



## Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome

Richard G Dorrell, Gillian Gile, Giselle Mccallum, Raphaël Méheust, Eric P Baptiste, Christen M. Klinger, Loraine Brillet-Guéguen, Katalina D Freeman, Daniel J Richter, Chris Bowler

### ► To cite this version:

Richard G Dorrell, Gillian Gile, Giselle Mccallum, Raphaël Méheust, Eric P Baptiste, et al.. Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. *eLife*, 2017, 6, pp.e23717. 10.7554/eLife.23717 . hal-01526828

HAL Id: hal-01526828

<https://hal.sorbonne-universite.fr/hal-01526828>

Submitted on 23 May 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

1   **Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid**  
2   **proteome**

3  
4   Richard G. Dorrell<sup>1\*</sup>, Gillian H. Gile<sup>2</sup>, Giselle McCallum<sup>1</sup>, Raphaël Méheust<sup>3</sup>, Eric P. Baptiste<sup>3</sup>,  
5   Christen M. Klinger<sup>4</sup>, Loraine Brillet-Guéguen<sup>5</sup>, Katalina D. Freeman<sup>2</sup>, Daniel J. Richter<sup>6</sup>, and  
6   Chris Bowler<sup>1\*</sup>

7  
8   <sup>1</sup>IBENS, Département de Biologie, École Normale Supérieure, CNRS, Inserm, PSL Research  
9   University, F-75005, Paris, France

10   <sup>2</sup>School of Life Sciences, Arizona State University, 427 E Tyler Mall, Tempe, AZ, 85287, USA

11   <sup>3</sup>Institut de Biologie Paris-Seine, Université Pierre et Marie Curie, Paris

12   <sup>4</sup>Department of Cell Biology, University of Alberta

13   <sup>5</sup>CNRS, UPMC, FR2424, ABiMS, Station Biologique, 29680, Roscoff, France

14   <sup>6</sup>Sorbonne Universités, UPMC Univ Paris 06, CNRS UMR 7144, Adaptation et Diversité en  
15   Milieu Marin, Equipe EPEP, Station Biologique de Roscoff, 29680 Roscoff, France

16  
17   \*To whom correspondence should be addressed: dorrell@biologie.ens.fr,  
18   cbowler@biologie.ens.fr

19   **Abstract**

20  
21  
22   Plastids are supported by a wide range of proteins encoded within the nucleus and  
23   imported from the cytoplasm. These plastid-targeted proteins may originate from the  
24   endosymbiont, the host, or other sources entirely. Here, we identify and characterise 770  
25   plastid-targeted proteins that are conserved across the ochrophytes, a major group of  
26   algae including diatoms, pelagophytes and kelps, that possess plastids derived from red  
27   algae. We show that the ancestral ochrophyte plastid proteome was an evolutionary  
28   chimera, with 25% of its phylogenetically tractable proteins deriving from green algae. We  
29   additionally show that functional mixing of host and plastid proteomes, such as through  
30   dual targeting, is an ancestral feature of plastid evolution. Finally, we detect a clear  
31   phylogenetic signal from one ochrophyte subgroup, the lineage containing pelagophytes  
32   and dictyochophytes, in plastid-targeted proteins from another major algal lineage, the  
33   haptophytes. This may represent a possible serial endosymbiosis event deep in eukaryotic  
34   evolutionary history.

35  
36   **Introduction**

37  
38   Since their origin, the eukaryotes have diversified into an extraordinary array of organisms,  
39   with different genome contents, physiological properties, and ecological adaptations<sup>1-3</sup>.  
40   Perhaps the most profound change that has occurred within individual eukaryotic cells is the  
41   acquisition of plastids via endosymbiosis, which has happened at least eleven times across  
42   the tree of life<sup>1</sup>. All but one characterized group of photosynthetic eukaryotes possess  
43   plastids resulting from a single ancient endosymbiosis of a beta-cyanobacterium by an  
44   ancestor of the archaeoplastid lineage (consisting of green algae and plants, red algae, and  
45   glaucophytes)<sup>1</sup>.

46  
47   Photosynthesis has subsequently spread outside of the archaeoplastids through secondary,  
48   tertiary, or more complex endosymbiosis events. By far the most ecologically successful of  
49   these lineages are those that possess plastids derived from secondary or more complex  
50   endosymbioses of a red alga<sup>1, 4, 5</sup>. These are the "CASH lineages", consisting of

52 photosynthetic members of the cryptomonads, alveolates (such as dinoflagellates),  
53 stramenopiles (also referred to as heterokonts) and haptophytes<sup>1, 4</sup> (see Table 1 and Fig. 1-  
54 figure supplement 1 for definitions). The most prominent of these are the photosynthetic  
55 members of the stramenopiles, termed the ochrophytes<sup>2, 6, 7</sup>. The ochrophytes include the  
56 diatoms, which are major primary producers in the ocean<sup>8, 9</sup>, multicellular kelps, which serve  
57 as spawning grounds for marine animals<sup>10</sup>, and the pelagophytes, small free-living algae  
58 frequently associated with harmful blooms<sup>11</sup> (Fig. 1, panel A; Fig. 1- figure supplement 1).  
59 The stramenopiles also contain many aplastidic and non-photosynthetic lineages (e.g.,  
60 oomycetes), which diverge at the base of the ochrophytes and play important roles as  
61 pathogens and in microbial food webs<sup>6, 12</sup> (Fig. 1- figure supplement 1).

62 Following their acquisition, plastids have undergone a number of evolutionary changes that  
63 bound them more intricately with the biology of the host. These include the transfer of  
64 plastid-derived genes to the host nucleus<sup>3, 13, 14</sup> and the targeting of proteins encoded within  
65 the nucleus to the plastid<sup>15, 16</sup>. Previous studies have shown that many plastid-targeted  
66 proteins are not derived from the endosymbiont genome<sup>17</sup>. Proteins encoded by genes  
67 acquired from other sources, such as laterally acquired genes<sup>18, 19</sup> or previous endosymbiotic  
68 organelles historically possessed by the host<sup>20, 21</sup>, or proteins that have been repurposed  
69 from endogenous host organelles<sup>22, 23</sup> have important roles in supporting the biology of  
70 plastid lineages. Other gene transfer events, e.g. from food sources<sup>24</sup>, bacterial symbionts<sup>25</sup>,  
71 viruses<sup>26</sup>, or diazotrophic non-plastid cyanobacterial endosymbionts<sup>27, 28</sup> have also played  
72 major roles in the evolution of photosynthetic eukaryotes, and it remains to be determined  
73 which of these have contributed to the diverse range of plastid proteins observed today. It  
74 nonetheless remains largely unknown which proteins had the most fundamental roles in  
75 establishing current plastid lineages<sup>3</sup>, i.e., which plastid proteins represent the ancestral  
76 components of plastid-targeted proteomes.

77  
78 Ochrophytes represent an excellent system in which to reconstruct the origins of plastid  
79 proteomes. Firstly, plastid-targeting sequences in different ochrophytes are relatively well  
80 conserved, enabling *in silico* prediction of plastid-targeted proteins from a wide range of  
81 different species<sup>29, 30</sup>, in contrast to plastid-targeting sequences within archaeoplastid  
82 lineages, which are extremely variable<sup>31, 32</sup>. Secondly, compared to other CASH lineages  
83 (haptophytes, cryptomonads, and dinoflagellates), ochrophytes represent an extremely well  
84 characterised system for experimental and bioinformatic investigation, with (to date) eleven  
85 complete genomes, and transcriptome libraries available for over 150 species through  
86 MMETSP<sup>33, 34</sup>. Reliable transformation and other manipulation strategies are also available  
87 for multiple species, such as the model diatom *Phaeodactylum tricornutum*<sup>35-37</sup>.

88  
89 Thirdly, the origin of the ochrophyte plastid is an evolutionarily valuable topic to  
90 understand. It is currently not known when the ochrophyte plastid was acquired: whether it  
91 originated recently, predates the radiation of aplastidic stramenopile relatives<sup>5, 6, 12</sup>, or was  
92 acquired prior to the divergence of stramenopiles from their closest relatives, the  
93 alveolates<sup>38</sup>. Verifying a late origin for the ochrophyte plastid would thus enable insights into  
94 the cellular changes that accompany the transition from a solely heterotrophic to a  
95 phototrophic lifestyle<sup>6, 12</sup>, which is currently not possible for archaeoplastids<sup>39, 40</sup>, and difficult  
96 for haptophytes and cryptomonads, in which these relatives respectively remain unknown  
97 or understudied at a genomic level<sup>39, 41</sup>. It has additionally been proposed, based on the  
98 presence of large numbers of genes of putative green algal origin in diatom genomes<sup>42, 43</sup>,  
99 that the ancestor of ochrophytes once possessed a green algal endosymbiont, which was  
100 subsequently replaced via the serial endosymbiosis of a red algal-derived plastid<sup>1, 44</sup>. This  
101 hypothesis remains controversial<sup>45-47</sup>, in particular due to issues associated with the

102

103 distinction of genes of red and green algal origins in ochrophyte genomes<sup>48-50</sup>. A final  
104 evolutionary suggestion regarding ochrophytes is that they have acted as endosymbiotic  
105 donors into other CASH lineages. One recent study proposed that haptophytes possess  
106 plastids acquired via the endosymbiosis of an ochrophyte<sup>5</sup>, although the exact identity of  
107 this endosymbiotic acquisition remain unresolved. Characterising the ancestral ochrophyte  
108 plastid proteome might therefore help answer major questions about the ways in which  
109 plastids become established in the host cell, and provide valuable insights into the origins  
110 and diversification of other ecologically important algal lineages.

111  
112 In this study, we present an experimentally verified *in silico* reconstruction of the proteins  
113 targeted to the plastid of the last common ochrophyte ancestor. We show that this ancestral  
114 plastid proteome was an evolutionary mosaic, containing 770 proteins from a range of  
115 different sources. Our dataset indicates that the ochrophyte plastid was acquired late in  
116 stramenopile evolution, following the divergence of extant aplastidic relatives, that plastid-  
117 targeted proteins of green algal origin played a significant role in its origin, and that there  
118 has been bidirectional integration of the biology of the ochrophyte host and plastid  
119 proteomes, such as the ancient recruitment of proteins from both host and endosymbiont  
120 to dually support the biology of the plastid and mitochondria. Finally, we show evidence for  
121 an ancient endosymbiosis of a specific ochrophyte lineage, an ancestor of the pelagophytes  
122 and dictyochophytes, by a common ancestor of the haptophytes, which we propose- based  
123 on discrepancies between the origins of the haptophyte plastid proteome and genome-  
124 reveals a possible serial endosymbiosis event early in haptophyte evolution, preceding the  
125 origins of the current haptophyte plastid. Our work resolves several long-standing questions  
126 of ochrophyte evolution, and provides new insights into the origins and diversification of  
127 CASH lineages as a whole.

128  
129 **Results**  
130

131 **1. *In silico* reconstruction of an ancestral plastid proteome**  
132

133 We developed an *in silico* pipeline for identifying putatively ancestral plastid-targeted  
134 proteins across the ochrophytes (Fig. 1). We screened a large composite library, comprising  
135 eleven different ochrophyte genomes, together with transcriptome data from a further 158  
136 ochrophyte species (Table S1- sheet 1<sup>145</sup>) using the ochrophyte plastid targeting predictors  
137 ASAFind (Table S2- sheet 1<sup>145</sup>)<sup>29</sup> and HECTAR (Table S3- sheet 1<sup>145</sup>)<sup>30</sup>. Sequences with  
138 predicted plastid localisation were binned into eleven taxonomic sub-categories within three  
139 major groups (chrysista, hypogyrista, and diatoms) based on recent multigene phylogenies<sup>12</sup>  
140 (Fig. 1, panel A; Fig. 1- figure supplement 1), then assembled by sequence similarity into  
141 homologous plastid-targeted protein groups (HPPGs, Materials and Methods).

142  
143 We next tested the level of conservation best able to identify truly ancestral HPPGs. We  
144 selected three patterns of conservation that identified the largest number of HPPGs from a  
145 positive control dataset of proteins with previously identified plastid-associated functions,  
146 and minimised the number identified from a negative control dataset of HPPGs generated  
147 using seed sequences from three other published CASH lineage genomes, for which no  
148 plastid-targeted orthologues were detected in any ochrophyte genome sequence (Materials  
149 and Methods; Table S2- sheet 2, sections 1-2; Table S3- sheet 2, sections 1-2<sup>145</sup>). The  
150 selected conservation patterns were: the presence of the protein in a majority of chrysistan  
151 sub-categories and a majority of either diatom or hypogyristeal sub-categories; or presence  
152 in at least one chrysistan sub-category and a majority of both diatoms and hypogyristeal (Fig.  
153 1, panel B). We extracted HPPGs matching the conservation patterns defined above and

154 verified their monophyly within ochrophytes via alignment and single-gene trees (Fig. 1,  
155 panel C; Table S4- sheet 1<sup>145</sup>). From this, we identified 770 proteins that were probably  
156 targeted to the ancestral ochrophyte plastid (Fig. 1, panel D; Table S4- sheet 2<sup>145</sup>). This  
157 dataset is significantly enriched in proteins from within the positive control dataset and  
158 contains significantly fewer proteins from the negative control dataset than would be  
159 expected through random assortment (chi-squared test,  $P < 1 \times 10^{-10}$ ; Fig. 1), confirming its  
160 specificity towards probable ancestral plastid-targeted proteins.

161

## 162 2. Experimental verification of ancestral ochrophyte HPPGs

163

164 We wished to verify that the ancestral ochrophyte plastid-targeted proteins inferred from  
165 the *in silico* pipeline are genuinely plastid-targeted. 106 of our inferred ancestral HPPGs  
166 include a *P. tricornutum* protein with prior experimental plastid localization, or unambiguous  
167 plastid function (Fig. 1, panel D), but the remainder do not. We selected ten proteins for  
168 experimental localisation (Fig. 2, panel A; Table S5<sup>145</sup>). These were chosen on the basis of  
169 having only non-plastid annotations on the first 50 BLAST hits against the NCBI nr database  
170 excluding ochrophytes, thus arguing against their predicted plastid localization beyond these  
171 organisms. In each case, all of the ochrophyte protein sequences within the alignment had a  
172 well conserved central domain, and a highly variable N-terminal domain of between 30 and  
173 50 amino acids containing an ASAFAF motif, consistent with a conserved plastid targeting  
174 sequence<sup>29</sup> (Fig. 2- figure supplement 1).

175 The selected proteins included five aminoacyl-tRNA synthetases that yielded BLAST top hits  
176 only against enzymes with cytoplasmic annotations, or of probable prokaryotic origin (Fig. 2-  
177 figure supplement 2). Also included were a GroES-type chaperonin of inferred mitochondrial  
178 origin, an Hsp90-type chaperonin of inferred endoplasmic reticulum origin and a  
179 pyrophosphate-dependent phosphofructokinase, which is related to cytosolic enzymes from  
180 other lineages (Fig. 2- figure supplement 3), and is distinct from the ATP-dependent  
181 phosphofructokinases used by primary plastid lineages<sup>51</sup>. The Mpv17 membrane protein is  
182 most closely related to enzymes with peroxisomal functions and localisation<sup>52, 53</sup>, but lacks  
183 any identifiable peroxisomal targeting sequence (PSL, KRR, or a PTS1 motif)<sup>54</sup> in its C-  
184 terminus. Novel protein 1 lacks any conserved domains, and yielded no BLAST matches  
185 outside of the ochrophytes below an expect value of  $1 \times 10^{-05}$  (except for one dinoflagellate  
186 sequence), and hence might constitute an entirely novel plastid-targeted protein (Fig. 2-  
187 figure supplement 4; Table S5<sup>145</sup>).

188 We generated C-terminal GFP-fusion constructs for each of these proteins using *P.*  
189 *tricornutum* genes and transformed wild-type *P. tricornutum* (Fig. 2, panel B; Fig. 2- figure  
190 supplement 5; Table S5<sup>145</sup>). In each case, we identified GFP fluorescence associated with the  
191 plastid. In one case (the peroxisomal membrane protein; Fig. 2, panel B), the GFP  
192 accumulated in a ring around the plastid equator, consistent with a periplastid compartment  
193 (PPC) localisation<sup>88, 55</sup>. In other cases (such as the five aminoacyl-tRNA synthetases, Fig. 2-  
194 figure supplement 5), the GFP signal localised both within and external to the plastid,  
195 consistent with a multipartite localisation within the cell. However, in all cases the proteins  
196 tested were at least partially targeted to the plastid.

197 We additionally generated heterologous GFP fusion constructs for five of the proteins using  
198 sequences from the "dinotom" *Glenodinium foliaceum*, a dinoflagellate alga that harbours  
199 permanent endosymbionts of diatom origin<sup>20, 56</sup>, and the eustigmatophyte *Nannochloropsis*  
200 *gaditana*, which as a member of the "PESC clade" is distantly related to *P. tricornutum* on  
201 the ochrophyte tree<sup>12</sup>. We expressed these constructs in *P. tricornutum* (Fig. 2, panel B; Fig.  
202 2- figure supplement 6), and, in each case, detected plastid-localized GFP fluorescence

203 similar to the patterns observed with the *P. tricornutum* gene constructs. Overall, our data  
204 therefore supports that the ancestral HPPG dataset consists of genuinely conserved plastid-  
205 targeted proteins, rather than misidentified proteins of non-plastid function.

206

### 207 **3. Evolutionary origins of the ochrophyte plastid**

208

#### 209 *The ochrophyte plastid is an evolutionary mosaic*

210

211 We wished to identify the evolutionary affinity of each ancestral HPPG in our dataset. In  
212 particular, we assessed whether proteins that are of unconventional origin, such as the  
213 products of genes endogenous to the host, or genes that have been acquired from other  
214 sources such as prokaryotes and green algae, have significantly contributed to the origins of  
215 the ochrophyte plastid<sup>1, 44</sup>.

216

217 We accordingly determined the closest relative of each ancestral HPPG (Materials and  
218 Methods). Due to ongoing controversies regarding the evolutionary composition of  
219 ochrophyte genomes<sup>46, 47</sup>, we utilised a combined phylogenetic and BLAST top hit approach  
220 to robustly infer the most probable origin of each HPPG (Materials and Methods; Table S4-  
221 sheet 2<sup>45</sup>). For both the BLAST and phylogenetic analyses, stringent criteria were applied to  
222 avoid misidentification due to topological ambiguity, or contamination within individual  
223 sequence datasets<sup>57, 58</sup> (Materials and Methods). We took the union of these two analyses to  
224 produce a dataset of 263 HPPGs for which both phylogenetic and BLAST top hit analyses  
225 indicated the same clear evolutionary origin. These origins were grouped into six  
226 evolutionary categories, red algae, green algae, aplastidic stramenopiles, other eukaryotes,  
227 prokaryotes, and viruses (Fig. 3, panel A).

228

229 Of the 263 HPPGs that were resolved from the combined analysis, 149 (57%) were of red  
230 algal, i.e. endosymbiont origin (Fig. 3, panel A; Table S4- sheet 3<sup>45</sup>). This is analogous to  
231 results from studies of archaeoplastid plastid proteomes, in which approximately half of the  
232 plastid-targeted proteins are of endosymbiont origin<sup>18, 32</sup>. The remaining 114 HPPGs resolved  
233 with other sister-groups, consistent with a mosaic origin of the ochrophyte plastid  
234 proteome. The most significant of these lineages was green algae (67 HPPGs, 25%), followed  
235 by aplastidic stramenopiles (26 HPPGs, 10%), and prokaryotes (21 HPPGs, 8%) (Fig. 3, panel  
236 A). None of the HPPGs were clearly assigned to other eukaryotes or to viruses, consistent  
237 with previous assertions that these lineages have contributed very little to ochrophyte  
238 evolution<sup>59</sup> (Fig. 3, panel A).

239

#### 240 *Late origin of ochrophyte plastids*

241

242 We wished to determine whether the ochrophyte plastid was acquired by a common  
243 ancestor of all stramenopiles or later in ochrophyte evolution. We reasoned that if the  
244 ochrophyte plastid was acquired early, i.e., before the divergence of aplastidic relatives,  
245 endosymbiotic gene transfer from the red algal symbiont to the host nucleus would have  
246 commenced prior to the radiation of the stramenopiles<sup>60</sup>. Based on the primary evolutionary  
247 affinities of each ancestral HPPG (Fig. 3, panel A), we would expect at least half of the  
248 aplastidic stramenopile-derived proteins to show a deeper red algal origin. We accordingly  
249 profiled the deeper evolutionary affinity of each ancestral HPPG of aplastidic stramenopile  
250 origin by a combined phylogenetic and BLAST top hit analysis, as before.

251

252 First, we noted that the majority (20/26) of the ochrophyte HPPGs with aplastidic  
253 stramenopile origins specifically resolved as a sister-group to oomycetes, as opposed to the

254 deeper-branching labyrinthulomycetes or slopalinids (Fig. 3, panel B; Table S4- sheet 3<sup>145</sup>).  
255 Because oomycetes are the sister-group of ochrophytes<sup>6,12</sup>, this suggests that our dataset  
256 retains useful phylogenetic signal.

257  
258 Next, from the 26 ancestral HPPGs of aplastidic stramenopile origin, we identified a clear  
259 sister-group to the stramenopile clade for 16 HPPGs using BLAST, and for 18 HPPGs using  
260 single-gene trees (Fig. 3, panel B). However, only one BLAST top hit and four trees showed a  
261 deeper red algal affinity (Fig. 3, panel B). These proportions are significantly smaller than the  
262 proportions of ochrophyte proteins of red origin in the entire ancestral HPPG dataset  
263 (expected frequencies: 9.54 BLAST top hits, 10.7 sister-groups; chi-squared-test, P≤ 0.01; Fig.  
264 3, panels A, B). In five cases we identified the same deeper affinity through combined BLAST  
265 top hit and tree sister-group analysis, but none of these were of red algal origin (Fig. 3, panel  
266 B). We conclude that plastid-targeted proteins in ochrophytes that are related to aplastidic  
267 stramenopile proteins are predominantly not of red origin. This is consistent with a late  
268 origin for the ochrophyte plastid, following the divergence of the ochrophytes and  
269 oomycetes.

270

#### 271 *A significant green algal contribution to ochrophyte plastid evolution*

272

273 Previous reports of green genes in ochrophyte genomes have been controversial due to a  
274 paucity of red algal sequence data<sup>44, 47, 59</sup>. We were able to avail in our pipeline of sequence  
275 information from five complete red algal genomes<sup>48, 49, 61-63</sup> and twelve red algal  
276 transcriptomes<sup>34, 64</sup>, allowing us to more clearly infer the reliability of the green signal in  
277 ochrophytes. We tested whether the inferred green algal origin could be due to a protein  
278 family's absence from red algal lineages (Fig. 4, panel A). For the majority of our green  
279 HPPGs (40/67), an orthologue was identified in at least four of the five major red algal sub-  
280 categories considered (cyanidiales, bangophytes and florideophytes, compsopogonophytes  
281 and stylonematophytes, porphyridiophytes, and rhodellophytes; Fig. 4, panel B; Fig. 4- figure  
282 supplement 1; Table S4- sheet 4<sup>145</sup>). We therefore conclude that these green genes were not  
283 misidentified as the result of undersampling within red sequence libraries, or secondary  
284 gene loss events in the red algae<sup>45, 50</sup>.

285

286 We then considered whether the green genes in our dataset originate from a specific source  
287 within the green algae. Phylogenetic analyses of the HPPGs of verified green origin exhibited  
288 a strong bias toward chlorophyte origins. Ochrophytes branched as sister-groups to  
289 individual or multiple chlorophyte lineages in 51 of the 67 trees (Fig. 4, panel C; Fig. 4- figure  
290 supplement 2). Similarly, we noted a strong predominance of chlorophyte lineages amongst  
291 BLAST top hits (56/67) despite the fact that these lineages only correspond to approximately  
292 25% of the green sequences present in our libraries (Fig. 4- figure supplement 3; Table S4-  
293 sheet 3<sup>145</sup>). In contrast, only 16 of the single-gene trees recovered a sister-group relationship  
294 between ochrophytes and all green lineages (chlorophytes and streptophytes), none  
295 recovered a specific sister-group relationship between ochrophytes and streptophytes (Fig.  
296 4, panel C), and only 11 of the BLAST top hits were to streptophyte sequences (Fig. 4- figure  
297 supplement 2; Table S4- sheet 3<sup>145</sup>). This bias is inconsistent with the green ancestral HPPGs  
298 being of misidentified red origin, or originating at a deeper position within the green algae,  
299 in which case they should show a more stochastic distribution of evolutionary affinities  
300 across all green lineages<sup>46</sup>.

301

302 Next, we tested whether our data supported a single origin for the green genes within the  
303 chlorophytes, or whether the HPPGs of green origin arose through gene transfer events  
304 from multiple chlorophyte lineages. We identified all amino acids that were uniquely shared

305 between ochrophytes and chlorophytes in the 31 green HPPGs for which we found no  
306 evidence of gene duplication or subsequent lateral gene transfer into green algae,  
307 ochrophytes, or other major photosynthetic eukaryotes (Table S6- sheets 1, 2<sup>145</sup>; Materials  
308 and Methods). We then inferred the most probable origin in the green algal tree for each  
309 uniquely shared residue as well as the earliest possible origin, taking into account gapped  
310 and missing positions (Fig. 4, panel D; Fig. 4- figure supplement 4; Table S7- sheets 1, 3<sup>145</sup>). In  
311 both analyses the majority of the uniquely shared residues were inferred to have originated  
312 in a common ancestor of all chlorophytes, or of all chlorophyte lineages excluding the basal  
313 *Prasinoderma/Nephroselmis* sub-category (189/289 positions in observed analysis; 100/147  
314 positions in the earliest possible analysis; Fig. 4, panel D; Fig. 4- figure supplement 4; Table  
315 S7- sheets 1, 3<sup>145</sup>). All other nodes within the green tree, including all specific green sub-  
316 categories, shared much smaller numbers of residues with ochrophytes (Fig. 4, panel D; Fig.  
317 4- figure supplement 4; Table S7- sheets 1, 3<sup>145</sup>). Thus, our data is congruent with the  
318 majority of the ochrophyte green genes originating from deep within the chlorophyte  
319 lineage.

320  
321 Finally, we considered whether the green genes that function in ochrophyte plastids were  
322 more likely to have been acquired through endosymbiosis, or through lateral gene transfers,  
323 for example from a food organism<sup>65, 66</sup> or other intracellular symbiont<sup>3</sup>. We reasoned that if  
324 the green genes in ochrophytes were predominantly of endosymbiotic origin, they should  
325 encode more plastid-targeted proteins than genes of alternative origin, in the same manner  
326 as genes of cyanobacterial origin retained in archaeoplastid genomes are biased towards  
327 encoding proteins with plastid functions<sup>20</sup>. We accordingly constructed a secondary dataset,  
328 consisting of 7140 non-redundant gene families that are broadly distributed across the  
329 ochrophytes, and tested the targeting preferences of proteins from each HPPG (Fig. 4, panel  
330 E; Fig. 4- figure supplement 5; Table S8- sheet 1<sup>145</sup>). 871 gene families resolved with the  
331 green algae per BLAST top hit analysis (Fig. 4- figure supplement 6; Table S8- sheet 2<sup>145</sup>).  
332 Using both ASAFind<sup>29</sup> and HECTAR<sup>30</sup>, gene families of predicted green algal origin were  
333 significantly more likely to encode proteins with plastid-targeting predictions than the  
334 dataset as a whole (chi-squared, P < 1E<sup>-03</sup>; Fig. 4, panel E; Fig. 4- figure supplement 5; Table  
335 S8- sheet 3<sup>145</sup>). We also observed a similar, though stronger, bias towards plastid-targeted  
336 proteins among the proteins of red algal origin (chi-squared, P < 1E<sup>-40</sup>; Fig. 4, panel E; Fig. 4-  
337 figure supplement 5; Table S8- sheet 3<sup>145</sup>). Collectively, our data support the presence of  
338 genes of chlorophyte origin in the last common ochrophyte ancestor, the majority of which  
339 have predicted plastid localisations, consistent with an acquisition through a plastid  
340 endosymbiosis event.

341

#### 342 **4. Functional consequences of mosaic origins for the ochrophyte plastid**

343

##### 344 *Metabolic completeness of the ochrophyte plastid*

345

346 We identified effectively complete core plastid metabolism pathways within the ancestral  
347 HPPG dataset (Fig. 5, panel A; Fig. 5- figure supplement 1; Table S9- sheet 1<sup>145</sup>). The majority  
348 of the remaining proteins remain plastid-encoded in some ochrophyte lineages, or are  
349 dispensible for the metabolic pathway (Fig. 5- figure supplements 1, 2)<sup>67-69</sup>. In four cases  
350 (isopropylmalate synthase, sedoheptulose bisphosphatase, 3-dehydroquinate synthase, and  
351 shikimate kinase) lateral gene transfer and replacement events have occurred into individual  
352 ochrophyte lineages since their radiation, preventing identification of a single HPPG within  
353 the ancestral dataset (Fig. 5, panel A; Fig. 5- figure supplements 2-6). Taking these  
354 exceptions into account, we conclude that the ancestral ochrophyte plastid proteome  
355 contained the fundamental components of core plastid metabolism.

356

357 *Mosaic origins of ochrophyte plastid metabolism*

358

359 Given the mosaic evolutionary origins of ancestral ochrophyte plastid-targeted proteins, we  
360 wondered whether certain evolutionary affinities might correlate with specific metabolic  
361 functions. It has previously been speculated, for example, that genes acquired by diatoms  
362 from green algae might have a specific role in tolerating variable light regimes<sup>42, 70, 71</sup> or  
363 eliminating toxic substances from diatom plastids<sup>72</sup>. We noted that many of the pathways in  
364 the ochrophyte plastid utilise a mixture of genes of red, green, host and prokaryotic origin  
365 (Fig. 5- figure supplement 1), which would suggest a converse scenario: that the mosaic  
366 origins of the ochrophyte plastid have led to the functional mixing of enzymes with disparate  
367 evolutionary origins.

368

369 Consistent with this latter idea, we found very little evidence that individual categories of  
370 HPPG (i.e., red algal, green algal, prokaryotic or host origin) are associated with particular  
371 KOG annotations, as inferred by chi-squared testing ( $P < 0.05$ ) against a null hypothesis that  
372 all KOG families and classes are homogenously distributed across the ancestral HPPG  
373 dataset, independent of evolutionary origin (Fig. 5, panel B; Fig. 5 – figure supplement 7;  
374 Table S9- sheet 2<sup>145</sup>). The notable exceptions are prokaryotic HPPGs being elevated in  
375 information storage and processing proteins, particularly those involved in translation, while  
376 HPPGs of host origin were enriched in proteins involved in cellular processes and signalling  
377 relative to the ancestral HPPG set as a whole (Fig. 5, panel B; Fig. 5 – figure supplement 7;  
378 Table S9- sheet 2<sup>145</sup>). In contrast, several KOG categories were more highly represented in  
379 the ancestral HPPG set than in HPPGs as a whole (Fig. 5, panel B; Fig. 5 – figure supplement  
380 7; Table S9- sheet 2<sup>145</sup>).

381

382 A related question is whether proteins that catalyse adjacent steps of a biochemical  
383 pathway tend to have shared or different evolutionary affinities. Multiple sets of non-native  
384 proteins might be preferentially utilised by ochrophyte plastids, over homologous proteins  
385 of endosymbiont origin, due to performing concerted steps in individual metabolic pathways  
386 or cellular processes<sup>1, 42, 73</sup>. In this instance, pairs of proteins that interact with one another  
387 would be more likely to come from the same evolutionary origin than would be expected by  
388 random association. Alternatively, early ochrophyte plastids might have had no preference  
389 for utilising interacting proteins of the same evolutionary origin, in which case proteins  
390 involved in specific metabolic pathways might frequently have different evolutionary origins  
391 to adjacent enzymes in the same pathway. Of the 313 pairs of such biochemical neighbours  
392 identified in the ancestral HPPGs, only 44 shared the same evolutionary origin, which is no  
393 different than that which would be expected by chance (expected number 41.05; chi-  
394 squared,  $P=0.541$ ; Fig. 5, panel C; Table S9- sheet 3<sup>145</sup>). Thus, interactions between proteins  
395 of different evolutionary origin were forged early in the evolution of the ochrophyte plastid.  
396

397

398 Finally, we sought correlations between expression dynamics and evolutionary affinity,  
399 taking advantage of microarray data from *P. tricornutum* and *T. pseudonana*<sup>74</sup> (Table S10-  
400 sheets 1-4<sup>145</sup>). We found no evidence that ancestral HPPG genes of any evolutionary origin  
401 had more similar expression dynamics to each other than to those of other evolutionary  
402 origins (ANOVA,  $P \leq 0.05$ ; Fig. 5, panel D; Fig. 5- figure supplements 8, 9; Table S10- sheet  
403 5<sup>145</sup>). For example, in both species, genes of green origin show a weaker average positive  
404 coregulation with one another than they do to genes from the same species of red or of  
405 prokaryotic origin (Fig. 5, panel D). Thus, the chimeric origins of the ochrophyte plastid has  
406 enabled extraordinary functional mixing of proteins from early in its evolution, with each of

406 the different donors contributing proteins with a broad range of biochemical functions and  
407 transcriptional patterns in response to changing physiological conditions.

408

409 *Ancient origins of chimeric plastid-targeted proteins*

410

411 We considered whether the mixing of proteins from different evolutionary sources might  
412 have more substantially changed the biology of the ochrophyte plastid. It has been reported  
413 by Méheust et al.<sup>75</sup> that proteins of chimeric evolutionary origin, generated by the fusion of  
414 domains from different evolutionary sources, form a significant component of plastid  
415 proteomes. Thus, the chimeric origins of the ochrophyte plastid might have enabled the  
416 creation of syncretic proteins not found in the endosymbiont or host ancestors. We  
417 identified orthologues of seven chimeric proteins identified in this study within our dataset,  
418 underlining their importance for the establishment of the ochrophyte plastid (Fig. 6, panel  
419 A)<sup>75</sup>.

420

421 Next, we assessed whether the mosaic composition of the ochrophyte plastid proteome had  
422 also enabled the establishment of novel chimeric fusion proteins, unique to ochrophyte  
423 plastids. Using the taxonomic subdivisions erected for this study, we identified further  
424 chimerism events in members of 42 ancestral HPPGs (Fig. 6, panel B; Table S9- sheet 1,  
425 sections 4, 5; Table S11<sup>145</sup>). These include three HPPGs (e.g. NADH-ubiquinone  
426 dehydrogenase) in which chimeric proteins have formed through the fusion of modules of  
427 prokaryotic origin to others of eukaryotic origin, and seven HPPGs (e.g. translation factor EF-  
428 3b, and an N6-adenine DNA methyltransferase) in which fusion events have occurred  
429 between modules of red origin and modules of green origin (Fig. 6, panel B). To our  
430 knowledge, neither of these types of fusion event have previously been reported for plastid-  
431 targeted proteins<sup>75</sup>. The chimeric proteins contain domains from a wide range of  
432 evolutionary origins: 20 (47.6%) contain a domain of inferred green origin and 18 (43.8%)  
433 contain a domain of host origin.

434

435 Amongst the chimeric proteins identified, we found two that probably fused in the  
436 ochrophyte ancestor (Fig. 6, panels A, B). In one case, a bifunctional protein containing an N-  
437 terminal 3,4-dihydroxy-2-butanone 4-phosphate (DHBP) synthase and C-terminal GTP  
438 cyclohydrolase II protein, which performs two consecutive steps of riboflavin biosynthesis<sup>76</sup>,  
439 has formed through the fusion of a cyclohydrolase domain of probable host origin to a  
440 synthase domain of probable red algal or actinobacterial origin (Fig. 6- figure supplements  
441 1,2). While bifunctional DHBP synthase/ GTP cyclohydrolase proteins are known in bacteria,  
442 red algae and plants (Fig. 6- figure supplement 1)<sup>48, 76</sup>, in these taxa the DHBP synthase  
443 domain is located at the protein C-terminus; thus, an analogous but topographically distinct  
444 fusion protein has evolved in ochrophytes. In a second, previously reported case<sup>75</sup>, a C-  
445 terminal plastid-targeted Tic20 subunit of red algal origin has become fused to an N-terminal  
446 EF-hand motif, for which no clear evolutionary outgroup (to an e value of below  $1 \times 10^{-05}$ )  
447 could be found (Fig. 6- figure supplement 3). Thus, the fusion of proteins of different  
448 evolutionary origins has generated new functions in the ochrophyte plastid proteome.

449

450 *Ancestral and bidirectional origins of dual targeting in ochrophytes*

451

452 Finally, we considered whether the acquisition of the ochrophyte plastid might have also  
453 fundamentally altered the biology of the host cell, by contributing proteins to host processes  
454 and structures outside the plastid. As an exemplar system, we considered dual targeting of  
455 proteins to plastids and mitochondria, which is known to occur extensively in plants<sup>77, 78</sup>, and  
456 has recently been documented in diatoms<sup>79</sup> and in other complex plastid lineages<sup>79, 80</sup>.

457 Previous studies have speculated that dual targeting may arise early in plastid evolution, for  
458 example through the retargeting of proteins from the host mitochondria to the plastid, or  
459 equally via the adaptation of proteins of plastid origin to the mitochondria<sup>18, 77</sup>.  
460

461 We indeed identified proteins that appeared to be dual targeted to the plastid and a  
462 secondary organelle (Fig. 2- figure supplements 5, 6), which we verified to be the  
463 mitochondria using Mitotracker orange (Fig. 7 panel A). In at least two cases (histidyl- and  
464 prolyl-tRNA synthetase) this dual targeting is a conserved feature, as we identified the same  
465 fluorescence patterns both in *P. tricornutum* and using heterologous expression constructs  
466 from *G. foliaceum* and *N. gaditana* (Fig. 7, panel A; Fig. 7- figure supplement 1). To determine  
467 whether dual targeted proteins were ancestrally present in the ochrophyte plastid, we  
468 developed an *in silico* pipeline, based on experimental data, to identify probable dual  
469 targeted proteins from within the HPPG dataset (Fig. 7- figure supplement 2; Table S12-  
470 sheet 1<sup>145</sup>). In total, we identified 1103 HPPGs that included at least one member that was  
471 probably dual targeted to plastids and mitochondria (Table S12- sheet 1<sup>145</sup>). 34 of these  
472 HPPGs passed the conservation thresholds previously inferred to signify an ancestral origin  
473 (Table S12- sheet 1<sup>145</sup>). Thus, dual targeting is an ancestral feature of the ochrophyte plastid.  
474

475 We then considered the origins of the ancestrally dual targeted ochrophyte proteins. 15 of  
476 the 34 putative ancestrally dual targeted HPPGs were orthologous to HPPGs of clear  
477 evolutionary origin; of these, the majority (11/15; 73%) were of red algal, i.e., probable  
478 endosymbiont origin (Fig. 7, panel B; Table S12- sheet 2<sup>145</sup>). To determine how these dual  
479 targeted HPPGs have altered the biology of the host, we searched for gene families  
480 corresponding to aminoacyl-tRNA synthetases within the 7140 non-redundant gene families  
481 previously identified to be shared across the ochrophytes (Table S8- sheet 1<sup>145</sup>). To enable  
482 function of the translational machinery, each genome within the ochrophyte cell (i.e.,  
483 nucleus, mitochondrion, and plastid) requires aminoacyl-tRNA synthetase activity for each  
484 amino acid<sup>79</sup>; thus, if any class of aminoacyl-tRNA synthetase is represented by fewer than  
485 three genes, then individual tRNA synthetases must support the biology of multiple  
486 organelles through dual targeting. We identified seven classes of tRNA synthetase for which  
487 there were only two gene families in the ochrophyte ancestor, one corresponding to a  
488 cytosolic enzyme, and the other to an enzyme that was probably dual targeted to both the  
489 mitochondria and plastid. These include five cases in which the dual targeted tRNA  
490 synthetase was of apparent red algal, i.e., endosymbiont origin (Fig. 7, panel C). Thus, the  
491 acquisition of the ochrophyte plastid also altered the biology of the mitochondria, with dual  
492 targeted proteins of endosymbiont origin functionally replacing endogenous mitochondrial-  
493 targeted homologues.  
494

## 495 **5. Complex evolutionary origins of CASH lineage plastids**

### 496 *A pelagophyte/dictyochophyte origin of the haptophyte plastid proteome*

497 We considered whether our dataset provides evidence for any of the other CASH lineage  
498 plastids (cryptomonads, haptophytes, or photosynthetic alveolates) originating within the  
499 ochrophytes<sup>1, 5, 7</sup>, or evidence for gene transfer from ochrophytes into lineages with complex  
500 plastids of green algal origin (chlorarachniophytes and euglenids)<sup>81, 82</sup>. In a majority  
501 (243/437) of trees in which they could be assigned a clear origin, plastid-targeted proteins  
502 from haptophytes resolved at a position within the ochrophyte clade (Materials and  
503 Methods; Fig. 8, panel A; Table S4- sheet 5<sup>145</sup>). All other groups (except for dinatoms, which  
504 have well-defined plastids of diatom origin<sup>20, 56</sup>) generally branched externally rather than  
505 within the ochrophyte clade (Fig. 8, panel A). Indeed, the proportion of haptophyte proteins  
506

508 that resolved within the ochrophytes was found to be significantly greater than any of the  
509 other groups except for dinotoms (chi-squared,  $P < 1 \times 10^{-5}$ ; Table S4- sheet 5<sup>145</sup>).

510  
511 We noted that the plastid-targeted haptophyte proteins of ochrophyte origin were biased  
512 towards specific origins, with over half of the proteins that grouped with a specific  
513 ochrophyte lineage (100/178) resolving with members of the hypogyristea (i.e.,  
514 pelagophytes, dictyochophytes, and bolidophytes; Fig. 8- figure supplement 1; Table S4-  
515 sheet 5<sup>145</sup>). No such bias could be observed in any other CASH lineage, in which invariably a  
516 significantly smaller proportion of proteins were found to resolve with hypogyristeans  
517 lineages (chi-squared  $P < 0.01$ ; Fig. 8- figure supplement 1; Table S4- sheet 5<sup>145</sup>). We  
518 additionally explored whether there might be unique synapomorphies shared between one  
519 ochrophyte lineage and the haptophytes. We found 53 ASAFind-generated HPPGs that  
520 contained a majority ( $\geq 2/3$ ) of the haptophyte sub-categories and contained at least one  
521 member of the hypogyristea, but contained no other ochrophyte orthologues (Fig. 8, panel  
522 B; Table S2- sheet 2, section 3<sup>145</sup>). This was significantly more than would be expected (28.3,  
523 chi-squared  $P = 0.00013$ ) through a random assortment of all HPPGs that were uniquely  
524 shared between haptophytes and one ochrophyte lineage, corrected for the relative size of  
525 each dataset (Materials and Methods). We similarly found a significantly larger number of  
526 HPPGs to be uniquely shared between a majority of both the haptophytes and a majority  
527 ( $\geq 2/3$ ) of the hypogyristeans sub-categories (15, expected number 8.0,  $P = 0.034$ ; Fig. 8, panel  
528 B) or shared between a majority of hypogyristea and at least one haptophyte sub-category  
529 (28, expected number 12.9,  $P = 0.00073$ ; Table S2- sheet 2, section 3<sup>145</sup>; Fig. 8, panel B). Thus,  
530 our data supports a specific gene transfer event between the hypogyristea and the  
531 haptophytes.

532  
533 We investigated whether there is a more specific origin for the ochrophyte sequences in  
534 haptophyte plastids. First, we tabulated the individual ochrophyte sub-categories identified  
535 in the first sister group to haptophyte sequences, of which the greatest number (94)  
536 resolved specifically with pelagophyte and dictyochophyte sequences, rather than with  
537 bolidophytes, non-hypogyristeans lineages, or more ancestral nodes (Fig. 8, panel C; Fig. 8-  
538 figure supplement 2). Next, we extracted all of the haptophyte plastid-targeted sequences  
539 assembled into each ancestral ochrophyte HPPG, performed BLAST top hit analysis (Table  
540 S13- sheets 1-3<sup>145</sup>), and identified sequences for which the best hit was from the same  
541 ochrophyte lineage (diatoms, hypogyristea, or chrysista) as the tree sister group (Table S13-  
542 sheet 4<sup>145</sup>). We performed separate analyses for query sequences from each of the three  
543 haptophyte sub-categories considered in our analysis (pavlovophytes, prymnesiales, or  
544 isochrysidales). In each case, at least 50% of the sequences that produced an evolutionarily  
545 consistent series of top hits resolved either with the pelagophytes or dictyochophytes (Fig.  
546 8- figure supplement 3; Table S13- sheet 4<sup>145</sup>). Thus, these proteins originated within an  
547 ancestor of the pelagophyte/ dictyochophyte lineage.

548  
549 We next tested the probable direction of the gene transfer events. We reasoned that if the  
550 genes identified within our study had been transferred from an ancestor of pelagophytes  
551 and dictyochophytes into the haptophytes, then we should also see a strong secondary  
552 signal linking the haptophytes to earlier ancestors of the pelagophyte/ dictyochophyte clade,  
553 for example the common ancestor of hypogyristea and diatoms. We inspected the  
554 secondary BLAST top hits associated with genes shared between haptophytes and  
555 hypogyristea (Fig. 8- figure supplement 4; Table S13- sheet 5<sup>145</sup>), and the next deepest sister-  
556 groups to haptophyte proteins that are of probable pelagophyte or dictyochophyte origin in  
557 each single-gene tree (Fig. 8- figure supplement 4; Table S4- sheet 2, section 6<sup>145</sup>). The  
558 majority of haptophyte proteins of hypogyristeans origin in single-gene trees (65/100) clearly

resolved within a broader HPPG containing multiple ochrophyte lineages, and this bias was corroborated by the specific sister groups associated with each protein as inferred by heat map analysis (Fig. 8- figure supplement 4, panel A). Moreover, the majority of haptophyte proteins with hypogyristeal BLAST top hits, and hypogyristeal proteins with haptophyte BLAST top hits (48/ 86 sequences total) had next best BLAST hits against diatoms (Fig. 8- figure supplement 4, panel B). We additionally tabulated the earliest and latest possible origin points of amino acid residues that were uniquely shared between haptophytes and some but not all ochrophyte lineages, from a dataset of 37 HPPGs for which there was a clear evolutionary affinity between haptophytes and ochrophytes and strict subsequent vertical inheritance (Fig. 8, panel D; Fig. 8- figure supplement 5; Table S6- sheets 3, 4<sup>145</sup>). A greater number of the uniquely shared residues were found to be conserved between the haptophytes and the common ancestor of hypogyristea and diatoms, than were specifically only shared with pelagophyte and dictyochophyte sequences, both per the latest possible origin (139 residues shared with hypogyristea and diatoms; 99 residues with pelagophytes and dictyochophytes; Fig. 8, panel D; Table S7- sheets 2, 3<sup>145</sup>) and per the earliest possible origin (46 residues shared with hypogyristea and diatoms; 41 residues with pelagophytes and dictyochophytes; Fig. 8- figure supplement 5; Table S7- sheets 2, 3<sup>145</sup>). This specifically supports a transfer of plastid-targeted proteins from an ancestor of the pelagophyte/ dictyochophyte clade into the haptophytes, rather than the other way around.

Finally, we tested whether these proteins were likely to have been acquired through an endosymbiotic event. We reasoned that the genes acquired by haptophytes through endosymbiotic events should encode a greater proportion of plastid-targeted proteins than would be observed with genes of alternative origin. We accordingly constructed a dataset of 12,728 non-redundant gene families that were broadly distributed across the haptophytes (Table S14- sheet 1<sup>145</sup>), of which 772 were of probable hypogyristeal origin (Fig. 8- figure supplement 6; Table S14- sheet 2<sup>145</sup>). A significantly larger proportion of the ancestral haptophyte gene families of hypogyristeal origin were predicted by ASAFind to be targeted to the plastid than would be expected by random distribution of the data (observed number 43, expected number 22.8, chi-squared P= 2.2 x 10<sup>-05</sup>; Fig. 8, panel E; Table S14- sheet 3<sup>145</sup>), consistent with an endosymbiotic origin. Thus, our data support an endosymbiotic uptake of an ancestor of the pelagophytes and dictyochophytes by an ancestor of the haptophytes.

#### 591 592 *Phylogenetic discrepancies between the haptophyte plastid proteome and genome* 593

594 The transfer of plastid-targeted proteins from the pelagophyte/dictyochophyte clade into  
595 the haptophytes is surprising, as previous studies have indicated that the haptophyte plastid  
596 genome originates either as a sister-group to the entire ochrophyte lineage<sup>5</sup> or to the  
597 cryptomonads<sup>83, 84</sup>. To verify this discrepancy we constructed two plastid trees, one using 54  
598 conserved proteins that are encoded in all sequenced red lineage and glaucophyte plastids  
599 (Fig. 9, panel A; Table S15- sheet 1<sup>145</sup>), and one using a smaller subset of 10 plastid-encoded  
600 proteins that were detected in many of the transcriptome libraries used in this study (Fig. 9,  
601 panel B; Table S15- sheet 1<sup>145</sup>).

602 A specific sister-group relationship between the cryptomonads and haptophytes was  
603 recovered, with moderate to strong bootstrap support, in both the gene-rich tree (Fig. 9,  
604 panel A) and the taxon-rich tree (Fig. 9, panel B). Both trees also strongly supported the  
605 monophyly of ochrophyte plastid genomes (Fig. 9). Alternative topology tests rejected any  
606 possibility that the haptophyte plastid originated within the ochrophytes (Fig. 9- figure  
607 supplement 1; P≤ 0.05). Similarly, trees calculated from alignments in which fast-evolving  
608 sites and clades had been serially removed, and in which the alignment had been recoded to

610 minimise amino acid composition biases (Fig. 9- figure supplement 2; Table S15- sheet 2;  
611 Table S16<sup>145</sup>) either recovered a sister-group relationship between haptophytes and  
612 cryptomonads, or placed haptophytes as the sister group to all ochrophytes. We additionally  
613 generated and inspected single-gene tree topologies for each of the constituent genes used  
614 to generate each concatenated multigene alignment, and could not find any that confidently  
615 resolved a sister-group relationship between haptophytes and the pelagophyte/  
616 dictyochophyte clade (Fig. 9- figure supplement 3; Table S15- sheet 3<sup>145</sup>). Finally, we found  
617 only three residues in the alignment that were uniquely shared among all four haptophytes  
618 and the sole representative of pelagophytes and dictyochophytes (*Aureococcus*) in the gene-  
619 rich dataset, and no residues that were shared between a majority of the haptophytes and  
620 at least one pelagophyte or dictyochophyte sequence in the taxon-rich dataset (Fig. 8, panel  
621 C; Table S17- sheet 4<sup>145</sup>). In contrast, we found large numbers of residues that were shared  
622 uniquely by haptophytes and other lineages (Fig. 9, panel C; Table S17- sheet 4<sup>145</sup>). This  
623 strong support for a relationship between haptophytes and cryptomonads is inconsistent  
624 with phylogenetic artifacts such as coevolution between specific protein complexes<sup>58, 85</sup> or  
625 gene duplication and differential loss of paralogues<sup>86</sup>, in which case there should still be a  
626 detectable underlying signal linking it to the pelagophytes and dictyochophytes. We  
627 conclude that while many plastid-targeted haptophyte proteins originate from an ancestor  
628 of the pelagophytes and dictyochophytes, the haptophyte plastid genome does not.  
629

### 630 Discussion

631 In this study, we have reconstructed an experimentally verified dataset of 770 plastid-  
632 targeted proteins that were present in the last common ancestor of all ochrophytes (Figs. 1,  
633 2). Our dataset accordingly provides windows into the evolutionary origins of the  
634 ochrophyte plastid lineage. These include evidence for a green algal contribution to  
635 ochrophyte plastid evolution and a late acquisition of the ochrophyte plastid following  
636 divergence of the ochrophyte lineage from oomycetes (Figs. 3, 4). This latter finding is  
637 particularly interesting as molecular divergence estimates place the ochrophytes as  
638 diverging from the oomycetes no more than 90 million years prior to the radiation of  
639 ochrophyte lineages<sup>87, 88</sup>. Assuming that these estimates are reliable, our dataset represents  
640 some of the earliest proteins to support the ochrophyte plastid following its endosymbiotic  
641 uptake. We also provide evidence for widespread mixing of proteins of different  
642 evolutionary origin in the ancestral ochrophyte plastid (Fig. 5), including evidence for the  
643 formation of new fusion proteins through the recombination of domains of different  
644 evolutionary origins (Fig. 6), and a bidirectional mixing of proteins derived from the  
645 endosymbiont with proteins from host organelles via dual targeting (Fig. 7). A schematic  
646 outline of these results is shown in Fig. 10.

647  
648 Many questions nonetheless remain to be answered. It remains to be determined whether  
649 the *in silico* prediction facilitated by programmes such as ASAFind and HECTAR are sufficient  
650 to enable the identification of all ochrophyte plastid proteins<sup>29, 30</sup>. This is particularly  
651 pertinent in the context of dual targeted proteins, insofar as the dataset of 34 potentially  
652 ancestrally dual targeted proteins identified in this study may not include proteins that are  
653 dual targeted to the plastid and other cellular organelles, such as the ER<sup>89</sup>, cytoplasm<sup>90</sup>, or  
654 nucleus<sup>91</sup>. We note also that, based on the fluorescence patterns observed with the  
655 exemplar proteins within this study (Figs. 2, 7), ASAFind and HECTAR may identify proteins  
656 targeted to the periplastid compartment, as well as to the plastid stroma. While these  
657 periplastid and multipartite proteins probably form an important part of plastid physiology,  
658 it will be interesting to dissect the specific signals associated with the targeting of proteins to  
659 individual sub-compartments within CASH lineage plastids<sup>55, 92</sup>.  
660

661  
662 Another major question concerns the origins of plastid-targeted proteins of green algal  
663 origin in ochrophytes. Overall, our data supports the targeting of a significant complement  
664 of proteins of chlorophyte origin to the ochrophyte plastid (Fig. 4). It remains to be  
665 determined, however, what the exact chlorophyte donor was, and how these genes may  
666 have been acquired. It is possible that the green genes were transferred into the ochrophyte  
667 lineage via lateral gene transfer, either from a range of different green algal sources or  
668 repeatedly from one lineage (for example, a semi-permanent intracellular symbiont<sup>3</sup>),  
669 although neither scenario would explain the bias in green algal genes in ochrophyte  
670 genomes towards encoding proteins of plastid function (Fig. 4, panel D). An alternative  
671 possibility might be a cryptic green algal endosymbiosis in the evolutionary history of the  
672 host, as has been previously suggested<sup>1, 44</sup> (Fig. 10), or a more convoluted pattern of  
673 acquisition. We note, for example, that the green genes identified in our study are not only  
674 plastid-targeted across the ochrophytes, but are apparently shared with haptophytes and  
675 cryptomonads (Fig. 10- figure supplement 1), which would be equally consistent with them  
676 having been present in a common ancestor of the CASH lineage plastid, and relocated to  
677 each host nuclear lineage following endosymbiosis (Fig. 10). Thus, pinpointing the exact  
678 nature and timing of the green gene transfer into ochrophytes rests not only on more  
679 extensive sequencing of deep-branching chlorophyte lineages, but also on characterising the  
680 genome composition of the closest aplastidic relatives of extant ochrophytes (e.g.,  
681 *Develorapax*, *Pirsonia*<sup>6</sup>), and the closest red algal relative of CASH lineage plastids, which  
682 remains unknown<sup>1, 4</sup>.  
683  
684 We also provide evidence for a chimeric origin of the haptophyte plastid (Figs. 8, 9). A  
685 schematic outline of these results is shown in Fig. 10- figure supplement 2. We have shown  
686 that a significant number of plastid-targeted proteins found in haptophytes originate from  
687 an ancestor of the pelagophytes and dictyochophytes (Fig. 8). This relationship is supported  
688 by multiple lines of evidence- i.e., uniquely shared proteins, single-gene tree topologies,  
689 BLAST top hit analysis, and analysis of synapomorphies in multigene alignments (Fig. 8 and  
690 supplements). Alongside the bias of haptophyte genes of hypogyristeal origin encoding  
691 proteins of plastid function (Fig. 8- panel E), these observations argue against these genes  
692 having been acquired through multiple independent lateral gene transfer events, and  
693 instead support an endosymbiosis event. We note that other studies have shown strong  
694 evidence for gene transfers between haptophytes and individual members of the  
695 hypogyristeal: for example, Stiller *et al.* have demonstrated a strong enrichment in BLAST top  
696 hits against haptophytes, from the genome of the pelagophyte *Aureococcus*  
697 *anophageferrens*, compared to other ochrophyte genomes<sup>5</sup>. We additionally note that an  
698 ancestral gene transfer from a pelagophyte/ dictyochophyte ancestor into the haptophytes  
699 is a chronologically realistic scenario: molecular clock estimates place the pelagophytes and  
700 dictyochophytes diverging between 300 and 700 million years before present<sup>87, 93</sup>, which  
701 broadly overlaps with the molecular dates estimated for the radiation of the haptophytes in  
702 the same studies<sup>87, 93</sup>, and precedes the first haptophyte microfossils, identified ca. 220  
703 million years before the present<sup>94</sup>.  
704 Finally, we verify that the evolutionary links between haptophyte and the pelagophyte/  
705 dictyochophyte clade in terms of plastid-targeted proteins are not supported by phylogenies  
706 of the haptophyte plastid genome (Fig. 9). Other multigene phylogenies of red lineage  
707 plastid genomes have similarly demonstrated that the haptophyte plastid genome instead  
708 resolves as a sister-lineage either to cryptomonads or to all ochrophytes<sup>5, 38, 83, 84</sup>.  
709 Furthermore, the structure and content of haptophyte and hypogyristeal plastid genomes  
710 are dissimilar: for example, haptophyte plastids possess an *rp/36* gene that has been laterally  
711 acquired from a bacterial donor and is shared with cryptomonad plastids but absent from

712 ochrophytes<sup>95</sup>, and ochrophyte plastids no longer retain genes encoding the plastid division  
713 machinery proteins *minD* and *minE*, which remain plastid-encoded in haptophytes and  
714 cryptomonads<sup>96</sup>. Similarly, extant haptophyte plastids have comparatively large plastid  
715 genomes and possess a conventional quadripartite structure<sup>97</sup>, whereas extant pelagophyte  
716 plastids have a reduced coding content compared to other photosynthetic ochrophytes,  
717 cryptomonads and haptophytes, and have secondarily lost the plastid inverted repeat<sup>98, 99</sup>,  
718 although it is not yet known whether dictyochophyte plastids share this reduced structure.

719 The discrepancy between the pelagophyte/ dictyochophyte origin of the haptophyte plastid  
720 proteome and the clear non-ochrophyte origin of its plastid genome might be explained by  
721 several different evolutionary scenarios. One possibility would be a serial endosymbiosis  
722 event deep in haptophyte evolutionary history, in which an ancient plastid derived from a  
723 pelagophyte/ dictyochophyte ancestor was acquired by the haptophyte common ancestor,  
724 then replaced subsequently by a plastid of non-ochrophyte origin (Fig. 10- Figure  
725 supplement 2). Verifying this scenario, or its alternatives (such as lateral gene transfer from  
726 pelagophyte or dictyochophyte algae into the algal ancestors of the haptophyte plastid)  
727 rests on identifying the exact origin of the current haptophyte plastid genome, and in  
728 particular demonstrating that the haptophyte plastid genome originates from within (rather  
729 than forms a sister-group to) a major lineage of eukaryotic algae other than ochrophytes  
730 (Fig. 10- Figure supplement 2). For this, sequence data from early-diverging members of the  
731 cryptomonads and haptophytes will be particularly important<sup>41, 100, 101</sup>. It also remains to be  
732 determined whether other CASH lineage plastids, such as the peridinin-type plastids found  
733 in most photosynthetic alveolates, originate within the ochrophytes<sup>7, 20</sup>. Similar plastid  
734 proteome reconstructions, using bespoke datasets for these species, will be particularly  
735 useful in unravelling their disparate evolutionary origins.

736 Overall, our dataset provides valuable and deep insights into the chimeric origins and  
737 complex fates of a major group of eukaryotic algae. Further studies using more sensitive  
738 pipelines, or using analogous datasets from other major CASH lineages, may elucidate the  
739 evolutionary and physiological diversification of plastids in the open ocean.

740

## 741 Materials and Methods

742

### 743 Identification of ancestral plastid-targeted ochrophyte proteins

744

745 Ancestral plastid-targeted proteins in ochrophytes were identified via a composite pathway,  
746 consisting of *in silico* prediction, identification of conserved proteins using BLAST, alignment,  
747 and single-gene tree building. First, the complete protein libraries annotated from eleven  
748 ochrophyte genomes (the diatoms *Phaeodactylum tricornutum*<sup>59</sup>, *Thalassiosira*  
749 *pseudonana*<sup>9</sup>, *Thalassiosira oceanica*<sup>102</sup>, *Fistulifera solaris*<sup>103</sup>, *Fragilariaopsis cylindrus*, *Synedra*  
750 *acus*<sup>104</sup>, and *Pseudonitzschia multiseries*; the pelagophyte *Aureococcus anophagefferens*<sup>11</sup>;  
751 the eustigmatophytes *Nannochloropsis gaditana* and *Nannochloropsis salina*<sup>37, 105</sup>; and the  
752 kelp *Ectocarpus siliculosus*<sup>10</sup>; Table S1- sheet 1<sup>145</sup>), were screened using the ochrophyte  
753 plastid-targeting predictors ASAFind<sup>29</sup> (used in conjunction with SignalP version 3.0<sup>106</sup>; Table  
754 S2<sup>145</sup>) and HECTAR<sup>30</sup> (integrated into a Galaxy<sup>107</sup> instance available at <http://webtools.sbs-roscff.fr>; Table S3<sup>145</sup>). All proteins that were deemed to possess plastid-targeting sequences  
755 (regardless of the confidence score applied by ASAFind<sup>29</sup>) were retained for further  
756 inspection.

757

758 Possible conserved plastid-targeted sequences (i.e. homologous plastid-targeted protein  
759 groups, or HPPGs) were next identified using a customised BLAST protocol. First, a library of  
760 non-redundant proteins was generated to serve as seed sequences for further searches.

761

762 Each plastid-targeted protein identified from ochrophyte genome sequences was searched  
763 by BLASTp against a modified Uniref<sup>108</sup> library, and the expect values for all top hits were  
764 extracted, to yield a floating BLAST threshold below which orthologous proteins were  
765 identified. All sequences from lineages with a history of secondary endosymbiosis were first  
766 removed from the Uniref library in order to avoid the confounding effects of gene transfer  
767 from current and former symbionts<sup>5, 7, 81, 82</sup>. The removed lineages included cryptomonads,  
768 centrohelids, telonemids, haptophytes, alveolates, rhizaria, euglenids, and plastid-bearing  
769 stramenopiles. All of the ochrophyte genome-derived plastid-targeted proteins were  
770 searched against one another by BLAST, and proteins that matched one another with an  
771 expect score lower than the first outgroup hit (or were retrieved as a stronger match than  
772 the outgroup hit if the expected values of both were zero), and thus likely correspond to  
773 different proteins within the same monophyletic plastid protein cluster, were merged. Only  
774 one protein was retained as the seed sequence for subsequent growth of each cluster: this  
775 was defined first via organism (in order of preference: *P. tricornutum*, *T. pseudonana*, *P.*  
776 *multiseries*, *F. cylindrus*, *S. acus*, *A. anophagefferens*, *E. siliculosus*, *N. gaditana*, *N. salina*, *T.*  
777 *oceanica*, *F. solaris*) and, where more than one protein was available for a given organism,  
778 the protein with the lowest BLAST expect value against the corresponding uniref top hit.  
779

780 Next, plastid-targeted protein sequences were sought from all available ochrophyte  
781 sequence data. A search database was built from all eleven completed ochrophyte genomes,  
782 147 ochrophyte sequence libraries from the Marine Microeukaryote Transcriptome  
783 Sequence Project<sup>34</sup>, eleven further ochrophyte transcriptome sequencing projects<sup>64, 109, 110</sup>  
784 and uniref. Cross-contamination was removed from MMETSP transcriptomes as previously  
785 described<sup>57</sup>. Briefly, this procedure compares the nucleotide sequences of contigs assembled  
786 from each MMETSP library by pairwise BLAST, and defines a separate cross-contamination  
787 threshold for each pair of MMETSP libraries based on their distribution of BLAST percent  
788 identities. These distributions should each contain a peak centered on the average  
789 nucleotide percent identity of transcripts between the two species. In addition, in the  
790 presence of cross-contamination, there should be a second peak at 100% identity. The  
791 procedure defines the cross-contamination threshold as the minimum between these two  
792 peaks; above the threshold, contigs (and the proteins predicted from them) are considered  
793 to be potentially cross-contaminated. In total, 2.5% of the MMETSP contigs were discarded  
794 through this method. A summary of the number of contigs discarded is provided in Table S1-  
795 sheet 2, section 1<sup>145</sup>.

796  
797 Each decontaminated sequence was trimmed at the N-terminus to the first methionine  
798 present, and binned into one of eleven different evolutionary categories, based on recent  
799 multigene phylogenetic trees for ochrophytes and diatoms<sup>12, 111-113</sup> (fig. 1, panel A; Table S1-  
800 sheet 1<sup>145</sup>). These consisted of: three chrysitan lineages (the "PX clade" of phaeophytes,  
801 xanthophytes and related lineages; raphidophytes; and the "PESC clade" of pinguiphyses,  
802 eustigmatophytes, synchromophytes, and synurophytes/chrysophytes), three hypogyristeian  
803 lineages (pelagophytes; dictyochophytes; and bolidophytes), and five diatom lineages (the  
804 basally divergent genus *Corethron*; radial centric lineages such as Coscinodiscophytes and  
805 Rhizosoleniaceae; the polar centric Thalassiosirales and Skeletonemataceae, which appear  
806 to be relatively distantly related to pennate diatoms<sup>111,113</sup>; polar centric lineages such as  
807 Odontellids and Chaetocerotales that appear to be more closely related to pennate  
808 diatoms<sup>111,113</sup>; and finally all pennate lineages). These binned sequences were then searched  
809 for plastid-targeted proteins by ASAFind and HECTAR as before.

810  
811 The seed sequences for the resulting non-redundant HPPGs were searched against the  
812 enlarged plastid sequence library using BLASTp. Proteins that matched against seed

813 sequences with a lower expect value than the outgroup best hit (or were retrieved as a  
814 stronger match than the outgroup hit if the expected values of both were zero), were added  
815 to each HPPG. Next, three custom thresholds were defined that were particularly successful  
816 in distinguishing probable proteins of true plastid localisation from false positives (fig. 1,  
817 panel B). For this, conservation patterns were selected that maximised the relative  
818 enrichment in proteins with unambiguous plastid functions (i.e., were annotated to function  
819 in photosynthesis, to constitute integral parts of the plastid thylakoid or inner membranes,  
820 or corresponded to the expression products of genes that are plastid-encoded in red algae  
821 but have been apparently relocated to the ochrophyte nucleus<sup>97</sup> or that corresponded to  
822 proteins previously verified experimentally to localise to ochrophyte plastids<sup>29, 30, 114, 115</sup>), and  
823 thus should contain relatively fewer examples of mispredicted proteins within the dataset.  
824 At the same time, conservation patterns were selected that minimised the number of HPPGs  
825 identified as conserved from a negative control dataset (consisting of HPPGs assembled  
826 using seed sequences from the published genome sequences of the cryptomonad *Guillardia*  
827 *theta*<sup>17</sup> or the haptophytes *Emiliania huxleyi*<sup>116</sup> and *Chrysochromulina tobin*<sup>117</sup>, and for which  
828 no plastid-targeted orthologues were detected in any of the ochrophyte genome sequences  
829 used in this study). The thresholds corresponded to: orthologues in a majority ( $\geq 2/3$ ) of  
830 chrysitan and a majority ( $\geq 3/5$ ) of diatom lineages; a majority of chrysitan and a majority  
831 ( $\geq 2/3$ ) of hypogyristeian lineages; and at least one chrysitan, and a majority of both  
832 hypogyristeian and diatom lineages (fig. 1).

833 All of the HPPGs that passed at least one threshold were extracted, and homology for each  
834 HPPG was confirmed individually (Table S4- sheet 1<sup>145</sup>). First, each HPPG was aligned using  
835 20 iterations of MUSCLE v8<sup>118</sup>, followed by the in-built alignment programme integrated into  
836 GeneIOUS v 4.76<sup>119</sup>, under the default criteria. Each HPPG alignment was manually  
837 inspected, and proteins that failed to align with the genomic sequences, clearly terminated  
838 within the conserved region of the protein, or were truncated at the N-terminus by a length  
839 of greater than 50 amino acids (i.e. the approximate length of an ochrophyte plastid-  
840 targeting sequence<sup>29, 114</sup>) were removed, following which HPPGs that no longer passed the  
841 taxonomic criteria defined for conservation were eliminated (Table S4- sheet 1<sup>145</sup>). Next,  
842 each HPPG was enriched with the sequences for the top 50 hits obtained when the seed  
843 sequence was searched against the modified uniref library as detailed above, alongside the  
844 single best hit for composite transcriptome and genome libraries constructed for 36  
845 eukaryotic sub-categories (Table S1- sheet 1<sup>145</sup>), and realigned against this reference. The  
846 transcriptome components of the reference sequence libraries were cleaned of residual  
847 contamination as defined above, and 23 individual MMETSP libraries were additionally  
848 excluded due to evidence of further contamination (Table S1- sheet 2<sup>145</sup>). Sequences that  
849 failed to align were removed, and HPPGs that failed to meet the criteria for conservation  
850 following alignment were eliminated (Table S4- sheet 1<sup>145</sup>).

851 Finally, each HPPG was trimmed at the N- and C-termini to (respectively) the first residue  
852 and last residue visually identified to be conserved in > 70% of the sequences in the  
853 alignment, corresponding to the probable conserved domain of the protein. Each HPPG was  
854 then trimmed with trimAl using the -gt 0.5 option<sup>120</sup>. 100 trees were calculated for each  
855 trimmed alignment using RAxML, with the JTT substitution model + gamma correction<sup>121</sup>.  
856 The consensus tree from the 100 bootstrap replicates was manually inspected for the  
857 presence of a clade of ochrophyte proteins, containing sufficient sequences to pass the  
858 criteria for conservation defined above, that was either monophyletic, or paraphyletic to the  
859 inclusion of only one of five different non-ochrophyte groups (prokaryotes, red algae, green  
860 algae, aplastidic stramenopiles, and all other eukaryotes excluding CASH lineages, rhizaria  
861

863 and euglenids; Table S4- sheet 1<sup>145</sup>). HPPGs that passed this final stage of analysis were  
864 deemed to correspond to ancestrally plastid-targeted proteins (Table S4- sheet 2<sup>145</sup>).  
865

866 All identified plastid-targeted proteins, HPPGs, full aligned HPPGs, and single-gene trees  
867 have been made publically accessible through the University of Cambridge dSpace server  
868 (<https://www.repository.cam.ac.uk/handle/1810/261421><sup>145</sup>).  
869

#### 870 **Generation of fluorescence expression constructs for *Phaeodactylum tricornutum***

871 *Phaeodactylum tricornutum* 1.86 (CCMP2561), *Nannochloropsis gaditana* CCMP526, and  
872 *Glenodinium foliaceum* PCC499 were maintained in liquid cultures of f/2 medium  
873 supplemented with vitamins, and 100 µg/ ml each of ampicillin, streptomycin, kanamycin  
874 and neomycin, in a constant 19°C environment in a 12h: 12h cycle of 150 µE m<sup>-2</sup> s<sup>-1</sup> light:  
875 dark. *P. tricornutum* was maintained on an orbital shaker at 100 rpm, while *N. gaditana* and  
876 *G. foliaceum* were maintained as stationary cultures. Large volume cultures of *P.*  
877 *tricornutum* (e.g. cultures grown for transformation by bombardment) were grown in  
878 artificial seawater, supplemented with vitamins but without antibiotics.  
879

880 Total cellular RNA was extracted from c. 30 ml volumes of late log phase culture from each  
881 species using a modified Trizol phase extraction and DNase treatment protocol as described  
882 elsewhere<sup>21</sup>. Each RNA sample was tested for integrity by gel electrophoresis and quantified  
883 by a nanodrop spectrophotometer, and confirmed to be free of residual DNA contamination  
884 by direct PCR using universal eukaryotic 18S rDNA primers<sup>122</sup>. Approximately 200 ng purified  
885 RNA from each species was used as the template for cDNA synthesis, using a Maxima First  
886 Strand cDNA Synthesis Kit (Thermo), following the manufacturer's instructions.  
887

888 Nucleotide sequences encoding plastid-targeted proteins of unusual provenance were  
889 identified using the complete genome sequences of *Phaeodactylum tricornutum* and  
890 *Nannochloropsis gaditana*<sup>37, 59</sup>, and the *Glenodinium foliaceum* CCAP1116/3 transcriptome  
891 library assembled as part of MMETSP<sup>34, 123</sup> (Table S5<sup>145</sup>). Two primers were designed for each  
892 sequence: a PCR forward primer corresponding to the 5' end of the ORF, and a  
893 translationally in-frame PCR reverse primer positioned a minimum of 45 bp into conserved  
894 domain of the protein sequence (Table S5<sup>145</sup>). These primers were respectively fused to 5'  
895 fragments complementing the 3' end of the *P. tricornutum* FcpA promoter, and the 5' end of  
896 the GFP CDS. For one gene (the novel plastid protein), PCR reverse primers were designed  
897 complementary to the 3' end of the CDS of each gene due to the lack of a verifiable CDD; a  
898 full-length PCR reverse primer was additionally designed against the histidyl-tRNA  
899 synthetase sequence from *Nannochloropsis gaditana* due to failure to obtain functional  
900 expression from N-terminal constructs (data not shown).  
901

902 High-fidelity PCR products were amplified with each primer pair from the corresponding  
903 cDNA product using Pfu DNA polymerase (Thermo), per the manufacturer's instructions. In  
904 two cases (*Nannochloropsis gaditana* peroxisomal membrane protein, and the novel plastid  
905 protein) inserts were amplified from synthetic, codon-optimised constructs, designed to  
906 maximise expression levels in *Phaeodactylum tricornutum* (Eurofins). Each product was  
907 separated by DNA gel electrophoresis, cut, purified using a PCR gel extraction column kit  
908 (Macherey-Nagel), quantified using a nanodrop spectrophotometer, and verified by Sanger  
909 sequencing (GATC Biotech). The purified products were then used for Gibson ligation  
910 reactions<sup>124</sup> (NEB), following the manufacturer's instructions, using linearised and DpnI-  
911 treated vector sequence generated from the pPhat-eGFP vector<sup>35</sup>, and transformed into  
912 chemically competent Top10 *E. coli* cells, prior to selection on LB-1% agar plates containing  
913

914 100 µg/ ml ampicillin. Individual colonies were picked, verified to contain the insert  
915 sequence by PCR, and grown as overnight liquid cultures on LB medium supplemented with  
916 100 µg/ ml ampicillin, prior to purification of the plasmids by alkaline lysis and isopropanol  
917 precipitation<sup>125</sup>. Purified plasmids were integrated into *P. tricornutum* cells via biolistic  
918 transformation, using the Biolistic PDS-1000/He Particle Delivery System (BioRad),  
919 essentially as previously described<sup>35, 126</sup>.

920  
921 Colonies obtained from each transformation were transferred to liquid f/2 supplemented  
922 with vitamins and 100 µg/ ml zeocin, and were left to recover under the same growth  
923 conditions as used for liquid cultures of untransformed cells. Expression of GFP was  
924 visualised using a TCS SP8 confocal microscope (Leica), an excitation wavelength of 488 nm  
925 and emission wavelength interval of c. 510-540 nm. Chlorophyll fluorescence (using an  
926 emission interval of 650-700 nm) and bright field images were simultaneously visualised for  
927 each cell. Wild-type cells that did not express GFP were used to identify the maximum  
928 exposure length possible without false detection of chlorophyll in the GFP channel (Fig. 2-  
929 figure supplement 7).

930  
931 Possible mitochondrial localisations of dual targeted proteins were identified by staining  
932 cells with approximately 100 mM Mitotracker orange, dissolved in filtered seawater, for 25  
933 minutes under standard culture conditions<sup>55</sup>. Cells were rinsed and resuspended in fresh  
934 filtered seawater prior to visualisation, using the same conditions as stated above for GFP,  
935 and a 548 nm excitation laser and 575-585 nm absorbance window for the Mitotracker  
936 signal. To ensure that there was no possible crosstalk between the two signals, negative  
937 controls consisting of an unstained GFP-expressing wild-type line, and stained wild-type  
938 cells, were used respectively to determine the maximum exposure length possible without  
939 (respectively) false detection of GFP in the Mitotracker channel, and false detection of  
940 Mitotracker in the GFP channel (Fig. 7- figure supplement 1).

941  
942 **Reconstruction of evolutionary origins of ancestral plastid-targeted proteins**

943  
944 The most probable evolutionary origins of individual plastid-targeted proteins were  
945 identified via the combined products of BLAST top hit analysis and phylogenetic sister-group  
946 inference. First, a composite reference sequence library was generated by appending the  
947 uniref outgroup library previously used for BLAST-based assembly of ancestral HPPGs, with  
948 twenty-two combined eukaryotic transcriptome and genomic libraries of taxa with no  
949 suspected history of serial endosymbiosis, which was previously used to enrich each single-  
950 gene tree (Table S1- sheet 1<sup>145</sup>). Each sequence within the library was then assigned a  
951 taxonomic affinity consisting of one of six lineages (green algae, red algae, aplastidic  
952 stramenopiles, all other eukaryotes, prokaryotes, and viruses) and one of 48 sub-categories,  
953 (Table S1- sheet 1, section 1<sup>145</sup>). Next, each seed protein sequence within each ancestral  
954 HPPG was searched by BLASTp against the composite library, with a threshold e-value of 1 x  
955 10<sup>-05</sup>. Sequences were annotated by the lineage and sub-category of the first hit obtained,  
956 and by the number of consecutive top hits obtained within the same lineage (Table S4- sheet  
957 2, section 2<sup>145</sup>). To minimise misidentification due to any residual contamination in individual  
958 sequence libraries, only sequences for which the first three or more BLAST hits resolved  
959 within the same lineage were deemed to be unambiguously related to that lineage.

960  
961 Sister-group relationships were additionally inferred for each ancestral HPPG from the  
962 previously generated single-gene trees (Table S4- sheet 2, section 3<sup>145</sup>). To ensure that only  
963 true sister-group relationships were recorded, and to avoid potential misidentifications of  
964 individual sister-group relationships due to species-specific gene transfer or contaminants

965 that had not previously been excluded by screening individual species libraries, only trees in  
966 which ochrophytes were monophyletic, (i.e., not paraphyletic with regard to any one of the  
967 five outgroups), for which a single sister-group could be identified (using the most  
968 phylogenetically complex node as the outgroup), and for which the sister-group contained at  
969 least two monophyletic or paraphyletic sequences, from different sub-categories of the  
970 same lineage, were used for subsequent analysis.

971

972 **Reconstruction of evolutionary relationships between ochrophytes and other CASH**  
973 **lineage plastids**

974

975 To identify the probable relationships between ochrophytes and other CASH lineage  
976 plastids, each ancestral HPPG tree was enriched with sequences from six different groups of  
977 organisms with histories of serial endosymbiosis (cryptomonads, haptophytes, dinotoms,  
978 other alveolates, euglenids, and chlorarachniophytes), subdivided into thirteen sub-  
979 categories (Table S1<sup>145</sup>). For the cryptomonad, haptophyte and dinotom sequences, as  
980 plastid-targeted proteins from these lineages may be identified using targeting predictors  
981 trained on diatoms such as HECTAR<sup>6</sup> and ASAFind<sup>29, 30</sup>, each of the HPPGs initially generated  
982 was enriched with plastid-targeted sequences from each cryptomonad, haptophyte and  
983 dinotom sub-category identified by *in silico* prediction with these programmes (Table S2-  
984 sheet 1; Table S3- sheet 1<sup>145</sup>).

985

986 The position of each group of organisms within the tree was then annotated as falling into  
987 one of eight different categories, four of which were internal to the ochrophytes (diatoms;  
988 hypogyrstea; chrysista; or an ambiguous internal position) and four of which were external  
989 to the ochrophytes (as an immediate sister-group to all ochrophytes prior to the first  
990 outgroup lineage previously identified; within the red algae; within the green algae; and at  
991 any other position external to the ochrophytes; Table S4- sheet 2, sections 5-6<sup>145</sup>). To  
992 minimise the incorporation of contaminant and non-plastid sequences, tree positions were  
993 only recorded if the branch containing sequences from that particular lineage included at  
994 least two of the sub-categories considered (for alveolates, cryptomonads, and haptophytes),  
995 contained at least one predicted plastid-targeted sequence (for dinotoms, cryptomonads  
996 and haptophytes), and for which only one category could be applied (i.e., the tree only  
997 contained one evolutionarily distinct group for each lineage, which could be unambiguously  
998 allocated one category over all others). Each tree annotation was repeated three times  
999 independently, and only tree annotations that were recorded consistently in each case were  
1000 retained for further analysis.

1001

1002 To identify proteins that were uniquely shared between haptophytes and other lineages,  
1003 every HPPG initially generated was screened for the inclusion of only two of five different  
1004 lineages (diatoms including dinotoms, hypogyrstea, chrysista, haptophytes, and  
1005 cryptomonads; Table S2- sheet 2, section 3; Table S3- sheet 2, section 3<sup>145</sup>). The frequencies  
1006 of these proteins were then compared to the numbers expected in a random distribution of  
1007 all uniquely shared HPPGs across the entire dataset: for example, if half of all uniquely  
1008 shared HPPGs were shared with diatoms and one other lineage, and half were shared with  
1009 haptophytes and one other lineage, then one-quarter of all uniquely shared HPPGs should  
1010 be shared between haptophytes and diatoms.

1011

1012 The specific evolutionary relationships associated with haptophyte plastid-targeted proteins  
1013 incorporated into ancestral HPPGs were investigated using a modified BLAST top hit  
1014 technique. Firstly, all of the plastid-targeted proteins assembled into each ancestral HPPG  
1015 were extracted and separated into each separate sub-category (Table S13- sheet 1<sup>145</sup>). Each

1016 sub-category list was then reduced to only leave one, randomly selected sequence per HPPG  
1017 (Table S13- sheet 2<sup>145</sup>). Finally, each sequence retained in the reduced list was searched by  
1018 BLAST against a composite library, consisting of the library previously used for outgroup top  
1019 hit analysis, enriched with all of the plastid-targeted proteins identified for ochrophytes,  
1020 haptophytes and cryptomonads , except for those that corresponded to the same particular  
1021 lineage as the query sequence (Table S13- sheets 1,3<sup>145</sup>). For example, in the case of  
1022 haptophytes, plastid-targeted sequences that had been separated into three individual  
1023 categories (pavlovophytes, prymnesiales, and isochrysidales<sup>127</sup>) were searched against a  
1024 composite library consisting of all outgroup sequences, and plastid-targeted sequences from  
1025 diatoms, hypogyrstea, chrysista, and cryptomonads, but excluding haptophytes. BLAST top  
1026 hit analysis was then performed as described above (Table S13- sheets 1, 3<sup>145</sup>). Finally, to  
1027 enable the identification of genes with consistent results from multiple analyses, the lineage  
1028 of the BLAST top hit was compared to the lineage of the haptophyte sister-group in the  
1029 single-gene tree analysis (Table S4- sheet 2, section 5; Table S13- sheet 4<sup>145</sup>).  
1030

### 1031 **Identification of uniquely shared residues in multigene HPPG datasets**

1032

1033 To identify residues that are uniquely shared between ochrophytes and other lineages,  
1034 multigene datasets were constructed of a) ancestral HPPGs of green algal origin, and b)  
1035 ancestral HPPGs for which haptophytes show origins within the ochrophytes. To minimise  
1036 the incorporation of sequences of misidentified origin, in each case only the HPPGs for  
1037 which the proposed evolutionary origin were identified both by BLAST top hit and single-  
1038 gene tree analysis were included. To avoid introducing artifacts due to lineage-specific gene  
1039 transfers, paralogy events, or other phylogenetic incongruencies that could otherwise bias  
1040 the eventual results<sup>86, 128</sup>, the single-gene tree generated for each HPPG was manually  
1041 inspected to exclude any that contain multiple clades (defined as monophyletic groups  
1042 containing more than one sequence from a particular lineage, separated from one another  
1043 by at least two sequences from outside that particular lineage) for each of the major  
1044 lineages of interest within the tree:

1045

1046

- 1047 • For the green gene dataset, HPPG trees containing more than one clade of  
1048 ochrophyte, cryptomonad, haptophyte, red algal, or green algal sequences were  
1049 excluded. To account for the possibility that CASH lineage sequences might  
1050 originate from within the green algae, the green algae were allowed to be  
1051 paraphyletic with regard to the cryptomonad, haptophyte and ochrophyte  
1052 sequences, but were not allowed to incorporate sequences from other lineages.  
1053 Similarly, to account for the possibility that subsequent gene transfers may have  
1054 occurred from ochrophytes into other CASH lineages, the ochrophytes were  
1055 allowed to be paraphyletic with regard to cryptomonad and haptophyte sequences,  
1056 but not to any other lineages.
- 1057 • For the haptophyte gene dataset, HPPG trees containing more than one clade of  
1058 ochrophyte, haptophyte, diatom, hypogyrstean, or chrysistan sequences were  
1059 excluded. To account for the possibility that haptophytes arose within the  
1060 ochrophytes, the ochrophyte, diatom, hypogyrstean and chrysistan sequences  
1061 were allowed to incorporate sequences from haptophytes. Similarly, due to the  
1062 paraphyly of hypogyrstea with regard to diatoms, the hypogyrstean sequences  
1063 were allowed to incorporate sequences from diatoms, but not from other lineages.  
1064 • In all cases, sequences from chlorarachniophytes, euglenids, and alveolates were  
1065 not incorporated into any of the clade assessments, due to uncertainty over the  
1066 gene transfer events that have occurred in each lineage<sup>7, 81, 82</sup>.

1067  
1068 This left datasets consisting of 32 HPPGs for which the ochrophytes were of clear green  
1069 algal origin, and 37 HPPGs in which the haptophytes were of clear ochrophyte origin, with  
1070 no conflicting phylogenetic signal. The rationale for inclusion and exclusion of each HPPG in  
1071 each analysis is presented in Table S6, sheets 1 and 3<sup>145</sup>.  
1072  
1073 Next, to eliminate individual sequences remaining within each HPPG that might have arisen  
1074 through species-specific gene transfer or contamination events, each trimmed sequence  
1075 within each approved alignment was inspected using a composite BLAST approach. First,  
1076 each sequence was searched against a composite library containing all uniref, jgi and  
1077 MMETSP sequences from every lineage within the tree of life, and the top ten hits were  
1078 tabulated for each sequence. In each case, only sequences for which at least the first three  
1079 hits were of the same lineage as that of the query were retained. For the haptophyte  
1080 multigene alignment, the ochrophytes were separately analysed as each of the three  
1081 component lineages (chrysista, hypogyristeae, and diatoms), which is to say that a query  
1082 obtained from a member of the hypogyristeae would only be retained if the first three BLAST  
1083 top hits originated from other hypogyristeal sequences, rather than other ochrophytes.  
1084  
1085 Next, each of the component sequences within each cleaned alignment were searched  
1086 against all other component sequences within the same alignment using BLASTp, and the  
1087 top ten hits within the alignment were ranked. In each case, sequences were only approved  
1088 for incorporation into the multigene dataset if the first non-self hit was to a different sub-  
1089 category within the same lineage, e.g. if a query sequence from a red alga yielded a top hit  
1090 against a red algal sequence from a different red sub-category. To allow for possible cases  
1091 of paraphyly and/or absence of sequences within each alignment, the following  
1092 modifications were applied:  
1093  
1094     

- Green algal sequences within the confirmed green origin alignments were allowed  
1095       to yield top hits against ochrophytes, cryptomonads, and haptophytes, but were  
1096       required to yield a best hit against another green alga with an expect value lower  
1097       than the top hit against red algal or glaucophyte sequences.
- Glaucophyte sequences were deemed to be of correct origin if they yielded a top hit  
1098       against cyanobacteria, red algae, or green algae, due to the incorporation (in  
1099       general) of only one glaucophyte sequence in each alignment.
- Ochrophyte sequences were deemed to be of correct origin if they yielded a top hit  
1100       against any other ochrophyte sub-category (regardless of whether this was of  
1101       diatom, hypogyristeal or chrysista origin). Ochrophyte sequences were  
1102       additionally allowed to yield top hits against cryptomonads (in the green gene  
1103       alignments), and haptophytes (in both green and haptophyte gene alignments), but  
1104       were required to yield a best hit against another ochrophyte with an expect value  
1105       lower than the best hit against green algal, red algal or glaucophyte sequences.
- Sequences for which no top hits were found for a different sub-category within the  
1106       same lineage, but for which at least one top hit were found within the same sub-  
1107       category within the lineage, and for which the first ten BLAST hits did not directly  
1108       indicate a contamination event, were deemed to be of correct origin.

  
1113 Tabulated outputs for each BLAST analysis are provided in Table S6, sheets 2 and 4. Finally,  
1114 each dataset was reduced to leave only one randomly selected sequence for each given sub-  
1115 category within each HPPG alignment.  
1116

1117 The number of residues that were uniquely shared between ochrophytes and green algae in  
1118 the green gene dataset, and haptophytes and ochrophytes in the haptophyte dataset, were  
1119 then tabulated (Table S7<sup>145</sup>). Briefly, residues were inferred to be uniquely shared between  
1120 ochrophytes and green algae if they were present in at least 2/3 of the ungapped  
1121 ochrophyte sequences, one or more green algal sequence, and if none of the red algal or  
1122 glaucophyte sequences shared the residue in question, but at least one of these sequences  
1123 had a non-matching (i.e. non-gapped) residue at that position (Table S7- sheet 1, section  
1124 2<sup>145</sup>). Similarly, residues were inferred to be uniquely shared between ochrophytes and  
1125 haptophytes if they were present in at least 2/3 of the ungapped haptophyte sequences,  
1126 one or more ochrophyte sequence, and if none of the green algal, red algal, glaucophyte or  
1127 cyanobacterial sequences shared the residue in question, but at least one of these  
1128 sequences had a non-matching (i.e., non-gapped) residue at that position (Table S7- sheet 2,  
1129 section 2<sup>145</sup>). The origin point of each uniquely shared residue was then inferred by  
1130 comparison to reference topologies respectively of green algae<sup>129</sup> and of ochrophytes (per  
1131 Fig. 1). Residues were assumed to have originated in a common ancestor of a particular  
1132 clade if that clade contained more lineages with matching than non-matching or gapped  
1133 residues (Table S7- sheets 1-2, section 5<sup>145</sup>). A second analysis was additionally performed in  
1134 which all gapped residues were deemed to be matching, to identify the earliest possible  
1135 origin point for each uniquely shared residue, taking into account secondary loss<sup>45, 50</sup> and  
1136 absence of sequences from each alignment<sup>46,47</sup>.  
1137

### 1138 **Analysis of targeting preferences of ancestral ochrophyte and haptophyte genes.**

1139 Two libraries of non-redundant gene families that were broadly conserved across  
1140 ochrophytes or haptophytes, and thus might represent gene products of the ancestral  
1141 genomes of these lineages, were generated using a similar BLAST-based assembly pipeline  
1142 as used to construct HPPGs (Table S8; Table S14<sup>145</sup>). Ochrophyte gene families were deemed  
1143 to be conserved if orthologues were detected in one of three different patterns of  
1144 ochrophyte sub-categories previously defined to correspond to ancestral plastid-targeted  
1145 proteins (Fig. 1, panel B; Table S8- sheet 1, section 3<sup>145</sup>). Haptophyte gene families, built  
1146 through a similar pipeline using seed sequences from the *Chrysochromulina tobin* and  
1147 *Emiliania huxleyi* genomes<sup>116, 117</sup>, were deemed to be ancestral if orthologues were identified  
1148 in at least two of the three haptophyte sub-categories considered (pavlovophytes,  
1149 prymnesiales, and isochrysidales; Table S14- sheet 1, section 3<sup>145</sup>).  
1151

1152 The most probable evolutionary origin of each gene family was inferred by BLAST top hit  
1153 analysis of the seed sequence (Table S8- sheets 1, 2; Table S14- sheets 1, 2<sup>145</sup>). Ochrophyte  
1154 sequences were searched against the composite uniref + MMETSP library used to previously  
1155 identify the most likely outgroup to each ancestral plastid-targeted protein (Table S8- sheet  
1156 1, section 6<sup>145</sup>), while haptophyte sequences were searched against the enriched library that  
1157 also contained all ochrophyte and cryptomonad sequences, to enable the distinction of  
1158 proteins of probable CASH lineage plastid origin from proteins that had evolved through  
1159 independent gene transfer events between haptophytes and non-CASH lineage organisms  
1160 (Table S14- sheet 1, section 6<sup>145</sup>). Targeting preferences for each protein encoded within  
1161 each gene family were identified using SignalP v 3.0 and ASAFind v 2.0<sup>29, 106</sup>, and with  
1162 HECTAR<sup>30</sup>, as previously discussed (Table S8- sheet 3; Table S14- sheet 3<sup>145</sup>). Targeting  
1163 preferences that were identified in a plurality of sequences and in ≥2/3 of the sequences  
1164 within each ochrophyte gene family were recorded (Table S8- sheet 2, sections 4-5<sup>145</sup>). As  
1165 only three haptophyte sequences were assembled for each ancestral haptophyte gene  
1166 family, only targeting predictions that were identified in ≥2/3 of the sequences within the  
1167 HPPG were inferred to be genuine (Table S14- sheet 2, sections 4-5<sup>145</sup>).  
1168

1168  
1169     **Functional and physiological annotation of ancestral plastid-targeted proteins**  
1170  
1171     Core plastid metabolism pathways were identified using recent reviews of ochrophyte  
1172     metabolism, or reviews of homologous plant plastid metabolic pathways where ochrophyte-  
1173     specific reviews have not yet been published<sup>51, 97, 115, 130-136</sup>. The probable function and KOG  
1174     classification of each HPPG were annotated using the pre-existing annotations associated  
1175     with seed protein sequence (if these existed), or if not the annotated function of the top  
1176     uniref hit previously identified by BLAST searches of the seed sequence (Table S9<sup>145</sup>).  
1177     Expression dynamics for each ancestral HPPG within the genomes of the model diatoms  
1178     *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* were inferred using microarray  
1179     data integrated into the DiatomPortal server<sup>74</sup> (Table S10- sheets 1,2<sup>145</sup>). Correlation  
1180     coefficients were calculated between each pair of *P. tricornutum* and *T. pseudonana* genes  
1181     that were incorporated into an ancestral HPPG, across all microarray libraries within the  
1182     dataset (Table S10- sheets 3,4<sup>145</sup>), with average values being calculated from all pairwise  
1183     correlations for different evolutionary categories of protein (Table S10- sheet 5<sup>145</sup>).  
1184  
1185     Possible chimeric proteins, resulting from the fusion of proteins of different evolutionary  
1186     origins, were identified in the dataset using a modified version of a previously published  
1187     protocol<sup>75</sup> (Table S9- sheet 1, sections 4,5; Table S11<sup>145</sup>). Each protein within each HPPG was  
1188     searched using BLASTp against the composite outgroup MMETSP-enriched library, using the  
1189     same taxonomic classification used for the identification of the evolutionary origin of each  
1190     seed protein within the dataset, and all hits with an expect value of  $1 \times 10^{-05}$ . Component  
1191     sequences were then grouped into component families according to the following rule: if  
1192     two component sequences overlapped by more than 70% of their lengths on the protein  
1193     composite, they belonged to the same component family. Overlapping and/ or nested  
1194     component families were additionally merged if one family was included by more than 70%  
1195     of its length into the other one. Component families were then assigned a broad  
1196     evolutionary origin corresponding to their taxonomic composition. If the three best  
1197     component sequences, according to their BLAST bitscore against the composite gene,  
1198     matched with the same lineage (e.g., green algae, red algae, aplastidic stramenopiles, or  
1199     other eukaryotes), the component was considered to have originated from that lineage.  
1200  
1201     Possible dual targeted proteins were identified within the dataset by screening all possible  
1202     plastid-targeted proteins with Mitofates, using a cut-off targeting threshold of 0.35<sup>137</sup>, which  
1203     was inferred to be more effective in identifying experimentally verified ochrophyte  
1204     mitochondria-targeted proteins (Fig. 7- figure supplement 2)<sup>29</sup> than other threshold values  
1205     or targeting prediction programmes such as TargetP<sup>138</sup> or Mitoprot<sup>139</sup>. The default Mitofates  
1206     positive cutoff value was modified from 0.38 to 0.35 in order to maximise the capture of  
1207     experimentally localised mitochondrial proteins, without admitting proteins with  
1208     unambiguous plastid localisation (Fig. 7- figure supplement 2). As dual targeting to plastids  
1209     and mitochondria may be achieved either by distinct protein isoforms resulting from  
1210     ambiguous targeting peptides or alternative internal translation initiation sites that allow  
1211     production of mitochondrial targeting sequences<sup>77, 80</sup>, each protein was screened with  
1212     Mitofates using both the full-length N-termini, and N-termini predicted to result from the  
1213     next downstream methionine within 30 residues. Possible conserved dual targeted proteins  
1214     were then identified via the same BLAST-based assembly pipeline and stringency thresholds  
1215     used to identify probable ancestral HPPGs (Table S12- sheet 1<sup>145</sup>). All putative dual targeted  
1216     proteins have been made publically accessible through the University of Cambridge dSpace  
1217     server (<https://www.repository.cam.ac.uk/handle/1810/261421>)<sup>145</sup>.  
1218

1219 **Construction and inspection of concatenated and exemplar phylogenetic trees**

1220

1221 For the plastid genome phylogenetic analysis, single-gene alignments were constructed by  
1222 BLAST searches of published red lineage and glaucophyte plastid genomes (for the gene rich  
1223 analysis) or of these genomes plus all MMETSP libraries for the same lineages (for the taxon  
1224 rich analysis), using the *Phaeodactylum tricornutum* protein sequence as query and a  
1225 threshold e-value of  $1 \times 10^{-05}$ , followed by alignment using GeneIOUS v 4.76<sup>119</sup>, as before.  
1226 The gene rich analysis included protein sequences from 54 genes that were identified in 22  
1227 different non-green lineage plastid genomes while the taxon-rich analysis included 10  
1228 different plastid genes that were identified in all 22 plastid genomes and at least 30 different  
1229 MMETSP libraries<sup>34</sup> (Table S15- sheet 1<sup>145</sup>). For the taxon-rich analysis, only species that  
1230 were represented in  $\geq 6/12$  of the single-gene alignments were included in the concatenated  
1231 alignment. Each concatenated alignment was trimmed using trimAl<sup>120</sup> using the -gt 0.8  
1232 option.

1233

1234 Single-gene alignments for four plastid-targeted proteins predicted to be of polyphyletic  
1235 origin in ochrophytes (3-dehydroquinate synthase, isopropylmalate dehydratase,  
1236 sedoheptulose bisphosphatase, and shikimate kinase) were generated using a similar BLAST-  
1237 based assembly and alignment pipeline as used to verify ancestral plastid-targeted proteins.  
1238 In this case, all non-redundant (as inferred by BLAST top hit evalue) plastid-targeted  
1239 sequences for each protein identified from ochrophyte genomes were used as independent  
1240 queries for the identification of plastid-targeted orthologues, 50 uniref top hits, and top hits  
1241 from the combined MMETSP and genomic libraries from 36 eukaryotic sub-categories, as  
1242 before. HPPGs were independently generated, aligned and trimmed for each seed sequence;  
1243 all HPPGs generated for each protein were then merged, realigned and retrimmed using  
1244 trimAl to generate a single-gene alignment. Single-gene alignments for each of the  
1245 constituent genes in each concatenated plastid genome tree were generated by splitting the  
1246 alignment into its component genes. All alignments have been made publically accessible  
1247 through the University of Cambridge dSpace server  
1248 (<https://www.repository.cam.ac.uk/handle/1810/261421>)<sup>145</sup>.

1249

1250 Trees were inferred for each concatenated and exemplar single-gene alignment (Table S15-  
1251 sheet 2<sup>145</sup>) using the MrBayes and RAxML programmes in-built into the CIPRES web-  
1252 server<sup>121, 140, 141</sup>. Bayesian trees were inferred using three substitution models (GTR, Jones,  
1253 and WAG), a minimum of 600000 generations, and an initial burn-in discard value of 0.5.  
1254 Trees were only utilised if the final convergence statistic between the two chains run was  $\leq$   
1255 0.1, and tree calculation was automatically stopped if the convergence statistic fell below  
1256 0.01. RAxML trees were inferred using three substitution models (GTR, JTT, and WAG) with  
1257 automatic bootstrapping, as previously described<sup>58</sup>. The best tree topology for each RAxML  
1258 tree was inferred, and bootstrapping was performed using a burnin value of 0.03.  
1259 Alternative tree topologies were tested for the RAxML + JTT tree inferred from each  
1260 concatenated alignment using CONSEL<sup>142</sup>, under the default conditions. Tree outputs have  
1261 been made publically accessible through the University of Cambridge dSpace server  
1262 (<https://www.repository.cam.ac.uk/handle/1810/261421>)<sup>145</sup>.

1263

1264 Modified alignments were generated for both of the plastid concatenated multigene  
1265 datasets from which individual clades of organisms (diatoms, hypoglycista, chrysista,  
1266 haptophytes, cryptomonads, red algae, and different combinations of green algae) had been  
1267 removed (Table S15- sheet 2<sup>145</sup>). Fast-site removal was performed using TIGER<sup>143</sup>. Site rate  
1268 evolution characteristics were calculated for each alignment using the -b 100 option, and  
1269 modified alignments were constructed from which the rate categories corresponding to the

1270 fastest evolving 40-50% of sites were serially removed (Table S15- sheet 2<sup>145</sup>). Amino acid  
1271 composition for each plastid alignment were calculated, and two modified alignments were  
1272 generated from which glycines (which in all alignments occur at significantly lower  
1273 frequencies in ochrophytes than in haptophytes or cryptomonads; chi-squared, P≤ 0.05;  
1274 Table S16- sheet 3<sup>145</sup>), and from which seven amino acids (alanine, aspartate, glycine,  
1275 histidine, leucine, asparagine, threonine and valine) which were found in at least one  
1276 alignment to occur at significantly different frequencies in ochrophytes compared to  
1277 haptophytes or to cryptomonads (P≤ 0.05; Table S16- sheet 3<sup>145</sup>) had been removed. Trees  
1278 were inferred for each modified alignment using RAxML with the JTT substitution, and  
1279 MrBayes with the Jones substitution, and bootstrap calculation as previously described.  
1280 Modified alignments and tree outputs have been made publically accessible through the  
1281 University of Cambridge dSpace server  
1282 (<https://www.repository.cam.ac.uk/handle/1810/261421>)<sup>145</sup>.  
1283

1284 Uniquely shared residues were manually tabulated for both of the plastid genome multigene  
1285 alignments (Table S17<sup>145</sup>). For the gene-rich plastid multigene alignment, residues that were  
1286 present in all haptophyte sequences and only found in a maximum of one other lineage (red  
1287 algae, glaucophytes, cryptomonads, diatoms, hypogyristera, or chrysista) were tabulated  
1288 (Table S17- sheet 1<sup>145</sup>). For the taxon-rich alignment, to take into account gaps and missing  
1289 characters, residues were tabulated if they were found in a majority of haptophyte  
1290 sequences, and one other lineage, as before (Table S17- sheet 2<sup>145</sup>). The total number of  
1291 residues shared, and uniquely shared, with each non-haptophyte species and lineage are  
1292 respectively tabulated in Table S17, sheets 3 and 4<sup>145</sup>.  
1293

#### 1294 **Data deposition**

1295 All supporting datasets for this study, including supplementary tables predicted plastid-  
1296 targeted and dual targeted protein libraries, single gene and multigene alignments, and tree  
1297 outputs, have been made publically and freely accessible through the University of  
1298 Cambridge dSpace server (<https://www.repository.cam.ac.uk/handle/1810/261421>)<sup>145</sup>.  
1299

#### 1300 **Acknowledgments**

1301 The authors would like to thank Achal Rastogi (École Normale Supérieure), Neal Clarke (Yale  
1302 University), Michael Melkonian (University of Koln), Gane Ka-Shu Wong (University of  
1303 Alberta) and Jun Yu (Beijing Institute of Genomics) for early access to sequence data used in  
1304 this study, and Catherine Cantrel, Anne-Flore Deton-Cabanillas, Zhanru Shao, Leïla Tirichine  
1305 and Javier Paz-Yepes (École Normale Supérieure) for assistance with generation of  
1306 transgenic expression constructs for *Phaeodactylum tricornutum*. Funding is acknowledged  
1307 from the ERC Advanced Award “Diatomite”, the Louis D Foundation of the Institut de  
1308 France, the French Government “Investissements d’Avenir” programmes MEMO LIFE (ANR-  
1309 10-LABX-54) and PSL\* Research University (ANR-11-IDEX-0001-02), and the Gordon and  
1310 Betty Moore Foundation (all to CB), and from FP7 (grant number 2007-2013 Grant  
1311 Agreement 615274, to EPB). RGD is supported by an EMBO early career fellowship (ALTF  
1312 1124-2014). DJR was supported by a postdoctoral fellowship from the Conseil Régional de  
1313 Bretagne and the French Government “Investissements d’Avenir” programme OCEANOMICS  
1314 (ANR-11-BTBR-0008). The authors would like to thank the handling reviewer and two  
1315 anonymous editors for constructive comments on the manuscript.  
1316

#### 1317 **Competing interests.**

1318

1321 The authors declare no competing financial or non-financial interests in this project.

1322

1323 **References**

1324

- 1325 1. Dorrell, R.G. & Smith, A.G. Do red and green make brown?: Perspectives on plastid  
1326 acquisitions within chromalveolates. *Eukaryot Cell* **10**, 856-868 (2011).
- 1327 2. de Vargas, C. *et al.* Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean.  
*Science* **348**, 1261605 (2015).
- 1328 3. Dorrell, R.G. & Howe, C.J. What makes a chloroplast? Reconstructing the  
1329 establishment of photosynthetic symbioses. *J Cell Sci* **125**, 1865-1875 (2012).
- 1330 4. Baurain, D. *et al.* Phylogenomic evidence for separate acquisition of plastids in  
1331 cryptophytes, haptophytes, and stramenopiles. *Mol Biol Evol* **27**, 1698-1709 (2010).
- 1332 5. Stiller, J.W. *et al.* The evolution of photosynthesis in chromist algae through serial  
1333 endosymbioses. *Nat Commun* **5**, 5764 (2014).
- 1334 6. Aleoshin, V.V., Mylnikov, A.P., Mirzaeva, G.S., Mikhailov, K.V. & Karpov, S.A.  
1335 Heterokont predator *Develorapax marinus* gen. et sp. nov. - a model of the  
1336 ochrophyte ancestor. *Front Microbiol* **7**, 1194 (2016).
- 1337 7. Ševčíková, T. *et al.* Updating algal evolutionary relationships through plastid genome  
1338 sequencing: did alveolate plastids emerge through endosymbiosis of an ochrophyte?  
*Sci Rep* **5**, 10134 (2015).
- 1339 8. Bowler, C., Vardi, A. & Allen, A.E. Oceanographic and biogeochemical insights from  
1340 diatom genomes. *Ann Rev Mar Sci* **2**, 333-365 (2010).
- 1341 9. Armbrust, E.V. *et al.* The genome of the diatom *Thalassiosira pseudonana*: ecology,  
1342 evolution, and metabolism. *Science* **306**, 79-86 (2004).
- 1343 10. Cock, J.M. *et al.* The *Ectocarpus* genome and the independent evolution of  
1344 multicellularity in brown algae. *Nature* **465**, 617-621 (2010).
- 1345 11. Gobler, C.J. *et al.* Niche of harmful alga *Aureococcus anophagefferens* revealed  
1346 through ecogenomics. *Proc Natl Acad Sci USA* **108**, 4352-4357 (2011).
- 1347 12. Derelle, R., López-García, P., Timpano, H. & Moreira, D. A phylogenomic framework  
1348 to study the diversity and evolution of stramenopiles (=Heterokonts). *Mol Biol Evol*  
1349 **33**, 2890-2898 (2016).
- 1350 13. Ruck, E.C., Nakov, T., Jansen, R.K., Theriot, E.C. & Alverson, A.J. Serial gene losses  
1351 and foreign DNA underlie size and sequence variation in the plastid genomes of  
1352 diatoms. *Genom Biol Evol* (2014).
- 1353 14. Stegemann, S., Hartmann, S., Ruf, S. & Bock, R. High-frequency gene transfer from  
1354 the chloroplast genome to the nucleus. *Proc Natl Acad Sci USA* **100**, 8828-8833  
1355 (2003).
- 1356 15. Nowack, E.C. & Grossman, A.R. Trafficking of protein into the recently established  
1357 photosynthetic organelles of *Paulinella chromatophora*. *Proc Natl Acad Sci USA* **109**,  
1358 5340-5345 (2012).
- 1359 16. Kleffmann, T. *et al.* The *Arabidopsis thaliana* chloroplast proteome reveals pathway  
1360 abundance and novel protein functions. *Curr Biol* **14**, 354-362 (2004).
- 1361 17. Curtis, B.A. *et al.* Algal genomes reveal evolutionary mosaicism and the fate of  
1362 nucleomorphs. *Nature* **492**, 59-65 (2012).
- 1363 18. Qiu, H. *et al.* Assessing the bacterial contribution to the plastid proteome. *Trends  
1364 Plant Sci* **18**, 680-687 (2013).
- 1365 19. Morse, D., Salois, P., Markovic, P. & Hastings, J.W. A nuclear-encoded form II  
1366 RuBisCO in dinoflagellates. *Science* **268**, 1622-1624 (1995).
- 1367 20. Dorrell, R.G. & Howe, C.J. Integration of plastids with their hosts: lessons learnt from  
1368 dinoflagellates *Proc Natl Acad Sci USA* **112**, 10247-10254 (2015).

1369

1370

- 1371 21. Dorrell, R.G. & Howe, C.J. Functional remodeling of RNA processing in replacement  
1372 chloroplasts by pathways retained from their predecessors. *Proc Natl Acad Sci USA*  
1373 **109**, 18879-18884 (2012).
- 1374 22. Fast, N.M., Kissinger, J.C., Roos, D.S. & Keeling, P.J. Nuclear-encoded, plastid-  
1375 targeted genes suggest a single common origin for apicomplexan and dinoflagellate  
1376 plastids. *Mol Biol Evol* **18**, 418-426 (2001).
- 1377 23. Harper, J.T. & Keeling, P.J. Nucleus-encoded, plastid-targeted glyceraldehyde-3-  
1378 phosphate dehydrogenase (GAPDH) indicates a single origin for chromalveolate  
1379 plastids. *Mol Biol Evol* **20**, 1730-1735 (2003).
- 1380 24. Nowack, E.C.M. *et al.* Gene transfers from diverse bacteria compensate for reductive  
1381 genome evolution in the chromatophore of *Paulinella chromatophora*. *Proc Natl*  
1382 *Acad Sci USA* **113**, 12214-12219 (2016).
- 1383 25. Dunning Hotopp, J.C. *et al.* Widespread lateral gene transfer from intracellular  
1384 bacteria to multicellular eukaryotes. *Science* **317**, 1753-1756 (2007).
- 1385 26. Gornik, S.G. *et al.* Loss of nucleosomal DNA condensation coincides with appearance  
1386 of a novel nuclear protein in dinoflagellates. *Curr Biol* **22**, 2303-2312 (2012).
- 1387 27. Prechtl, J., Kneip, C., Lockhart, P., Wenderoth, K. & Maier, U.G. Intracellular spheroid  
1388 bodies of *Rhopalodia gibba* have nitrogen-fixing apparatus of cyanobacterial origin.  
1389 *Mol Biol Evol* **21**, 1477-1481 (2004).
- 1390 28. Thompson, A.W. *et al.* Unicellular cyanobacterium symbiotic with a single-celled  
1391 eukaryotic alga. *Science* **337**, 1546-1550 (2012).
- 1392 29. Gruber, A., Rocap, G., Kroth, P.G., Armbrust, E.V. & Mock, T. Plastid proteome  
1393 prediction for diatoms and other algae with secondary plastids of the red lineage.  
1394 *Plant J* **81**, 519-528 (2015).
- 1395 30. Gschloessl, B., Guermeur, Y. & Cock, J.M. HECTAR: a method to predict subcellular  
1396 targeting in heterokonts. *BMC Bioinformat* **9**, 393 (2008).
- 1397 31. Fuss, J., Liegmann, O., Krause, K. & Rensing, S.A. Green Targeting Predictor and  
1398 Ambiguous Targeting Predictor 2: the pitfalls of plant protein targeting prediction  
1399 and of transient protein expression in heterologous systems. *New Phytol* **200**, 1022-  
1400 1033 (2013).
- 1401 32. Suzuki, K. & Miyagishima, S. Eukaryotic and eubacterial contributions to the  
1402 establishment of plastid proteome estimated by large-scale phylogenetic analyses.  
1403 *Mol Biol Evol* **27**, 581-590 (2010).
- 1404 33. Mock, T. *et al.* Evolutionary genomics of the cold-adapted diatom *Fragilariaopsis*  
1405 *cylindrus*. *Nature* **541**, 536-540 (2017).
- 1406 34. Keeling, P.J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing  
1407 Project (MMETSP): illuminating the functional diversity of eukaryotic life in the  
1408 oceans through transcriptome sequencing. *PLoS Biol* **12**, 1001889 (2014).
- 1409 35. Siaut, M. *et al.* Molecular toolbox for studying diatom biology in *Phaeodactylum*  
1410 *tricornutum*. *Gene* **406**, 23-35 (2007).
- 1411 36. Takahashi, F. *et al.* AUREOCHROME, a photoreceptor required for  
1412 photomorphogenesis in stramenopiles. *Proc Natl Acad Sci USA* **104**, 19625-19630  
1413 (2007).
- 1414 37. Radakovits, R. *et al.* Draft genome sequence and genetic transformation of the  
1415 oleaginous alga *Nannochloropsis gaditana*. *Nat Comms* **4**, 686 (2013).
- 1416 38. Janouskovec, J., Horák, A., Oborník, M., Lukes, J. & Keeling, P.J. A common red algal  
1417 origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proc Natl Acad*  
1418 *Sci USA* **107**, 10949-10954 (2010).
- 1419 39. Burki, F. *et al.* Untangling the early diversification of eukaryotes: a phylogenomic  
1420 study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. *Proc*  
1421 *Biol Sci* **283** (2016).

- 1422 40. Cavalier-Smith, T., Chao, E.E. & Lewis, R. Multiple origins of Heliozoa from flagellate  
1423 ancestors: New cryptist subphylum Corbihelia, superclass Corbistoma, and  
1424 monophyly of Haptista, Cryptista, Hacrobia and Chromista. *Mol Phylogenet Evol* **93**,  
1425 331-362 (2015).
- 1426 41. Yabuki, A. *et al.* *Palpitomonas bilix* represents a basal cryptist lineage: insight into  
1427 the character evolution in Cryptista. *Sci Rep* **4**, 4641 (2014).
- 1428 42. Frommolt, R. *et al.* Ancient recruitment by chromists of green algal genes encoding  
1429 enzymes for carotenoid biosynthesis. *Mol Biol Evol* **25**, 2653-2667 (2008).
- 1430 43. Petersen, J., Teich, R., Brinkmann, H. & Cerff, R. A "green" phosphoribulokinase in  
1431 complex algae with red plastids: evidence for a single secondary endosymbiosis  
1432 leading to haptophytes, cryptophytes, heterokonts, and dinoflagellates. *J Mol Evol*  
1433 **62**, U143-U142 (2006).
- 1434 44. Moustafa, A. *et al.* Genomic footprints of a cryptic plastid endosymbiosis in diatoms.  
1435 *Science* **324**, 1724-1726 (2009).
- 1436 45. Ku, C. *et al.* Endosymbiotic origin and differential loss of eukaryotic genes. *Nature*  
1437 **524**, 427-432 (2015).
- 1438 46. Woehle, C., Dagan, T., Martin, W.F. & Gould, S.B. Red and problematic green  
1439 phylogenetic signals among thousands of nuclear genes from the photosynthetic  
1440 and apicomplexa-related *Chromera velia*. *Genom Biol Evol* **3**, 1220-1230 (2011).
- 1441 47. Deschamps, P. & Moreira, D. Re-evaluating the green contribution to diatom  
1442 genomes. *Genom Biol Evol* **4**, 683-688 (2012).
- 1443 48. Matsuzaki, M. *et al.* Genome sequence of the ultrasmall unicellular red alga  
1444 *Cyanidioschyzon merolae* 10D. *Nature* **428**, 653-657 (2004).
- 1445 49. Collén, J. *et al.* Genome structure and metabolic features in the red seaweed  
1446 *Chondrus crispus* shed light on evolution of the Archaeplastida. *Proc Natl Acad Sci  
1447 USA* **110**, 5247-5252 (2013).
- 1448 50. Qiu, H., Price, D.C., Yang, E.C., Yoon, H.S. & Bhattacharya, D. Evidence of ancient  
1449 genome reduction in red algae (Rhodophyta). *J Phycol* **51**, 624-636 (2015).
- 1450 51. Smith, S.R., Abbriano, R.M. & Hildebrand, M. Comparative analysis of diatom  
1451 genomes reveals substantial differences in the organization of carbon partitioning  
1452 pathways. *Algal Res* **1**, 2-16 (2012).
- 1453 52. Wolfe-Simon, F., Starovoytov, V., Reinfelder, J.R., Schofield, O. & Falkowski, P.G.  
1454 Localization and role of manganese superoxide dismutase in a marine diatom. *Plant  
1455 Physiol* **142**, 1701-1709 (2006).
- 1456 53. Gillard, J. *et al.* Physiological and transcriptomic evidence for a close coupling  
1457 between chloroplast ontogeny and cell cycle progression in the pennate diatom  
1458 *Seminavis robusta*. *Plant Physiol* **148**, 1394-1411 (2008).
- 1459 54. Ramirez, R.A., Espinoza, B. & Kwok, E.Y. Identification of two novel type 1  
1460 peroxisomal targeting signals in *Arabidopsis thaliana*. *Acta Histochem* **116**, 1307-  
1461 1312 (2014).
- 1462 55. Tanaka, A. *et al.* Ultrastructure and membrane traffic during cell division in the  
1463 marine pennate diatom *Phaeodactylum tricornutum*. *Protist* **166**, 506-521 (2015).
- 1464 56. Imanian, B., Pombert, J.F. & Keeling, P.J. The complete plastid genomes of the two  
1465 'Dinotoms' *Durinskia baltica* and *Kryptoperidinium foliaceum*. *PLoS One* **5**, 10711  
1466 (2010).
- 1467 57. Marron, A.O. *et al.* The evolution of silicon transport in eukaryotes. *Mol Biol Evol* **33**,  
1468 3226-3248 (2016).
- 1469 58. Dorrell, R.G. *et al.* Progressive and biased divergent evolution underpins the origin  
1470 and diversification of peridinin dinoflagellate plastids. *Mol Biol Evol* **34**, 361-379  
1471 (2017).

- 1472 59. Bowler, C. *et al.* The *Phaeodactylum* genome reveals the evolutionary history of  
1473 diatom genomes. *Nature* **456**, 239-244 (2008).
- 1474 60. Stiller, J.W., Huang, J.L., Ding, Q., Tian, J. & Goodwillie, C. Are algal genes in  
1475 nonphotosynthetic protists evidence of historical plastid endosymbioses? *BMC*  
1476 *Genom* **10**, 484 (2009).
- 1477 61. Nakamura, Y. *et al.* The first symbiont-free genome sequence of marine red alga,  
1478 *Susabi-nori* (*Pyropia yezoensis*). *PLoS One* **8**, 57122 (2013).
- 1479 62. Bhattacharya, D. *et al.* Genome of the red alga *Porphyridium purpureum*. *Nat*  
1480 *Comms* **4** (2013).
- 1481 63. Schöcknecht, G. *et al.* Gene transfer from bacteria and archaea facilitated evolution  
1482 of an extremophilic eukaryote. *Science* **339**, 1207-1210 (2013).
- 1483 64. Matasci, N. *et al.* Data access for the 1,000 Plants (1KP) project. *Gigascience* **3**, 17  
1484 (2014).
- 1485 65. Keeling, P.J. & Palmer, J.D. Horizontal gene transfer in eukaryotic evolution. *Nat Rev*  
1486 *Genet* **9**, 605-618 (2008).
- 1487 66. Doolittle, W.E. You are what you eat: a gene transfer ratchet could account for  
1488 bacterial genes in eukaryotic nuclear genomes. *Trends Genet* **14**, 307-311 (1998).
- 1489 67. Ershov, Y., Gantt, R.R., Cunningham, F.X. & Gantt, E. Isopentenyl diphosphate  
1490 isomerase deficiency in *Synechocystis* sp. strain PCC6803. *FEBS Lett* **473**, 337-340  
1491 (2000).
- 1492 68. Rohdich, F. *et al.* Studies on the nonmevalonate terpene biosynthetic pathway:  
1493 metabolic role of IspH (LytB) protein. *Proc Natl Acad Sci USA* **99**, 1158-1163 (2002).
- 1494 69. Gutierrez-Marcos, J.F., Roberts, M.A., Campbell, E.I. & Wray, J.L. Three members of a  
1495 novel small gene-family from *Arabidopsis thaliana* able to complement functionally  
1496 an *Escherichia coli* mutant defective in PAPS reductase activity encode proteins with  
1497 a thioredoxin-like domain and "APS reductase" activity. *Proc Natl Acad Sci USA* **93**,  
1498 13377-13382 (1996).
- 1499 70. Dittami, S.M., Michel, G., Collén, J., Boyen, C. & Tonon, T. Chlorophyll-binding  
1500 proteins revisited--a multigenic family of light-harvesting and stress proteins from a  
1501 brown algal perspective. *BMC Evol Biol* **10**, 365 (2010).
- 1502 71. Coesel, S., Obornik, M., Varela, J., Falciatore, A. & Bowler, C. Evolutionary origins and  
1503 functions of the carotenoid biosynthetic pathway in marine diatoms. *Plos One* **3**,  
1504 2896 (2008).
- 1505 72. Chan, C.X. *et al.* Red and green algal monophyly and extensive gene sharing found in  
1506 a rich repertoire of red algal genes. *Curr Biol* **21**, 328-333 (2011).
- 1507 73. Yurchenko, T., Sevcikova, T., Strnad, H., Butenko, A. & Elias, M. The plastid genome  
1508 of some eustigmatophyte algae harbours a bacteria-derived six-gene cluster for  
1509 biosynthesis of a novel secondary metabolite. *Open Biol* **6**, 11 (2016).
- 1510 74. Ashworth, J., Turkarslan, S., Harris, M., Orellana, M.V. & Baliga, N.S. Pan-  
1511 transcriptomic analysis identifies coordinated and orthologous functional modules in  
1512 the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*. *Mar Genom*  
1513 **26**, 21-28 (2016).
- 1514 75. Méheust, R., Zelzion, E., Bhattacharya, D., Lopez, P. & Bapteste, E. Protein networks  
1515 identify novel symbiogenetic genes resulting from plastid endosymbiosis. *Proc Natl*  
1516 *Acad Sci USA* **113**, 3579-3584 (2016).
- 1517 76. Herz, S., Eberhardt, S. & Bacher, A. Biosynthesis of riboflavin in plants. The *ribA* gene  
1518 of *Arabidopsis thaliana* specifies a bifunctional GTP cyclohydrolase II/3,4-dihydroxy-  
1519 2-butanone 4-phosphate synthase. *Phytochem* **53**, 723-731 (2000).
- 1520 77. Xu, L., Carrie, C., Law, S.R., Murcha, M.W. & Whelan, J. Acquisition, conservation,  
1521 and loss of dual-targeted proteins in land plants. *Plant Physiol* **161**, 644-662 (2013).

- 1522 78. Duchene, A.M. *et al.* Dual targeting is the rule for organellar aminoacyl-tRNA  
1523 synthetases in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* **102**, 16484-16489  
1524 (2005).
- 1525 79. Gile, G.H., Moog, D., Slamovits, C.H., Maier, U.G. & Archibald, J.M. Dual organellar  
1526 targeting of aminoacyl-tRNA synthetases in diatoms and cryptophytes. *Genom Biol  
1527 Evol* **7**, 1728-1742 (2015).
- 1528 80. Hirakawa, Y., Burki, F. & Keeling, P.J. Dual targeting of aminoacyl-tRNA synthetases  
1529 to the mitochondrion and complex plastid in chlorarachniophytes. *J Cell Sci* **125**,  
1530 6176-6184 (2012).
- 1531 81. Maruyama, S., Suzuki, T., Weber, A.P.M., Archibald, J.M. & Nozaki, H. Eukaryote-to-  
1532 eukaryote gene transfer gives rise to genome mosaicism in euglenids. *BMC Evol Biol*  
1533 **11**, 105 (2011).
- 1534 82. Archibald, J.M., Rogers, M.B., Toop, M., Ishida, K. & Keeling, P.J. Lateral gene  
1535 transfer and the evolution of plastid-targeted proteins in the secondary plastid-  
1536 containing alga *Bigelowiella natans*. *Proc Natl Acad Sci USA* **100**, 7678-7683 (2003).
- 1537 83. Khan, H. *et al.* Plastid genome sequence of the cryptophyte alga *Rhodomonas salina*  
1538 CCMP1319: lateral transfer of putative DNA replication machinery and a test of  
1539 chromist plastid phylogeny. *Mol Biol Evol* **24**, 1832-1842 (2007).
- 1540 84. Le Corguillé, G. *et al.* Plastid genomes of two brown algae, *Ectocarpus siliculosus* and  
1541 *Fucus vesiculosus*: further insights on the evolution of red-algal derived plastids.  
1542 *BMC Evol Biol* **9**, 253 (2009).
- 1543 85. Guo, Z.H. & Stiller, J.W. Comparative genomics and evolution of proteins associated  
1544 with RNA polymerase IIC-terminal domain. *Mol Biol Evol* **22**, 2166-2178 (2005).
- 1545 86. Qiu, H., Yang, E.C., Bhattacharya, D. & Yoon, H.S. Ancient gene paralogy may mislead  
1546 inference of plastid phylogeny. *Mol Biol Evol* **29**, 3333-3343 (2012).
- 1547 87. Brown, J.W. & Sorhannus, U. A molecular genetic timescale for the diversification of  
1548 autotrophic stramenopiles (Ochrophyta): substantive underestimation of putative  
1549 fossil ages. *PLoS One* **5**, 12759 (2010).
- 1550 88. Matari, N.H. & Blair, J.E. A multilocus timescale for oomycete evolution estimated  
1551 under three distinct molecular clock models. *BMC Evol Biol* **14**, 101 (2014).
- 1552 89. Porter, B.W., Yuen, C.Y.L. & Christopher, D.A. Dual protein trafficking to secretory  
1553 and non-secretory cell compartments: clear or double vision? *Plant Sci* **234**, 174-179  
1554 (2015).
- 1555 90. Pham, J.S. *et al.* A dual-targeted aminoacyl-tRNA synthetase in *Plasmodium*  
1556 *falciparum* charges cytosolic and apicoplast tRNA(Cys). *Biochem J* **458**, 513-523  
1557 (2014).
- 1558 91. Krause, K., Oetke, S. & Krupinska, K. Dual targeting and Retrograde Translocation:  
1559 Regulators of Plant Nuclear Gene Expression Can Be Sequestered by Plastids. *Int J  
1560 Mol Sci* **13**, 11085-11101 (2012).
- 1561 92. Liu, X.J. *et al.* Addressing various compartments of the diatom model organism  
1562 *Phaeodactylum tricornutum* via sub-cellular marker proteins. *Algal Res* **20**, 249-257  
1563 (2016).
- 1564 93. Parfrey, L.W., Lahr, D.J., Knoll, A.H. & Katz, L.A. Estimating the timing of early  
1565 eukaryotic diversification with multigene molecular clocks. *Proc Natl Acad Sci USA*  
1566 **108**, 13624-13629 (2011).
- 1567 94. Bown, P.R. *Calcareous Nannofossil Biostratigraphy*. (Kluwer Academic, London;  
1568 1998).
- 1569 95. Rice, D.W. & Palmer, J.D. An exceptional horizontal gene transfer in plastids: gene  
1570 replacement by a distant bacterial paralog and evidence that haptophyte and  
1571 cryptophyte plastids are sisters. *BMC Biol* **4**, 31 (2006).

- 1572 96. de Vries, J. & Gould, S.B. The monoplasmidic bottleneck in algae and plant evolution. *bioRxiv* (2017).
- 1573 97. Green, B.R. Chloroplast genomes of photosynthetic eukaryotes. *Plant J* **66**, 34-44 (2011).
- 1574 98. Worden, A.Z. *et al.* Global distribution of a wild alga revealed by targeted metagenomics. *Curr Biol* **22**, R675-677 (2012).
- 1575 99. Ong, H.C. *et al.* Analyses of the complete chloroplast genomes of two members of the pelagophyceae: *Aureococcus anophagefferens* CCMP1984 and *Aureoumbra lagunensis* CCMP1507. *J Phycol* **46**, 602-615 (2010).
- 1576 100. Choi, C.J. *et al.* Newly discovered deep-branching marine plastid lineages are numerically rare but globally distributed. *Curr Biol* **27**, 15-16 (2017).
- 1577 101. Kim, E. *et al.* Newly identified and diverse plastid-bearing branch on the eukaryotic tree of life. *Proc Natl Acad Sci USA* **108**, 1496-1500 (2011).
- 1578 102. Lommer, M. *et al.* Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genom Biol* **13** (2012).
- 1579 103. Tanaka, T. *et al.* Oil accumulation by the oleaginous diatom *Fistulifera solaris* as revealed by the genome and transcriptome. *Plant Cell* **27**, 162-176 (2015).
- 1580 104. Galachyants, Y.P. *et al.* Sequencing of the complete genome of an araphid pennate diatom *Synedra acus* subsp. radians from Lake Baikal. *Dokl Biochem Biophys* **461**, 84-88 (2015).
- 1581 105. Wang, D.M. *et al.* *Nannochloropsis* genomes reveal evolution of microalgal oleaginous traits. *PLoS Genet* **10**, 1004094 (2014).
- 1582 106. Bendtsen, J.D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 783-795 (2004).
- 1583 107. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucl Acids Res* **44**, W3-W10 (2016).
- 1584 108. Suzek, B.E., Huang, H.Z., McGarvey, P., Mazumder, R. & Wu, C.H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatic* **23**, 1282-1288 (2007).
- 1585 109. Mangot, J.F. *et al.* Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Scient Reports* **7** (2017).
- 1586 110. Kessenich, C.R., Ruck, E.C., Schurko, A.M., Wickett, N.J. & Alverson, A.J. Transcriptomic insights into the life history of bolidophytes, the sister lineage to diatoms. *J Phycol* **50**, 977-983 (2014).
- 1587 111. Sorhannus, U. & Fox, M.G. Phylogenetic analyses of a combined data set suggest that the *Attheya* lineage is the closest living relative of the pennate diatoms (Bacillariophyceae). *Protist* **163**, 252-262 (2012).
- 1588 112. Yang, E.C. *et al.* Supermatrix data highlight the phylogenetic relationships of photosynthetic stramenopiles. *Protist* **163**, 217-231 (2012).
- 1589 113. Theriot, E.C., Ashworth, M.P., Nakov, T., Ruck, E. & Jansen, R.K. Dissecting signal and noise in diatom chloroplast protein encoding genes with phylogenetic information profiling. *Mol Phylogenetic Evol* **89**, 28-36 (2015).
- 1590 114. Huesgen, P.F. *et al.* Proteomic amino-termini profiling reveals targeting information for protein import into complex plastids. *PLoS One* **8**, 74483 (2013).
- 1591 115. Grouneva, I., Rokka, A. & Aro, E.M. The thylakoid membrane proteome of two marine diatoms outlines both diatom-specific and species-specific features of the photosynthetic machinery. *J Proteom Res* **10**, 5338-5353 (2011).
- 1592 116. Read, B.A. *et al.* Pan genome of the phytoplankton *Emiliania* underpins its global distribution. *Nature* **499**, 209-213 (2013).

- 1621 117. Hovde, B.T. *et al.* Genome sequence and transcriptome analyses of  
1622 *Chrysochromulina tobin*: metabolic tools for enhanced algal fitness in the prominent  
1623 order Prymnesiales (Haptophyceae). *PLoS Genet* **11**, 1005469 (2015).
- 1624 118. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high  
1625 throughput. *Nucl Acids Res* **32**, 1792-1797 (2004).
- 1626 119. Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software  
1627 platform for the organization and analysis of sequence data. *Bioinformat* **28**, 1647-  
1628 1649 (2012).
- 1629 120. Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. trimAl: a tool for  
1630 automated alignment trimming in large-scale phylogenetic analyses. *Bioinformat* **25**,  
1631 1972-1973 (2009).
- 1632 121. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of  
1633 large phylogenies. *Bioinformat* **30**, 1312-1313 (2014).
- 1634 122. Gachon, C.M.M. *et al.* The CCAP KnowledgeBase: linking protistan and  
1635 cyanobacterial biological resources with taxonomic and molecular data. *Systemat*  
1636 *Biodiv* **11**, 407-413 (2013).
- 1637 123. Hehenberger, E., Burki, F., Kolisko, M. & Keeling, P.J. Functional relationship  
1638 between a dinoflagellate host and its diatom endosymbiont. *Mol Biol Evol* **33**, 2376-  
1639 2390 (2016).
- 1640 124. Gibson, D.G. *et al.* Enzymatic assembly of DNA molecules up to several hundred  
1641 kilobases. *Nat Methods* **6**, 343-345 (2009).
- 1642 125. Feliciello, I. & Chinali, G. A modified alkaline lysis method for the preparation of  
1643 highly purified plasmid DNA from *Escherichia coli*. *Anal Biochem* **212**, 394-401  
1644 (1993).
- 1645 126. Falciatore, A., Casotti, R., Leblanc, C., Abrescia, C. & Bowler, C. Transformation of  
1646 nonselectable reporter genes in marine diatoms. *Mar Biotechnol (NY)* **1**, 239-251  
1647 (1999).
- 1648 127. Simon, M., Lopez-Garcia, P., Moreira, D. & Jardillier, L. New haptophyte lineages and  
1649 multiple independent colonizations of freshwater ecosystems. *Env Microbiol Rep* **5**,  
1650 322-332 (2013).
- 1651 128. Leigh, J.W., Susko, E., Baumgartner, M. & Roger, A.J. Testing congruence in  
1652 phylogenomic analysis. *Systemat Biol* **57**, 104-115 (2008).
- 1653 129. Leliaert, F., Verbruggen, H. & Zechman, F.W. Into the deep: new discoveries at the  
1654 base of the green plant phylogeny. *Bioessays* **33**, 683-692 (2011).
- 1655 130. Allen, J.F., de Paula, W.B.M., Puthiyaveetil, S. & Nield, J. A structural phylogenetic  
1656 map for chloroplast photosynthesis. *Trends Plant Sci* **16**, 645-655 (2011).
- 1657 131. Kroth, P.G. *et al.* A model for carbohydrate metabolism in the diatom  
1658 *Phaeodactylum tricornutum* deduced from comparative whole genome analysis.  
*PLoS One* **3**, 1426 (2008).
- 1660 132. Bromke, M.A. Amino acid biosynthesis pathways in diatoms. *Metabolites* **3**, 294-311  
1661 (2013).
- 1662 133. Bertrand, M. Carotenoid biosynthesis in diatoms. *Photosynthesis Res* **106**, 89-102  
1663 (2010).
- 1664 134. Miret, J.A. & Munne-Bosch, S. Plant amino acid-derived vitamins: biosynthesis and  
1665 function. *Amino Acids* **46**, 809-824 (2014).
- 1666 135. Bandyopadhyay, S., Chandramouli, K. & Johnson, M.K. Iron-sulfur cluster  
1667 biosynthesis. *Biochem Soc Trans* **36**, 1112-1119 (2008).
- 1668 136. Shtaida, N., Khozin-Goldberg, I. & Boussiba, S. The role of pyruvate hub enzymes in  
1669 supplying carbon precursors for fatty acid synthesis in photosynthetic microalgae.  
*Photosynthesis Res* **125**, 407-422 (2015).

- 1671 137. Fukasawa, Y. *et al.* MitoFates: improved prediction of mitochondrial targeting  
 1672 sequences and their cleavage sites. *Mol Cell Proteom* **14**, 1113-1126 (2015).
- 1673 138. Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. Locating proteins in the cell  
 1674 using TargetP, SignalP and related tools. *Nat Protocol* **2**, 953-971 (2007).
- 1675 139. Claros, M.G. Mitroprot, a Macintosh application for studying mitochondrial proteins.  
 1676 *Comp App Biosci* **11**, 441-447 (1995).
- 1677 140. Miller, M.A. *et al.* A RESTful API for access to phylogenetic tools via the CIPRES  
 1678 science gateway. *Evol Bioinform Online* **11**, 43-48 (2015).
- 1679 141. Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model  
 1680 choice across a large model space. *Syst Biol* **61**, 539-542 (2012).
- 1681 142. Shimodaira, H. & Hasegawa, M. CONSEL: for assessing the confidence of  
 1682 phylogenetic tree selection. *Bioinform* **17**, 1246-1247 (2001).
- 1683 143. Cummins, C.A. & McInerney, J.O. A method for inferring the rate of evolution of  
 1684 homologous characters that can potentially improve phylogenetic inference, resolve  
 1685 deep divergence and correct systematic biases. *Syst Biol* **60**, 833-844 (2011).
- 1686 144. Yoon, H.S., Muller, K.M., Sheath, R.G., Ott, F.D. & Bhattacharya, D. Defining the  
 1687 major lineages of red algae (Rhodophyta). *J Phycol* **42**, 482-492 (2006).
- 1688 145. Dorrell, R.G., Gile, G.H., McCallum, G., Brillet-Guegen, L., Klinger, C.M., Meheust, R.,  
 1689 Peterson, K., Richter, D., Bowler, C., Bapteste, E.P. Research data supporting "The  
 1690 ancestral ochrophyte plastid proteome",  
 1691 <https://www.repository.cam.ac.uk/handle/1810/261421>
- 1692

**Table 1- Glossary Box**

1693 A schematic figure of eukaryotic taxonomy, showing the evolutionary origins of nuclear and  
 1694 plastid lineages, adapted from previous reviews<sup>3</sup>, is shown in Fig. 1- figure supplement 1.

1695

<b>Complex plastids</b>	Plastids acquired through the endosymbiosis of a eukaryotic alga. These include secondary plastids of ultimate red algal origin (such as those found in ochrophytes, haptophytes and cryptomonads), secondary plastids derived from green algae (such as those found in euglenids or chlorarachniophytes), or tertiary plastids such as those found in dinotoms and certain other dinoflagellates (resulting from the endosymbioses of eukaryotic algae that themselves contain plastids of secondary endosymbiotic origin).
<b>CASH lineages</b>	The four major lineages of algae with plastids of secondary or higher red origin, that is to say <u>Cryptomonads</u> , <u>Alveolates</u> (dinoflagellates, and apicomplexans), <u>Stramenopiles</u> , and <u>Haptophytes</u> .
<b>Stramenopiles</b>	A diverse and ecologically major component of the eukaryotic tree, containing both photosynthetic members (the ochrophytes), which possess complex plastids of red algal origin, and aplastidic and non-photosynthetic members (e.g. oomycetes, labyrinthulomycetes, and the human pathogen <i>Blastocystis</i> ), which form the earliest-diverging branches. It is debated when within stramenopile evolution the extant ochrophyte plastid was acquired.
<b>Ochrophytes</b>	Photosynthetic and plastid-bearing members of the stramenopiles, including many ecologically important lineages (diatoms, kelps, pelagophytes) and potential model lineages for biofuels research ( <i>Nannochloropsis</i> ). Ochrophytes form the most significant component of eukaryotic marine phytoplankton <sup>1,2</sup> .
<b>Haptophytes</b>	Single-celled, photosynthetic eukaryotes, possessing complex plastids of

	ultimate red origin. Some haptophytes (the coccolithophorids) are renowned for their ability to form large blooms (visible from space), and to form intricate calcareous shells <sup>1,94</sup> , which if deposited on the ocean floor go on to form a major component of limestone and other sedimentary rocks.
<b>HPPG</b>	"Homologous plastid protein group". Proteins identified in this study to possess plastid-targeting sequences that are homologous to one another, as defined by BLAST-based HPPG assembly and single gene phylogenetic analysis.

1698

1699

1700 **Figure Legends**

1701

1702 **Fig. 1. Procedure for identification of conserved plastid-targeted proteins in ochrophytes.**

1703 **Panel A** shows a schematic unrooted ochrophyte tree, with the three major ochrophyte  
 1704 lineages (chrysista, hypogyrstea, and diatoms) denoted by different coloured labels. "PX"  
 1705 refers to the combined clade of phaeophytes, xanthophytes and related taxa, and "PESC" to  
 1706 pinguiphyses, eustigmatophytes, synchromophytes, chrysophytes and relatives. A global  
 1707 overview of the eukaryotic tree of life, including the position of ochrophytes relative to  
 1708 other lineages is shown in figure supplement 1. **Panel B** shows the number of inferred  
 1709 positive control HPPGs (i.e., HPPGs encoding proteins with experimentally confirmed plastid  
 1710 localisation, or unambiguously plastid function) and negative control HPPGs (i.e., HPPGs  
 1711 encoding proteins with no obvious orthologues in ochrophyte genomes, but found in  
 1712 haptophyte and cryptomonad genomes) detected as plastid-targeted in different numbers  
 1713 of ochrophyte lineages using ASAFind (**i**) and HECTAR (**ii**). The blue bars show the number of  
 1714 positive controls identified to pass a specific conservation threshold, plotted against the left  
 1715 hand vertical axis of the graph, while the red bars show the number of negative controls that  
 1716 pass the same conservation threshold, plotted against the right hand vertical axis of the  
 1717 graph. The number of different sub-categories included in each conservation threshold is  
 1718 shown in a heatmap below the two graphs, with the specific distribution for each bar in the  
 1719 graph shown in the aligned cells directly beneath it. Each shaded cell corresponds to an  
 1720 identified orthologue in one sub-category of a particular ochrophyte lineage: orange cells  
 1721 indicate presence of chrysistan sub-categories; light brown cells the presence of  
 1722 hypogyrstean sub-categories; and dark brown cells the presence of diatom sub-categories.  
 1723 In each graph, black arrows label the conservation thresholds inferred to give the strongest  
 1724 separation (as inferred by chi-squared P-value) between positive and negative control  
 1725 sequences. The table (**iii**) tabulates the three conservation patterns identified as appropriate  
 1726 for distinguishing probable ancestral HPPGs from false positives. **Panel C** shows the  
 1727 complete HPPG assembly, alignment and phylogenetic pathway used to identify conserved-  
 1728 targeted proteins. **Panel D** tabulates the number of HPPGs built using ASAFind and HECTAR  
 1729 predictions, and the number of non-redundant HPPGs identified in the final dataset. The  
 1730 final total represents the pooled total of non-redundant HPPGs identified with both ASAFind  
 1731 and HECTAR.

1732

1733 **Fig. 2. Verification of unusual ancestral plastid-targeted proteins.** **Panel A** lists the ten  
 1734 proteins selected for experimental characterisation and their most probable previous  
 1735 localisation prior to their establishment in the ochrophyte plastid, based on the first 50 nr  
 1736 BLAST hits. Exemplar alignments and single-gene tree topologies for some of these proteins  
 1737 are shown in figure supplements 1-4. **Panel B** shows the localisation of GFP constructs for  
 1738 copies of two proteins with an unambiguous plastid localisation (a pyrophosphate-  
 1739 dependent PFK, which localises to the pyrenoid, and a novel plastid protein, with

1740 cosmopolitan distribution across the plastid) and one protein with a periplastid localisation  
1741 (a predicted peroxisomal membrane protein) from the diatom *Phaeodactylum tricornutum*,  
1742 the diatom endosymbiont of the dinoflagellate *Glenodinium foliaceum* and the  
1743 eustigmatophyte *Nannochloropsis gaditana*, expressed in *P. tricornutum*. All scale bars = 10  
1744 µm. Expression constructs for seven additional *P. tricornutum* proteins and three additional  
1745 *N. gaditana* proteins with multipartite plastid localisations are shown in figure supplements  
1746 5 and 6, and control images (wild-type cells, and cells expressing untargeted eGFP) are  
1747 shown in figure supplement 7.

1748  
1749 **Fig. 3. Evolutionary origins of the ochrophyte plastid proteome.** Panel A displays the origins  
1750 inferred by BLAST top hit, phylogenetic analysis, and combined analysis for all ancestral  
1751 HPPGs. Panel B shows (i) a schematic diagram of stramenopile taxonomy, with the  
1752 evolutionary relationships between labyrinthulomycetes, oomycetes, slopalinids and  
1753 ochrophytes proposed by recent multigene studies<sup>12</sup>, and the probable closest stramenopile  
1754 relative (as inferred by BLAST top hit analysis) of the 26 ancestral HPPGs verified by  
1755 combined analysis to be of aplastidic stramenopile origin, and (ii) the next nearest relative,  
1756 as inferred through BLAST top hit, phylogenetic and combined analysis, of the 26 aplastidic  
1757 stramenopile HPPGs verified by combined analysis. The evolutionary categories in this graph  
1758 are shaded as per in panel A.  
1759

1760 **Fig. 4. Verification and origins of the green signal in ochrophyte plastids.** Panel A shows a  
1761 schematic tree of the 11 archaeoplastid sub-categories with which each green HPPG  
1762 alignment was enriched prior to phylogenetic analysis. The topology of the red and green  
1763 algae are shown according to previously published phylogenies<sup>129, 144</sup>. Green sub-categories  
1764 are in green text; red algal sub-categories in red text; and other sub-categories are in blue  
1765 text. Five ancestral positions within the green algal tree inspected in subsequent analyses  
1766 are labelled with coloured boxes. Panel B shows the number of HPPGs of verified red (red  
1767 bars) or green origin (green bars) for which orthologues were identified in different numbers  
1768 green sub-categories (plotted on the x-axis) and red sub-categories (plotted on the z-axis).  
1769 An equivalent graph showing only HPPGs for which a glaucophyte orthologue was detected  
1770 is shown in figure supplement 1. Panel C compares the number of trees in which HPPGs of  
1771 verified green origin resolve as a sister group to all green lineages (including chlorophytes  
1772 and streptophytes); to multiple chlorophyte sub-categories but to the exclusion of  
1773 streptophytes; and to individual chlorophyte sub-categories only. A detailed heatmap of the  
1774 evolutionary distribution of the green sub-categories detected in each sister-group is shown  
1775 in figure supplement 2, and the distribution of BLAST top hits within each sub-category is  
1776 shown in figure supplement 3. Panel D lists the number of residues inferred from a dataset  
1777 of 32 ochrophyte HPPGs of verified green origin, which have been subsequently entirely  
1778 vertically inherited in all major photosynthetic eukaryotic lineages, to be uniquely shared  
1779 between ochrophytes and some but not all green lineages, hence might represent specific  
1780 synapomorphic residues. Residues are categorized by inferred origin point within the tree  
1781 topology shown in panel A, i.e., each of the five ancestral nodes labelled. A final category  
1782 shows all of the residues inferred to be specifically shared with one green sub-category, and  
1783 not with any other. The distribution of residues based on the earliest possible origin point  
1784 (taking into account gapped and missing residues in each HPPG alignment) is shown in figure  
1785 supplement 4. Panel E shows the number of the 7140 conserved gene families inferred to  
1786 have been present in the last common ochrophyte ancestor that are predicted by ASAFind  
1787 to encode proteins targeted to the plastid, subdivided by probable evolutionary origin, and  
1788 the number expected to be present in each category assuming a random distribution of  
1789 plastid-targeted proteins across the entire dataset, independent of evolutionary origin.  
1790 Evolutionary categories of proteins found to be significantly more likely (chi-squared test,

1791 P=0.05) to encode plastid-targeted proteins than would be expected are labelled with black  
1792 arrows. An equivalent distribution of plastid-targeted proteins inferred using HECTAR is  
1793 shown in figure supplement 5.

1794

1795 **Fig. 5. Functional mixing of the ancestral ochrophyte HPPGs. Panel A** tabulates nineteen  
1796 different fundamental plastid metabolism pathways and biological processes recovered in  
1797 the ancestral HPPG dataset. Detailed information concerning the origin and identity of each  
1798 component of each pathway is provided in figure supplement 1, and an overview and  
1799 phylogenetic trees of each of the non-vertically inherited enzymes identified are provided in  
1800 figure supplements 2-6. **Panel B** compares the distribution of individual KOG families in the  
1801 complete HPPG library, the ancestral HPPG dataset, and HPPGs of verified prokaryotic origin.  
1802 KOG families pertaining to metabolism are shown in shades of green, families pertaining to  
1803 information storage are shown in shades of red, and families pertaining to cellular processes  
1804 are shown in shades of blue. Families with unknown KOG classification or general function  
1805 predictions only are not shown. KOG classes that are enriched in the ancestral HPPG dataset  
1806 compared to relative proportions of each KOG class found in the full HPPG dataset, or in  
1807 individual ancestral HPPGs of prokaryotic origin compared to the ancestral HPPG dataset (as  
1808 inferred by chi-squared test,  $P < 0.05$ ), are labelled with black horizontal arrows. No such  
1809 enrichments were observed in any evolutionary category of ancestral HPPGs other than  
1810 prokaryotes, hence analogous distributions of HPPGs of red algal, green algal and host origin  
1811 are not shown. Overviews of the broader KOG classes that are enriched either in the  
1812 ancestral HPPG dataset, or in specific evolutionary categories of ancestral HPPG, are shown  
1813 in figure supplement 7. **Panel C** tabulates the number of ancestral HPPGs performing  
1814 consecutive metabolic functions, or that are likely to have direct regulatory interactions,  
1815 alongside the number of these protein pairs in which both members are of verified  
1816 evolutionary origin; the number observed where both members possess the same  
1817 evolutionary origin; the expected number of protein pairs where both members possess the  
1818 same evolutionary origin; and the chi-squared probability of similarity between the observed  
1819 and expected values. **Panel D** shows heatmaps for the pairwise correlation coefficients of  
1820 expression for genes encoding different evolutionary categories, as verified using combined  
1821 BLAST top hit and single-gene tree analysis, of ancestral HPPGs in the model diatoms  
1822 *Phaeodactylum tricornutum* (i) and *Thalassiosira pseudonana* (ii). A scale bar showing the  
1823 relationship between shading and correlation coefficient is shown to the right of the  
1824 heatmaps. Boxplots comparing the individual expression profiles of different categories of  
1825 ancestral HPPG, and the associated ANOVA P values calculated, are shown in figure  
1826 supplements 8 (for *P. tricornutum*) and 9 (for *T. pseudonana*).  
1827

1828 **Fig. 6. Origins of chimeric proteins in the ochrophyte plastid. Panel A** tabulates eight  
1829 ancestral HPPGs containing domains of cyanobacterial and non-cyanobacterial origin, as  
1830 previously identified by Méheust et al<sup>75</sup> that were inherited by the ochrophyte plastid, and  
1831 two chimeric ancestral HPPGs which are probably of specific ochrophyte origin. **Panel B**  
1832 shows a complete tabulated list of all ancestral HPPGs (listed by identifier, with the  
1833 predicted function given in brackets) in which at least one chimerism event between  
1834 domains of red algal, green algal, aplastidic stramenopile, other eukaryotic, and prokaryotic  
1835 origin was detected. In each case, the inferred evolutionary origins of the N-terminal (NTD)  
1836 and C-terminal (CTD) components of the chimeric members of the HPPG are given,  
1837 according to the colour key within the figure, followed by its distribution across all  
1838 ochrophyte lineages. The two chimeric HPPGs inferred to have arisen in the ochrophyte  
1839 ancestor are shown in bold text and labelled with horizontal arrows. Exemplar alignments  
1840 and phylogenies of the two chimeric proteins inferred to have originated in the ochrophyte  
1841 ancestor are shown in figure supplements 1-3.

1842  
1843 **Fig. 7. Ancient and bidirectional connections between the ochrophyte plastid and**  
1844 **mitochondria.** Panel A shows Mitotracker-Orange stained *P. tricornutum* lines expressing  
1845 GFP fusion constructs for the N-terminal regions of histidyl- and prolyl-tRNA synthetase  
1846 sequences from *P. tricornutum* and the eustigmatophyte *Nannochloropsis gaditana*.  
1847 Targeting constructs for an additional four dual targeted proteins in *P. tricornutum* and one  
1848 dual targeted protein in *G. foliaceum*, alongside Mitotracker-negative and wild type control  
1849 images, are shown in figure supplement 1. Panel B profiles the predicted evolutionary  
1850 origins of the 34 ancestral dual targeted HPPGs, as inferred by BLAST top hit and single-gene  
1851 tree analysis. Data supporting the thresholds used to identify probable dual targeted HPPGs  
1852 *in silico* are supplied in figure supplement 2. Panel C shows seven classes of tRNA synthetase  
1853 for which only two copies were inferred in the genome of the last common ochrophyte  
1854 ancestor. Evolutionary origins are inferred from combined BLAST top hit and single-gene  
1855 tree analysis for dual targeted proteins, and from BLAST top hit analysis alone for  
1856 cytoplasmic proteins. In five cases the dual targeted isoform is inferred to be of ultimate red  
1857 algal origin, indicating that a protein derived from the endosymbiont has functionally  
1858 replaced the endogenous host mitochondria-targeted copy.  
1859  
1860 **Fig. 8. Footprints of an ancient endosymbiosis in the haptophyte plastid proteome.** Panel  
1861 A indicates the number of ancestral ochrophyte HPPGs that included sequences from other  
1862 algal lineages in single-gene tree analyses, and whether those algal lineages branched within  
1863 or external to ochrophytes. An overview of the specific origins of proteins of ochrophyte  
1864 origin in each lineage is shown in figure supplement 1. Panel B compares the number of  
1865 ASAFind-derived HPPGs that are uniquely shared between hypogyristeal (i) or haptophytes  
1866 (ii) and one other CASH lineage. Values are given for proteins found in a majority of sub-  
1867 categories in hypogyristeal/ haptophytes and at least one sub-category from only one other  
1868 lineage (light bars), and proteins found in a majority of sub-categories in hypogyristeal/  
1869 haptophytes and a majority of sub-categories from only one other lineage (dark bars).  
1870 Values that are significantly greater than would be expected through random distribution  
1871 are labelled with black arrows. Panel C shows a schematic ochrophyte tree, with six different  
1872 ancestral nodes within this tree labelled with coloured boxes, and the most probable origin  
1873 point for each of the 243 haptophyte plastid-targeted proteins of probable ochrophyte  
1874 origin within this tree, as inferred by inspection of the nearest ochrophyte sister-group in  
1875 single-gene trees. A detailed heatmap of the ochrophyte sub-categories contained in each  
1876 lineage is shown in figure supplement 2, and BLAST top hit analyses corresponding to each  
1877 plastid-targeted protein are shown in figure supplement 3. Panel D shows the number of  
1878 residues that are uniquely shared between haptophytes and each node of the ochrophyte  
1879 tree for 37 genes in which there has been a clear transfer from ochrophytes to haptophytes,  
1880 and entirely vertical subsequent inheritance. A similar graph, showing the earliest possible  
1881 inferred origin of each uniquely shared residue, is shown in figure supplement 4. Panel E  
1882 shows the number of the 12728 conserved gene families inferred to have been present in  
1883 the last common haptophyte ancestor that are predicted by ASAFind to encode proteins  
1884 targeted to the plastid, subdivided by probable evolutionary origin, and the number  
1885 expected to be present in each category assuming a random distribution of plastid-targeted  
1886 proteins across the entire dataset, independent of evolutionary origin. Evolutionary  
1887 categories of proteins found to be significantly more likely (chi-squared test, P=0.05) to  
1888 encode plastid-targeted proteins than would be expected by random distribution are  
1889 labelled with black arrows. The evolutionary origins of the ancestral gene families are shown  
1890 in figure supplement 5.  
1891

1892 **Fig. 9. Non-ochrophyte origins of the haptophyte plastid genome. Panels A and B,**  
1893 respectively, show gene-rich and taxon-rich phylogenies of plastid-encoded proteins from  
1894 red algae and plastids of red algal origin with the glaucophyte *Cyanophora paradoxa* as  
1895 outgroup. **Panel A:** Combined Bayesian and Maximum Likelihood analysis (MrBayes +  
1896 RAxML, GTR, JTT, WAG) of a 22 taxa x 12103 aa alignment of 54 proteins encoded by all  
1897 published red and red-derived plastid genomes. **Panel B:** analysis of a 75 taxa x 3737 aa  
1898 alignment of 10 conserved plastid-encoded proteins detectable in a broad range of red  
1899 lineage MMETSP libraries. Nodes resolve with robust support (posterior probabilities of 1 for  
1900 all Bayesian trees and > 80% bootstrap support for all ML trees) are shown with filled circles;  
1901 individual support values for each analysis are shown for the remaining nodes are shown as  
1902 detailed in the box below panel B. Alternative topology tests, the results of fast-site and  
1903 clade deduction analysis for each tree, and heatmap comparisons of sister-group  
1904 relationships identified for single-gene trees of each constituent gene within each  
1905 concatenated alignment are shown in figure supplements 1-3. **Panel C** shows the number of  
1906 residues in each alignment that are uniquely shared between haptophytes and only one  
1907 other lineage. For the gene-rich alignment (i), which is gap-free, residues are included that  
1908 are found in all four haptophyte sequences and at least one sequence from the lineage  
1909 under consideration. For the taxon-rich alignment (ii), to account for the presence of gapped  
1910 positions, residues are included that are found in at least 11 of the 22 haptophyte sequences  
1911 and at least one sequence from the lineage under consideration.

1912  
1913 **Fig. 10. Schematic diagram of events giving rise to the ancestral ochrophyte plastid**  
1914 **proteome.** Each cell diagram depicts a different stage in the ochrophyte plastid  
1915 endosymbiosis; each protein depicted represents on or more proteins inferred in this study  
1916 to have been nucleus-encoded and plastid-targeted in the last common ancestor of all  
1917 ochrophytes. An ancient ochrophyte ancestor, which had already diverged from oomycetes  
1918 and other aplastidic stramenopile relatives, and which may have possessed a green algal  
1919 plastid (**A**), acquired a red lineage plastid via secondary or higher endosymbiosis (**B**). Both  
1920 the host and the endosymbiont are likely to have been evolutionary chimeras, possessing  
1921 proteins encoded by genes acquired from endosymbiotic and/or lateral gene transfer  
1922 events. Both host and symbiont are additionally likely to have possessed chimeric proteins,  
1923 generated through the fusion of genes of different evolutionary origins, and a large number  
1924 of mitochondrial-, ER- and (in the case of the red endosymbiont) potentially dual targeted  
1925 proteins. Following genetic integration of the red endosymbiont with its stramenopile host,  
1926 the first ochrophytes (**C**) thus possessed a wide range of proteins of plastid function  
1927 acquired from different sources, with no apparent functional bias in the types of proteins  
1928 that were retained from different sources. Chimeric proteins and dual targeted proteins,  
1929 either acquired directly from the endosymbiont, or generated de novo, were also  
1930 widespread features of this ancestral plastid proteome. Detailed information regarding the  
1931 relationship between ultimate the evolutionary origins of each HPPG, and its presence or  
1932 absence in other CASH lineages, is provided in figure supplement 1. A schematic diagram of  
1933 possible models through which the haptophyte plastid may have originated is shown in  
1934 figure supplement 2.

1935  
1936 **Supporting figure and dataset legends.**  
1937

1938 **Fig. 1- figure supplement 1. Overview of eukaryotic diversity.** This figure, adapted from a  
1939 previous review<sup>3</sup>, profiles the diversity of different eukaryotic nuclear lineages. Each grey  
1940 ellipse corresponds to one major clade, or “supergroup” of eukaryotes. A brown ellipse  
1941 within the stramenopile clade delineates the ochrophyte lineages. Dashed lines denote  
1942 uncertain taxonomic relationship. For each taxon, a type species (defined either by the

1943 presence of a complete genome, extensive transcriptome library, or of particular anthropic  
1944 significance) is given in brackets. Taxa that lack plastids are labelled in grey, and taxa with  
1945 plastids are shaded according to the evolutionary origin of that plastid lineage.

1946

1947 **Fig. 2- figure supplement 1- Exemplar ochrophyte plastid protein alignments.** This figure  
1948 shows untrimmed GeneIOUS alignments for two ancestral HPPGs of unusual provenance. In  
1949 each case the full length of the protein (labelled **i**) and N-terminal region only (**ii**) are shown,  
1950 demonstrating the broad conservation of the N-terminus position. Sequences for which  
1951 exemplar targeting constructs (*Phaeodactylum tricornutum*, *Nannochloropsis gaditana*,  
1952 *Glenodinium foliaceum*) are shown at the top of each alignment.

1953

1954 **Fig. 2- figure supplement 2. Tree of ochrophyte glycyl-tRNA synthetase sequences.** This  
1955 tree shows the consensus unrooted Bayesian topology for a 95 taxa x 487 aa alignment of  
1956 glycyl tRNA synthetase sequences. The font colour of each sequence corresponds to the  
1957 taxonomic origin (see legend below for details) and are labelled with the taxonomic  
1958 identifiers previously defined in Table S1. Sequences labelled with chl\_ possess apparent  
1959 plastid targeting sequences recognisable by CASH lineage plastids. The ancestral ochrophyte  
1960 plastidic isoform, of apparent chlamydiobacterial origin, is labelled with a blue ellipse. Black  
1961 circles at each node denote posterior probabilities of 1.0 in Bayesian inferences with three  
1962 different substitution matrices (GTR, Jones, and WAG), and grey circles indicate posterior  
1963 probabilities of 0.8 with at least two of these matrices. Support values for all remaining  
1964 nodes, using both Bayesian and RAxML analysis, is provided in the form MrBayes posterior  
1965 probabilities: GTR/Jones/WAG RAxML best tree likelihoods: GTR/ JTT/ WAG

1966

1967 **Fig. 2- figure supplement 3. Tree of ochrophyte pyrophosphate dependent phosphofructo-**  
1968 **1- kinase sequences.** This tree shows the consensus Bayesian topology inferred for a 94 taxa  
1969 x 449 aa alignment of pyrophosphate-dependent PFK, with taxa and support values shown  
1970 as per Fig. 2, figure supplement 2. The ancestral ochrophyte plastid isoform, of probable  
1971 aplastidic stramenopile origin, is labelled with a cyan ellipse.

1972

1973 **Fig. 2- figure supplement 4. Tree of a novel ochrophyte plastid-targeted protein.** This tree  
1974 shows the consensus Bayesian topology inferred for a 16 taxa x 103 aa alignment of a  
1975 plastid-targeted protein seemingly restricted to ochrophytes and one dinoflagellate lineage.  
1976 Taxa are labelled and support values are shown as per fig. 2- figure supplement 2.

1977

1978 **Fig. 2- figure supplement 5. Multipartite *Phaeodactylum* plastid-targeted proteins.** This  
1979 figure shows the localisation of GFP overexpression constructs for copies of seven proteins  
1980 from the diatom *Phaeodactylum tricornutum* that are of non-plastid origin, but show  
1981 multipartite localization to the plastid and one other organelle (the mitochondria, or in the  
1982 case of the “ER heat shock protein” to the endoplasmic reticulum).

1983

1984 **Fig. 2- figure supplement 6. Heterologous expression constructs of multipartite plastid-**  
1985 **targeted proteins.** This figure shows the localisation of GFP overexpression constructs for  
1986 copies of two proteins from the dinotom *Glenodinium foliaceum* (**Panel A**), and three  
1987 proteins from the eustigmatophyte *Nannochloropsis gaditana* (**Panel B**) that are of non-  
1988 plastid origin, but show multipartite localisation to the plastid and one other organelle, per  
1989 Fig. 2, figure supplement 5.

1990

1991 **Fig. 2- figure supplement 7. Exemplar control images for confocal microscopy.** This figure  
1992 shows fluorescence patterns for wild-type *Phaeodactylum tricornutum* cells (**i**), and  
1993 transformant *Phaeodactylum* cells expressing GFP that has not been fused to any N-terminal

1994 targeting sequence (ii), both visualised under the same conditions used for all other  
1995 transformant cultures.

1996

1997 **Fig. 4- figure supplement 1. Sampling richness associated with ancestral HPPGs of green**  
1998 **algal origin.** This figure shows the number of sub-different archaeoplastid orthologues for  
1999 ancestral HPPGs verified by combined BLAST top hit and single-gene tree analysis to be of  
2000 either green algal origin (green bars) or red algal origin (red bars), for which glaucophyte  
2001 orthologues could also be identified.

2002

2003 **Fig.4- figure supplement 2. Heatmaps of nearest sister-groups of ancestral HPPGs of**  
2004 **verified green origin.** This figure shows the specific topologies of single gene trees for HPPGs  
2005 verified to be of green origin by combined BLAST and phylogenetic analysis. **Panel A** shows a  
2006 reference topology of evolutionary relationships between green lineages, defined as per  
2007 Leliaert et al. 2011. Six ancestral nodes that might correspond to the origin point of  
2008 ochrophyte HPPGs are labelled with coloured boxes. **Panel B** shows the presence and  
2009 absence of each green subcategory in the immediate sister-group to the ochrophyte HPPG in  
2010 each single tree of HPPGs of verified origin. HPPGs are grouped by the inferred origin point  
2011 within the green algae, with the number of HPPGs identified for each origin point given with  
2012 round brackets.

2013

2014 **Fig. 4- figure supplement 3. Specific origins of green HPPGs as inferred from BLAST top hit**  
2015 **analyses.** These charts show (i) the number of BLAST top hits against each of the individual  
2016 green sub-categories from HPPGs for which a green origin was identified both from BLAST  
2017 top hit and single-gene tree analysis, and (ii) the total number of non-redundant sequences  
2018 from each green sub-category included in the BLAST library.

2019

2020 **Fig. 4- figure supplement 4. Earliest evolutionary origins of shared plastid residues.** This  
2021 figure shows the number of residues in the concatenated alignment of HPPGs of verified  
2022 green origin, which have been subsequently vertically inherited in all major photosynthetic  
2023 eukaryotes that are present in green algae and ochrophytes, and are not found in red algae  
2024 and glaucophytes. Residues are divided by inferred origin point, and are shown as per fig. 4,  
2025 panel D. The values here are calculated as the earliest possible origin point for each uniquely  
2026 shared residue, in which all gapped and missing positions within the alignment are treated  
2027 as potential identities. 100 of the 147 residues inferred to have originated within green algae  
2028 in this analysis originated either within a common ancestor of all chlorophytes, or in a  
2029 common ancestor of all chlorophytes excluding the basally divergent lineages *Prasinoderma*,  
2030 *Prasinococcus* and *Nephroselmis*.

2031

2032 **Fig. 4- figure supplement 5. Origins and HECTAR based targeting tests of proteins encoded**  
2033 **by conserved ochrophyte gene clusters.** **Panel A** shows the most probable evolutionary  
2034 origin, identified using BLAST top hit analysis, for 7140 conserved gene clusters inferred to  
2035 have been present in the last common ochrophyte ancestor. **Panel B** shows the number of  
2036 these gene families that are predicted by HECTAR to encode proteins targeted to the plastid,  
2037 subdivided by probable evolutionary origin, and the number expected to be present in each  
2038 category assuming a random distribution of plastid-targeted proteins across the entire  
2039 dataset, independent of evolutionary origin. Categories inferred to be significantly enriched  
2040 above the expected values are labelled with black arrows.

2041

2042 **Fig. 5- figure supplement 1. Reconstructed metabolism pathways and core biological**  
2043 **processes in the ancestral ochrophyte plastid.** This figure tabulates each of the ancestral  
2044 ochrophyte HPPGs corresponding to 350 central plastid metabolism and other biological

processes. The "origin" column shows the probable evolutionary source for each HPPG as defined by combined BLAST tophit and single-gene tree analysis. The origin of each ancestral HPPG is either assigned a "high confidence" value (in which the same origin was robustly supported both by single-gene tree and by BLAST tophit analysis) or a "low confidence" value (in the absence of robust and consistent support through both techniques; corresponding to the tree sister-group if one could be clearly assigned, or the BLAST tophit identity if not). A dash indicates the corresponding protein was not identified in the ancestral HPPG dataset due to either being plastid-encoded or alternative reasons; detailed explanations for the enzymes that are neither plastid-encoded nor detected in the ancestral HPPG dataset are provided in figure supplement 2.

**Fig. 5- figure supplement 2. Core plastid metabolism proteins not identified within the ancestral HPPG dataset.**

**Fig. 5 - figure supplement 3. Tree of ochrophyte sedoheptulose- 7-bisphosphatase sequences.** This figure shows the consensus Bayesian topology inferred for a 218 taxa x 303 aa alignment of sedoheptulose-7-bisphosphatase sequences, shown as per fig. 2, figure supplement 2. Two different ochrophyte plastid isoforms- one restricted to chrysista, and of probable red algal origin, and one found in hypogyristeae and diatoms, of probable green algal origin- are shown respectively by red and green ellipses.

**Fig. 5- figure supplement 4. Tree of ochrophyte 3-dehydroquinate synthase sequences.**

This figure shows the consensus Bayesian topology inferred for a 324 taxa x 387 aa alignment of 3-dehydroquinate synthase, shown as per fig. 2, figure supplement 2. Three ochrophyte plastid isoforms are shown with coloured ellipses: a probable bacterial isoform restricted to pelagophytes and dictyochophytes (blue ellipse), and two isoforms of ambiguous red/ green origin found respectively in raphidophytes and eustigmatophytes, and in diatoms (green ellipses with red borders).

**Fig. 5 - figure supplement 5. Tree of ochrophyte isopropylmalate dehydrogenase sequences.**

This tree shows the consensus Bayesian phylogeny inferred for a 202 taxa x 592 aa alignment of isopropyl malate dehydrogenase sequences, shown as per fig. 2- figure supplement 2. Two ochrophyte plastid isoforms are shown with coloured ellipses: an isoform of green algal origin restricted to diatoms and hypogyristeae (green ellipse), and a red algal isoform found in diatoms, pelagophytes and xanthophytes (red ellipse).

**Fig. 5- figure supplement 6. Tree of ochrophyte shikimate kinase sequences.** This figure

shows the consensus Bayesian topology inferred for a 127 taxa x 262 aa alignment of shikimate kinase sequences. The WAG Bayesian topology was excluded from the consensus due to non-convergence between the two chains, hence the tree is produced from the consensus of GTR and Jones substitution matrices only, but is otherwise presented identically to fig. 2, figure supplement 2. Two distinct ochrophyte plastid isoforms are shown with coloured ellipses: a green algal isoform conserved across diatoms, dictyochophytes and raphidophytes (red ellipse), and a pelagophyte isoform of uncertain origin (grey ellipse).

**Fig. 5- figure supplement 7. KOG classes associated with different categories of HPPGs.**

These pie charts profile the distribution of different KOG classes across (i) all HPPGs except for those with general function predictions only, or without any clear KOG function, (ii) the same, but restricted to ancestral HPPGs and (iii) the same, for ancestral HPPGs of unambiguous red, green, prokaryotic and aplastidic stramenopile origin as identified by combined BLAST top hit and single-gene tree analysis. KOG classes that occur at elevated

frequency in the ancestral HPPG dataset compared to the complete HPPG dataset, and one KOG class enriched in the prokaryotic HPPG dataset compared to the ancestral HPPG dataset (chi-squared test, P< 0.05) are labelled with horizontal arrows.

**Fig. 5- figure supplement 8. Coregulation of genes incorporated into HPPGs of different origin in the model diatom *Phaeodactylum tricornutum*.** Panel A shows boxplots of the correlation coefficients between the expression profiles of genes encoding members of ancestral HPPGs of red algal origin (i), green algal origin (ii), prokaryotic origin (iii) or host origin (iv), compared to genes encoding members of other HPPGs. Each HPPG is separated by evolutionary origin on the x-axis of each graph: for example, the box labelled “green algae” on the “red algae” graph shows the correlation coefficients between genes encoding members of ancestral HPPGs of red origin, and ancestral HPPGs of green origin. Panel B shows the P value statistics of mean separation calculated when comparing genes encoding members of ancestral HPPGs of the same origin (shown by row) to members of ancestral HPPGs of different origin (shown by column). For example, the intersect between the “red” row and “green” column shows the difference in mean correlation coefficient between pairs of genes that both encode members of ancestral HPPGs of red origin, and gene pairs of which one encodes an ancestral HPPG member of red origin, and the other an ancestral HPPG member of green origin. None of the P values calculated are significant, i.e. there are no categories of ancestral HPPG in which the internal correlation coefficients of gene expression are any different to those observed across the dataset as a whole.

**Fig. 5- figure supplement 9. Coregulation of genes incorporated into HPPGs of different origin in the model diatom *Thalassiosira pseudonana*.** Boxplots (Panel A) and P value statistics (Panel B) are shown as per Fig. 5- figure supplement 8. Only two of the correlation value ANOVA tests (comparison of red-red and red-host correlations, and prokaryotic-prokaryotic and prokaryotic-host correlations, shaded in green) reveal a significantly higher correlation coefficient between pairs of genes encoding members of HPPG of the same evolutionary origin than pairs of genes encoding members of HPPGs with different evolutionary origins. These differences most probably reflect the extremely weak correlation coefficients associated with genes encoding HPPGs of host origin to all other genes considered (compare “Host” category on boxplots i, ii and iii to all other categories); however, detailed comparison of the correlation values between genes encoding ancestral HPPGs of host origin and genes encoding ancestral HPPGs of different evolutionary origin (Panel A, boxplot iv; Panel B, bottom row) reveals no specific difference in the pairwise correlation values observed between genes encoding ancestral HPPGs of host origin, and genes encoding ancestral HPPGs of all other origins within the dataset.

**Fig. 6- figure supplement 1. Alignments of an ochrophyte-specific riboflavin biosynthesis fusion protein.** Panel A shows alignments of the full length (i) and cyclohydrolase domain only (ii) of a plastid-targeted GTP cyclohydrolase II/ 3,4-dihydroxy-2-butanone 4-phosphate synthase protein conserved across the ochrophytes. Coloured bars adjacent to each sequence correspond to the phylogenetic identity of the sequence. The cyclohydrolase domain of the ochrophyte protein is positioned in the N-terminal region, and the synthase domain in the C-terminal region. Three uniquely shared residues at the N-terminus of the cyclohydrolase domain confirm that it has been inherited from the aplastidic stramenopile ancestor of the ochrophytes.

**Fig. 6- figure supplement 2. Origins of ochrophyte plastid 3,4-dihydroxy-2-butanone 4-phosphate synthase.** This figure shows the consensus Bayesian topology inferred for a 22 taxa x 206 aa alignment of 3,4-dihydroxy-2-butanone 4-phosphate synthase domains from

2147 different lineages, inferred using Jones and WAG matrices, and shown as per fig. 2, figure  
2148 supplement 2. The ochrophyte plastid isoforms branch with red algal and actinobacterial  
2149 sequences.

2150

2151 **Fig. 6- figure supplement 3. An ochrophyte-specific Tic20 fusion protein.** This figure shows  
2152 alignments of the full length (**i**) and conserved region only (**ii**) of plastid Tic20 sequences,  
2153 displayed as per figure supplement 1.

2154

2155 **Fig. 7- figure supplement 1. Experimental verification of additional ochrophyte dual-**  
2156 **targeted proteins. Panel A** shows Mitotracker-orange stained *Phaeodactylum tricornutum*  
2157 lines expressing four additional dual-targeted proteins (glycyl-, leucyl-, and methionyl-tRNA  
2158 synthetases, and a predicted mitochondrial GroES-type chaperone) from *Phaeodactylum*  
2159 *tricornutum*, and a dual-targeted histidyl-tRNA synthetase from *Glenodinium foliaceum*.  
2160 **Panel B** shows control images that confirm an absence of crosstalk between GFP and  
2161 Mitotracker: wild-type *Phaeodactylum* cells stained with Mitotracker, and cells expressing  
2162 the *Glenodinium* histidyl-tRNA synthetase–GFP fusion construct and visualised with the  
2163 Mitotracker laser and channel in the absence of Mitotracker stain.

2164

2165 **Fig. 7- figure supplement 2. Comparison of different in silico targeting prediction**  
2166 **programmes for the identification of dual-targeted ochrophyte proteins. Panel A** shows  
2167 Mitofates scores for ochrophyte proteins verified experimentally to be dual targeted in this  
2168 and a previous study<sup>9</sup>. **Panel B** shows Mitofates scores for all ochrophyte proteins for which  
2169 a subcellular localisation has been identified in previous studies. The red lines in each graph  
2170 show the Mitofates default cutoff (0.385) and the green lines indicate our chosen cutoff  
2171 (0.35). **Panel C** compares different in silico targeting prediction algorithms with respect to  
2172 predicted mitochondrial localization by experimentally validated localization. Mitofates  
2173 strikes the best balance between high true positives and low false positives.

2174

2175 **Fig. 8- figure supplement 1. Origin of proteins of ochrophyte origin in different CASH**  
2176 **lineages.** This figure profiles the evolutionary origins of proteins inferred by single-gene  
2177 phylogenetic analysis to have been transferred from the ochrophytes into other lineages  
2178 that have acquired plastids through secondary or more complex endosymbioses. Proteins  
2179 are divided into the three major ochrophyte lineages (i.e. diatoms, chrysista, and  
2180 hypogyrstea); all remaining proteins (inferred to have been acquired from an ancestor of  
2181 multiple ochrophyte lineages, or of ambiguous but clearly ochrophyte origin) are grouped as  
2182 a final category. The haptophyte proteins that could be attributed to a specific ochrophyte  
2183 lineage are particularly skewed (100/178 proteins) to origins within the hypogyrstea.

2184

2185 **Fig.8- figure supplement 2. Heatmaps of nearest sister-groups to haptophytes in ancestral**  
2186 **ochrophyte HPPG trees.** This figure shows the specific ochrophyte lineages implicated in the  
2187 origin of haptophyte plastid-targeted proteins, as inferred from the nearest ochrophyte  
2188 sister-groups to haptophytes in trees of 242 haptophyte proteins of probable ochrophyte  
2189 origin from combined BLAST top hit and single-gene tree analysis. At the top a schematic  
2190 tree diagram of the ochrophytes is shown as per fig. 1, with six major nodes in ochrophyte  
2191 evolution labelled with coloured boxes. The heatmap below shows the specific distribution  
2192 of sister-groups in each tree, shown as per figure 4- figure supplement 2.

2193

2194 **Fig. 8- figure supplement 3. Internal evolutionary affinities of haptophyte plastid-targeted**  
2195 **proteins incorporated into ancestral ochrophyte HPPGs.** This figure profiles the  
2196 evolutionary origins of haptophyte plastid-targeted proteins incorporated into ancestral  
2197 ochrophyte HPPGs by BLAST top hit analysis. Separate values are provided for query

2198 sequences from each of the three haptophyte sub-categories (pavlovophytes,  
2199 prymnesiophytes, and isochrysidales) considered within the analysis. Only sequences for  
2200 which a consistent origin could be identified by both BLAST top hit and single-gene tree  
2201 analysis are included. For each haptophyte lineage > 50% of the sequences verified by  
2202 combined analysis to be of a specific ochrophyte origin have either pelagophyte or  
2203 dictyochophyte top hits.

2204  
2205 **Fig. 8- figure supplement 4. Evidence for gene transfer from pelagophytes and**  
2206 **dictyochophytes into haptophytes. Panel A** shows the next deepest sister groups identified  
2207 for haptophyte proteins of hypogyristeal origin in single-gene trees. The pie chart (i)  
2208 compares the number of single-gene trees in which the combined clade of haptophyte and  
2209 hypogyristeal proteins resolves within a larger clade comprising the ochrophyte HPPG,  
2210 compared to the number that resolves in external positions, either with other lineages or as  
2211 a sister-group to all other sequences within the HPPG clade. Sequences for which no clear  
2212 next deepest sister group affinity could be identified are listed as “not determined”. The  
2213 heatmap (ii) shows the specific sister-group sequences associated with 65 HPPGs in which  
2214 the haptophyte sequences specifically resolve with the pelagophyte/ dictyochophyte clade  
2215 and for which a clear internal or external position for the haptophyte/ hypogyristeal group  
2216 relative to the remaining ochrophyte HPPG clade could be identified. Both analyses indicate  
2217 a clear bias for haptophyte sequences branching within a deeper ochrophyte clade, not just  
2218 restricted to the immediate sister-groups. **Panel B** tabulates the BLAST next best hits for  
2219 haptophyte sequences for which a phylogenetically consistent (>3 consecutive top hits) top  
2220 hit to hypogyrista could be identified, and pelagophyte/ dictyochophyte sequences for  
2221 which a phylogenetically consistent top hit to haptophytes could be identified. In each case  
2222 either the largest number of sequences, or (in the case of pavlovophytes) the joint largest  
2223 number of sequences for which a phylogenetically consistent next best hit could be  
2224 identified resolved with diatoms, indicating that these sequences were probably present in  
2225 the common ancestor of diatoms and hypogyrista, and subsequently transferred to the  
2226 haptophytes.  
2227

2228 **Fig. 8- figure supplement 5. Earliest possible origin points of uniquely conserved sites in**  
2229 **haptophyte plastid-targeted proteins. This** figure shows the total number of residues that  
2230 are uniquely shared between a 37 proteins that have clearly been transferred between the  
2231 ochrophytes and haptophytes, and are of subsequently entirely vertical origin, assuming the  
2232 earliest possible origin point for each residue (i.e. in which gapped or missing positions were  
2233 interpreted as identities). 87/ 128 of the uniquely shared residues inferred to originate  
2234 within the ochrophytes were congruent to gene transfers between the haptophytes and  
2235 pelagophyte and dictyochophyte clade; of these, slightly more than half (46) are inferred to  
2236 have originated in a common ancestor of all hypogyrista and diatoms, consistent with the  
2237 gene transfer having occurred from an ancestor of the pelagophytes and dictyochophytes  
2238 into the haptophytes, rather than the converse.  
2239

2240 **Fig. 8- figure supplement 6. Evolutionary origin of ancestral haptophyte genes.** This figure  
2241 shows the most likely evolutionary origin assigned by BLAST top hit analysis to the 12728  
2242 conserved gene families inferred to have been present in the last common haptophyte  
2243 ancestor.  
2244

2245 **Fig. 9- figure supplement 1. Alternative topology tests of plastid genome trees.** Tests were  
2246 performed with the RAxML + JTT trees inferred for the gene-rich (**panel A**) and taxon-rich  
2247 (**panel B**) plastid-encoded protein alignments. In each case, a schematic diagram of the tree  
2248 topology obtained is given (i). The black box corresponds to the branch position of

2249 haptophytes in the consensus tree; alternative branching positions for the haptophyte  
2250 sequences are labelled with numbered boxes. The table below (ii) lists the probabilities for  
2251 each alternative position under eight different tests performed with CONSEL. Alternative  
2252 positions that are not rejected by a topology test are shaded. All possible trees in which the  
2253 haptophyte sequences branch within the ochrophytes are clearly rejected under all  
2254 conditions, confirming that its plastid genome is of non-ochrophyte origin. The legend at the  
2255 bottom of panel B gives full names for each test performed.

2256

2257 **Fig. 9- figure supplement 2. Fast site removal and clade deduction analysis of plastid**  
2258 **genome trees. Panel A** shows the support values obtained for Bayesian + Jones trees  
2259 inferred from modified versions of the taxon-rich plastid multigene alignment from which  
2260 the 13 fastest evolving site categories had been removed for four different branching  
2261 relationships pertaining to the placements of haptophyte and hypogyristeian sequences. The  
2262 % of residues from the original alignment retained in each modified alignment are shown  
2263 with grey bars. **Panel B** tabulates the support obtained for two different evolutionary  
2264 relationships (haptophytes as a sister group to all cryptomonads, and as a sister group to all  
2265 ochrophytes) in gene-rich (i) and taxon-rich (ii) alignments modified to remove all amino  
2266 acids that occur at different frequencies in haptophytes to ochrophyte lineages, and  
2267 modified to remove individual or pairs of CASH lineages. "x" indicates that the topology in  
2268 question was not obtained.

2269

2270 **Fig. 9- figure supplement 3. Single-gene tree topologies associated with individual plastid-**  
2271 **encoded genes.** These heatmaps show the first sister-groups identified to haptophytes, and  
2272 members of the pelagophyte/ dictyochophyte clade, in single-gene trees of component  
2273 genes included in concatenated trees of plastid-encoded proteins using both the gene-rich  
2274 (i) and taxon-rich (ii) alignments. Topologies are given for trees inferred with MrBayes using  
2275 the Jones substitution matrix, and RAxML trees inferred using JTT, under the same  
2276 conditions as the multigene trees. The identity of the first sister-group is shaded according  
2277 to the legend given below. Only three single-gene trees (labelled with black arrows) support  
2278 any sister-group relationship between haptophytes and the pelagophyte/ dictyochophyte  
2279 clade; however, in each case (explained beneath the legend) this topology is not robustly  
2280 supported, either due to polyphyly of one of the constituent lineages, or conflicting  
2281 topologies identified via alternative methods.

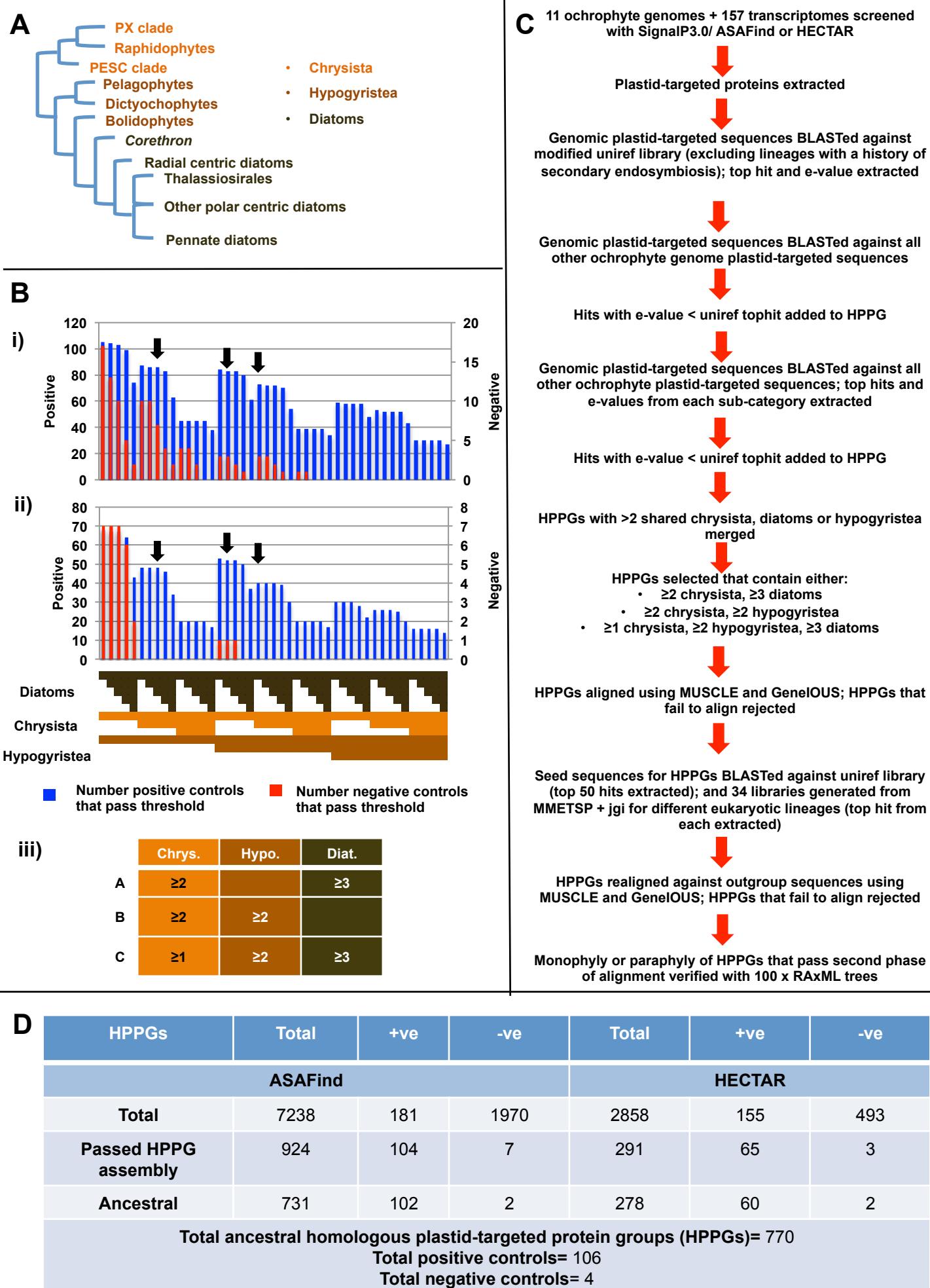
2282

2283 **Fig. 10- figure supplement 1. Complex origins of different ancestral ochrophyte HPPGs**  
2284 **Panel A** shows the evolutionary positions of lineages with histories of secondary  
2285 endosymbiosis in trees of ancestral ochrophyte HPPGs verified by combined BLAST top hit  
2286 and single-gene tree analysis to be either of red algal (i) or green algal origin (ii). In both  
2287 cases, in more than half of the constituent trees, haptophyte and cryptomonad sequences  
2288 resolve as closer relatives to the ochrophytes than the red or green algal evolutionary  
2289 outgroup, either due to resolving in the ochrophyte HPPG or forming a specific sister-group  
2290 to the ochrophyte lineages. **Panel B** plots the distribution of cryptomonads (i) and  
2291 haptophytes (ii) in trees for different categories of ancestral ochrophyte HPPG of verified  
2292 evolutionary origin. HPPGs of green algal origin more frequently show internal or sister  
2293 positions for the cryptomonad sequences than all other categories of HPPG, and in more  
2294 than 50% of cases resolve internal or sister positions for the haptophyte sequences. This  
2295 might be consistent with a green algal contribution in the endosymbiotic ancestor of  
2296 cryptomonad, haptophyte and ochrophyte plastids.

2297

2298 **Fig. 10 –figure supplement 2. Different scenarios for the origins of haptophyte plastids.**  
2299 This schematic tree diagram shows different possibilities for the origins of the haptophyte

2300 plastid as predicted from the data within this study. No inference is made here regarding the  
2301 ultimate origin of the ochrophyte plastid, although it is noted that the ochrophyte,  
2302 cryptomonad and haptophyte plastids are likely to be closely related to one another within  
2303 the red plastid lineages. First, a common ancestor of the pelagophytes and dictyochophytes  
2304 was taken up by a common ancestor of the haptophytes (point **1**), yielding a permanent  
2305 plastid that contributed genes for a large number of plastid-targeted proteins in extant  
2306 haptophytes. This plastid was subsequently replaced via serial endosymbiosis (point **2**)  
2307 yielding the current haptophyte plastid and plastid genome. This serial endosymbiosis event  
2308 either involved a close relative of extant cryptomonads (**2A**) or a currently unidentified  
2309 species that forms a sister-group in plastid gene trees to all extant ochrophytes, but is  
2310 evolutionarily distinct from the pelagophytes (**2B**). It is possible that the haptophyte plastid  
2311 may have been acquired through the secondary endosymbiosis of a different lineage of red  
2312 algae to the ochrophyte, either via a cryptomonad intermediate (**2C**) or directly (**2D**).  
2313  
2314



**A**

Protein	Probable origin
ER Heat Shock Protein	Host ER
Glycyl tRNA synthetase	Bacterial
Histidyl tRNA synthetase	Host cytoplasm
Methionyl tRNA synthetase	Bacterial
Leucyl tRNA synthetase	Host cytoplasm
Mitochondrial GroES chaperonin	Host mitochondria
Pyrophosphate-dependent phosphofructokinase	Symbiont cytoplasm
Peroxisomal membrane protein MPV17	Symbiont peroxisome
Prolyl tRNA-synthetase	Symbiont cytoplasm
Novel protein 1	Unknown

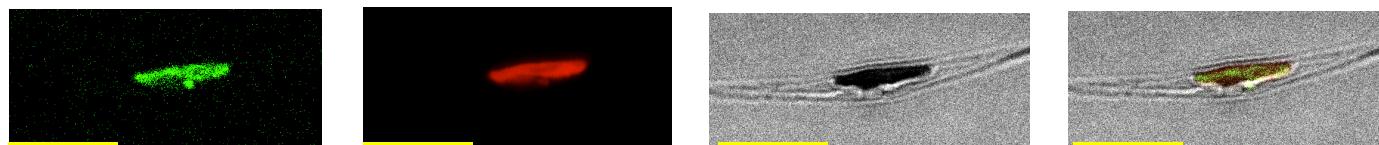
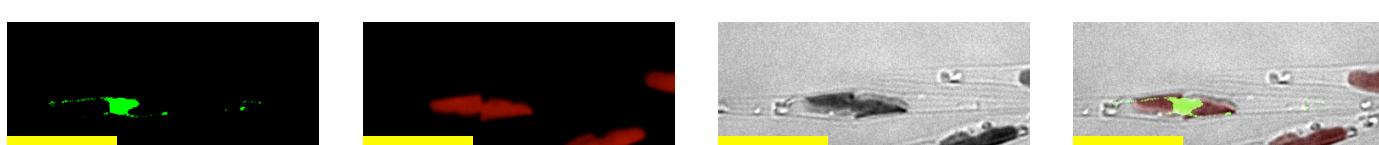
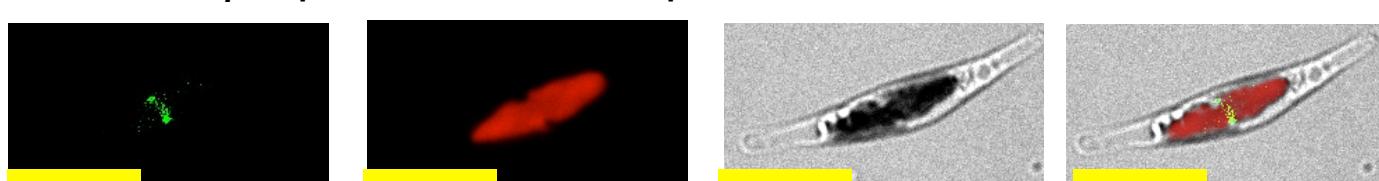
**B**

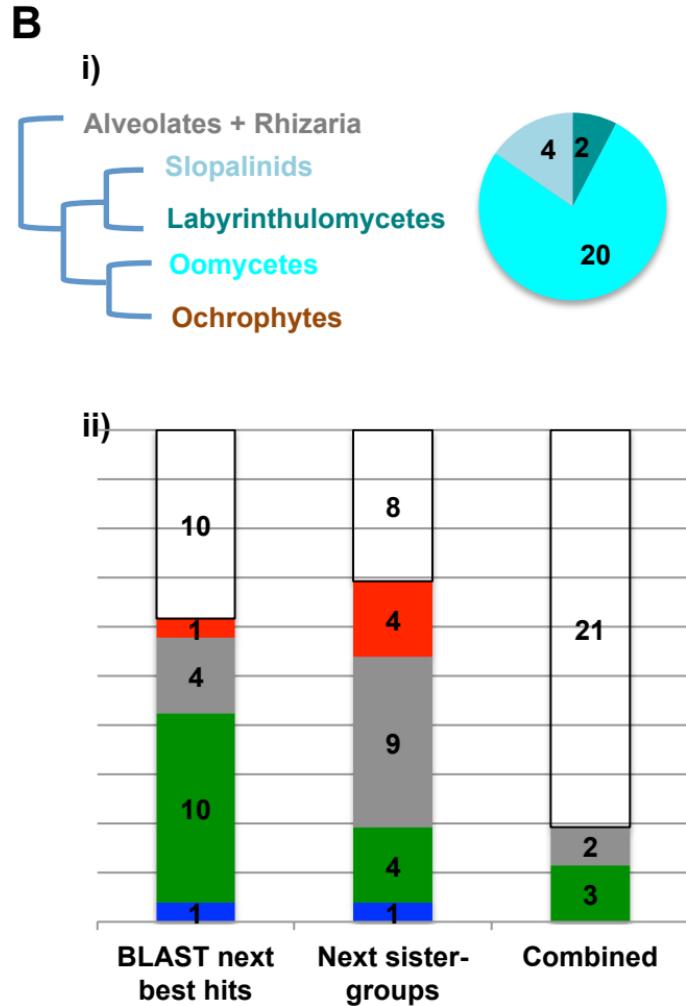
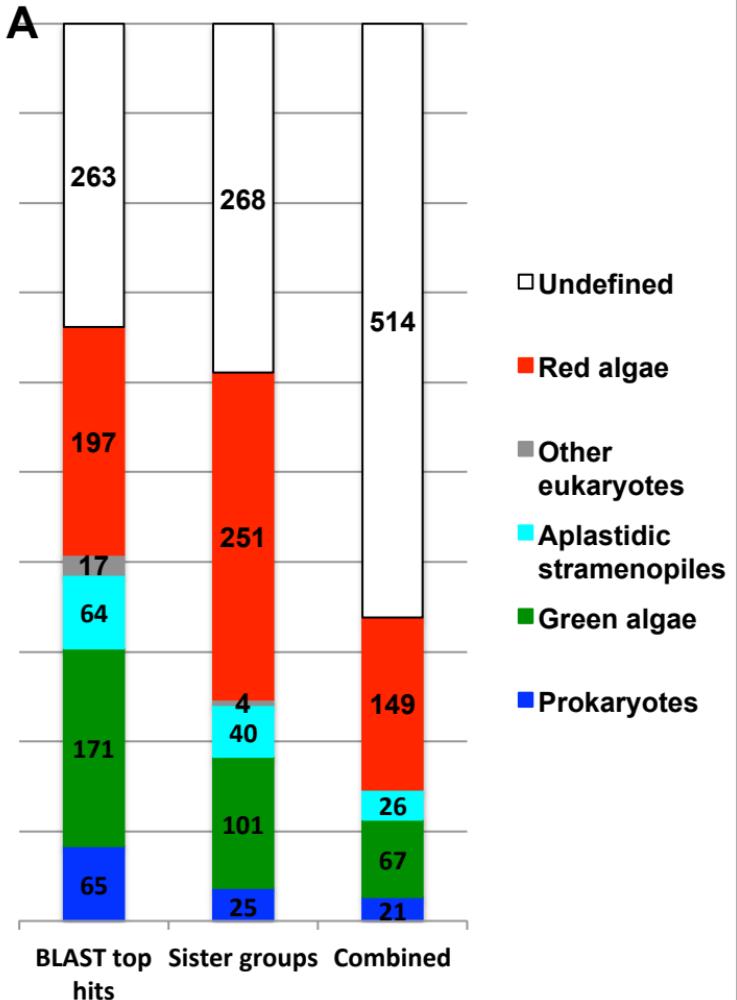
GFP

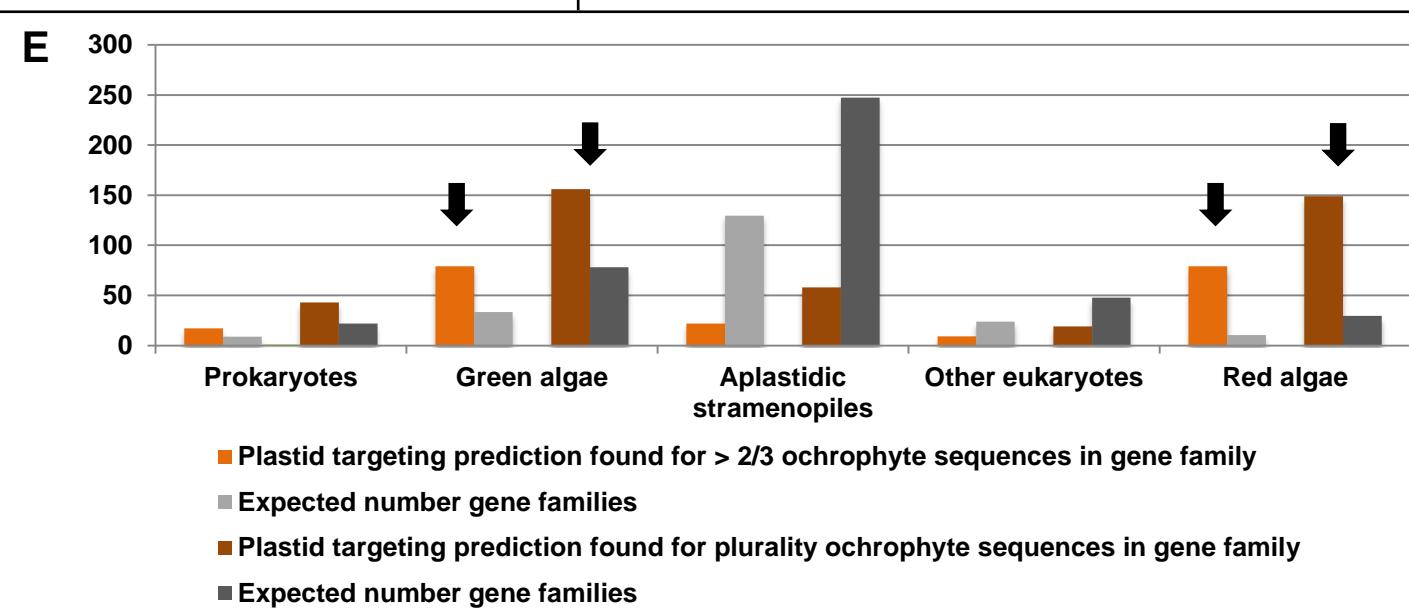
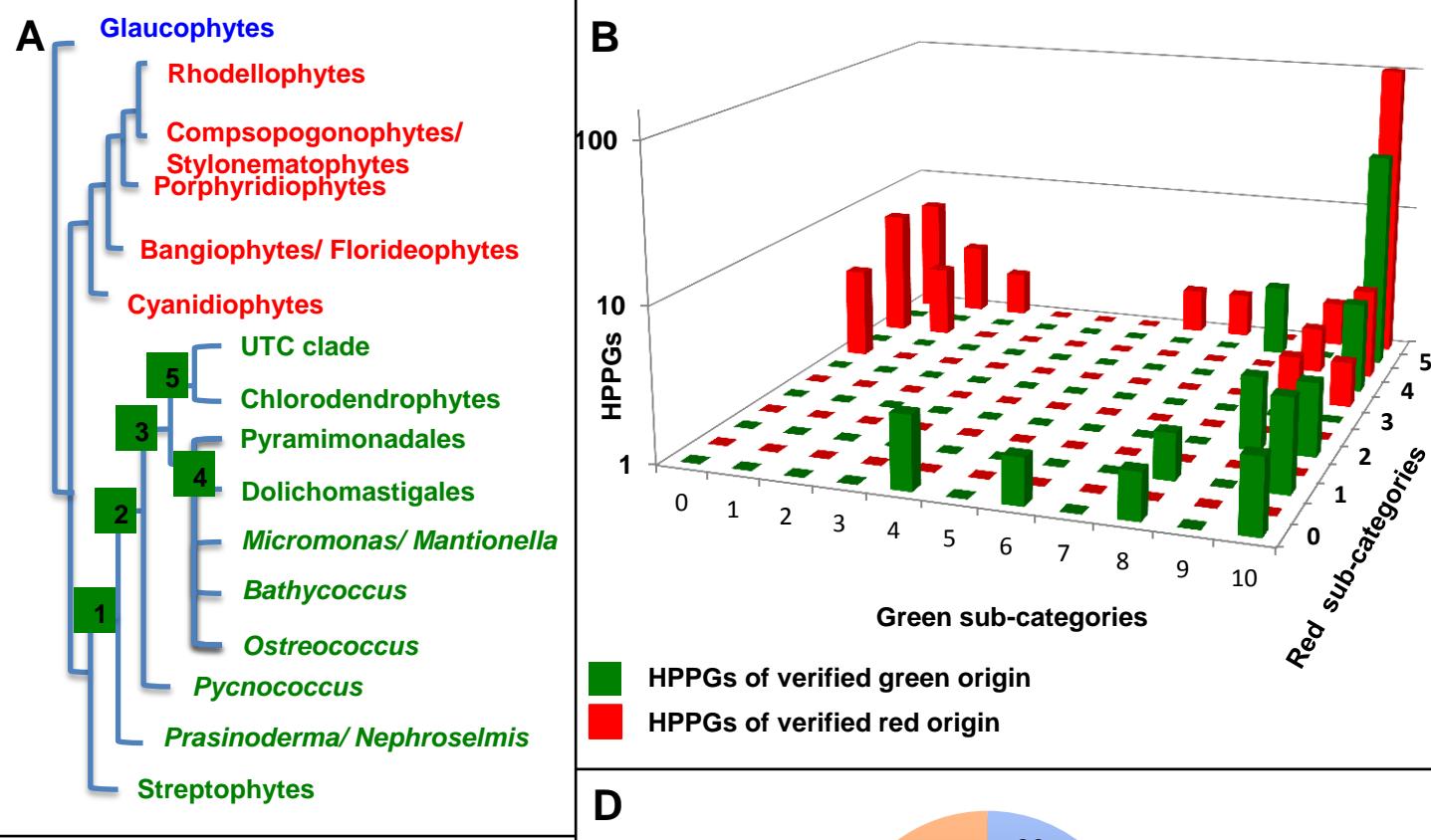
Chlorophyll

Bright-field

Merge

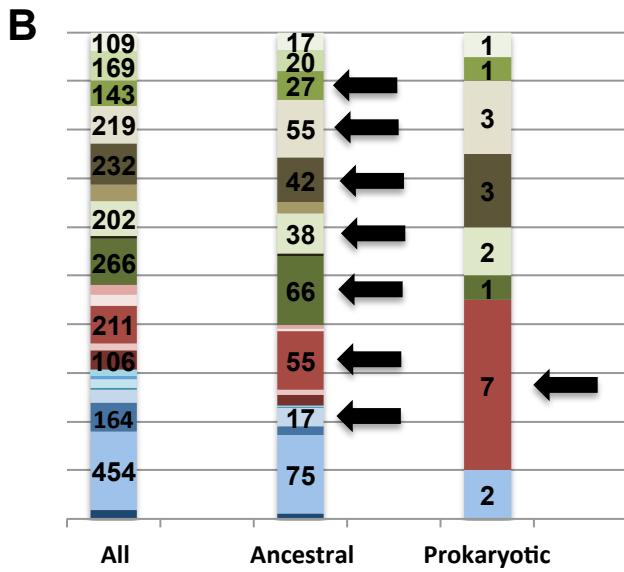
*Phaeodactylum* pyrophosphate-dependent PFK*Phaeodactylum* novel plastid protein*Phaeodactylum* peroxisomal membrane protein*Glenodinium* novel plastid protein*Glenodinium* peroxisomal membrane protein*Nannochloropsis* novel plastid protein*Nannochloropsis* peroxisomal membrane protein





**A**

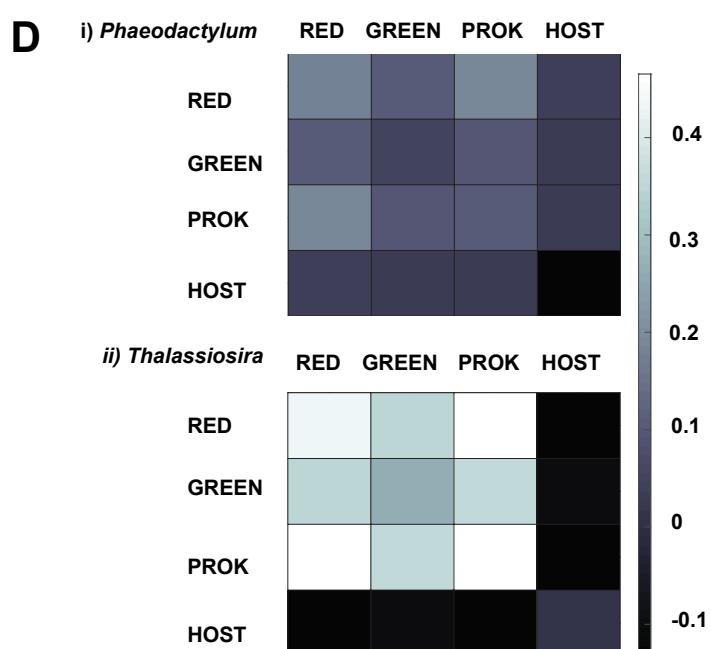
Biological process	Identified	Plastid-encoded	Disposable	Non-vertical
Light-harvesting proteins	14	-	-	-
Photosynthesis	28	45	-	-
Central carbon metabolism	27	2	-	1
Lipid synthesis	16	-	-	-
Tetrapyrrole synthesis	24	1	1	-
Carotenoid synthesis	18	-	1	-
Fe-S cluster synthesis	8	2	-	-
Riboflavin synthesis	2	-	-	-
Glu/Gln/Asp/Lys synthesis	16	-	-	-
Phe/Trp/Tyr synthesis	13	-	-	2
Ile/Leu/Val synthesis	6	1	-	1
Ser/Cys synthesis	8	-	1	-
tRNA synthesis	22	-	-	-
Nucleotide synthesis	4	-	-	-
Ribosomal proteins	8	45	1	-
Translation initiation	7	2	-	-
Protein import complexes	8	4	-	-
Division	2	0	-	-
Clp protease complex	8	1	-	-
Total	239	103	4	4



- Secondary metabolite biosynthesis
- Inorganic ion transport and metabolism
- Lipid transport and metabolism
- Coenzyme transport and metabolism
- Carbohydrate transport and metabolism
- Nucleotide transport and metabolism
- Amino acid transport and metabolism
- Cell cycle control
- Energy production and conversion
- Replication
- Transcription
- Translation
- Chromatin structure and dynamics
- RNA processing and modification
- Cytoskeleton
- Nuclear structure
- Extracellular structures
- Defense mechanisms
- Intracellular trafficking
- Signal transduction mechanisms
- Posttranslational modification

**C**

Number protein pairs	313
Number protein pairs between HPPGs of clear evolutionary origin	95
Observed number protein pairs between HPPGs of same origin	44
Expected number protein pairs between HPPGs of same origin	41.05
Chi-squared P	0.541



**A**

i) Chimeras inherited by the ochrophyte ancestor	Origin in ochrophytes	NTD	CTD
PpiC-type peptidyl-prolyl cis-trans isomerase	Ambiguous	Firmicutes/ Proteobacteria	Cyanobacteria
Hypothetical protein	Red	Firmicutes	Cyanobacteria
Rieske 2Fe-2S region	Green	Cyanobacteria	Proteobacteria
Probable heme-binding protein	Red	Cyanobacteria	Proteobacteria
Acyl-CoA:diacylglycerol acyltransferase (DGAT)	Host	Cyanobacteria	Actinobacteria
Phenylalanyl-tRNA synthetase	Red	Cyanobacteria	Proteobacteria

ii) Chimeras endogenous to ochrophytes

Calmodulin and related proteins/ Tic20

NTD

CTD

Unknown

Red algae

DHBP synthase/ GTP cyclohydrolase

Aplastidic stramenopiles

Actinobacteria/ Red algae

**B**

**HPPG**

xbq (14 kDa zinc-binding protein)  
xmw (3,8-divinyl protochlorophyllide a 8-vinyl reductase)  
xqu (Asparaginyl-tRNA ligase)  
2fn (Translation elongation factor EF-3b)  
2ia (Calmodulin and related proteins)  
2nl (Carboxy-terminal-processing peptidase)  
xin (Delta-aminolevulinic acid dehydratase)  
**2oe (DHBP synthase/ GTP cyclohydrolase)**  
**2ce (EF-hand protein/ Tic20)**  
2ik (Fibrillin family protein)  
xjq (FKBP-type peptidyl-prolyl cis-trans isomerase)  
xko (Formate/ nitrite transporter)  
2jw (fucoxanthin chlorophyll a/c protein)  
2qh (Galactosyltransferases)  
xbz (Glutathione reductase)  
xes (Glycine-rich protein 2)  
xou (Hypothetical protein)  
xsi (Hypothetical protein)  
2bu (Hypothetical protein)  
2ii (Hypothetical protein)  
2pi (Hypothetical protein)  
xrh (IMP-GMP specific 5'-nucleotidase)  
2db (Kynurenine 3-monoxygenase)  
xii (Mitochondrial chaperonin)  
2ju (Molecular chaperone (HSP90 family))  
2eo (N-6 Adenine-specific DNA methylase)  
xmq (NADH-dehydrogenase (ubiquinone))  
2ir (Peptide methionine sulfoxide reductase)  
2mr (Phenylalanyl-tRNA synthetase 1)  
2lx (Phenylalanyl-tRNA ligase 2)  
2pv (Phosphatidate cytidyltransferase)  
xpn (Phosphoglycerate mutase)  
2ga (Phytanoyl-CoA dicarboxylase)  
2da (Plastid lipid-associated protein)  
2ra (Predicted K+/H+-antiporter)  
4gv (Protein disulfide-isomerase)  
2fd (Psb29)  
2dx (Psb31)  
3ac (Puromycin-sensitive aminopeptidase)  
2gx (Putative aminopeptidase)  
3at (Ribosomal RNA adenine dimethylase)  
2fj (Rieske 2Fe-2S region)

NTD CTD

PX clade

Raphidophytes  
PESC clade

Bolidophytes

Dictyochophytes

Pelagophytes

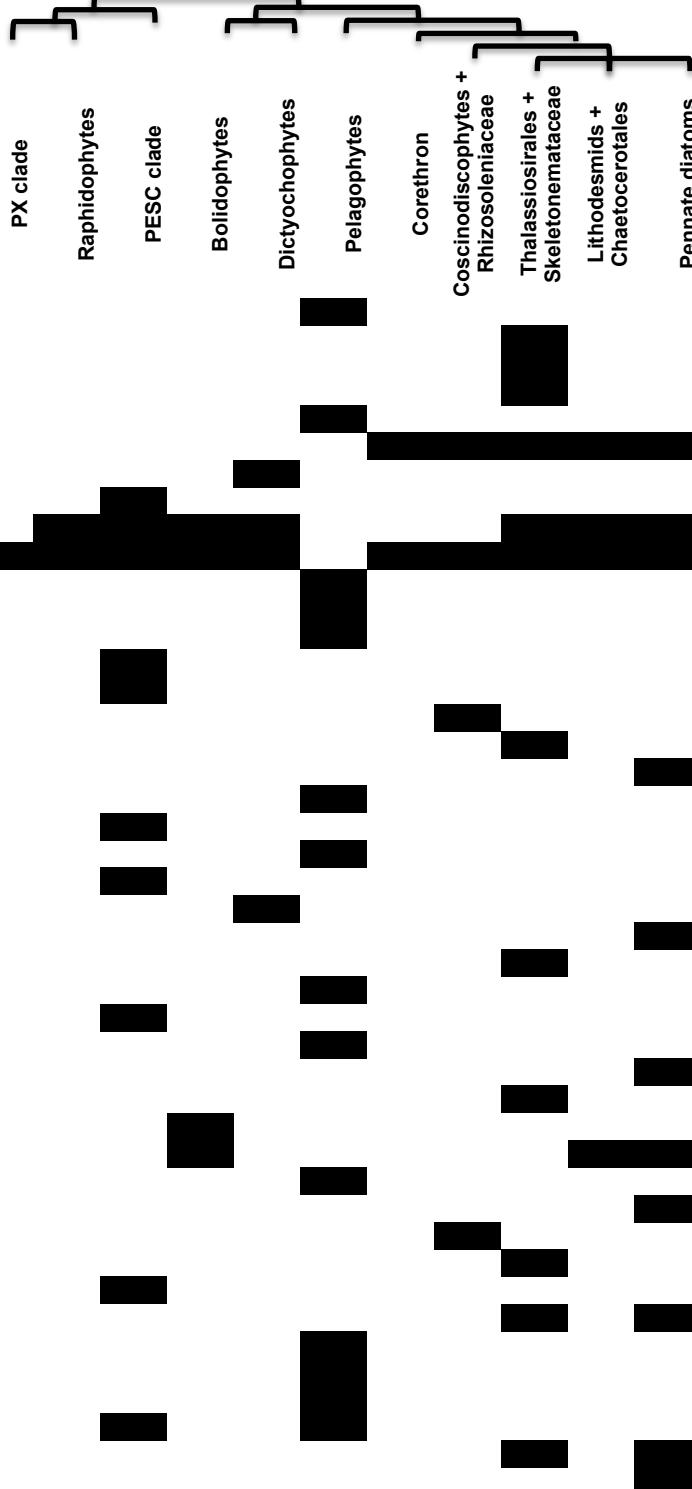
Corethron

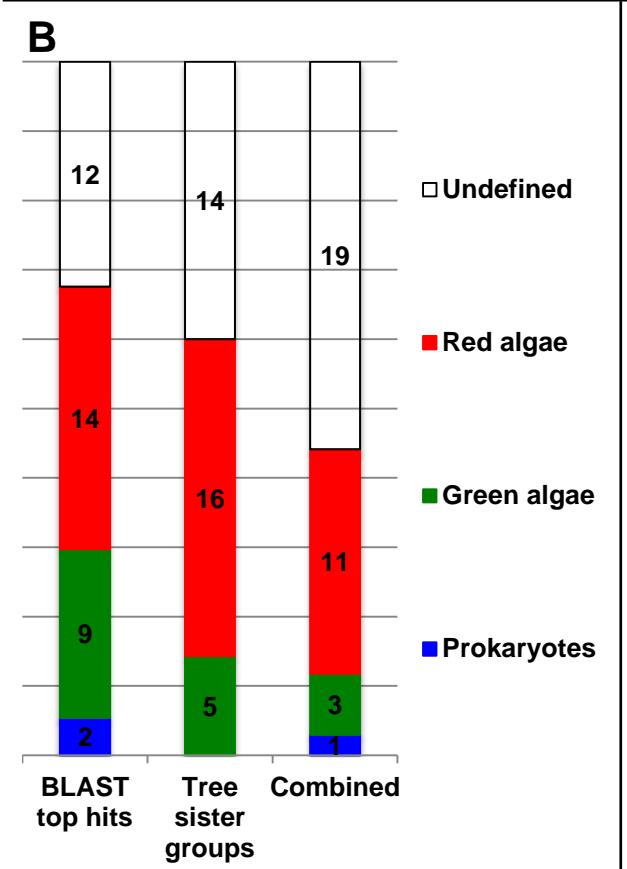
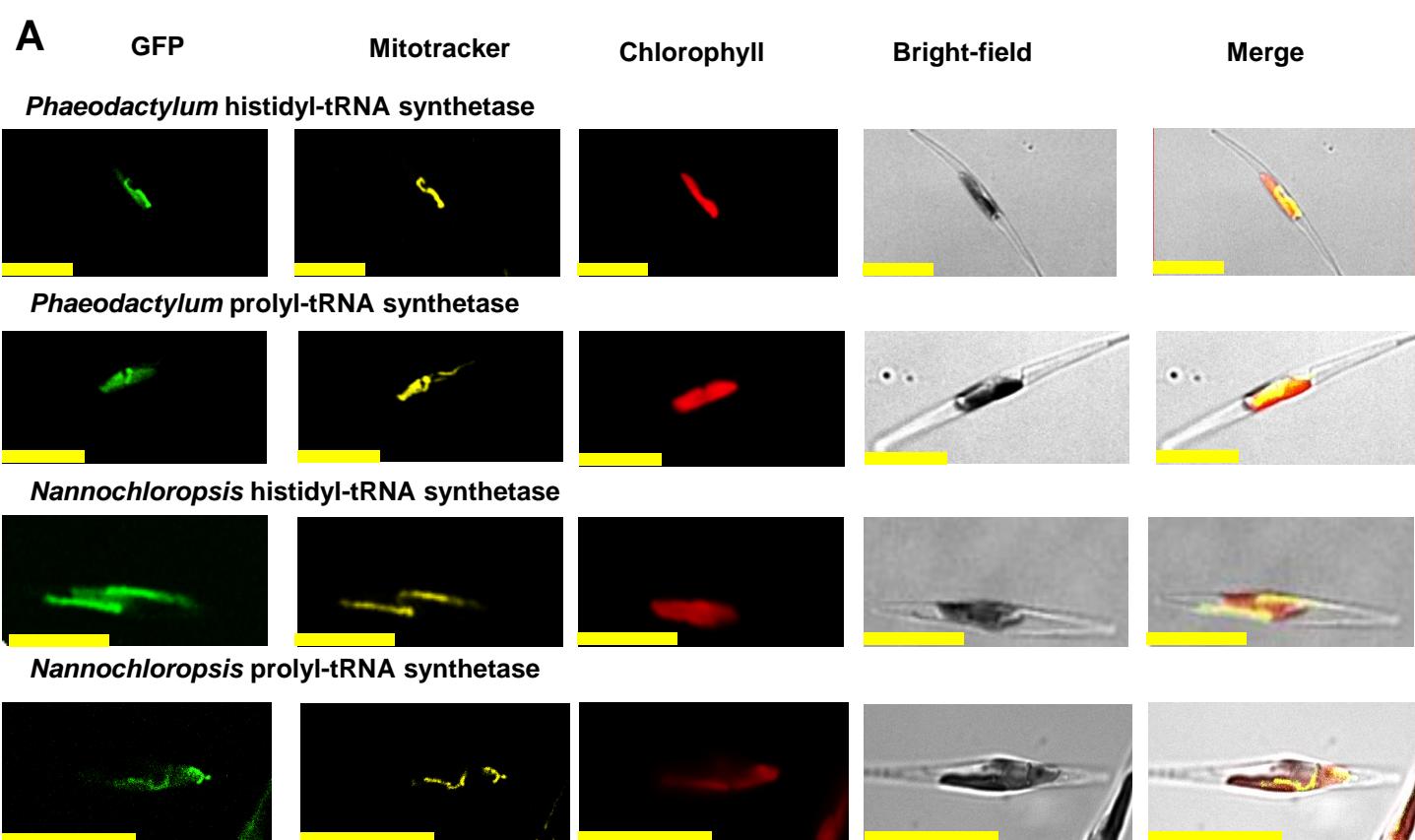
Coscinodiscophytes + Rhizosoleniaceae

Thalassiosirales + Skeletonemataceae

Lithodesmids + Chaetocerotales

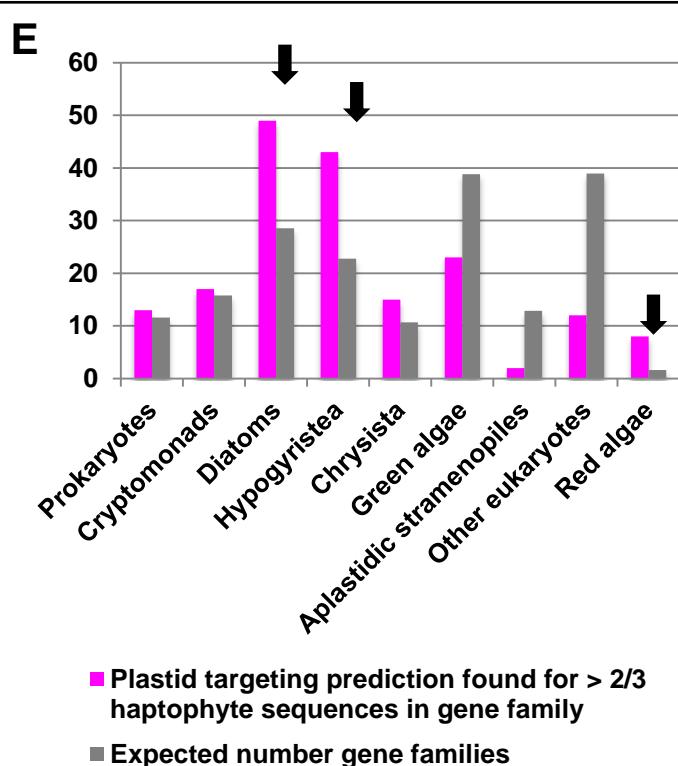
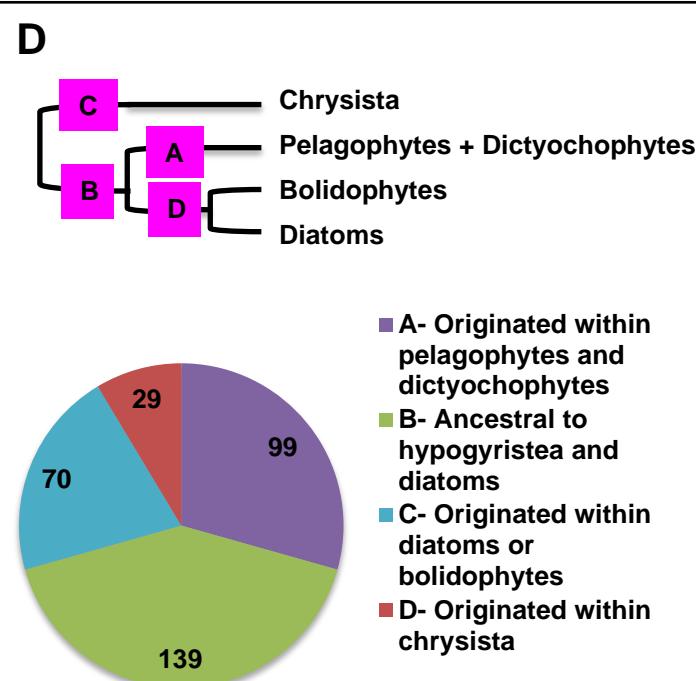
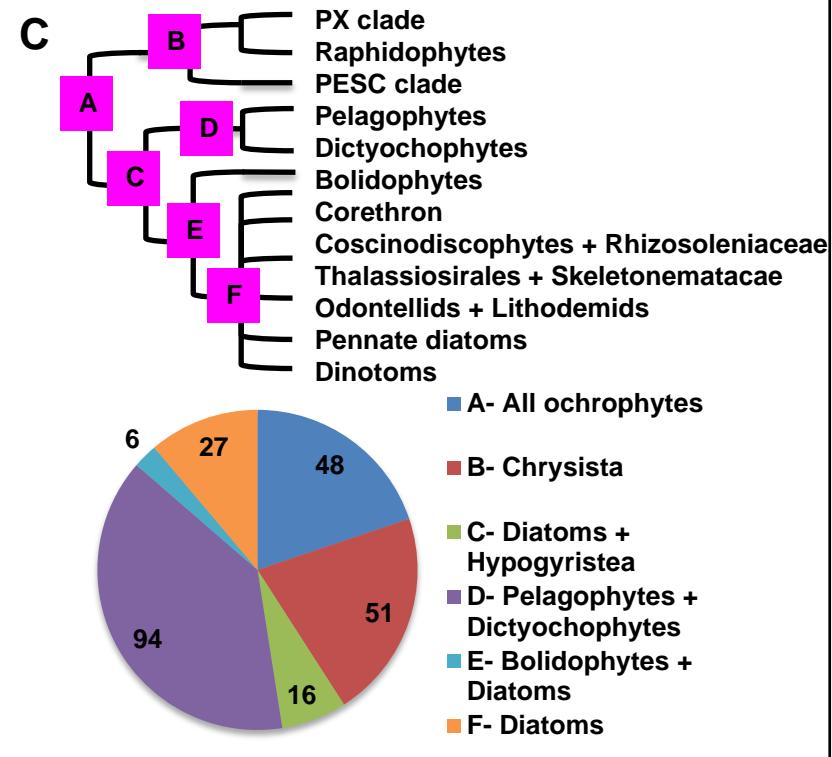
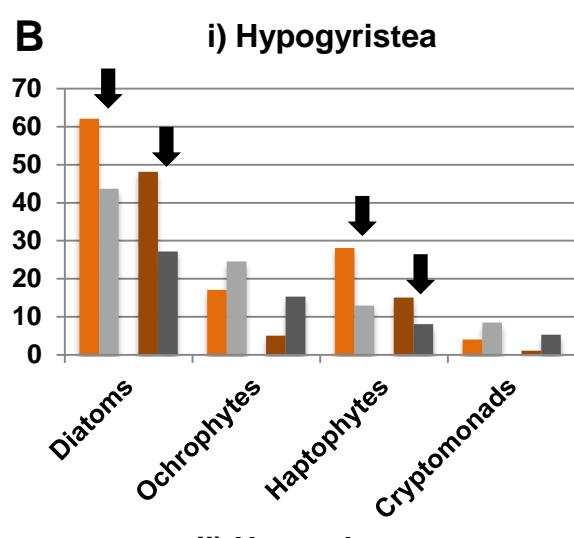
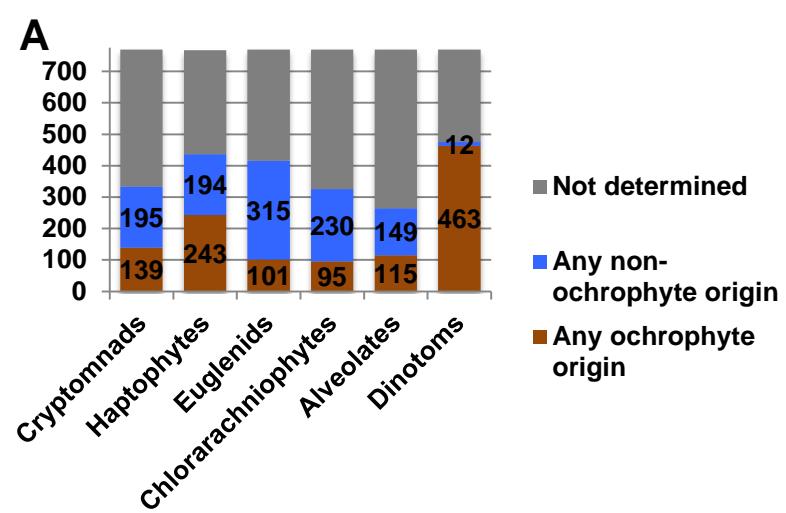
Pennate diatoms

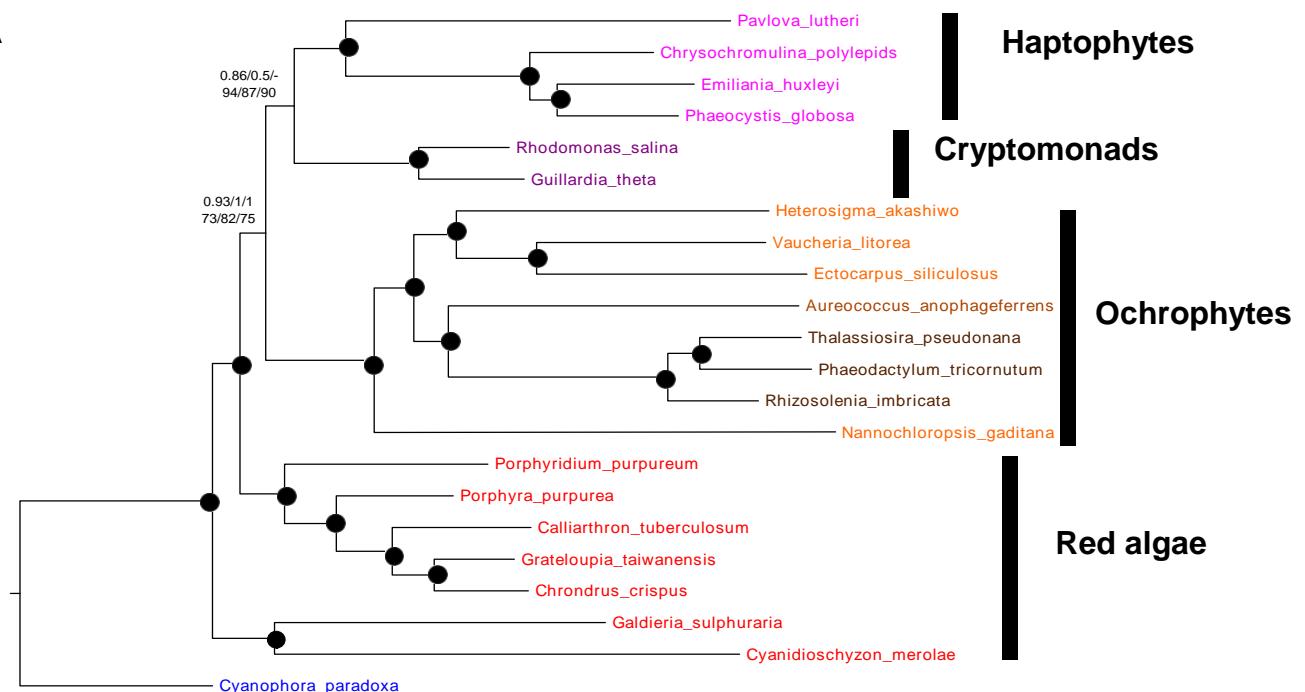
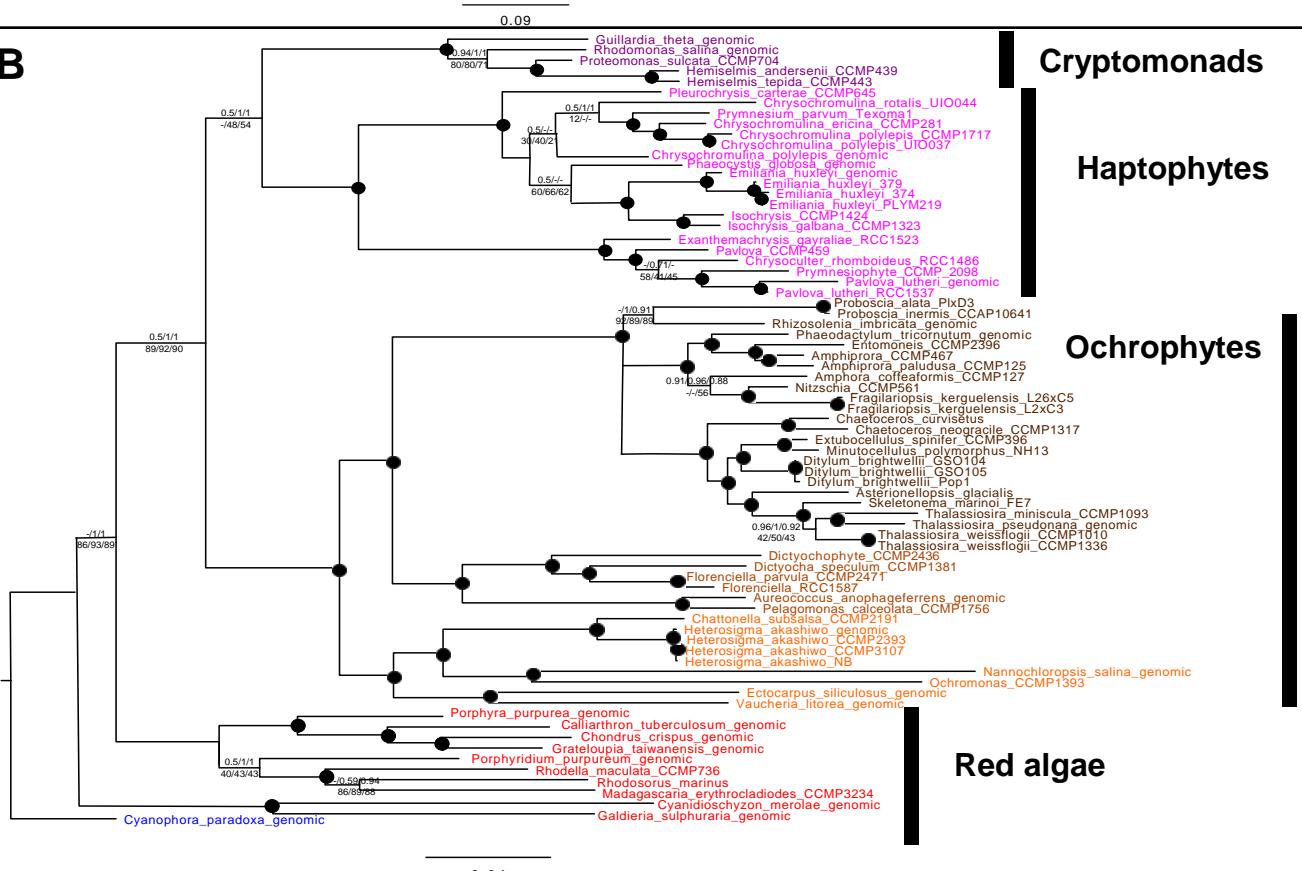




**C**

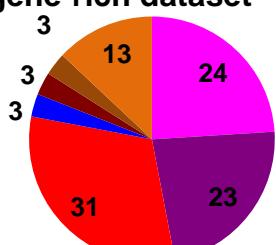
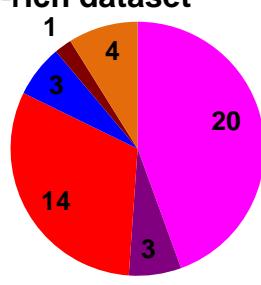
tRNA synthetase	Cytoplasmic isoform	Dual-targeted isoform
Ser	A plastidic stromal	Prokaryotic
Ala	A plastidic stromal	A plastidic stromal
Trp, Arg, Asn, Asp, Val	A plastidic stromal	Red algal



**A****Haptophytes****Cryptomonads****Ochrophytes****Red algae****B****Cryptomonads****Haptophytes****Ochrophytes****Red algae****Key to support values**

● Support value 1.0 in all MrBayes consensus trees; > 80% all ML best trees

Other nodes  
MrBayes: GTR/ Jones/ WAG  
RAxML: GTR/ JTT/ WAG

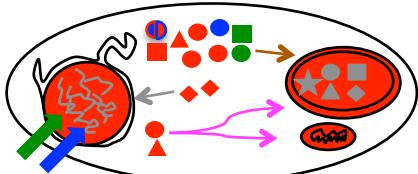
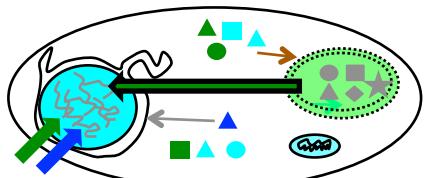
**C****i) gene-rich dataset****ii) taxon-rich dataset**

- Exclusive to haptophytes
- Cryptomonads
- Red algae
- Glaucoophytes
- Diatoms
- Hypogyristeria

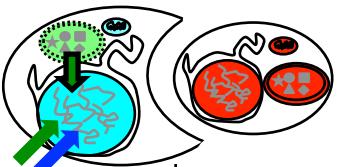
### Stramenopile host

### Red lineage symbiont

A



B



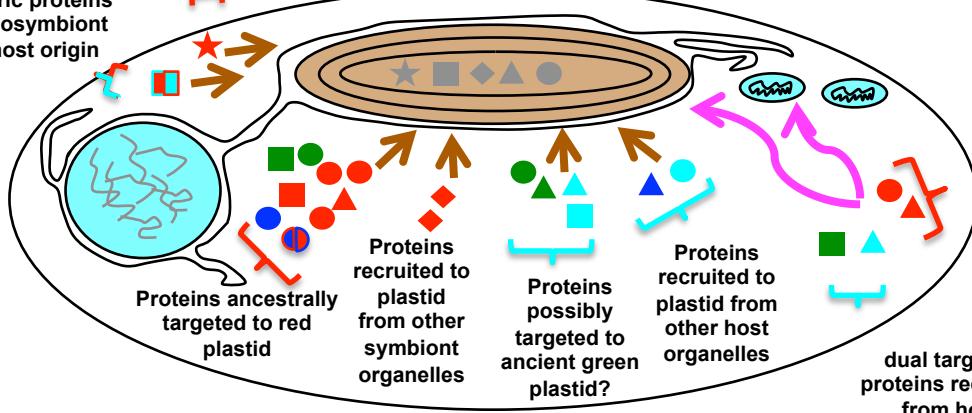
### Endosymbiotic intermediates

C

Proteins previously encoded in red plastid

### Ochrophyte ancestor

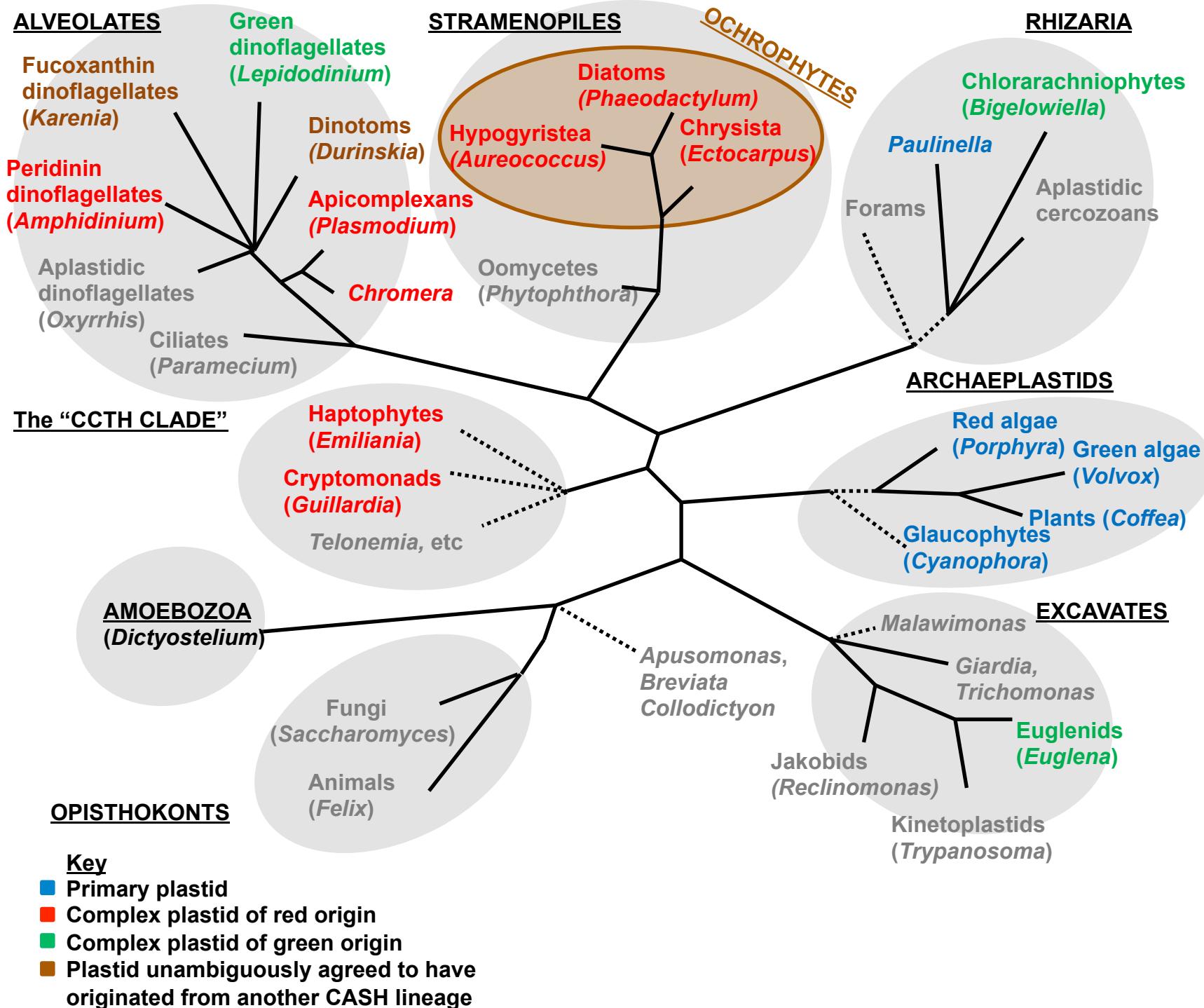
Chimeric proteins of endosymbiont and host origin



Proteins ancestrally dual targeted to red plastid and mitochondria

- Nucleus
- Mitochondrion
- Red lineage plastid
- Putative green plastid
- LGT from green algae
- LGT from prokaryotes
- Uniquely plastid-targeted proteins
- Uniquely mitochondria-targeted proteins
- Uniquely other (e.g. ER)-targeted proteins

- Dual mitochondria and plastid-targeted proteins
- Different proteins that are nucleus-encoded and plastid-targeted in ochrophytes
- Chimeric/ fusion plastid proteins
- Proteins of acquired from endosymbiont (regardless of origin)
- Proteins acquired from host (regardless of origin)
- Proteins endogenous to red algae
- Proteins endogenous to stramenopile host
- Proteins acquired from green algae
- Proteins acquired from prokaryotes



**Fig. 1- figure supplement 1. Overview of eukaryotic diversity.**

This figure, adapted from a previous review<sup>3</sup>, profiles the diversity of different eukaryotic nuclear lineages. Each grey ellipse corresponds to one major clade, or “super-group” of eukaryotes. A brown ellipse within the stramenopile clade delineates the ochrophyte lineages. Dashed lines denote uncertain taxonomic relationship. For each taxon, a type species (defined either by the presence of a complete genome, extensive transcriptome library, or of particular anthropic significance) is given in brackets. Taxa that lack plastids are labelled in grey, and taxa with plastids are shaded according to the evolutionary origin of that plastid lineage.

**Fig. 2-figure supplement 1- Exemplar ochrophyte plastid protein alignments.** This figure shows untrimmed Geneious alignments for two ancestral HPPGs of unusual provenance. In each case the full length of the protein (labelled i) and N-terminal region only (ii) are shown, demonstrating the broad conservation of the N-terminus position. Sequences for which exemplar targeting constructs (*Phaeodactylum tricornutum*, *Nannochloropsis gaditana*, *Glenodinium foliaceum*) are shown at the top of each alignment.

### A i) full length

### ER heat-shock protein

1. *Phaeodactylum tricornutum*

2. *Nannochloropsis gaditana*

3. *Glenodinium foliaceum*

4. PESC *Chromulina nebula*

5. PESC *Nannochloropsis oceanica*

6. PX *Vaucheria litorea*

7. PX *Ectocarpus siliculosus*

8. Raph *Chattonella subsalsa*

9. Bol *Bolidomonas pacifica*

10. Dict *Rhizochromulina marina*

11. Pen *Fragilariaopsis cylindrus*

12. Pen *Pseudonitzschia multiseries*

### ii) NTD

1. *Phaeodactylum tricornutum*

2. *Nannochloropsis gaditana*

3. *Glenodinium foliaceum*

4. PESC *Chromulina nebula*

5. PESC *Nannochloropsis\_oce...*

6. PX *Vaucheria litorea*

7. PX *Ectocarpus siliculosus*

8. Raph *Chattonella subsalsa*

9. Bol *Bolidomonas pacifica*

10. Dict *Rhizochromulina marina*

11. Pen *Aureococcus anophage...*

12. Pen *CCMP2097*

13. D0 *Coretrohon hystrix*

14. D1 *Aulacoseira subartica*

15. D2 *Detonula conferacea*

16. D3 *Ditylum brightwelli*

17. Pen *Fragilariaopsis cylindrus*

18. Pen *Pseudonitzschia multi...*

### B i) full length

### Histidyl tRNA-synthetase

1. *Phaeodactylum tricornutum*

2. *Nannochloropsis gaditana*

3. *Glenodinium foliaceum*

4. Raph *Chattonella subsalsa*

5. Bol *Bolidomonas pacifica*

6. Dict *Florenceella parvula*

7. D1 *Dactyliosolen fragilissimus*

8. D2 *Thalassiosira pseudonana*

10. D3 *Chaetoceros affinis*

11. Pen *Synedra acus*

### ii) NTD

1. *Phaeodactylum tricornutum*

2. *Nannochloropsis gaditana*

3. *Glenodinium foliaceum*

4. Raph *Chattonella subsalsa*

5. Bol *Bolidomonas pacifica*

6. Dict *Florenceella parvula*

7. D1 *Dactyliosolen fragilissimus*

8. D2 *Thalassiosira oceanica*

9. D2 *Thalassiosira pseudonana*

10. D3 *Chaetoceros affinis*

11. Pen *Synedra acus*

### Key

→ ASAFAF motif

→ Conserved domain

← Position of PCR reverse primer



**Fig. 2- figure supplement 2. Tree of ochrophyte glycyl-tRNA synthetase sequences.**

This tree shows the consensus unrooted Bayesian topology for a 95 taxa x 487 aa alignment of glycyl tRNA synthetase sequences. The font colour of each sequences corresponds to the taxonomic origin (see legend below for details) and are labelled with the taxonomic identifiers previously defined in Table S1. Sequences labelled with chl\_ possess apparent plastid targeting sequences recognisable by CASH lineage plastids. The ancestral ochrophyte plastidic isoform, of apparent chlamydiobacterial origin, is labelled with a blue ellipse. Black circles at each node denote posterior probabilities of 1.0 in Bayesian inferences with three different substitution matrices (GTR, Jones, and WAG), and grey circles indicate posterior probabilities of 0.8 with at least two of these matrices. Support values for all remaining nodes, using both Bayesian and RAxML analysis, is provided in the form

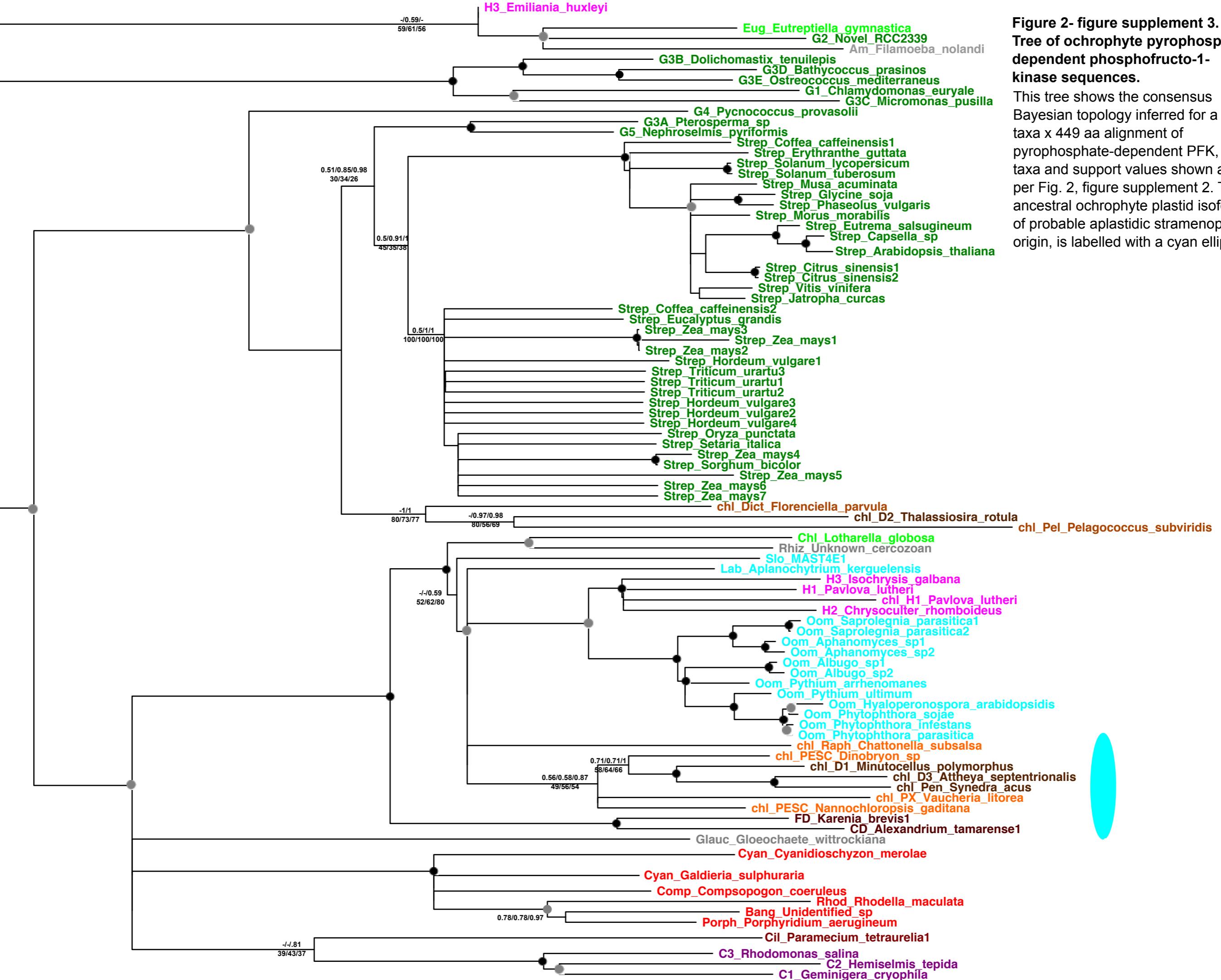
MrBayes posterior probabilities: GTR/Jones/WAG  
RAxML best tree likelihoods: GTR/ JTT/ WAG

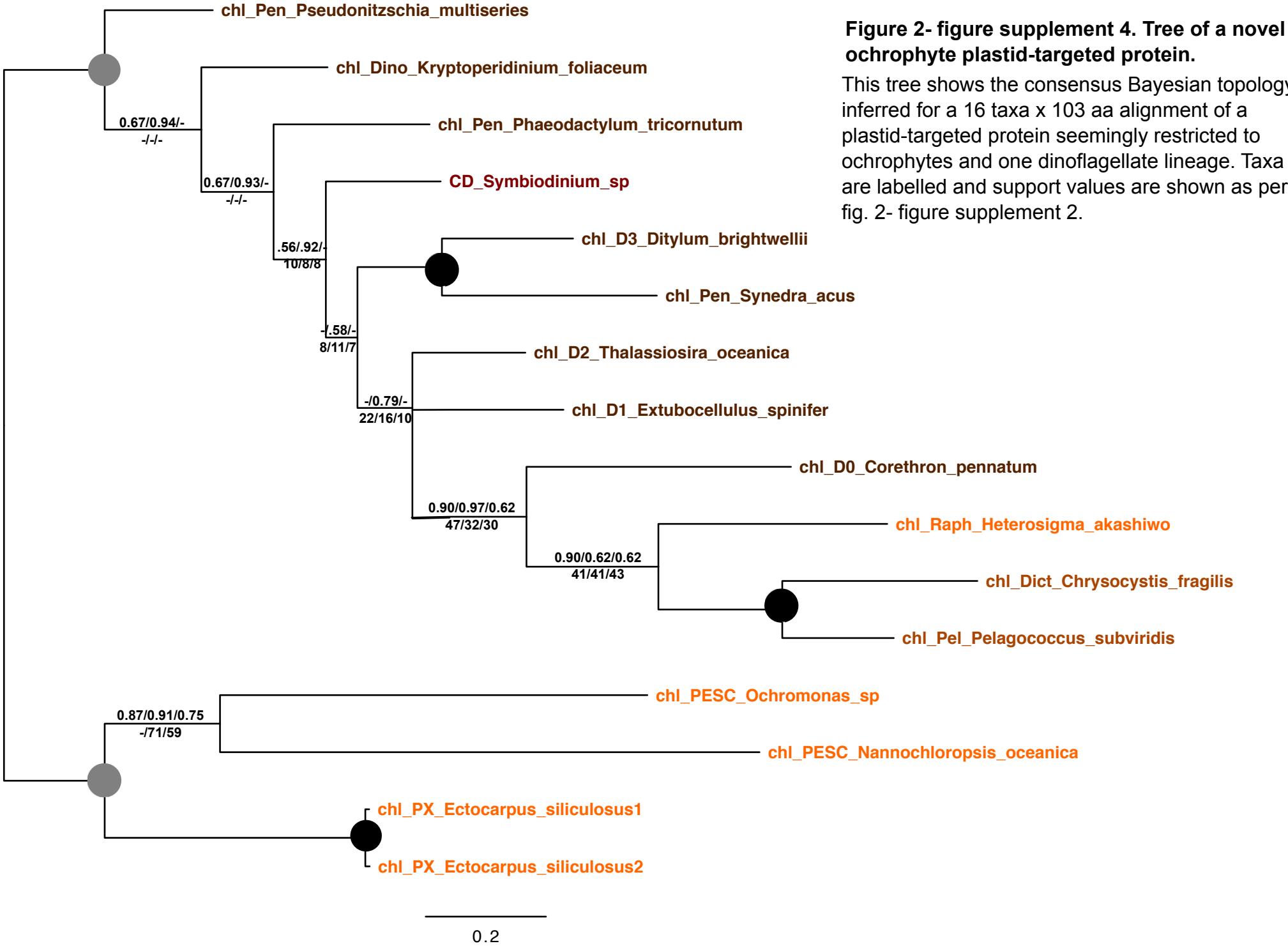
#### Taxonomic key

- Prokaryotes
- Red algae
- Green algae
- Aplastidic stramenopiles
- Haptophytes
- Cryptomonads
- Alveolates
- Chlorarachniophytes/Euglenids
- Ochrophytes-Chrysista
- Ochrophytes-Hypogyristeria
- Ochrophytes-Diatoms
- Other eukaryotes



**Figure 2- figure supplement 3.**  
**Tree of ochrophyte pyrophosphate dependent phosphofructo-1-kinase sequences.**  
This tree shows the consensus Bayesian topology inferred for a 94 taxa x 449 aa alignment of pyrophosphate-dependent PFK, with taxa and support values shown as per Fig. 2, figure supplement 2. The ancestral ochrophyte plastid isoform, of probable apoplastidic stramenopile origin, is labelled with a cyan ellipse.

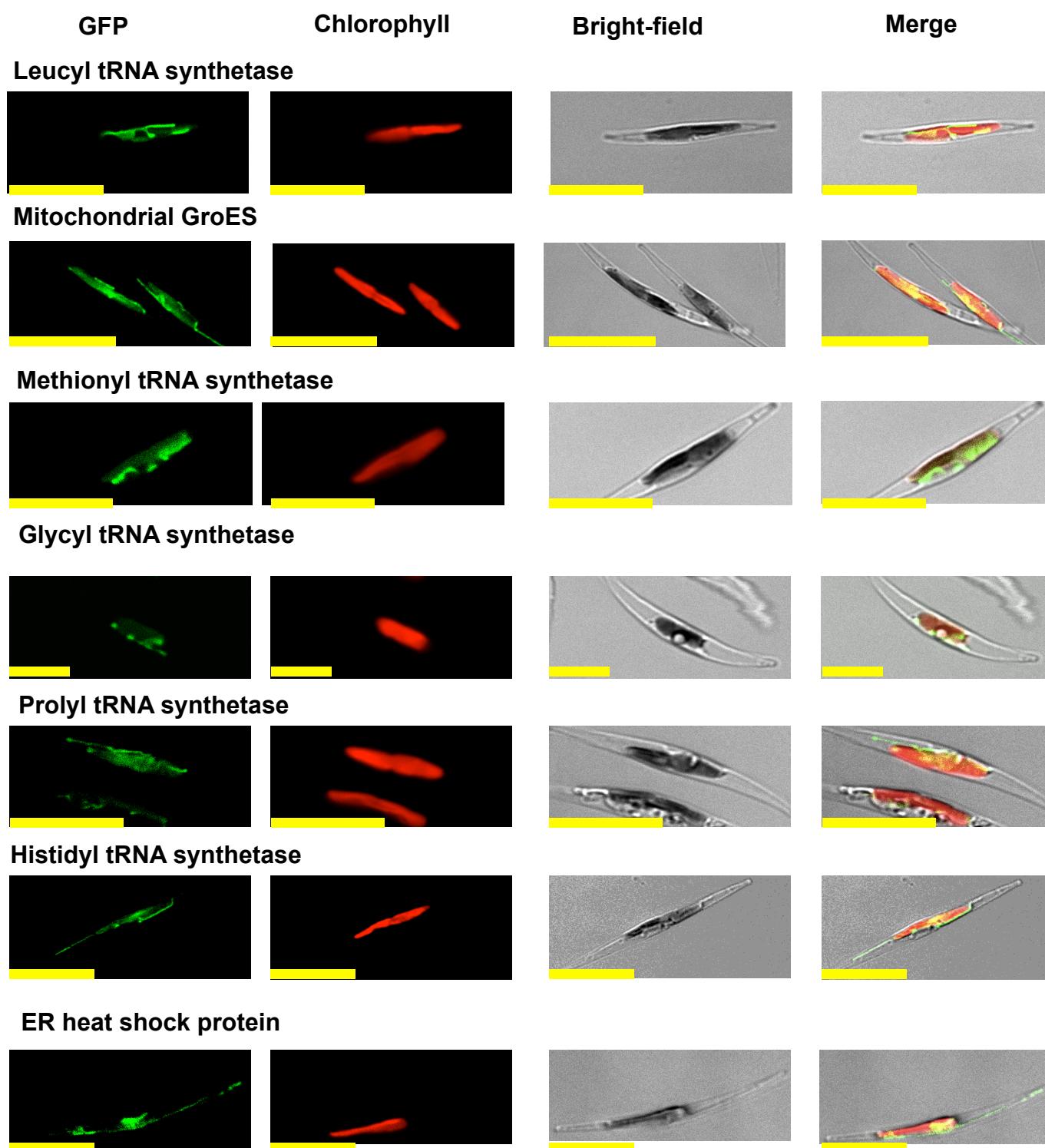




## **Figure 2- figure supplement 4. Tree of a novel ochrophyte plastid-targeted protein.**

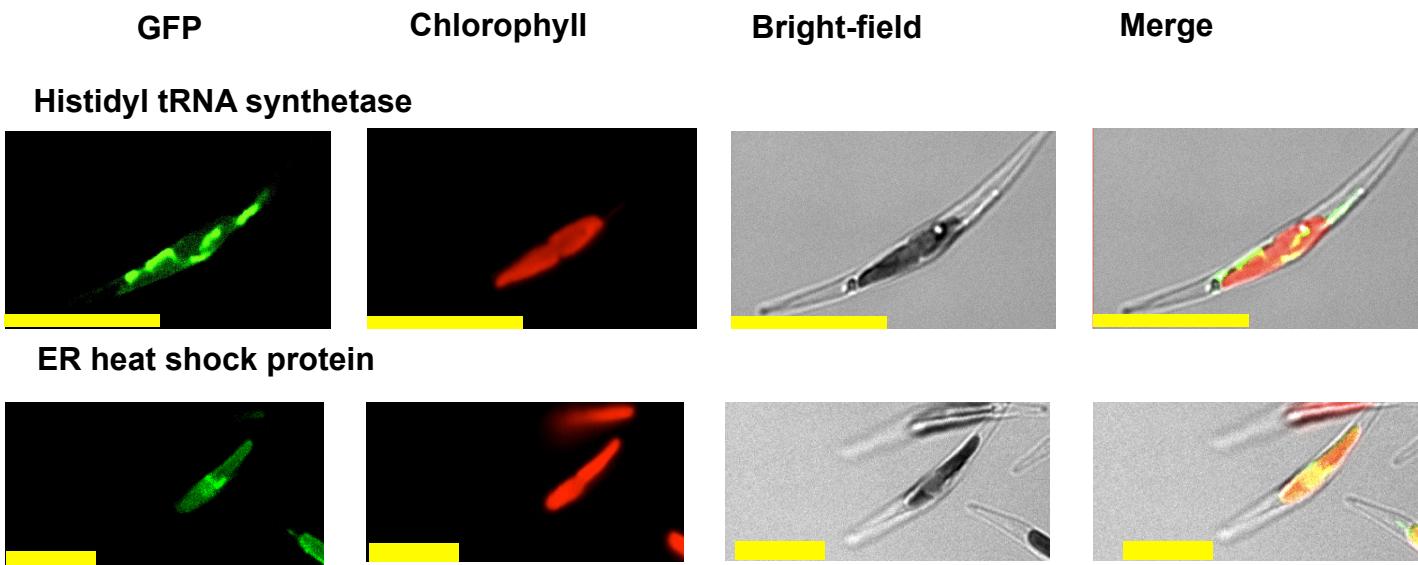
This tree shows the consensus Bayesian topology inferred for a 16 taxa x 103 aa alignment of a plastid-targeted protein seemingly restricted to ochrophytes and one dinoflagellate lineage. Taxa are labelled and support values are shown as per fig. 2- figure supplement 2.

**Fig. 2- figure supplement 5. Multipartite *Phaeodactylum* plastid-targeted proteins.** This figure shows the localisation of GFP overexpression constructs for copies of seven proteins from the diatom *Phaeodactylum tricornutum* that are of non-plastid origin, but show multipartite localisation to the plastid and one other organelle (the mitochondria, or in the case of the “ER heat shock protein” to the endoplasmic reticulum).

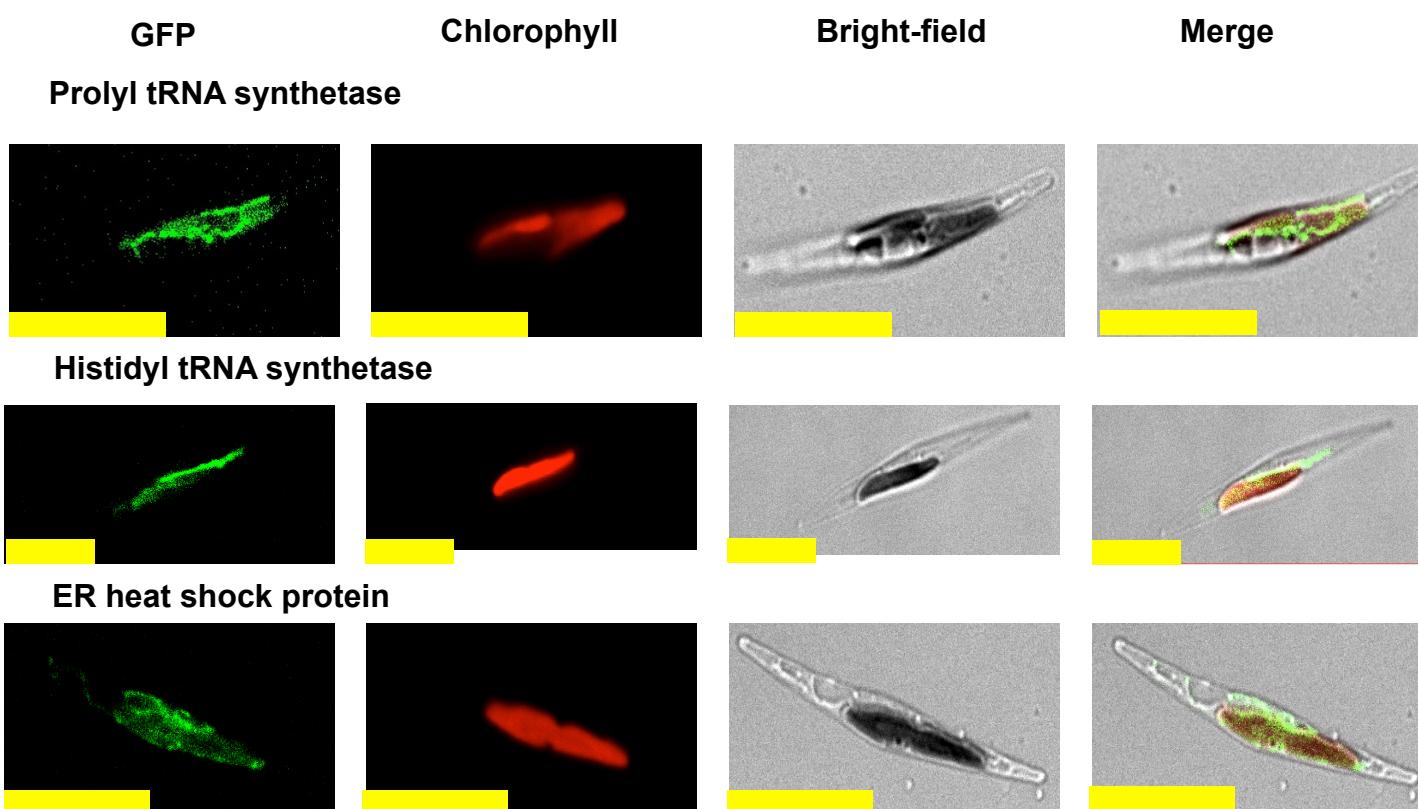


**Fig. 2-figure supplement 6. Heterologous expression constructs of multipartite plastid-targeted proteins.** This figure shows the localisation of GFP overexpression constructs for copies of two proteins from the dinotom *Glenodinium foliaceum* (**Panel A**), and three proteins from the eustigmatophyte *Nannochloropsis gaditana* (**Panel B**) that are of non-plastid origin, but show multipartite localisation to the plastid and one other organelle, per Fig. 2, figure supplement 5.

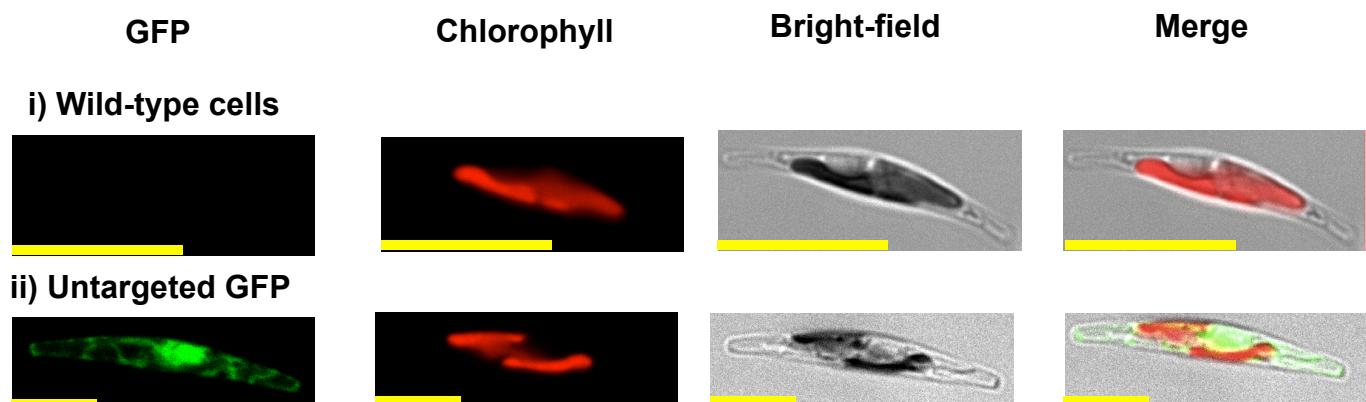
**A** *Glenodinium foliaceum*



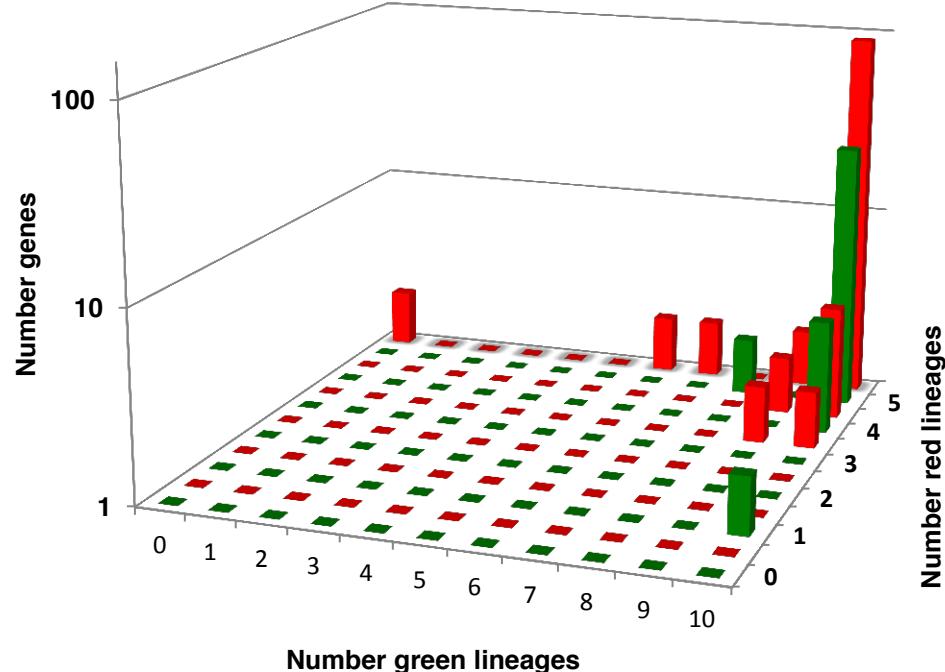
**B** *Nannochloropsis gaditana*



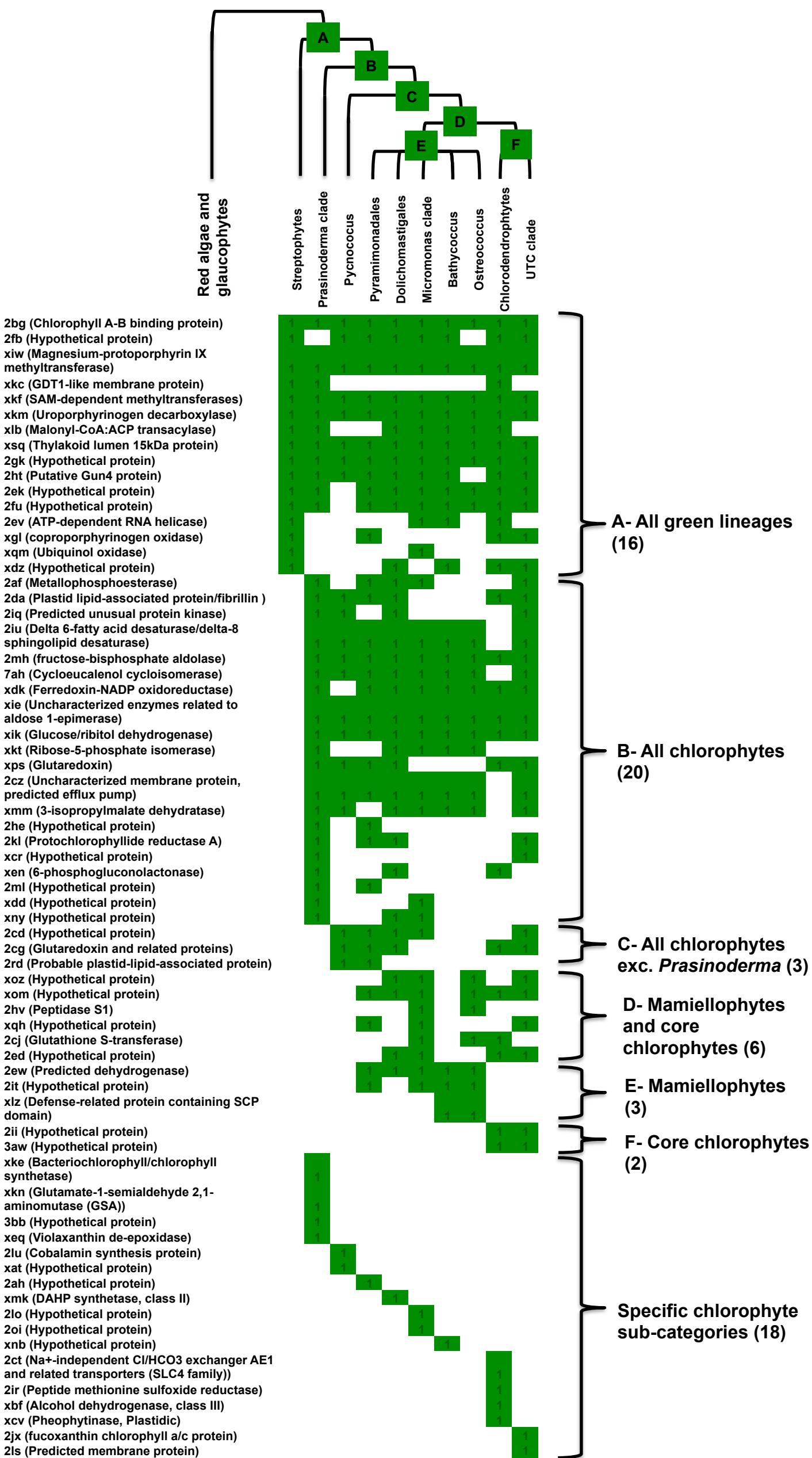
**Fig. 2-figure supplement 7. Exemplar control images for confocal microscopy.** This figure shows fluorescence patterns for wild-type *Phaeodactylum tricornutum* cells (i), and transformant *Phaeodactylum* cells expressing GFP that has not been fused to any N-terminal targeting sequence (ii), both visualised under the same conditions used for all other transformant cultures.



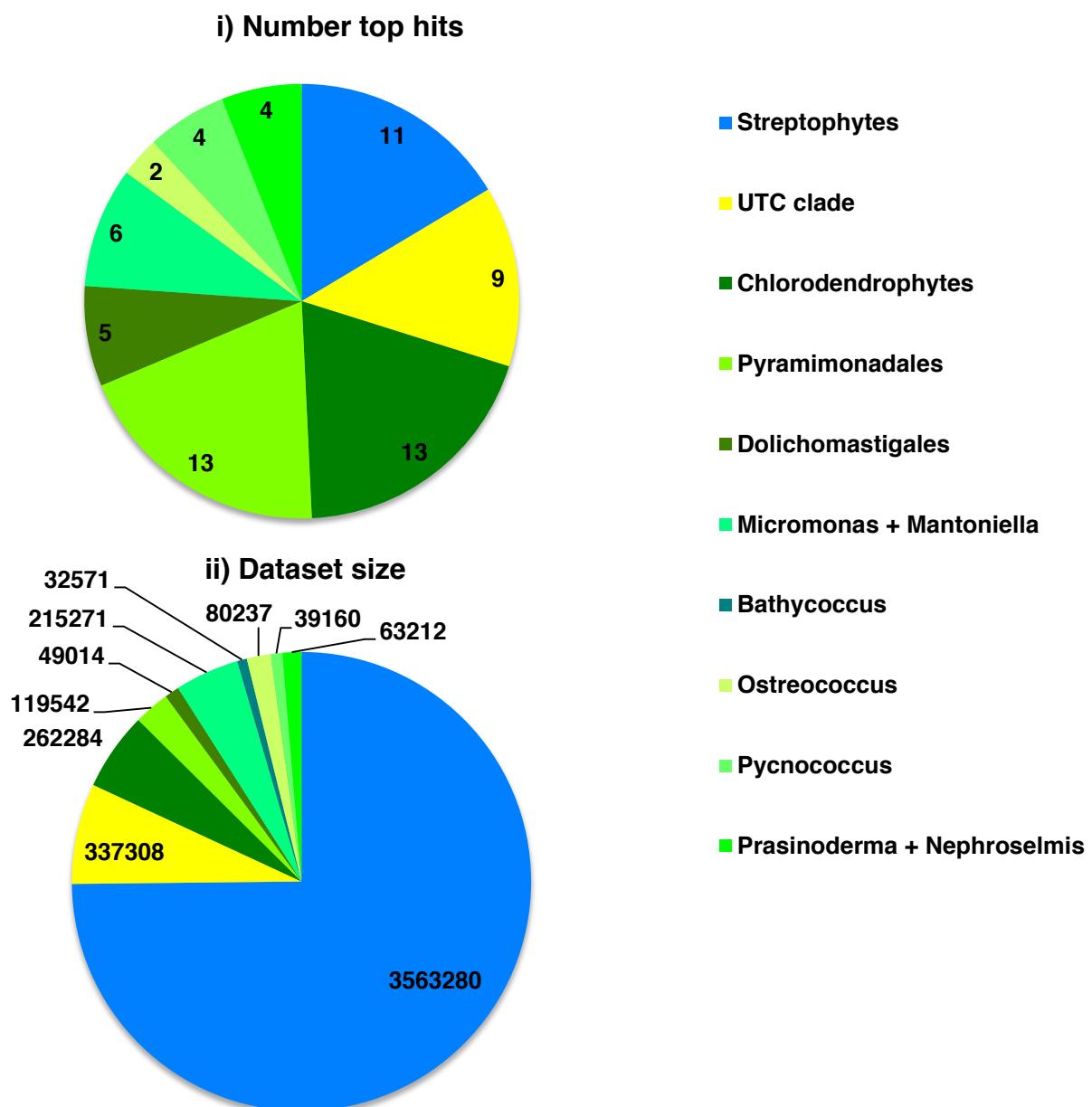
**Fig. 4- figure supplement 1. Sampling richness associated with ancestral HPPGs of green algal origin.** This figure shows the number of sub-different archaeoplastid orthologues for ancestral HPPGs verified by combined BLAST top hit and single-gene tree analysis to be of either green algal origin (green bars) or red algal origin (red bars), for which glaucophyte orthologues could also be identified.



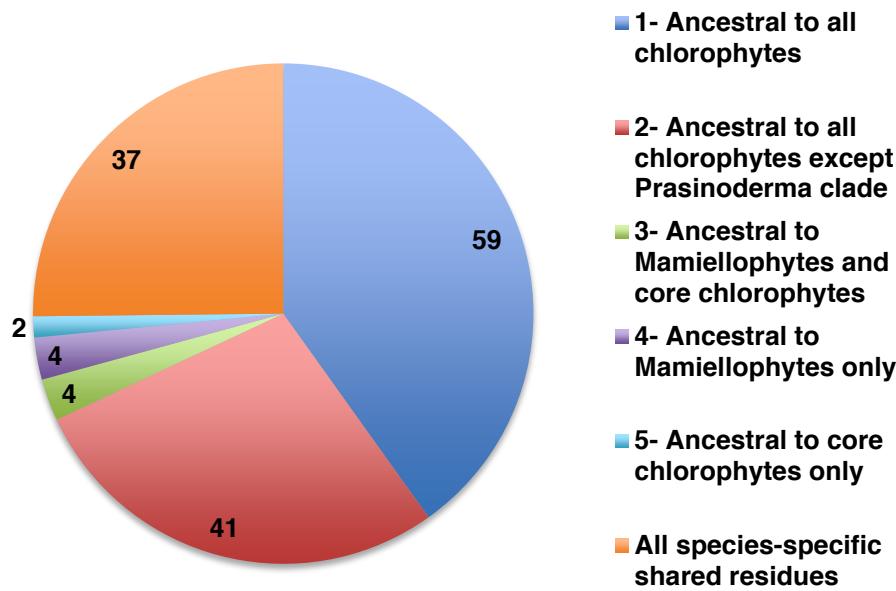
**Fig.4- figure supplement 2. Heatmaps of nearest sister-groups of ancestral HPPGs of verified green origin.** This figure shows the specific topologies of single gene trees for HPPGs verified to be of green origin by combined BLAST and phylogenetic analysis. **Panel A** shows a reference topology of evolutionary relationships between green lineages, defined as per Leliaert et al. 2011. Six ancestral nodes that might correspond to the origin point of ochrophyte HPPGs are labelled with coloured boxes. **Panel B** shows the presence and absence of each green sub-category in the immediate sister-group to the ochrophyte HPPG in each single tree of HPPGs of verified origin. HPPGs are grouped by the inferred origin point within the green algae, with the number of HPPGs identified for each origin point given with round brackets.



**Fig. 4-figure supplement 3. Specific origins of green HPPGs as inferred from BLAST top hit analyses.** These charts show (i) the number of BLAST top hits against each of the individual green sub-categories from HPPGs for which a green origin was identified both from BLAST top hit and single-gene tree analysis, and (ii) the total number of non-redundant sequences from each green sub-category included in the BLAST library.

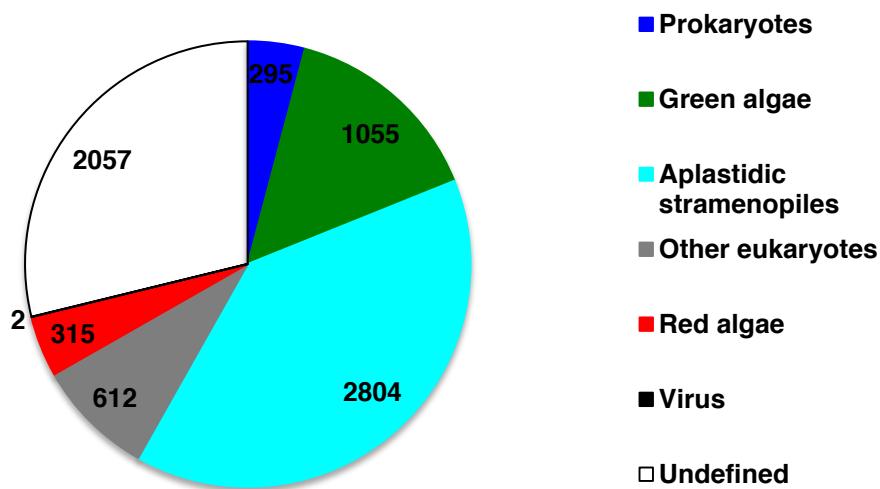


**Fig. 4- figure supplement 4. Earliest evolutionary origins of shared plastid residues.** This figure shows the number of residues in the concatenated alignment of HPPGs of verified green origin, which have been subsequently vertically inherited in all major photosynthetic eukaryotes that are present in green algae and ochrophytes, and are not found in red algae and glaucophytes. Residues are divided by inferred origin point, and are shown as per fig. 4, panel D. The values here are calculated as the earliest possible origin point for each uniquely shared residue, in which all gapped and missing positions within the alignment are treated as potential identities. 100 of the 147 residues inferred to have originated within green algae in this analysis originated either within a common ancestor of all chlorophytes, or in a common ancestor of all chlorophytes excluding the basally divergent lineages *Prasinoderma*, *Prasinococcus* and *Nephroselmis*.

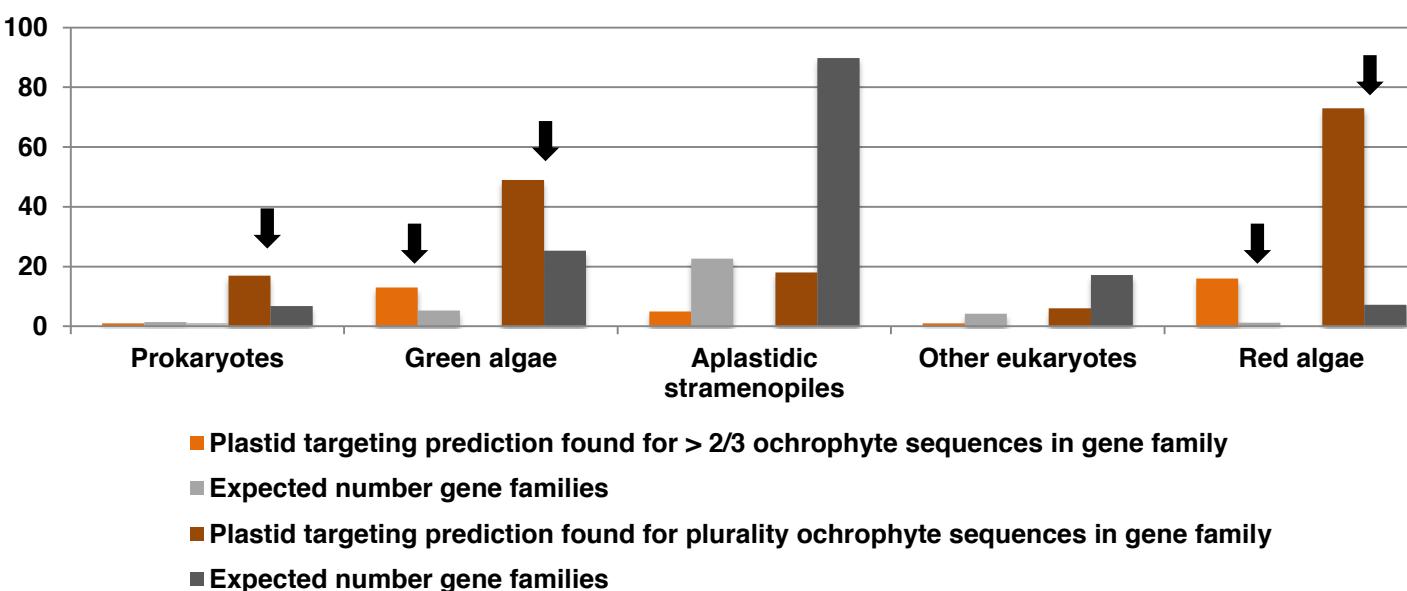


**Fig. 4- figure supplement 5. Origins and HECTAR based targeting tests of proteins encoded by conserved ochrophyte gene clusters.** Panel A shows the most probable evolutionary origin, identified using BLAST top hit analysis, for 7140 conserved gene clusters inferred to have been present in the last common ochrophyte ancestor. Panel B shows the number of these gene families that are predicted by HECTAR to encode proteins targeted to the plastid, subdivided by probable evolutionary origin, and the number expected to be present in each category assuming a random distribution of plastid-targeted proteins across the entire dataset, independent of evolutionary origin. Categories inferred to be significantly enriched above the expected values are labelled with black arrows.

**A**

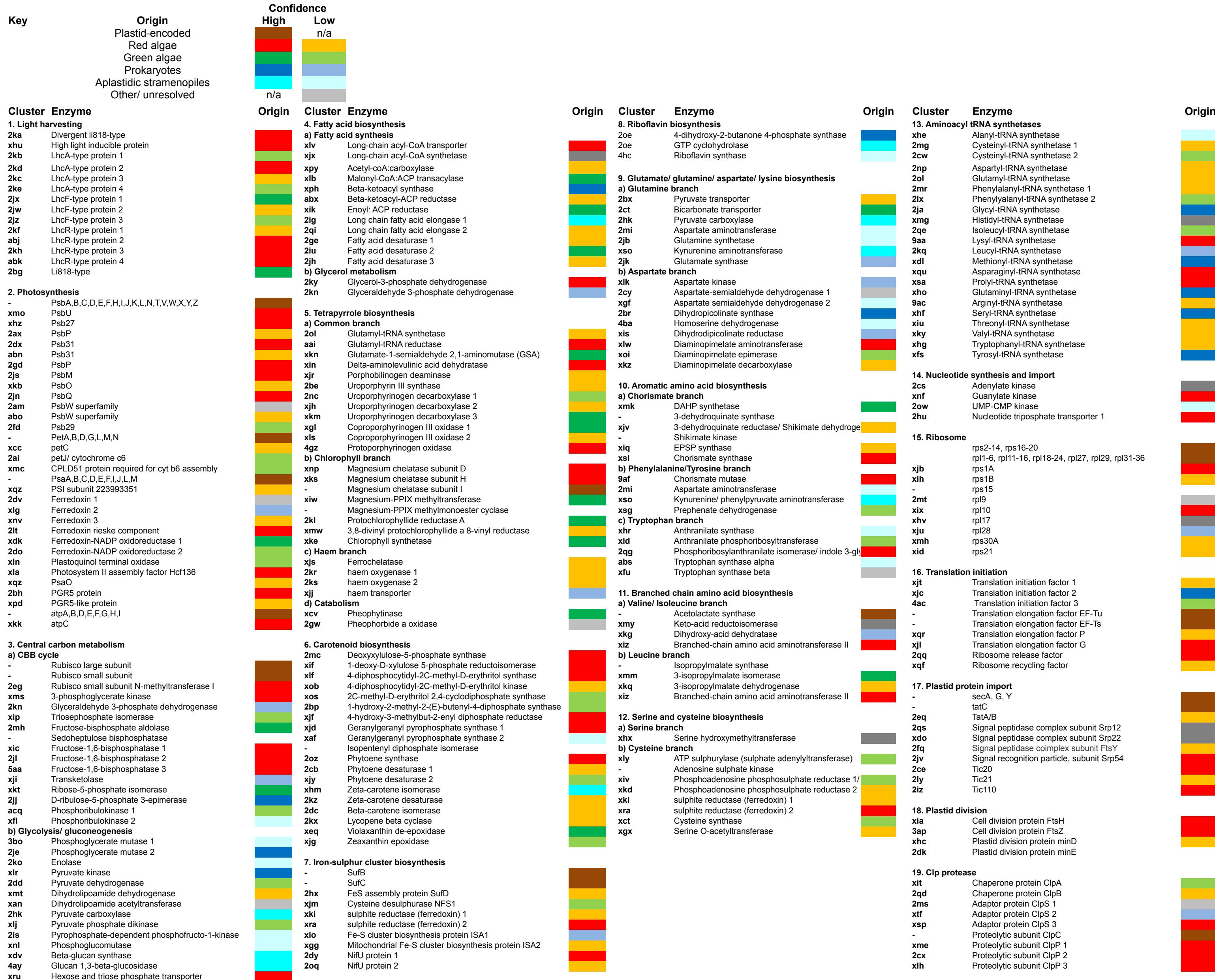


**B**



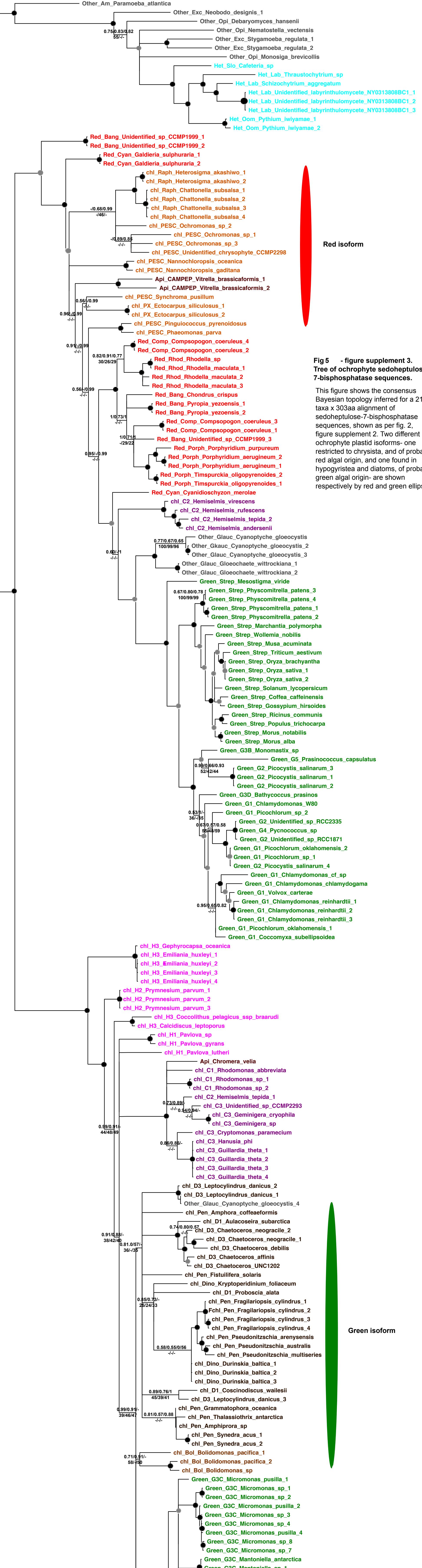
### Fig. 5-figure supplement 1. Reconstructed metabolism pathways and core biological processes in the ancestral ochrophyte plastid.

This figure tabulates each of the ancestral ochrophyte HPPGs corresponding to 350 central plastid metabolism and other biological processes. The "origin" column shows the probable evolutionary source for each HPPG as defined by combined BLAST tophit and single-gene tree analysis. The origin of each ancestral HPPG is either assigned a "high confidence" value (in which the same origin was robustly supported both by single-gene tree and by BLAST tophit analysis) or a "low confidence" value (in the absence of robust and consistent support through both techniques; corresponding to the tree sister-group if one could be clearly assigned, or the BLAST tophit identity if not). A dash indicates the corresponding protein was not identified in the ancestral HPPG dataset due to either being plastid-encoded or alternative reasons; detailed explanations for the enzymes that are neither plastid-encoded nor detected in the ancestral HPPG dataset are provided in figure supplement 2.



**Fig. 5- figure supplement 2. Core plastid metabolism proteins not identified within the ancestral HPPG dataset.**

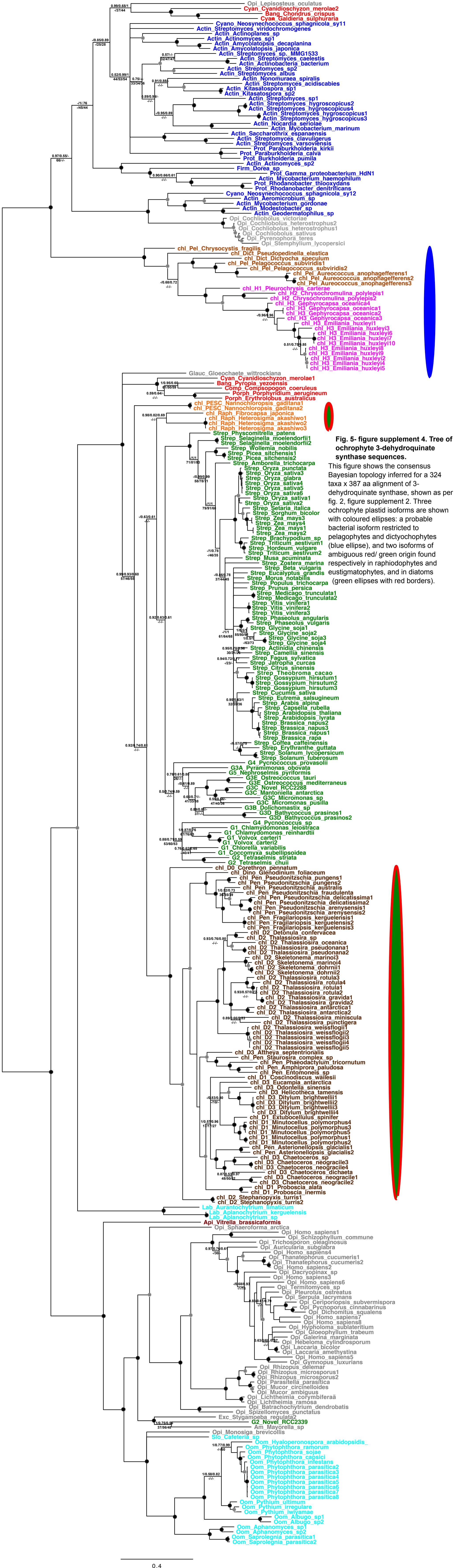
Enzyme	Pathway	Distribution	Probable explanation	References
Sedoheptulose-bis-phosphatase	CBB cycle	Multiple isoforms	Functionally conserved, but with different LGT events in different ochrophyte lineages	Fig. supplement 3
Transaldolase	CBB cycle	Hypogyristeae and diatoms	Functionally complemented by sedoheptulose-bis-phosphatase/fructose-bisphosphate aldolase	Kroth et al., 2008
Isopropylmalate dehydrogenase	Leucine biosynthesis	Multiple isoforms	Functionally conserved, but with different LGT events in different ochrophyte lineages	Fig. supplement 4
3-dehydroquinate synthase	Shikimate biosynthesis	Multiple isoforms	Functionally conserved, but with different LGT events in different ochrophyte lineages	Fig. supplement 5
Shikimate kinase	Shikimate biosynthesis	Multiple isoforms	Functionally conserved, but with different LGT events in different ochrophyte lineages	Fig. supplement 6
APS kinase	Fe-S cluster biosynthesis	Not found	Functionally dispensible; may be complemented by PAPS reductase	Gutierrez-Marcos et al. 1996
Magnesium protoporphyrin IX methylmonoester cyclase	Chlorophyll biosynthesis	Not found	Not known to be essential for chlorophyll metabolism outside of Tanaka and Tanaka green lineage	Tanaka and Tanaka 2007
Isopentenyl diphosphate isomerase	Carotenoid biosynthesis	Not found	Dispensable for isoprenoid metabolism	Ershov et al. 2000; Rohdich et al. 2002
rps15	Ribosomal small subunit	Not found	Not known outside of green lineage	Green 2011

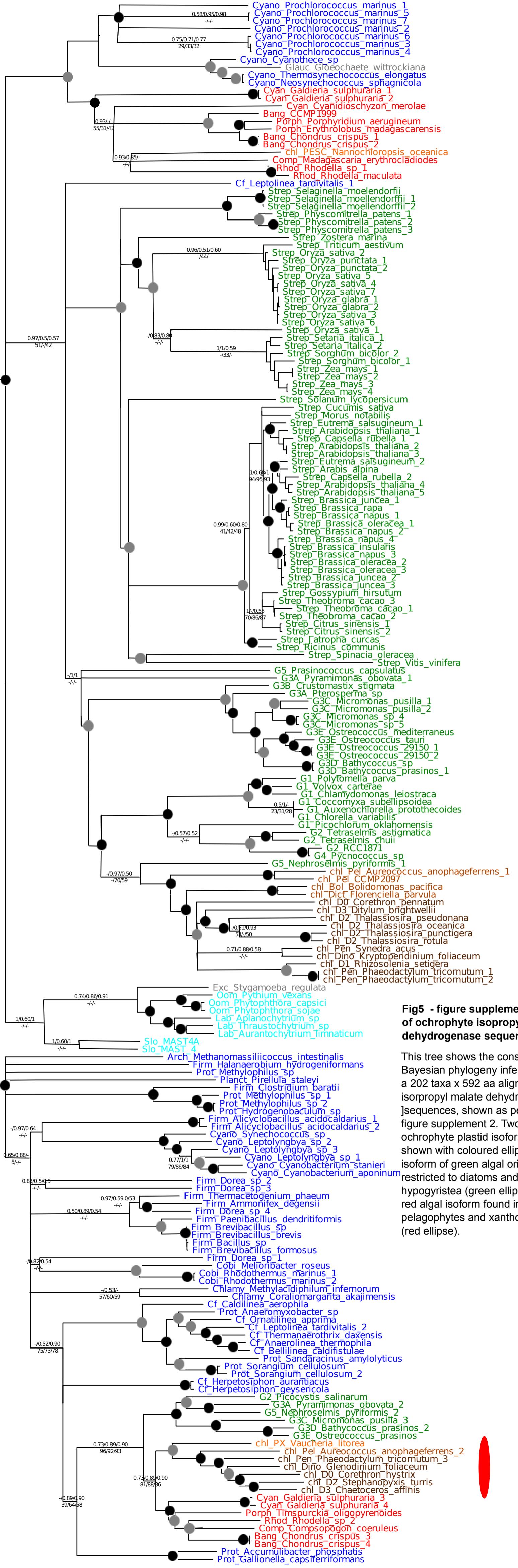


**Fig 5 - figure supplement 3.**  
Tree of ochrophyte sedoheptulose-7-bisphosphatase sequences.

This figure shows the consensus Bayesian topology inferred for a 218 taxa x 303aa alignment of sedoheptulose-7-bisphosphatase sequences, shown as per fig. 2, figure supplement 2. Two different ochrophyte plastid isoforms- one restricted to chrysista, and of probable red algal origin, and one found in hypogyrstea and diatoms, of probable green algal origin- are shown respectively by red and green ellipses.

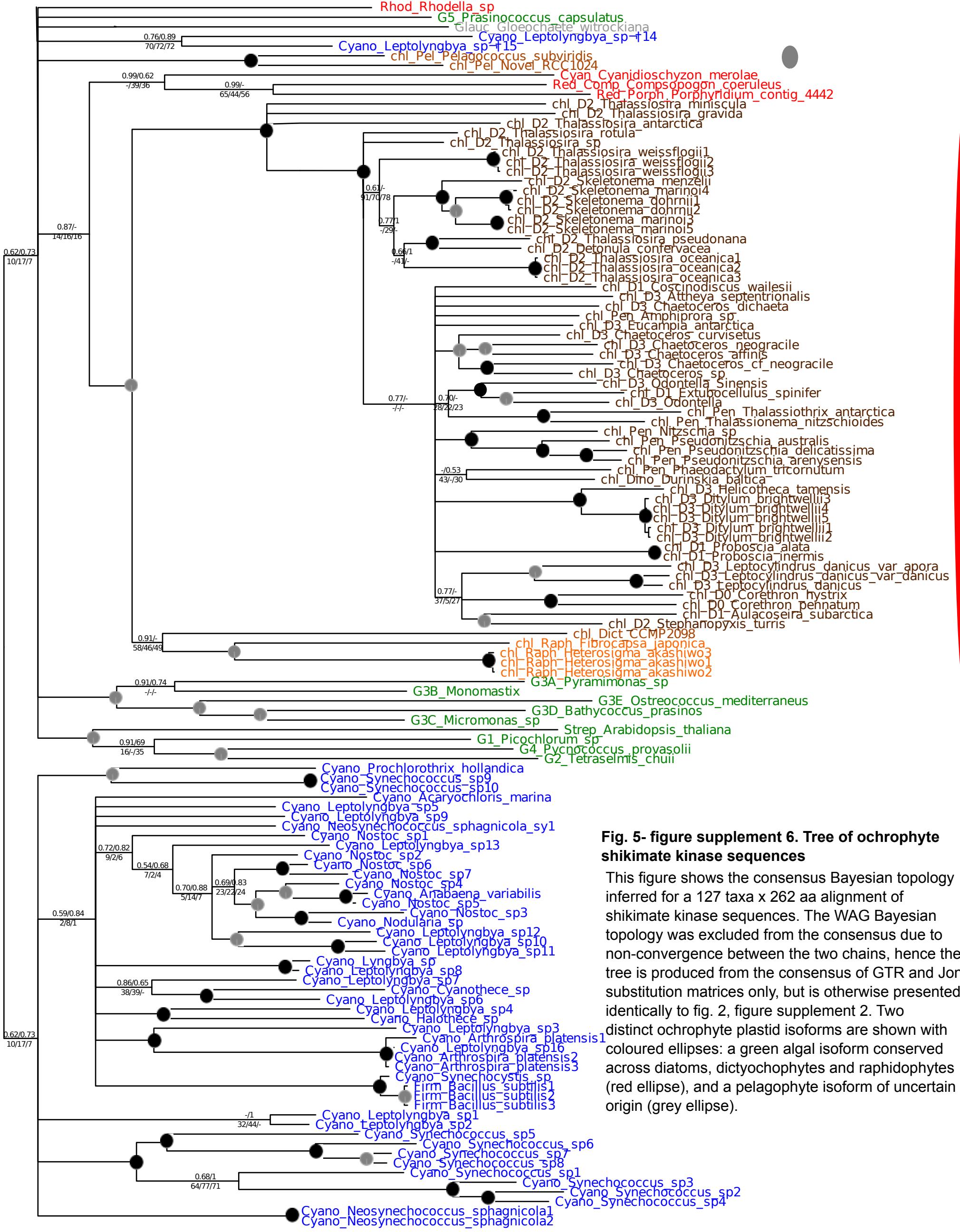
Green isoform





**Fig5 - figure supplement 5. Tree of ochrophyte isopropylmalate dehydrogenase sequences.**

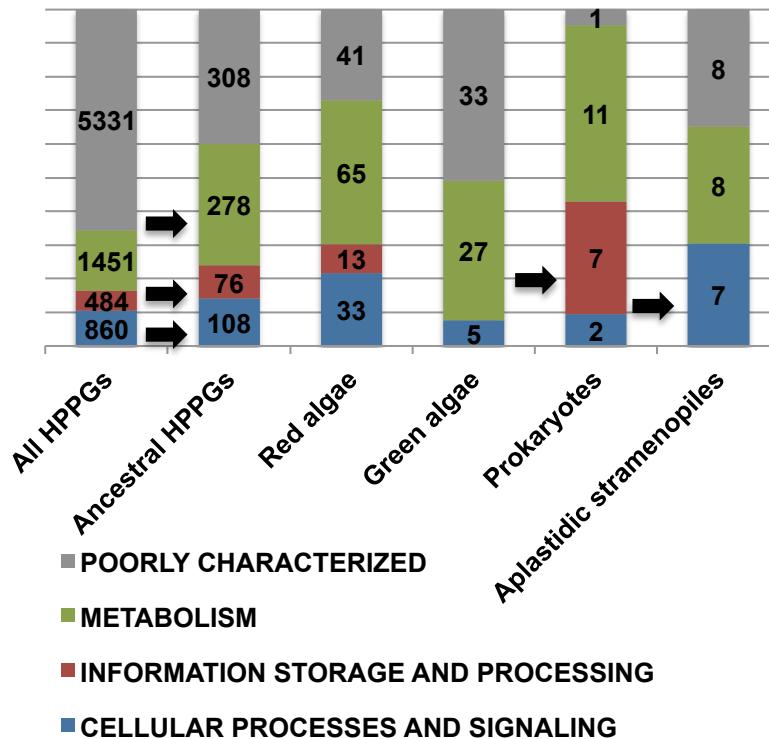
This tree shows the consensus Bayesian phylogeny inferred for a 202 taxa x 592 aa alignment of isopropyl malate dehydrogenase sequences, shown as per fig. 2-figure supplement 2. Two ochrophyte plastid isoforms are shown with coloured ellipses: an isoform of green algal origin restricted to diatoms and hypoglycophytes (green ellipse), and a red algal isoform found in diatoms, pelagophytes and xanthophytes (red ellipse).



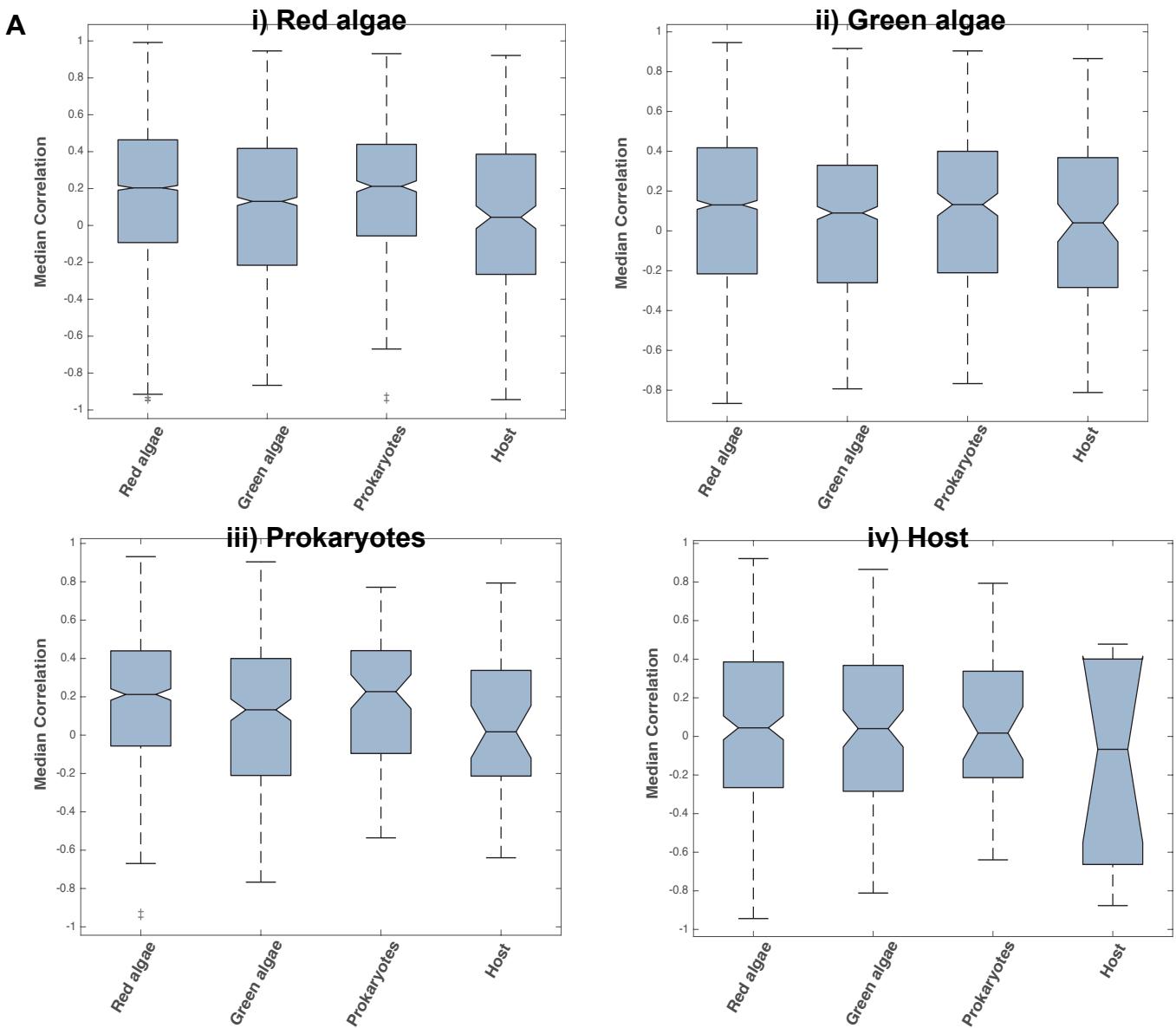
**Fig. 5- figure supplement 6. Tree of ochrophyte shikimate kinase sequences**

This figure shows the consensus Bayesian topology inferred for a 127 taxa x 262 aa alignment of shikimate kinase sequences. The WAG Bayesian topology was excluded from the consensus due to non-convergence between the two chains, hence the tree is produced from the consensus of GTR and Jones substitution matrices only, but is otherwise presented identically to fig. 2, figure supplement 2. Two distinct ochrophyte plastid isoforms are shown with coloured ellipses: a green algal isoform conserved across diatoms, dictyochophytes and raphidophytes (red ellipse), and a pelagophyte isoform of uncertain origin (grey ellipse).

**Fig. 5- figure supplement 7. KOG classes associated with different categories of HPPGs.**  
 These pie charts profile the distribution of different KOG classes across (i) all HPPGs except for those with general function predictions only, or without any clear KOG function, (ii) the same, but restricted to ancestral HPPGs and (iii) the same, for ancestral HPPGs of unambiguous red, green, prokaryotic and aplastidic stramenopile origin as identified by combined BLAST tophit and single-gene tree analysis. KOG classes that occur at elevated frequency in the ancestral HPPG dataset compared to the complete HPPG dataset, and one KOG class enriched in the prokaryotic HPPG dataset compared to the ancestral HPPG dataset (chi-squared test,  $P < 0.05$ ) are labelled with horizontal arrows.



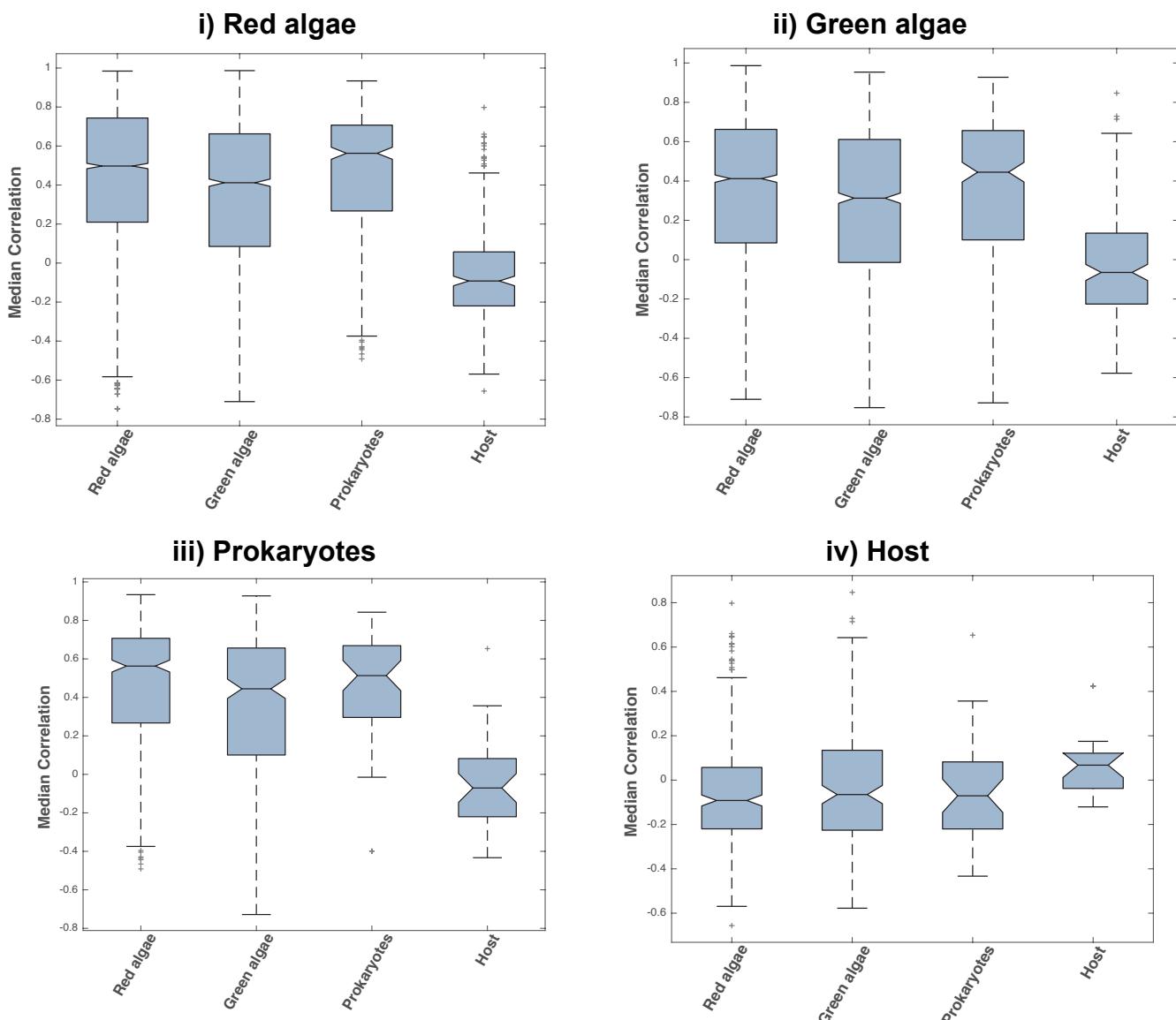
**Fig. 5- figure supplement 8. Coregulation of genes incorporated into HPPGs of different origin in the model diatom *Phaeodactylum tricornutum*.** Panel A shows boxplots of the correlation coefficients between the expression profiles of genes encoding members of ancestral HPPGs of red algal origin (i), green algal origin (ii), prokaryotic origin (iii) or host origin (iv), compared to genes encoding members of other HPPGs. Each HPPG is separated by evolutionary origin on the x-axis of each graph: for example, the box labelled “green algae” on the “red algae” graph shows the correlation coefficients between genes encoding members of ancestral HPPGs of red origin, and ancestral HPPGs of green origin. **Panel B** shows the P value statistics of mean separation calculated when comparing genes encoding members of ancestral HPPGs of the same origin (shown by row) to members of ancestral HPPGs of different origin (shown by column). For example, the intersect between the “red” row and “green” column shows the difference in mean correlation coefficient between pairs of genes that both encode members of ancestral HPPGs of red origin, and gene pairs of which one encodes an ancestral HPPG member of red origin, and the other an ancestral HPPG member of green origin. None of the P values calculated are significant, i.e. there are no categories of ancestral HPPG in which the internal correlation coefficients of gene expression are any different to those observed across the dataset as a whole.



	Red	Green	Prokaryotic	Host
<b>Red</b>		0.393	-0.945	0.491
<b>Green</b>	-0.555		-0.780	0.905
<b>Prokaryotic</b>	-0.358	-0.432		0.564
<b>Host</b>	-0.925	0.598	-0.475	

**Fig. 5- figure supplement 9. Coregulation of genes incorporated into HPPGs of different origin in the model diatom *Thalassiosira pseudonana*.** Boxplots (**Panel A**) and P value statistics (**Panel B**) are shown as per Fig. 5- figure supplement 8. Only two of the correlation value ANOVA tests (comparison of red-red and red-host correlations, and prokaryotic-prokaryotic and prokaryotic-host correlations, shaded in green) reveal a significantly higher correlation coefficient between pairs of genes encoding members of HPPG of the same evolutionary origin than pairs of genes encoding members of HPPGs with different evolutionary origins. These differences most probably reflect the extremely weak correlation coefficients associated with genes encoding HPPGs of host origin to all other genes considered (compare “Host” category on boxplots **i**, **ii** and **iii** to all other categories); however, detailed comparison of the correlation values between genes encoding ancestral HPPGs of host origin and genes encoding ancestral HPPGs of different evolutionary origin (**Panel A**, boxplot **iv**; **Panel B**, bottom row) reveals no specific difference in the pairwise correlation values observed between genes encoding ancestral HPPGs of host origin, and genes encoding ancestral HPPGs of all other origins within the dataset.

**A**



**B**

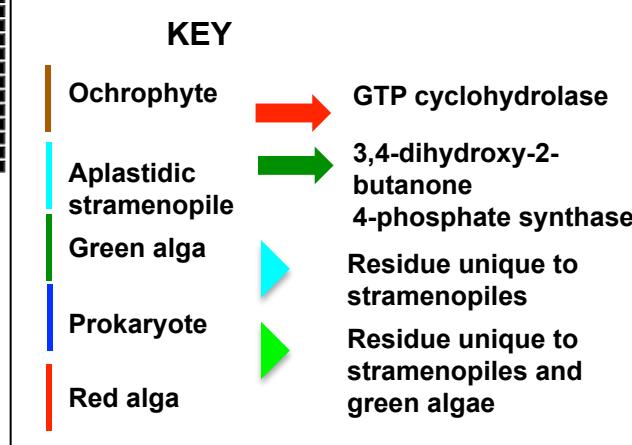
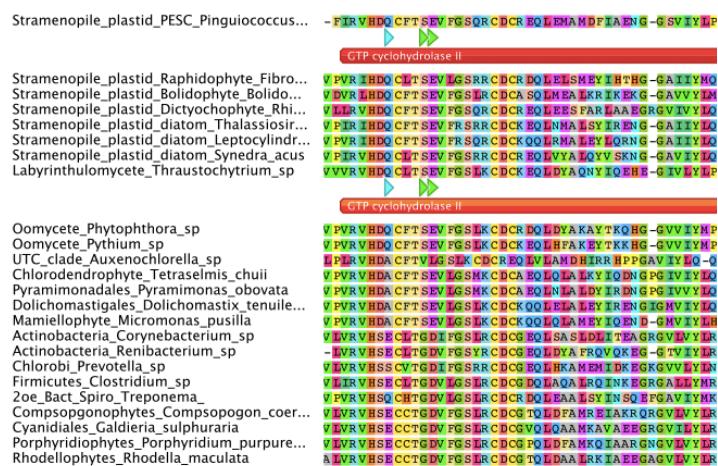
	Red	Green	Prokaryotic	Host
<b>Red</b>				
<b>Green</b>	-0.376			
<b>Prokaryotic</b>	0.279	0.473		
<b>Host</b>	-0.951	0.478	0.323	

**Fig. 6-figure supplement 1. Alignments of an ochrophyte-specific riboflavin biosynthesis fusion protein.** Panel A shows alignments of the full length (i) and cyclohydrolase domain only (ii) of a plastid-targeted GTP cyclohydrolase II/ 3,4-dihydroxy-2-butanone 4-phosphate synthase protein conserved across the ochrophytes. Coloured bars adjacent to each sequence correspond to the phylogenetic identity of the sequence. The cyclohydrolase domain of the ochrophyte protein is positioned in the N-terminal region, and the synthase domain in the C-terminal region. Three uniquely shared residues at the N-terminus of the cyclohydrolase domain confirm that it has been inherited from the aplastidic stramenopile ancestor of the ochrophytes.

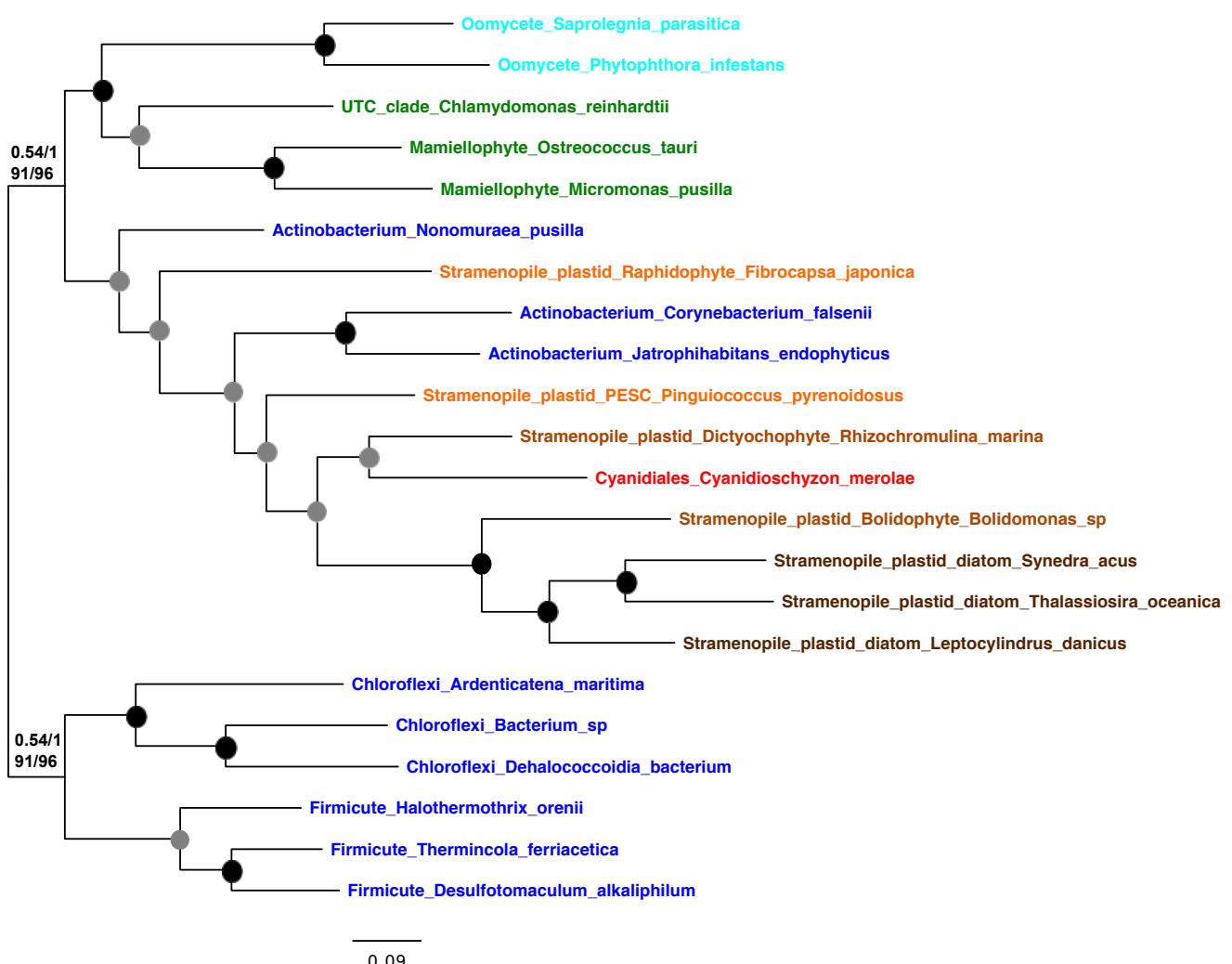
### A) i) Full sequence length



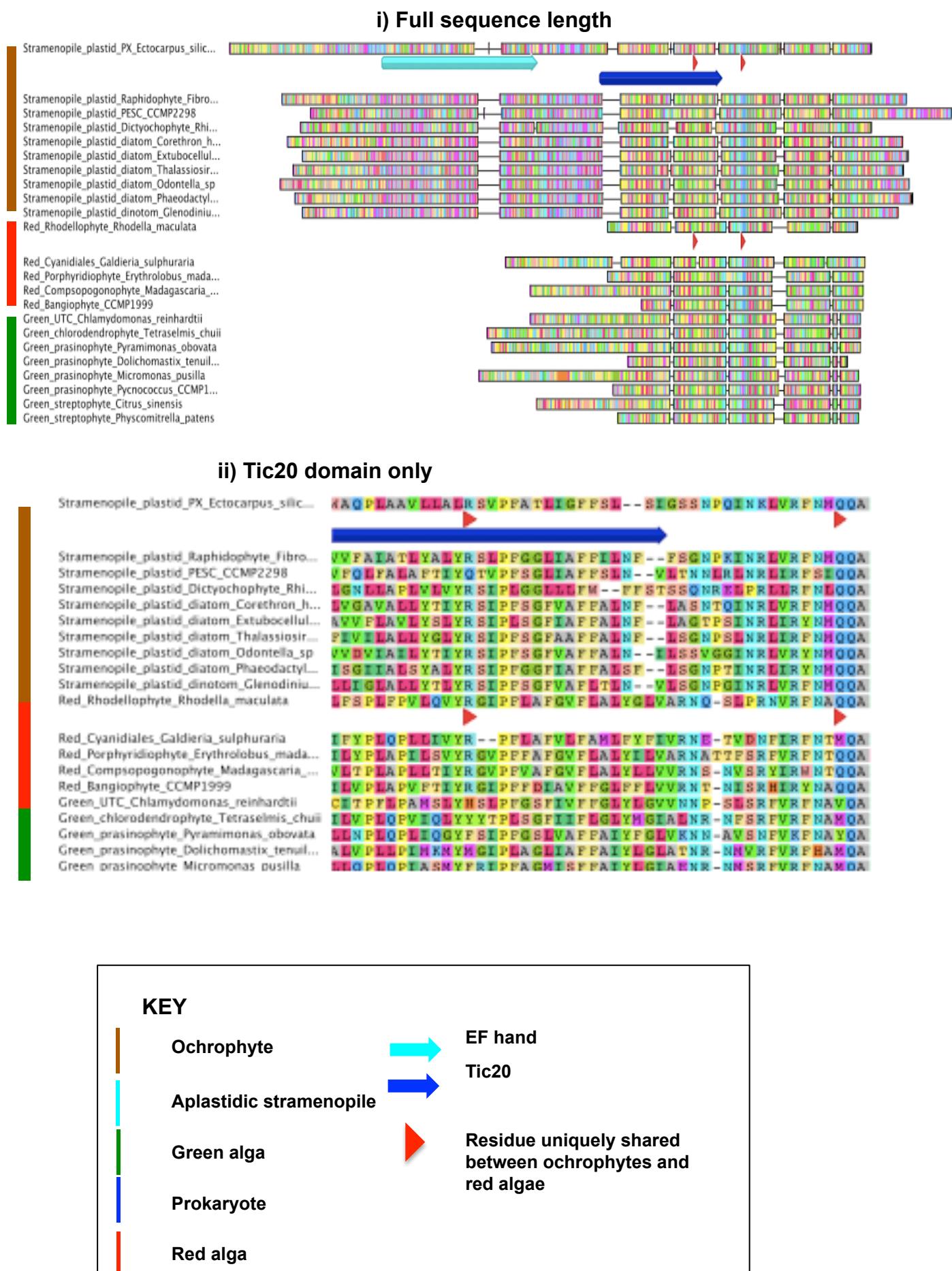
### ii) Cyclohydrolase domain only



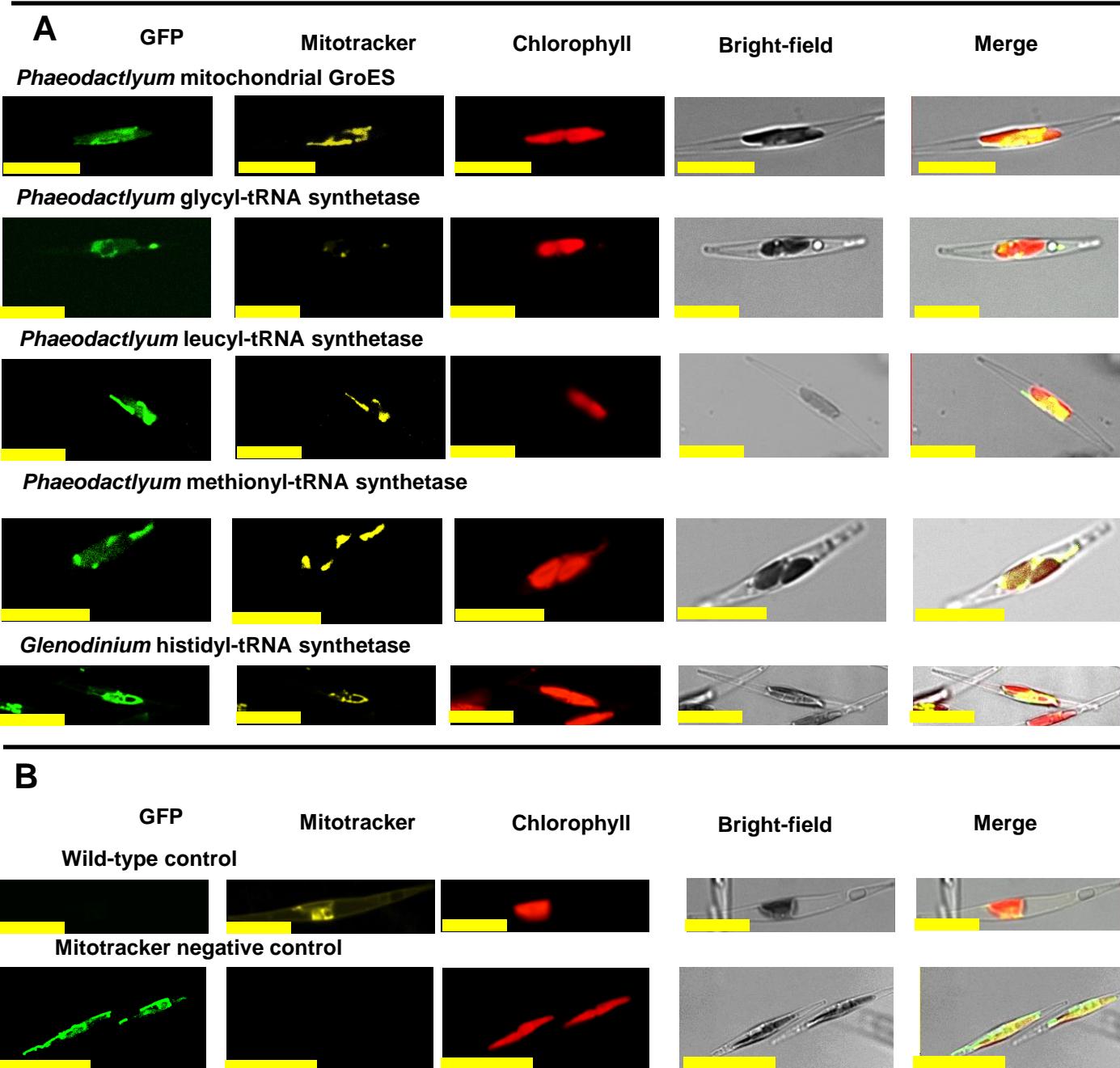
**Fig. 6-figure supplement 2. Origins of ochrophyte plastid 3,4-dihydroxy-2-butanone 4-phosphate synthase.** This figure shows the consensus Bayesian topology inferred for a 22 taxa x 206 aa alignment of 3,4-dihydroxy-2-butanone 4-phosphate synthase domains from different lineages, inferred using Jones and WAG matrices, and shown as per fig. 2, figure supplement 2. The ochrophyte plastid isoforms branch with red algal and actinobacterial sequences.



**Fig. 6- figure supplement 3. An ochrophyte-specific Tic20 fusion protein.** This figure shows alignments of the full length (i) and conserved region only (ii) of plastid Tic20 sequences, displayed as per figure supplement 9.

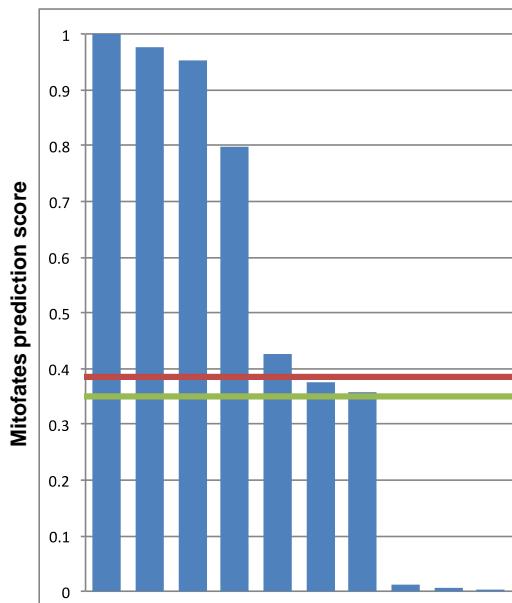


**Fig. 7-figure supplement 1. Experimental verification of additional ochrophyte dual-targeted proteins.** Panel A shows Mitotracker-orange stained *Phaeodactylum tricornutum* lines expressing four additional dual-targeted proteins (glycyl-, leucyl-, and methionyl-tRNA synthetases, and a predicted mitochondrial GroES-type chaperone) from *Phaeodactylum tricornutum*, and a dual-targeted histidyl-tRNA synthetase from *Glenodinium foliaceum*. Panel B shows control images that confirm an absence of crosstalk between GFP and mitotracker: wild-type *Phaeodactylum* cells stained with mitotracker, and cells expressing the *Glenodinium* histidyl-tRNA synthetase–GFP fusion construct and visualised with the mitotracker laser and channel in the absence of mitotracker stain.

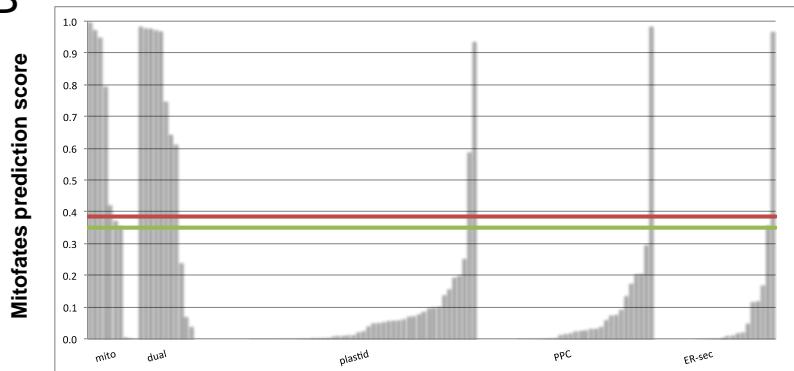


**Fig. 7- figure supplement 2. Comparison of different in silico targeting prediction programmes for the identification of dual-targeted ochrophyte proteins.** **Panel A** shows Mitofates scores for ochrophyte proteins verified experimentally to be dual targeted in this and a previous study<sup>9</sup>. **Panel B** shows Mitofates scores for all ochrophyte proteins for which a subcellular localisation has been identified in previous studies. The red lines in each graph show the Mitofates default cutoff (0.385) and the green lines indicate our chosen cutoff (0.35). **Panel C** compares different in silico targeting prediction algorithms with respect to predicted mitochondrial localization by experimentally validated localization. Mitofates strikes the best balance between high true positives and low false positives.

A



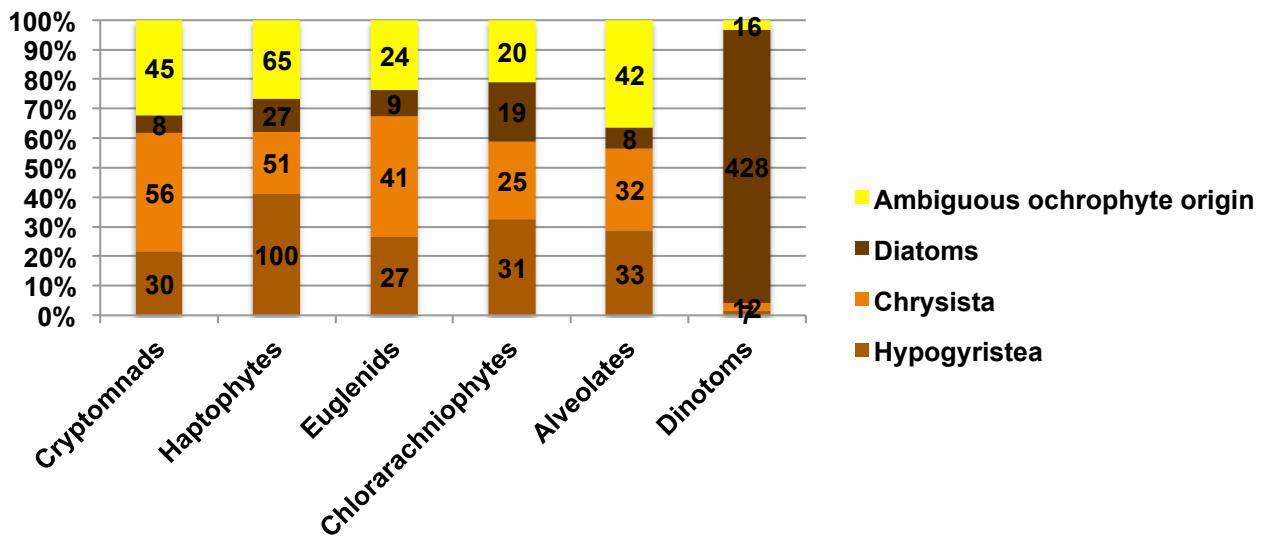
B



C

validated localization	% predicted mitochondrially targeted			
	mitofates > 0.35	mitoprot > 0.9	targetP > 0.9	targetP > 0.7
mitochondrion	70	70	40	60
mitochondrion and plastid	67	83	42	83
plastid	4	11	0	2
PPC	3	9	6	9
endomembrane or secreted	8	17	4	4

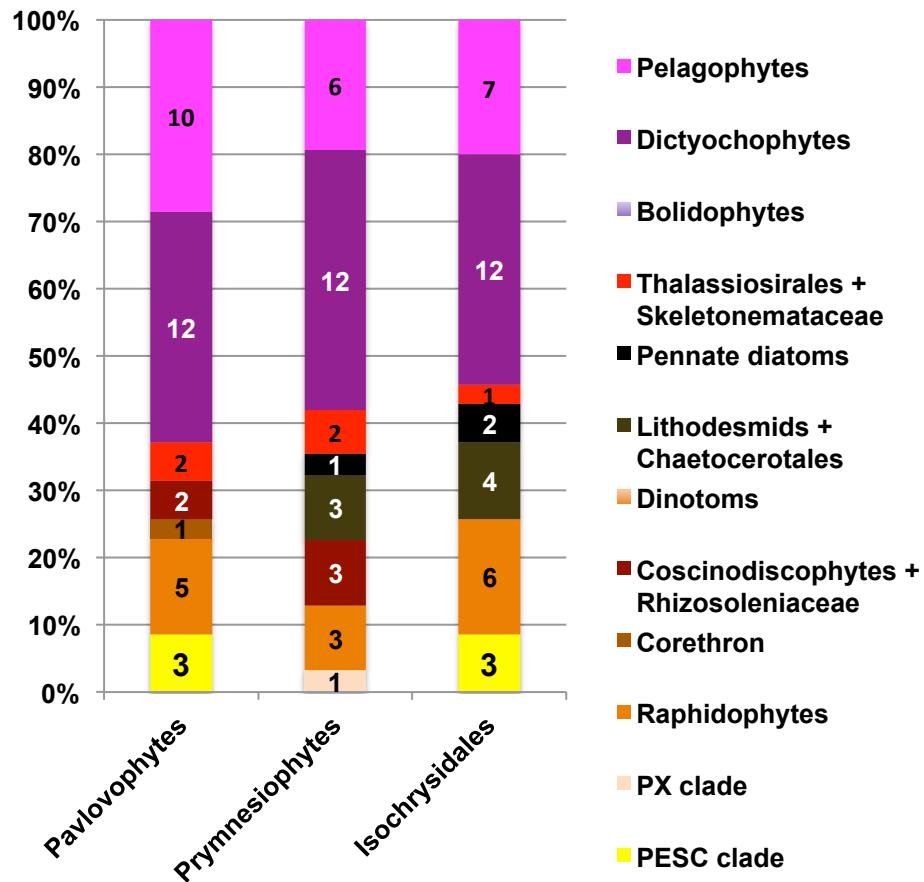
**Fig. 8- figure supplement 1. Origin of proteins of ochrophyte origin in different CASH lineages.** This figure profiles the evolutionary origins of proteins inferred by single-gene phylogenetic analysis to have been transferred from the ochrophytes into other lineages that have acquired plastids through secondary or more complex endosymbioses. Proteins are divided into the three major ochrophyte lineages (i.e. diatoms, chrysista, and hypogyristera); all remaining proteins (inferred to have been acquired from an ancestor of multiple ochrophyte lineages, or of ambiguous but clearly ochrophyte origin) are grouped as a final category. The haptophyte proteins that could be attributed to a specific ochrophyte lineage are particularly skewed (100/178 proteins) to origins within the hypogyristera.



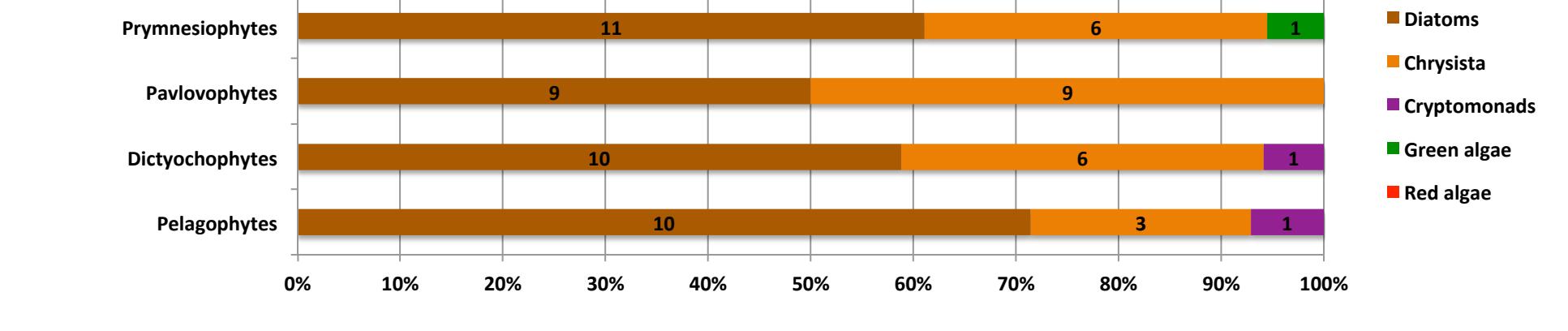
**Fig.8- figure supplement 2. Heatmaps of nearest sister-groups to haptophytes in ancestral ochrophyte HPPG trees.** This figure shows the specific ochrophyte lineages implicated in the origin of haptophyte plastid-targeted proteins, as inferred from the nearest ochrophyte sister-groups to haptophytes in trees of 242 haptophyte proteins of probable ochrophyte origin from combined BLAST top hit and single-gene tree analysis. At the top a schematic tree diagram of the ochrophytes is shown as per fig. 1, with six major nodes in ochrophyte evolution labelled with coloured boxes. The heatmap below shows the specific distribution of sister-groups in each tree, shown as per figure 4- figure supplement 2.



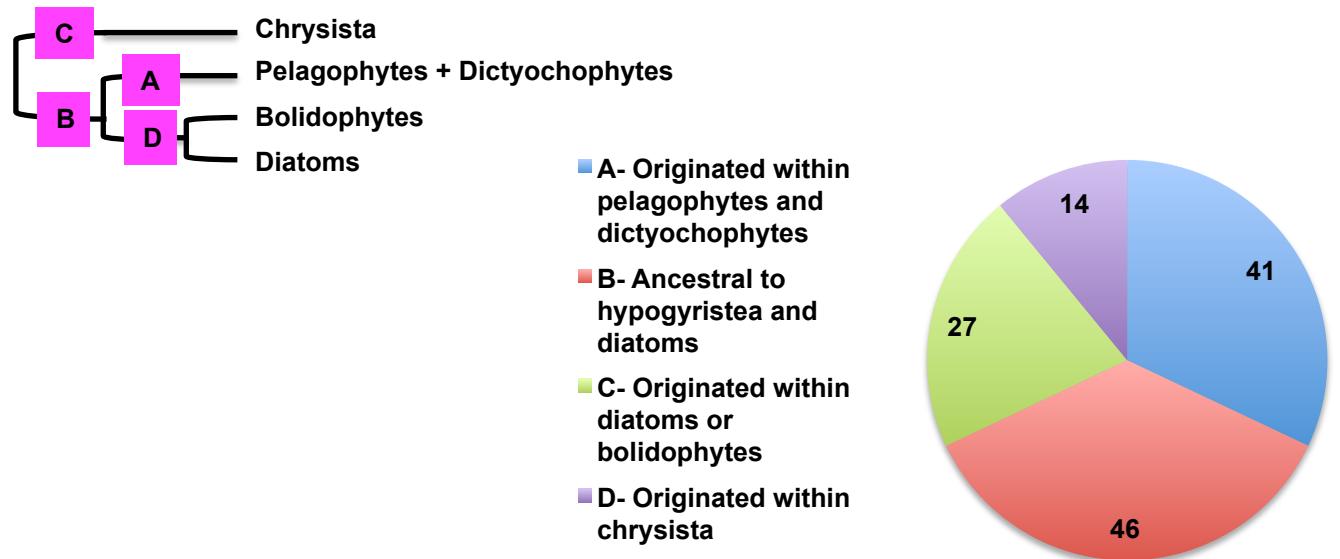
**Fig. 8- figure supplement 3. Internal evolutionary affinities of haptophyte plastid-targeted proteins incorporated into ancestral ochrophyte HPPGs.** This figure profiles the evolutionary origins of haptophyte plastid-targeted proteins incorporated into ancestral ochrophyte HPPGs by BLAST top hit analysis. Separate values are provided for query sequences from each of the three haptophyte sub-categories (pavlovophytes, prymnesiophytes, isochrysidales) considered within the analysis. Only sequences for which a consistent origin could be identified by both BLAST top hit and single-gene tree analysis are included. For each haptophyte lineage > 50% of the sequences verified by combined analysis to be of a specific ochrophyte origin have either pelagophyte or dictyochophyte top hits.



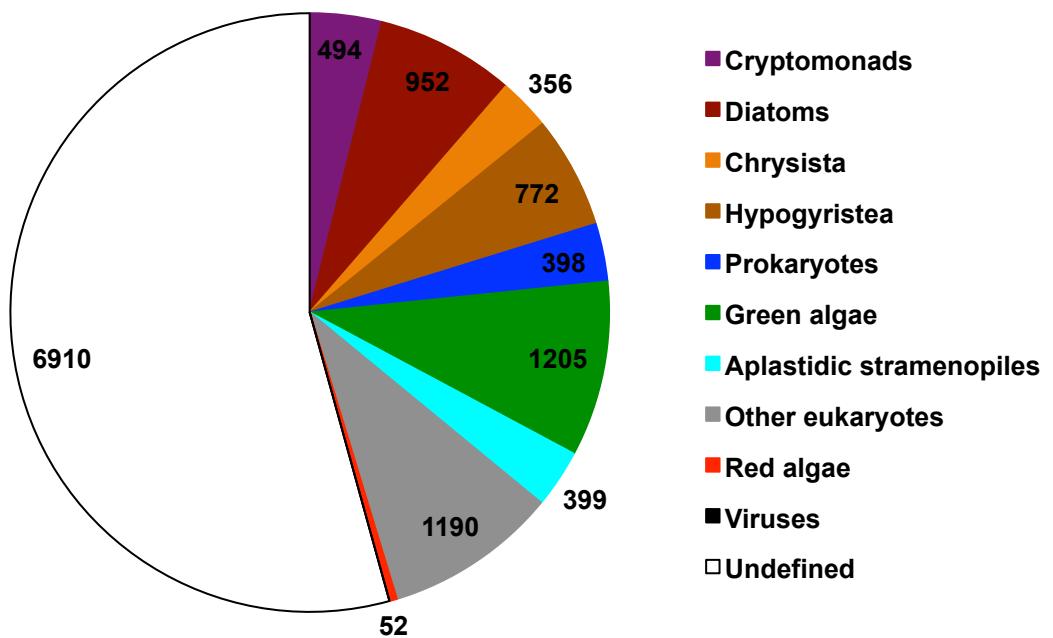
**Fig. 8-figure supplement 4. Evidence for gene transfer from pelagophytes and dictyochophytes into haptophytes.** Panel A shows the next deepest sister groups identified for haptophyte proteins of hypoglycine origin in single-gene trees. The pie chart (i) compares the number of single-gene trees in which the combined clade of haptophyte and hypoglycine proteins resolves within a larger clade comprising the ochrophyte HPPG, compared to the number that resolves in external positions, either with other lineages or as a sister-group to all other sequences within the HPPG clade. Sequences for which no clear next deepest sister group affinity could be identified are listed as “not determined”. The heatmap (ii) shows the specific sister-group sequences associated with 65 HPPGs in which the haptophyte sequences specifically resolve with the pelagophyte/ dictyochophyte clade and for which a clear internal or external position for the haptophyte/ hypoglycine group relative to the remaining ochrophyte HPPG clade could be identified. Both analyses indicate a clear bias for haptophyte sequences branching within a deeper ochrophyte clade, not just restricted to the immediate sister-groups. Panel B tabulates the BLAST next best hits for haptophyte sequences for which a phylogenetically consistent (>3 consecutive top hits) top hit to hypoglycine could be identified, and pelagophyte/ dictyochophyte sequences for which a phylogenetically consistent top hit to haptophytes could be identified. In each case either the largest number of sequences, or (in the case of pavlovophytes) the joint largest number of sequences for which a phylogenetically consistent next best hit could be identified resolved with diatoms, indicating that these sequences were probably present in a common ancestor of diatoms and hypoglycine, and subsequently transferred to haptophytes.



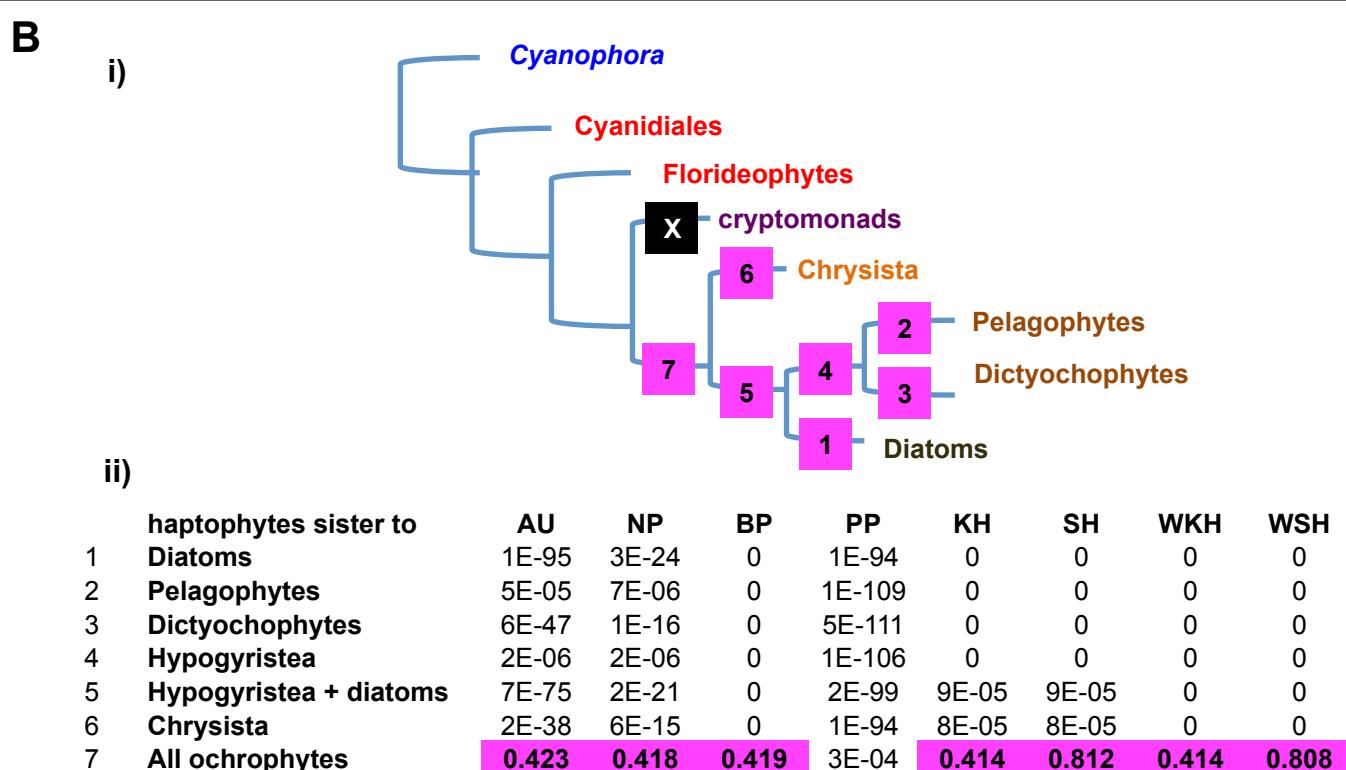
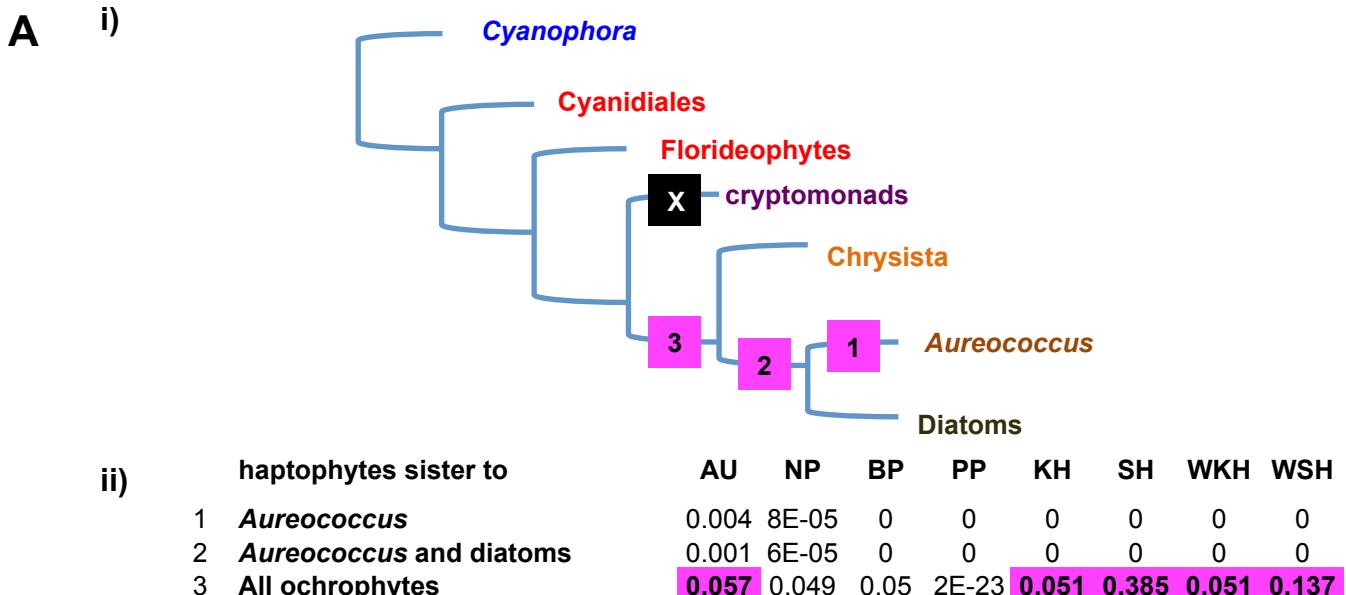
**Fig. 8- figure supplement 5. Earliest possible origin points of uniquely conserved sites in haptophyte plastid-targeted proteins.** This figure shows the total number of residues that are uniquely shared between a 37 proteins that have clearly been transferred between the ochrophytes and haptophytes, and are of subsequently entirely vertical origin, assuming the earliest possible origin point for each residue (i.e. in which gapped or missing positions were interpreted as identities). 87/ 128 of the uniquely shared residues inferred to originate within the ochrophytes were congruent to gene transfers between the haptophytes and pelagophyte and dictyochophyte clade; of these, slightly more than half (46) are inferred to have originated in a common ancestor of all hypogyristeae and diatoms, consistent with the gene transfer having occurred from an ancestor of the pelagophytes and dictyochophytes into the haptophytes, rather than the converse.



**Fig. 8-figure supplement 6. Evolutionary origin of ancestral haptophyte genes.** This figure shows the most likely evolutionary origin assigned by BLAST top hit analysis to the 12728 conserved gene families inferred to have been present in the last common haptophyte ancestor.

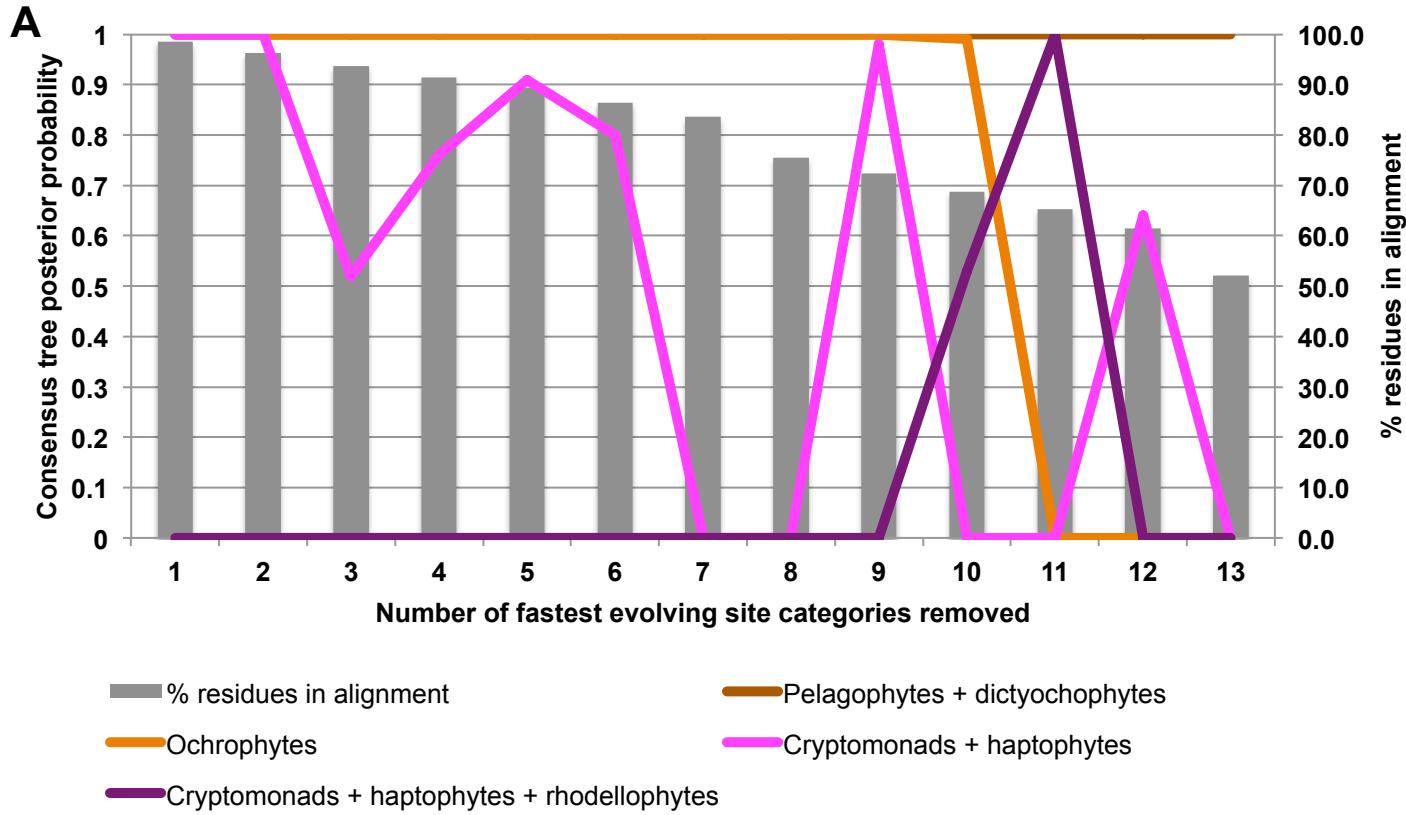


**Fig. 9- figure supplement 1. Alternative topology tests of plastid genome trees.** Tests were performed with the RAxML + JTT trees inferred for the gene-rich (**panel A**) and taxon-rich (**panel B**) plastid-encoded protein alignments. In each case, a schematic diagram of the tree topology obtained is given (**i**). The black box corresponds to the branch position of haptophytes in the consensus tree; alternative branching positions for the haptophyte sequences are labelled with numbered boxes. The table below (**ii**) lists the probabilities for each alternative position under eight different tests performed with CONSEL. Alternative positions that are not rejected by a topology test are shaded. All possible trees in which the haptophyte sequences branch within the ochrophytes are clearly rejected under all conditions, confirming that its plastid genome is of non-ochrophyte origin. The legend at the bottom of panel B gives full names for each test performed.



- AU - approximately unbiased test
- NP & BP - bootstrap probabilities for the selection
- PP - bayesian posterior probability (using BIC)
- KH - Kishino-Hasegawa test
- SH - Shimodaira-Hasegawa test
- WKh & WSH - weighted versions of the above two tests

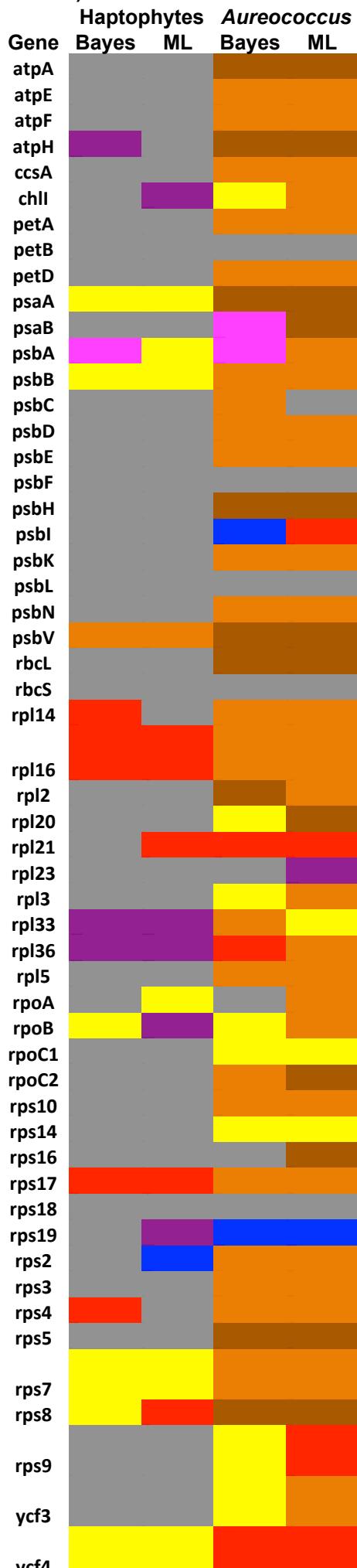
**Fig. 9-figure supplement 2. Fast site removal and clade deduction analysis of plastid genome trees.** **Panel A** shows the support values obtained for Bayesian + Jones trees inferred from modified versions of the taxon-rich plastid multigene alignment from which the 13 fastest-evolving site categories had been removed for four different branching relationships pertaining to the placements of haptophyte and hypogyristeian sequences. The % of residues from the original alignment retained in each modified alignment are shown with grey bars. **Panel B** tabulates the support obtained for two different evolutionary relationships (haptophytes as a sister group to all cryptomonads, and as a sister group to all ochrophytes) in gene-rich (i) and taxon-rich (ii) alignments modified to remove all amino acids that occur at different frequencies in haptophytes to ochrophyte lineages, and modified to remove individual or pairs of CASH lineages. “x” indicates that the topology in question was not obtained.



**B**

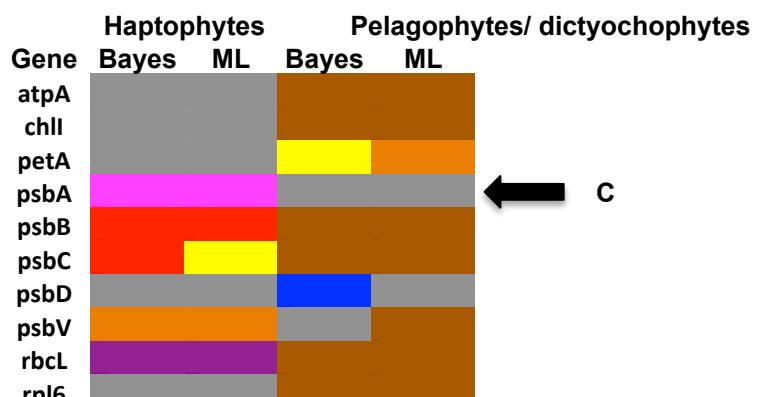
Topology	Tree	i) Gene-rich alignment									
		No glycines	No variant aa	No diatoms	No chrysista	No cryptomonads	No diatoms + chrysista	No diatoms + cryptomonads	No chrysista + cryptomonads		
cryptomonads + haptophytes	MrBayes	1	1	1	1	x	x	x	x	x	x
cryptomonads + haptophytes	RAxML	95	97	98	62	x	30	x	x	x	x
haptophytes + ochrophytes	MrBayes	x	x	x	x	1	x	1	1	1	1
haptophytes + ochrophytes	RAxML	x	x	x	x	100	x	100	100	100	100
ii) Taxon-rich alignment											
cryptomonads + haptophytes	MrBayes	1	0.84	1	1	x	x	x	x	x	x
cryptomonads + haptophytes	RAxML	35	x	x	x	x	x	x	x	x	x
haptophytes + ochrophytes	MrBayes	x	x	x	x	1	1	1	1	1	1
haptophytes + ochrophytes	RAxML	x	x	43	73	100	69	100	100	100	100

### i) Gene-rich dataset



**Fig. 9- figure supplement 3. Single-gene tree topologies associated with individual plastid-encoded genes.** These heatmaps show the first sister-groups identified to haptophytes, and members of the pelagophyte/dictyochophyte clade, in single-gene trees of component genes included in concatenated trees of plastid-encoded proteins using both the gene-rich (i) and taxon-rich (ii) alignments. Topologies are given for trees inferred with MrBayes using the Jones substitution matrix, and RAxML trees inferred using JTT, under the same conditions as the multigene trees. The identity of the first sister-group is shaded according to the legend given below. Only three single-gene trees (labelled with black arrows) support any sister-group relationship between haptophytes and the pelagophyte/dictyochophyte clade; however, in each case (explained beneath the legend) this topology is not robustly supported, either due to polyphyly of one of the constituent lineages, or conflicting topologies identified via alternative methods.

### ii) Taxon-rich dataset



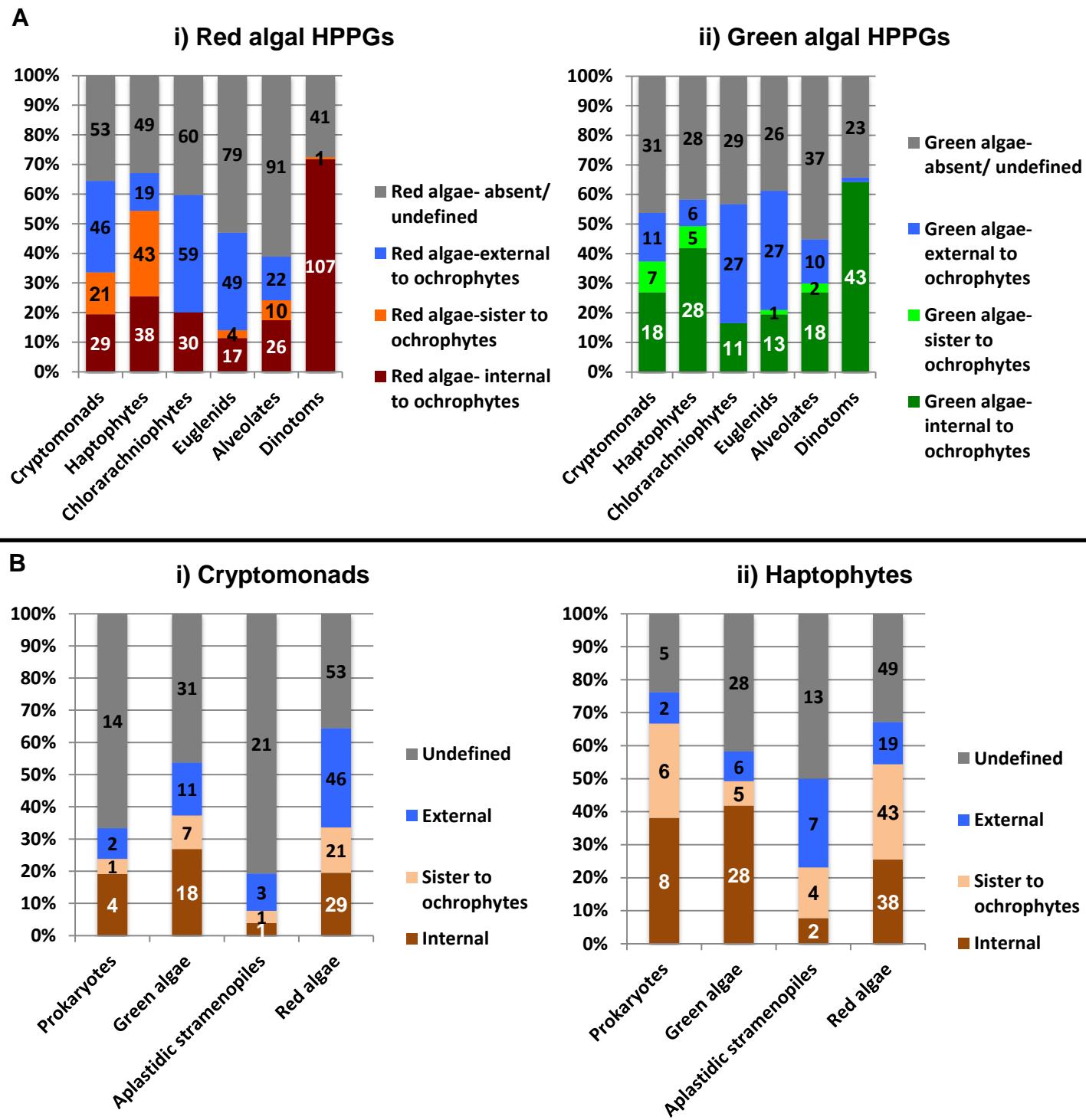
**Key**

- Diatoms
- Chrysista
- Multiple ochrophyte lineages
- Cryptomonads
- Red algae
- Glaucophytes
- Ambiguous topology
- Sister-group relationship between haptophytes and pelagophytes/ dictyochophytes

**Detailed topological explanations of labelled trees**

- A *Aureococcus* resolves with *Pavlova*, but not other haptophytes
- B *Aureococcus* and haptophytes resolve as sister-groups under Bayes only
- C Haptophytes resolve with pelagophytes, but not dictyochophytes

**Fig. 10- figure supplement 1. Complex origins of different ancestral ochrophyte HPPGs** Panel A shows the evolutionary positions of lineages with histories of secondary endosymbiosis in trees of ancestral ochrophyte HPPGs verified by combined BLAST top hit and single-gene tree analysis to be either of red algal (i) or green algal origin (ii). In both cases, in more than half of the constituent trees, haptophyte and cryptomonad sequences resolve as closer relatives to the ochrophytes than the red or green algal evolutionary outgroup, either due to resolving in the ochrophyte HPPG or forming a specific sister-group to the ochrophyte lineages. Panel B plots the distribution of cryptomonads (i) and haptophytes (ii) in trees for different categories of ancestral ochrophyte HPPG of verified evolutionary origin. HPPGs of green algal origin more frequently show internal or sister positions for the cryptomonad sequences than all other categories of HPPG, and in more than 50% of cases resolve internal or sister positions for the haptophyte sequences. This might be consistent with a green algal contribution in the endosymbiotic ancestor of cryptomonad, haptophyte and ochrophyte plastids.



**Fig. 10 –figure supplement 2. Different scenarios for the origins of haptophyte plastids.** This schematic tree diagram shows different possibilities for the origins of the haptophyte plastid as predicted from the data within this study. No inference is made here regarding the ultimate origin of the ochrophyte plastid, although it is noted that the ochrophyte, cryptomonad and haptophyte plastids are likely to be closely related to one another within the red plastid lineages. First, a common ancestor of the pelagophytes and dictyochophytes was taken up by a common ancestor of the haptophytes (point 1), yielding a permanent plastid that contributed genes for a large number of plastid-targeted proteins in extant haptophytes. This plastid was subsequently replaced via serial endosymbiosis (point 2) yielding the current haptophyte plastid and plastid genome. This serial endosymbiosis event either involved a close relative of extant cryptomonads (**2A**) or a currently unidentified species that forms a sister-group in plastid gene trees to all extant ochrophytes, but is evolutionarily distinct from the pelagophytes (**2B**). It is possible that the haptophyte plastid may have been acquired through the secondary endosymbiosis of a different lineage of red algae to the ochrophyte, either via a cryptomonad intermediate (**2C**) or directly (**2D**).

