

1 **Intraspecies genomic divergence of coral algal symbionts shaped**

2 **by gene duplication**

3 Sarah Shah¹, Katherine E. Dougan¹, Yibi Chen¹, Debasish Bhattacharya², Cheong Xin
4 Chan¹

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

7 ²Rutgers University, Department of Biochemistry and Microbiology, New Brunswick, NJ,
8 08901, USA

9 **Abstract**

10 Dinoflagellates of Order Suessiales include the diverse Family Symbiodiniaceae known for
11 their role as essential coral reef symbionts, and the cold-adapted *Polarella glacialis*. These
12 taxa inhabit a broad range of ecological niches and exhibit extensive genomic divergence,
13 although their genomes are in the smaller size ranges (haploid size < 3 Gbp) compared to
14 most other dinoflagellates. Different isolates of a species are known to form symbiosis with
15 distinct hosts and exhibit different regimes of gene expression, but intraspecies whole-
16 genome divergence remains little known. Focusing on three Symbiodiniaceae species (the
17 free-living *Effrenium voratum*, and the symbiotic *Symbiodinium microadriaticum* and
18 *Durusdinium trenchii*) and the free-living outgroup *P. glacialis*, all for which whole-genome
19 data from multiple isolates are available, we assessed intraspecies genomic divergence at
20 sequence and structural levels. Our analysis based on alignment and alignment-free methods
21 revealed greater extent of intraspecies sequence divergence in symbiodiniacean species than
22 in *P. glacialis*. Our results also reveal the implications of gene duplication in generating
23 functional innovation and diversification of Symbiodiniaceae, particularly in *D. trenchii* for
24 which whole-genome duplication was involved. Interestingly, tandem duplication of single-
25 exon genes was found to be more prevalent in genomes of free-living species than in those of
26 symbiotic species. These results in combination demonstrate the remarkable intraspecies
27 genomic divergence in dinoflagellates under the constraint of reduced genome sizes, shaped
28 by genetic duplications and symbiogenesis events during diversification of Symbiodiniaceae.

29 Introduction

30 Dinoflagellates of the Order Suessiales include the Family Symbiodiniaceae, which
31 predominantly consists of symbiotic lineages essential to coral reef organisms.
32 Symbiodiniaceae taxa collectively exhibit a broad spectrum of symbiotic associations (i.e.,
33 facultativeness) and variable degrees of host specificity (i.e., host-specialist vs host-
34 generalist), although some are described as solely free-living (Thornhill et al. 2014;
35 LaJeunesse et al. 2018). A comparative analysis of whole-genome sequences from 15 taxa
36 revealed extensive sequence and structural divergence among Symbiodiniaceae taxa, which
37 was more prevalent in isolates of the symbiotic species, *Symbiodinium microadriaticum*
38 (González-Pech et al. 2021). This was supported by a metagenomics survey of single-
39 nucleotide polymorphisms in the genomes of symbiotic *Symbiodinium fitti* from different
40 coral taxa and biogeographical origins, revealing intraspecies sequence divergence correlated
41 to coral host taxa (Reich et al. 2021).

42 A recent comparative genomic analysis incorporating genomes from three isolates of
43 the free-living species *E. voratum* revealed genome features representative of the
44 Symbiodiniaceae progenitor, due to the absence of symbiogenesis in the *Effrenium* lineage
45 (Shah et al. 2023). These features include longer introns, more extensive RNA editing, less
46 pseudogenisation, and, perhaps most surprisingly, similar genome sizes when compared to
47 symbiotic counterparts. The genome size of *E. voratum* suggests that genome reduction (to
48 haploid genome size < 3Gbp) occurred in symbiodiniacean dinoflagellates before
49 diversification of Order Suessiales (Shah et al. 2023). These results further hint at a role of
50 symbiotic lifestyle in shaping intraspecies genomic divergence and the evolution of these
51 taxa. Intrageneric variation of the ITS2 phylogenetic marker sequences is known among
52 Symbiodiniaceae taxa (Wilkinson et al. 2015; Hume et al. 2019). However, intraspecies
53 whole-genome divergence in these taxa relative to symbiotic versus free-living lifestyle
54 remains little known. Whole-genome data from multiple isolates of a species provide an
55 excellent analysis platform to address this knowledge gap.

56 Here, we investigate intraspecies genomic divergence in four Suessiales species (of
57 which three are Symbiodiniaceae); these taxa represent two free-living species and two
58 symbiotic species, for which whole-genome data from multiple isolates are available. We
59 focus specifically on sequence and structural conservation, gene family dynamics, and gene
60 duplication, and how these features may reflect adaptation to the distinct lifestyles.

61 Results and Discussion

62 We used four Suessiales species for which multi-isolate genome data are publicly available,
63 to investigate patterns of intraspecies genomic divergence related to facultative lifestyle. The
64 two symbiotic symbiodiniacean species, *S. microadriaticum* (González-Pech et al. 2021;
65 Nand et al. 2021) and *Durusdinium trenchii* (Dougan et al. 2022a), represent taxa that arose
66 from independent origins of symbiogenesis (Figure 1 and Supplementary Table S1). The
67 remaining two are free-living species, the symbiodiniacean *E. voratum* (Shah et al. 2023) and
68 *Polarella glacialis* that is sister to the Symbiodiniaceae in the Order Suessiales (Stephens et
69 al. 2020). The available genome data were generated from isolates collected over vast
70 geographic areas: the thermotolerant symbiont *D. trenchii* from the Caribbean Sea and
71 Pacific Ocean, the free-living *E. voratum* from the Mediterranean Sea and both sides of the
72 Pacific Ocean, the symbiotic *S. microadriaticum* from the Red Sea, Pacific Ocean, and the
73 Caribbean Sea, and the psychrophilic *P. glacialis* from the Antarctic and Arctic oceans
74 (Figure 1). Collectively, these data provide a robust analytic framework for interrogating
75 intraspecies genomic divergence.

76

77 Genomes of facultative symbionts exhibit higher sequence divergence

78 We investigated divergence of genome sequence following the approach of González-Pech et
79 al. (2021). For each pairwise comparison of genome sequences, we calculated the percentage
80 of aligned bases, Q , and overall sequence identity of aligned regions, ID . Genome sequences
81 from isolates of the same species are highly similar ($Q > 70.2\%$, $ID > 98.6\%$ with minimum
82 alignment length 100 bp; Figure 2A, see Supplementary Figure S1 for detail), compared to
83 those between species ($Q < 10.0\%$, $ID < 98.6\%$). High intraspecies sequence similarity was
84 observed despite the diverse geographic origins for isolates from each species (Figure 1).
85 Genome sequences of the free-living *P. glacialis* were the most similar ($Q = 95.5\%$, $ID =$
86 98.7%; CCMP1383 against CCMP2088), followed by the symbiotic *D. trenchii* ($Q = 93.3\%$,
87 $ID = 99.8\%$; CCMP2556 against SCF082), the free-living *E. voratum* ($Q = 92.0\%$, $ID = 99.4\%$;
88 RCC1521 against rt-383), and the symbiotic *S. microadriaticum* ($Q = 78.5\%$, $ID = 99.7\%$;
89 CCMP2467 against CassKB8). Among the three *E. voratum* isolates, CCMP421 showed
90 smaller percentage of aligned genome bases against rt-383 ($Q = 70.2\%$) and against
91 RCC1521 ($Q = 79.2\%$), compared to $Q = 92.0\%$ observed between RCC1521 and rt383; this
92 is likely due to the more-fragmented CCMP421 genome assembly, also reflected in the low

93 percentage of mapped sequence reads (Supplementary Table S2). Between the two symbiotic
94 species, the greater divergence observed in *S. microadriaticum* might represent its much
95 earlier emergence and diversification (LaJeunesse et al. 2018). Alternatively, the lower
96 divergence in *D. trenchii* may be due to the recent whole-genome duplication (WGD) in this
97 lineage (Dougan et al. 2022a). Genome data of multiple isolates from a broader taxon
98 representation of Symbiodiniaceae lineages will help clarify the possible link between
99 intraspecies divergence and facultative lifestyle of these symbionts.

100 To extend genome comparisons beyond alignable sequence regions, we further
101 assessed sequence divergence using an alignment-free k -mer-based approach. This approach
102 was found to be robust against the contiguity of genome assemblies (Dougan et al. 2022c),
103 and has been applied successfully to discover distinct phylogenetic signals in different
104 genomic regions of Symbiodiniaceae (Lo et al. 2022; Shah et al. 2023). We followed Lo et al.
105 (2022) to derive pairwise D_2^S distances, d , based on shared k -mer profiles at $k = 23$ observed
106 in whole-genome sequences (see Methods). As shown in Figure 2B, the lowest sequence
107 divergence was seen in *P. glacialis* ($d = 0.30$), followed by *E. voratum* ($d = 0.53$ between
108 RCC1521 and rt-383; $d = 0.9$ when implicating the more-fragmented CCMP421 assembly),
109 *D. trenchii* (0.54), and the three *S. microadriaticum* isolates (0.72-0.76). This pattern of
110 divergence is consistent with our observations based on Q and ID in Figure 2A.

111 We further assessed the conserved core 23-mers in each species (i.e., k -mers common
112 in genomes of all isolates within a species). For each species, we assessed the extent of
113 genome content shared among the isolates based on x , the percentage of core 23-mers relative
114 to all distinct 23-mers; in the perfect scenario where genomes of all isolates are identical, $x =$
115 100%. Using this approach, *E. voratum* and *S. microadriaticum* show similar extent of shared
116 genome content among their corresponding isolates (x ranges between 19.5% and 25.2%;
117 Supplementary Table S3). Approximately two-fold greater x was observed for *P. glacialis*
118 (52.3-54.9%) and *D. trenchii* (55.6-55.7%); this observation likely reflects the impact of a
119 diploid genome assembly in the former (Stephens et al. 2020) and WGD in the latter (Dougan
120 et al. 2022a). Duplicated genomic regions arising from WGD are resolved over long
121 evolutionary time scales of hundreds of millions of years (Carretero-Paulet and Van de Peer
122 2020). Given the recent (~1 MYA) WGD in *D. trenchii*, this species likely has not had
123 sufficient time to resolve genetic redundancy. Regardless, our results here lend support to the
124 general utility of k -mer-derived distances in clarifying genome-sequence divergence beyond
125 gene boundaries, which may serve as evidence to guide or complement taxonomic

126 classification of Symbiodiniaceae, and potentially of other dinoflagellates (Dougan et al.
127 2022c).

128

129 **Intraspecies structural divergence in the genomes of Symbiodiniaceae**

130 To assess intraspecies structural genomic divergence, we identified collinear gene blocks in
131 all possible pairwise genome comparisons for each species (see Methods); the greater
132 recovery of these blocks and their implicated genes indicates a greater conserved synteny
133 among the isolates in a species. As expected, due to recent WGD, the two symbiotic *D.*
134 *trenchii* isolates CCMP2556 and SCF082 displayed the greatest conserved synteny (1,613
135 blocks implicating ~22% of total genes spanning 181-199 Mbp; Supplementary Table S4).
136 On the other hand, genomes of the symbiotic *S. microadriaticum* (100-196 blocks, 2.7-3.6%
137 of genes, 8.1-16 Mbp) showed less conserved synteny than the free-living *E. voratum*
138 RCC1521 and rt383 (344 blocks, 6.6-8.1% of genes, 51-60 Mbp; Supplementary Table S4);
139 at first glance this result appears to support observations in an earlier study (González-Pech et
140 al. 2021) that the extent of structural rearrangements is greater in genomes of facultative
141 symbionts than those of free-living taxa. However, the greater contiguity of the *E. voratum*
142 assemblies (scaffold N50 length = 720 Kbp for RCC1521, 252 Kbp for rt-383) than that of *S.*
143 *microadriaticum* assemblies (e.g., scaffold N50 length = 43 Kbp for CassKB8 and 50 Kbp for
144 04-503SCI.03) represents a systematic bias that would affect recovery of collinear gene
145 blocks. *S. microadriaticum* CCMP2467 (N50 length 9.96 Mbp) (Supplementary Table S1),
146 the sole representation of a chromosome-level assembly, lacks comparative power in this
147 instance. As a case in point, the inclusion of the fragmented assembly of *E. voratum*
148 CCMP421 (N50 length 304 Kbp; 38,022 scaffolds) lowers the extent of conserved synteny
149 identified in *E. voratum* (195-331 blocks, 4.4-7.9% of genes, 30-65 Mbp; Supplementary
150 Table S4), and we identified no collinear gene blocks between the outgroup *P. glacialis*
151 isolates due in part to sparsity of genes on the assembled genome scaffolds (Stephens et al.
152 2020). These results in combination suggest that while structural rearrangements contribute to
153 structural divergence of Symbiodiniaceae genomes as postulated in those of facultative
154 symbionts (González-Pech et al. 2019) even at intraspecies level, such an analysis based on
155 collinear gene blocks is sensitive to contiguity of assembled genome sequences. An in-depth
156 assessment of structural divergence would require genome assemblies of comparably high
157 quality.

158

159 **Genetic duplication enables functional innovation**

160 We assessed the evolution of protein families for evidence of functional innovation and
161 divergence within species, and its relation to lifestyle. For each species, we inferred
162 homologous protein sets with OrthoFinder using sequences predicted from all corresponding
163 isolates (see Methods); the homologous sets that are specific to an isolate may reflect
164 instances of contrasting divergence in and/or specialisation of protein functions (e.g., putative
165 remote homologs), occurring at distinct evolutionary rates. First, we assessed number of
166 isolate-specific sets for each species based on OrthoFinder results ran at default parameters
167 (i.e., inflation parameter $I = 1.5$). The highest percentage of isolate-specific sets was observed
168 in *D. trenchii* (17.2% of total sets), followed by *P. glacialis* (16.0%); these numbers are
169 nearly four-fold greater than that observed in *S. microadriaticum* (4.0%) and *E. voratum*
170 (4.1%; Figure 3). To investigate the robustness of this result, we increased the inflation
171 parameter (I) for clustering within OrthoFinder that controls the granularity (i.e., higher
172 inflation parameter produces smaller clusters). As expected in all cases, the increase of I
173 resulted in an increase of isolate-specific protein sets; at $I = 10$, the percentage of these sets is
174 37.8% (*D. trenchii*), 32.4% (*P. glacialis*), 15.6% (*S. microadriaticum*), and 10.8% (*E.*
175 *voratum*). Despite the high synteny and sequence conservation in *D. trenchii*, the substantial
176 number of protein families retained in duplicate after WGD show evidence of isolate-specific
177 divergence and/or specialization in *D. trenchii* where facultative lifestyle has been
178 hypothesized to be the main driver of post-WGD adaptation (Dougan et al. 2022a). On the
179 other hand, the comparable extent of isolate-specific protein sets in *P. glacialis* may represent
180 heterozygosity inherent to a diploid representation of the genome assembly (Stephens et al.
181 2020), distinct from the haploid genome assemblies among the Symbiodiniaceae taxa. None
182 of the *E. voratum* and *S. microadriaticum* isolates showed evidence of WGD (Supplementary
183 Table S5), and thus the similar level of isolate-specific divergence in these species supports
184 the notion of massive genome reduction in the Suessiales ancestor, with WGD a mechanism
185 for escaping this process to generate functional innovation, as observed in *D. trenchii*
186 (Dougan et al. 2022a).

187

188 **Genomes of free-living species exhibit greater extent of tandemly duplicated single-exon
189 genes**

190 Tandemly duplicated (TD) genes, i.e., duplicated genes found next to each other on the
191 genome, are part of unidirectional gene clusters commonly found in dinoflagellates, thought

192 to facilitate their expression (Nand et al. 2021; Chen et al. 2022). In an earlier study
193 (Stephens et al. 2020), ~40% of the gene repertoire in *P. glacialis* genomes were located in
194 unidirectional gene clusters, many of which encoded functions associated with cold and low-
195 light adaptation. Here we defined a TD block as a block comprising two or more consecutive
196 genes with high sequence identity on a genome scaffold (see Methods). In our independent
197 survey of TD genes in all 19 available Suessiales genomes, we found the largest number and
198 proportion of TD genes in the free-living lineages of *P. glacialis* (7.8% in CCMP1383, 9.2%
199 in CCMP2088) and *S. natans* (7.1%), followed by the symbiotic *S. tridacnidorum*
200 CCMP2592 (6.5%) and *C. goreau* SCF055 (6.0%), with smaller proportions observed in the
201 free-living *E. voratum* (3.9% in rt-383, 4.4% in RCC1521), and the smallest in *S.*
202 *microadriaticum* (1.0-2.2%) (Table 1). Some of the largest TD blocks consisted of 13-16
203 genes, found in genomes of free-living lineages (*S. natans*, and the *P. glacialis* CCMP1383
204 and CCMP2088). Among the free-living *E. voratum* isolates, the TD block sizes were slightly
205 smaller, implicating genes encoding ribulose bisphosphate carboxylase (the largest block of 9
206 genes in RCC1521), HECT and RLD domain-containing E3 ubiquitin protein ligase 4 (rt-
207 383, 7 genes), calmodulin (rt-383, 7 genes), and solute carrier family 4 (rt-383, 7 genes)
208 (Supplementary Table S6); these implicated functions are essential for photosynthesis, ion
209 binding, and transmembrane transport. However, we cannot dismiss the possibility of
210 genome-assembly contiguity in affecting recovery of TD blocks. For instance, the recovery of
211 TD genes in the chromosome-level assembly of *S. microadriaticum* CCMP2467 is 2.2%
212 versus ~1.0% in the other two assemblies, and the recovery of 1.5% in *E. voratum* CCMP421
213 contrasts to 3.9-4.4% in the other two *E. voratum* genomes. Despite this, a greater extent of
214 TD genes in free-living lineages (*P. glacialis*: 55.2-59.4%; *E. voratum* RCC1521: 23.1% and
215 rt-383: 22.5%; *S. natans*: 21.8%) were single-exon genes, in contrast to the symbiotic *D.*
216 *trenchii* and *S. microadriaticum* (4.2-9.2%) (Table 1). Our results lend support to the notion
217 that tandem duplication may facilitate transcription of genes encoding essential functions
218 implicating single-exon genes, and is potentially more prominent in genomes of free-living
219 taxa than those of symbiotic lineages (Stephens et al. 2020).

220 Introner elements (IE) are non-autonomous mobile elements characterised by inverted
221 repeat motifs within introns that are hypothesised to propagate introns into genes (Worden et
222 al. 2009; van der Burgt et al. 2012; Huff et al. 2016), which have been found to be more
223 prevalent in genomes of free-living dinoflagellate species (Farhat et al. 2021; Dougan et al.
224 2022b; Shah et al. 2023). We examined the presence of these elements in the assembled
225 genomes and TD genes for the multi-isolate Suessiales species (Supplementary Table 1). We

226 found the proportion of IE-containing genes overall to be less in Symbiodiniaceae (3.2-6.3%)
227 than *P. glacialis* (10.7-11.5%), a trend also observed in the genome of bloom-forming
228 dinoflagellate species, *Prorocentrum cordatum* (10.4%) (Dougan et al. 2022b). Nonetheless,
229 IEs were only found in a small proportion of TD genes (2.5-5.7%) per Suessiales isolate,
230 suggesting they are neither connected to lifestyle nor play a major role in propagating TD
231 genes in Suessiales (Supplementary Table S1).

232

233 **Most tandemly duplicated genes undergo purifying selection**

234 To assess selection acting on TD genes, we focused on the two best-quality genome
235 assemblies (based on number of scaffolds and N50 length) from each species (i.e., total of
236 eight isolates), excluding the fragmented assemblies of *E. voratum* CCMP421 and *S.*
237 *microadriaticum* CassKB8. We calculated the ratio ω as the nonsynonymous substitution rate
238 (K_a) to synonymous substitution rate (K_s) between all possible gene pairs within each TD
239 block (Supplementary Table S6; see Methods); in general, $\omega > 1.0$ indicates positive
240 selection, $\omega = 1.0$ indicates neutral selection, whereas $\omega < 1.0$ indicates purifying selection
241 (Yang and Bielawski 2000) among TD genes within a block. Based on this analysis,
242 compared to genomes of symbiotic species, those of free-living species yielded larger
243 proportions of TD blocks with mean $\omega < 1.0$, indicating purifying selection, i.e., 71.7% in *P.*
244 *glacialis* and 67.7% in *E. voratum*, compared to 64.2% in *D. trenchii* and 49.1% in *S.*
245 *microadriaticum* (Figure 4A; Supplementary Table S7). In all cases, the mean K_s value per
246 TD block is less than 0.5 (Figure 4B). The observed mean ω values are similar between two
247 isolates of a species, e.g., mean variance of $\omega = 0.26$ for both *P. glacialis* isolates
248 (Supplementary Figure S2), suggesting a common pattern of selective pressures acting on TD
249 genes for the species. An exception is the symbiotic *S. microadriaticum* (mean variance of ω
250 = 0.16 for 04-503SCI.03 and 0.95 for CCMP2467; Supplementary Figure S2), but more
251 genome data from other multi-isolate symbiotic species will enable the systematic
252 investigation of the possible links between selection acting on TD genes and lifestyles.

253 To assess functions encoded by TD genes, we focused on TD gene blocks that were
254 recovered in genomes of both isolates in one or more species. Functional annotation of these
255 gene blocks is shown in Figure 4C, and the mean ω value for the corresponding block is
256 shown in Figure 4D. Genes encoding calmodulin, sulfotransfer domain-containing proteins,
257 and disulfide-isomerase proteins were recovered in TD blocks in all eight isolates. Fructose-
258 bisphosphate aldolase, dinoflagellate viral nucleoproteins, and caltractin were recovered in at

259 least 7 of the 8 isolates. Genes in TD blocks recovered only in free-living *P. glacialis* and *E.*
260 *voratum* encode functions related to photosynthesis (i.e., photosystem I reaction centre
261 subunit III, chloroplast TIC 20-II protein, PS II complex 12 kDa extrinsic protein, and
262 peridinin-chlorophyll *a*-binding protein). In comparison, those in TD blocks found only in the
263 two symbiotic species encode for Nek1 protein that is involved in maintaining centrosomes,
264 and NaCP60E, a sodium channel protein. Most of these functions were encoded by no more
265 than 50 TD genes per isolate (Figure 4C) in which the mean ω per gene block was < 1
266 (Figure 4D). These results do not speak directly to the specificity of gene functions to tandem
267 duplication in the genomes we analysed, given that some gene copies may also occur
268 elsewhere in the genomes. However, our results suggest a tendency for TD genes within a
269 block to undergo purifying selection, regardless of lifestyle.

270

271 **Concluding remarks**

272 Our results, based on multi-isolate whole-genome data from representative species,
273 demonstrate how facultative lifestyle or the lack thereof has shaped the genome evolution of
274 Symbiodiniaceae dinoflagellates. Generation of genetic and functional diversity at the
275 intraspecies level implicates genetic duplication, including tandem duplication of genes. All
276 these evolutionary regimes are under the constraint of genome reduction that is hypothesised
277 to pre-date the diversification of Order Suessiales (Shah et al. 2023). Although our results
278 hint at the potential linkages of facultative lifestyles to some of the varying features observed
279 between free-living versus symbiotic species, whole-genome data from a broader taxonomic
280 representation (and from multiple isolates) will enable a more-systematic investigation to
281 establish these linkages.

282

283 **Methods**

284 **Data**

285 For this study, we used publicly available genome assemblies and gene models of *D. trenchii*
286 CCMP2556 and SCF082 (Dougan et al. 2022a), *E. voratum* isolates RCC1521, rt-383, and
287 CCMP421 (Shah et al. 2023), *S. microadriaticum* CCMP2467 (Nand et al. 2021), 04-
288 503SCI.03 and CassKB8 (González-Pech et al. 2021), and *P. glacialis* CCMP1383 and
289 CCMP2088 (Stephens et al. 2020) (Supplementary Table S1). To contrast the contiguity of

290 these genome assemblies, we obtained chromosome numbers from cytological observations
291 (Blank and Trench 1985; Jeong et al. 2014; Wham et al. 2017). For tandem gene duplication
292 analysis, we used genomic datasets from 9 more Symbiodiniaceae isolates (Supplementary
293 Table S1) generated in Chen et al. (2020; 2022), González-Pech et al. (2021), and Shoguchi
294 et al. (2013; 2018). To determine the intraspecific identity of the three *E. voratum* genome
295 datasets, we mapped the short-read gDNA of each isolate obtained from (Shah et al. 2023) to
296 each other using Bowtie2 v2.4.4 (Langmead and Salzberg 2012) with the *--very-fast*
297 algorithm.

298 **Assessment of genome-sequence similarity based on alignment**

299 To assess genome-sequence similarity of the four target species based on sequence
300 alignment, we used nucmer (*--mum*) implemented in MUMmer 4.0.0beta2 (Marçais et al.
301 2018) at minimum alignment lengths of 100 bp, 1 Kb, and 10 Kb to align assembled genome
302 sequences for every possible pair of isolates in each species. For each pairwise comparison,
303 we calculated the percentage of aligned bases, Q , and overall sequence identity of aligned
304 regions, ID . Maximum values of for both Q and ID at 100% indicate that two genome
305 sequences are identical. We then used mummerplot (*-f --layout*) and dnadiff to generate
306 figures and reports for these alignments.

307 **Assessment of genome-sequence similarity using an alignment-free approach**

308 Adopting the same approach described in Lo et al. (2022), we calculated D_2^S statistic based on
309 shared k -mers for each pair of genomes, from which a distance (d) was derived. Briefly,
310 Jellyfish v2.3.0 (Marçais and Kingsford 2011) was used to derive k -mers (at $k = 23$) from
311 each genome assembly, from which distances were calculated using *d2ssect*
312 (<https://github.com/bakeronit/d2ssect>) from all possible pairs of genomes. Following the
313 earlier studies (Lo et al. 2022; Shah et al. 2023), core 23-mers among isolates of each species
314 were identified from the extracted 23-mers, using the bash command *comm* (-12). BEDtools
315 (Quinlan and Hall 2010) *intersect* was used to find regions of overlap between the core k -
316 mers and different genomic features.

317 **Gene family evolution and introner element search**

318 To infer homologous protein sets among isolates for a species, all protein sequences predicted
319 from all isolates were used as input for OrthoFinder v2.5.4 (Emms and Kelly 2019). The
320 analysis was conducted at different inflation parameters ($I = 1.5, 2.0, 4.0, 6.0, 8.0$, or 10.0).

321 From the generated homologous protein sets, the proportion of isolate-specific sets was
322 identified. To identify introner elements, we used the introner element sequences identified in
323 Shah et al. (2023) from eight Suessiales isolates as a reference for Pattern Locator (Mrázeck
324 and Xie 2006) to search for inverted and direct repeat motifs within introns.

325 **Identification of collinear gene blocks and types of gene duplication**

326 To identify collinear gene blocks shared by isolates of a species, we first identified
327 homologous protein sequences using BLASTp (e-value < 10^{-5} , query or subject cover > 50%,
328 filtered for top five hits for each query). This output was used as input for MCScanX (Wang
329 et al. 2012) (-b 2) to search for collinear gene blocks between all possible pairs of isolates.
330 For *D. trenchii*, we filtered out duplicated genes (Dougan et al. 2022a) from the MCScanX
331 output by selecting gene pairs that were more similar to each other (i.e., low nonsynonymous
332 (K_a) + synonymous (K_s) substitution score), then chose gene blocks that still contained ≥ 5
333 genes. Gene Ontology (GO) terms were assigned to all gene sets via UniProt (version
334 2022_01) to GO (version December 2022) ID mapping on the UniProt website
335 (uniprot.org/id-mapping). The *duplicate_gene_classifier* implemented in MCScanX was used
336 to assess five distinct type of gene duplications: 1) singleton = not duplicated, 2) dispersed =
337 duplicated with > 10 genes in between, 3) proximal = duplicated with < 10 genes in between,
338 4) WGD = whole or segmental genome duplication inferred by anchor genes in collinear gene
339 blocks comprising at least 5 genes, 5) tandem = duplicated one after the other, i.e., two or
340 more consecutive genes on the same scaffold.

341 **Analysis of tandemly duplicated genes**

342 Tandemly duplicated (TD) genes were identified based on the results of MCScanX above.
343 For this analysis, we focused on two best-quality genome assemblies from each species, i.e.,
344 for a total of eight genomes. For each TD block, we calculated the nonsynonymous
345 substitution rate (K_a) and synonymous rate (K_s) between all possible pairs of genes within the
346 block, using the *add_ka_and_ks_to_collinearity.pl* script implemented in MCScanX (Wang
347 et al. 2012). The ratio ω was defined as K_a/K_s . When assessing mean ω for each TD block,
348 instances of infinity values, e.g., due to $K_s = 0$, were ignored.

349 **Competing interests**

350 Authors declare that they have no competing interests.

351 **Author contributions**

352 Conceptualization, SS, KED, DB and CXC; methodology, SS, KED, YC, and CXC; formal
353 analysis, SS, KED, and YC; investigation, SS, KED; writing—original draft preparation, SS;
354 writing—review and editing, SS, KED, DB, and CXC; visualisation, SS; supervision, KED,
355 DB, CXC; funding acquisition, DB and CXC. All authors have read and agreed to the
356 published version of the manuscript.

357 **Funding**

358 This research was supported by the University of Queensland Research Training Program
359 scholarship (SS and YC), and the Australian Research Council grant DP19012474 awarded to
360 CXC and DB. DB was also supported by NSF grant NSF-OCE 1756616 and a NIFA-USDA
361 Hatch grant (NJ01180).

362 **Acknowledgements**

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

368
369

370 References

371

372 Allen Coral Atlas. 2022. Imagery, maps and monitoring of the world's tropical coral reefs.
373 Available from: <https://doi.org/10.5281/zenodo.3833242>

374 Blank RJ, Trench RK. 1985. Speciation and symbiotic dinoflagellates. *Science*. 229:656-658.

375 Carretero-Palet L, Van de Peer Y. 2020. The evolutionary conundrum of whole-genome
376 duplication. *American Journal of Botany*. 107:1101-1105.

377 Chen Y, González-Pech RA, Stephens TG, Bhattacharya D, Chan CX. 2020. Evidence that
378 inconsistent gene prediction can mislead analysis of dinoflagellate genomes. *Journal of*
379 *Phycology*. 56:6-10.

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

383 Dougan KE, Bellantuono AJ, Kahlke T, Abbriano RM, Chen Y, Shah S, Granados-Cifuentes
384 C, van Oppen MJH, Bhattacharya D, Suggett DJ, et al. 2022a. Whole-genome duplication in
385 an algal symbiont serendipitously confers thermal tolerance to corals. *bioRxiv*.
386 2022.04.10.487810.

387 Dougan KE, Deng Z-L, Wöhlbrand L, Reuse C, Bunk B, Chen Y, Hartlich J, Hiller K, John
388 U, Kalvelage J, et al. 2022b. Multi-omics analysis reveals the molecular response to heat
389 stress in a “red tide” dinoflagellate. *bioRxiv*. 2022.07.25.501386.

390 Dougan KE, González-Pech RA, Stephens TG, Shah S, Chen Y, Ragan MA, Bhattacharya D,
391 Chan CX. 2022c. Genome-powered classification: insights gained from coral algal
392 symbionts. *Trends in Microbiology*. 30:831-840.

393 Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative
394 genomics. *Genome Biology*. 20:238.

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

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

400 González-Pech RA, Stephens TG, Chen Y, Mohamed AR, Cheng Y, Shah S, Dougan KE,
401 Fortuin MDA, Lagorce R, Burt DW, et al. 2021. Comparison of 15 dinoflagellate genomes

402 reveals extensive sequence and structural divergence in family Symbiodiniaceae and genus
403 *Symbiodinium*. BMC Biology. 19:73.

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

406 Hume BCC, Smith EG, Ziegler M, Warrington HJM, Burt JA, LaJeunesse TC, Wiedenmann
407 J, Voolstra CR. 2019. SymPortal: A novel analytical framework and platform for coral algal
408 symbiont next-generation sequencing ITS2 profiling. Molecular Ecology Resources.
409 19:1063-1080.

410 Jeong HJ, Lee SY, Kang N, Yoo Y, Lim AS, Lee MJ, Yih W, Yamashita H, LaJeunesse T.
411 2014. Genetics and morphology characterize the dinoflagellate *Symbiodinium voratum*, n. sp.,
412 (Dinophyceae) as the sole representative of *Symbiodinium* Clade E. The Journal of
413 Eukaryotic Microbiology. 61:75-94.

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

417 Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature
418 Methods. 9:357-359.

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

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

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

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

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

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

433 Reich HG, Kitchen SA, Stankiewicz KH, Devlin-Durante M, Fogarty ND, Baums IB. 2021.
434 Genomic variation of an endosymbiotic dinoflagellate (*Symbiodinium* 'fitti') among closely
435 related coral hosts. Molecular Ecology. 30:3500-3514.

436 Shah S, Dougan KE, Chen Y, Lo R, Laird G, Fortuin MDA, Rai SK, Murigneux V,
437 Bellantuono AJ, Rodriguez-Lanetty M, et al. 2023. Massive genome reduction occurred prior
438 to the origin of coral algal symbionts. bioRxiv. 2023.03.24.534093.

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

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

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

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

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

452 Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee T-h, Jin H, Marler B, Guo H, et al.
453 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and
454 collinearity. Nucleic Acids Research. 40:e49.

455 Wham DC, Ning G, LaJeunesse TC. 2017. *Symbiodinium glynnii* sp. nov., a species of stress-
456 tolerant symbiotic dinoflagellates from pocilloporid and montiporid corals in the Pacific
457 Ocean. Phycologia. 56:396-409.

458 Wilkinson SP, Fisher PL, van Oppen MJH, Davy SK. 2015. Intra-genomic variation in
459 symbiotic dinoflagellates: recent divergence or recombination between lineages? BMC
460 Evolutionary Biology. 15:46.

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

464 Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends in
465 Ecology & Evolution. 15:496-503.

466

467

468 **Table**

469 **Table 1. Tandemly duplicated (TD) genes within 19 Suessiales isolates.**

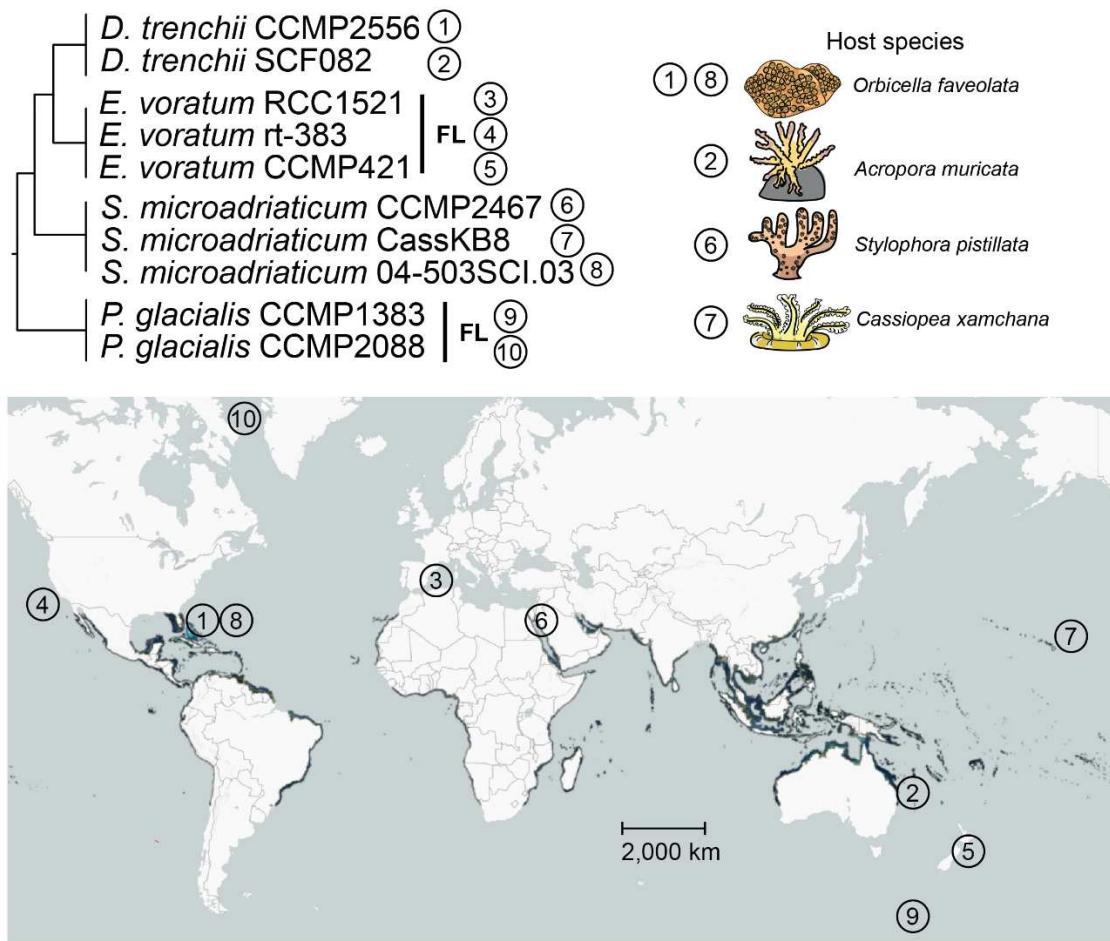
470 TD genes were defined as ≥ 2 consecutive genes on the same scaffold making up a “block”,
471 with its size represented by the total number of consecutive TD genes.

Species and isolate	Number of TD genes	Number of TD blocks	Median of TD block size	Maximum TD block size	Number of single-exon genes in the genome	% of single-exon genes among TD genes
<i>B. minutum</i> Mf1.05b.01	1,225 (3.7%)	569	2	7	2,054 (6.3%)	9.9
<i>Cladocopium</i> sp. C92	1,148 (2.5%)	536	2	8	789 (1.7%)	2.2
<i>C. goreau</i> SCF055	2,017 (6.0%)	937	2	7	1,870 (5.6%)	9.6
<i>D. trenchii</i> CCMP2556	1,031 (1.8%)	745	2	6	3,828 (6.9%)	9.2
<i>D. trenchii</i> SCF082	1,045 (2.0%)	645	2	6	5,677 (10.6%)	7.5
<i>E. voratum</i> CCMP421	495 (1.5%)	233	2	4	1,420 (4.4%)	5.1
<i>E. voratum</i> RCC1521	1,405 (4.4%)	559	3	9	3,983 (12.0%)	23.1
<i>E. voratum</i> rt-383	1,567 (3.9%)	635	3	7	3,574 (9.0%)	22.5
<i>S. linucheae</i> CCMP2456	737 (2.3%)	348	2	6	255 (0.8%)	8.4
<i>S. microadriaticum</i> 04-503SCI.03	437 (1.1%)	206	2	4	2,734 (7.1%)	5.9
<i>S. microadriaticum</i> CassKB8	418 (1.0%)	200	2	4	3,074 (7.2%)	5.7
<i>S. microadriaticum</i> CCMP2467	1,060 (2.2%)	475	2	7	2,770 (5.7%)	4.2
<i>S. natans</i> CCMP2548	2,499 (7.1%)	1,021	2	13	5,099 (14.5%)	21.8
<i>S. necroappetens</i> CCMP2469	577 (1.6%)	274	2	6	3,187 (8.9%)	14.9
<i>S. pilosum</i> CCMP2461	496 (2.1%)	236	2	4	1,431 (6.1%)	8.3
<i>S. tridacnidorum</i> CCMP2592	2,491 (6.5%)	1,254	2	10	5,192 (11.4%)	19.2
<i>S. tridacnidorum</i> Sh18	581 (2.3%)	272	2	5	3,033 (11.8%)	9
<i>P. glacialis</i> CCMP1383	5,376 (9.2%)	2,095	2	16	15,263 (26.2%)	59.4
<i>P. glacialis</i> CCMP2088	4,028 (7.8%)	1,634	2	14	12,619 (24.4%)	55.2

472

473 **Figures**

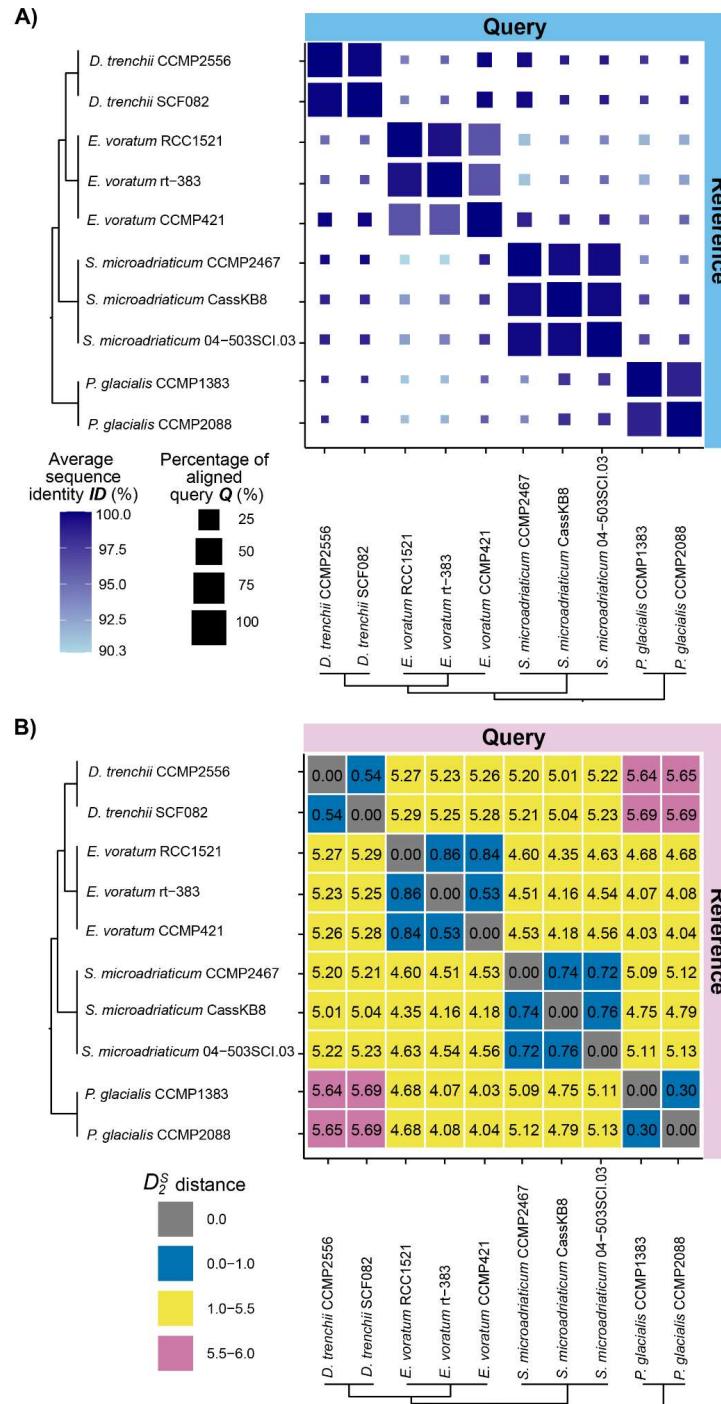
474



475
476 **Figure 1. Suessiales species, following LSU rDNA phylogeny (LaJeunesse et al. 2018),**
477 **for which genome data of multiple isolates are available.**

478 Coral reef (in dark blue and cyan) world map by Allen Coral Atlas (2022). Those not marked
479 FL (free-living) are symbiotic and their host species are represented on the top right.

480
481

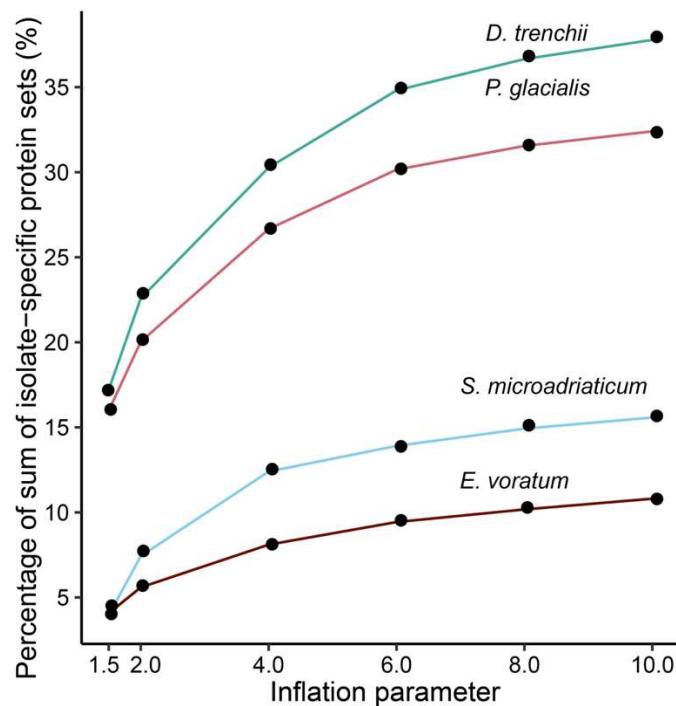


482

483 **Figure 2. Intra/interspecies genome sequence identity among the four Suessiales species.**

484 (A) Alignment-based identity (minimum alignment length = 100 bp) with query genome
 485 sequences (y-axis) aligned to the references (x-axis). The colour of the squares corresponds to
 486 percent sequence identity *ID* (darker blue = higher identity) and the sizes represent the
 487 percentage of the query genome sequence *Q* aligned to the reference. (B) Alignment-free D_2^S
 488 distances (*d*) showing delineation between species (*d* < 1 in blue), Family (*d* between 1.0 and
 489 5.5 in yellow), and the longest evolutionary distance across the Order (*d* > 5.5 in pink).

490



491

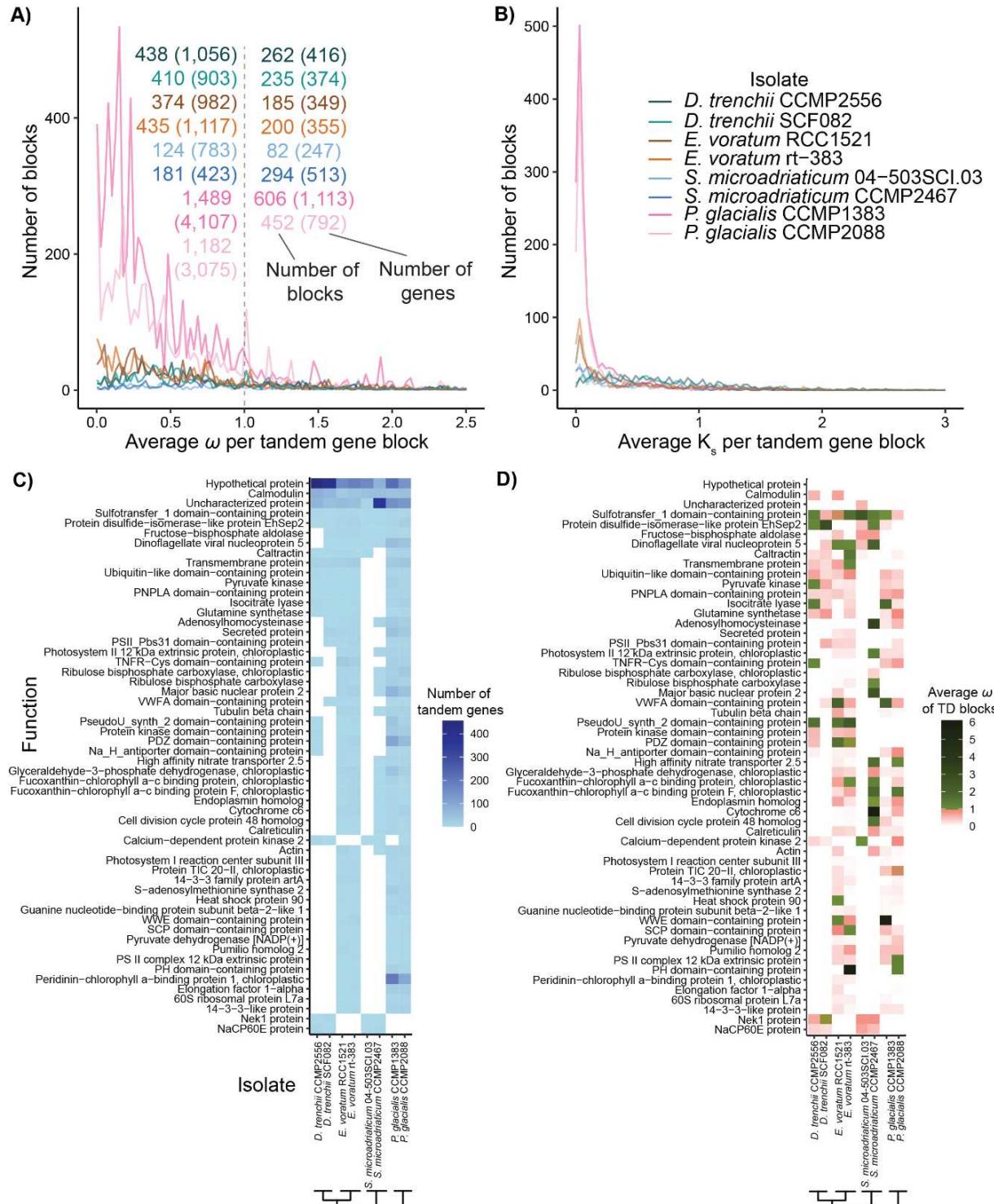
492 **Figure 3. The percentage of isolate-specific protein sets in each Suessiales species.**

493 Protein sequences were clustered at inflation parameter I between 1.5 and 10 using

494 OrthoFinder.

495

496



497

498 **Figure 4. TD genes and their functions in eight Suessiales isolates.**

499 The number of TD blocks showing distribution respectively for (A) mean ω and (B) mean K_s
500 of each TD block and its associated TD genes with $\omega < 1$ or > 1 . Functions encoded by TD
501 blocks that were recovered in genomes of both isolates in one or more species, showing the
502 (C) sum of TD genes, (D) mean ω .