Genetic biomarkers of the sterol-biosynthetic pathway in microalgae

Laura Villanueva*, W. Irene C. Rijpstra, Stefan Schouten, and Jaap S. Sinninghe Damsté

NIOZ Royal Netherlands Institute for Sea Research, Department of Marine Organic Biogeochemistry, PO Box 59, 179AB Den Burg, The Netherlands

To be submitted to *Environmental Microbiology*

***To whom correspondence should be addressed. Royal Netherlands Institute for Sea Research P.O. Box 59, NL-1790 AB Den Burg, The Netherlands. E-mail: laura.villanueva@nioz.nl, phone number: +31 (0)222-369-428, fax number: +31 (0)222-319-674.

Running title: Biomarkers of sterol biosynthesis in microalgae

Keywords: Cycloartenol synthase, sterols, diatom, phytoplankton, phylogenomics, lipid biomarker, microalgae

Summary

Sterols are isoprenoid lipids present in all eukaryotes. These compounds have been used to determine the composition of algal communities in marine and lake environments, and because of their preservation potential, have been used to reconstruct past eukaryotic presence and diversity in the geological record. In the last years there have been major advances in understanding the sterol biosynthetic pathways and the enzymes involved. Here, we have explored the diversity and phylogenetic distribution of the gene coding the cycloartenol synthase protein, a key enzyme of the phytosterol biosynthetic pathway. The cycloartenol synthase gene (CSG) was annotated in genomes of diatoms and other microalgae using protein homology with previously annotated CSG sequences. Based on this, primers for the detection of CSG sequences were designed and evaluated in cultures and environmental samples. A comparison of the phylogeny of the recovered CSG sequences in combination with sequence data of Rubisco gene sequences demonstrates the potential of CSG sequences as phylogenetic marker, as well as an indicator for the identity of sterol-producing organisms in the environment. The proposed gene-based approach can be used to assess the sterol-forming potential of algal groups independent of physiological conditions.

Introduction

Biomarker lipids have been extensively used for determining the composition and function of microbial communities in past and modern environments (e.g. Hinrichs et al. 1999; Sinninghe Damsté et al., 2002; Kuypers et al., 2003; Talbot et al., 2003; Brocks et al., 2005, Volkman et al., 1998; Volkman 2003, etc). Lipids make excellent molecular fossils because of their relative resistance to degradation, and because some have structures unique to certain taxonomic groups. The combination of DNA-based diversity studies (mainly based on ribosomal rRNA gene taxonomy) and chemotaxonomic characterization of lipids has been shown to be a powerful approach to constrain the diversity of microbial communities (e.g. Stefen et al., 1999; Sinninghe Damsté et al., 2004; Villanueva et al., 2004; Rampen et al., 2010). Some studies have also compared biomarker lipids with functional/metabolic genes to assess both the diversity of certain microbial groups as well as their potential ability to perform an activity (e.g. Ertefai et al., 2008, Pitcher et al., 2011).

Sterols are important lipid biomarkers and are present in all eukaryotic organisms. These lipids have been considered as important tools for molecular paleontologists because sterols can be preserved as e.g. steranes in the fossil record for billions of years (Summons et al., 1999; Peters et al., 2005; Brocks and Pearson, 2005, among others). Steranes are thus molecular fossils for the presence sterol-producing organisms and their distribution can give taxonomic information (Moldowan et al., 1990; Brocks et al., 1999; Peters et al., 2005; Kodner et al., 2008). Furthermore, the cyclization of squalene to sterols and some of the following steps in the sterol biosynthetic pathway require

74 molecular oxygen, and the presence of steroids in the fossil record are thus indicators of
75 oxygenation of the atmosphere and oceans (Summons et al., 1999; 2006).
76 The diversity of sterols and their synthetic pathways has been studied extensively and
77 revealed a wide variety of structures (A-ring and side chain alkylation, cyclopropane
78 rings, unsaturations, etc) some of which can be specific for certain eukaroytic groups
79 (Volkman et al., 1998; Volkman, 2003; Rampen et al., 2010). However, most sterols are
80 usually not exclusive of a specific group or genera and, in addition, a change in sterol
81 distributions may also result from changes in environmental and growing conditions
82 rather than community composition changes (Fabregas et al., 1997; Shifrin & Chisholm,
83 1980; Rampen et al., 2009b). Furthermore, the taxonomic distribution of sterols in
84 microalgae is fully based on culture analysis and thus may not reflect environmental
85 diversity since large numbers of environmental gene sequences are different from those
86 of cultivated species. Finally, the algal taxonomy has been mostly based on morphology,
87 or more recently on genetic data based on the 18S rRNA gene and plastid-encoded large
88 subunit of Rubisco (ribulose 1,5-bisphosphate carboxylase, *rbc*L) enzyme, that do not
89 necessary reflect the structural diversity of sterols (Moniz & Kaczmarska 2009; Rampen
90 et al., 2010).
91 One approach to solve the above issues is to examine the presence and diversity of
92 genes involved in the biosynthesis of biomarker lipids as an evidence of the potential
93 ability to biosynthesize the compound of interest, as well as a phylogenetic marker. For
94 example, Pearson and collaborators (2007, 2009) investigated the phylogeny of the
95 producers of hopanoids, isoprenoid bacterial lipids, by analyzing the sequence diversity
96 and distribution of the squalene-hopane cyclase (*sqhC*) gene, concluding that the ability

97 of hopanoid production is not as widespread among bacteria as previously thought.
98 Following the same approach, a recent study by Welander and collaborators (2010)
99 investigated the genes involved in the synthesis of 2-methylhopanoid and showed that the
100 gene required for the C-2 methylation in hopanoids was found in bacterial taxa other than
101 cyanobacteria, invalidating the use of 2-methylhopanes as biomarkers of the appearance
102 of oxygenic photosynthesis on Earth (Welander et al., 2010).
103 In this study, we have made use of recent advances on the phylogenomics of the
104 sterol biosynthetic pathway (Desmond & Gribaldo, 2009) and the growing availability of
105 complete or draft genomes of microalgae, allowing the identification of key genes of the
106 phytosterol biosynthetic pathway. Among all the enzymes of the sterol pathway,
107 oxidosqualene cyclases (OSCs) are one of the most conserved at the sequence level and
108 homologues have been detected in all species capable of sterol synthesis (Desmond &
109 Gribaldo, 2009). There are two main types of OSCs based on the end product of the
110 cyclization: lanosterol synthases (found in animals, fungi, choanozoa, trypanosomatids
111 and dinoflagellates), and cycloartenol synthases (found in higher plants, red and green
112 algae, amoebozoa, diatoms, euglenids and heterolobosea). Previous studies have also
113 identified conserved active sites and specific amino acid residues responsible for
114 particular steps in the cyclization cascade (see Summons et al, 2006 for a review).
115 We targeted the gene encoding the cycloartenol synthase enzyme (CSG) because it is
116 the first specific step in the phytosterol biosynthetic pathway (Fig. 1) and because it is
117 possible to detect homologues of this gene in different organisms due to its conservation
118 at the sequence level (Summons et al., 2006). We have focused on the characterization of
119 the CSG of diatoms as these unicellular algae are thought to be the most common group

120 of eukaryotic phytoplankton in modern oceans and responsible for approximately 40% of
121 marine primary productivity (Falkowski et al., 1998; Moniz&Kaczmarska 2009). Thus,
122 they are likely one of the most important steroid- producing organisms in marine
123 environments.
124 We searched for conserved areas of the cycloartenol synthase amino acid sequences
125 in diatom representatives and designed specific primers to recover a fragment sequence
126 with phylogenetic value (i.e. a variable sequence comprised between conserved motifs).
127 We also explored the diversity and distribution of the key CSG of the phytosterol
128 biosynthetic pathway, and evaluated the potential to link the analysis of lipid biosynthetic
129 genes to patterns of distribution and abundance of their specific biomarker (i.e. sterol) in
130 natural environments.
131
132 **Results and Discussion**
133 *Annotation and evolutionary analysis of CSG sequences in diatoms*
134 Only three genomes of the phylum Bacillariophyta (diatoms) are currently available
135 in public databases, either annotated or in draft, i.e. *Thalassiosira pseudonana*,
136 *Phaeodactylum tricornutum,* and *Fragilariopsis cylindricus*. *T. pseudonana* protein
137 GI:223995517 has been previously annotated as cycloartenol synthase;-2,3-
138 epoxysqualene mutase-like protein (Armbrust et al., 2004). A protein blast against non-
139 redundant protein sequences revealed a 71% identity with the *P.tricornutum* protein
140 GI:219120893 formerly annotated as acetyl-coenzyme A synthetase (Bowler et al.,
141 2008). A protein alignment by ClustalW (multiple alignment Gap penalty 10, gap
142 extension 0.2, and Gonnet protein weight matrix) of the two sequences was submitted to

143 the JGI Genome portal (http://genome.jgi-psf.org) protein blast tool using as a reference
144 database the draft genome of the diatom *F. cylindricus*. A protein ortholog to the putative
145 cycloartenol synthase protein (CSP) detected in *T. pseudonana* and *P. tricornutum* was
146 assigned in the scaffold 9 of the draft genome of *F. cylindricus* based on a 63% and 61%
147 protein identity, respectively. In this way, putative CSG sequences in all three diatom
148 whole-genome sequences available in public databases were assigned.

149 The evolutionary divergence between the *T. pseudonana*, *F. cylindricus* and *P.*
150 *tricornutum* open reading frames (ORFs) of the CSG was computed by using the Jukes-
151 Cantor model (codon positions included 1st+2nd+3rd+Noncoding and all ambiguous
152 positions were removed for each sequence pair) showing between 0.45 to 0.49 base
153 substitutions per site and a sequence identity between 25–31%, which suggests important
154 sequence divergence (diversity) even between CSG sequences of diatoms.

155 *CSG sequences in other diatoms*

156 We developed primers for detection of CSG sequences in other diatom genera. These
157 primers were designed to match conserved amino acid sites of the CSP (see Table 1 and
158 supplementary Fig. 1–2). The areas of conserved amino acid positions in the three ORF
159 of the CSG were investigated in an alignment performed by ClustalW (Thompson et al.,
160 1994). For the reverse primer Cycloart_R and CycloR_TPF (Table 1), an area comprising
161 the amino acid motif GYNGSQC was chosen because it is conserved across different
162 phyla following the cycloartenol branch of the sterol biosynthesis, and also because it
163 includes the tyrosine (Y) amino acid position 381 (Y381 motif) that is one of the
164 conserved amino acid residues responsible for particular steps in the cyclization cascade
165 (Summons et al., 2006).

166 The designed primer pairs (Table 1) were tested on DNA extracted from 14 cultures
167 of diatoms, including the three diatoms, *T. pseudonana*, *P. tricornutum*, and *F.*
168 *cylindricus*, for which whole genome data were available and based on which the primers
169 were designed (supplementary table 1). The primer pairs Cycloart_F/R and CycloF_TPF
170 F&R gave the expected size PCR product. Half of the diatom cultures tested showed
171 positive amplification, including *Skeletonema costatum* CCMP 1281, *Skeletonema*
172 *subsalsum* CCAP 1077/8, *Pseudo-nitzschia seriata* CCMP1309, *Extubocellulus spinifer*
173 CCMP 393, and, as expected, the three diatom strains that were used to design the
174 primers, *T. pseudonana* CCMP 1335, *P. tricornutum*, and *F. cylindricus* CCMP 1102.
175 Encouragingly, different genera showed positive results, suggesting that the primers may
176 be generally applicable within the diatom group. However, in some other cases the
177 designed primers failed to amplify members of the same genus (*Thalassiosira* and
178 *Extubocellulus*). This suggests that the sequence of the CSG is more diverse than inferred
179 based on the three diatom whole genome sequences available and our primers may not be
180 generally applicable to all diatoms.
181 To check the identity of the partial CSG sequences obtained from the diatom cultures,
182 other than those of *T. pseudonana*, *P. tricornutum*, and *F. cylindricus*, we performed a
183 translated nucleotide query against protein database blast (xblast). The first subject
184 homolog entry was *T. pseudonana* protein GI:223995517 for all the partial CSG
185 sequences obtained from the diatom cultures, confirming a high homology with
186 annotated CSG sequences and thus their identity as putative CSG sequences as well. The
187 translated CSG sequences were aligned with *T. pseudonana*, *F. cylindricus* and *P.*
188 *tricornutum* putative CSP sequences in order to investigate the phylogenetic diversity

189 between protein sequences (supplementary figure 2). The alignment revealed several
190 amino acid (aa) changes between diatom genera, also between closely related species,
191 such as in the case of *S. costatum* and *S. subsalsum* with a 96% identity and 3 aa changes
192 in the 193 aa-fragment sequences analyzed (aa residues 91, 105 and 144 of the
193 alignment). It is also important to highlight that the CSP sequences of *Pseudo-nitzschia*
194 *seriata* and *F. cylindricus* were the sequences with a lower percentage of identity in
195 comparison to the rest of the diatom sequences but relatively close compared to each
196 other (82% identity; Suppl Fig 2). The same aa change was observed in these two
197 sequences with respect to the others in several positions of the fragment, but only in two
198 of the positions the amino acid was replaced by another of the same nature (see
199 supplementary Figure 2): i.e. position aa 113 in the alignment, threonine/serine (T/S) in
200 *P. seriata* and *F. cylindricus* respectively *vs*. cysteine (C) in the rest (all of them
201 nucleophilic amino acids); position 118, lysine (K) in *P. seriata* and *F. cylindricus*
202 respectively vs. arginine (R) in the rest (all of them basic amino acids). Generally, the
203 conservation of specific motifs of the CSP has allowed its annotation in different phyla
204 (see Summons et al., 2006 for a review). However, a close look at the amino acid
205 sequence of the three putative CSP sequences annotated in *T. pseudonana*, *P.*
206 *tricornutum*, and *F. cylindricus* demonstrates some flexibility in the protein sequence that
207 would suggest that CSP has been influenced by evolutionary pressures even within the
208 phylum Bacillariophyta.

209 *Phylogeny and evolution of CSG sequences in other microalgal groups*

210 In order to construct the phylogenetic diversity and evolution of CSG with respect to
211 other phytoplanktonic groups, we searched for related microalgal sequences in genomic

212 databases. To this end, the putative CSP sequences of *T. pseudonana*, *F. cylindricus* and
213 *P. tricornutum* were considered query sequences in a protein blast in NCBI (BlastP;
214 Altschul et al., 1990) against non-redundant protein sequences and complete/draft
215 genomes available of microalgae. Draft genomes of microalgae available in JGI DOE,
216 such as *Emiliania huxleyi*, were also screened for protein orthologs by pBLAST.

217 The obtained annotated CSP sequences in microalgae (Table 2) were used to
218 construct a maximum likelihood tree (Figure 2A) based on the entire nucleotide
219 translated sequence. In addition, the partial CSG sequences comprised by the primers
220 used in this study were translated and the obtained protein sequences were used to build a
221 CSP tree (Figure 2B). This tree included the partial CSP sequences deduced by xblast
222 obtained from the diatom cultures as well as the annotated sequences are listed in Table
223 2. The phylogeny based on CSP in microalgae was compared with a phylogeny by
224 maximum likelihood method based on the *rbcL* protein sequence (Figure 2C), as well as
225 the 18S rRNA gene sequence (supplementary Fig. 3) of the same algae, as previous
226 studies have suggested that the higher variability of the *rbcL* of the chloroplast with
227 respect to the 18S rRNA might make it more suited for phylogenetic studies (Evans et al.,
228 2007).

229 The topology of the CSP, *rbcL* protein and the 18S rRNA gene sequence trees (Figure
230 2, Suppl Fig 3) was similar compared to each other, with distinctive clustering between
231 Heterokontophyta (Bacillariophyceae, Pelagophyceae and Phaeophyceae) and Clorophyta
232 (Chlorophyceae, Trebouxiophyceae and Mamiellophyceae). However, both 18S rRNA
233 gene and *rbcL* protein trees showed a more clear separation of the Bacillariophyceae
234 group (diatoms) from the other microalgae, while the CSG tree did not indicate a clear

235 divergence of Bacillariophyceae and the other two sequences of Heterokontophyta
236 (*Aureococcus anophafferens* and *Ectocarpus siliculosus*). Other groups such as
237 Prymnesiophyceae (*Emiliania huxley*) are clearly separated in the tree topology of the
238 CSP and 18S rRNA gene trees from the other microalgae, while in the *rbcL* protein tree
239 the *E.huxleyi* protein sequence is clustered with *A. anophafferens* and *E. siliculosus*
240 (Figure 2C). In both the CSP and the 18S rRNA gene trees the divergence of the
241 Chlorophyceae and Trebouxiophyceae groups (both Chlorophyta) looks similar in
242 comparison with the *rbcL* protein tree. In conclusion, the CSP-based phylogeny follows
243 the same distribution of groups of the *rbcL* protein and 18S rRNA gene-based
244 reconstruction, being able to cluster major groups (e.g. Haptophyta, Chlorophyta and
245 Heterokontophyta), as well as families (e.g. Bacillariophyceae), and genera. This
246 demonstrates the potential of the CSG sequences to be used as a phylogenetic marker.
247 However, the different clustering generated by CSP, *rbcL* protein and 18S rRNA gene
248 observed in some cases, such as in *E. huxleyi, E. siliculosus* and *A. anophagefferens*
249 requires further attention as it might help to clarify the evolutionary history of these three
250 groups.
251 When we focus on the diatom sequences clustering, a comparison between the protein
252 tree based on the entire CSP and the studied fragment displays the same topology (Figure
253 2A and 2B), which supports the phylogenetic value of the CSG fragment amplified by the
254 primers introduced in this study. On the other hand, the phylogenetic clustering of the
255 diatom group by the 18S rRNA gene (Suppl. Fig. 3) and *rbcL* protein sequences (Figure
256 2C) separated the two main groups of centric and pennate diatoms: order Thalassiosirales
257 (*Thalassiosira/Skeletonema*) plus order Cymatosyrales (*Extubocellulus*), and order

258 Bacillariales (*P. seriata/F. cylindricus)* plus order Naviculales (*Phaeodactylum*),
259 respectively. These differences in the phylogenetic distribution of diatoms based on these
260 three genetic markers (i.e. CSP, *rbcL* protein, and 18S rRNA gene) will be more clear
261 once other CSG sequences become available.
262 The alignment of the putative CSP sequences of *T. pseudonana*, *P. tricornutum*, and
263 *F. cylindricus* revealed amino acid changes, suggesting that the CSG is more diverse than
264 previously thought even between members of the same phyla (see below). Thus, in order
265 to test the occurrence of evolutionary events of positive and purifying selection in the
266 codifying region of the CSP sequence we applied the Nei-Gojobori method (based on
267 computing the numbers of synonymous and non-synonymous substitutions and the
268 numbers of potentially synonymous and non-synonymous sites; Nei and Gojobori 1986)
269 to the annotated sequences in Bacillariophyceae (diatoms; Table 3) and Chlorophyta
270 (green algae; Table 4). The identity matrix for the Bacillariophyceae class (Table 3A)
271 revealed the highest percentage of identity between CSP sequences of the same genus (*S.*
272 *costatum* and S. *subsalsum*, 92%). CSP sequences of diatoms belonging to the same order
273 such as Thalassiosirales (*Skeletonema & T. pseudonana*) and Bacillariales (*Pseudo-*
274 *nitzschia seriata & F. cylindricus*) had a percentage of identity of 81% and 88%,
275 respectively. The lowest percentage of identity was observed between *P. tricornutum*
276 (order Naviculales) and *P. seriata/F. cylindricus* (Bacillariales; 60%). In the Z-test of
277 selection for Bacillariophyceae (Table 3B), the test statistic (ds–dN; number of
278 synonymous substitutions per synonymous site minus number of non-synonymous
279 substitutions per non-synonymous site) or probability of rejecting the null hypothesis of
280 neutrality (ds=dN) in favor of the alternative hypothesis (ds>dN, purifying selection) is

281 shown above the diagonal. In this case, the test of selection clearly stated the role of
282 purifying selection with a significant *p*-value in all cases. As expected, the lower value of
283 the test statistic is found between species of the same genus (*Skeletonema*) as the
284 members of the same genus have very similar CSP sequences and thus, any amino acid
285 change in the protein would result in purifying selection (favoring synonymous
286 substitutions that code for the same amino acid).

287 For Chlorophyta (green algae), CSP sequences of members of the same class, e.g.
288 *Chlamydomonas/Volvox* (class Chlorophyceae) and *Ostreococcus/Micromonas* (class
289 Mamiellophyceae) had a protein percent identity 61–71% (Table 4A). For the class
290 Bacillariophyceae the following CSP sequence identity percentages can be found in the
291 different taxonomic divisions: Phylum (40–50%) < Class (60–70%) < Order (80–90%) <
292 Genus (more than 95%).Thus, although there is a high degree of conservation of CSP
293 sequences (see Summons et al, 2006 for a review), the CSG sequence is diverse enough
294 to distinguish species of the same genus supporting its value as a phylogenetic marker.

295 *Detection of CSG sequences in environmental samples*

296 We tested the CSG primers (Table 1) on several environmental samples for the
297 detection of diatom-related CSG sequences. These samples were a microbial mat from
298 the island Schiermonnikoog that is characterized by the presence of pennate diatoms (e.g.
299 *Navicula, Diploneis, Amphora* and *Cylindrotheca*) as shown using microscopic methods
300 (Dijkman et al., 2010). Furthermore, we also analyzed suspended particulate matter
301 (SPM) from North Sea surface water which has high contents of diatom pigments
302 (diatoxanthin, diadinoxanthin, data not shown) and in which *Thalassiosira*, *Chaetoceros*

303 and *Skeletonema* species were previously observed (Cadee and Hegeman, 2002;
304 Brandsma et al., 2012).
305 Primer pairs cycloart_F/R and cycloF_TPF F&R designed based on the diatom
306 sequences available gave positive results on the environmental samples tested in this
307 study. Thus, CSG sequences were amplified and sequenced from the microbial mat and
308 North Sea SPM and analyzed phylogenetically by the maximum likelihood method. The
309 sequences recovered from the North Sea water column mainly clustered with the order
310 Thalassiosirales (*Skeletonema* and *Thalassiosira pseudonana*), while the majority of the
311 diatom mat CSG sequences were closer to the representative sequence of *Extubocellulus*
312 *spinifer* (order Cymatosyrales; centric diatom) with some also closer to the
313 Thalassiosirales order cluster (Figure 3). As a comparison we also sequenced the *rbcL*
314 gene: its distribution was more diverse than the CSG sequence-based tree (Figure 4). The
315 majority of the *rbcL* gene sequences retrieved from the North Sea SPM clustered with the
316 order Thalassiosirales but also to sequences such as *Phaeodactylum tricornutum* (order
317 Naviculales). *rbcL* gene sequences retrieved from the microbial mat were also more
318 diverse and clustered with *Extubocellulus spinifer* as well as with the orders
319 Thalassiosirales, Bacillariales (*Pseudo-nitzschia* & *Fragilariopsis*) and Naviculales.

320 In addition, we also characterized the sterols in these samples to obtain an idea of the
321 sterol diversity and potentially already identify specific diatom sterols. The main sterols
322 in the Schiermonnikoog mat were cholest-5-en-3β-ol (cholesterol), cholesta-5,24-dien-
323 3β-ol (desmosterol), 24-methylcholest-5,24(28)-dien-3β-ol, 24-methylcholest-5-en-3β-ol
324 (campesterol), and 24-ethylcholest-5-en-3β-ol (sitosterol) (see Suppl Table 2 for details).
325 These sterols have been reported for diatom cultures (Rampen et al. 2010), specifically,

326 24-methylcholesta-5,24(28)-dien-3β-ol is found in high abundances in some centric
327 diatoms such as *Thalassiosira* and *Skeletonema* genera (Rampen et al., 2010). In the
328 North Sea water SPM the main sterols were cholesterol, desmosterol and 24-
329 methylcholest-5,24(28)-dien-3β-ol. The diversity of phytosterols detected in the North
330 Sea SPM was thus lower than in the diatom mat and only one sterol, 5α(H)-24-methyl-
331 cholest-22en-3β-ol, was uniquely found in the North Sea water but not in the diatom mat.
332 The abundant presence of 24-methylcholesta-5,24(28)-dien-3β-ol in both environments
333 may indicate the importance of diatoms in these two systems. In addition, other sterols
334 such as brassicasterol, also known as diatomsterol, 24-methylcholest-5,22E-dien-3β-ol,
335 and 24-ethylcholesta-5,22E-dien-3β-ol (stigmasterol) have also been associated with
336 diatoms (Rontani & Volkman, 2005). However, the presence of these sterols does not
337 provide conclusive for their origin of diatoms as all these sterols are also found in other
338 algae while the sterol composition also does not indicate which diatom genera are present
339 (Volkman et al., 1998; Volkman, 2003; Rampen et al., 2010 and references cited therein).
340 A comparison of the CSG tree with those of the sterol distributions support the idea
341 that diatoms from the order Thalassiosirales in both samples may be the source of the 24-
342 methylcholest-5,24(28)-dien-3β-ol. In addition, the higher diversity of diatom CSG
343 sequences detected in the microbial mat compared to the North Sea SPM also
344 corresponds to a higher diversity of detected sterols. The detection of sequences
345 homologous to the Bacillariales and Naviculales orders in the microbial mat suggests that
346 cholesta-5,22-dien-3β-ol and 24-ethylcholesta-5,22-dien-3β-ol may be sourced by these
347 diatoms as they are also dominant sterols in cultivated relatives (Rampen et al., 2010).

348 The fact that CSP appear to be less conserved than other phylogenetic markers has the
349 disadvantage that this protein might be more difficult to annotate based on protein
350 homology. On the other hand, it might provide clues about the evolutionary placement of
351 certain organisms better than more conserved proteins. In general, CSG and *rbcL* gene
352 analysis in environmental samples has proven effective for surveying the sterol-forming
353 diatom community. However, the fact that *rbcL* gene analysis has elucidated more
354 diversity indicates that CSG primers are still limited in their diversity coverage.
355 *Implications*
356 The approach presented here based on specific gene searching in whole-genome and
357 metagenomic databases has allowed the design of effective primers for the detection of a
358 key gene in the sterol biosynthetic pathway (i.e. CSG) in microalgae. The comparison
359 between the phylogenetic reconstruction of microalgae 18S rRNA gene, *rbcL* protein and
360 CSP sequences supports the value of CSG sequences as marker of the presence and
361 phylogeny of sterol-producing microalgae. However, further studies are needed to
362 improve the diversity coverage of the CSG marker by sequencing more whole genomes
363 of diatoms and other microalgae. Through this it will be possible to redesign the
364 developed primers and assign a more refined taxonomic identification, which will lead to
365 a more accurate the association between CSG sequences, their producer, and the sterol
366 composition in environmental samples .
367 The genomic characterization of enzymes involved in lipid biosynthetic pathways
368 opens a new chapter in organic geochemistry studies. Genomic approaches provide an
369 independent assessment of the organism ability to produce a molecule of interest without
370 extensive screening of cultures. Our study has expanded the range of CSG sequences

371 available, introduced a quick screening of environmental samples for the diversity of
372 sterol-forming microorganisms and may, once the sequence coverage has increased,
373 provide a link between sterols and their main sources. This is also the starting point of
374 other studies involving determination of the abundance and expression of this key gene of
375 the phytosterol biosynthetic pathway in environmental samples. Ultimately, the CSG has
376 potential to elucidate the origin of certain microalgae groups and as a molecular clock to
377 track the appearance of the sterol biosynthetic pathway.
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393

Material and Methods

Sampling

A diatom-dominated microbial mat was sampled in the sandy beach of the Dutch barrier island Schiermonnikoog (53°29´N and 6°08´E; for a more detailed description see Stal et al., 1985) in January 2010, transported to the lab at 4°C and then stored at –80°C until further analysis. North Sea surface water was sampled at a Jetty platform at the NIOZ at the western entrance of the North Sea into the Wadden Sea at the Island Texel (53°0´2´´N, 4°7´2´´E). With each incoming-tide, water from the coastal North Sea moves as far as 25 km into the Wadden Sea (Potsma, 1954). At high tide, water collected represents Dutch coastal North Sea waters since the estuarine influence is minimal. Strong tidal currents assure that the water is vertically mixed. Therefore, surface water samples taken during high tide are representative of the entire water column. Suspended particulate matter (SPM) sample was taken on 24th March 2010. For DNA analysis, measured volumes (ca. 1 L) of water were filtered through a 142 mm diameter, 0.2 µm pore size polycarbonate filter (Millipore, Billerica, MA) and stored at –80°C until extraction. For lipid analyses, a measured volume (ca. 20 L) of water was filtered sequentially through pre-ashed 3 µm, and 0.7 µm-pore-size, glass fiber filters (GF/F, Pall, 142 mm filter diameter). GF/F filters were stored at –40°C until extraction. Diatom cultures were obtained from culture collection, grown in batch cultures and harvested at the end of the logarithmic growth phase by filtration on pre-ashed 47 mm, 0.7 µm GF/F filters (see Rampen et al., 2010).

417 *DNA extraction*

418 Approximately 0.2 g of wet weight of diatomaceous-mat was homogenized by a DNA-
419 clean spatula and extracted by the Power Biofilm DNA extraction kit from MoBio
420 (Carlsbad, CA) according to the manufacturer's instructions. North Sea water samples on
421 0.2 μm PCC filters were extracted by a bead-beating protocol followed by DNeasy
422 extraction kit of Qiagen (Valencia, CA). Diatom cultures were extracted as described by
423 Rampen et al. (2009a). DNA quality and concentration were evaluated by gel agarose
424 electrophoresis and Nanodrop (Wilmington, DE).

425 *PCR amplification, cloning and sequencing*

426 Partial CSG sequences were amplified by using the primers listed in Table 1 on diatom
427 pure cultures DNA extracts and environmental samples. PCR reaction mixture was the
428 following (final concentration): Q-solution 1×; PCR buffer 1×; BSA (200 μg/ml); dNTPs
429 (20 μM); primers (0.2 pmol/μl); $MgCl_2$ (1.5 mM); 1.25 U Taq polymerase (Qiagen,
430 Valencia, CA, USA) or BioThermD Taq DNA polymerase (Semiramis Genetics Ltd.,
431 Manchester, UK). PCR conditions for these amplifications were the following: 95°C, 5
432 min; 40× [95°C, 1 min; Tm, 1 min; 72°C, 1 min]; final extension 72°C, 5 min. A gradient
433 PCR cycle was performed for each set of primers and samples from 48 to 55°C melting
434 temperature. Amplification of Rubisco gene was performed as described by Rampen et
435 al., 2009a. Positive amplification bands were excised from agarose gel and gel or PCR
436 purified (QIAquick gel/PCR purification kit, Qiagen) and cloned in the TOPO-TA
437 cloning® kit from Invitrogen (Carlsbad, CA, USA) and transformed in *E. coli* TOP10
438 cells following the manufacturer's recommendations. Recombinant clones plasmid DNAs
439 were purified by Qiagen Miniprep kit and screening by sequencing using M13F (-20) (5'-

440 GTA AAA CGA CGG CCA G-3') and M13R (5'-CAG GAA ACA GCT ATG AC-3')
441 primers with BigDye® v1.1 sequencing kit in house on a ABI PRISM® 310 Genetic
442 analyzer (Applied Biosystems, Foster city, CA, USA) or sequenced in Macrogen Europe
443 Inc.
444 *Alignments, tree reconstruction and evolutionary analyses*
445 Putative CSG partial sequences obtained from diatom pure cultures and environmental
446 samples were translated to protein by submitting them as query sequences in translated
447 blast (xblast: Find similar proteins to translated query in a protein database) and reviewed
448 by manual annotation. DNA/Protein alignments were performed by ClustalW (multiple
449 alignment Gap penalty 15, gap extension 6.66, and IUB DNA weight matrix) (Thompson
450 et al., 1994). Mega5 software (Tamura et al., 2011) was used to estimate the best
451 DNA/protein models for maximum likelihood analysis (automatic neighbor joining tree;
452 statistical method, maximum likelihood; substitution type, nucleotide/amino acid; use all
453 sites). In case of protein alignments, the choice of protein model by Mega5 (with models
454 ranked by Bayesian information criterion, BIC) was contrasted with the choice of model
455 of evolution for protein phylogeny given by ProtTest 2.4 (Abascal et al., 2005) with
456 model selection criterion of Akaike Information Criterion (AIC) and (Bayesian
457 Information Criterion) BIC. Maximum likelihood phylogenetic reconstruction of the 18S
458 rRNA gene, partial CSG sequences and partial *rbcL* gene sequences was performed by
459 Mega5 using the model with higher ranking (see figure legends for details). For the 18S
460 rDNA, partial CSG sequences and partial *rbcL* gene sequences, the genera; time
461 reversible model plus gamma distribution (GTR + G) was used (plus invariant sites, + I
462 for the case of the *rbcL* gene sequences tree). *rbcL* protein and CSP sequence trees were

463 generated with the WAG+G+F model. Bootstrap analysis was performed in all cases with
464 1000 replicates. All sites were considered for the calculations and the maximum
465 likelihood heuristic method chosen was nearest neighbor interchange (NNI).
466 Evolutionary analyses for annotated partial CSP sequences were performed with
467 MEGA5. Codon-based Z-test Test of Purifying Selection was also performed. Analyses
468 were conducted using the Nei-Gojobori method and bootstrap methods with 1000
469 replicates.
470 *Data submission*
471 Partial CSG sequences were deposited in GenBank under the accession numbers
472 (accession number pending to be assigned). Partial *rbcL* gene sequences were deposited
473 under accession numbers (accession number pending to be assigned).
474 *Lipid extraction, separation and GC/MS detection*
475 Freeze-dried microbial mats were homogenized by lipid-free mortar and pestle and the
476 GF/F filters were freeze-dried and cut into small pieces with sterile scissors before being
477 ultrasonically extracted four times using dichloromethane (DCM)/methanol (MeOH)
478 (1:1, v/v) as described in Rampen et al 2009a. An aliquot of the total lipid extracts was
479 separated over a pipette-column filled with Al_2O_3, using hexane:dichloromethane (DCM;
480 9:1, v:v) and DCM: methanol (MeOH; 1:1, v:v) to elute the apolar and sterol fractions,
481 respectively. Prior to analysis by gas chromatography (GC) and gas chromatography–
482 mass spectrometry (GC–MS), the sterol fractions were silylated by adding 25 µl BSTFA
483 [N,O-bis(trimethylsilyl)trifluoroacetamine] and 25 µl pyridine and heating the mixtures
484 at 60°C for 20 min. GC and GC–MS analyses were performed as described by Rampen et

al. (2009a). Sterols (as their TMS derivatives) were identified based on their mass spectra and relative retention times in comparison with literature data.

506 **References**

507 Abascal, F., Zardoya R., and Posada D. (2005) ProtTest: Selection of best-fit models of
508 protein evolution. Bioinformatics 21: 2104-2105.

509 Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local
510 alignment search tool. J Mol Biol 215: 403-410.

511 Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H.,
512 Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., and Brzezinski, M.A. et al. (2004) The
513 genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism.
514 Science 306: 79-86.

515 Bowler, C. et al. (2008) The Phaeodactylum genome reveals the evolutionary history of
516 diatom genomes. Nature 456:239-244.

517 Brandsma, J., Hopmans, E.C., Phillippart, K.J.M., Vedhuis, M.J.W., Schouten, S., and
518 Sinninghe Damste, J.S. (2012) Low temporal variations in the intact polar lipid
519 composition of North Sea coastal marine water reveals limited chemotaxonomic value.
520 Biogeosciences 9: 1073-1084.

521 Brocks, J.J., Logan, G.A., Buick, R., and Summons, R.E. (1999) Archean molecular
522 fossils and early rise of eukaryotes. Science 285: 1033-1036.

523 Brocks, J.J., and Pearson, A.P. (2005) Building the biomarker tree of life. Rev Mineral
524 Geochem 59: 233-258.

525 Brocks, J.J., Love, G.D., Summmons, R.E., Knoll, A.H., Logan, G.A., and Bowden, S.A.
526 (2005) Biomarker evidence for green and purple sulphur bacteria in a stratified
527 Palaeoproterozoic sea. Nature 437: 866-870.

528 Cadée, G.C., and Hegeman, J. (2002) Phytoplankton in the Marsdiep at the end of the
529 20th century; 30 years monitoring biomass, primary production, and Phaeocystis
530 blooms. J Sea Res 48: 97-110.
531 Desmond, E., and Gribaldo, S. (2009) Phylogenomics of sterol biosynthesis: Insights into
532 the origin, evolution and diversity of a key eukaryotic feature. Genome Biol Evol 1:
533 364-381.
534 Dijkman, N.A., Boschker, H.T.S., Stal, L.J., and Kromkamp, J.C. (2010) Composition
535 and heterogeneity of the microbial community in a coastal microbial mat as revealed
536 by the analysis of pigments and phospholipid-derived fatty acids. J Sea Res 63:.62-70.
537 Ertefai, T.F., Fisher, M.C., Fredricks, H.F., Lipp, J.S., Pearson, A., Birgel, D., Udert,
538 K.M., Cavanaugh, C.M., Gschwend, P.M., Hinrichs, K.U. (2008) Vertical distribution
539 of microbial lipids and functional genes in chemically distinct layers of a highly
540 polluted meromictic lake. Org Geochem 39: 1572-1588.
541 Evans, K.M., Wortley, A.H., and Mann, D.G. (2007) An assessment of potential diatom
542 "barcode" genes (cox1, rbcL, 18S, and ITS rDNA) and their reffectiveness in
543 determining relationships in Sellaphora (Bacillariophyta). Protist 158: 349-364.
544 Fabregas, J., Aran, J., Morales, E.D., Lamela, T., and Otero, A. (1997) Modification of
545 sterol concentration in marine microalgae. Phytochemistry 46: 1189-1191.
546 Falkowski, P.G., Barber, R.T., and Smetacek, V. (1998) Biogeochemical controls and
547 feedbacks on ocean primary production. Science 281: 200-206.
548 Fischer,W.W., and Pearson, A. (2007) Hypotheses for the origin and early evolution of
549 triterpenoid cyclases. Geobiol 5: 19-34.

550 Hinrichs, K.-U., Hayes, J.M., Sylva, S.P., Brewer, P.G., and DeLong, E.F. (1999)
551 Methane-consuming archaebacteria in marine sediments. Nature 398: 802-805.
552 Kodner, R.B., Pearson, A., Summons, R.E., and Knoll, A.H. (2008) Sterols in red and
553 green algae: quantification, phylogeny, and relevance for the interpretation of geologic
554 steranes. Geobiol 6: 411-420.
555 Kuypers, M.M.M., Sliekers, A.E., Lavik, G., Schmid, M., Jorgensen, B.B., Kuenen, J.G.,
556 Sinninghe Damsté, J.S., Strous, M., and Jetten, M.S.M. (2003) Anaerobic ammonium
557 oxidation by anammox bacteria in the Black Sea. Nature 422: 608-611.
558 Kuypers, M.M.M., van Breugel,Y., Schoten, S., Erba, E., and Sinninghe Damste, J.
559 (2004) N_2-fixing cyanobacteria supplied nutrient N for Cretaceous oceanic anoxic
560 events. Geology 32: 853-856.
561 Moldowan, J.M., Fago, F.J., Lee, C.Y., Jacobson, S.R., Watt, D.A., Slougui,
562 N.,Jeganathan, A., and Young, D.C. (1990) Sedimentary 24-n propylcholestanes,
563 molecular fossils diagnostic of marine algae. Science 247: 309-312.
564 Moniz, M.B.J., and Kaczmarska, I. (2009) Barcoding diatoms: Is there a good marker?
565 Mol Ecol Resources 9: 65-74.
566 Nei, M., and Gojobori, T. (1986) Simple methods for estimating the numbers of
567 synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418-426.
568 Pearson, A., Flood Page, S.R., Jorgenson, T.L., Fischer, W.W., and Higgins, M.B. (2007)
569 Novel hopanoid cyclases from the environment. Environ Microbiol 9: 2175-2188.
570 Pearson, A., Leavitt, W.D., Saenz, J.P., Summons, R.E., Tam, M.C.-M., and Close, H.G.
571 (2009) Diversity of hopanoids and squalene-hopene cyclasghes across a tropical land-
572 sea gradient. Environ Microbiol 11: 1208-1223.

573 Peters, K.E., Walters, C.C., and Moldowan, J.M. (2005) The biomarker guide.
574 Cambridge University Press, Cambridge, UK.
575 Pitcher, A., Villanueva, L., Hopmans, E.C., Schouten, S., Reichart, G.J., and Sinninghe
576 Damsté, J.S. (2011) Niche segregation of ammonia-oxidizing archaea and anammox
577 bacteria in the Arabian Sea oxygen minimum zone. ISME J 5: 1896-1904.
578 Potsma, H. (1954) Hydrography of the Dutch Wadden Sea. Archives Néerlandaises de
579 Zoologie 10: 405-511.
580 Rampen, S.W., Abbas, B.A., Schouten, S., and Sinninghe Damsté, J.S. 2010. A
581 comprehensive study of sterols in marine diatoms (Bacillariophyta): Implications for
582 their use as tracers for diatom productivity. Limnol Oceanograp 55: 91-105.
583 Rampen, S.W., Schouten, S., Koning, E., Brummer, G-J.A., and Sinninghe Damsté, J.S.
584 (2008) A 90 kyr upwelling record from the northwestern Indian Ocean using a novel
585 long-chain diol index. Earth Planet Sci Letters 276: 207-213.
586 Rampen, S.W., Schouten, S., Panoto, E., Brink, M., Andersen, R.A., Muyzer, G., Abbas,
587 B., and Sinninghe Damsté, J.S. (2009a) Phylogenetic position of Attheya longicornis
588 and Attheya septentrionalis (Bacillariophyta). J Phycol 45: 444-453.
589 Rampen, S.W., Schouten, S., Schefuß, E., and Sinninghe Damsté, J.S. (2009b) Impact of
590 temperature on long chain diol and mid-chain hydroxyl methyl alkanoate composition
591 in Proboscia diatoms: Results from culture and field studies. Org Geochem 40: 1124-
592 1131.
593 Rontani, J-F., and Volkman, J.K. (2005) Lipid characterization of coastal cyanobacterial
594 mats from the Camargue (France). Org Geochem 36: 251-272.

595 Shifrin, N.S., Chisholm, S.W. (1980) Phytoplankton lipids: Enviornmental influences on
596 production and possible commercial applications. In Algae biomass. Shelef, G., and
597 Soeder, C.J. (eds). Elsevier, Amsterdam, pp 623-645.
598 Sinninghe Damsté, J.S., Muyzer, G., Abbas, B., Rampen, S.W., Masse, G., Guyallard,
599 W., Belt, S.T., Robert, J-M., Rowland, S.J., Moldowan, J.M., Barbati, S.M., Fago,
600 F.J., Denisevich, P., Dahl, J., Trindade, L.A.F., and Schouten, S. (2004) The rise of
601 rhisolenid diatoms. Nature 304: 584-587.
602 Sinninghe Damsté, J.S., Schouten, S., Hopmans, E.C., van Duin, A.C.T., and
603 Geenevasen, J.A.J. (2002) Crenarchaeol: the characteristic core glycerol dibiphytanyl
604 glycerol tetraether membrane lipid of cosmopolitan pelagic crenarchaeota. J Lipid Res
605 43: 1641-1651.
606 Stal, L.J., van Gemerden, H., and Krumbein, W.E. (1985) Structure and development of
607 a benthic marine microbial mat. FEMS Microbiol Ecol 31: 111-125.
608 Stephen, J.R., Chang, Y-J., Gan, Y.D., Peacock, A., Pfiffner, S.M., Barcelona, M.J.,
609 White, D.C., and Macnaughton, S.J. (1999) Microbial characterization of a JP-4 fuel –
610 contaminated site using a combined lipid biomarker/polymerase chain reaction-
611 denaturing gradient gel electrophoresis (PCR-DGGE)-based approach. Environ
612 Microbiol 1: 231-241.
613 Summons, R.E., Jahnke, L.L., Hope, J.M., and Logan, G.A. (1999) 2-methylhopanoids as
614 biomarkers for cyanobacterial oxygenic photosynthesis. Nature 400: 554-557.
615 Sumons, R.E., Bradley, A.S., Jahnke, L.L., and Waldbauer, J.R. (2006) Steroids,
616 triterpenoids and molecular oxygen. Phil Trans R Soc B 361: 951-968.

617 Talbot, H.M., Watson, D.F., Pearson, E.J., and Farrimond, P.(2003) Diverse biohopanoid
618 compositions of non-marine sediments. Org Geochem 34: 1353-1371.
619 Tamura, K., Peterson, P., Peterson, N., Stecher, G., Nei, M., and Kumar, S . (2011)
620 MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood,
621 Evolutionary Distance, and Maximum Parsimony Methods. Mol Biol Evol 28: 2731-
622 2739.
623 Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: improving the
624 sensitivity of progressive multiple sequence alignment through sequence weighting,
625 position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-
626 4680.
627 Villanueva, L., Navarrete, A., Urmeneta, J., White, D.C., and Guerrero, R. (2004)
628 Combined phospholipid biomarker-16S rRNA gene denaturing gradient gel
629 electrophoresis analysis of bacterial diversity and physiological status in an intertidal
630 microbial mat. Appl Environ Microbiol 70: 6920-6926.
631 Volkman, J.K., Barrett, S.M., Blackburn, S.I., Mansour, M.P., Sikes, E.L., Gelin, F.
632 (1998) Microalgal biomarkers: a review of recent research developments. Org
633 Geochem 29: 1163-1179.
634 Volkman, J.K. (2003) Sterols in microorganisms. Appl Microbiol Biotechnol 60: 495-
635 506.
636 Welander, P.V., Coleman, M.L., Sessions, A.L., Summons, R.E., and Newman, D.K.
637 (2010) Identification of a methylase required for 2-methylhopanoid production and
638 implications for the interpretation of sedimentary hopanes. Proc Natl Acad Sci USA
639 107: 8537-8542.

Figure Legends

Figure 1. Sterol biosynthetic pathway. Isoprenoid precursor isopentenyl diphosphate (IPP) is the precursor of squalene. Hopanoids are synthesized from the cyclization of squalene by a squalene-hopene cyclase (SHC) in a process independent of oxygen. For sterol synthesis squalene is transformed by a squalene monooxygenase (SQMO) that requires O_2. Squalene epoxide is then cyclized either to lanosterol or to cycloartenol by lanosterol synthase or cycloartenol synthase.

Figure 2. Phylogenetic tree of cycloartenol synthase protein (CSP) and *rbcL* (Rubisco) protein sequences.

(A) Sequences annotated in microalgal genomes inferred by using the Maximum Likelihood method based on the Whelan And Goldman + Freq. model. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+*G*, parameter = 0.9821). The scale indicates number of substitutions per site. The analysis involved 13 amino acid sequences. There were a total of 993 positions in the final dataset. Bootstrap values (1000 replicates) are indicated on the nodes.

(B) Phylogenetic tree of CSP fragment comprised by the primers applied in this study annotated in microalgal genomes inferred by using the Maximum Likelihood method based on the General Reverse Transcriptase + Freq. model. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+*G*, parameter = 0.9267). The rate variation model allowed for some sites to be evolutionarily invariable ([+*I*], 0.0000% sites). The analysis involved 17

663 amino acid sequences. All ambiguous positions were removed for each sequence
664 pair. There were a total of 228 positions in the final dataset.
665 (C) Phylogenetic tree of *rbc*L protein sequences in the microalgae under study inferred
666 by using the Maximum Likelihood method based on the General Reverse
667 Transcriptase (GRT) + Freq. model. A discrete Gamma distribution was used to
668 model evolutionary rate differences among sites (5 categories (+*G*, parameter =
669 0.8751). The analysis involved 15 amino acid sequences. There were a total of 491
670 positions in the final dataset.
671
672 Figure 3. Phylogenetic tree of CSG sequences obtained from the environmental samples
673 under study and inferred by using the Maximum Likelihood method based on the GTR
674 model. A discrete Gamma distribution was used to model evolutionary rate differences
675 among sites (5 categories (+*G*, parameter = 1.3092). The analysis involved 45 nucleotide
676 sequences. Codon positions included were 1st+2nd+3rd+Noncoding. There were a total
677 of 721 positions in the final dataset.
678
679 Figure 4. Phylogenetic tree of *rbc*L gene sequences obtained from the environmental
680 samples under study and inferred by using the Maximum Likelihood method based on the
681 GTR model. Gamma distribution was used to model evolutionary rate differences among
682 sites (5 categories (+*G*, parameter = 0.2904). The rate variation model allowed for some
683 sites to be evolutionarily invariable ([+*I*], 59.5472% sites). The analysis involved 54
684 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. There
685 were a total of 1427 positions in the final dataset.

Table 1. List of primers for the detection of CSG sequences

Name	AA sequence	Primer (5'-3')	Name	AA sequence	Primer
Cycloart_F	WLLPNWF (146–152 aa)*	TGGCTKCTMCCMAACTGGTTT	Cycloart_R	GYNGSQC (310–316 aa)	G[CATTGGCTKCCGTTRTAGCC]
CycloF_TPF	PNW(F/I)PFHP (149–156 aa)	CCMAACTGGWTTCCTTTYCATC	CycloR_TPF	GYNGSQC (310–316 aa)	[CAYTGGCTKCCGTTRTAKCC]†

*aa (amino acid) positions of the cycloartenol synthase *T.pseudonana* (protein GI:223995517).
†Primer CycloR_TPF has a nucleotide less than Cycloart_R primer but we indicate the same aa motif as a reference.

Table 2. Cycloartenol synthase ortolog proteins deduced from database.

Organism	Ortholog	Phylogeny
Ostreococcus tauri OTH95 Green Algae	XM_003077949 XP_003077997.1*	Viridiplantae; Chlorophyta; Mamiellophyceae; Mamiellales
Ostreococcus lucimarinus CCE9901 Green Algae	XM_001416533 XP_001416570.1	Viridiplantae; Chlorophyta; Mamiellophyceae; Mamiellales
Micromonas pusilla CCMP1545 Green Algae	XM_003055917 XP_003055963.1	Viridiplantae; Chlorophyta; Mamiellophyceae; Mamiellales
Micromonas sp. RCC299 Green Algae	XM_002507604 XP_002507650.1	Viridiplantae; Chlorophyta; Mamiellophyceae; Mamiellales
Chlamydomonas reinhardtii Green Algae	XM_001689822 XP_001689874.1*	Viridiplantae; Chlorophyta; Chlorophyceae; Chlamydomonadales; Chlamydomonadaceae
Chlorella variabilis Green algae	GL433842 EFN56189.1	Viridiplantae; Chlorophyta; Trebouxiophyceae; Chlorellales; Chlorellaceae
Volvox carteri ft. nagariensis Green algae	XM_002955246 XP_002955292.1	Viridiplantae; Chlorophyta; Chlorophyceae; Chlamydomonadales; Volvocaceae; Volvox
Thalassiosira pseudonana Diatom	XM_002287396 XP_002287432.1*	Stramenopiles; Bacillariophyta; Coscinodiscophyceae; Thalassiosirophycidae; Thalassiosirales; Thalassiosiraceae
Phaeodactylum tricornutum Diatom	XM_002185642 XP_002185678.1†	Stramenopiles; Bacillariophyta; Bacillariophyceae; Bacillariophycidae; Naviculales; Phaeodactylaceae
Fragilariopsis cylindricus Diatom	Scaffold 9	Stramenopiles; Bacillariophyta; Bacillariophyceae; Bacillariophycidae; Bacillariales; Bacillariaceae
Aureococcus anophagefferens	GL833121.1 EGB12462.1	Stramenopiles; Pelagophyceae; Aureococcus
Ectocarpus siliculosus	Scaffold sctg_148 CBN75619.1	Stramenopiles; PX clade; Phaeophyceae; Ectocarpales; Ectocarpaceae; Ectocarpus.
Emiliania huxleyi 1516	Scaffold 360	Haptophyceae; Isochrysidales; Noelaerhabdaceae

*annotated as putative or cycloartenol synthase protein. †Annotated as acetyl-coenzyme A synthetase. The rest of the proteins were annotated as hypothetical/predicted protein and assigned as putative cycloartenol synthases in this study based on percentage of identity with annotated orthologs in other taxa. Peptide sequences matching query sequences with E-value $<1e^{-150}$ and aligning over 55% of the query protein length were considered as significant.

Table 3. Evolutionary analysis of annotated partial cycloartenol synthase proteins in class Bacillariophyceae (diatoms)

(A)

Identity matrix						
S. costatum						
S. subsalsum	0.96					
Extubocellulus	0.76	0.75				
Pseudo-nitzchia	0.67	0.66	0.62			
Fragilariopsis	0.64	0.64	0.64	0.82		
Thalassiosira	0.81	0.81	0.73	0.64	0.63	
Phaeodactylum	0.70	0.69	0.71	0.6	0.61	0.67

(B)

Z test of selection	*S. costatum*	*S. subsalsum*	*Extubocellulus*	*Pseudo-nitzchia*	*Fragilariopsis*	*Thalassiosira*	*Phaeodactylum*
S. costatum		6.98	10.82	8.93	9.52	11.09	9.97
S. subsalsum	0.00		12.26	9.87	4.66	11.53	10.81
Extubocellulus	0.00	0.00		7.53	9.21	9.75	10.75
Pseudo-nitzchia	0.00	0.00	0.00		10.34	9.87	9.23
Fragilariopsis	0.00	0.00	0.00	0.00		9.34	9.16
Thalassiosira	0.00	0.00	0.00	0.00	0.00		11.12
Phaeodactylum	0.00	0.00	0.00	0.00	0.00	0.00	

(A) Identity matrix: Identity values between sequences being 1, 100% identical. (B) Codon-based Test of Purifying Selection for analysis between sequences: The probability of rejecting the null hypothesis of strict-neutrality ($d_N = d_S$) in favor of the alternative hypothesis ($d_N < d_S$) (above diagonal) is shown. Below the diagonal the p-values are shown (less than 0.05 are considered significant at the 5% level). The test statistic ($d_S - d_N$) is shown below the diagonal. d_S and d_N are the numbers of synonymous and nonsynonymous substitutions per site, respectively. The variance of the difference was computed using the bootstrap method (1000 replicates). Analyses were conducted using the Nei-Gojobori method (Nei & Gojori, 1986). The analysis involved 7 nucleotide sequences. All ambiguous positions were removed for each sequence pair. There were a total of 193 positions in the final dataset.

Table 4. Evolutionary analysis of annotated cycloartenol synthase proteins in phylum Chlorophyta (green algae)

(A)

Identity matrix					
Volvox					
Ostreococcus	0.467				
Chlamydomonas	0.715	0.496			
Micromonas	0.510	0.607	0.527		
Chlorella	0.427	0.395	0.464	0.415	

(B)

Z test of selection	*Volvox*	*Ostreococcus*	*Chlamydomonas*	*Micromonas*	*Chlorella*
Volvox		12.04	13.11	17.15	10.47
Ostreococcus	0.00		11.33	19.91	11.38
Chlamydomonas	0.00	0.00		17.02	6.88
Micromonas	0.00	0.00	0.00		15.96
Chlorella	0.00	0.00	0.00	0.00	

(A) Identity matrix: Identity values between sequences being 1, 100% identical. (B) Codon-based Test of Purifying Selection for analysis between sequences: The probability of rejecting the null hypothesis of strict-neutrality ($d_N = d_S$) in favor of the alternative hypothesis ($d_N < d_S$) (above diagonal) is shown. Below the diagonal the p-values are shown (less than 0.05 are considered significant at the 5% level). The test statistic ($d_S - d_N$) is shown below the diagonal. d_S and d_N are the numbers of synonymous and nonsynonymous substitutions per site, respectively. The variance of the difference was computed using the bootstrap method (1000 replicates). Analyses were conducted using the Nei-Gojobori method (Nei & Gojori, 1986). The analysis involved 7 nucleotide sequences. All ambiguous positions were removed for each sequence pair. There were a total of 826 positions in the final dataset.

Suppl Table 1. Diatom cultures investigated in this study.

Species	Sterol composition*	PCR result
Thalassiosira pseudonana CCMP 1335	C_{28} 11(85), 15(6); C_{29} 25(5), 26(4)	+
Thalassiosira gravida CCMP 986	C_{28} 11(95); C_{29} 25(5)	-
Thalassiosira aff antarctica CCAP 1085/9	C_{27} 2(4); C_{28} 11(83); C_{29} 25(3)	-
Fragilaropsis cylindricus CCMP 1102	C_{27} 2(68), 5(32)	+
Pseudo-nitzschia seriata CCMP 1309	C_{27} 2(82), 5(18)	+
Extubocellulus spinifer CCMP 393	C_{27} 2(79), 5(2); C_{28} 10(13), 11(1), 13(1), 14(1); C_{29} 29(3)	+
Extubocellulus cribiger CCAP 1026/1	C_{27} 2(74), 5(4); C_{28} 10(12), 11(4), 13(1), 15(2); C_{29} 29(3)	-
Phaeodactylum tricornutum	C_{28} 10(99), 15(1)	+
Attheya septentrionalis CS 425/03	C_{27} 5(7), 5(2); C_{28} 11(37), 15(19); C_{29} 26(tr), 35(38)	-
Skeletonema costatum CCMP 1281	C_{27} 3(5), 5(8), 8(1); C_{28} 11(73), 18(11); C_{29} 25(1), 26(1)	+
Skeletonema subsalsum CCAP 1077/8	C_{27} 3(10), 5(44); C_{28} 11(24), 15(5); C_{29} 25(tr), 26(11), 35(6)	+
Proboscia indica CCMP 1896	C_{28} 11(90), 18(10)	-
Proboscia inermis CCAP 1064/1	C_{27} 5(44); C_{28} 11(56)	-
Proboscia alata	C_{27} 3(34), 5(29); C_{28} 11(35); C_{29} 24(2)	-

*Data from Rampen et al., 2010.: Numbers in front of parentheses correspond to the sterol numbers as it follows: [2 (cholesta-5,22E-dien-3β-ol); 3 (cholesta-5,24-dien-3β-ol); 5 (cholest-5-en-3β-ol); 8 (5α-cholestan-3β-ol); 10 (24-methylcholesta-5,22E-dien-3β-ol); 11 (24-methylcholesta-5,24(28)-dien-3β-ol); 13 (23-methylcholesta-5,22E-dien-3β-ol); 14 (23-methylcholesta-5,23(28)-dien-3β-ol); 15 (24-methylcholesta-5-en-3β-ol); 18 (24-methylcholesta-24(28)-en-3β-ol); 24 (24-ethylcholesta-5,22E-dien-3β-ol); 25 (24-ethylcholesta-5,24(28E)-dien-3β-ol); 26 (24-ethylcholesta-5,24(28Z)-dien-3β-ol); 29 (23,24-dimethylcholesta-5,22E-dien-3β-ol); 35 (24-ethylcholest-5-en-3β-ol)]. Values in parentheses represent the concentration of the individual sterol, as a percentage of the total sterols. tr indicates relative abundances of <0.5%.

Suppl Table 2. Sterol composition of North Sea SPM and microbial mat samples.

Diatomaceous microbial mat sterol composition	North Sea water column sterol composition
- Cholesta-5,22E-dien-3β-ol - 5α-cholest-22-en-3β-ol - Cholest-5en-3β-ol (cholesterol) - 5α-cholestan-3β-ol (cholestanol) - Cholesta-5,24-dien-3β-ol (desmosterol) - 24-methylcholest-5,22E-dien-3β-ol (brassicasterol/diatomsterol) - 5α(H)-cholest-7en-3β-ol - 24-methylcholest-5,24(28)-dien-3β-ol - 24-methylcholest-5-en-3β-ol (campesterol) - 23,24-dimethylcholesta-5,22E-dien-3β-ol - 24-ethylcholesta-5,22E-dien-3β-ol (stigmasterol) - 24-ethylcholesta-5-en-3β-ol (β-sitosterol) - 5α(H)-23,24-dimethylcholestanol (5αH-C29:0) - 24-ethylcholesta-5,24Z-Zden-3β-ol (isofucosterol) - 4α, 23, 24-trimethyl-5α(H)cholest-22-en-3β-ol (dinosterol)	- Cholesta-5,22E-dien-3β-ol - Cholest-5en-3β-ol (cholesterol) - 5α-cholestan-3β-ol (cholestanol) - Cholesta-5,24-dien-3β-ol (desmosterol) - 24-methylcholest-5,22E-dien-3β-ol (brassicasterol/diatomsterol) - 5α(H)-cholest-7en-3β-ol - 5α(H)-24-methyl-cholest-22en-3β-ol - 24-methylcholest-5,24(28)-dien-3β-ol - 24-methylcholest-5-en-3β-ol (campesterol) - 24-ethylcholesta-5-en-3β-ol (β-sitosterol) - 24-ethylcholesta-5,24Z-Zden-3β-ol (isofucosterol)

Figure 1

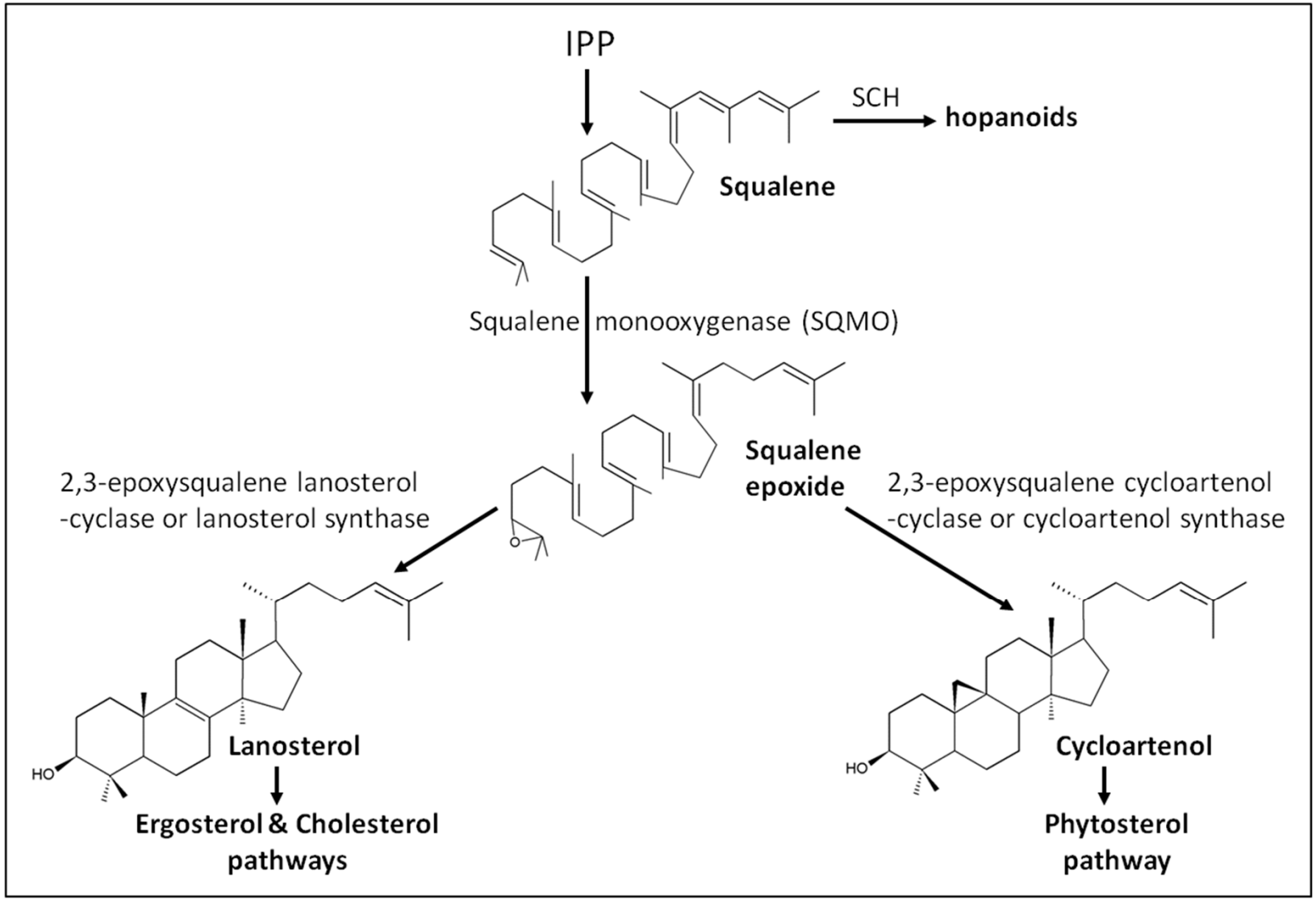

IPP
SCH
hopanoids
Squalene
Squalene monooxygenase (SQMO)
Squalene epoxide
2,3-epoxysqualene lanosterol -cyclase or lanosterol synthase
2,3-epoxysqualene cycloartenol -cyclase or cycloartenol synthase
HO
Lanosterol
Ergosterol & Cholesterol pathways
HO
Cycloartenol
Phytosterol pathway

Figure 2A

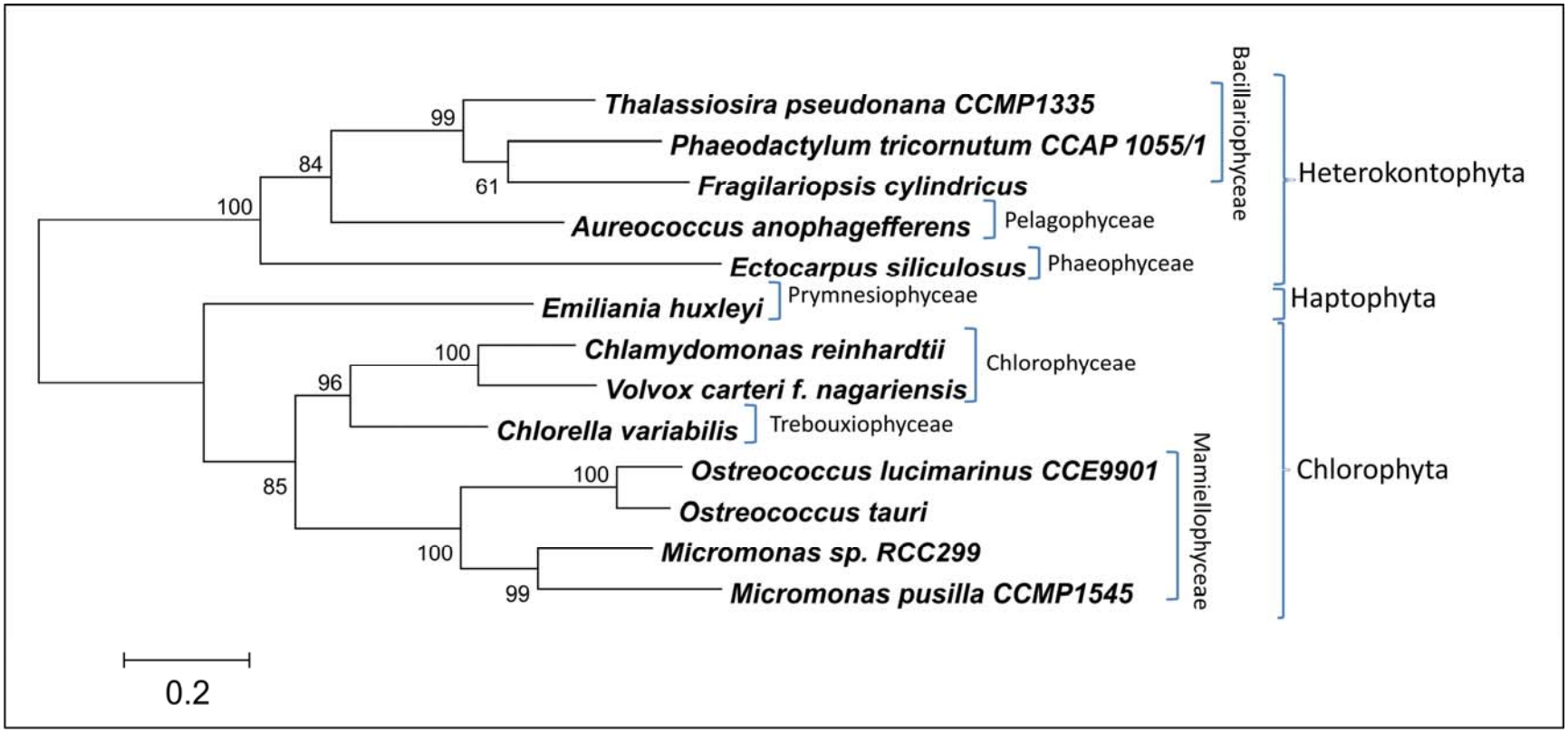

Thalassiosira pseudonana CCMP1335
Phaeodactylum tricornutum CCAP 1055/1
Fragilariopsis cylindricus
Aureococcus anophagefferens
Ectocarpus siliculosus
Emiliania huxleyi
Chlamydomonas reinhardtii
Volvox carteri f. nagariensis
Chlorella variabilis
Ostreococcus lucimarinus CCE9901
Ostreococcus tauri
Micromonas sp. RCC299
Micromonas pusilla CCMP1545
99
84
61
100
100
96
85
100
100
99
Bacillariophyceae
Pelagophyceae
Phaeophyceae
Prymnesiophyceae
Chlorophyceae
Trebouxiophyceae
Mamiellophyceae
Heterokontophyta
Haptophyta
Chlorophyta
0.2

Figure 2B

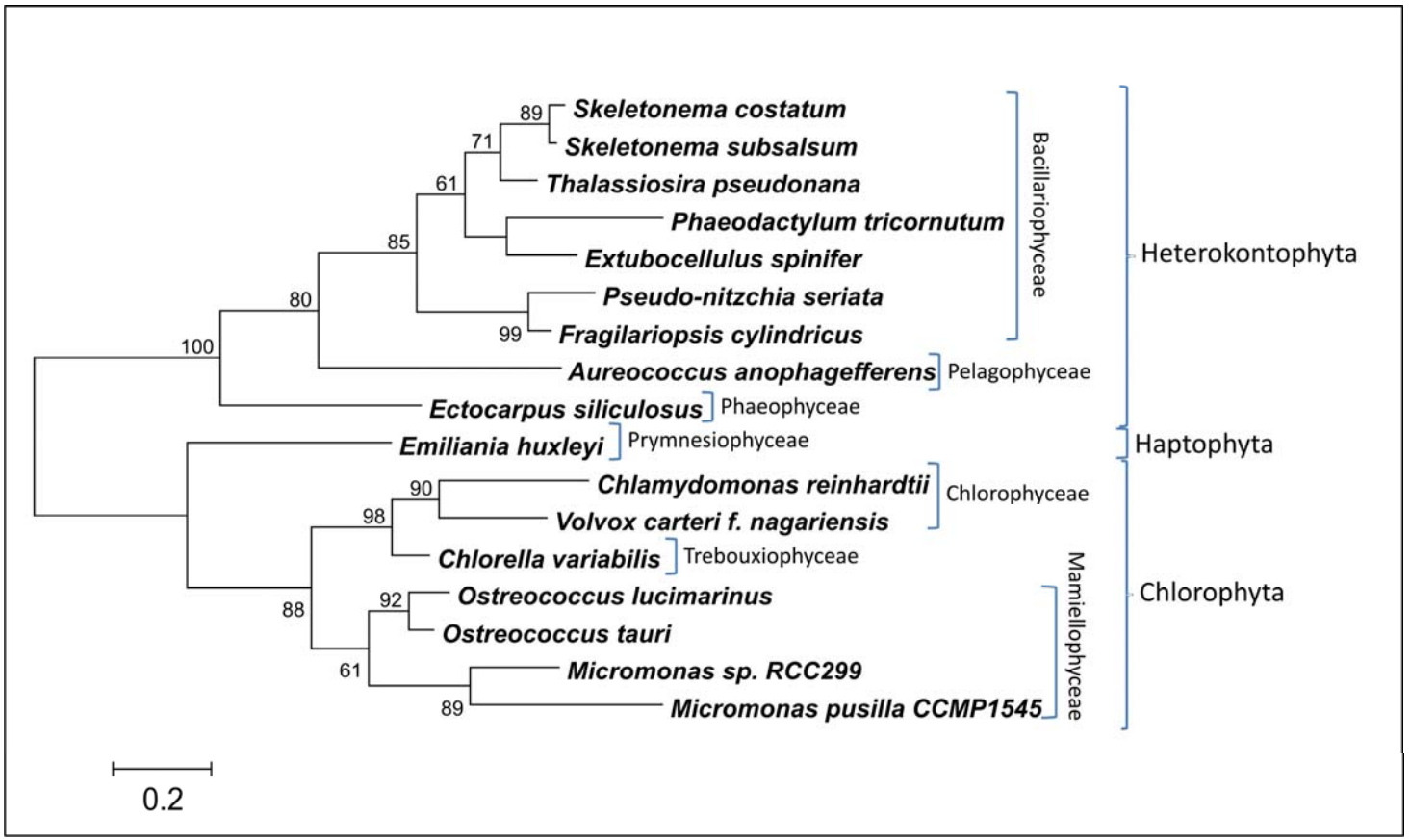

Skeletonema costatum
Skeletonema subsalsum
Thalassiosira pseudonana
Phaeodactylum tricornutum
Extubocellulus spinifer
Pseudo-nitzchia seriata
Fragilariopsis cylindricus
Aureococcus anophagefferens
Ectocarpus siliculosus
Emiliania huxleyi
Chlamydomonas reinhardtii
Volvox carteri f. nagariensis
Chlorella variabilis
Ostreococcus lucimarinus
Ostreococcus tauri
Micromonas sp. RCC299
Micromonas pusilla CCMP1545
89
71
61
85
80
99
100
90
98
92
88
61
89
Bacillariophyceae
Pelagophyceae
Phaeophyceae
Prymnesiophyceae
Chlorophyceae
Trebouxiophyceae
Mamiellophyceae
Heterokontophyta
Haptophyta
Chlorophyta
0.2

Figure 2C

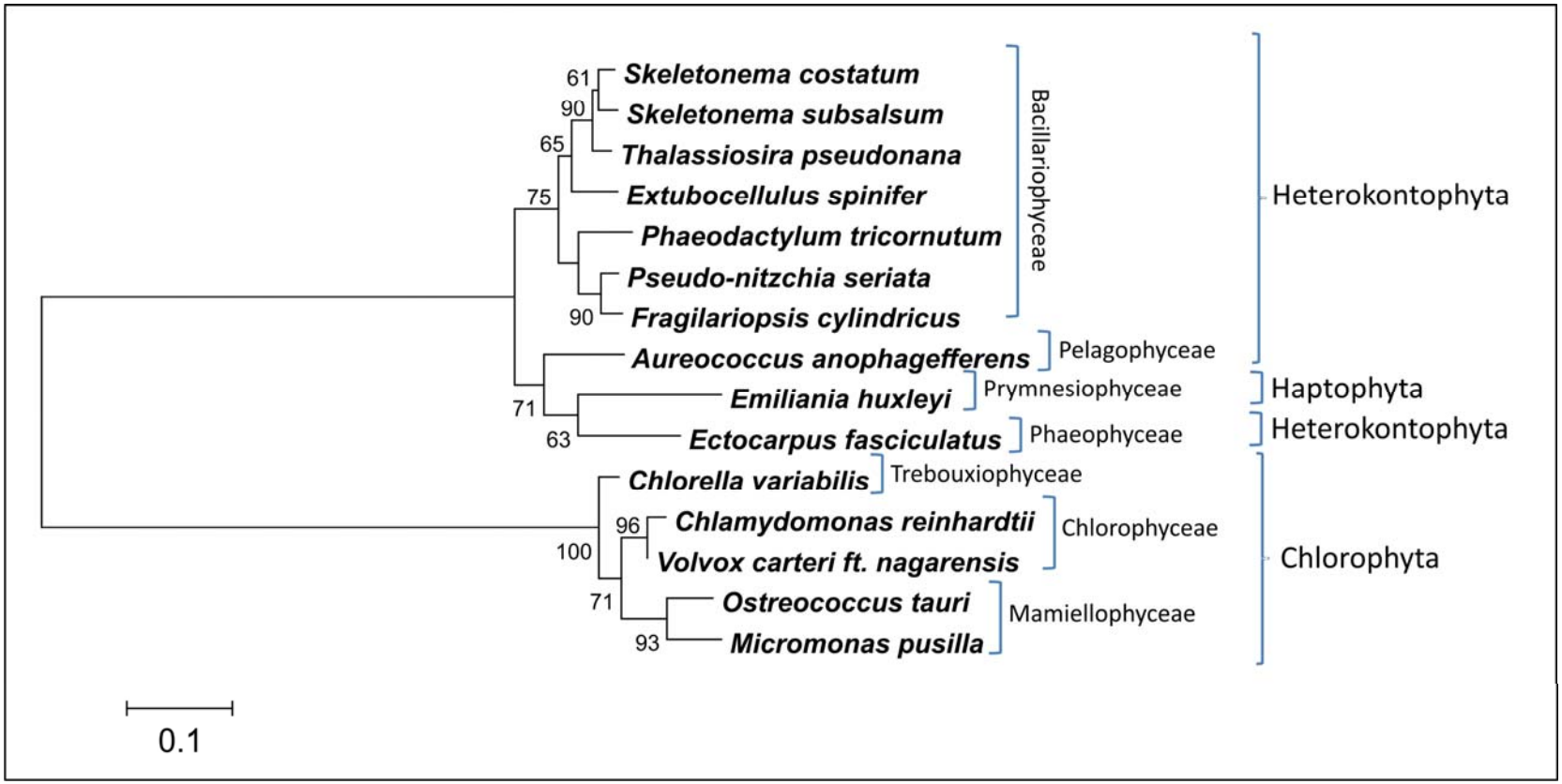
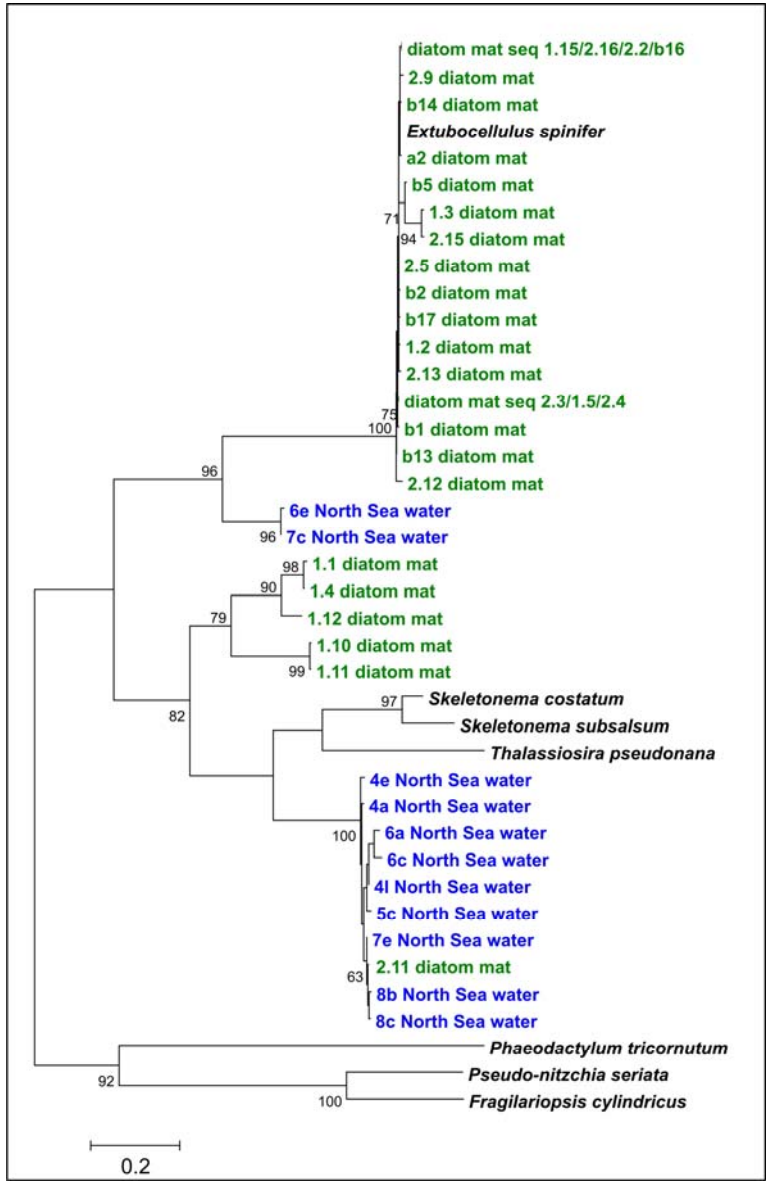

61
90
65
75
90
71
63
96
100
71
93
Skeletonema costatum
Skeletonema subsalsum
Thalassiosira pseudonana
Extubocellulus spinifer
Phaeodactylum tricornutum
Pseudo-nitzchia seriata
Fragilariopsis cylindricus
Aureococcus anophagefferens
Emiliania huxleyi
Ectocarpus fasciculatus
Chlorella variabilis
Chlamydomonas reinhardtii
Volvox carteri ft. nagarensis
Ostreococcus tauri
Micromonas pusilla
Bacillariophyceae
Pelagophyceae
Prymnesiophyceae
Phaeophyceae
Trebouxiophyceae
Chlorophyceae
Mamiellophyceae
Heterokontophyta
Haptophyta
Heterokontophyta
Chlorophyta
0.1

Figure 3


diatom mat seq 1.15/2.16/2.2/b16
2.9 diatom mat
b14 diatom mat
Extubocellulus spinifer
a2 diatom mat
b5 diatom mat
1.3 diatom mat
2.15 diatom mat
2.5 diatom mat
b2 diatom mat
b17 diatom mat
1.2 diatom mat
2.13 diatom mat
diatom mat seq 2.3/1.5/2.4
b1 diatom mat
b13 diatom mat
2.12 diatom mat
6e North Sea water
7c North Sea water
1.1 diatom mat
1.4 diatom mat
1.12 diatom mat
1.10 diatom mat
1.11 diatom mat
Skeletonema costatum
Skeletonema subsalsum
Thalassiosira pseudonana
4e North Sea water
4a North Sea water
6a North Sea water
6c North Sea water
4l North Sea water
5c North Sea water
7e North Sea water
2.11 diatom mat
8b North Sea water
8c North Sea water
Phaeodactylum tricornutum
Pseudo-nitzchia seriata
Fragilariopsis cylindricus
71
94
75
100
96
96
98
90
79
99
97
82
100
63
92
100
0.2

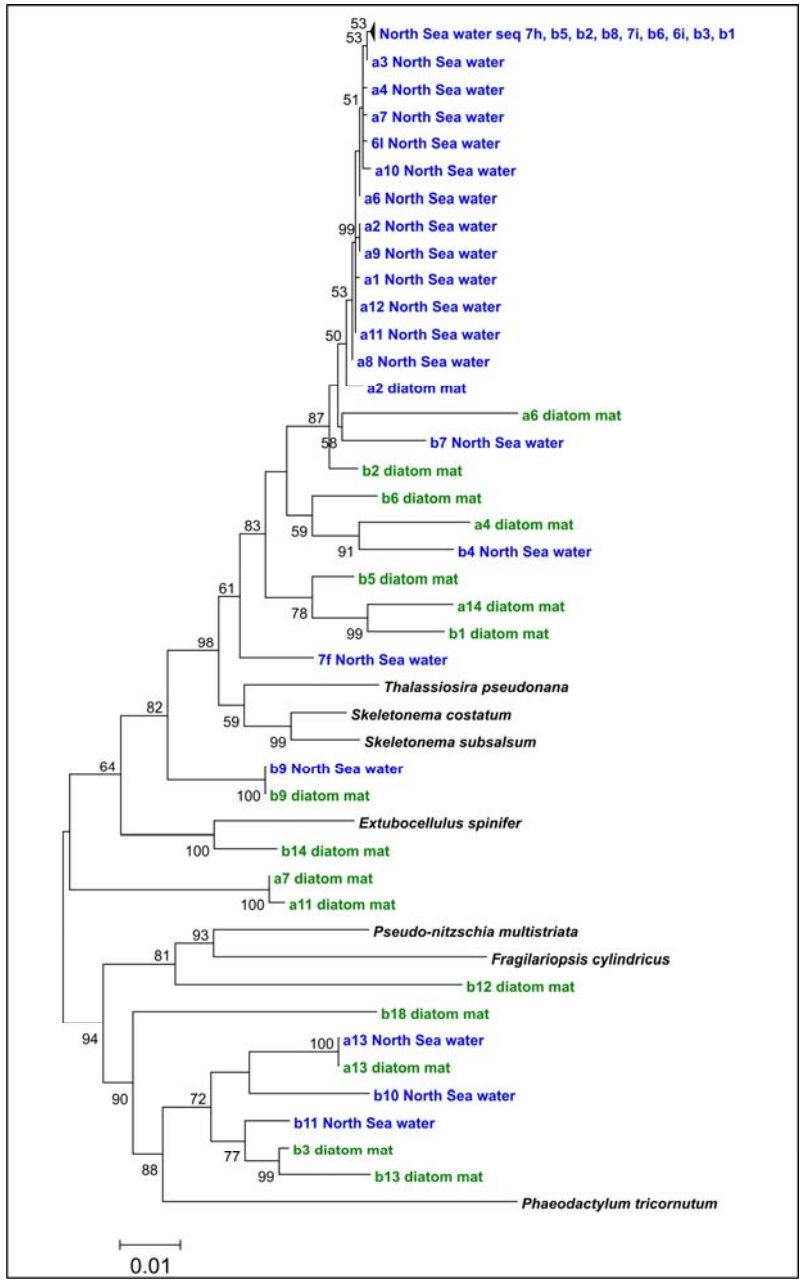
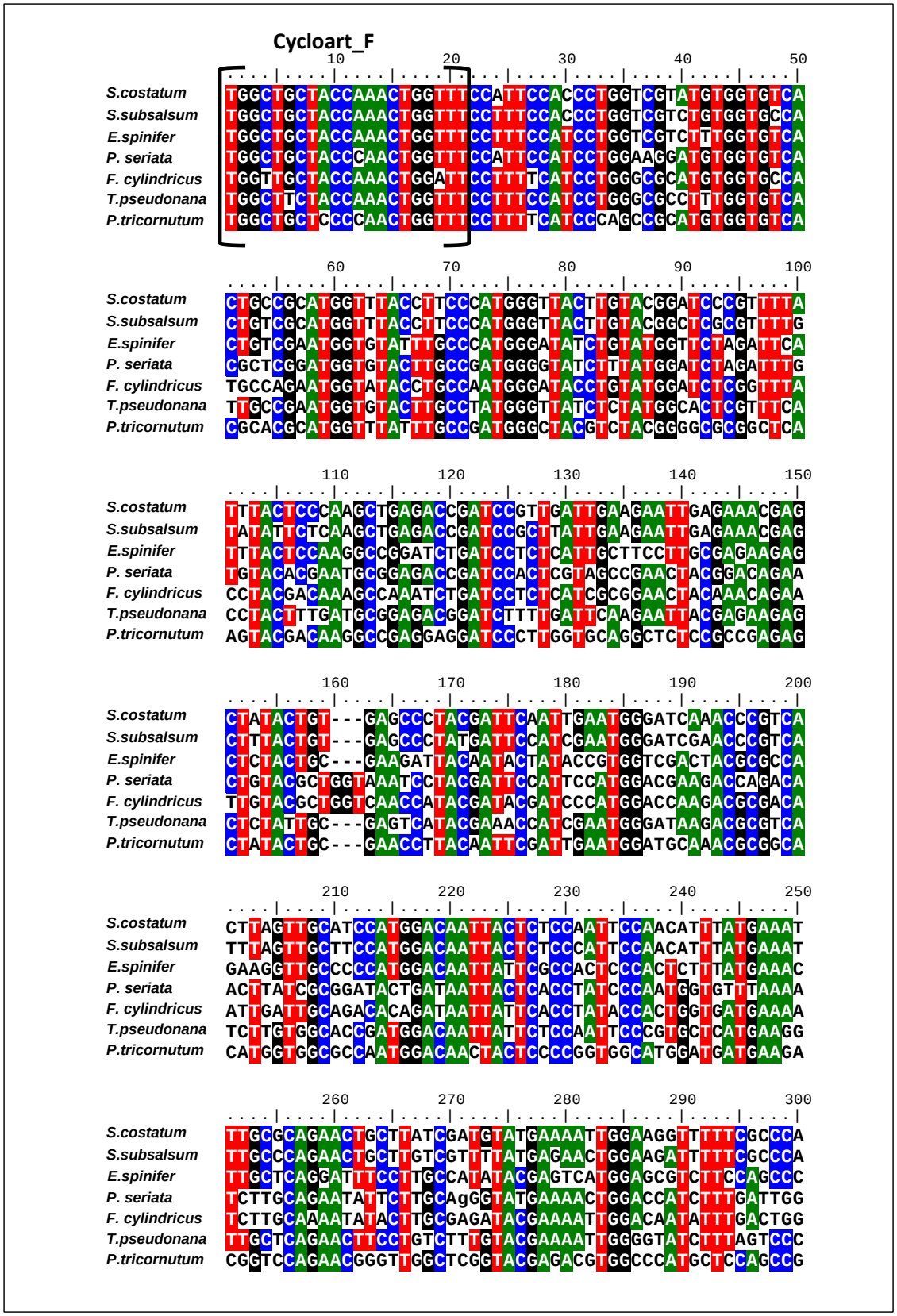
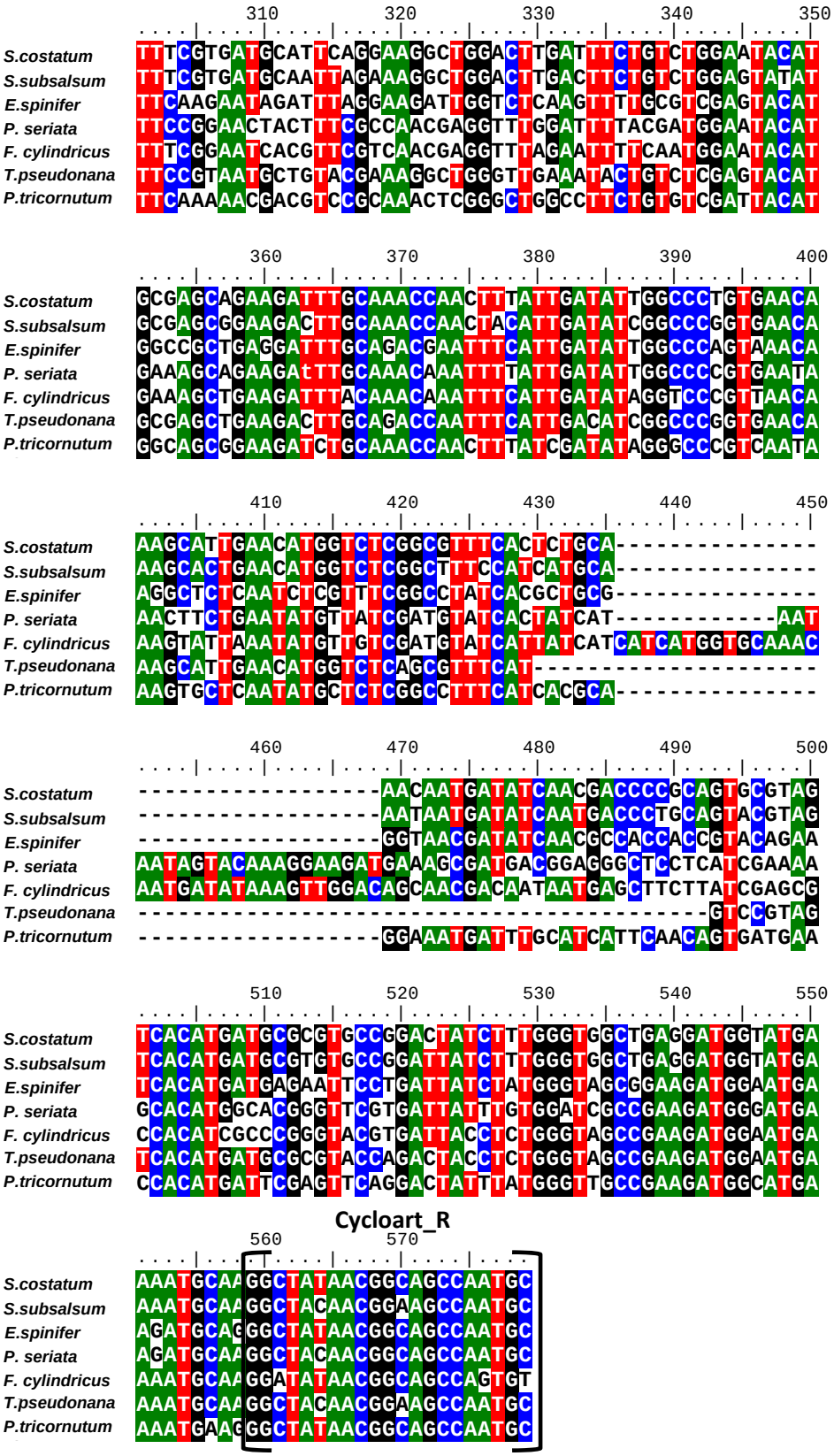

North Sea water seq 7h, b5, b2, b8, 7i, b6, 6i, b3, b1
a3 North Sea water
a4 North Sea water
a7 North Sea water
6l North Sea water
a10 North Sea water
a6 North Sea water
a2 North Sea water
a9 North Sea water
a1 North Sea water
a12 North Sea water
a11 North Sea water
a8 North Sea water
a2 diatom mat
a6 diatom mat
b7 North Sea water
b2 diatom mat
b6 diatom mat
a4 diatom mat
b4 North Sea water
b5 diatom mat
a14 diatom mat
b1 diatom mat
7f North Sea water
Thalassiosira pseudonana
Skeletonema costatum
Skeletonema subsalsum
b9 North Sea water
b9 diatom mat
Extubocellulus spinifer
b14 diatom mat
a7 diatom mat
a11 diatom mat
Pseudo-nitzschia multistriata
Fragilariopsis cylindricus
b12 diatom mat
b18 diatom mat
a13 North Sea water
a13 diatom mat
b10 North Sea water
b11 North Sea water
b3 diatom mat
b13 diatom mat
Phaeodactylum tricornutum
0.01

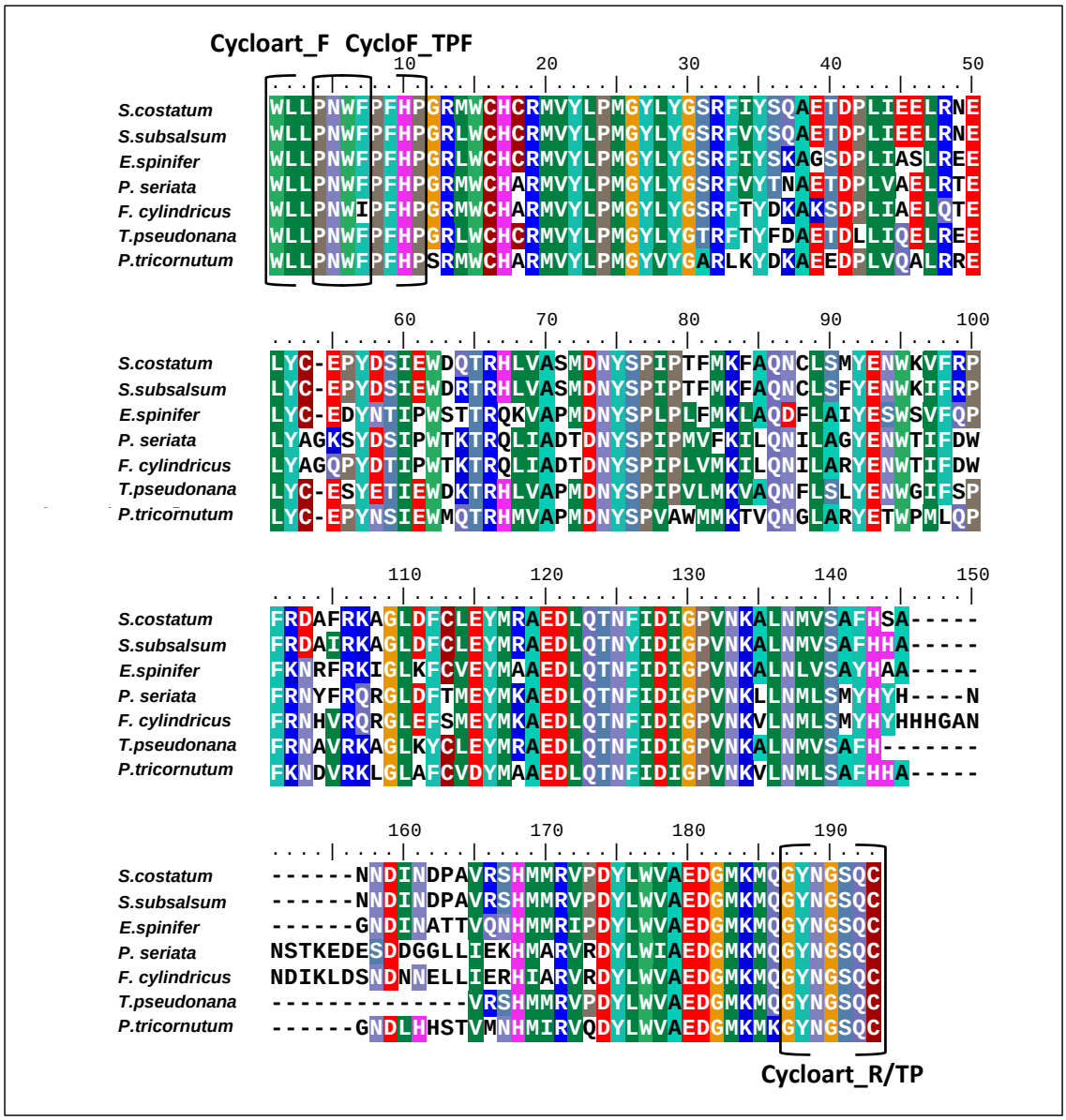
Figure 4

Suppl. Figure 1. Alignment cycloartenol synthase partial gene sequences involved in this analysis




310
320
330
340
350
S.costatum
S.subsalsum
E.spinifer
P. seriata
F. cylindricus
T.pseudonana
P.tricornutum
360
370
380
390
400
410
420
430
440
450
460
470
480
490
500
510
520
530
540
550
Cycloart_R
560
570

Suppl. Figure 2. Protein alignment sequences of partial cycloartenol synthase (CSG) sequences from diatom pure cultures inferred by xblast



Suppl. Figure 3. Phylogenetic tree of 18S rRNA sequences (DNA) in the microalgae under study inferred by using the Maximum Likelihood method based on the GTR model. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+*G*, parameter = 0.2831). The analysis involved 15 nucleotide sequences. All ambiguous positions were removed for each sequence pair. There were a total of 3589 positions in the final dataset.