Research Article

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) are 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, which is often associated with a symbiotic lifestyle, predates the origin of Symbiodiniaceae. 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 was retained (or reduced) among all extant algae in this lineage. Nonetheless, our data reveal that the free-living lifestyle distinguishes *Effrenium* from symbiotic Symbiodiniaceae vis-à-vis their 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.

41 **Introduction**

42 Dinoflagellate algae in the Family Symbiodiniaceae comprise taxa that form symbioses with
43 diverse marine organisms. Of particular importance to modern coral reefs, Symbiodiniaceae
44 provide photosynthates *via* fixed carbon and essential nutrients to corals while resident in
45 these cnidarians. The Symbiodiniaceae ancestor is believed to have been free-living (*1*) with
46 members of this group forming symbiotic associations with corals as early as 230 million
47 years ago (MYA) (*2*). Symbiogenesis, or the establishment of a symbiotic relationship
48 between two or more taxa (*3*), can drastically influence lineage evolution, adaptation, and
49 speciation as observed in obligate parasites and diverse symbiotic taxa (*4-8*). This
50 phenomenon is termed the resident genome syndrome and was previously hypothesised to
51 explain the observed patterns of Symbiodiniaceae genome evolution (*9*).

52 Based on current divergence time estimates for Symbiodiniaceae, the split between
53 the basal genera (*Symbiodinium* and *Philozoon*) and the rest of the family occurred 166
54 MYA, whereas the more-recently branching symbiotic lineages diversified ~109 MYA (*1*). If
55 the emergence of symbiogenesis coincides with the earliest fossil evidence from 230 MYA
56 (*2*), then different Symbiodiniaceae lineages would have arose and diversified during major
57 global geological events. These events include the switch from aragonite to calcite seas (~190
58 MYA; (*10*), the breakup of Pangea (150-230 MYA; (*11*), the diversification or extinction of
59 potential hosts, e.g., the extinction of rudists 66 MYA (*12, 13*), and the overall change in
60 coral morphology from the Triassic (201-252 MYA) to the Cretaceous (66-145 MYA) (*14,*
61 *15*). More-recent examples include the rapid radiation of the genus *Cladocopium* 4-6 MYA
62 due to geographic isolation (*16*), and the co-diversification of *Symbiodinium fitti* and their
63 coral hosts (*17*); see (*18*) for latest systematic revision of *Cladocopium* species.

64 As described above, symbiogenesis is expected to impact the genome evolution of
65 symbionts to varying extents within a broad spectrum of "facultativeness" that reflects the
66 nature of the host association (i.e., with obligate free-living and obligate symbiont at
67 opposing extremes) and underpins evolutionary processes such as genome streamlining,
68 genetic drift, expansion/contraction of mobile elements, pseudogenisation, gene loss, and
69 varying mutation rates (*9*). Previous studies investigating the effects of symbiogenesis on
70 Symbiodiniaceae genomes have focused almost entirely on symbiotic genera, with the polar-

71 dwelling, highly specialised *Polarella glacialis* (*19*), a sister of Symbiodiniaceae and within
72 the same Order Suessiales, providing the only free-living outgroup.

73 Haploid genome sizes of Symbiodiniaceae taxa and *P. glacialis* are estimated to be <
74 2 Gbp, based on available sequencing data (*19-22*), and < 5Gbp based on DNA staining and
75 qPCR analysis of marker sequences (*23, 24*). Estimates based on DNA staining are generally
76 larger than those based on sequencing data, likely due to the permanently condensed
77 chromosomal structures of dinoflagellates that result in overestimation of DNA content (*25,*
78 *26*). The diverse dinoflagellate taxa external to the Symbiodiniaceae are predominantly free-
79 living and, in comparison, have massive genome sizes, e.g., 4.8 Gbp estimated from
80 sequencing data for the bloom-forming *Prorocentrum cordatum* (*27*), and 200 Gbp based on
81 DNA staining for *Alexandrium tamarense* (*28*).

82 Among Symbiodiniaceae taxa, *Effrenium* is the early-diverging, exclusively free-
83 living genus. The sole species, *E. voratum*, is globally distributed in temperate and
84 subtropical waters (*1, 29*). Attempts to establish a symbiotic relationship between *E. voratum*
85 and the anemone *Exaiptasia pallida* have been unsuccessful (*30, 31*). Current understanding
86 of Symbiodiniaceae evolutionary history suggests that *E. voratum* diverged 147 MYA from
87 the basal, largely symbiotic genera of *Symbiodinium* and *Philozoon*, and prior to the other
88 later-diverging symbiotic genera (*1*). Whereas genomes of other free-living species such as
89 *Symbiodinium natans* (*32*) have been generated, these taxa belong to genera that also include
90 symbiotic species and thus might have experienced a symbiotic lifestyle at some point in their
91 history. We expect the genus *Effrenium* to have remained unaffected by the influence of
92 symbiogenesis, and thus retain the ancestral free-living lifestyle (and genome features) of
93 Symbiodiniaceae.

94 In this study, we present *de novo* assembled genome and transcriptome data for three
95 isolates of *E. voratum*. Incorporating publicly available genome-scale data from 16
96 Symbiodiniaceae taxa plus four free-living taxa external to the Symbiodiniaceae in a
97 comparative genomic analysis, we examine genomic features in *E. voratum*. These include
98 mobile elements, gene structures, gene-families, and pseudogenisation to gain insights into
99 ancestral features of Symbiodiniaceae, and more broadly, Suessiales genome evolution.

100 ## Results

101 **Genome-size reduction pre-dated divergence of Order Suessiales and Family**
102 **Symbiodiniaceae**

103 Using a mix of short- and long-read sequencing data (Table S1), we generated *de novo*
104 genome assemblies for three isolates of *E. voratum* (assembly sizes 1.1-1.3 Gbp), with
105 estimated haploid genome sizes of 1.2-1.9 Gbp (Table S2 and Fig. S1), completeness
106 (BUSCO recovery 67.2-77.2%), and number of predicted genes (32,102-39,878) (Table 1,
107 Tables S3 and S4) comparable to the other Symbiodiniaceae genomes (*19, 21, 22, 33-36*). We
108 obtained all available genomic data from 23 dinoflagellate taxa: 19 from Symbiodiniaceae
109 (Order Suessiales), two sister taxa of *P. glacialis* (Order Suessiales), and two distantly related
110 free-living dinoflagellate taxa, *Prorocentrum cordatum* (Order Prorocentrales) and
111 *Amphidinium gibbosum* (Order Amphidiniales). The 21 Suessiales taxa were grouped into:
112 (a) the earlier-branching, largely symbiotic genus *Symbiodinium* (S1), (b) the three
113 exclusively free-living *E. voratum* isolates (Ev), (c) the later-branching symbiotic
114 Symbiodiniaceae lineages (S2), and (d) the free-living outgroup *P. glacialis* (Po) sister to
115 Family Symbiodiniaceae (Table S3). The phylogenetic positions of these groups relative to
116 other dinoflagellates are shown in Fig. 1A, along with light micrographs of representative
117 species in S1, S2, and Ev (Fig. 1B). Cell size of *E. voratum* (12.2-13.3 µm (*29*)) is generally
118 larger than S1 (e.g., *Symbiodinium microadriaticum* CassKB8; 8.0-11.0 µm (*23, 37*)) or S2
119 cells (e.g., *D. trenchii* CCMP2556; 7.5-10.0 µm (*23*)).

120 The most striking feature of genome evolution among Suessiales is the marked
121 reduction in genome size that occurred in the common ancestor of this lineage. Whereas free-
122 living dinoflagellates external to the Suessiales have genomes that range widely in size from
123 ca. 5-200 Gbp, except the parasitic *Amoebophrya ceratii* that has a highly reduced genome of
124 size 0.1 Gbp, all Suessiales genomes have a much narrower size range from 0.7-2.0 Gbp,
125 estimated using sequencing data (Fig. 1A). This is accompanied by a ~40% loss in gene
126 families prior to the diversification of Suessiales, when compared to their closest
127 dinoflagellate relatives (node N1 in Fig. 1A). This pattern is reminiscent of the red algae
128 (Rhodophyta), whose common ancestor underwent massive genome reduction, precipitating
129 the loss of canonical eukaryotic features such as flagella-based motility, phytochromes, and
130 autophagy (*38*). These algae split into two monophyletic lineages, the extremophilic

131 Cyanidiophytina that specialised to life in hot spring environments, and the species-rich
132 mesophilic lineages (e.g., red seaweeds) that inhabit a variety of aquatic environments (*38*).
133 Most red algae have therefore smaller genomes when compared to the green lineage and have
134 adapted to diverse habitats through gene family evolution and horizontal gene transfer. In an
135 analogous fashion in dinoflagellates, it appears that the Suessiales common ancestor
136 underwent significant genome reduction, likely due to life in extreme habitats (e.g., the
137 psychrophilic *Polarella glacialis*). This streamlining of the gene inventory may have played
138 an important role in driving symbiotic associations with cnidarians that offered nutrient-rich
139 and protected habitats within the animal tissues. The facultative lifestyle was likely retained
140 in most Symbiodiniaceae because it offers the benefit of sexual reproduction during the free-
141 living stage. The most substantial recovery from genome streamlining is offered by whole
142 genome duplication, which has occurred in the *Durusdinium* lineage (*33*).

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

144 Genome sequences of the three *E. voratum* isolates share high similarity with 96% sequence
145 identity over 93% of bases (Fig. 2A). Repetitive regions containing protein-coding genes
146 were highly conserved relative those of other Suessiales (the Order containing
147 Symbiodiniaceae and the earlier branching sister *P. glacialis*). We identified 98,344 core *k*-
148 mers ($k = 23$; all possible 23-base sequences) that are common in genomes of all Suessiales
149 taxa following Lo et al. (*39*), and recovered 95% of core 23-mers in repetitive regions of *E.*
150 *voratum* (Fig. S2; see Materials and Methods). Among the three *E. voratum* genome datasets,
151 we recovered 5-32 putative mitochondrial genome sequences in each dataset that encode the
152 marker genes of *cob*, *cox1* and/or *cox3* (Table S5), and 6-32 putative plastid genome
153 sequences in each dataset that encode the 16S/23S rRNA or one of the 11 plastid-encoded
154 genes (*40*), including a putative empty minicircle sequence (Table S6).

155 Given the history of genome reduction, we investigated the traits that may
156 differentiate *Effrenium* from the symbiotic lineages of Symbiodiniaceae. We studied genome
157 size, intron evolution, gene family evolution, pseudogenisation, RNA editing, and
158 phylogenetic signals in Ev, S1 and S2, relative to the outgroup Po. The GC content of coding
159 regions varied among the Symbiodiniaceae lineages (Fig. 2B) and were significantly lower in
160 S2 (mean 54.2%; $p < 0.05$) relative to Ev (61.0%) and S1 (57.7%); in intronic regions, the
161 mean GC is 44.6% (S2), 50.1% (Ev) and 50.4% (S1), whereas that of whole-genome

162 sequences is 45.8% (S2), 50.7% (Ev), and 50.6% (S1). GC-rich genome sequences confer
163 bendability to DNA helices (*41*) and may prevent cell freezing or desiccation (*42*). Variation
164 of GC content in dinoflagellate genomes does not appear to correlate to lifestyle; among the
165 free-living species external to Symbiodiniaceae, genomes of *P. glacialis* and *A. gibbosum* has
166 a mean GC content of 46.4%, similar to S2, whereas the genome of *Pr. cordatum* has the
167 highest GC content described thus far for any dinoflagellate, at 59.7% (*27*). Intracellular
168 bacteria have a mutational bias towards low genomic GC content, e.g. ~20% (*43*), but
169 intracellular eukaryotes display both low and high extreme GC content patterns, ranging as
170 low as 24% in the malaria parasite *Plasmodium falciparum* (*44*) to 67% in the green algal
171 symbiont *Chlorella variabilis* in the ciliate *Paramecium* (*45*). The lower GC content in S2
172 genomes than the S1 counterparts underscores the dynamic nature of genomic GC content
173 evolution in intracellular eukaryotes.

174 Mobile elements, particularly transposable elements (TEs), can influence genomic
175 architecture and base composition, and have been used to reconstruct the evolutionary history
176 of many species (*46-48*). Although facultative Symbiodiniaceae symbionts (i.e., S1 and S2)
177 are expected to contain a larger proportion of mobile elements in their genomes when
178 compared to free-living (Ev) lineages (*9*), no significant difference among the groups was
179 observed in the overall abundance (Table S7 and Fig. S3A) or conservation of TEs (Fig.
180 S3B). Although more contiguous genome assemblies generally contain more conserved
181 repeats (Kimura substitution values centred around 3 for *D. trenchii* CCMP2556, *C.*
182 *proliferum* SCF055 [formerly described as *C. goreaui* SCF055 (*18*)], the three *E. voratum*
183 isolates, *S. natans* CCMP2548, and *S. tridacnidorum* CCMP2592) compared to the others
184 (i.e., values centred around 20), we note diverged repeats (Kimura values centred around 25)
185 in the chromosome-scale assembly of *S. microadriaticum* CCMP2467 (Fig. S3B). This result
186 suggests a potential technical bias in the recovery of mobile elements. Despite this issue,
187 based on the proportions for distinct types of mobile elements in each group, we found
188 significantly ($p < 0.05$) more long interspersed nuclear elements (LINEs) in S1 (3.9%)
189 compared to Ev (1.8%) and to S2 (1.9%), with the proportion of the outgroup Po at 8% (Fig.
190 2C). This result suggests that S1 likely retains ancestral LINEs, lending support to the notion
191 of loss of LINEs after the diversification of basal Symbiodiniaceae genera (*21*).

192 Introner elements (IEs) are a type of mobile element consisting of inverted and direct
193 repeat motifs found at 5′- and 3′-end of introns in diverse eukaryotes (*49-52*). Recent research

194 on dinoflagellates revealed that IEs are more abundant in free-living (within 10-12% of
195 genes) than in symbiotic/parasitic species (0.8-6.0% of genes) (*27, 53*). Although not as high
196 as in other free-living dinoflagellates, the Ev genomes exhibit more IE-containing genes (5%)
197 than do the S1 (4%) and S2 genomes (3%) (Fig. 2D). IEs have been postulated to be non-
198 autonomous and their mobility is dependent on transposases encoded in dinoflagellate
199 genomes (*53*). We recovered transposase protein sequences from most of the
200 Symbiodiniaceae genomes in this study (Table S8), suggesting a capacity for IEs to be
201 mobile.

202 Significantly ($p < 0.05$) longer genes were observed in Ev (mean 20 Kb) than S1 (8
203 Kb) and S2 (11 Kb) (Fig. 2E), primarily driven by longer intron sizes (introns make up, on
204 average, 92% of a gene [Fig. 2F]; sizes peak at 1 Kb [Fig. 2G]) and higher intron density per
205 gene (mean of 18 for Ev, 14 for S1, 14 for S2) (Fig. 2H). The driving mechanism for this
206 trend may reflect one of two evolutionary scenarios: (a) intron expansion in Ev, or (b) intron
207 contraction in S1/S2. In examples of the first scenario, TE-mediated insertions drive intron
208 expansion and are biased toward the 5′ end of genes to prevent disruption of functional
209 elements (*54*), yielding larger intron sizes at 5′ ends. We did not observe this trend in Ev (Fig.
210 2I). In the second scenario, which has been observed in endosymbiotic/parasitic organisms,
211 the reduction of intron size and density occurs as a result of genome reduction and/or
212 streamlining induced by spatial confinement in the host organism or cell (*46, 55*).
213 Considering the evolutionary history of Symbiodiniaceae, intron contraction in S1/S2 taxa
214 due to their symbiotic lifestyle is more plausible than intron expansion in Ev. Large introns
215 observed in Po (mean 1.4 Kb) (Fig. 2G), albeit at a lower intron density per gene (Fig. 2H),
216 lend further support to this notion.

217 **Symbiogenesis shaped evolution of gene families and post-transcriptional processing in**
218 **Symbiodiniaceae**

219 To examine the effect of symbiotic lifestyle on protein family evolution in Symbiodiniaceae,
220 we first inferred 53,173 homologous protein families from all 811,611 protein sequences
221 predicted from the 21 Suessiales genomes (see Materials and Methods). Most protein families
222 (47,353 of 53,173 [89%]) were shared among the four groups (S1+S2+Ev+Po). With respect
223 to functions annotated in all families, these families were enriched in functions such as
224 cellular motility, biosynthetic processes for rRNA, antibiotics, and glycosides (Table S9).

225 There were more lineage-specific protein families in S1 (6,389) and S2 (4,056) than in Ev
226 (1,734), and the two symbiotic groups (S1+S2) shared 3,357 protein families not found in the
227 other groups (Fig. 3). The high number of protein families present only in symbiotic lineages
228 that split from each other over 40 million years of evolution suggests convergent evolution
229 due to the symbiotic lifestyle. These protein families were enriched for diverse functions
230 including signalling, apoptosis, protein splicing, photosynthesis, cell adhesion, and various
231 transferase activities (Fig. 3). Incidentally, Po shared more protein families with symbiotic
232 lineages (1,290; S1+S2+Po) than with Ev (221; Ev+Po); these families were enriched in
233 functions such as autophagy and microtubule organisation.

234 Facultative symbionts in Symbiodiniaceae are expected to display higher levels of
235 pseudogenisation, a major feature of the resident genome syndrome (*9*). To investigate this
236 issue, we identified putative pseudogenes in the Symbiodiniaceae genomes (Table S10)
237 following González-Pech et al. (*21*). We defined the level of pseudogenisation, Ψ, as a ratio
238 of the number of putative pseudogenes to the number of putative functional genes per
239 homologous family; see Materials and Methods. We compared Ψ independently for Ev (Ψ_{Ev})
240 against that for S1 (Ψ_{S1}), S2 (Ψ_{S2}), and the combined S1 and S2 (Ψ_{S1+S2}), then identified
241 protein families that exhibited significant difference ($p < 0.05$) of this value. More protein
242 families display $\Psi_{S1} > \Psi_{Ev}$ (336) and $\Psi_{S2} > \Psi_{Ev}$ (273; Fig. 4A), compared to $\Psi_{S1} < \Psi_{Ev}$ (300)
243 and $\Psi_{S2} < \Psi_{Ev}$ (126; Fig. 4B). There was nine-fold more protein families exhibiting $\Psi_{S1+S2} >$
244 Ψ_{Ev} (229; Fig. 4A) than vice versa (25; Fig. 4B). These pseudogenes are associated with a
245 wide range of functions, including cell cycle processes and stimuli response (Figures 4C and
246 4D). The protein families that display significantly higher Ψ in the symbiotic lineages are
247 mostly mutually exclusive from the 3,357 families that putatively experienced convergent
248 evolution (only 16–23 families are represented in Ψ_{S1}, Ψ_{S2}, Ψ_{S1+S2}). We found negligible
249 technical biases in the clustering of homologous sequences that may affect our inference of
250 pseudogenes, i.e., protein families displaying Ψ remained stable at different clustering
251 parameters (Fig. S4; see Materials and Methods). These results suggest that in addition to
252 convergent evolution in the symbiotic lineages, these lineages of Symbiodiniaceae have
253 experienced a greater extent of pseudogenisation than has the free-living Ev.

254 Editing of mRNAs allows Symbiodiniaceae to increase the variability in protein
255 isoforms (*56*). To assess mRNA editing in *E. voratum*, we focused on RCC1521, which has

256 the most contiguous genome assembly, and compared the rates and types of mRNA editing
257 against a representative from S1 (*S. microadriaticum* CCMP2467 (*56*)) and S2 (*D. trenchii*
258 CCMP2556 (*33*)). We identified 45,009 unique mRNA edited sites in *E. voratum*, about 13-
259 fold and 4-fold greater than in *S. microadriaticum* and *D. trenchii*, respectively (Table S11).
260 A larger proportion of genes in *E. voratum* (9,158, 28.5%) contain mRNA edits, compared to
261 *S. microadriaticum* (774, 1.6%) and *D. trenchii* (4,227, 7.6%), although the distribution of
262 substitution types is similar among the three species (Fig. S5A-D). Most edits in *D. trenchii*
263 and *E. voratum* (> 60%) are located in exons (Fig. S5E), whereas in *S. microadriaticum* edits
264 are evenly split between exons and introns, although this may be due to the more-fragmented
265 genome assembly of Liew et al. (*56*). As also observed in *S. microadriaticum* (*56*), there is a
266 slight bias for edits to occur near 5′ ends of genes (within the first ~10% of gene length, $p <$
267 0.05; Fig. S5F) and the edits tend to be located within 1 Kb of each other (versus the
268 distribution expected at random, $p < 0.05$; Fig. S5G). The high level of mRNA editing in Ev
269 is consistent with data from another free-living dinoflagellate, *Pr. cordatum* (42,969 edited
270 sites; 32,067 within 10,169 [12%] genes) (*27*). Together with our observation of fewer Ev-
271 specific protein families than those specific to S1 or S2 (Fig. 2), these results suggest a more
272 pronounced role of mRNA editing in generating functional diversity in free-living versus
273 symbiotic dinoflagellate taxa. The relationship and impact of this RNA editing on the
274 encoded proteins remain to be investigated using proteomics.

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

276 Our alignment-based phylogenetic trees inferred using multiple protein families (Fig. 1A),
277 the standard molecular marker of 18S rDNA (Fig. 5A), and the ITS2 (Fig. S6A) were
278 consistent with previously established phylogenies (*1, 57, 58*). We then analysed the
279 phylogenetic signal of whole-genome sequence data using a *k*-mer-based alignment-free (AF)
280 approach (see Materials and Methods), focusing on distinct genome-sequence regions
281 following Lo et al. (*39*). Interestingly, the inferred AF phylogenies of introns (Fig. 5B),
282 repetitive regions (Fig. S6B), repeat-masked whole-genome sequences (Fig. 5C), and entire
283 whole-genome sequences (Fig. S6C) placed Ev as the basal group, branching earlier than
284 S1/S2. In comparison, the AF phylogenies for coding regions (i.e., coding sequences [CDS;
285 Fig. 5D] and protein sequences [Fig. S6D]) were largely congruent with the 18S rDNA
286 phylogeny (Fig. 5A), placing S1 as earlier branching than Ev. The branching orders of Ev vs
287 S1/S2 in AF trees were supported by a robust jackknife support of ≥ 96% based on 100

288 subsampled replicates (Fig. 5B-D; see Materials and Methods). This trend is consistent with
289 visualisation of the AF distances as a phylogenomic network (Fig. S6E), in which *E. voratum*
290 is more closely related to *P. glacialis* based on intron sequences, and to *Symbiodinium* spp.
291 based on CDS.

292 The incongruence between phylogenies of coding and non-coding regions with robust
293 support of the distinct clades clearly indicate differential selective pressure acting on these
294 two regions in Symbiodiniaceae genomes, as demonstrated in an earlier study (*39*). This is
295 likely explained by incomplete lineage sorting, horizontal gene transfer, hybridisation, or
296 convergent GC-biased gene conversion (*59-61*). This result may also reflect the retention of
297 ancestral non-coding regions in the *E. voratum* genomes, and/or the loss of some non-coding
298 regions in symbiotic lineages due to genome streamlining.

299 **Symbiogenesis differentially affected early- versus late-branching symbiotic lineages of**
300 **Symbiodiniaceae**

301 Following a possible genome reduction in the free-living ancestor of the Suessiales, some
302 lineages may have developed symbiogenesis with a range of hosts, giving rise to the
303 Symbiodiniaceae Family. Common and distinct genomic features we observed between early
304 and late-branching symbiotic lineages of Symbiodiniaceae suggest an interplay between the
305 geological eras during which they arose, and the corresponding coral morphology and ocean
306 chemistry (Fig. 6). Ancestral Symbiodiniaceae inhabited stony corals presumably as early as
307 230 MYA in the late Triassic (*2*) and may have driven the Norian-Rhaetian reef bloom (*62*).
308 These early Scleractinian corals (e.g., *Retiophyllia*) tended to be uniserial, i.e., possessing one
309 corallite per branch, and phaceloid with thick walls (*63, 64*), and thus were less efficient at
310 harvesting light (*65*). The ability of the early-diverging extant *Symbiodinium* to thrive under
311 high or variable light (*1*) may be a trait inherited from their ancestor living in these ancient
312 corals. Because these early Symbiodiniaceae adapted to different hosts, they likely underwent
313 genome streamlining (*9*), experienced high pseudogenisation and a reduction in mRNA
314 editing and intron sizes (Fig. 6), as observed here and in other studies (*21, 56*). Although
315 these trends were also observed in the later-branching symbionts, *Symbiodinium* uniquely
316 retained ancestral LINE repeats (Fig. 2C) which were lost in later-branching
317 Symbiodiniaceae including *E. voratum.*

318 *E. voratum* is estimated to have diverged ~147 MYA in the late Jurassic (*1*). We did
319 not find evidence for genome streamlining in *E. voratum*, but instead found genomic
320 hallmarks associated with earlier-branching free-living lineages external to Family
321 Symbiodiniaceae, e.g., more IE-containing genes and larger intron size than in S1 and S2.
322 Given that *Effrenium* is also the sole genus within the Family that is exclusively free-living,
323 we postulate that this lineage has not been impacted by symbiogenesis, in contrast to S1 or
324 S2. Since the lineage diverged from S1, global events such as the breakup of Pangea (*11, 66*)
325 and the Cretaceous-Paleogene mass extinction have occurred, along with changes in coral
326 reef biomass (Fig. 6). Without additional evidence from the fossil record or ancient DNA
327 analysis (oldest evidence of Suessiales is *P. glacialis* from just 9,000 years ago (*67*)), we
328 cannot explain why *Effrenium* retained a free-living lifestyle. *Effrenium* and most other
329 Symbiodiniaceae lineages (including free-living species from *Symbiodinium*) are capable of
330 forming endolithic relationships with bacteria, e.g., as calcified biofilms (*68, 69*), but why
331 *Effrenium* cannot form endosymbiosis with a host is unknown.

332 At the estimated time when S2 lineages diversified (~109 MYA), shallow-water
333 corals were described to be multiserial, flatter, and more efficient at harvesting light (*65*).
334 This is coincident with a rise in ocean temperature (i.e., "greenhouse Earth", when no
335 continental glaciers exited) and the switch in ocean chemistry from an aragonite sea to a
336 calcite sea, which would have made it difficult for corals to build their aragonite skeletons
337 (*14*) (Fig. 6). In contrast, bivalve rudists that could build aragonite or calcite shells (*70*)
338 radiated and flourished (*13, 15*). These taxa likely harboured photosymbionts (*12*),
339 presumably ancestral Symbiodiniaceae, given that extant Symbiodiniaceae (e.g., *S.*
340 *tridacnidorum*) can inhabit modern bivalves (*71*). Although genomes of S2 exhibit lower GC
341 content than those of S1, our results clearly indicate that both S1 and S2 underwent genome
342 streamlining, which led to convergent evolution of protein families associated with functions
343 relevant to forming a symbiotic association, such as cell signalling, apoptosis, and
344 photosynthesis (Fig. 3). We posit that symbiogenesis drove genome evolution in
345 Symbiodiniaceae and elicited common features such as pseudogenisation, lowered mRNA
346 editing and intron contraction, but some features (such as LINE retention and GC content)
347 were affected differently in earlier vs. later-branching symbiotic lineages.

348 ## Discussion

349 Our results provide strong evidence for a phase of genome reduction that occurred in the
350 Suessiales ancestor. Therefore, the Symbiodiniaceae have smaller genome sizes than most
351 free-living dinoflagellates, not because of the coral symbiosis, but due to more ancient
352 selective constraints. These results are consistent with (but do not prove) the appealing idea
353 that symbiosis offered an "escape" from reduced functional capacity due to genome reduction
354 during the early stages of Symbiodiniaceae evolution. Regardless, the mixture of obligate
355 free-living to facultative lifestyles among extant Symbiodiniaceae has resulted in divergent
356 paths of genome evolution. Our results demonstrate the retention of ancestral
357 Symbiodiniaceae genome features in *E. voratum* (in contrast to symbiotic lineages) despite
358 multiple emergences of symbiogenesis over the past 200 million years. These observations
359 support the notion that evolution favoured a free-living lifestyle for *E. voratum* (and by
360 extension the genus *Effrenium*), likely due to local selective pressures. Therefore, *Effrenium*
361 presents a useful free-living outgroup for studying the structural and functional genome
362 features of symbiotic Symbiodiniaceae, and the implications of these features on ecology and
363 evolution, including but not limited to host specificity and the "facultativeness" of symbiotic
364 associations.

365 ## Materials and Methods

366 ### *E. voratum* cultures

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

371 ### Microscopy

372 Cell cultures of *B. minutum* Mf1.05b.01, *D. trenchii* CCMP2556, *E. voratum* CCMP421, *S.*
373 *microadriaticum* Cass KB8, and *Symbiodinium linucheae* SSA01 were grown in f/2 medium
374 (*72, 73*) prepared using artificial seawater. The cells were fixed for 30 minutes in 4% v/v
375 formaldehyde:f/2 medium (salinity 35 ppt). Cells were then pelleted by centrifugation (300 *g*,

376 5 min, room temperature [RT]) and washed twice using the f/2medium. All micrographs were
377 captured at 100x (with a 0.63x adapter) on an Olympus BX 63 in brightfield mode.

378 **Extraction of genomic DNA**

379 To extract genomic DNA (gDNA) for RCC1521 and rt-383, cells were pelleted by spinning
380 the culture at 300 *g* (5 min, RT). Then the cell pellet was resuspended in 100-500 µL (higher
381 volumes for more cells) prewarmed (at 60°C) lysis buffer (100mM Tris-HCl, 20mM EDTA,
382 4% CTAB (w/v), 1.4 NaCl, 1% 360,000 $g \cdot mol^{-1}$ PVP (w/v), 2% β-mercaptoethanol). This
383 mixture was transferred to liquid nitrogen pre-chilled mortar pestle and ground to a fine
384 powder and high molecular weight (HMW) gDNA was extracted as described in Rai (*74*).
385 Briefly, Proteinase K (200 µL of stock at 20 mg/mL) was added, multiple rounds of
386 chloroform:isoamyl alcohol (24:1 v/v) extractions were performed, the CTAB-DNA pellet
387 was washed using ethanol, RNase A (4 µL of stock at 20 mg/mL) was added, and HMW
388 gDNA was captured by incubating with 7.5 M ammonium acetate. The final HMW gDNA
389 was resuspended in Tris-HCl (10 mM, pH 8) prewarmed at 50 °C, then was stored at -20°C
390 until sequencing.

391 *E. voratum* CCMP421 cells were pelleted and snap frozen in liquid nitrogen and
392 ground along with glass beads (diameter 425-600 µm), and the 2×CTAB method described in
393 Stephens et al. (*19*) was followed. Briefly, the ground powder was transferred into a lysis
394 buffer (100 mM Tris-HCl pH 8, 20 mM EDTA pH 8, 1.4 M NaCl, 2% w/v CTAB), mixed
395 with RNAse A (final concentration 20 µg/mL) and incubated at 37°C (30 min), and then with
396 Proteinase K (final concentration 120 µg/mL) at 65°C (2h). The standard extractions using
397 phenol-chloroform-isoamyl alcohol (25:24:1 v/v; centrifuged at 14,000 *g*, 5 min, RT) and
398 chloroform-isoamyl alcohol (24:1 v/v; centrifuged at 14,000 *g*, 5 min, RT) were then
399 performed. The DNA was precipitated using chilled isopropanol (18,000 *g*, 5 min, 4°C),
400 washed using chilled 70% ethanol, then stored in Tris-HCl (10 mM, pH 8) until sequencing.

401 **Extraction of total RNA**

402 To extract total RNA from *E. voratum* RCC1521 for Iso-Seq sequencing, cell pellets were
403 lysed via five freeze-thaw cycles with 425-600 µm diameter glass beads added to cell pellet,
404 dipped in liquid nitrogen, vortexed, and thawed at 37°C), and then the QIAGEN RNeasy kit

405 was used following the plant tissue protocol. To increase transcriptome diversity, we
406 extracted more RNA using a different method. Total RNA was extracted for RNA-Seq
407 (RCC1521 and rt-383) and Iso-Seq (from rt-383) following Acosta-Maspons et al. (*75*) with
408 slight modifications. Cells were first pelleted by centrifugation (300 *g*, 5 min). The cell pellet
409 was suspended in 100–500 µL lysis buffer (100mM Tris-HCl, 25 mM EDTA, 2% CTAB
410 w/v, 2M NaCl, 0.75 g/L spermidine trihydrochloride, 4% β-mercaptoethanol). This mixture
411 was transferred to a mortar and pestle that were prechilled in liquid nitrogen, and ground to a
412 fine powder. The powder was swirled with liquid nitrogen and transferred to a Falcon tube
413 (15 mL) standing in dry ice, then the liquid nitrogen was allowed to evaporate. Lysis buffer
414 (5 mL; prewarmed at 65°C) was added, mixed by vortexing, and the tube was kept on ice for
415 subsequent steps. An equal volume of chloroform:isoamyl alcohol (24:1 v/v) was mixed in;
416 the mixture was divided into aliquots (2mL) in Eppendorf tubes and centrifuged (10,000 *g*, 10
417 min, 4°C). The clear supernatants were transferred to new tubes (2mL each) before the
418 extraction with chloroform:isoamyl alcohol (24:1 v/v) was repeated. Nucleic acids were
419 precipitated using LiCl (final concentration 2M) overnight at 4°C, and centrifuged (17,000 *g*,
420 30 min, 4°C). The resulting pellet was resuspended in the SSTE buffer (50 µL; 1 M NaCl,
421 0.5% SDS w/v, 10mM Tris-HCl pH 8, 1mM EDTA; prewarmed at 65°C); the samples were
422 not set on ice to avoid undesirable precipitates. Further chloroform:isoamyl alcohol (24:1 v/v)
423 extraction was performed (centrifuged 10,000 *g*, 10 min, 4°C), and the RNA pellet was
424 precipitated using an equal volume of 2-propanol (incubated 10 min at RT, then centrifuged
425 at 17,000 *g*, 10 min, 4°C). The pellet was washed with 80% ethanol (500 µL), dislodged by
426 pulse vortexing (2 s), and the tube centrifuged (17,000 *g*, 5 min, 4°C); this step was repeated.
427 The final pellet was air-dried (10 min) in a fume hood, resuspended nuclease-free water (25
428 µL), and stored at -80°C.

429 **Generation of genome data**

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

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

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

452 **Generation of transcriptome data**

453 Transcriptome data were generated for RCC1521 and rt-383. Illumina RNA-Seq libraries
454 were generated using polyA-selection with the Dynabeads mRNA purification Kit and the
455 Illumina Stranded mRNA Prep following standard protocols. Sequencing was performed on
456 the Illumina NovaSeq 6000 platform at the Australian Genome Research Facility
457 (Melbourne, Australia). PacBio Iso-Seq libraries were prepared using the NEBNext® Single
458 Cell/Low Input cDNA Synthesis and Amplification Module (New England BioLabs) and the
459 SMRTbell Express Template Prep Kit 2.0 following standard protocol and sequenced on the
460 PacBio Sequel II at the University of Queensland Sequencing Facility.

461 **Transcriptome assembly and processing**

462 For RNA-Seq data from RCC1521 and rt-383, adapters and unique molecular identifiers were
463 removed using Illumina's bcl2fastq v2.20.0.422, and the reads were trimmed (for polyG and
464 polyA) and filtered using fastp v0.20.0 (*-A -L 35 -g -x --cut_front --cut_window_size 4 --*
465 *cut_mean_quality 15*). The processed reads were assembled using Trinity v2.9.1 (*76*) in both

466 "de novo" (*--SS_lib_type RF --trimmomatic*) and "genome-guided" modes; for the latter,
467 RNA-Seq reads were first mapped to the assembled genome using HISAT2 (*77*) before
468 Trinity was run using *--SS_lib_type RF --genome_guided_bam --genome_guided_max_intron*
469 *70000*. We assessed the completeness of the each transcript set using BUSCO v5.1.2 (*78*)
470 against the alveolata_odb10 database (Table S1).

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

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

477 Illumina short reads were used for estimating genome size based on *k*-mers. First, the reads
478 from each genome dataset were processed to remove potential adapters (for 10X linked-reads
479 of CCMP421, specifically the first 23 bases) and polyG tails using fastp v0.20.0 (*79*).
480 Jellyfish v2.3.0 (*80*) was used to obtain *k*-mers of sizes 17-31. The *k*-mer distribution was
481 then plotted for each *k,* and the genome size was estimated as the sum of observed *k*-mers
482 divided by the mean coverage (corresponding to the peak of the curve), averaged from the
483 different *k*-mer sizes (Table S2). The ploidy of the genome datasets was assessed using
484 GenomeScope2 (*81*) based on 21-mer distribution and a better fit was observed using a
485 haploid (p=1) model in all three cases. Genome size estimations of other dinoflagellates
486 (Table S12) were taken from Rizzo et al. (*82*), Hou and Lin (*28*), Sano and Kato (*83*), and
487 Kohli et al. (*84*).
488

489 ***De novo* genome assembly**

490 *De novo* genome assemblies were generated for RCC1521 and rt-383 adopting a hybrid-data
491 approach, combining Illumina (short reads), PacBio (long reads), and Nanopore (long reads)
492 data using MaSuRCA (*85, 86*) v4.0.1 for RCC1521, and v3.4.2 for rt-383, with the built-in
493 CABOG as the final assembler; the key distinction between these two versions of MaSuRCA
494 is the six-fold decrease in run-time in v4.0.1 relative to that for v3.4.2, with negligible impact

495 on the yielded assemblies. For CCMP421, the *de novo* genome assembly was generated from
496 10X Genomics linked-read sequencing data using Supernova v2.1.1 (*87*).

497 RCC1521 and rt-383 assemblies were further scaffolded with L_RNA_scaffolder
498 (*88*), using IsoSeq transcripts and *de novo* assembled transcripts from RNA-Seq (above). For
499 CCMP421, linked-read distance information was first used to refine the assembly with
500 ARBitR v0.2 (*-m 27k -s 10k*) (*89*) prior to scaffolding with L_RNA_scaffolder (*88*). Due to
501 the low quality of the publicly available transcriptome data of CCMP421 (i.e.,
502 MMETSP1110 (*90*) with only 54% mapped to the corresponding assembled genome; Table
503 S13), we used the *de novo* assembled transcripts from RCC1521 and rt-383 to scaffold the
504 CCMP421 genome assembly via L_RNA_scaffolder.

505 To ensure high quality of each genome assembly, we identified and removed potential
506 contaminant sequences of bacterial or archaeal sources following a decision tree based on
507 analysis using BlobTools v1.1 (*91*) as described in earlier studies of algal genomes (*22, 92*);
508 this step yielded the final assembly for each isolate. For each assembly, we assessed data
509 completeness using BUSCO v5.1.2 (*78*) against the alveolata_odb10 database. Genome-
510 sequence similarity among the three *E. voratum* isolates was assessed using nucmer
511 implemented in the MUMmer package v4.0.0beta2 (*--mum*) (*93*).

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

513 We identified putative mitochondrial and plastid genome sequences from the three isolates of
514 *E. voratum*. To search for mitochondrial scaffolds, we followed Stephens et al. (*19*) by
515 adopting BLASTn v2.10.0+ (*94*) search against the assembled genomes, using the protein-
516 coding sequences of *B. minutum* mitochondrial genes as queries (GenBank accessions
517 LC002801.1 and LC002802.1; $E \leq 10^{-10}$). We used BEDtools merge (*95*) to merge
518 overlapping BLAST hits of the same gene (*-s -o collapse -c 1,2,3,4,5,6*), then used BEDtools
519 intersect (*-wa -wb*) to check for overlaps with the predicted gene annotations. Genome
520 scaffolds with BLAST hits that did not have any other predicted genes were considered
521 putative mitochondrial genomes. We manually annotated the genes on the putative genome
522 scaffolds using Artemis (*96*) with translation table 4.

523 To identify plastid genome fragments that are known to be shorter and may not be
524 recovered in a hybrid genome assembly combing long- and short-read sequence data, we
525 performed an independent short-read only genome assembly for each isolate using CLC
526 Genomics Workbench v21.0.4 (Table S14). We used BLASTn search ($E \leq 10^{-10}$) using
527 protein-coding sequences of plastid-encoded genes for *Cladocopium* sp. C3 (GenBank
528 accessions HG515015.1-HG515028.1) as query, and annotated the putative genome scaffolds
529 in Artemis (translation table 11). To look for empty minicircles, we first determined the core
530 region of all plastid genome scaffolds. We masked coding regions of the scaffolds using
531 BEDtools *maskfasta* and did pairwise BLASTn searches. The region that was common to
532 most scaffolds was considered the core region. We then used this as query in BLASTn
533 searches among unannotated *E. voratum* short-read genome scaffolds. A scaffold was
534 considered empty if they did not have any hits in the NCBI nr database (November 2020).
535 Finally, we looked for evidence of circularisation in all of the recovered organellar genome
536 scaffolds using nucmer (*--mum -l 0*) and mummerplot (*--layout -- png*) from MUMmer
537 4.0.0beta2 (*93*) to self-align the genome sequences and visualise the alignments.

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

539 To predict protein-coding genes, we used an integrated, multi-method workflow customised
540 for dinoflagellates (*22*) incorporating transcriptome and protein evidence (pipeline available
541 at https://github.com/TimothyStephens/Dinoflagellate_Annotation_Workflow). We first
542 predicted repetitive elements from the genome assembly, followed by gene predictions based
543 on a) the genome only, b) protein sequence similarity, and c) mRNA transcripts.

544 *De novo* repeat families were predicted from the genome assembly using
545 RepeatModeler v2.0.1 (*97*), and these repeats were added to the Dfam database (dfam.org;
546 downloaded June 2019) to guide RepeatMasker v4.1.0 (*98*) in repeat-masking the genome
547 assembly. Next, GeneMark-ES v4.65 (*99*) was used for *ab initio* gene prediction on the
548 masked genome assembly. Protein-based gene prediction was performed using MAKER
549 v2.31.10 (*protein2genome*) (*100*) modified to recognise dinoflagellate alternative splice sites.
550 It integrated the custom repeat library from the repeat analysis step, and BLASTn/x (*101*)
551 searches were performed against the combined protein sequence database of SwissProt
552 (released March 2020) and the Suessiales sequences, hereby "Suessiales_pep", listed in Table
553 S7 of Chen et al. (*22*).

554 For transcript-based gene prediction, Iso-Seq transcripts where available (i.e., for
555 RCC1521 and rt-383) were mapped on the corresponding genome assembly using minimap2
556 v2.18 (*102*) for which the code was modified to recognise dinoflagellate alternative splice
557 sites, using options *--secondary=no -ax splice:hq -uf --splice-flank=no.* The assembled
558 transcripts from RNA-Seq (both *de novo* and genome-guided) for each isolate were mapped
559 to the corresponding genome assembly using BLAT (*103*); for CCMP421, *de novo* assembled
560 transcripts from RCC1521 and rt-383, plus the assembly of these reads guided by the
561 CCMP421 genome, were used in this step (due to poor quality of the publicly available
562 CCMP421 transcriptome data; see Tables S1 and S13). The resulting GFF3 files were input
563 into PASA v2.4.1 (*--IMPORT_CUSTOM_ALIGNMENTS_GFF3 --*
564 *transcribed_is_aligned_orient -C -R --MAX_INTRON_LENGTH 70000*) (*104*) which was
565 modified to recognise dinoflagellate alternative splice sites.

566 The PASA-predicted genes were filtered in the following steps: they were searched
567 against the combined RefSeq (release 98) and the Suessiales_pep database using BLASTp
568 v2.2.26 (e-value < 10^{-20}, both query and subject coverage > 80%) (*101*), putative transposon
569 sequences were removed via running HHBLITS v3.3.0 (*105*) and TransposonPSI v1.0.0
570 (*106*) against the UniRef30_3030_03 database from Uniclust (uniclust.mmseqs.com) (*107*),
571 redundant sequences were removed using CD-HIT v4.8.1 (*-c 0.75 -n 5*) (*108*), and the script
572 *Prepare_golden_genes_for_predictors.pl* from the JAMg pipeline
573 (https://github.com/genomecuration/JAMg) was used to produce a highly curated set of
574 "golden genes". These golden genes were used to guide the *ab initio* gene prediction tools
575 SNAP (*109*) and AUGUSTUS v3.4.0 (*110*) on the repeat-masked genome assembly. Gene
576 models from the five tools (MAKER, GeneMark-ES, PASA, SNAP, AUGUSTUS) were
577 integrated using EVidenceModeler v1.1.1 (*111*). Finally, the resulting gene models were
578 refined to correct exon boundaries, add UTRs, and incorporate gene models from alternative
579 splicing using the *Load_Current_Gene_Annotations.dbi* and *Launch_PASA_pipeline.pl* from
580 the PASA pipeline (*112, 113*) iteratively for three rounds to yield the final gene models.

581 We assessed the completeness of the predicted protein sequences using BUSCO
582 v5.1.2 (*78*) against the alveolata_odb10 database in "protein" mode. Genes with transcript
583 support were identified by aligning transcripts for each isolate to their predicted coding
584 sequences using BLASTn (e-value < 10^{-5}, percent identity ≥ 90 %, subject cover ≥ 50 %).

585 **Functional annotation of predicted genes**

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

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

593 To infer phylogenies for the 18S rDNA and ITS2 marker sequences, we first recovered 18S
594 and ITS2 sequences from the three genome assemblies of *E. voratum* using BLASTn.
595 Reference sequences for the 18S rDNA were downloaded from
596 https://doi.org/10.5061/dryad.1717129 (79 sequences) (*1*) and ITS2 from SymPortal
597 (https://symportal.org; "published post-MED sequences" of 8,409 sequences downloaded 10
598 September 2021) (*114*), respectively. For each marker sequence set, multiple sequence
599 alignment was generated using MAFFT v7.471 in *mafft-linsi* mode (*115*), trimmed using
600 trimAl v1.4.rev15 *(-automated1*) (*116*), from which a maximum-likelihood phylogenetic tree
601 was inferred using IQ-TREE v2.1.3 *(-nm 2000 -bb 2000 -m MFP*) (*117*). The 18S rDNA tree
602 based on LaJeunesse et al. (*1*) and the associated divergence times was overlayed with major
603 geological events and visualised using tvBOT (*118*).

604 To reconstruct a reference species tree of dinoflagellates based on strictly orthologous
605 protein sequences, we incorporated 1,603,073 predicted protein sequences from 33
606 dinoflagellate taxa, comprising 21 Suessiales taxa (including the three *E. voratum* isolates in
607 this study) and 12 other taxa external to Suessiales (Table S12) (*90, 119, 120*). These
608 sequences were clustered into homologous sets using OrthoFinder v2.5.4 (*121*), from which a
609 species tree was estimated from strictly orthologous sets.

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

611 We used an alignment-free approach to infer phylogenetic relationships from (a) whole-
612 genome sequences (WGS) and from distinct genomic regions of (b) repeat-masked WGS, (c)

613 coding sequences (CDS), (d) introns, (e) annotated repeats, and (f) predicted protein
614 sequences. Each of these distinct regions were extracted from assembled genome sequences
615 using *gff3_file_to_feature_files.pl* implemented in PASA (*104*). We followed Lo et al. (*39*) to
616 identify optimal *k*-mer length (*k*) for each of these datasets. Briefly, for each dataset, *k*-mers
617 at varied length *k* were enumerated using Jellyfish v2.3.0 (*80*); for all datasets, odd-numbered
618 *k* between 13 and 27 were used, except for the repeats dataset (odd-numbered *k* values
619 between 13 and 51 were used) and protein sequences (odd-numbered *k* values between 3 and
620 9 were used). For each dataset except the protein sequences, the optimal *k* was determined
621 based on the cumulative proportion of unique *k*-mers and the cumulative proportion of
622 distinct *k*-mers, at the point when distributions of both proportions reached a plateau (Fig.
623 S7); *k* value determined this way was found to yield the greatest distinguishing power for
624 phylogenetic analysis (*122*). For protein sequences, we inferred alignment-free phylogenies
625 (see below) from each *k* and chose the *k* with a topology that best matched the 18S rDNA
626 alignment-based tree as implemented in Lo et al. (*39*). The optimal *k* was identified as 23 for
627 WGS and repeat-masked WGS, 19 for CDS, 21 for introns, 51 for annotated repeats, and 9
628 for protein sequences. Jellyfish v2.3.0 was used to extract *k*-mers at the corresponding
629 optimal *k* length for each dataset, with the option *-C* used to enforce strand-specific
630 directionality for the intron and CDS datasets.
631 To infer alignment-free (AF) phylogenies based on *k*-mers (i.e. using the optimal *k* for
632 each corresponding dataset identified above), we derived pairwise distance based on D_2^S
633 statistic (*123*) following Chan et al. (*124*), using *d2ssect*
634 (https://github.com/bakeronit/d2ssect). These pairwise distances were then used to infer a
635 phylogenetic tree using *neighbor* implemented in PHYLIP v3.698 (*125*). For each tree, we
636 assessed node support based on jackknife analysis of 100 "pseudo-replicates" following
637 Bernard et al. (*126*). In each pseudo-replicate, 40% of the data, in iteratively subsampled 100-
638 bp regions at random, were deleted using the Python script *jackknife.py* from which an AF
639 tree was inferred; the R script *Jackknife.r* was then used to calculate jackknife support value
640 in percentage, among the pseudo-replicate trees, for each node in the original AF tree. These
641 scripts are available at https://github.com/chanlab-genomics/alignment-free-tools.
642 To identify core *k*-mers (*127*) that are shared by all 21 Suessiales genomes used in
643 this study, we used the optimal $k = 23$ for the WGS dataset. Using the extracted 23-mers from
644 the entire WGS dataset as input, core 23-mers were identified using the bash command *comm*
645 (-12). BEDtools (*95*) *intersect* was used to find regions of overlap between the core *k*-mers
646 and different genomic features.

647 **Analysis of gene family evolution**

648 To examine the gene family evolution between *E. voratum* and the earlier-/later-branching
649 symbiotic lineages of Symbiodiniaceae, we grouped 21 Suessiales protein sequence datasets
650 into three groups: Ev, S1, and S2 (Table S3), with Po (*P. glacialis* CCMP1383 and
651 CCMP2088) as the outgroup. The 811,661 protein sequences from the 21 Suessiales taxa
652 were clustered into homologous families using OrthoFinder v2.5.4 (*121*). We then subset the
653 shared/exclusive protein families among the different groups (Ev, S1, S2, and Po). GO
654 enrichment was performed using the topGO package in R (algorithm = "elim", statistic =
655 "fisher") for six comparisons: (a) shared genes in Ev+Po (test set) versus all genes in Ev+Po
656 (background), (b) shared genes in S1+S2 versus all genes in S1 and S2, (c) shared genes in
657 S1+S2+Po versus all genes in S1, S2, and Po, (d) shared genes in S1+S2+Ev+Po versus all
658 genes in the 21 taxa, (e) exclusive genes to S1 versus all S1 genes, and (f) exclusive genes to
659 Ev versus all Ev genes.

660 **Identification of pseudogenes**

661 Pseudogenes were identified following the method described in González-Pech et al. (*21*)
662 based on tBLASTn search using the predicted protein sequences as query against the
663 corresponding genome sequences for which the predicted gene model sequences were
664 masked. Matched regions (≥ 75% identity) were considered fragments of pseudogenes, and
665 fragments at no more than 1 Kb apart and in the same orientation were considered
666 collectively as a pseudogene.

667 In this analysis, we focused on 752,954 protein sequences from the 19 Suessiales taxa,
668 specifically excluding *S. natans* and *S. pilosum* from S1 to avoid signatures of free-living
669 lifestyle in these taxa interfering with potential signatures of symbiogenesis. The protein
670 sequences were first clustered into homologous families using OrthoFinder v2.5.4 (*121*). We
671 define the extent of pseudogenisation, Ψ, as the ratio of the number of putative pseudogenes
672 to the number of putative functional genes in a homologous family. We determined this value
673 independently for Ev (Ψ_{Ev}) against that for S1 (Ψ_{S1}), S2 (Ψ_{S2}), and the combined S1 and S2
674 (Ψ_{S1+S2}); a protein family with $\Psi_{S1} > \Psi_{Ev}$ indicates a greater extent of pseudogenisation in S1
675 than in Ev. We then ran a one-way ANOVA test and Tukey's test using the R package *rstatix*

676 (https://cran.r-project.org/package=rstatix) to assess correlation among the Ψ values of Ev,
677 S1, S2, and the combined S1+S2 groups.

678 To assess the robustness of our approach to the inflation parameter (*I*) in single-
679 linkage clustering within OrthoFinder that modulates granularity (i.e., sizes and numbers) of
680 the resulting protein families, we ran OrthoFinder independently using *I* = 1.1, 1.3, 1.7 and
681 2.0 in addition to the default value of 1.5, and performed the analysis of pseudogenisation per
682 above.

683 **Analysis of mRNA editing**

684 For the analysis of mRNA editing, we focused on *E. voratum* RCC1521, and the
685 representative genomes for S1 (*S. microadriaticum* CCMP2467) and S2 (*D. trenchii*
686 CCMP2556), for which high-quality genome and transcriptome data are available. Editing of
687 mRNAs was identified using JACUSA2 (*128*), based on observed nucleotide variants in the
688 mapping of transcripts onto the genome, relative to the mapping of genome sequence reads
689 onto the genome. First, gDNA reads were mapped on each genome using BWA-mem
690 v0.7.17-r1198 (*129*) using default settings. Then, for each genome, RNA-Seq reads were
691 mapped using HISAT2 v2.2.0 *(--rna-strandness RF*) (*77*) to the assembled genome
692 sequences with a HGFM HISAT2 index (*hisat2-build --exon --ss*) informed by the annotated
693 splice sites. To generate the index, the gene annotation file in GFF3 was converted to the
694 GTF format using Gffread (*130*), and splice site and exon locations extracted with the
695 *hisat2_extract_splice_sites.py* and *hisat2_extract_exons.py* scripts. Iso-Seq reads, where
696 available, were mapped using minimap2 v2.18 (*102*) for which the code was modified to
697 recognise alternative splice sites of dinoflagellates, with options *--splice-flank=no –*
698 *secondary=no -ax splice:hq -uf –junc-bed*.

699 Duplicate mappings were removed from the gDNA BAM files using Picard
700 MarkDuplicates (*ASSUME_SORTED=true REMOVE_DUPLICATES=true*
701 *CREATE_INDEX=TRUE VALIDATION_STRINGENCY=LENIENT*). The MD field
702 documenting mismatched and deleted bases was added to the gDNA BAM files with
703 samtools calmd (*-b*), as required as input for JACUSA2. JACUSA2 analysis was performed
704 on the gDNA, RNA-Seq, and Iso-Seq BAM files using option *-a D,Y,H* to remove false
705 positives caused by read starts/ends, indels, splice sites, and homopolymers. Due to the

706 different strand directionality, *-P2 RF-FIRSTSTRAND* was specified for the runs
707 incorporating RNA-Seq data, and *-P2 FR-SECONDSTRAND* for the runs using Iso-Seq data.
708 The results from JACUSA2 were overlayed using BEDtools intersect (*95*) at *-s -wo* with
709 predicted genes (including isoforms), and the edited sites were visualised using
710 JACUSA2helper (https://github.com/dieterich-lab/JACUSA2helper).

711 We followed Liew et al. (*56*) to assess 5′ bias in the location of RNA editing and the
712 propensity for edits to occur together. We calculated the frequency of the locations of each
713 edit with respect to the features they were in (exon, intron, gene), normalised by the length of
714 each feature. For each edit, its distance (in bp) to the closest upstream edit, and that to the
715 closest downstream edit where all within the same gene, were determined. The average of
716 these two values was used as the per-edit observed distance to neighbouring edits.

717 **Analysis of introner elements**

718 To identify introner elements (IEs), we used Pattern Locator (*131*) to search for the patterns
719 described in Farhat et al. (*53*): inverted repeats of 8-20 nucleotides within 30 bases of the 5′
720 and 3′ ends of each intron, flanked by direct repeats of 3-5 nucleotides. We first used Seqkit
721 (*132*) to obtain the first and last 30 bases at each end of introns, then used Pattern Locator to
722 identify the IEs.

723 ## References

724 1. T. C. LaJeunesse, J. E. Parkinson, P. W. Gabrielson, H. J. Jeong, J. D. Reimer, C. R.
725 Voolstra, S. R. Santos, Systematic revision of Symbiodiniaceae highlights the
726 antiquity and diversity of coral endosymbionts. *Curr. Biol.* **28**, 2570-2580 (2018).

727 2. K. Frankowiak, E. Roniewicz, J. Stolarski, Photosymbiosis in Late Triassic
728 scleractinian corals from the Italian Dolomites. *PeerJ* **9**, e11062 (2021).

729 3. R. Guerrero, L. Margulis, M. Berlanga, Symbiogenesis: the holobiont as a unit of
730 evolution. *Int. Microbiol.* **16**, 133-143 (2013).

731 4. J. G. Stephen, H. J. Tripp, S. Givan, M. Podar, L. Vergin Kevin, D. Baptista, L.
732 Bibbs, J. Eads, H. Richardson Toby, M. Noordewier, S. Rappé Michael, M. Short Jay,
733 C. Carrington James, J. Mathur Eric, Genome streamlining in a cosmopolitan oceanic
734 bacterium. *Science* **309**, 1242-1245 (2005).

5. J. C. Kwan, S. Donia Mohamed, W. Han Andrew, E. Hirose, G. Haygood Margo, W. Schmidt Eric, Genome streamlining and chemical defense in a coral reef symbiosis. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 20655-20660 (2012).

6. H. G. Morrison, G. McArthur Andrew, D. Gillin Frances, B. Aley Stephen, D. Adam Rodney, J. Olsen Gary, A. Best Aaron, W. Z. Cande, F. Chen, J. Cipriano Michael, J. Davids Barbara, C. Dawson Scott, G. Elmendorf Heidi, B. Hehl Adrian, E. Holder Michael, M. Huse Susan, U. Kim Ulandt, E. Lasek-Nesselquist, G. Manning, A. Nigam, E. J. Nixon Julie, D. Palm, E. Passamaneck Nora, A. Prabhu, I. Reich Claudia, S. Reiner David, J. Samuelson, G. Svard Staffan, L. Sogin Mitchell, Genomic minimalism in the early diverging intestinal parasite *Giardia lamblia*. *Science* **317**, 1921-1926 (2007).

7. C. H. Opperman, M. Bird David, M. Williamson Valerie, S. Rokhsar Dan, M. Burke, J. Cohn, J. Cromer, S. Diener, J. Gajan, S. Graham, T. D. Houfek, Q. Liu, T. Mitros, J. Schaff, R. Schaffer, E. Scholl, R. Sosinski Bryon, P. Thomas Varghese, E. Windham, Sequence and genetic map of *Meloidogyne hapla*: a compact nematode genome for plant parasitism. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 14802-14807 (2008).

8. J. P. McCutcheon, N. A. Moran, Extreme genome reduction in symbiotic bacteria. *Nat. Rev. Microbiol.* **10**, 13-26 (2012).

9. R. A. González-Pech, D. Bhattacharya, M. A. Ragan, C. X. Chan, Genome evolution of coral reef symbionts as intracellular residents. *Trends Ecol. Evol.* **34**, 799-806 (2019).

10. S. Vulpius, W. Kiessling, New constraints on the last aragonite–calcite sea transition from early Jurassic ooids. *Facies* **64**, 3 (2017).

11. A. Schettino, E. Turco, Breakup of Pangaea and plate kinematics of the central Atlantic and Atlas regions. *Geophys. J. Int.* **178**, 1078-1097 (2009).

12. N. J. de Winter, S. Goderis, S. J. M. Van Malderen, M. Sinnesael, S. Vansteenberge, C. Snoeck, J. Belza, F. Vanhaecke, P. Claeys, Subdaily-scale chemical variability in a *Torreites sanchezi* rudist shell: implications for rudist paleobiology and the Cretaceous day-night cycle. *Paleoceanogr. Paleoclimatol.* **35**, e2019PA003723 (2020).

13. G. D. Stanley, B. van de Schootbrugge, in *Coral Bleaching: Patterns, Processes, Causes and Consequences,* M. J. H. van Oppen, J. M. Lough, Eds. (Springer International Publishing, Cham, 2018), pp. 9-26.

14. J. M. Pandolfi, W. Kiessling, Gaining insights from past reefs to inform understanding of coral reef response to global climate change. *Curr. Opin. Environ. Sustain.* **7**, 52-58 (2014).

771 15. J. E. N. Veron, in *Encyclopedia of Modern Coral Reefs: Structure, Form and Process,*
772 D. Hopley, Ed. (Springer Netherlands, Dordrecht, 2011), pp. 947-957.

773 16. D. J. Thornhill, A. M. Lewis, D. C. Wham, T. C. LaJeunesse, Host-specialist lineages
774 dominate the adaptive radiation of reef coral endosymbionts. *Evolution* **68**, 352-367
775 (2014).

776 17. K. E. Turnham, D. C. Wham, E. Sampayo, T. C. LaJeunesse, Mutualistic microalgae
777 co-diversify with reef corals that acquire symbionts during egg development. *ISME J.*
778 **15**, 3271-3285 (2021).

779 18. C. C. Butler, K. E. Turnham, A. M. Lewis, M. R. Nitschke, M. E. Warner, D. W.
780 Kemp, O. Hoegh-Guldberg, W. K. Fitt, M. J. H. van Oppen, T. C. LaJeunesse, Formal
781 recognition of host-generalist species of dinoflagellate (*Cladocopium*,
782 Symbiodiniaceae) mutualistic with Indo-Pacific reef corals. *J. Phycol.*,
783 doi:10.1111/jpy.13340 (2023).

784 19. T. G. Stephens, R. A. González-Pech, Y. Cheng, A. R. Mohamed, D. W. Burt, D.
785 Bhattacharya, M. A. Ragan, C. X. Chan, Genomes of the dinoflagellate *Polarella*
786 *glacialis* encode tandemly repeated single-exon genes with adaptive functions. *BMC*
787 *Biol.* **18**, 56 (2020).

788 20. T. Li, L. Yu, B. Song, Y. Song, L. Li, X. Lin, S. Lin, Genome improvement and core
789 gene set refinement of *Fugacium kawagutii*. *Microorganisms* **8**, 102 (2020).

790 21. R. A. González-Pech, T. G. Stephens, Y. Chen, A. R. Mohamed, Y. Cheng, S. Shah,
791 K. E. Dougan, M. D. A. Fortuin, R. Lagorce, D. W. Burt, D. Bhattacharya, M. A.
792 Ragan, C. X. Chan, Comparison of 15 dinoflagellate genomes reveals extensive
793 sequence and structural divergence in family Symbiodiniaceae and genus
794 *Symbiodinium*. *BMC Biol.* **19**, 73 (2021).

795 22. Y. Chen, S. Shah, K. E. Dougan, M. J. H. van Oppen, D. Bhattacharya, C. X. Chan,
796 Improved *Cladocopium goreaui* genome assembly reveals features of a facultative
797 coral symbiont and the complex evolutionary history of dinoflagellate genes.
798 *Microorganisms* **10**, 1662 (2022).

799 23. T. C. LaJeunesse, G. Lambert, R. A. Andersen, M. A. Coffroth, D. W. Galbraith,
800 *Symbiodinium* (Pyrrhophyta) genome sizes (DNA content) are smallest among
801 dinoflagellates. *J. Phycol.* **41**, 880-886 (2005).

802 24. O. S. Saad, X. Lin, T. Y. Ng, L. Li, P. Ang, S. Lin, Genome Size, rDNA Copy, and
803 qPCR Assays for Symbiodiniaceae. *Front. Microbiol.* **11**, 847 (2020).

804 25. P. J. Rizzo, The enigma of the dinoflagellate chromosome. *J. Protozool.* **38**, 246-252
805 (1991).

806 26. P. J. Rizzo, Those amazing dinoflagellate chromosomes. *Cell Res.* **13**, 215-217
807 (2003).

808 27. K. E. Dougan, Z.-L. Deng, L. Wöhlbrand, C. Reuse, B. Bunk, Y. Chen, J. Hartlich, K.
809 Hiller, U. John, J. Kalvelage, J. Mansky, M. Neumann-Schaal, J. Overmann, J.
810 Petersen, S. Sanchez-Garcia, K. Schmidt-Hohagen, S. Shah, C. Spröer, H. Sztajer, H.
811 Wang, D. Bhattacharya, R. Rabus, D. Jahn, C. X. Chan, I. Wagner-Döbler, Multi-
812 omics analysis reveals the molecular response to heat stress in a "red tide"
813 dinoflagellate. *bioRxiv*, 2022.2007.2025.501386 (2022).

814 28. Y. Hou, S. Lin, Distinct gene number-genome size relationships for eukaryotes and
815 non-eukaryotes: gene content estimation for dinoflagellate genomes. *PLoS ONE* **4**,
816 e6978 (2009).

817 29. H. J. Jeong, S. Y. Lee, N. Kang, Y. Yoo, A. S. Lim, M. J. Lee, W. Yih, H. Yamashita,
818 T. LaJeunesse, Genetics and morphology characterize the dinoflagellate
819 *Symbiodinium voratum*, n. sp., (Dinophyceae) as the sole representative of
820 *Symbiodinium* Clade E. *J. Eukaryot. Microbiol.* **61**, 75-94 (2014).

821 30. T. Xiang, E. A. Hambleton, J. C. DeNofrio, J. R. Pringle, A. R. Grossman, Isolation
822 of clonal axenic strains of the symbiotic dinoflagellate *Symbiodinium* and their growth
823 and host specificity. *J. Phycol.* **49**, 447-458 (2013).

824 31. Y. Gabay, V. M. Weis, S. K. Davy, Symbiont identity influences patterns of
825 symbiosis establishment, host growth, and asexual reproduction in a model cnidarian-
826 dinoflagellate symbiosis. *Biol. Bull.* **234**, 1-10 (2018).

827 32. G. Hansen, N. Daugbjerg, *Symbiodinium natans* sp. nov.: a "free-living"
828 dinoflagellate from tenerife (Northeast-Atlantic Ocean). *J. Phycol.* **45**, 251-263
829 (2009).

830 33. K. E. Dougan, A. J. Bellantuono, T. Kahlke, R. M. Abbriano, Y. Chen, S. Shah, C.
831 Granados-Cifuentes, M. J. H. van Oppen, D. Bhattacharya, D. J. Suggett, C. X. Chan,
832 M. Rodriguez-Lanetty, Whole-genome duplication in an algal symbiont
833 serendipitously confers thermal tolerance to corals. *bioRxiv*, 2022.2004.2010.487810
834 (2022).

835 34. S. J. Robbins, C. M. Singleton, C. X. Chan, L. F. Messer, A. U. Geers, H. Ying, A.
836 Baker, S. C. Bell, K. M. Morrow, M. A. Ragan, D. J. Miller, S. Forêt, E. Ball, R.
837 Beeden, M. Berumen, M. Aranda, T. Ravasi, P. Bongaerts, O. Hoegh-Guldberg, I.
838 Cooke, B. Leggat, S. Sprungala, A. Fitzgerald, C. Shang, P. Lundgren, T. Fyffe, F.
839 Rubino, M. van Oppen, K. Weynberg, S. J. Robbins, C. M. Singleton, C. Xin Chan,
840 L. F. Messer, A. U. Geers, H. Ying, A. Baker, S. C. Bell, K. M. Morrow, M. A.
841 Ragan, D. J. Miller, S. Foret, C. R. Voolstra, G. W. Tyson, D. G. Bourne, C. R.
842 Voolstra, G. W. Tyson, D. G. Bourne, C. ReFuGe, A genomic view of the reef-
843 building coral *Porites lutea* and its microbial symbionts. *Nat. Microbiol.* **4**, 2090-2100
844 (2019).

845 35. E. Shoguchi, C. Shinzato, T. Kawashima, F. Gyoja, S. Mungpakdee, R. Koyanagi, T.
846 Takeuchi, K. Hisata, M. Tanaka, M. Fujiwara, M. Hamada, A. Seidi, M. Fujie, T.
847 Usami, H. Goto, S. Yamasaki, N. Arakaki, Y. Suzuki, S. Sugano, A. Toyoda, Y.
848 Kuroki, A. Fujiyama, M. Medina, Mary A. Coffroth, D. Bhattacharya, N. Satoh, Draft
849 assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene
850 structure. *Curr. Biol.* **23**, 1399-1408 (2013).

851 36. A. Nand, Y. Zhan, O. R. Salazar, M. Aranda, C. R. Voolstra, J. Dekker, Genetic and
852 spatial organization of the unusual chromosomes of the dinoflagellate *Symbiodinium*
853 *microadriaticum*. *Nat. Genet.* **53**, 618-629 (2021).

854 37. E. Biquand, N. Okubo, Y. Aihara, V. Rolland, D. C. Hayward, M. Hatta, J.
855 Minagawa, T. Maruyama, S. Takahashi, Acceptable symbiont cell size differs among
856 cnidarian species and may limit symbiont diversity. *ISME J.* **11**, 1702-1712 (2017).

857 38. D. Bhattacharya, H. Qiu, J. Lee, H. Su Yoon, A. P. M. Weber, D. C. Price, When less
858 is more: red algae as models for studying gene loss and genome evolution in
859 eukaryotes. *Crit. Rev. Plant Sci.* **37**, 81-99 (2018).

860 39. R. Lo, K. E. Dougan, Y. Chen, S. Shah, D. Bhattacharya, C. X. Chan, Alignment-free
861 analysis of whole-genome sequences from Symbiodiniaceae reveals different
862 phylogenetic signals in distinct regions. *Front. Plant Sci.* **13**, 815714 (2022).

863 40. A. C. Barbrook, C. R. Voolstra, C. J. Howe, The chloroplast genome of a
864 *Symbiodinium* sp. clade C3 isolate. *Protist* **165**, 1-13 (2014).

865 41. A. E. Vinogradov, DNA helix: the importance of being GC-rich. *Nucleic Acids Res.*
866 **31**, 1838-1844 (2003).

867 42. P. Šmarda, P. Bureš, L. Horová, I. J. Leitch, L. Mucina, E. Pacini, L. Tichý, V.
868 Grulich, O. Rotreklová, Ecological and evolutionary significance of genomic GC
869 content diversity in monocots. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E4096-E4102
870 (2014).

871 43. R. Hershberg, D. A. Petrov, Evidence that mutation Is universally biased towards AT
872 in bacteria. *PLoS Genet.* **6**, e1001115 (2010).

873 44. H. Nikbakht, X. Xia, D. A. Hickey, The evolution of genomic GC content undergoes
874 a rapid reversal within the genus *Plasmodium*. *Genome* **57**, 507-511 (2014).

875 45. G. Blanc, G. Duncan, I. Agarkova, M. Borodovsky, J. Gurnon, A. Kuo, E. Lindquist,
876 S. Lucas, J. Pangilinan, J. Polle, A. Salamov, A. Terry, T. Yamada, D. D. Dunigan, I.
877 V. Grigoriev, J.-M. Claverie, J. L. Van Etten, The *Chlorella variabilis* NC64A
878 genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic
879 sex. *Plant Cell* **22**, 2943-2955 (2010).

880 46. L. Cai, B. J. Arnold, Z. Xi, D. E. Khost, N. Patel, C. B. Hartmann, S. Manickam, S.
881 Sasirat, L. A. Nikolov, S. Mathews, T. B. Sackton, C. C. Davis, Deeply altered
882 genome architecture in the endoparasitic flowering plant *Sapria himalayana* Griff.
883 (Rafflesiaceae). *Curr. Biol.* **31**, 1002-1011.e1009 (2021).

884 47. M. Osanai-Futahashi, Y. Suetsugu, K. Mita, H. Fujiwara, Genome-wide screening and
885 characterization of transposable elements and their distribution analysis in the
886 silkworm, *Bombyx mori*. *Insect Biochem. Mol. Biol.* **38**, 1046-1057 (2008).

887 48. N. A. Rayan, R. C. H. del Rosario, S. Prabhakar, Massive contribution of transposable
888 elements to mammalian regulatory sequences. *Semin. Cell Dev. Biol.* **57**, 51-56
889 (2016).

890 49. J. T. Huff, D. Zilberman, S. W. Roy, Mechanism for DNA transposons to generate
891 introns on genomic scales. *Nature* **538**, 533-536 (2016).

892 50. A. Z. Worden, J.-H. Lee, T. Mock, P. Rouzé, M. P. Simmons, A. L. Aerts, A. E.
893 Allen, M. L. Cuvelier, E. Derelle, M. V. Everett, E. Foulon, J. Grimwood, H.
894 Gundlach, B. Henrissat, C. Napoli, S. M. McDonald, M. S. Parker, S. Rombauts, A.
895 Salamov, P. Von Dassow, J. H. Badger, P. Coutinho, E. Demir, I. Dubchak, C.
896 Gentemann, W. Eikrem, J. E. Gready, U. John, W. Lanier, E. A. Lindquist, S. Lucas,
897 K. F. X. Mayer, H. Moreau, F. Not, R. Otillar, O. Panaud, J. Pangilinan, I. Paulsen, B.
898 Piegu, A. Poliakov, S. Robbens, J. Schmutz, E. Toulza, T. Wyss, A. Zelensky, K.
899 Zhou, E. V. Armbrust, D. Bhattacharya, U. W. Goodenough, Y. Van de Peer, I. V.
900 Grigoriev, Green evolution and dynamic adaptations revealed by genomes of the
901 marine picoeukaryotes *Micromonas*. *Science* **324**, 268-272 (2009).

902 51. A. van der Burgt, E. Severing, P. J. de Wit, J. Collemare, Birth of new spliceosomal
903 introns in fungi by multiplication of introner-like elements. *Curr. Biol.* **22**, 1260-1265
904 (2012).

905 52. L. Gozashti, S. W. Roy, B. Thornlow, A. Kramer, M. Ares, R. Corbett-Detig,
906 Transposable elements drive intron gain in diverse eukaryotes. *Proc. Natl. Acad. Sci.*
907 *U. S. A.* **119**, e2209766119 (2022).

908 53. S. Farhat, P. Le, E. Kayal, B. Noel, E. Bigeard, E. Corre, F. Maumus, I. Florent, A.
909 Alberti, J.-M. Aury, T. Barbeyron, R. Cai, C. Da Silva, B. Istace, K. Labadie, D.
910 Marie, J. Mercier, T. Rukwavu, J. Szymczak, T. Tonon, C. Alves-de-Souza, P. Rouzé,
911 Y. Van de Peer, P. Wincker, S. Rombauts, B. M. Porcel, L. Guillou, Rapid protein
912 evolution, organellar reductions, and invasive intronic elements in the marine aerobic
913 parasite dinoflagellate *Amoebophrya* spp. *BMC Biol.* **19**, 1 (2021).

914 54. M. J. McCoy, A. Z. Fire, Intron and gene size expansion during nervous system
915 evolution. *BMC Genomics* **21**, 360 (2020).

916 55. B. Wang, X. Liang, M. L. Gleason, T. Hsiang, R. Zhang, G. Sun, A chromosome-
917 scale assembly of the smallest Dothideomycete genome reveals a unique genome
918 compaction mechanism in filamentous fungi. *BMC Genomics* **21**, 321 (2020).

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

922 57. T. Shi, G. Niu, H. Kvitt, X. Zheng, Q. Qin, D. Sun, Z. Ji, D. Tchernov, Untangling
923 ITS2 genotypes of algal symbionts in zooxanthellate corals. *Mol. Ecol. Resour.* **21**,
924 137-152 (2021).

925 58. T. G. Stephens, M. A. Ragan, D. Bhattacharya, C. X. Chan, Core genes in diverse
926 dinoflagellate lineages include a wealth of conserved dark genes with unknown
927 functions. *Sci. Rep.* **8**, 17175 (2018).

928 59. E. D. Jarvis, S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li, S. Y. Ho, B. C.
929 Faircloth, B. Nabholz, J. T. Howard, A. Suh, C. C. Weber, R. R. da Fonseca, J. Li, F.
930 Zhang, H. Li, L. Zhou, N. Narula, L. Liu, G. Ganapathy, B. Boussau, M. S. Bayzid,
931 V. Zavidovych, S. Subramanian, T. Gabaldon, S. Capella-Gutierrez, J. Huerta-Cepas,
932 B. Rekepalli, K. Munch, M. Schierup, B. Lindow, W. C. Warren, D. Ray, R. E.
933 Green, M. W. Bruford, X. Zhan, A. Dixon, S. Li, N. Li, Y. Huang, E. P. Derryberry,
934 M. F. Bertelsen, F. H. Sheldon, R. T. Brumfield, C. V. Mello, P. V. Lovell, M.
935 Wirthlin, M. P. Schneider, F. Prosdocimi, J. A. Samaniego, A. M. Vargas Velazquez,
936 A. Alfaro-Nunez, P. F. Campos, B. Petersen, T. Sicheritz-Ponten, A. Pas, T. Bailey,
937 P. Scofield, M. Bunce, D. M. Lambert, Q. Zhou, P. Perelman, A. C. Driskell, B.
938 Shapiro, Z. Xiong, Y. Zeng, S. Liu, Z. Li, B. Liu, K. Wu, J. Xiao, X. Yinqi, Q. Zheng,
939 Y. Zhang, H. Yang, J. Wang, L. Smeds, F. E. Rheindt, M. Braun, J. Fjeldsa, L.
940 Orlando, F. K. Barker, K. A. Jonsson, W. Johnson, K. P. Koepfli, S. O'Brien, D.
941 Haussler, O. A. Ryder, C. Rahbek, E. Willerslev, G. R. Graves, T. C. Glenn, J.
942 McCormack, D. Burt, H. Ellegren, P. Alstrom, S. V. Edwards, A. Stamatakis, D. P.
943 Mindell, J. Cracraft, E. L. Braun, T. Warnow, W. Jun, M. T. Gilbert, G. Zhang,
944 Whole-genome analyses resolve early branches in the tree of life of modern birds.
945 *Science* **346**, 1320-1331 (2014).

946 60. K. O. Chan, C. R. Hutter, P. L. Wood, L. L. Grismer, R. M. Brown, Larger, unfiltered
947 datasets are more effective at resolving phylogenetic conflict: introns, exons, and
948 UCEs resolve ambiguities in golden-backed frogs (Anura: Ranidae; genus *Hylarana*).
949 *Mol. Phylogen. Evol.* **151**, 106899 (2020).

950 61. E. Pessia, A. Popa, S. Mousset, C. Rezvoy, L. Duret, G. A. B. Marais, Evidence for
951 widespread GC-biased gene conversion in eukaryotes. *Genome Biol. Evol.* **4**, 675-682
952 (2012).

953 62. W. Kiessling, E. Roniewicz, L. Villier, P. Leonide, U. Struck, An early Hettangian
954 coral reef in southern France implications for the end-Triassic reef crisis. *Palaios* **24**,
955 657-671 (2009).

63. A. G. Coates, B. C. J. Jeremy, Clonal growth, algal symbiosis, and reef formation by corals. *Paleobiology* **13**, 363-378 (1987).

64. G. D. Stanley, The evolution of modern corals and their early history. *Earth-Sci. Rev.* **60**, 195-225 (2003).

65. S. Enríquez, E. R. Méndez, O. Hoegh-Guldberg, R. Iglesias-Prieto, Key functional role of the optical properties of coral skeletons in coral ecology and evolution. *Proc. R. Soc. B* **284**, 20161667 (2017).

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

67. S. De Schepper, J. L. Ray, K. S. Skaar, H. Sadatzki, U. Z. Ijaz, R. Stein, A. Larsen, The potential of sedimentary ancient DNA for reconstructing past sea ice evolution. *ISME J.* **13**, 2566-2577 (2019).

68. M. R. Nitschke, C. Fidalgo, J. Simões, C. Brandão, A. Alves, J. Serôdio, J. C. Frommlet, Symbiolite formation: a powerful in vitro model to untangle the role of bacterial communities in the photosynthesis-induced formation of microbialites. *ISME J.* **14**, 1533-1546 (2020).

69. J. C. Frommlet, M. L. Sousa, A. Alves, S. I. Vieira, D. J. Suggett, J. Serôdio, Coral symbiotic algae calcify ex hospite in partnership with bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6158-6163 (2015).

70. T. Steuber, H. Löser, Species richness and abundance patterns of Tethyan Cretaceous rudist bivalves (Mollusca: Hippuritacea) in the central-eastern Mediterranean and Middle East, analysed from a palaeontological database. *Palaeogeogr., Palaeoclimatol., Palaeoecol.* **162**, 75-104 (2000).

71. E. Shoguchi, G. Beedessee, I. Tada, K. Hisata, T. Kawashima, T. Takeuchi, N. Arakaki, M. Fujie, R. Koyanagi, M. C. Roy, M. Kawachi, M. Hidaka, N. Satoh, C. Shinzato, Two divergent *Symbiodinium genomes* reveal conservation of a gene cluster for sunscreen biosynthesis and recently lost genes. *BMC Genomics* **19**, 458 (2018).

72. R. R. L. Guillard, J. H. Ryther, Studies of marine planktonic diatoms: I. *Cyclotella nana* hustedt, and *Detonula confervacea* (Cleve) Gran. *Can. J. Microbiol.* **8**, 229-239 (1962).

73. R. R. L. Guillard, in *Culture of Marine Invertebrate Animals: Proceedings — 1st Conference on Culture of Marine Invertebrate Animals Greenport,* W. L. Smith, M. H. Chanley, Eds. (Springer US, Boston, MA, 1975), pp. 29-60.

74. S. K. Rai, "Total High Molecular Weight DNA extraction from plant tissues for long read sequencing" (2022); https://dx.doi.org/10.17504/protocols.io.b5qyq5xw

992 75. A. Acosta-Maspons, I. González-Lemes, A. A. Covarrubias, Improved protocol for
993 isolation of high-quality total RNA from different organs of *Phaseolus vulgaris* L.
994 *BioTechniques* **66**, 96-98 (2019).

995 76. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X.
996 Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A.
997 Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N.
998 Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without
999 a reference genome. *Nat. Biotechnol.* **29**, 644 (2011).

1000 77. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome
1001 alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**,
1002 907-915 (2019).

1003 78. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov,
1004 BUSCO: assessing genome assembly and annotation completeness with single-copy
1005 orthologs. *Bioinformatics* **31**, 3210-3212 (2015).

1006 79. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor.
1007 *Bioinformatics* **34**, i884-i890 (2018).

1008 80. G. Marçais, C. Kingsford, A fast, lock-free approach for efficient parallel counting of
1009 occurrences of *k*-mers. *Bioinformatics* **27**, 764-770 (2011).

1010 81. T. R. Ranallo-Benavidez, K. S. Jaron, M. C. Schatz, GenomeScope 2.0 and
1011 Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**,
1012 1432 (2020).

1013 82. P. J. Rizzo, M. Jones, S. M. Ray, Isolation and properties of isolated nuclei from the
1014 Florida red tide dinoflagellate *Gymnodinium breve* (Davis). *J. Protozool.* **29**, 217-222
1015 (1982).

1016 83. J. Sano, K. H. Kato, Localization and copy number of the protein-coding genes actin,
1017 α-Tubulin, and HSP90 in the nucleus of a primitive dinoflagellate, *Oxyrrhis marina*.
1018 *Zool. Sci.* **26**, 745-753 (2009).

1019 84. G. S. Kohli, U. John, R. I. Figueroa, L. L. Rhodes, D. T. Harwood, M. Groth, C. J. S.
1020 Bolch, S. A. Murray, Polyketide synthesis genes associated with toxin production in
1021 two species of *Gambierdiscus* (Dinophyceae). *BMC Genomics* **16**, 410 (2015).

1022 85. A. V. Zimin, G. Marçais, D. Puiu, M. Roberts, S. L. Salzberg, J. A. Yorke, The
1023 MaSuRCA genome assembler. *Bioinformatics* **29**, 2669-2677 (2013).

1024 86. W. Wang, A. Das, D. Kainer, M. Schalamun, A. Morales-Suarez, B. Schwessinger, R.
1025 Lanfear, The draft nuclear genome assembly of *Eucalyptus pauciflora*: a pipeline for
1026 comparing *de novo* assemblies. *GigaScience* **9**, giz160 (2020).

1027 87. N. I. Weisenfeld, V. Kumar, P. Shah, D. M. Church, D. B. Jaffe, Direct determination
1028 of diploid genome sequences. *Genome Res.* **27**, 757-767 (2017).

1029 88. W. Xue, J.-T. Li, Y.-P. Zhu, G.-Y. Hou, X.-F. Kong, Y.-Y. Kuang, X.-W. Sun,
1030 L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC Genomics* **14**, 604
1031 (2013).

1032 89. M. Hiltunen, M. Ryberg, H. Johannesson, ARBitR: an overlap-aware genome
1033 assembly scaffolder for linked reads. *Bioinformatics* **37**, 2203-2205 (2020).

1034 90. L. K. Johnson, H. Alexander, C. T. Brown, Re-assembly, quality evaluation, and
1035 annotation of 678 microbial eukaryotic reference transcriptomes. *GigaScience* **8**,
1036 giy158 (2019).

1037 91. A. V. Zimin, D. Puiu, M.-C. Luo, T. Zhu, S. Koren, G. Marçais, J. A. Yorke, J.
1038 Dvořák, S. L. Salzberg, Hybrid assembly of the large and highly repetitive genome of
1039 *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads
1040 algorithm. *Genome Res.* **27**, 787-792 (2017).

1041 92. C. Iha, K. E. Dougan, J. A. Varela, V. Avila, C. J. Jackson, K. A. Bogaert, Y. Chen,
1042 L. M. Judd, R. Wick, K. E. Holt, M. M. Pasella, F. Ricci, S. I. Repetti, M. Medina, V.
1043 R. Marcelino, C. X. Chan, H. Verbruggen, Genomic adaptations to an endolithic
1044 lifestyle in the coral-associated alga *Ostreobium*. *Curr. Biol.* **31**, 1393-1402 (2021).

1045 93. G. Marçais, A. L. Delcher, A. M. Phillippy, R. Coston, S. L. Salzberg, A. Zimin,
1046 MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**,
1047 e1005944 (2018).

1048 94. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment
1049 search tool. *J. Mol. Biol.* **215**, 403-410 (1990).

1050 95. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing
1051 genomic features. *Bioinformatics* **26**, 841-842 (2010).

1052 96. T. Carver, S. R. Harris, M. Berriman, J. Parkhill, J. A. McQuillan, Artemis: an
1053 integrated platform for visualization and analysis of high-throughput sequence-based
1054 experimental data. *Bioinformatics* **28**, 464-469 (2011).

1055 97. J. M. Flynn, R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, A. F. Smit,
1056 RepeatModeler2 for automated genomic discovery of transposable element families.
1057 *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451 (2020).

1058 98. A. F. A. Smit, R. Hubley, P. Green, "RepeatMasker Open-4.0" (2017);
1059 https://www.repeatmasker.org.

1060 99. A. Lomsadze, V. Ter-Hovhannisyan, Y. O. Chernoff, M. Borodovsky, Gene
1061 identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids*
1062 *Res.* **33**, 6494-6506 (2005).

1063 100. B. L. Cantarel, I. Korf, S. M. C. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A.
1064 Sánchez Alvarado, M. Yandell, MAKER: an easy-to-use annotation pipeline designed
1065 for emerging model organism genomes. *Genome Res.* **18**, 188-196 (2008).

1066 101. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T. L.
1067 Madden, BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421
1068 (2009).

1069 102. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**,
1070 3094-3100 (2018).

1071 103. W. J. Kent, BLAT--the BLAST-like alignment tool. *Genome Res.* **12**, 656-664 (2002).

1072 104. B. J. Haas, A. L. Delcher, S. M. Mount, J. R. Wortman, R. K. Smith, Jr., L. I.
1073 Hannick, R. Maiti, C. M. Ronning, D. B. Rusch, C. D. Town, S. L. Salzberg, O.
1074 White, Improving the *Arabidopsis* genome annotation using maximal transcript
1075 alignment assemblies. *Nucleic Acids Res.* **31**, 5654-5666 (2003).

1076 105. M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: lightning-fast iterative
1077 protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173-175
1078 (2012).

1079 106. B. J. Haas, "TransposonPSI" (2010); https://sourceforge.net/projects/transposonpsi/.

1080 107. M. Mirdita, L. von den Driesch, C. Galiez, M. J. Martin, J. Söding, M. Steinegger,
1081 Uniclust databases of clustered and deeply annotated protein sequences and
1082 alignments. *Nucleic Acids Res.* **45**, D170-D176 (2017).

1083 108. W. Li, A. Godzik, CD-HIT: a fast program for clustering and comparing large sets of
1084 protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).

1085 109. I. Korf, Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).

1086 110. M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, B. Morgenstern, AUGUSTUS:
1087 ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435-W439
1088 (2006).

1089 111. B. J. Haas, S. L. Salzberg, W. Zhu, M. Pertea, J. E. Allen, J. Orvis, O. White, C. R.
1090 Buell, J. R. Wortman, Automated eukaryotic gene structure annotation using
1091 EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**,
1092 R7 (2008).

1093 112. M. A. Campbell, B. J. Haas, J. P. Hamilton, S. M. Mount, C. R. Buell,
1094 Comprehensive analysis of alternative splicing in rice and comparative analyses with
1095 *Arabidopsis*. *BMC Genomics* **7**, 327 (2006).

1096 113. B. J. Haas, in *Nuclear pre-mRNA Processing in Plants,* A. S. N. Reddy, M. Golovkin,
1097 Eds. (Springer Berlin Heidelberg, Berlin, Heidelberg, 2008), pp. 17-37.

1098 114. B. C. C. Hume, E. G. Smith, M. Ziegler, H. J. M. Warrington, J. A. Burt, T. C.
1099 LaJeunesse, J. Wiedenmann, C. R. Voolstra, SymPortal: A novel analytical
1100 framework and platform for coral algal symbiont next-generation sequencing ITS2
1101 profiling. *Mol. Ecol. Resour.* **19**, 1063-1080 (2019).

1102 115. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7:
1103 improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).

1104 116. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated
1105 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-
1106 1973 (2009).

1107 117. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: a fast and
1108 effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol.*
1109 *Biol. Evol.* **32**, 268-274 (2015).

1110 118. J. Xie, Y. Chen, G. Cai, R. Cai, Z. Hu, H. Wang, Tree Visualization By One Table
1111 (tvBOT): a web application for visualizing, modifying and annotating phylogenetic
1112 trees. *Nucleic Acids Res.*, doi:10.1093/nar/gkad1359 (2023).

1113 119. G. Beedessee, T. Kubota, A. Arimoto, K. Nishitsuji, R. F. Waller, K. Hisata, S.
1114 Yamasaki, N. Satoh, J. i. Kobayashi, E. Shoguchi, Integrated omics unveil the
1115 secondary metabolic landscape of a basal dinoflagellate. *BMC Biol.* **18**, 139 (2020).

1116 120. U. John, Y. Lu, S. Wohlrab, M. Groth, J. Janouškovec, G. S. Kohli, F. C. Mark, U.
1117 Bickmeyer, S. Farhat, M. Felder, S. Frickenhaus, L. Guillou, P. J. Keeling, A.
1118 Moustafa, B. M. Porcel, K. Valentin, G. Glöckner, An aerobic eukaryotic parasite
1119 with functional mitochondria that likely lacks a mitochondrial genome. *Sci. Adv.* **5**,
1120 eaav1110 (2019).

1121 121. D. M. Emms, S. Kelly, OrthoFinder: phylogenetic orthology inference for
1122 comparative genomics. *Genome Biol.* **20**, 238 (2019).

1123 122. P. Greenfield, U. Roehm, Answering biological questions by querying *k*-mer
1124 databases. *Concurr. Comput. Pract. Exper.* **25**, 497-509 (2013).

1125 123. G. Reinert, D. Chew, F. Sun, M. S. Waterman, Alignment-free sequence comparison
1126 (I): statistics and power. *J. Comput. Biol.* **16**, 1615-1634 (2009).

124. C. X. Chan, G. Bernard, O. Poirion, J. M. Hogan, M. A. Ragan, Inferring phylogenies of evolving sequences without multiple sequence alignment. *Sci. Rep.* **4**, 6504 (2014).

125. J. Felsenstein, "PHYLIP (Phylogeny Inference Package)" (2005); https://evolution.gs.washington.edu/phylip.html.

126. G. Bernard, C. X. Chan, M. A. Ragan, Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. *Sci. Rep.* **6**, 28970 (2016).

127. G. Bernard, P. Greenfiled, M. A. Ragan, C. X. Chan, M. J. Claesson, *k*-mer similarity, networks of microbial genomes, and taxonomic rank. *mSystems* **3**, e00257-00218 (2018).

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

129. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

130. G. Pertea, M. Pertea, GFF Utilities: GffRead and GffCompare [version 2; peer review: 3 approved]. *F1000Res.* **9**, 304 (2020).

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

132. W. Shen, S. Le, Y. Li, F. Hu, SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS ONE* **11**, e0163962 (2016).

133. M. Polne-Fuller, A novel technique for preparation of axenic cultures of *Symbiodinium* (Pyrrophyta) through selective digestion by amoebae. *J. Phycol.* **27**, 552-554 (1991).

134. F. H. Chang, Winter phytoplankton and microzooplankton populations off the coast of Westland, New Zealand, 1979. *N. Z. J. Mar. Freshwat. Res.* **17**, 279-304 (1983).

135. A. M. Quattrini, E. Rodríguez, B. C. Faircloth, P. F. Cowman, M. R. Brugler, G. A. Farfan, M. E. Hellberg, M. V. Kitahara, C. L. Morrison, D. A. Paz-García, J. D. Reimer, C. S. McFadden, Palaeoclimate ocean conditions shaped the evolution of corals and their skeletons through deep time. *Nat. Ecol. Evol.* **4**, 1531-1538 (2020).

136. E. L. Grossman, M. M. Joachimski, Ocean temperatures through the Phanerozoic reassessed. *Sci. Rep.* **12**, 8938 (2022).

Acknowledgements

We are grateful to Todd LaJeunesse and Hannah Reich who generously supplied the cell cultures of two *Effrenium voratum* strains (RCC1521 and rt-383) used in this study. 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.

Funding

This research was supported by the University of Queensland Research Training Program 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).

Author contributions

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

Competing interests

The authors declare that they have no competing interests.

1184 **Data and materials availability**

1185 All sequencing data generated from this study are available on NCBI GenBank *via*
1186 BioProject accession PRJEB61191. The assembled genome, predicted gene models and
1187 proteins, and the identified organellar genome sequences are available at
1188 https://doi.org/10.48610/1f0377a.

1189 **Table**

1190 **Table 1. Genome assemblies and gene predictions of *E. voratum* RCC1521, rt-383, and**
1191 **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	(*29*)	(*133*)	(*134*)
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

1192

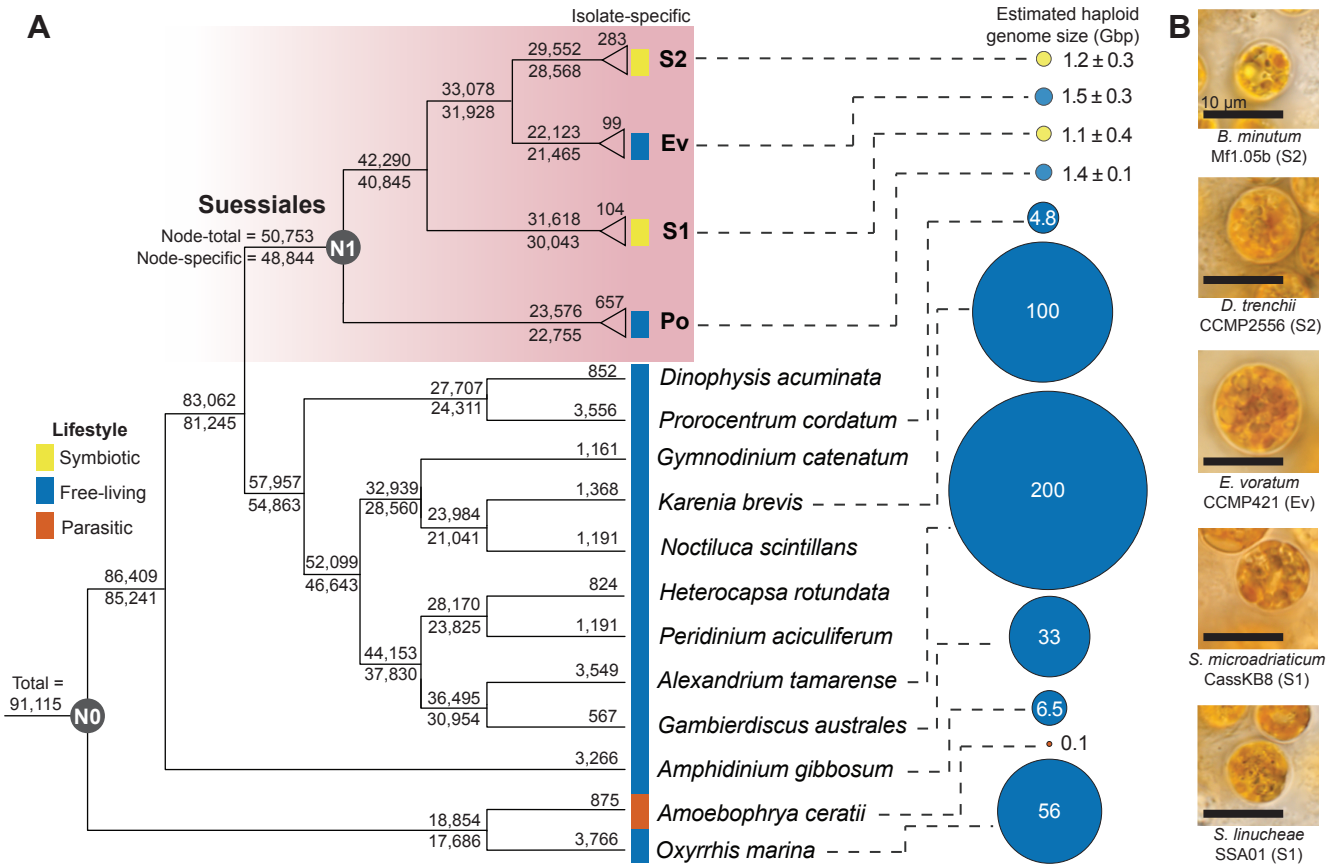

Figure 1. Species tree of dinoflagellates alongside estimated genome sizes. (**A**) The species tree was inferred using 91,115 orthologous protein sets derived from 1,420,328 protein sequences from 33 dinoflagellate taxa. Symbiotic lifestyles are shown in yellow, free-living in blue, and parasitic in 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. (**B**) Micrographs of representative taxa from the Symbiodiniaceae groups S1, S2, and Ev. Scale bar = 10 µm.

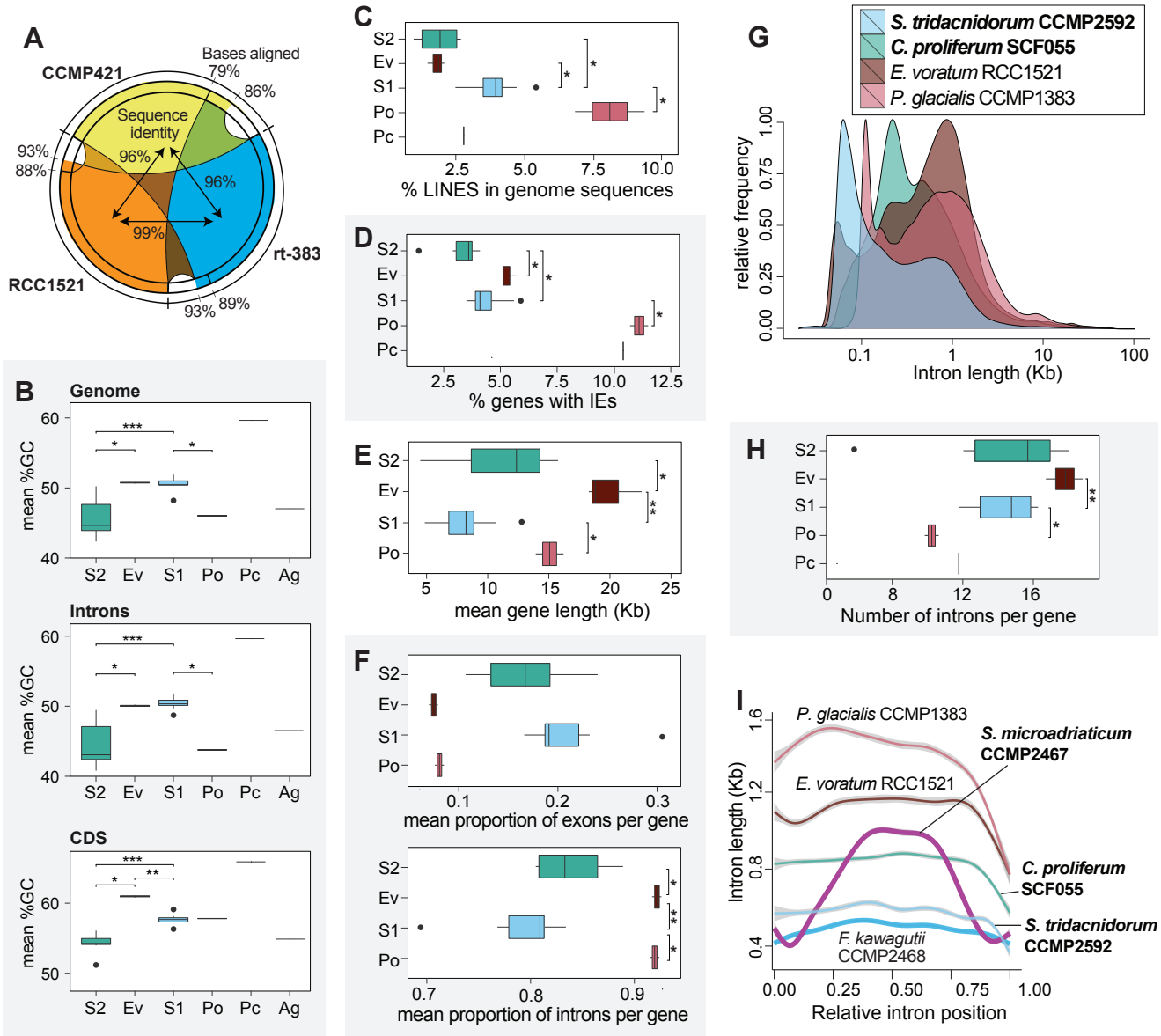

Figure 2. Genome features of *E. voratum* and other dinoflagellates. (**A**) Genome-sequence identity (italics within circle) and the percentage of aligned bases (outside circle) among the three *E. voratum* isolates. Features of representative genomes in the four Suessiales groups of S2, Ev, S1, and Po, plus *Prorocentrum cordatum* (Pc), in the order from the most-recent to most-ancient divergence, shown for (**B**) mean GC content in whole-genome, intronic, and CDS regions (*Amphidinium gibbosum* (Ag) was added for dinoflagellate-wide comparison), (**C**) percentage of mobile elements, and (**D**) percentage of genes containing introner elements (IEs). Gene features of Suessiales showing (**E**) mean gene length, (**F**) average proportions of exons and introns, (**G**) the relative frequency of introns by length (symbiotic lineages in boldface), (**H**) number of introns per gene, and (**I**) intron length versus relative intron position (symbiotic lineages in boldface). In all bar charts, *, **, *** represent $p < 0.05$, < 0.01, and < 0.001 respectively based on Wilcoxon rank sum test.

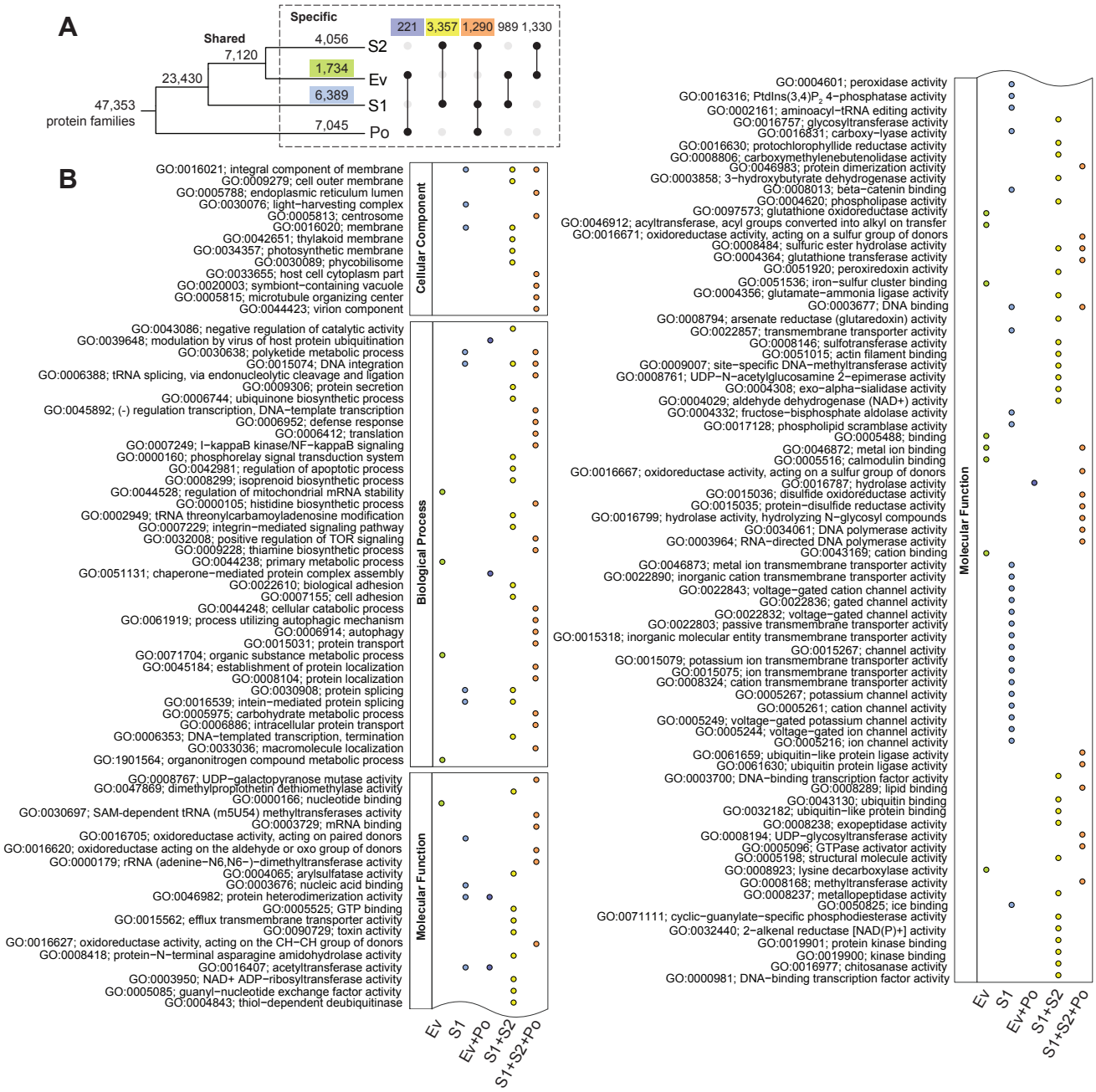

Figure 3. Gene family evolution in Suessiales. (**A**) Number of protein families is shown at each node and branch represents those that are shared among or specific to S1, Ev, S2, and/or Po. Number of families that are exclusive to Ev (green), to S1 (light blue), to Ev+Po (dark blue), to S2+S1 (yellow), 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.

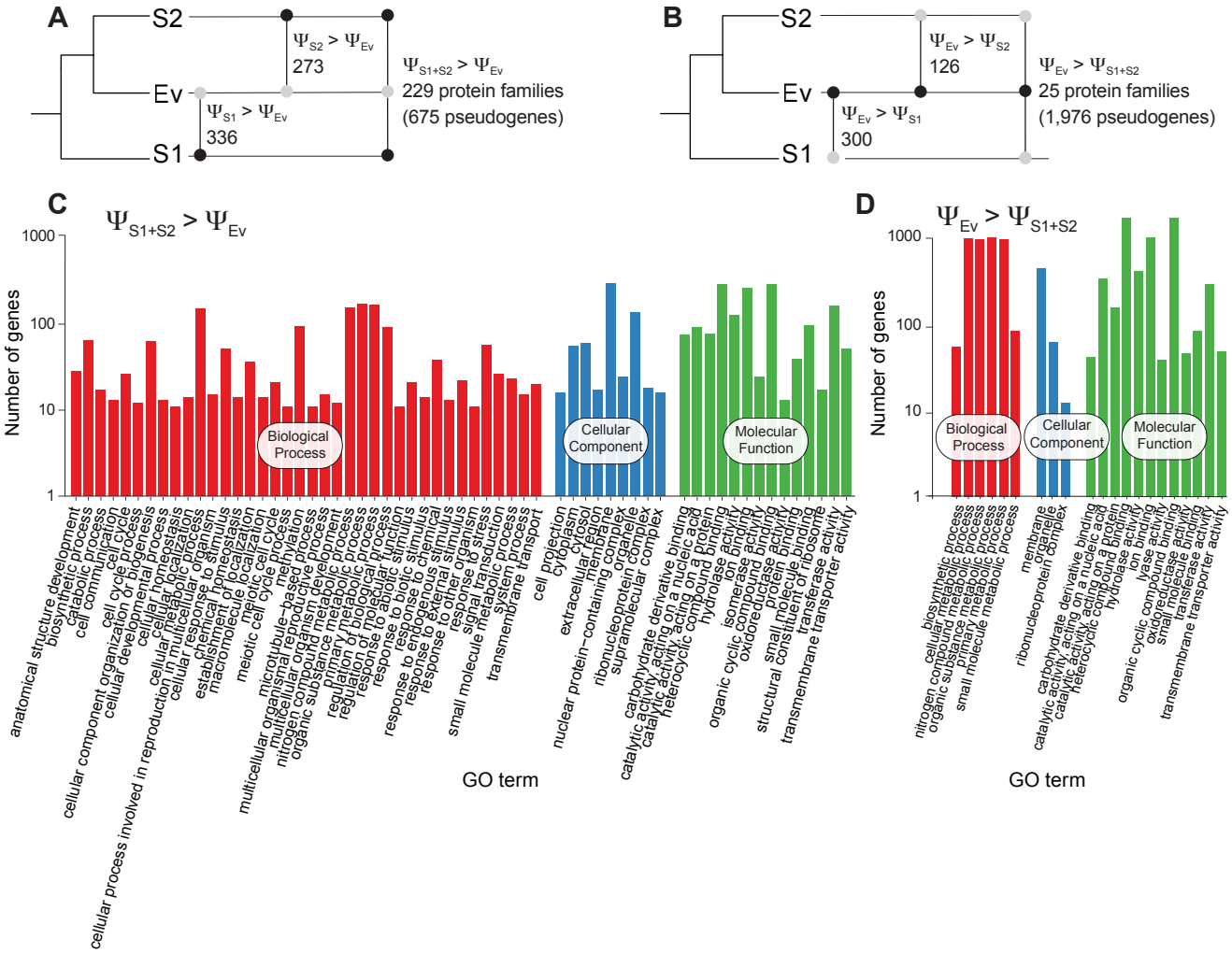

Figure 4. Pseudogenisation in Symbiodiniaceae. Number of protein families with significantly different level of pseudogenisation, Ψ, shown for (**A**) those with greater extent in symbiotic lineages (i.e., higher Ψ in S1, S2, or S1+S2, relative to Ev), and (**B**) those with greater extent in the free-living 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 $\Psi_{Ev} > \Psi_{S1+S2}$.

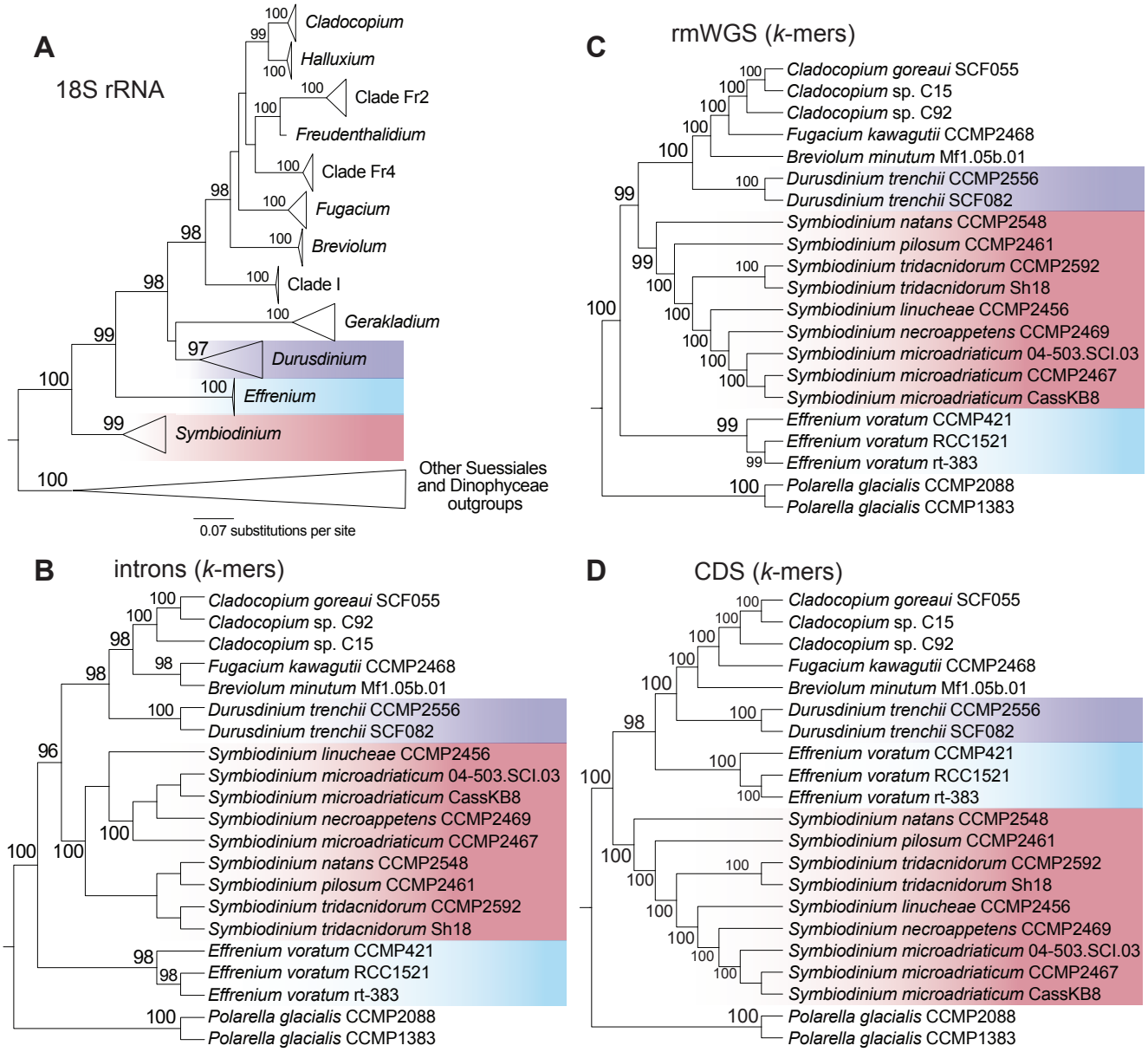

Figure 5. Phylogenetic relationship of *E. voratum* within Symbiodiniaceae and other taxa. (**A**) Maximum likelihood species tree of dinoflagellates inferred based on alignment of 18S rDNA sequences from 13 representative taxa, showing node support derived from ultrafast bootstrap of 2,000 samples; 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.

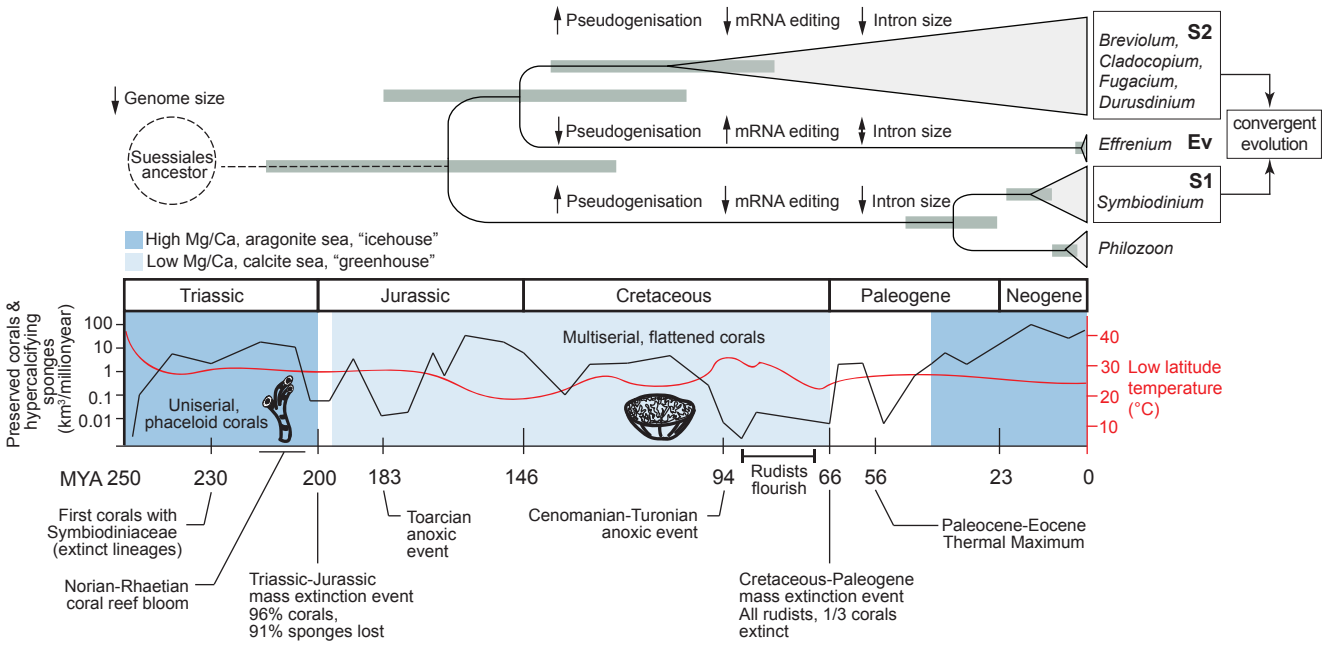

Figure 6. Timeline of Symbiodiniaceae genome evolution and coral evolution. The estimated divergence timeline of the Family Symbiodiniaceae (*1*) is shown at the top, indicating representative taxa of S1, Ev, and S2. Grey bars represent 95% confidence intervals of divergence times. Key genome signatures 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 (*1*, *2*, *13-15*, *135*, *136*).