1987. Clonal growth, algal symbiosis, and reef formation by corals.
833 Paleobiology. 13:363-378.

834 De Schepper S, Ray JL, Skaar KS, Sadatzki H, Ijaz UZ, Stein R, Larsen A. 2019. The
835 potential of sedimentary ancient DNA for reconstructing past sea ice evolution. The ISME
836 Journal. 13:2566-2577.

837 de Winter NJ, Goderis S, Van Malderen SJM, Sinnesael M, Vansteenberge S, Snoeck C,
838 Belza J, Vanhaecke F, Claeys P. 2020. Subdaily-scale chemical variability in a *Torreites*
839 *sanchezi* rudist shell: implications for rudist paleobiology and the Cretaceous day-night cycle.
840 Paleoceanography and Paleoclimatology. 35:e2019PA003723.

841 Dougan KE, Bellantuono AJ, Kahlke T, Abbriano RM, Chen Y, Shah S, Granados-Cifuentes
842 C, van Oppen MJH, Bhattacharya D, Suggett DJ, et al. 2022a. Whole-genome duplication in
843 an algal symbiont serendipitously confers thermal tolerance to corals. bioRxiv.

844 Dougan KE, Deng Z-L, Wöhlbrand L, Reuse C, Bunk B, Chen Y, Hartlich J, Hiller K, John
845 U, Kalvelage J, et al. 2022b. Multi-omics analysis reveals the molecular response to heat
846 stress in a "red tide" dinoflagellate. bioRxiv.

847 Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative
848 genomics. Genome Biology. 20:238.

849 Enríquez S, Méndez ER, Hoegh-Guldberg O, Iglesias-Prieto R. 2017. Key functional role of
850 the optical properties of coral skeletons in coral ecology and evolution. Proceedings of the
851 Royal Society B: Biological Sciences. 284:20161667.

852 Farhat S, Le P, Kayal E, Noel B, Bigeard E, Corre E, Maumus F, Florent I, Alberti A, Aury J-
853 M, et al. 2021. Rapid protein evolution, organellar reductions, and invasive intronic elements
854 in the marine aerobic parasite dinoflagellate *Amoebophrya* spp. BMC Biology. 19:1.

855 Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package). Department of Genome
856 Sciences, University of Washington, U.S.A. Available from:
857 https://evolution.gs.washington.edu/phylip.html

858 Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020.
859 RepeatModeler2 for automated genomic discovery of transposable element families.
860 Proceedings of the National Academy of Sciences of the United States of America. 117:9451.

861 Frankowiak K, Roniewicz E, Stolarski J. 2021. Photosymbiosis in Late Triassic scleractinian
862 corals from the Italian Dolomites. PeerJ. 9:e11062.

863 Gabay Y, Weis VM, Davy SK. 2018. Symbiont identity influences patterns of symbiosis
864 establishment, host growth, and asexual reproduction in a model cnidarian-dinoflagellate
865 symbiosis. The Biological Bulletin. 234:1-10.

866 González-Pech RA, Bhattacharya D, Ragan MA, Chan CX. 2019. Genome evolution of coral
867 reef symbionts as intracellular residents. Trends in Ecology & Evolution. 34:799-806.

868 González-Pech RA, Stephens TG, Chen Y, Mohamed AR, Cheng Y, Shah S, Dougan KE,
869 Fortuin MDA, Lagorce R, Burt DW, et al. 2021. Comparison of 15 dinoflagellate genomes
870 reveals extensive sequence and structural divergence in family Symbiodiniaceae and genus
871 *Symbiodinium*. BMC Biology. 19:73.

872 Gozashti L, Roy SW, Thornlow B, Kramer A, Ares M, Corbett-Detig R. 2022. Transposable
873 elements drive intron gain in diverse eukaryotes. Proceedings of the National Academy of
874 Sciences. 119:e2209766119.

875 Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L,
876 Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq
877 data without a reference genome. Nature Biotechnology. 29:644.

878 Greenfield P, Roehm U. 2013. Answering biological questions by querying *k*-mer databases.
879 Concurrency and Computation: Practice and Experience. 25:497-509.

880 Grossman EL, Joachimski MM. 2022. Ocean temperatures through the Phanerozoic
881 reassessed. Scientific Reports. 12:8938.

882 Guerrero R, Margulis L, Berlanga M. 2013. Symbiogenesis: the holobiont as a unit of
883 evolution. International Microbiology. 16:133-143.

884 Haas BJ. 2008. Analysis of alternative splicing in plants with bioinformatics tools. In: Reddy
885 ASN, Golovkin M, editors. Nuclear pre-mRNA Processing in Plants. Berlin, Heidelberg:
886 Springer Berlin Heidelberg. p. 17-37.

887 Haas BJ. 2010. TransposonPSI. Broad Institute. Available from:
888 https://sourceforge.net/projects/transposonpsi/

889 Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI, Maiti R,
890 Ronning CM, Rusch DB, Town CD, et al. 2003. Improving the *Arabidopsis* genome
891 annotation using maximal transcript alignment assemblies. Nucleic Acids Research. 31:5654-
892 5666.

893 Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR.
894 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the
895 Program to Assemble Spliced Alignments. Genome Biology. 9:R7.

896 Hansen G, Daugbjerg N. 2009. *Symbiodinium natans* sp. nov.: a "free-living" dinoflagellate
897 from tenerife (Northeast-Atlantic Ocean). Journal of Phycology. 45:251-263.

898 Hershberg R, Petrov DA. 2010. Evidence that mutation Is universally biased towards AT in
899 bacteria. PLOS Genetics. 6:e1001115.

900 Hiltunen M, Ryberg M, Johannesson H. 2020. ARBitR: an overlap-aware genome assembly
901 scaffolder for linked reads. Bioinformatics. 37:2203-2205.

902 Hou Y, Lin S. 2009. Distinct gene number-genome size relationships for eukaryotes and non-
903 eukaryotes: gene content estimation for dinoflagellate genomes. PLOS ONE. 4:e6978.

904 Huff JT, Zilberman D, Roy SW. 2016. Mechanism for DNA transposons to generate introns
905 on genomic scales. Nature. 538:533-536.

906 Hume BCC, Smith EG, Ziegler M, Warrington HJM, Burt JA, LaJeunesse TC, Wiedenmann
907 J, Voolstra CR. 2019. SymPortal: A novel analytical framework and platform for coral algal
908 symbiont next-generation sequencing ITS2 profiling. Molecular Ecology Resources.
909 19:1063-1080.

910 Iha C, Dougan KE, Varela JA, Avila V, Jackson CJ, Bogaert KA, Chen Y, Judd LM, Wick R,
911 Holt KE, et al. 2021. Genomic adaptations to an endolithic lifestyle in the coral-associated
912 alga *Ostreobium*. Current Biology. 31:1393-1402.

913 Jarvis Erich D, Mirarab S, Aberer Andre J, Li B, Houde P, Li C, Ho Simon YW, Faircloth
914 Brant C, Nabholz B, Howard Jason T, et al. 2014. Whole-genome analyses resolve early
915 branches in the tree of life of modern birds. Science. 346:1320-1331.

916 Jeong HJ, Lee SY, Kang N, Yoo Y, Lim AS, Lee MJ, Yih W, Yamashita H, LaJeunesse T.
917 2014. Genetics and morphology characterize the dinoflagellate *Symbiodinium voratum*, n. sp.,
918 (Dinophyceae) as the sole representative of *Symbiodinium* Clade E. The Journal of
919 Eukaryotic Microbiology. 61:75-94.

920 John U, Lu Y, Wohlrab S, Groth M, Janouškovec J, Kohli GS, Mark FC, Bickmeyer U,
921 Farhat S, Felder M, et al. 2019. An aerobic eukaryotic parasite with functional mitochondria
922 that likely lacks a mitochondrial genome. Science Advances. 5:eaav1110.

923 Johnson LK, Alexander H, Brown CT. 2019. Re-assembly, quality evaluation, and annotation
924 of 678 microbial eukaryotic reference transcriptomes. GigaScience. 8:giy158.

925 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
926 improvements in performance and usability. Molecular Biology and Evolution. 30:772-780.

927 Kent WJ. 2002. BLAT--the BLAST-like alignment tool. Genome Research. 12:656-664.

928 Kiessling W, Roniewicz E, Villier L, Leonide P, Struck U. 2009. An early Hettangian coral
929 reef in southern France implications for the end-Triassic reef crisis. Palaios. 24:657-671.

930 Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment
931 and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology. 37:907-915.

932 Kohli GS, John U, Figueroa RI, Rhodes LL, Harwood DT, Groth M, Bolch CJS, Murray SA.
933 2015. Polyketide synthesis genes associated with toxin production in two species of
934 *Gambierdiscus* (Dinophyceae). BMC Genomics. 16:410.

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

936 Kwan JC, Donia Mohamed S, Han Andrew W, Hirose E, Haygood Margo G, Schmidt Eric
937 W. 2012. Genome streamlining and chemical defense in a coral reef symbiosis. Proceedings
938 of the National Academy of Sciences of the United States of America. 109:20655-20660.

939 LaJeunesse TC, Lambert G, Andersen RA, Coffroth MA, Galbraith DW. 2005.
940 *Symbiodinium* (Pyrrhophyta) genome sizes (DNA content) are smallest among
941 dinoflagellates. Journal of Phycology. 41:880-886.

942 LaJeunesse TC, Parkinson JE, Gabrielson PW, Jeong HJ, Reimer JD, Voolstra CR, Santos
943 SR. 2018. Systematic revision of Symbiodiniaceae highlights the antiquity and diversity of
944 coral endosymbionts. Current Biology. 28:2570-2580.

945 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics.
946 34:3094-3100.

947 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
948 transform. Bioinformatics. 25:1754-1760.

949 Li T, Yu L, Song B, Song Y, Li L, Lin X, Lin S. 2020. Genome improvement and core gene
950 set refinement of *Fugacium kawagutii*. Microorganisms. 8:102.

951 Li W, Godzik A. 2006. CD-HIT: a fast program for clustering and comparing large sets of
952 protein or nucleotide sequences. Bioinformatics. 22:1658-1659.

953 Liew YJ, Li Y, Baumgarten S, Voolstra CR, Aranda M. 2017. Condition-specific RNA
954 editing in the coral symbiont *Symbiodinium microadriaticum*. PLOS Genetics. 13:e1006619.

955 Lo R, Dougan KE, Chen Y, Shah S, Bhattacharya D, Chan CX. 2022. Alignment-free
956 analysis of whole-genome sequences from Symbiodiniaceae reveals different phylogenetic
957 signals in distinct regions. Frontiers in Plant Science. 13:815714.

958 Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification
959 in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Research. 33:6494-
960 6506.

961 Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4:
962 A fast and versatile genome alignment system. PLOS Computational Biology. 14:e1005944.

963 Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of
964 occurrences of *k*-mers. Bioinformatics. 27:764-770.

965 McCoy MJ, Fire AZ. 2020. Intron and gene size expansion during nervous system evolution.
966 BMC Genomics. 21:360.

967 McCutcheon JP, Moran NA. 2012. Extreme genome reduction in symbiotic bacteria. Nature
968 Reviews Microbiology. 10:13-26.

969 Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. 2017. Uniclust
970 databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids
971 Research. 45:D170-D176.

972 Morrison HG, McArthur Andrew G, Gillin Frances D, Aley Stephen B, Adam Rodney D,
973 Olsen Gary J, Best Aaron A, Cande WZ, Chen F, Cipriano Michael J, et al. 2007. Genomic
974 minimalism in the early diverging intestinal parasite *Giardia lamblia*. Science. 317:1921-
975 1926.

976 Mrázek J, Xie S. 2006. Pattern locator: a new tool for finding local sequence patterns in
977 genomic DNA sequences. Bioinformatics. 22:3099-3100.

978 Nand A, Zhan Y, Salazar OR, Aranda M, Voolstra CR, Dekker J. 2021. Genetic and spatial
979 organization of the unusual chromosomes of the dinoflagellate *Symbiodinium*
980 *microadriaticum*. Nature Genetics. 53:618-629.

981 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective
982 stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and
983 Evolution. 32:268-274.

984 Nikbakht H, Xia X, Hickey DA. 2014. The evolution of genomic GC content undergoes a
985 rapid reversal within the genus *Plasmodium*. Genome. 57:507-511.

986 Opperman CH, Bird David M, Williamson Valerie M, Rokhsar Dan S, Burke M, Cohn J,
987 Cromer J, Diener S, Gajan J, Graham S, et al. 2008. Sequence and genetic map of
988 *Meloidogyne hapla*: a compact nematode genome for plant parasitism. Proceedings of the
989 National Academy of Sciences of the United States of America. 105:14802-14807.

990 Osanai-Futahashi M, Suetsugu Y, Mita K, Fujiwara H. 2008. Genome-wide screening and
991 characterization of transposable elements and their distribution analysis in the silkworm,
992 *Bombyx mori*. Insect Biochemistry and Molecular Biology. 38:1046-1057.

993 Pandolfi JM, Kiessling W. 2014. Gaining insights from past reefs to inform understanding of
994 coral reef response to global climate change. Current Opinion in Environmental
995 Sustainability. 7:52-58.

996 Pertea G, Pertea M. 2020. GFF Utilities: GffRead and GffCompare. F1000Research. 9

997 Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GAB. 2012. Evidence for
998 widespread GC-biased gene conversion in eukaryotes. Genome Biology and Evolution.
999 4:675-682.

1000 Piechotta M, Wyler E, Ohler U, Landthaler M, Dieterich C. 2017. JACUSA: site-specific
1001 identification of RNA editing events from replicate sequencing data. BMC Bioinformatics.
1002 18:7.

1003 Polne-Fuller M. 1991. A novel technique for preparation of axenic cultures of *Symbiodinium*
1004 (Pyrrophyta) through selective digestion by amoebae. Journal of Phycology. 27:552-554.

1005 Quattrini AM, Rodríguez E, Faircloth BC, Cowman PF, Brugler MR, Farfan GA, Hellberg
1006 ME, Kitahara MV, Morrison CL, Paz-García DA, et al. 2020. Palaeoclimate ocean conditions
1007 shaped the evolution of corals and their skeletons through deep time. Nature Ecology &
1008 Evolution. 4:1531-1538.

1009 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
1010 features. Bioinformatics. 26:841-842.

1011 Rai SK. 2022. Total High Molecular Weight DNA extraction from plant tissues for long read
1012 sequencing. Available from: 10.17504/protocols.io.b5qyq5xw

1013 Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for
1014 reference-free profiling of polyploid genomes. Nature Communications. 11:1432.

1015 Rayan NA, del Rosario RCH, Prabhakar S. 2016. Massive contribution of transposable
1016 elements to mammalian regulatory sequences. Seminars in Cell & Developmental Biology.
1017 57:51-56.

1018 Reinert G, Chew D, Sun F, Waterman MS. 2009. Alignment-free sequence comparison (I):
1019 statistics and power. Journal of Computational Biology. 16:1615-1634.

1020 Remmert M, Biegert A, Hauser A, Söding J. 2012. HHblits: lightning-fast iterative protein
1021 sequence searching by HMM-HMM alignment. Nature Methods. 9:173-175.

1022 Rizzo PJ. 1991. The enigma of the dinoflagellate chromosome. The Journal of Protozoology.
1023 38:246-252.

1024 Rizzo PJ. 2003. Those amazing dinoflagellate chromosomes. Cell Research. 13:215-217.

1025 Rizzo PJ, Jones M, Ray SM. 1982. Isolation and properties of isolated nuclei from the Florida
1026 red tide dinoflagellate *Gymnodinium breve* (Davis). The Journal of Protozoology. 29:217-
1027 222.

1028 Robbins SJ, Singleton CM, Chan CX, Messer LF, Geers AU, Ying H, Baker A, Bell SC,
1029 Morrow KM, Ragan MA, et al. 2019. A genomic view of the reef-building coral *Porites lutea*
1030 and its microbial symbionts. Nature Microbiology. 4:2090-2100.

1031 Saad OS, Lin X, Ng TY, Li L, Ang P, Lin S. 2020. Genome Size, rDNA Copy, and qPCR
1032 Assays for Symbiodiniaceae. Frontiers in Microbiology. 11:847.

1033 Sano J, Kato KH. 2009. Localization and copy number of the protein-coding genes actin, α-
1034 Tubulin, and HSP90 in the nucleus of a primitive dinoflagellate, *Oxyrrhis marina*. Zoological
1035 Science. 26:745-753.

1036 Schettino A, Turco E. 2009. Breakup of Pangaea and plate kinematics of the central Atlantic
1037 and Atlas regions. Geophysical Journal International. 178:1078-1097.

1038 Shen W, Le S, Li Y, Hu F. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for
1039 FASTA/Q File Manipulation. PLOS ONE. 11:e0163962.

1040 Shi T, Niu G, Kvitt H, Zheng X, Qin Q, Sun D, Ji Z, Tchernov D. 2021. Untangling ITS2
1041 genotypes of algal symbionts in zooxanthellate corals. Molecular Ecology Resources. 21:137-
1042 152.

1043 Shoguchi E, Beedessee G, Tada I, Hisata K, Kawashima T, Takeuchi T, Arakaki N, Fujie M,
1044 Koyanagi R, Roy MC, et al. 2018. Two divergent *Symbiodinium genomes* reveal conservation
1045 of a gene cluster for sunscreen biosynthesis and recently lost genes. BMC Genomics. 19:458.

1046 Shoguchi E, Shinzato C, Kawashima T, Gyoja F, Mungpakdee S, Koyanagi R, Takeuchi T,
1047 Hisata K, Tanaka M, Fujiwara M, et al. 2013. Draft assembly of the *Symbiodinium minutum*
1048 nuclear genome reveals dinoflagellate gene structure. Current Biology. 23:1399-1408.

1049 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO:
1050 assessing genome assembly and annotation completeness with single-copy orthologs.
1051 Bioinformatics. 31:3210-3212.

1052 Šmarda P, Bureš P, Horová L, Leitch IJ, Mucina L, Pacini E, Tichý L, Grulich V, Rotreklová
1053 O. 2014. Ecological and evolutionary significance of genomic GC content diversity in
1054 monocots. Proceedings of the National Academy of Sciences. 111:E4096-E4102.

1055 Smit AFA, Hubley R, Green P. 2017. RepeatMasker Open-4.0. Available from:
1056 http://www.repeatmasker.org

1057 Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab
1058 initio prediction of alternative transcripts. Nucleic Acids Research. 34:W435-W439.

1059 Stanley GD. 2003. The evolution of modern corals and their early history. Earth-Science
1060 Reviews. 60:195-225.

1061 Stanley GD, van de Schootbrugge B. 2018. The evolution of the coral–algal symbiosis and
1062 coral bleaching in the geologic past. In: van Oppen MJH, Lough JM, editors. Coral
1063 Bleaching: Patterns, Processes, Causes and Consequences. Cham: Springer International
1064 Publishing. p. 9-26.

1065 Stephen JG, Tripp HJ, Givan S, Podar M, Vergin Kevin L, Baptista D, Bibbs L, Eads J,
1066 Richardson Toby H, Noordewier M, et al. 2005. Genome streamlining in a cosmopolitan
1067 oceanic bacterium. Science. 309:1242-1245.

1068 Stephens TG, González-Pech RA, Cheng Y, Mohamed AR, Burt DW, Bhattacharya D,
1069 Ragan MA, Chan CX. 2020. Genomes of the dinoflagellate *Polarella glacialis* encode
1070 tandemly repeated single-exon genes with adaptive functions. BMC Biology. 18:56.

1071 Stephens TG, Ragan MA, Bhattacharya D, Chan CX. 2018. Core genes in diverse
1072 dinoflagellate lineages include a wealth of conserved dark genes with unknown functions.
1073 Scientific Reports. 8:17175.

1074 Steuber T, Löser H. 2000. Species richness and abundance patterns of Tethyan Cretaceous
1075 rudist bivalves (Mollusca: Hippuritacea) in the central-eastern Mediterranean and Middle
1076 East, analysed from a palaeontological database. Palaeogeography, Palaeoclimatology,
1077 Palaeoecology. 162:75-104.

1078 Thornhill DJ, Lewis AM, Wham DC, LaJeunesse TC. 2014. Host-specialist lineages
1079 dominate the adaptive radiation of reef coral endosymbionts. Evolution. 68:352-367.

1080 Turnham KE, Wham DC, Sampayo E, LaJeunesse TC. 2021. Mutualistic microalgae co-
1081 diversify with reef corals that acquire symbionts during egg development. The ISME Journal.

1082 van der Burgt A, Severing E, de Wit PJ, Collemare J. 2012. Birth of new spliceosomal
1083 introns in fungi by multiplication of introner-like elements. Current Biology. 22:1260-1265.

1084 Veron JEN. 2011. Scleractinia, evolution and taxonomy. In: Hopley D, editor. Encyclopedia
1085 of Modern Coral Reefs: Structure, Form and Process. Dordrecht: Springer Netherlands. p.
1086 947-957.

1087 Vinogradov AE. 2003. DNA helix: the importance of being GC-rich. Nucleic Acids
1088 Research. 31:1838-1844.

1089 Vulpius S, Kiessling W. 2017. New constraints on the last aragonite–calcite sea transition
1090 from early Jurassic ooids. Facies. 64:3.

1091 Wang B, Liang X, Gleason ML, Hsiang T, Zhang R, Sun G. 2020a. A chromosome-scale
1092 assembly of the smallest Dothideomycete genome reveals a unique genome compaction
1093 mechanism in filamentous fungi. BMC Genomics. 21:321.

1094 Wang W, Das A, Kainer D, Schalamun M, Morales-Suarez A, Schwessinger B, Lanfear R.
1095 2020b. The draft nuclear genome assembly of *Eucalyptus pauciflora*: a pipeline for
1096 comparing *de novo* assemblies. GigaScience. 9:giz160.

1097 Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of
1098 diploid genome sequences. Genome Research. 27:757-767.

1099 Worden AZ, Lee J-H, Mock T, Rouzé P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML,
1100 Derelle E, Everett MV, et al. 2009. Green evolution and dynamic adaptations revealed by
1101 genomes of the marine picoeukaryotes *Micromonas*. Science. 324:268-272.

1102 Xiang T, Hambleton EA, DeNofrio JC, Pringle JR, Grossman AR. 2013. Isolation of clonal
1103 axenic strains of the symbiotic dinoflagellate *Symbiodinium* and their growth and host
1104 specificity. Journal of Phycology. 49:447-458.

1105 Xue W, Li J-T, Zhu Y-P, Hou G-Y, Kong X-F, Kuang Y-Y, Sun X-W. 2013.
1106 L_RNA_scaffolder: scaffolding genomes with transcripts. BMC Genomics. 14:604.

1107 Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA
1108 genome assembler. Bioinformatics. 29:2669-2677.

1109 Zimin AV, Puiu D, Luo M-C, Zhu T, Koren S, Marçais G, Yorke JA, Dvořák J, Salzberg SL.
1110 2017. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a
1111 progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Research.
1112 27:787-792.

1113

1114

1115 # Table

1116 **Table 1**

1117 Genome assemblies and gene predictions of *E. voratum* RCC1521, rt-383, and CCMP421.

Isolate	RCC1521	rt-383 (=CCMP3420)	CCMP421
Location of isolation	Mediterranean Sea off Blanes, Spain	Eastern North Pacific off Santa Barbara, USA	Cooks Strait, New Zealand
Reference	(Jeong et al. 2014)	(Polne-Fuller 1991)	(Chang 1983)
Genome sequencing technologies	Illumina, PacBio, Nanopore	Illumina, PacBio, Nanopore	10X Linked-reads
Genome assembly size (Gb)	1.2	1.3	1.1
Estimated genome size (Gb)	1.4	1.2	1.9
GC-content of genome assembly (%)	50.8	50.6	50.9
Total read coverage	446×	212×	153×
Number of genome scaffolds	3,881	11,607	38,022
N50 of genome assembly (Kb)	720	252	304
Number of predicted genes	32,108	39,878	32,615
% BUSCO recovery (genome)	34.0	35.1	20.4
% BUSCO recovery (proteins)	76.7	77.2	67.2

1118

Figures

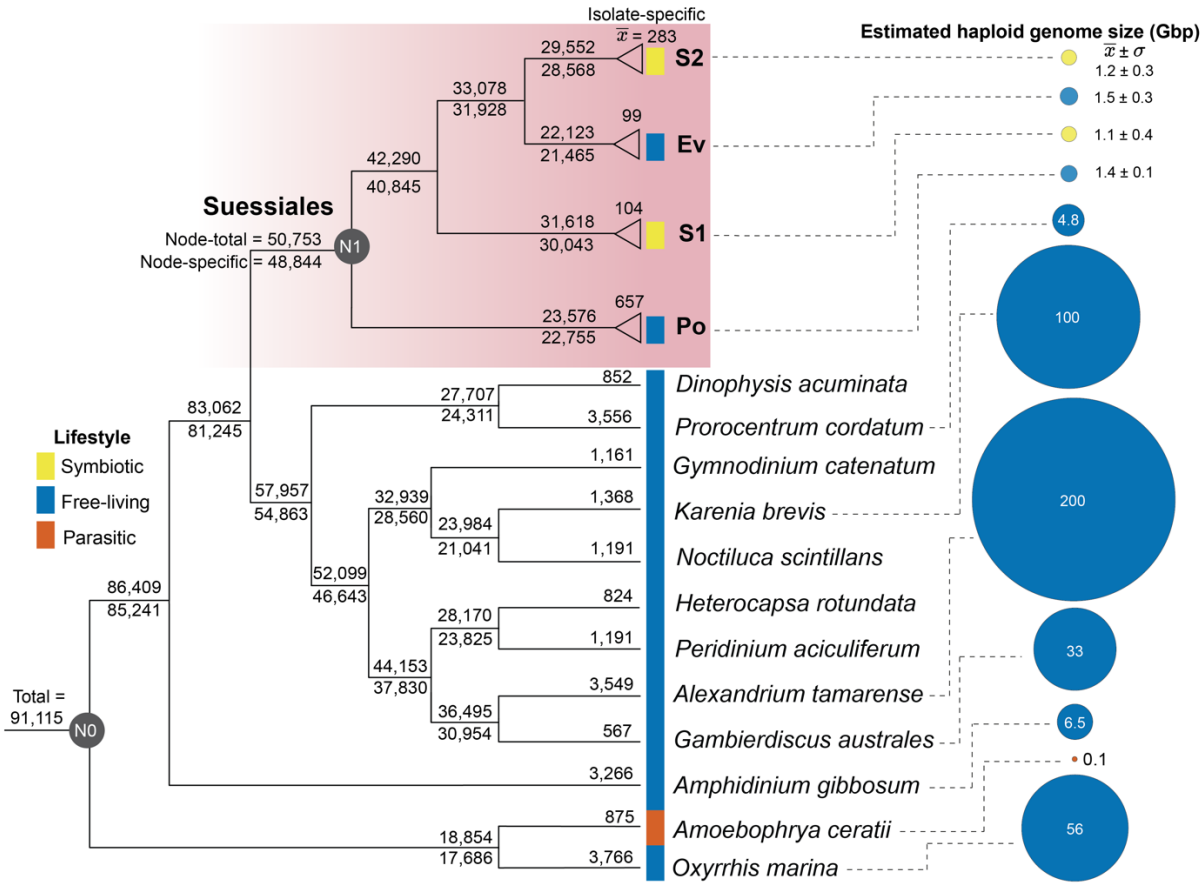
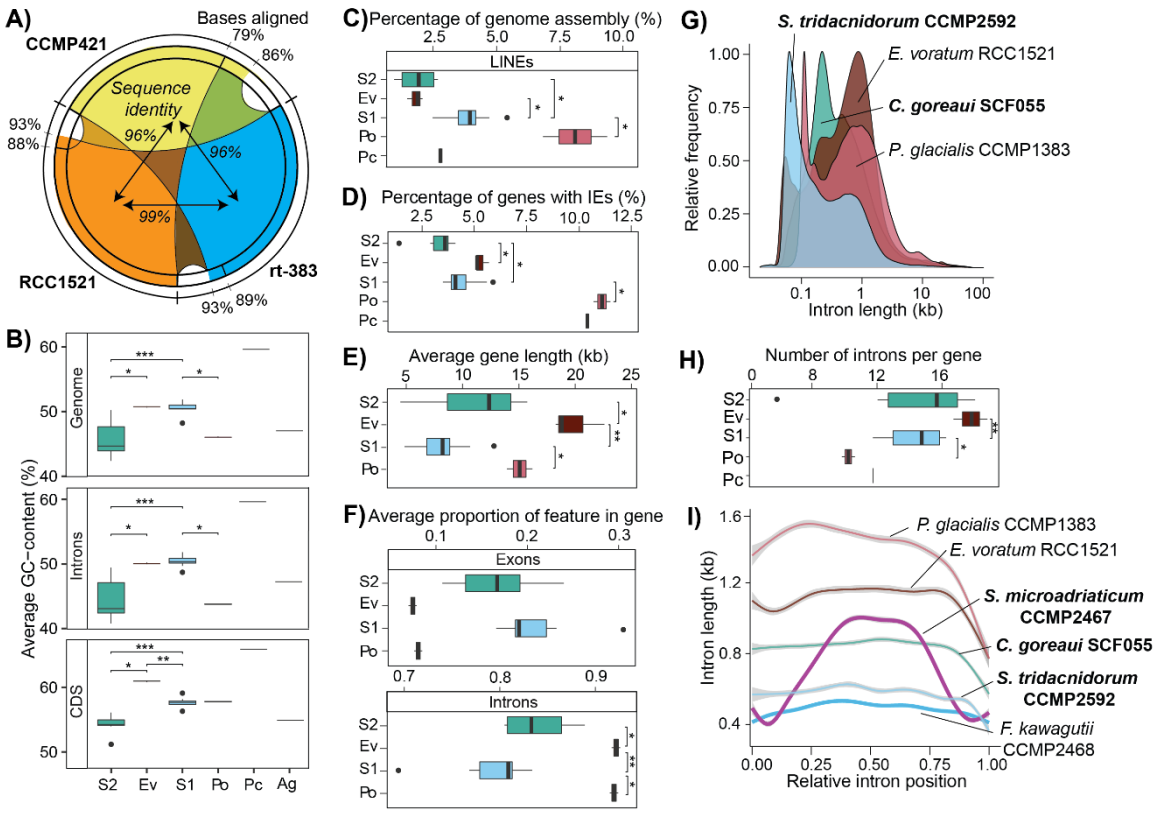


Figure 1

Species tree of dinoflagellates alongside estimated genome sizes. The species tree was inferred using 91,115 orthologous protein sets derived from 1,420,328 protein sequences from 33 dinoflagellate taxa. Symbiotic lifestyles are shown in yellow, free-living in blue, and parasitic in orange. Node-total refers to the number of protein families which contain one or more taxa at the node. Node-specific are protein families that exclude taxa outside the node. Numbers at the tips of each branch represent isolate-specific protein families. For the S1, S2, Ev, and Po groups of Suessiales, the mean of isolate-specific protein families and the mean ± standard deviation of estimated genome sizes are displayed. Node N0 contains homologous protein sets from all 33 taxa, whereas node N1 includes data from the 21 Suessiales.

1132



1133
1134 **Figure 2**

1135 **Genome features of *E. voratum* and other dinoflagellates.** (A) Genome-sequence identity
1136 (*italics* within circle) and the percentage of aligned bases (outside circle) among the three *E.*
1137 *voratum* isolates. Features of representative genomes in the four Suessiales groups of S2, Ev,
1138 S1, and Po, plus *Prorocentrum cordatum* (Pc), in the order from the most-recent to most-
1139 ancient divergence, shown for (B) mean GC content in whole-genome, intronic, and CDS
1140 regions (*Amphidinium gibbosum* (Ag) was added for dinoflagellate-wide comparison), (C)
1141 percentage of mobile elements, and (D) percentage of genes containing introner elements
1142 (IEs). Gene features of Suessiales showing (E) mean gene length, (F) average proportions of
1143 exons and introns, (G) the relative frequency of introns by length (symbiotic lineages in
1144 boldface), (H) number of introns per gene, and (I) intron length versus relative intron position
1145 (symbiotic lineages in boldface). In all bar charts, *, **, *** represent $p < 0.05$, < 0.01, and $<$
1146 0.001 respectively based on Wilcoxon rank sum test.

1147

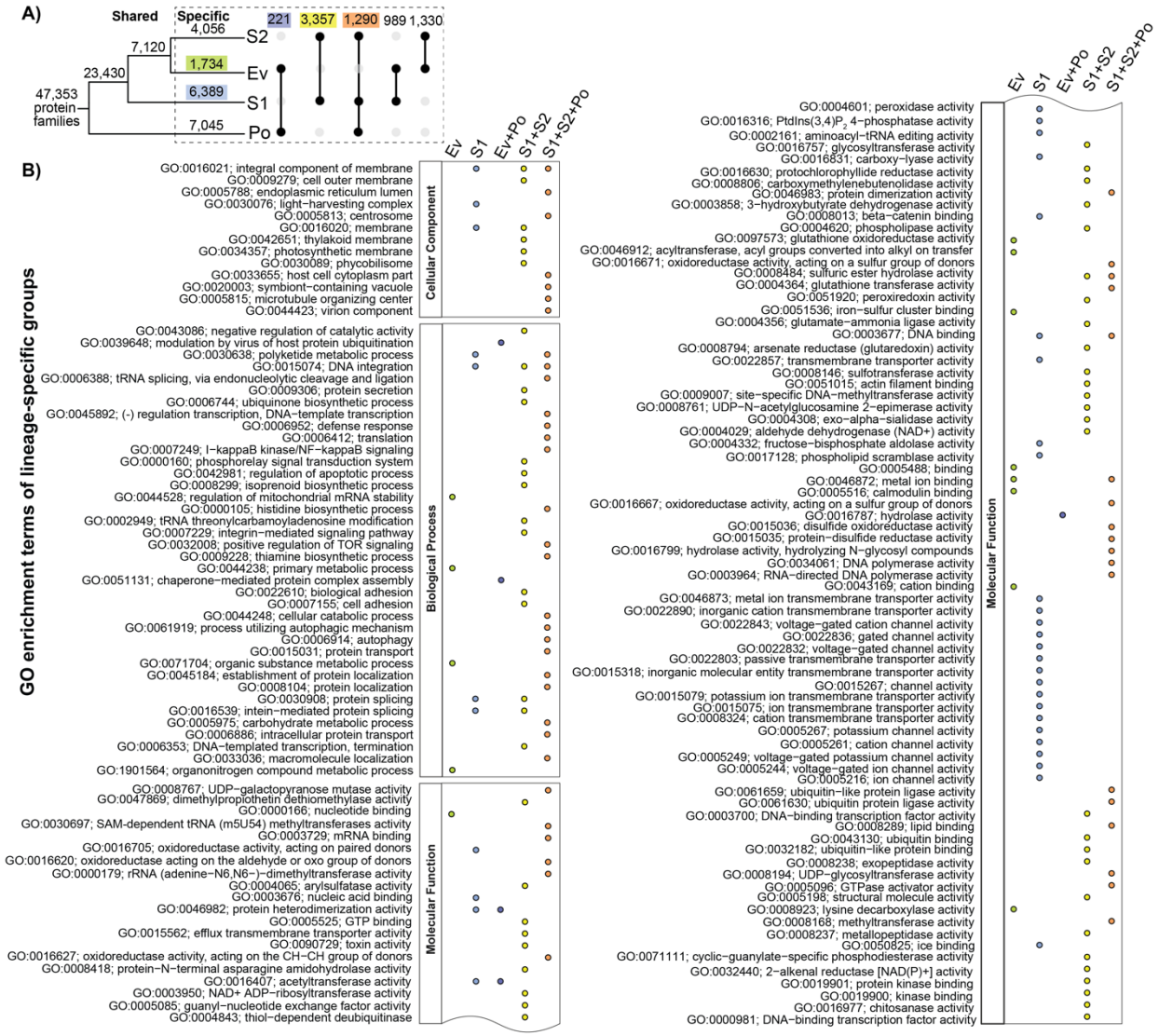


Figure 3

Gene family evolution in Suessiales. (A) Number of protein families is shown at each node and branch represents those that are shared among or specific to S1, Ev, S2, and/or Po. Number of families that are exclusive to Ev (green), to S1 (light blue), to Ev+Po (dark blue), to S2+S1 (yellow), and to S2+S1+Po (orange) were highlighted. (B) Enriched Gene Ontology (GO) terms for genes in the five distinct groups relative to all GO terms in the corresponding taxa, arranged in decreasing order of significance from top to bottom within the categories: Cellular Component, Biological Process, and Molecular Function.

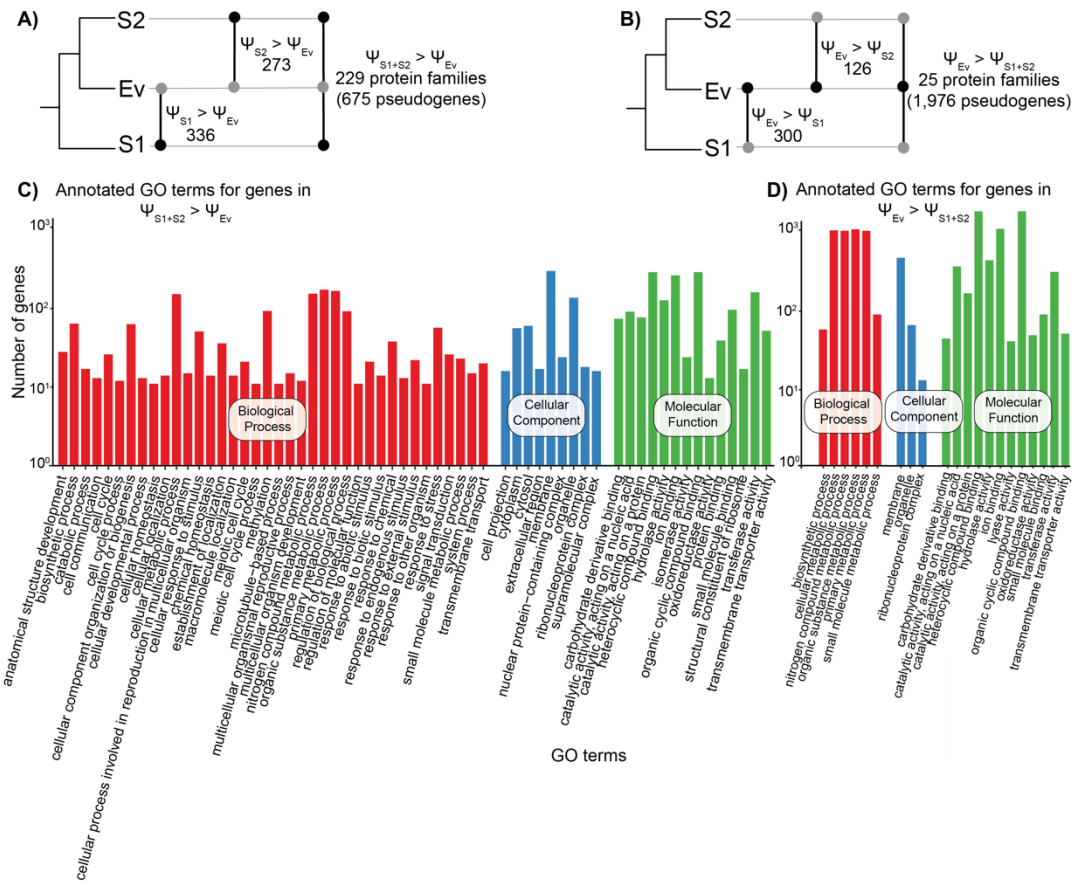


Figure 4

Pseudogenisation in Symbiodiniaceae. Number of protein families with significantly different level of pseudogenisation, Ψ, shown for (A) those with greater extent in symbiotic lineages (i.e., higher Ψ in S1, S2, or S1+S2, relative to Ev), and (B) those with greater extent in the free-living Ev (i.e., higher Ψ in Ev relative to S1, S2, or S1+S2). The black circles on the upset plots indicate taxa groups with higher Ψ than those with grey circles. The associated GO terms are shown for (C) those where $\Psi_{S1+S2} > \Psi_{Ev}$, and (D) those where $\Psi_{Ev} > \Psi_{S1+S2}$.

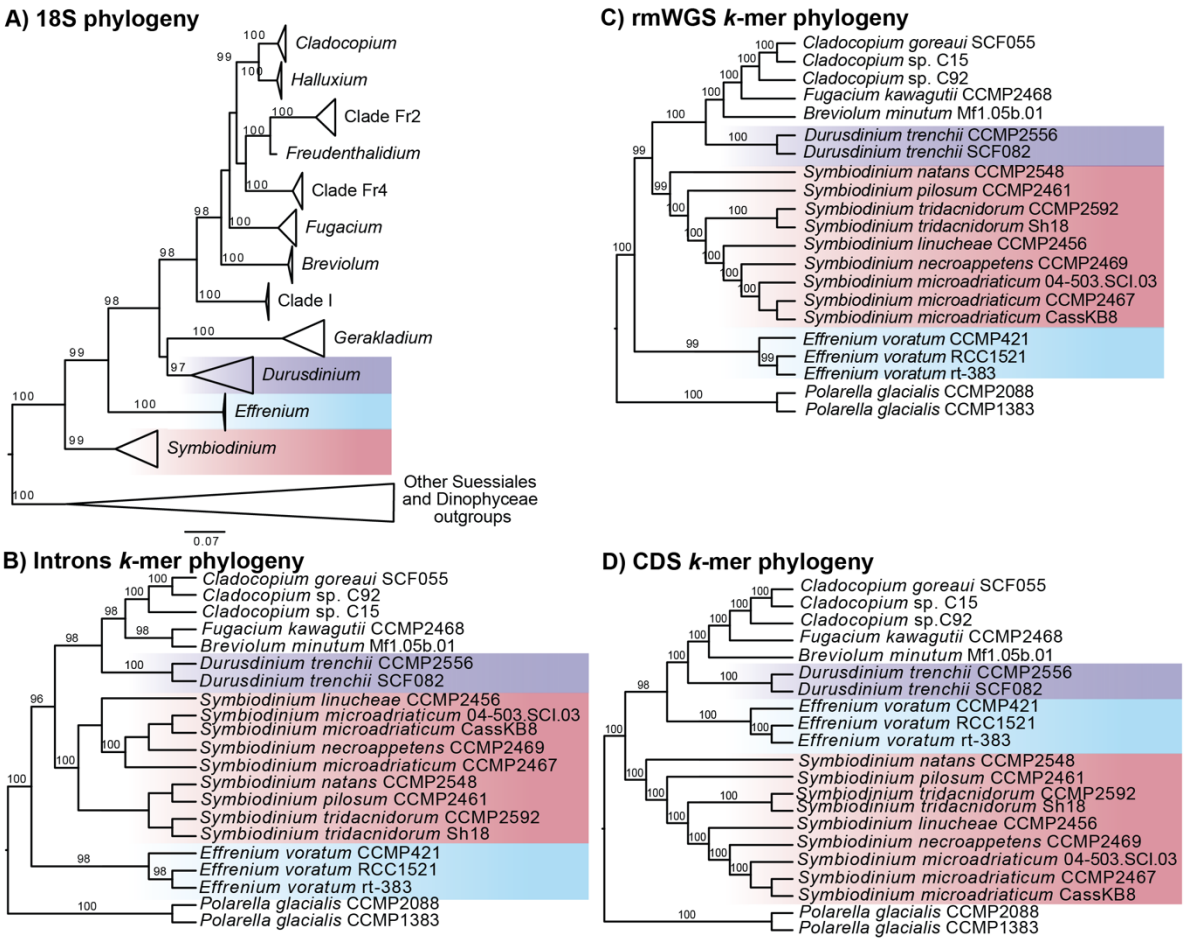


Figure 5

Phylogenetic relationship of *E. voratum* within Symbiodiniaceae and other taxa. (A) Maximum likelihood species tree of dinoflagellates inferred based on alignment of 18S rDNA sequences from 13 representative taxa, showing node support derived from ultrafast bootstrap of 2,000 samples; unit of branch length is number of substitutions per site. Tree topologies inferred from *k*-mers in 21 Suessiales genomes using our alignment-free (AF) approach are shown for (B) intronic regions ($k = 21$), (C) repeat-masked whole-genome sequences ($k = 23$), and (D) coding sequence regions ($k = 19$). Jackknife support of ≥ 96% based on 100 pseudo-replicates is shown on each node of the AF trees.

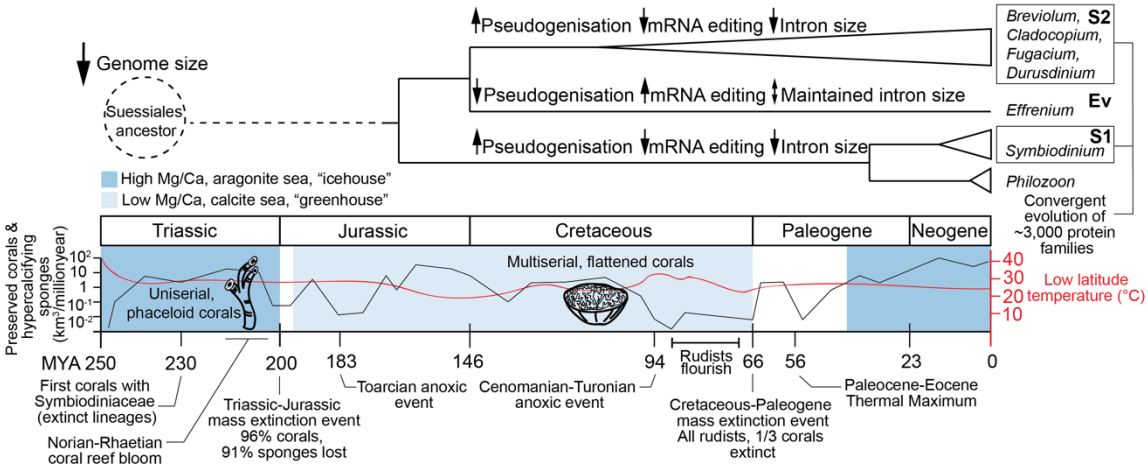


Figure 6

Timeline of Symbiodiniaceae genome evolution and coral evolution. The estimated divergence timeline of the Family Symbiodiniaceae (LaJeunesse et al. 2018) is shown at the top, indicating representative taxa of S1, Ev, and S2. Key genome signatures for each group related to pseudogenisation, mRNA editing, and intron size are shown along the branch. The dotted line represents the yet-unknown timeline of Suessiales divergence from the rest of dinoflagellates. Evolutionary timescale along the different eras highlighting key geological events relevant to coral evolution, aligning with Symbiodiniaceae divergence, is shown at the bottom. Mass of preserved corals and sponges represented by black line (left y-axis), low latitude (30°S-30°N) ocean temperature in red (right y-axis), ocean chemistry of Mg/Ca ratio showing aragonite (dark blue) vs calcite sea (light blue); data were sourced from earlier studies (Veron 2011; Pandolfi and Kiessling 2014; LaJeunesse et al. 2018; Stanley and van de Schootbrugge 2018; Quattrini et al. 2020; Frankowiak et al. 2021; Grossman and Joachimski 2022).