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

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

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

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

[1]IBENS, Département de Biologie, École Normale Supérieure, CNRS, Inserm, PSL Research University, F-75005, Paris, France
[2]School of Life Sciences, Arizona State University, 427 E Tyler Mall, Tempe, AZ, 85287, USA
[3]Institut de Biologie Paris-Seine, Université Pierre et Marie Curie, Paris
[4]Department of Cell Biology, University of Alberta
[5]CNRS, UPMC, FR2424, ABiMS, Station Biologique, 29680, Roscoff, France
[6]Sorbonne Universités, UPMC Univ Paris 06, CNRS UMR 7144, Adaptation et Diversité en Milieu Marin, Equipe EPEP, Station Biologique de Roscoff, 29680 Roscoff, France

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

Abstract

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

Introduction

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

Photosynthesis has subsequently spread outside of the archaeplastids through secondary, tertiary, or more complex endosymbiosis events. By far the most ecologically successful of these lineages are those that possess plastids derived from secondary or more complex endosymbioses of a red alga[1, 4, 5]. These are the "CASH lineages", consisting of

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

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

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

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

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

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

Results

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

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

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

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

2. Experimental verification of ancestral ochrophyte HPPGs

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

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

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

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

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

3. Evolutionary origins of the ochrophyte plastid

The ochrophyte plastid is an evolutionary mosaic

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

We accordingly determined the closest relative of each ancestral HPPG (Materials and Methods). Due to ongoing controversies regarding the evolutionary composition of ochrophyte genomes[46, 47], we utilised a combined phylogenetic and BLAST top hit approach to robustly infer the most probable origin of each HPPG (Materials and Methods; Table S4- sheet 2[145]). For both the BLAST and phylogenetic analyses, stringent criteria were applied to avoid misidentification due to topological ambiguity, or contamination within individual sequence datasets[57, 58] (Materials and Methods). We took the union of these two analyses to produce a dataset of 263 HPPGs for which both phylogenetic and BLAST top hit analyses indicated the same clear evolutionary origin. These origins were grouped into six evolutionary categories, red algae, green algae, aplastidic stramenopiles, other eukaryotes, prokaryotes, and viruses (Fig. 3, panel A).

Of the 263 HPPGs that were resolved from the combined analysis, 149 (57%) were of red algal, i.e. endosymbiont origin (Fig. 3, panel A; Table S4- sheet 3[145]). This is analogous to results from studies of archaeplastid plastid proteomes, in which approximately half of the plastid-targeted proteins are of endosymbiont origin[18, 32]. The remaining 114 HPPGs resolved with other sister-groups, consistent with a mosaic origin of the ochrophyte plastid proteome. The most significant of these lineages was green algae (67 HPPGs, 25%), followed by aplastidic stramenopiles (26 HPPGs, 10%), and prokaryotes (21 HPPGs, 8%) (Fig. 3, panel A). None of the HPPGs were clearly assigned to other eukaryotes or to viruses, consistent with previous assertions that these lineages have contributed very little to ochrophyte evolution[59] (Fig. 3, panel A).

Late origin of ochrophyte plastids

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

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

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

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

A significant green algal contribution to ochrophyte plastid evolution

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

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

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

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

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

4. Functional consequences of mosaic origins for the ochrophyte plastid

Metabolic completeness of the ochrophyte plastid

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

Mosaic origins of ochrophyte plastid metabolism

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

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

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

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

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

Ancient origins of chimeric plastid-targeted proteins

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

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

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

Ancestral and bidirectional origins of dual targeting in ochrophytes

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

Previous studies have speculated that dual targeting may arise early in plastid evolution, for example through the retargeting of proteins from the host mitochondria to the plastid, or equally via the adaptation of proteins of plastid origin to the mitochondria[18, 77].

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

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

5. Complex evolutionary origins of CASH lineage plastids

A pelagophyte/ dictyochophyte origin of the haptophyte plastid proteome

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

that resolved within the ochrophytes was found to be significantly greater than any of the other groups except for dinotoms (chi-squared, $P < 1 \times 10^{-05}$; Table S4- sheet 5[145]).

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

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

We next tested the probable direction of the gene transfer events. We reasoned that if the genes identified within our study had been transferred from an ancestor of pelagophytes and dictyochophytes into the haptophytes, then we should also see a strong secondary signal linking the haptophytes to earlier ancestors of the pelagophyte/ dictyochophyte clade, for example the common ancestor of hypogyristea and diatoms. We inspected the secondary BLAST top hits associated with genes shared between haptophytes and hypogyristea (Fig. 8- figure supplement 4; Table S13- sheet 5[145]), and the next deepest sister-groups to haptophyte proteins that are of probable pelagophyte or dictyochophyte origin in each single-gene tree (Fig. 8- figure supplement 4; Table S4- sheet 2, section 6[145]). The majority of haptophyte proteins of hypogyristean 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 hypogyristean BLAST top hits, and hypogyristean 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[145]). 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[145]) 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[145]). 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[145]), of which 772 were of probable hypogyristean origin (Fig. 8- figure supplement 6; Table S14- sheet 2[145]). A significantly larger proportion of the ancestral haptophyte gene families of hypogyristean 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^{-05}; Fig. 8, panel E; Table S14- sheet 3[145]), 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.

Phylogenetic discrepancies between the haptophyte plastid proteome and genome

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

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

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

Discussion

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

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

Another major question concerns the origins of plastid-targeted proteins of green algal origin in ochrophytes. Overall, our data supports the targeting of a significant complement of proteins of chlorophyte origin to the ochrophyte plastid (Fig. 4). It remains to be determined, however, what the exact chlorophyte donor was, and how these genes may have been acquired. It is possible that the green genes were transferred into the ochrophyte lineage via lateral gene transfer, either from a range of different green algal sources or repeatedly from one lineage (for example, a semi-permanent intracellular symbiont[3]), although neither scenario would explain the bias in green algal genes in ochrophyte genomes towards encoding proteins of plastid function (Fig. 4, panel D). An alternative possibility might be a cryptic green algal endosymbiosis in the evolutionary history of the host, as has been previously suggested[1, 44] (Fig. 10), or a more convoluted pattern of acquisition. We note, for example, that the green genes identified in our study are not only plastid-targeted across the ochrophytes, but are apparently shared with haptophytes and cryptomonads (Fig. 10- figure supplement 1), which would be equally consistent with them having been present in a common ancestor of the CASH lineage plastid, and relocated to each host nuclear lineage following endosymbiosis (Fig. 10). Thus, pinpointing the exact nature and timing of the green gene transfer into ochrophytes rests not only on more extensive sequencing of deep-branching chlorophyte lineages, but also on characterising the genome composition of the closest aplastidic relatives of extant ochrophytes (e.g., *Develorapax, Pirsonia*[6]), and the closest red algal relative of CASH lineage plastids, which remains unknown[1, 4].

We also provide evidence for a chimeric origin of the haptophyte plastid (Figs. 8, 9). A schematic outline of these results is shown in Fig. 10- figure supplement 2. We have shown that a significant number of plastid-targeted proteins found in haptophytes originate from an ancestor of the pelagophytes and dictyochophytes (Fig. 8). This relationship is supported by multiple lines of evidence- i.e., uniquely shared proteins, single-gene tree topologies, BLAST top hit analysis, and analysis of synapomorphies in multigene alignments (Fig. 8 and supplements). Alongside the bias of haptophyte genes of hypogyristean origin encoding proteins of plastid function (Fig. 8- panel E), these observations argue against these genes having been acquired through multiple independent lateral gene transfer events, and instead support an endosymbiosis event. We note that other studies have shown strong evidence for gene transfers between haptophytes and individual members of the hypogyristea: for example, Stiller *et al.* have demonstrated a strong enrichment in BLAST top hits against haptophytes, from the genome of the pelagophyte *Aureococcus anophageferrens,* compared to other ochrophyte genomes[5]. We additionally note that an ancestral gene transfer from a pelagophyte/ dictyochophyte ancestor into the haptophytes is a chronologically realistic scenario: molecular clock estimates place the pelagophytes and dictyochophytes diverging between 300 and 700 million years before present[87, 93], which broadly overlaps with the molecular dates estimated for the radiation of the haptophytes in the same studies[87, 93], and precedes the first haptophyte microfossils, identified ca. 220 million years before the present[94].

Finally, we verify that the evolutionary links between haptophyte and the pelagophyte/ dictyochophyte clade in terms of plastid-targeted proteins are not supported by phylogenies of the haptophyte plastid genome (Fig. 9). Other multigene phylogenies of red lineage plastid genomes have similarly demonstrated that the haptophyte plastid genome instead resolves as a sister-lineage either to cryptomonads or to all ochrophytes[5, 38, 83, 84]. Furthermore, the structure and content of haptophyte and hypogyristean plastid genomes are dissimilar: for example, haptophyte plastids possess an *rpl36* gene that has been laterally acquired from a bacterial donor and is shared with cryptomonad plastids but absent from

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

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

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

Materials and Methods

Identification of ancestral plastid-targeted ochrophyte proteins

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

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

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

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

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

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

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

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

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

and euglenids; Table S4- sheet 1[145]). HPPGs that passed this final stage of analysis were deemed to correspond to ancestrally plastid-targeted proteins (Table S4- sheet 2[145]).

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

Generation of fluorescence expression constructs for *Phaeodactylum tricornutum*

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

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

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

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

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

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

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

Reconstruction of evolutionary origins of ancestral plastid-targeted proteins

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

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

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

Reconstruction of evolutionary relationships between ochrophytes and other CASH lineage plastids

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

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

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

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

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

Identification of uniquely shared residues in multigene HPPG datasets

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

- For the green gene dataset, HPPG trees containing more than one clade of ochrophyte, cryptomonad, haptophyte, red algal, or green algal sequences were excluded. To account for the possibility that CASH lineage sequences might originate from within the green algae, the green algae were allowed to be paraphyletic with regard to the cryptomonad, haptophyte and ochrophyte sequences, but were not allowed to incorporate sequences from other lineages. Similarly, to account for the possibility that subsequent gene transfers may have occurred from ochrophytes into other CASH lineages, the ochrophytes were allowed to be paraphyletic with regard to cryptomonad and haptophyte sequences, but not to any other lineages.
- For the haptophyte gene dataset, HPPG trees containing more than one clade of ochrophyte, haptophyte, diatom, hypogyristean, or chrysistan sequences were excluded. To account for the possibility that haptophytes arose within the ochrophytes, the ochrophyte, diatom, hypogyristean and chrysistan sequences were allowed to incorporate sequences from haptophytes. Similarly, due to the paraphyly of hypogyristea with regard to diatoms, the hypogyristean sequences were allowed to incorporate sequences from diatoms, but not from other lineages.
- In all cases, sequences from chlorarachniophytes, euglenids, and alveolates were not incorporated into any of the clade assessments, due to uncertainty over the gene transfer events that have occurred in each lineage[7, 81, 82].

This left datasets consisting of 32 HPPGs for which the ochrophytes were of clear green algal origin, and 37 HPPGs in which the haptophytes were of clear ochrophyte origin, with no conflicting phylogenetic signal. The rationale for inclusion and exclusion of each HPPG in each analysis is presented in Table S6, sheets 1 and 3[145].

Next, to eliminate individual sequences remaining within each HPPG that might have arisen through species-specific gene transfer or contamination events, each trimmed sequence within each approved alignment was inspected using a composite BLAST approach. First, each sequence was searched against a composite library containing all uniref, jgi and MMETSP sequences from every lineage within the tree of life, and the top ten hits were tabulated for each sequence. In each case, only sequences for which at least the first three hits were of the same lineage as that of the query were retained. For the haptophyte multigene alignment, the ochrophytes were separately analysed as each of the three component lineages (chrysista, hypogyristea, and diatoms), which is to say that a query obtained from a member of the hypogyristea would only be retained if the first three BLAST top hits originated from other hypogyristean sequences, rather than other ochrophytes.

Next, each of the component sequences within each cleaned alignment were searched against all other component sequences within the same alignment using BLASTp, and the top ten hits within the alignment were ranked. In each case, sequences were only approved for incorporation into the multigene dataset if the first non-self hit was to a different sub-category within the same lineage, e.g. if a query sequence from a red alga yielded a top hit against a red algal sequence from a different red sub-category. To allow for possible cases of paraphyly and/or absence of sequences within each alignment, the following modifications were applied:

- Green algal sequences within the confirmed green origin alignments were allowed to yield top hits against ochrophytes, cryptomonads, and haptophytes, but were required to yield a best hit against another green alga with an expect value lower 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 against cyanobacteria, red algae, or green algae, due to the incorporation (in general) of only one glaucophyte sequence in each alignment.
- Ochrophyte sequences were deemed to be of correct origin if they yielded a top hit against any other ochrophyte sub-category (regardless of whether this was of diatom, hypogyristean or chrysistan origin). Ochrophyte sequences were additionally allowed to yield top hits against cryptomonads (in the green gene alignments), and haptophytes (in both green and haptophyte gene alignments), but were required to yield a best hit against another ochrophyte with an expect value 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 same lineage, but for which at least one top hit were found within the same sub-category within the lineage, and for which the first ten BLAST hits did not directly indicate a contamination event, were deemed to be of correct origin.

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

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

Analysis of targeting preferences of ancestral ochrophyte and haptophyte genes.

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

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

Functional and physiological annotation of ancestral plastid-targeted proteins

Core plastid metabolism pathways were identified using recent reviews of ochrophyte metabolism, or reviews of homologous plant plastid metabolic pathways where ochrophyte-specific reviews have not yet been published[51, 97, 115, 130-136]. The probable function and KOG classification of each HPPG were annotated using the pre-existing annotations associated with seed protein sequence (if these existed), or if not the annotated function of the top uniref hit previously identified by BLAST searches of the seed sequence (Table S9[145]). Expression dynamics for each ancestral HPPG within the genomes of the model diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* were inferred using microarray data integrated into the DiatomPortal server[74] (Table S10- sheets 1,2[145]). Correlation coefficients were calculated between each pair of *P. tricornutum* and *T. pseudonana* genes that were incorporated into an ancestral HPPG, across all microarray libraries within the dataset (Table S10- sheets 3,4[145]), with average values being calculated from all pairwise correlations for different evolutionary categories of protein (Table S10- sheet 5[145]).

Possible chimeric proteins, resulting from the fusion of proteins of different evolutionary origins, were identified in the dataset using a modified version of a previously published protocol[75] (Table S9- sheet 1, sections 4,5; Table S11[145]). Each protein within each HPPG was searched using BLASTp against the composite outgroup MMETSP-enriched library, using the same taxonomic classification used for the identification of the evolutionary origin of each seed protein within the dataset, and all hits with an expect value of 1×10^{-05}. Component sequences were then grouped into component families according to the following rule: if two component sequences overlapped by more than 70% of their lengths on the protein composite, they belonged to the same component family. Overlapping and/ or nested component families were additionally merged if one family was included by more than 70% of its length into the other one. Component families were then assigned a broad evolutionary origin corresponding to their taxonomic composition. If the three best component sequences, according to their BLAST bitscore against the composite gene, matched with the same lineage (e.g., green algae, red algae, aplastidic stramenopiles, or other eukaryotes), the component was considered to have originated from that lineage.

Possible dual targeted proteins were identified within the dataset by screening all possible plastid-targeted proteins with Mitofates, using a cut-off targeting threshold of 0.35[137], which was inferred to be more effective in identifying experimentally verified ochrophyte mitochondria-targeted proteins (Fig. 7- figure supplement 2)[29] than other threshold values or targeting prediction programmes such as TargetP[138] or Mitoprot[139]. The default Mitofates positive cutoff value was modified from 0.38 to 0.35 in order to maximise the capture of experimentally localised mitochondrial proteins, without admitting proteins with unambiguous plastid localisation (Fig. 7- figure supplement 2). As dual targeting to plastids and mitochondria may be achieved either by distinct protein isoforms resulting from ambiguous targeting peptides or alternative internal translation initiation sites that allow production of mitochondrial targeting sequences[77, 80], each protein was screened with Mitofates using both the full-length N-termini, and N-termini predicted to result from the next downstream methionine within 30 residues. Possible conserved dual targeted proteins were then identified via the same BLAST-based assembly pipeline and stringency thresholds used to identify probable ancestral HPPGs (Table S12- sheet 1[145]). All putative dual targeted proteins have been made publically accessible through the University of Cambridge dSpace server (https://www.repository.cam.ac.uk/handle/1810/261421) [145].

Construction and inspection of concatenated and exemplar phylogenetic trees

For the plastid genome phylogenetic analysis, single-gene alignments were constructed by BLAST searches of published red lineage and glaucophyte plastid genomes (for the gene rich analysis) or of these genomes plus all MMETSP libraries for the same lineages (for the taxon rich analysis), using the *Phaeodactylum tricornutum* protein sequence as query and a threshold e-value of 1×10^{-05}, followed by alignment using GeneIOUS v 4.76[119], as before. The gene rich analysis included protein sequences from 54 genes that were identified in 22 different non-green lineage plastid genomes while the taxon-rich analysis included 10 different plastid genes that were identified in all 22 plastid genomes and at least 30 different MMETSP libraries[34] (Table S15- sheet 1[145]). For the taxon-rich analysis, only species that were represented in ≥ 6/12 of the single-gene alignments were included in the concatenated alignment. Each concatenated alignment was trimmed using trimal[120] using the -gt 0.8 option.

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

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

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

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

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

Data deposition

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

Acknowledgments

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

Competing interests.

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

References

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

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

40. Cavalier-Smith, T., Chao, E.E. & Lewis, R. Multiple origins of Heliozoa from flagellate ancestors: New cryptist subphylum Corbihelia, superclass Corbistoma, and monophyly of Haptista, Cryptista, Hacrobia and Chromista. *Mol Phylogenet Evol* **93**, 331-362 (2015).
41. Yabuki, A. *et al. Palpitomonas bilix* represents a basal cryptist lineage: insight into the character evolution in Cryptista. *Sci Rep* **4**, 4641 (2014).
42. Frommolt, R. *et al.* Ancient recruitment by chromists of green algal genes encoding enzymes for carotenoid biosynthesis. *Mol Biol Evol* **25**, 2653-2667 (2008).
43. Petersen, J., Teich, R., Brinkmann, H. & Cerff, R. A "green" phosphoribulokinase in complex algae with red plastids: evidence for a single secondary endosymbiosis leading to haptophytes, cryptophytes, heterokonts, and dinoflagellates. *J Mol Evol* **62**, 143-U142 (2006).
44. Moustafa, A. *et al.* Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* **324**, 1724-1726 (2009).
45. Ku, C. *et al.* Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* **524**, 427-432 (2015).
46. Woehle, C., Dagan, T., Martin, W.F. & Gould, S.B. Red and problematic green phylogenetic signals among thousands of nuclear genes from the photosynthetic and apicomplexa-related *Chromera velia*. *Genom Biol Evol* **3**, 1220-1230 (2011).
47. Deschamps, P. & Moreira, D. Re-evaluating the green contribution to diatom genomes. *Genom Biol Evol* **4,** 683-688 (2012).
48. Matsuzaki, M. *et al.* Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature* **428**, 653-657 (2004).
49. Collén, J. *et al.* Genome structure and metabolic features in the red seaweed *Chondrus crispus* shed light on evolution of the Archaeplastida. *Proc Natl Acad Sci USA* **110**, 5247-5252 (2013).
50. Qiu, H., Price, D.C., Yang, E.C., Yoon, H.S. & Bhattacharya, D. Evidence of ancient genome reduction in red algae (Rhodophyta). *J Phycol* **51**, 624-636 (2015).
51. Smith, S.R., Abbriano, R.M. & Hildebrand, M. Comparative analysis of diatom genomes reveals substantial differences in the organization of carbon partitioning pathways. *Algal Res* **1**, 2-16 (2012).
52. Wolfe-Simon, F., Starovoytov, V., Reinfelder, J.R., Schofield, O. & Falkowski, P.G. Localization and role of manganese superoxide dismutase in a marine diatom. *Plant Physiol* **142**, 1701-1709 (2006).
53. Gillard, J. *et al.* Physiological and transcriptomic evidence for a close coupling between chloroplast ontogeny and cell cycle progression in the pennate diatom *Seminavis robusta*. *Plant Physiol* **148**, 1394-1411 (2008).
54. Ramirez, R.A., Espinoza, B. & Kwok, E.Y. Identification of two novel type 1 peroxisomal targeting signals in *Arabidopsis thaliana*. *Acta Histochem* **116**, 1307-1312 (2014).
55. Tanaka, A. *et al.* Ultrastructure and membrane traffic during cell division in the marine pennate diatom *Phaeodactylum tricornutum*. *Protist* **166**, 506-521 (2015).
56. Imanian, B., Pombert, J.F. & Keeling, P.J. The complete plastid genomes of the two 'Dinotoms' *Durinskia baltica* and *Kryptoperidinium foliaceum*. *PLoS One* **5,** 10711 (2010).
57. Marron, A.O. *et al.* The evolution of silicon transport in eukaryotes. *Mol Biol Evol* **33**, 3226-3248 (2016).
58. Dorrell, R.G. *et al.* Progressive and biased divergent evolution underpins the origin and diversification of peridinin dinoflagellate plastids. *Mol Biol Evol* **34**, 361-379 (2017).

59. Bowler, C. *et al.* The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**, 239-244 (2008).
60. Stiller, J.W., Huang, J.L., Ding, Q., Tian, J. & Goodwillie, C. Are algal genes in nonphotosynthetic protists evidence of historical plastid endosymbioses? *BMC Genom* **10**, 484 (2009).
61. Nakamura, Y. *et al.* The first symbiont-free genome sequence of marine red alga, Susabi-nori (*Pyropia yezoensis*). *PLoS One* **8**, 57122 (2013).
62. Bhattacharya, D. *et al.* Genome of the red alga *Porphyridium purpureum*. *Nat Comms* **4** (2013).
63. Schönknecht, G. *et al.* Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science* **339**, 1207-1210 (2013).
64. Matasci, N. *et al.* Data access for the 1,000 Plants (1KP) project. *Gigascience* **3**, 17 (2014).
65. Keeling, P.J. & Palmer, J.D. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet* **9**, 605-618 (2008).
66. Doolittle, W.E. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet* **14**, 307-311 (1998).
67. Ershov, Y., Gantt, R.R., Cunningham, F.X. & Gantt, E. Isopentenyl diphosphate isomerase deficiency in *Synechocystis* sp. strain PCC6803. *FEBS Lett* **473**, 337-340 (2000).
68. Rohdich, F. *et al.* Studies on the nonmevalonate terpene biosynthetic pathway: metabolic role of IspH (LytB) protein. *Proc Natl Acad Sci USA* **99**, 1158-1163 (2002).
69. Gutierrez-Marcos, J.F., Roberts, M.A., Campbell, E.I. & Wray, J.L. Three members of a novel small gene-family from *Arabidopsis thaliana* able to complement functionally an *Escherichia coli* mutant defective in PAPS reductase activity encode proteins with a thioredoxin-like domain and ''APS reductase'' activity. *Proc Natl Acad Sci USA* **93**, 13377-13382 (1996).
70. Dittami, S.M., Michel, G., Collén, J., Boyen, C. & Tonon, T. Chlorophyll-binding proteins revisited--a multigenic family of light-harvesting and stress proteins from a brown algal perspective. *BMC Evol Biol* **10**, 365 (2010).
71. Coesel, S., Obornik, M., Varela, J., Falciatore, A. & Bowler, C. Evolutionary origins and functions of the carotenoid biosynthetic pathway in marine diatoms. *Plos One* **3**, 2896 (2008).
72. Chan, C.X. *et al.* Red and green algal monophyly and extensive gene sharing found in a rich repertoire of red algal genes. *Curr Biol* **21**, 328-333 (2011).
73. Yurchenko, T., Sevcikova, T., Strnad, H., Butenko, A. & Elias, M. The plastid genome of some eustigmatophyte algae harbours a bacteria-derived six-gene cluster for biosynthesis of a novel secondary metabolite. *Open Biol* **6**, 11 (2016).
74. Ashworth, J., Turkarslan, S., Harris, M., Orellana, M.V. & Baliga, N.S. Pan-transcriptomic analysis identifies coordinated and orthologous functional modules in the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*. *Mar Genom* **26**, 21-28 (2016).
75. Méheust, R., Zelzion, E., Bhattacharya, D., Lopez, P. & Bapteste, E. Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. *Proc Natl Acad Sci USA* **113**, 3579-3584 (2016).
76. Herz, S., Eberhardt, S. & Bacher, A. Biosynthesis of riboflavin in plants. The *ribA* gene of *Arabidopsis thaliana* specifies a bifunctional GTP cyclohydrolase II/3,4-dihydroxy-2-butanone 4-phosphate synthase. *Phytochem* **53**, 723-731 (2000).
77. Xu, L., Carrie, C., Law, S.R., Murcha, M.W. & Whelan, J. Acquisition, conservation, and loss of dual-targeted proteins in land plants. *Plant Physiol* **161**, 644-662 (2013).

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

96. de Vries, J. & Gould, S.B. The monoplastidic bottleneck in algae and plant evolution. *bioRXiv* (2017).
97. Green, B.R. Chloroplast genomes of photosynthetic eukaryotes. *Plant J* **66**, 34-44 (2011).
98. Worden, A.Z. *et al.* Global distribution of a wild alga revealed by targeted metagenomics. *Curr Biol* **22**, R675-677 (2012).
99. Ong, H.C. *et al.* Analyses of the complete chloroplast genomes of two members of the pelagophyceae: *Aureococcus anophageferrens* CCMP1984 and *Aureoumbra lagunensis* CCMP1507. *J Phycol* **46**, 602-615 (2010).
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).
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).
102. Lommer, M. *et al.* Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genom Biol* **13** (2012).
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).
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).
105. Wang, D.M. *et al. Nannochloropsis* genomes reveal evolution of microalgal oleaginous traits. *PLoS Genet* **10**, 1004094 (2014).
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).
107. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucl Acids Res* **44**, W3-W10 (2016).
108. Suzek, B.E., Huang, H.Z., McGarvey, P., Mazumder, R. & Wu, C.H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformat* **23**, 1282-1288 (2007).
109. Mangot, J.F. *et al.* Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Scient Reports* **7** (2017).
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).
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).
112. Yang, E.C. *et al.* Supermatrix data highlight the phylogenetic relationships of photosynthetic stramenopiles. *Protist* **163**, 217-231 (2012).
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 Phylogenet Evol* **89**, 28-36 (2015).
114. Huesgen, P.F. *et al.* Proteomic amino-termini profiling reveals targeting information for protein import into complex plastids. *PLoS One* **8**, 74483 (2013).
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).
116. Read, B.A. *et al.* Pan genome of the phytoplankton *Emiliania* underpins its global distribution. *Nature* **499**, 209-213 (2013).

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

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

Table 1- Glossary Box

A schematic figure of eukaryotic taxonomy, showing the evolutionary origins of nuclear and plastid lineages, adapted from previous reviews[3], is shown in Fig. 1- figure supplement 1.

Complex plastids	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).
CASH lineages	The four major lineages of algae with plastids of secondary or higher red origin, that is to say Cryptomonads, Alveolates (dinoflagellates, and apicomplexans), Stramenopiles, and Haptophytes.
Stramenopiles	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 *Blastocystis*), which form the earliest-diverging branches. It is debated when within stramenopile evolution the extant ochrophyte plastid was acquired.
Ochrophytes	Photosynthetic and plastid-bearing members of the stramenopiles, including many ecologically important lineages (diatoms, kelps, pelagophytes) and potential model lineages for biofuels research (*Nannochloropsis*). Ochrophytes form the most significant component of eukaryotic marine phytoplankton[1,2].
Haptophytes	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[1,94], which if deposited on the ocean floor go on to form a major component of limestone and other sedimentary rocks.
HPPG	"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.

Figure Legends

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

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

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

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

Fig. 4. Verification and origins of the green signal in ochrophyte plastids. Panel A shows a schematic tree of the 11 archaeplastid sub-categories with which each green HPPG alignment was enriched prior to phylogenetic analysis. The topology of the red and green algae are shown according to previously published phylogenies[129, 144]. Green sub-categories are in green text; red algal sub-categories in red text; and other sub-categories are in blue text. Five ancestral positions within the green algal tree inspected in subsequent analyses are labelled with coloured boxes. **Panel B** shows the number of HPPGs of verified red (red bars) or green origin (green bars) for which orthologues were identified in different numbers green sub-categories (plotted on the x-axis) and red sub-categories (plotted on the z-axis). An equivalent graph showing only HPPGs for which a glaucophyte orthologue was detected is shown in figure supplement 1. **Panel C** compares the number of trees in which HPPGs of verified green origin resolve as a sister group to all green lineages (including chlorophytes and streptophytes); to multiple chlorophyte sub-categories but to the exclusion of streptophytes; and to individual chlorophyte sub-categories only. A detailed heatmap of the evolutionary distribution of the green sub-categories detected in each sister-group is shown in figure supplement 2, and the distribution of BLAST top hits within each sub-category is shown in figure supplement 3. **Panel D** lists the number of residues inferred from a dataset of 32 ochrophyte HPPGs of verified green origin, which have been subsequently entirely vertically inherited in all major photosynthetic eukaryotic lineages, to be uniquely shared between ochrophytes and some but not all green lineages, hence might represent specific synapomorphic residues. Residues are categorized by inferred origin point within the tree topology shown in panel A, i.e., each of the five ancestral nodes labelled. A final category shows all of the residues inferred to be specifically shared with one green sub-category, and not with any other. The distribution of residues based on the earliest possible origin point (taking into account gapped and missing residues in each HPPG alignment) is shown in figure supplement 4. **Panel E** shows the number of the 7140 conserved gene families inferred to have been present in the last common ochrophyte ancestor that are predicted by ASAFind 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. Evolutionary categories of proteins found to be significantly more likely (chi-squared test,

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

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

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

Fig. 7. Ancient and bidirectional connections between the ochrophyte plastid and mitochondria. Panel A shows Mitotracker-Orange stained *P. tricornutum* lines expressing GFP fusion constructs for the N-terminal regions of histidyl- and prolyl-tRNA synthetase sequences from *P. tricornutum* and the eustigmatophyte *Nannochloropsis gaditana.* Targeting constructs for an additional four dual targeted proteins in *P. tricornutum* and one dual targeted protein in *G. foliaceum*, alongside Mitotracker-negative and wild type control images, are shown in figure supplement 1. **Panel B** profiles the predicted evolutionary origins of the 34 ancestral dual targeted HPPGs, as inferred by BLAST top hit and single-gene tree analysis. Data supporting the thresholds used to identify probable dual targeted HPPGs *in silico* are supplied in figure supplement 2. **Panel C** shows seven classes of tRNA synthetase for which only two copies were inferred in the genome of the last common ochrophyte ancestor. Evolutionary origins are inferred from combined BLAST top hit and single-gene tree analysis for dual targeted proteins, and from BLAST top hit analysis alone for cytoplasmic proteins. In five cases the dual targeted isoform is inferred to be of ultimate red algal origin, indicating that a protein derived from the endosymbiont has functionally replaced the endogenous host mitochondria-targeted copy.

Fig. 8. Footprints of an ancient endosymbiosis in the haptophyte plastid proteome. Panel A indicates the number of ancestral ochrophyte HPPGs that included sequences from other algal lineages in single-gene tree analyses, and whether those algal lineages branched within or external to ochrophytes. An overview of the specific origins of proteins of ochrophyte origin in each lineage is shown in figure supplement 1. **Panel B** compares the number of ASAFind-derived HPPGs that are uniquely shared between hypogyristea (**i**) or haptophytes **(ii**) and one other CASH lineage. Values are given for proteins found in a majority of sub-categories in hypogyristea/ haptophytes and at least one sub-category from only one other lineage (light bars), and proteins found in a majority of sub-categories in hypogyristea/ haptophytes and a majority of sub-categories from only one other lineage (dark bars). Values that are significantly greater than would be expected through random distribution are labelled with black arrows. **Panel C** shows a schematic ochrophyte tree, with six different ancestral nodes within this tree labelled with coloured boxes, and the most probable origin point for each of the 243 haptophyte plastid-targeted proteins of probable ochrophyte origin within this tree, as inferred by inspection of the nearest ochrophyte sister-group in single-gene trees. A detailed heatmap of the ochrophyte sub-categories contained in each lineage is shown in figure supplement 2, and BLAST top hit analyses corresponding to each plastid-targeted protein are shown in figure supplement 3. **Panel D** shows the number of residues that are uniquely shared between haptophytes and each node of the ochrophyte tree for 37 genes in which there has been a clear transfer from ochrophytes to haptophytes, and entirely vertical subsequent inheritance. A similar graph, showing the earliest possible inferred origin of each uniquely shared residue, is shown in figure supplement 4. **Panel E** shows the number of the 12728 conserved gene families inferred to have been present in the last common haptophyte ancestor that are predicted by ASAFind 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. Evolutionary categories of proteins found to be significantly more likely (chi-squared test, P=0.05) to encode plastid-targeted proteins than would be expected by random distribution are labelled with black arrows. The evolutionary origins of the ancestral gene families are shown in figure supplement 5.

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

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

Supporting figure and dataset legends.

Fig. 1- figure supplement 1. Overview of eukaryotic diversity. This figure, adapted from a previous review[3], profiles the diversity of different eukaryotic nuclear lineages. Each grey ellipse corresponds to one major clade, or “supergroup” 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.

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 sequence 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

Fig. 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 aplastidic stramenopile origin, is labelled with a cyan ellipse.

Fig. 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 localization 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.

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 archaeplastid 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 subcategory 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 a 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 probably 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.

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.

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 hypogyristea 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 hypogyristea (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

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 1.

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 study9. **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.

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 hypogyristea); 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 hypogyristea.

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, and 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 hypogyristean origin in single-gene trees. The pie chart (i) compares the number of single-gene trees in which the combined clade of haptophyte and hypogyristean 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/ hypogyristean 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 hypogyristea 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 the common ancestor of diatoms and hypogyristea, and subsequently transferred to the 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 hypogyristea 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.

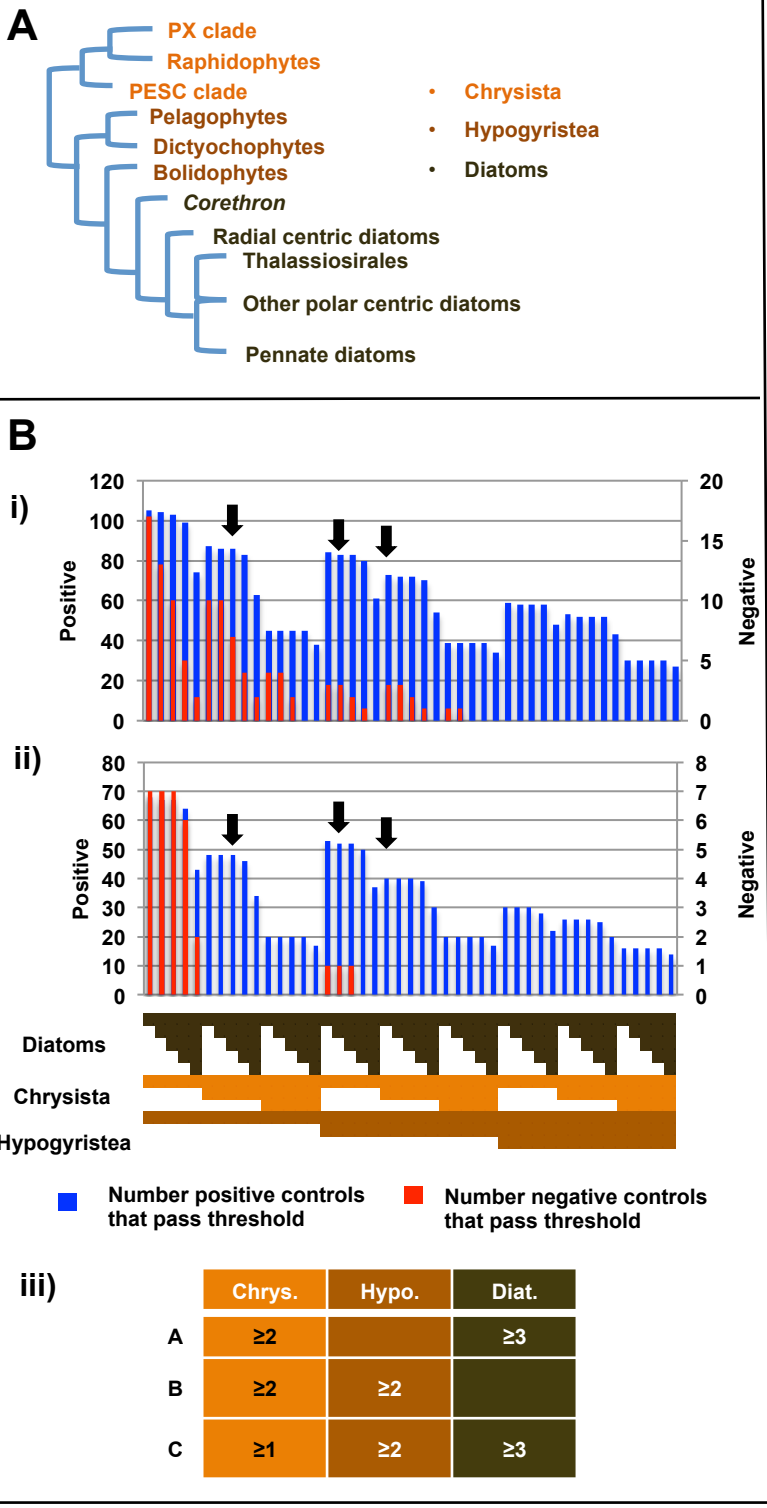
Fig. 9- figure supplement 2. Fast site removal and clade deduction analysis of plastid genome trees. Panel A shows the support values obtained for Bayesian + Jones trees inferred from modified versions of the taxon-rich plastid multigene alignment from which the 13 fastest evolving site categories had been removed for four different branching relationships pertaining to the placements of haptophyte and hypogyristean 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.

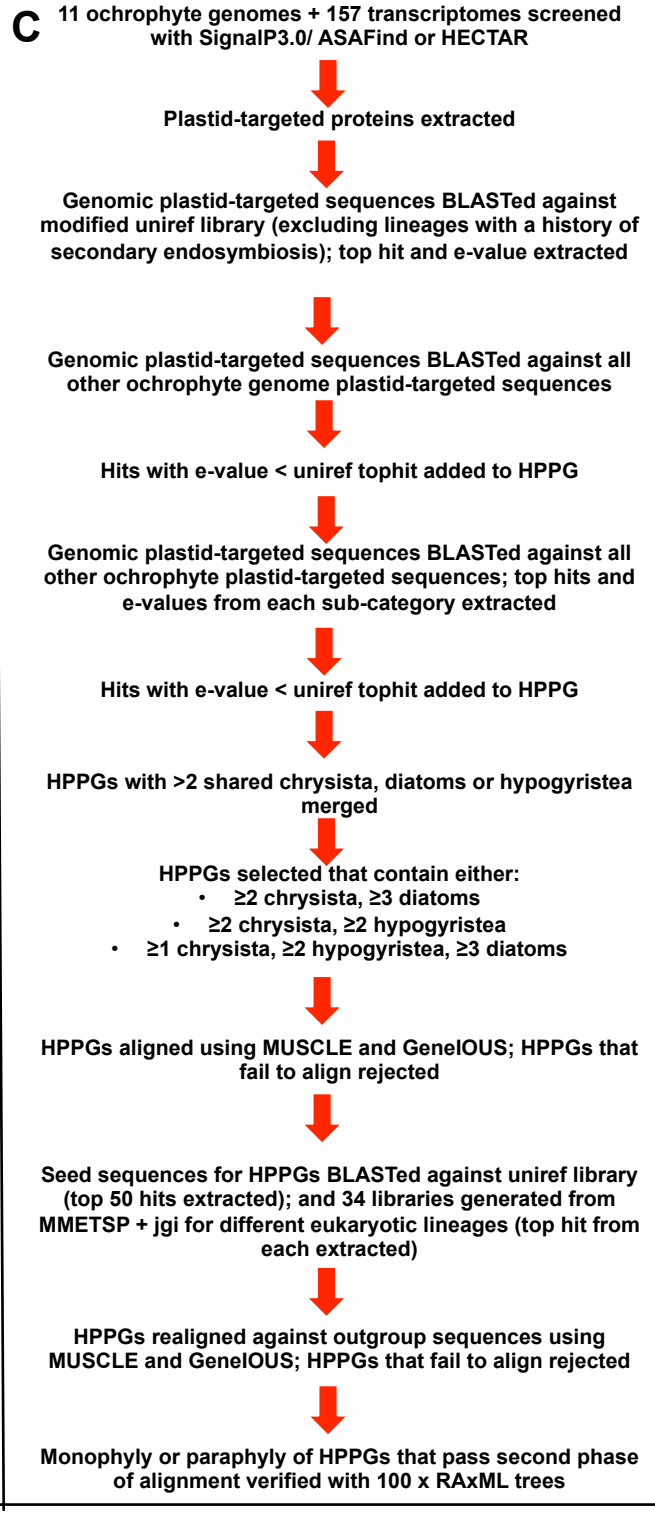
Fig. 9- figure supplement 3. Single-gene tree topologies associated with individual plastid-encoded genes. These heatmaps show the first sister-groups identified to haptophytes, and members of the pelagophyte/ dictyochophyte clade, in single-gene trees of component genes included in concatenated trees 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.

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**).





D

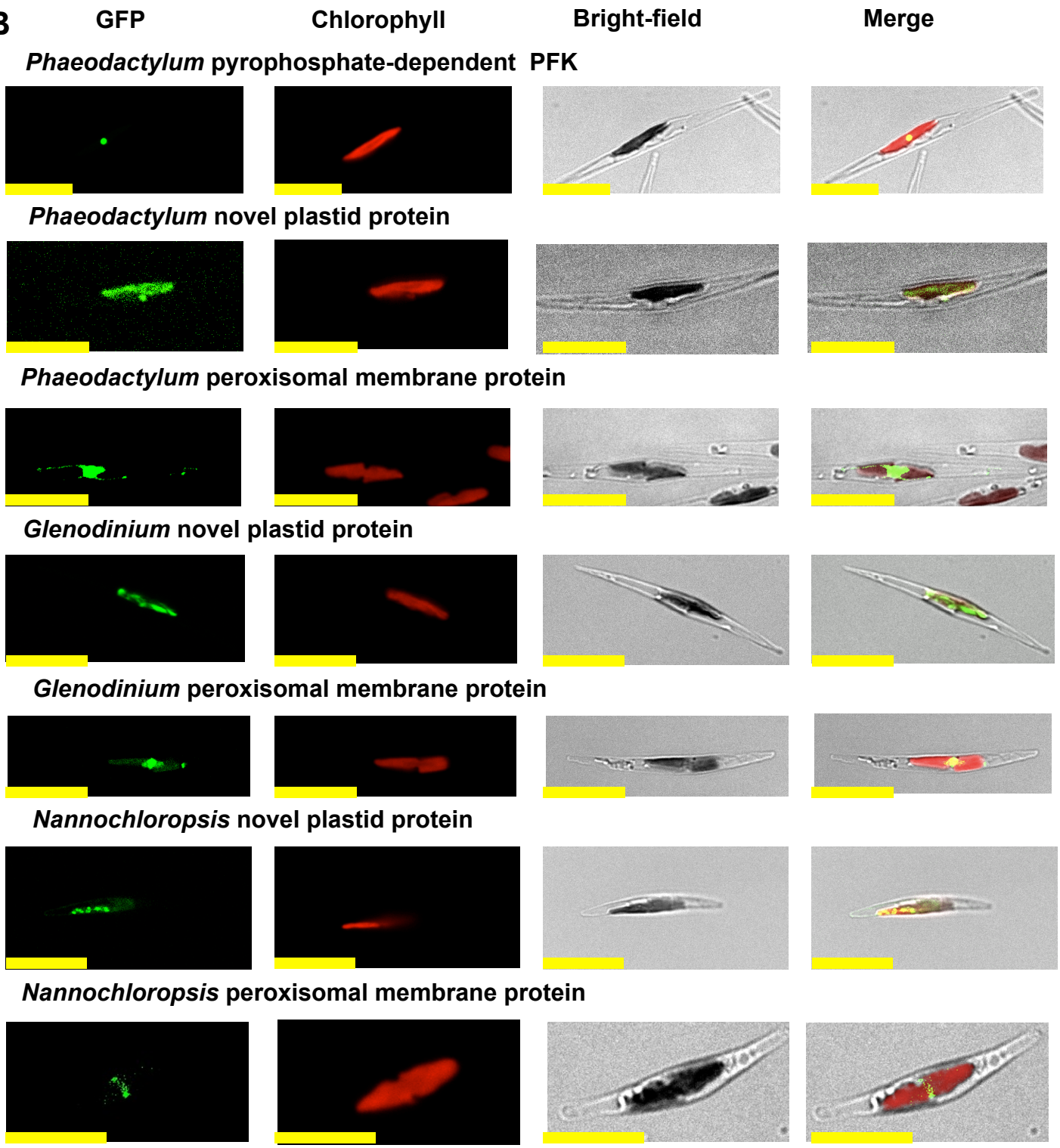
HPPGs	Total	+ve	-ve	Total	+ve	-ve
	ASAFind			HECTAR		
Total	7238	181	1970	2858	155	493
Passed HPPG assembly	924	104	7	291	65	3
Ancestral	731	102	2	278	60	2

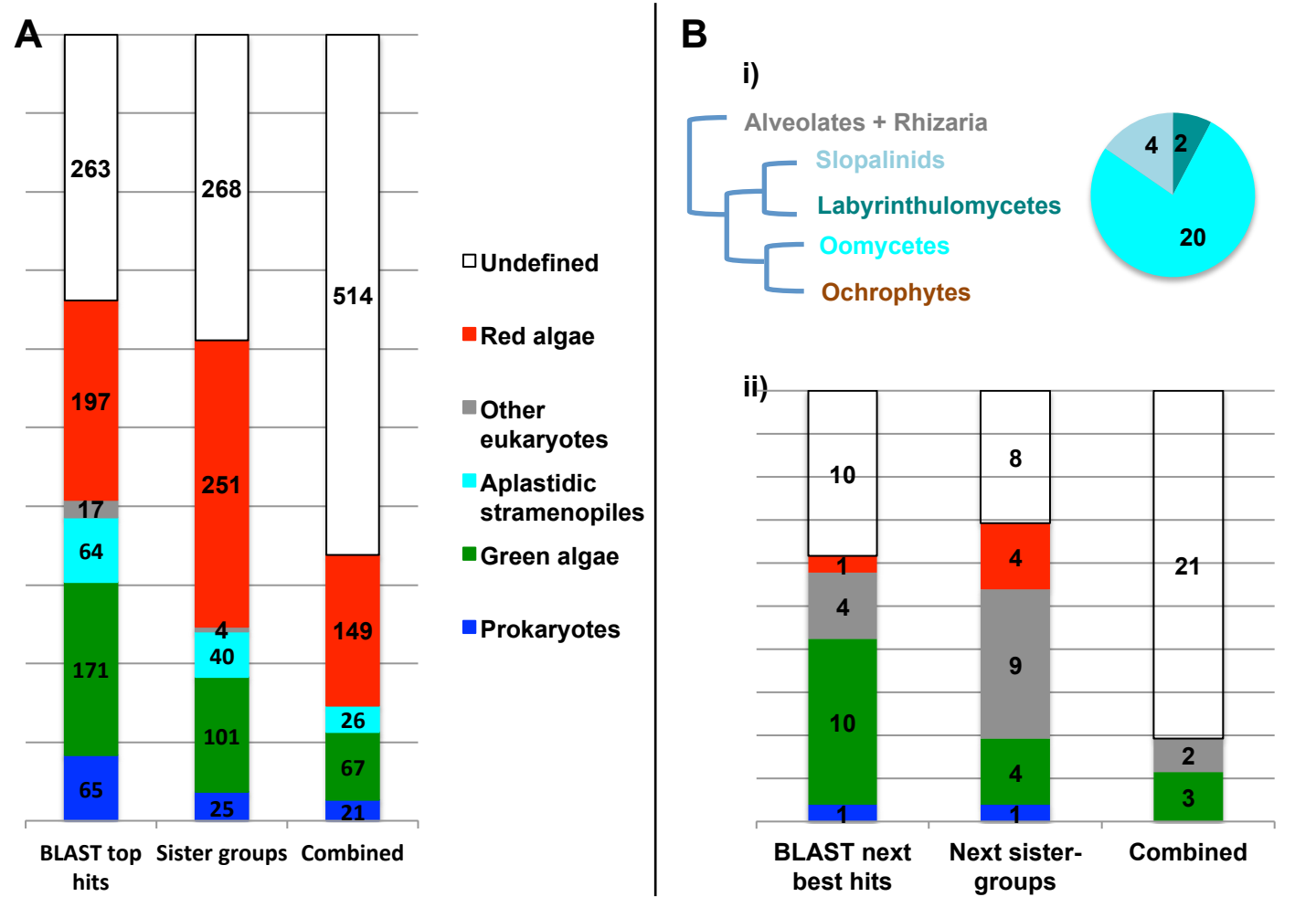
Total ancestral homologous plastid-targeted protein groups (HPPGs)= 770
Total positive controls= 106
Total negative controls= 4

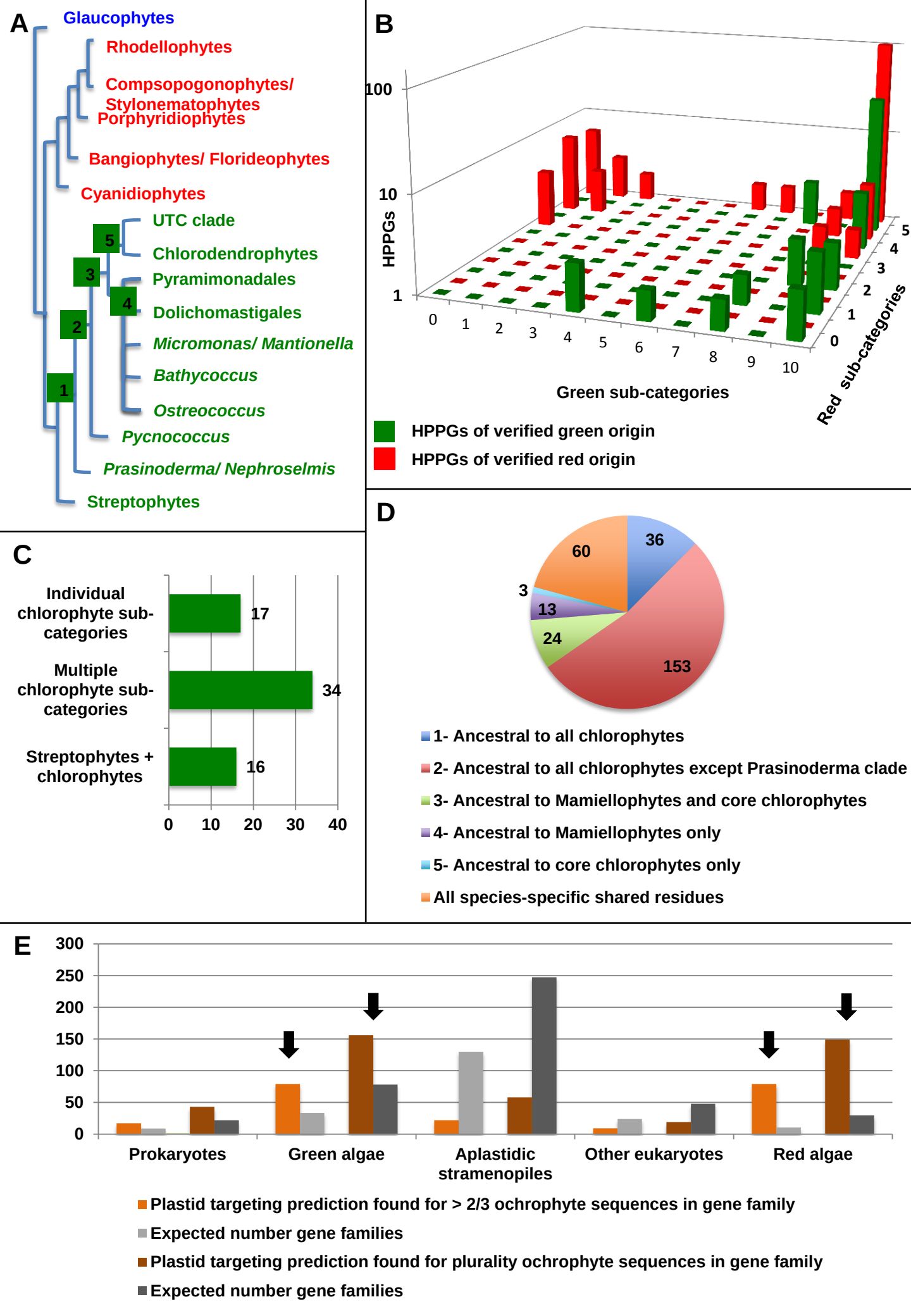
A

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



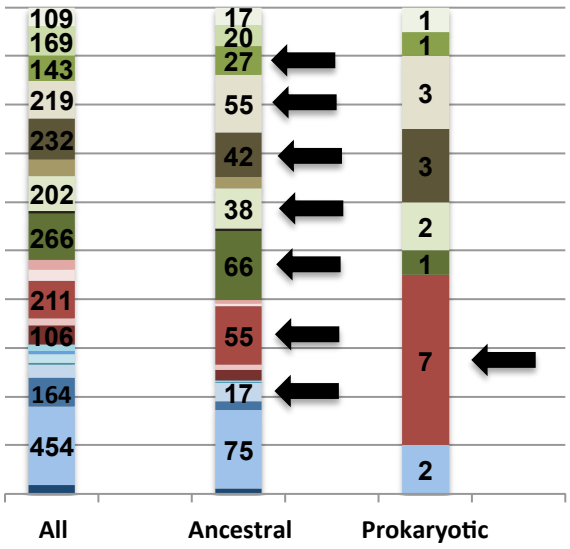

A
263
197
17
64
171
65
268
251
4
40
101
25
514
149
26
67
21
BLAST top hits
Sister groups
Combined
Undefined
Red algae
Other eukaryotes
Aplastidic stramenopiles
Green algae
Prokaryotes
B
i)
Alveolates + Rhizaria
Slopalinids
Labyrinthulomycetes
Oomycetes
Ochrophytes
4
2
20
ii)
10
1
4
10
1
8
4
9
4
1
21
2
3
BLAST next best hits
Next sister-groups
Combined

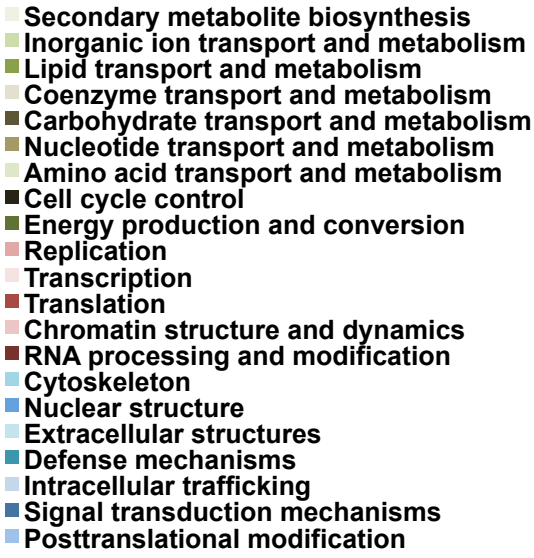

A
Glaucophytes
Rhodellophytes
Compsopogonophytes/
Stylonematophytes
Porphyridiophytes
Bangiophytes/ Florideophytes
Cyanidiophytes
UTC clade
Chlorodendrophytes
Pyramimonadales
Dolichomastigales
Micromonas/ Mantionella
Bathycoccus
Ostreococcus
Pycnococcus
Prasinoderma/ Nephroselmis
Streptophytes
5
3
4
2
1
B
100
10
1
HPPGs
0 1 2 3 4 5 6 7 8 9 10
Green sub-categories
0 1 2 3 4 5
Red sub-categories
HPPGs of verified green origin
HPPGs of verified red origin
C
Individual chlorophyte sub-categories
17
Multiple chlorophyte sub-categories
34
Streptophytes + chlorophytes
16
0 10 20 30 40
D
36
153
24
13
3
60
1- Ancestral to all chlorophytes
2- Ancestral to all chlorophytes except Prasinoderma clade
3- Ancestral to Mamiellophytes and core chlorophytes
4- Ancestral to Mamiellophytes only
5- Ancestral to core chlorophytes only
All species-specific shared residues
E
300
250
200
150
100
50
0
Prokaryotes
Green algae
Aplastidic stramenopiles
Other eukaryotes
Red algae
Plastid targeting prediction found for > 2/3 ochrophyte sequences in gene family
Expected number gene families
Plastid targeting prediction found for plurality ochrophyte sequences in gene family
Expected number gene families

A

Biological process	Identified	Plastid-encoded	Dispensible	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

B

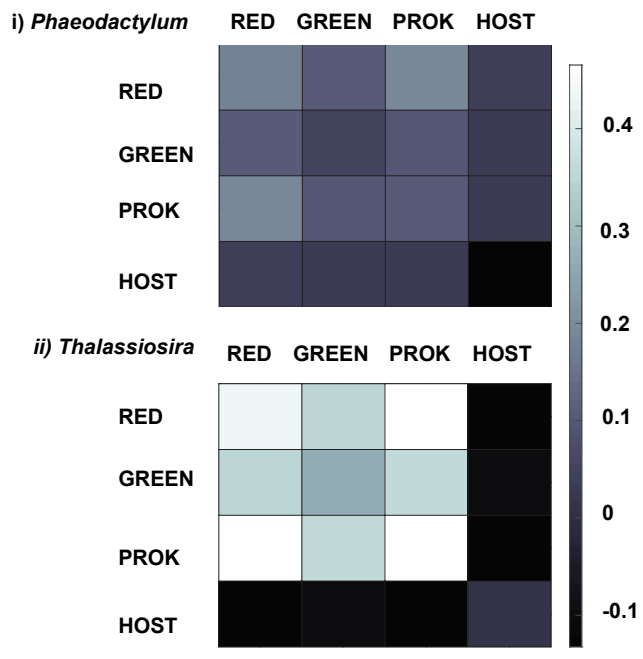




C

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

D

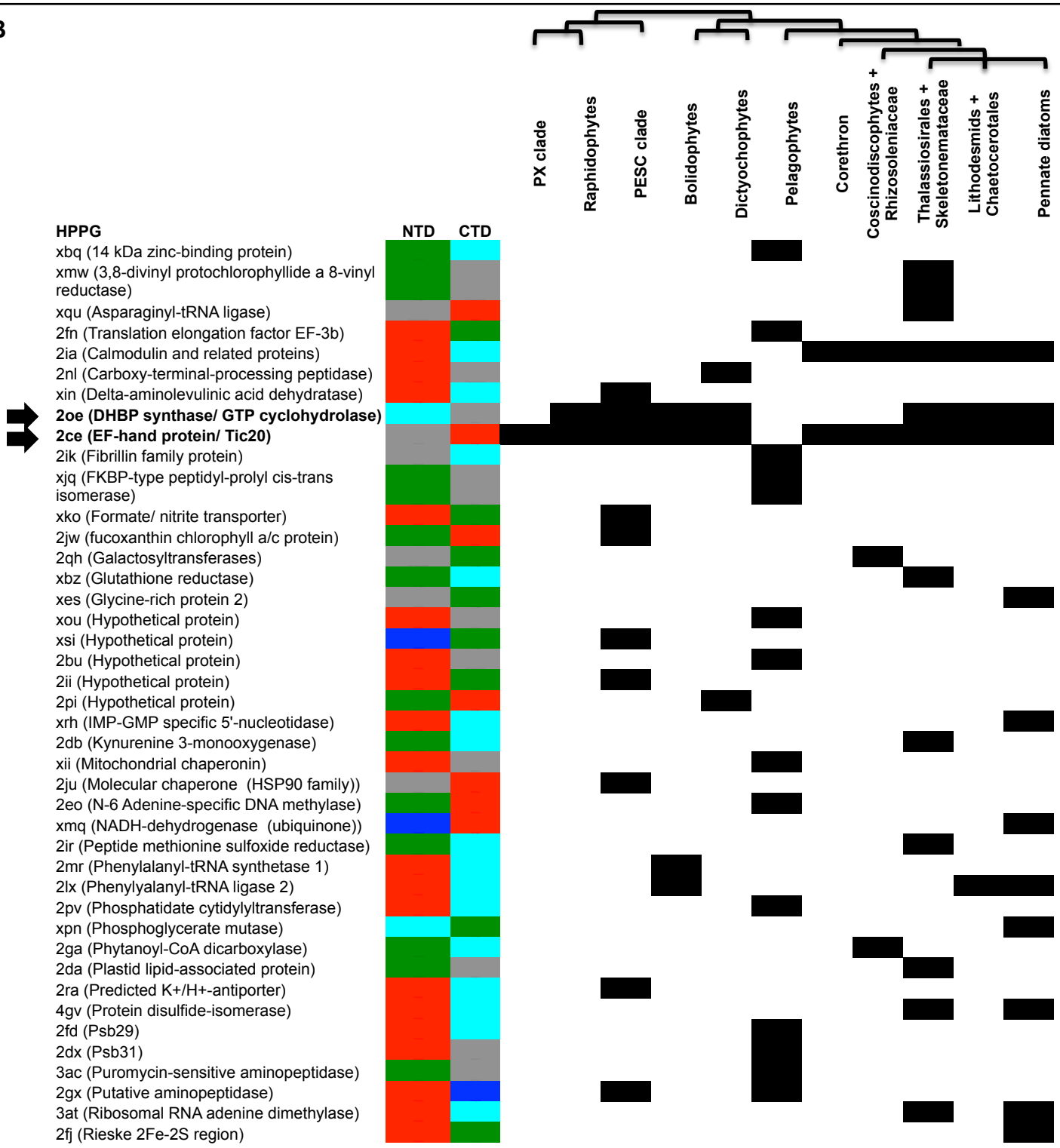


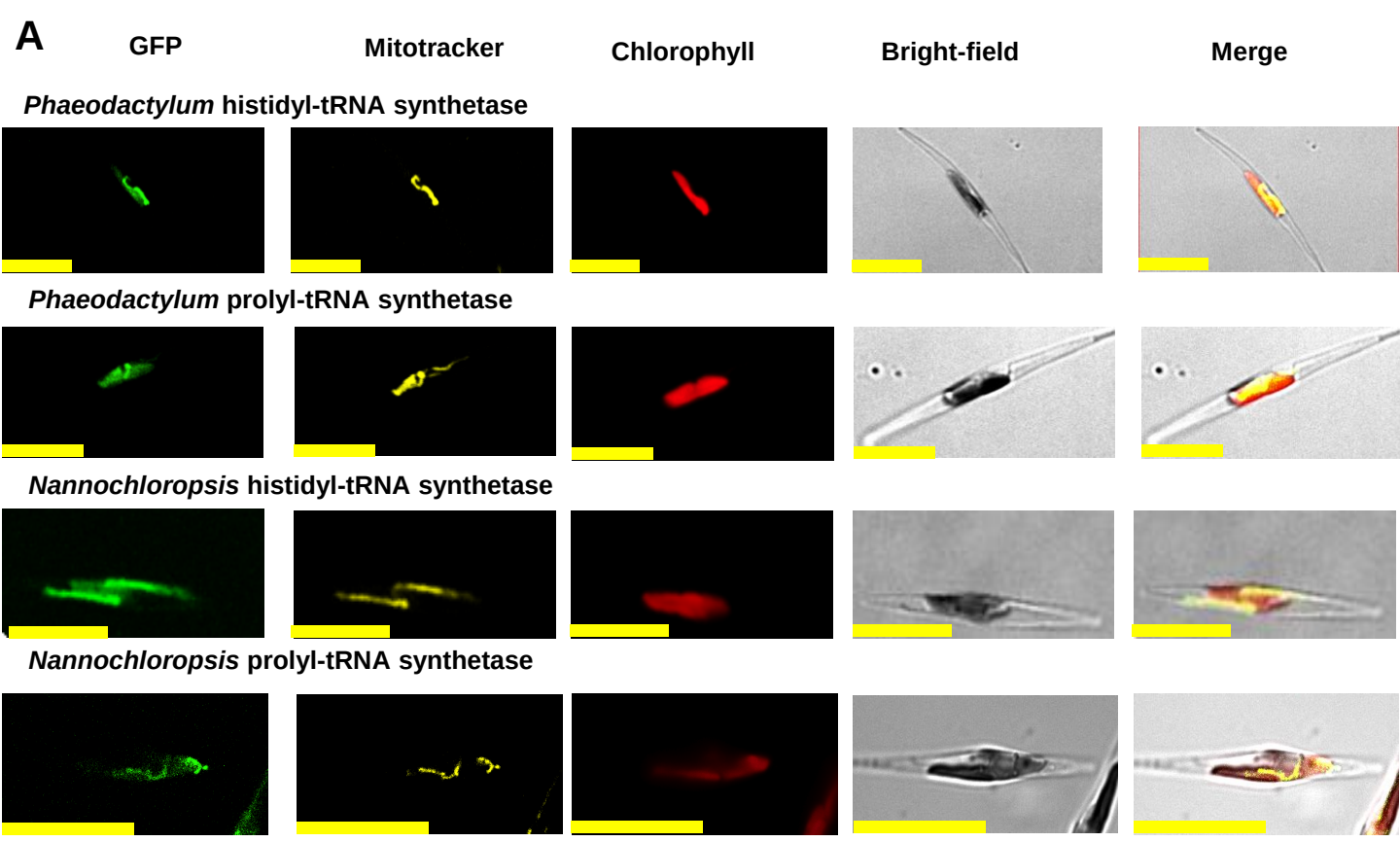
A

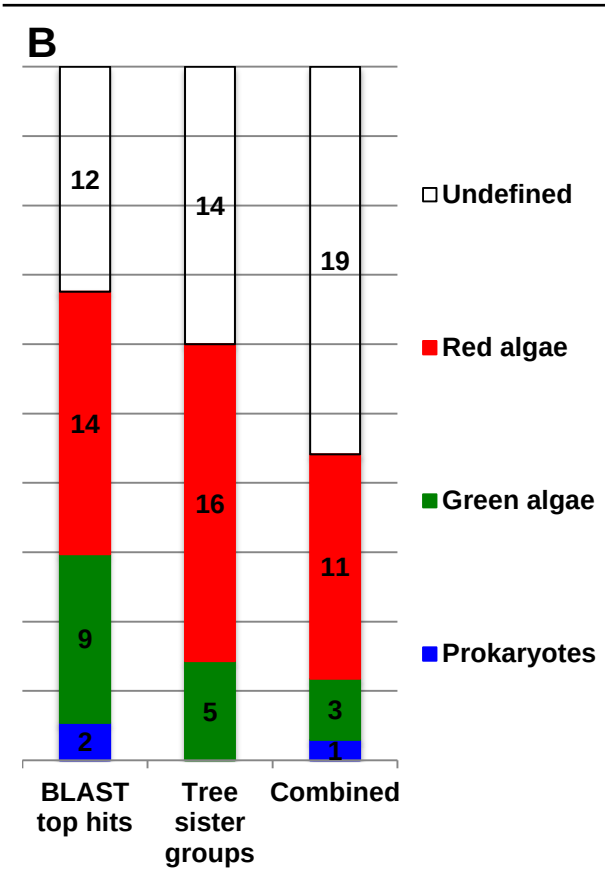
i) Chimeras inherited by the ochrophyte ancestor	Origin in ochrophytes	NTD	CTD
PpiC-type peptidyl-prolyl cis-trans isomerase	Ambiguous	Firmicutes/ Proteobacteria	Cyanobacteria
Hypothetical protein	Red	Firmicutes	Cyanobacteria
Rieske 2Fe-2S region	Green	Cyanobacteria	Proteobacteria
Probable heme-binding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

ii) Chimeras endogenous to ochrophytes	NTD	CTD
Calmodulin and related proteins/ Tic20	Unknown	Red algae
DHBP synthase/ GTP cyclohydrolase	Aplastidic stramenopiles	Actinobacteria/ Red algae

B

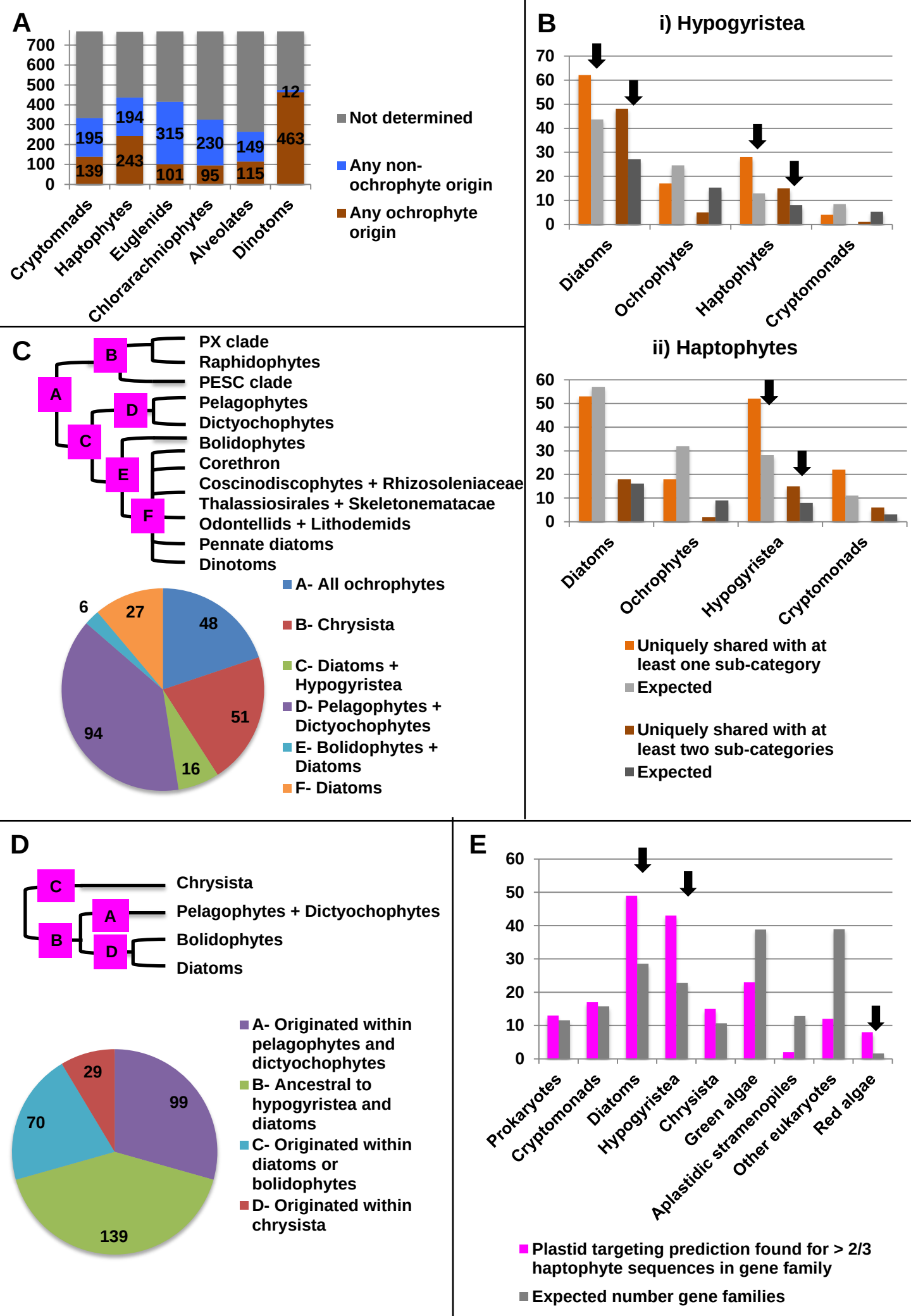


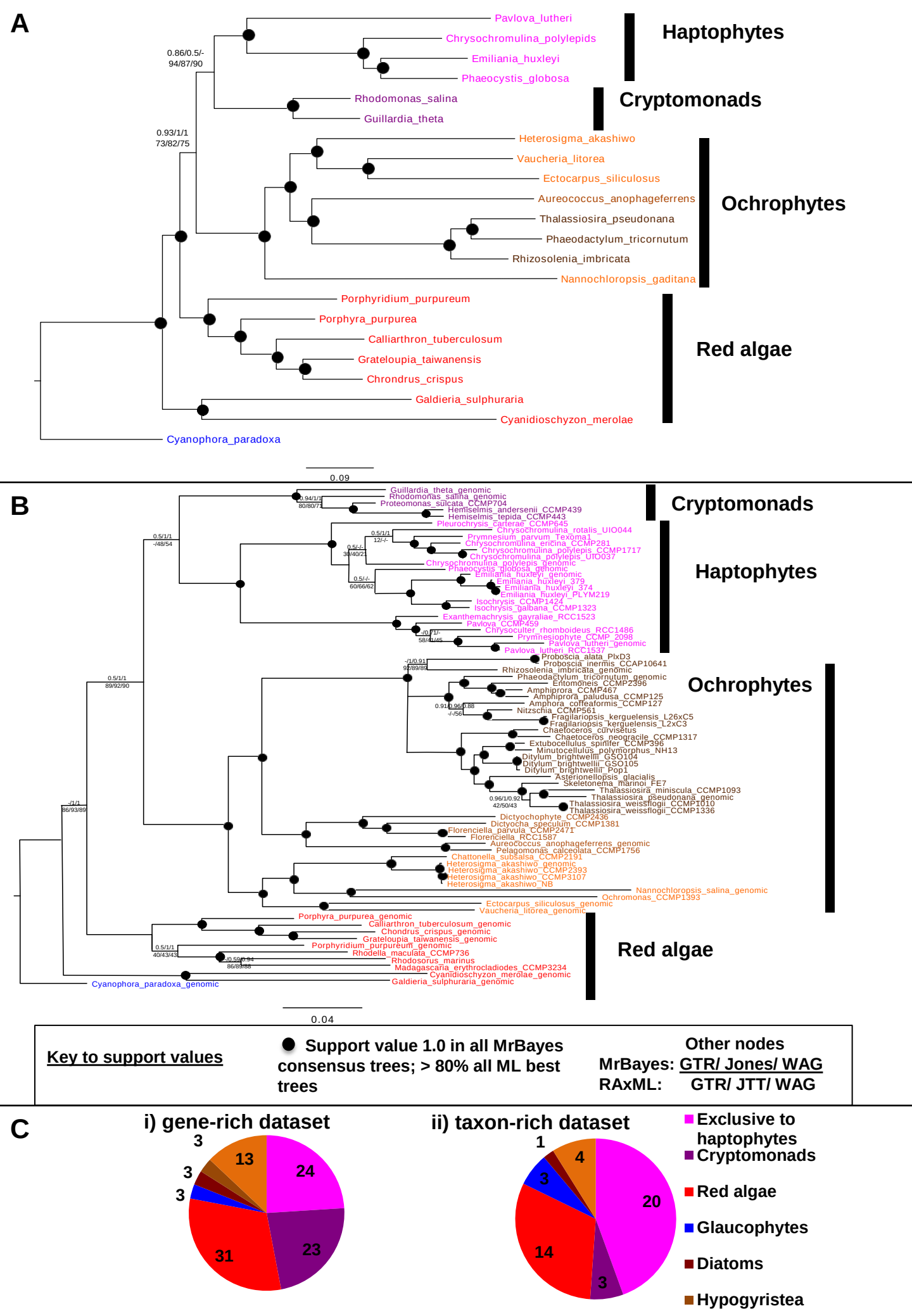


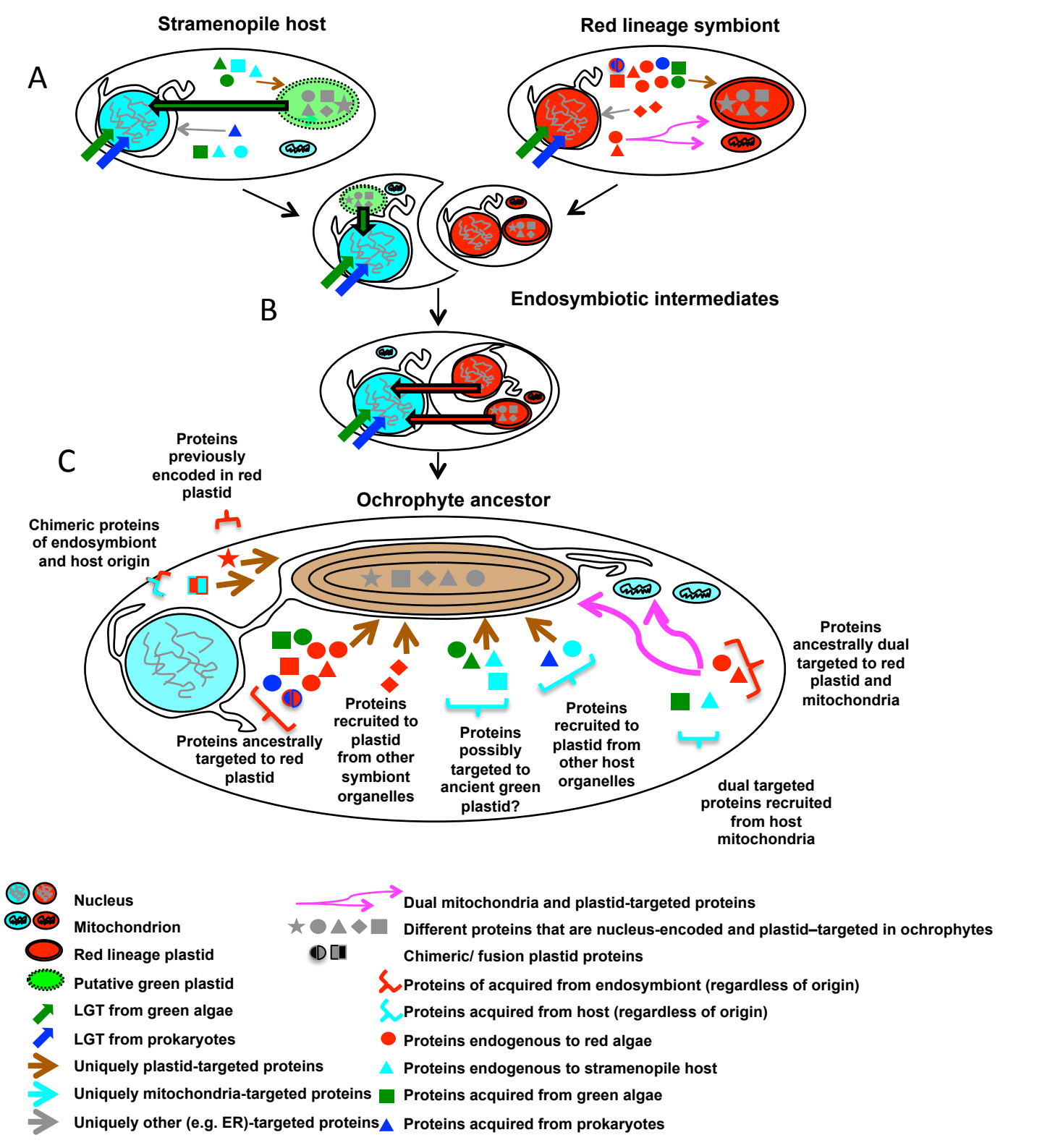


C

tRNA synthetase	Cytoplasmic isoform	Dual-targeted isoform
Ser	Aplastidic stram	Prokaryotic
Ala	Aplastidic stram	Aplastidic stram
Trp, Arg, Asn, Asp, Val	Aplastidic stram	Red algal


A
700
600
500
400
300
200
100
0
195
139
194
243
315
101
230
95
149
115
12
463
Cryptomnads
Haptophytes
Euglenids
Chlorarachniophytes
Alveolates
Dinotoms
Not determined
Any non-ochrophyte origin
Any ochrophyte origin
B
i) Hypogyristea
70
60
50
40
30
20
10
0
Diatoms
Ochrophytes
Haptophytes
Cryptomonads
ii) Haptophytes
60
50
40
30
20
10
0
Diatoms
Ochrophytes
Hypogyristea
Cryptomonads
Uniquely shared with at least one sub-category
Expected
Uniquely shared with at least two sub-categories
Expected
C
A
B
C
D
E
F
PX clade
Raphidophytes
PESC clade
Pelagophytes
Dictyochophytes
Bolidophytes
Corethron
Coscinodiscophytes + Rhizosoleniaceae
Thalassiosirales + Skeletonematacae
Odontellids + Lithodemids
Pennate diatoms
Dinotoms
6
27
48
51
94
16
A- All ochrophytes
B- Chrysista
C- Diatoms + Hypogyristea
D- Pelagophytes + Dictyochophytes
E- Bolidophytes + Diatoms
F- Diatoms
D
C
A
B
D
Chrysista
Pelagophytes + Dictyochophytes
Bolidophytes
Diatoms
29
99
70
139
A- Originated within pelagophytes and dictyochophytes
B- Ancestral to hypogyristea and diatoms
C- Originated within diatoms or bolidophytes
D- Originated within chrysista
E
60
50
40
30
20
10
0
Prokaryotes
Cryptomonads
Diatoms
Hypogyristea
Chrysista
Green algae
Aplastidic stramenopiles
Other eukaryotes
Red algae
Plastid targeting prediction found for > 2/3 haptophyte sequences in gene family
Expected number gene families


A
Pavlova_lutheri
Chrysochromulina_polylepids
Emiliania_huxleyi
Phaeocystis_globosa
Haptophytes
0.86/0.5/-
94/87/90
Rhodomonas_salina
Guillardia_theta
Cryptomonads
0.93/1/1
73/82/75
Heterosigma_akashiwo
Vaucheria_litorea
Ectocarpus_siliculosus
Aureococcus_anophageferrens
Thalassiosira_pseudonana
Phaeodactylum_tricornutum
Rhizosolenia_imbricata
Nannochloropsis_gaditana
Ochrophytes
Porphyridium_purpureum
Porphyra_purpurea
Calliarthron_tuberculosum
Grateloupia_taiwanensis
Chrondrus_crispus
Galdieria_sulphuraria
Cyanidioschyzon_merolae
Red algae
Cyanophora_paradoxa
0.09
B
Guillardia_theta_genomic
Rhodomonas_salina_genomic
Proteomonas_sulcata_CCMP704
Hemiselmis_andersenii_CCMP439
Hemiselmis_tepida_CCMP443
Cryptomonads
Pleurochrysis_carterae_CCMP645
Chrysochromulina_rotalis_UIO044
Prymnesium_parvum_Texoma1
Chrysochromulina_ericina_CCMP281
Chrysochromulina_polylepis_CCMP1717
Chrysochromulina_polylepis_UIO037
Chrysochromulina_polylepis_genomic
Phaeocystis_globosa_genomic
Emiliania_huxleyi_genomic
Emiliania_huxleyi_379
Emiliania_huxleyi_374
Emiliania_huxleyi_PLYM219
Isochrysis_CCMP1424
Isochrysis_galbana_CCMP1323
Exanthemachrysis_gayraliae_RCC1523
Pavlova_CCMP459
Chrysoculter_rhomboideus_RCC1486
Prymnesiophyte_CCMP_2098
Pavlova_lutheri_genomic
Pavlova_lutheri_RCC1537
Haptophytes
Proboscia_alata_PlxD3
Proboscia_inermis_CCAP10641
Rhizosolenia_imbricata_genomic
Phaeodactylum_tricornutum_genomic
Entomoneis_CCMP2396
Amphiprora_CCMP467
Amphiprora_paludusa_CCMP125
Amphora_coffeaformis_CCMP127
Nitzschia_CCMP561
Fragilariopsis_kerguelensis_L26xC5
Fragilariopsis_kerguelensis_L2xC3
Chaetoceros_curvisetus
Chaetoceros_neogracile_CCMP1317
Extubocellulus_spinifer_CCMP396
Minutocellulus_polymorphus_NH13
Ditylum_brightwellii_GSO104
Ditylum_brightwellii_GSO105
Ditylum_brightwellii_Pop1
Asterionellopsis_glacialis
Skeletonema_marinoi_FE7
Thalassiosira_miniscula_CCMP1093
Thalassiosira_pseudonana_genomic
Thalassiosira_weissflogii_CCMP1010
Thalassiosira_weissflogii_CCMP1336
Dictyochophyte_CCMP2436
Dictyocha_speculum_CCMP1381
Florenciella_parvula_CCMP2471
Florenciella_RCC1587
Aureococcus_anophageferrens_genomic
Pelagomonas_calceolata_CCMP1756
Chattonella_subsalsa_CCMP2191
Heterosigma_akashiwo_genomic
Heterosigma_akashiwo_CCMP2393
Heterosigma_akashiwo_CCMP3107
Heterosigma_akashiwo_NB
Nannochloropsis_salina_genomic
Ochromonas_CCMP1393
Ectocarpus_siliculosus_genomic
Vaucheria_litorea_genomic
Ochrophytes
Porphyra_purpurea_genomic
Calliarthron_tuberculosum_genomic
Chondrus_crispus_genomic
Grateloupia_taiwanensis_genomic
Porphyridium_purpureum_genomic
Rhodella_maculata_CCMP736
Rhodosorus_marinus
Madagascaria_erythrocladiodes_CCMP3234
Cyanidioschyzon_merolae_genomic
Galdieria_sulphuraria_genomic
Red algae
Cyanophora_paradoxa_genomic
0.04
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
24
23
31
3
3
3
13
ii) taxon-rich dataset
20
3
14
3
1
4
Exclusive to haptophytes
Cryptomonads
Red algae
Glaucophytes
Diatoms
Hypogyristea


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 targeted to red plastid
Proteins recruited to plastid from other symbiont organelles
Proteins possibly targeted to ancient green plastid?
Proteins recruited to plastid from other host organelles
Proteins ancestrally dual targeted to red plastid and mitochondria
dual targeted proteins recruited from host 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

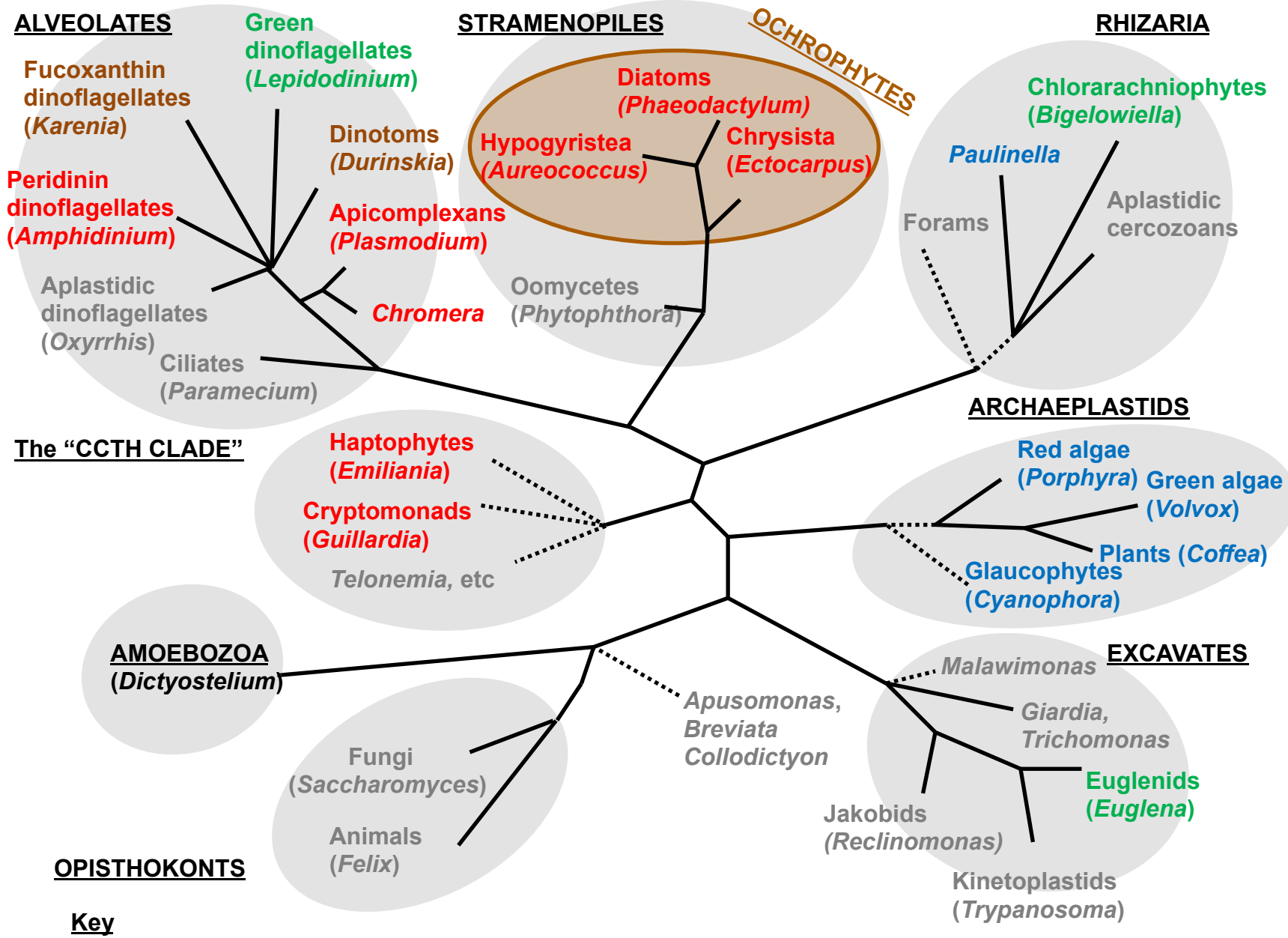
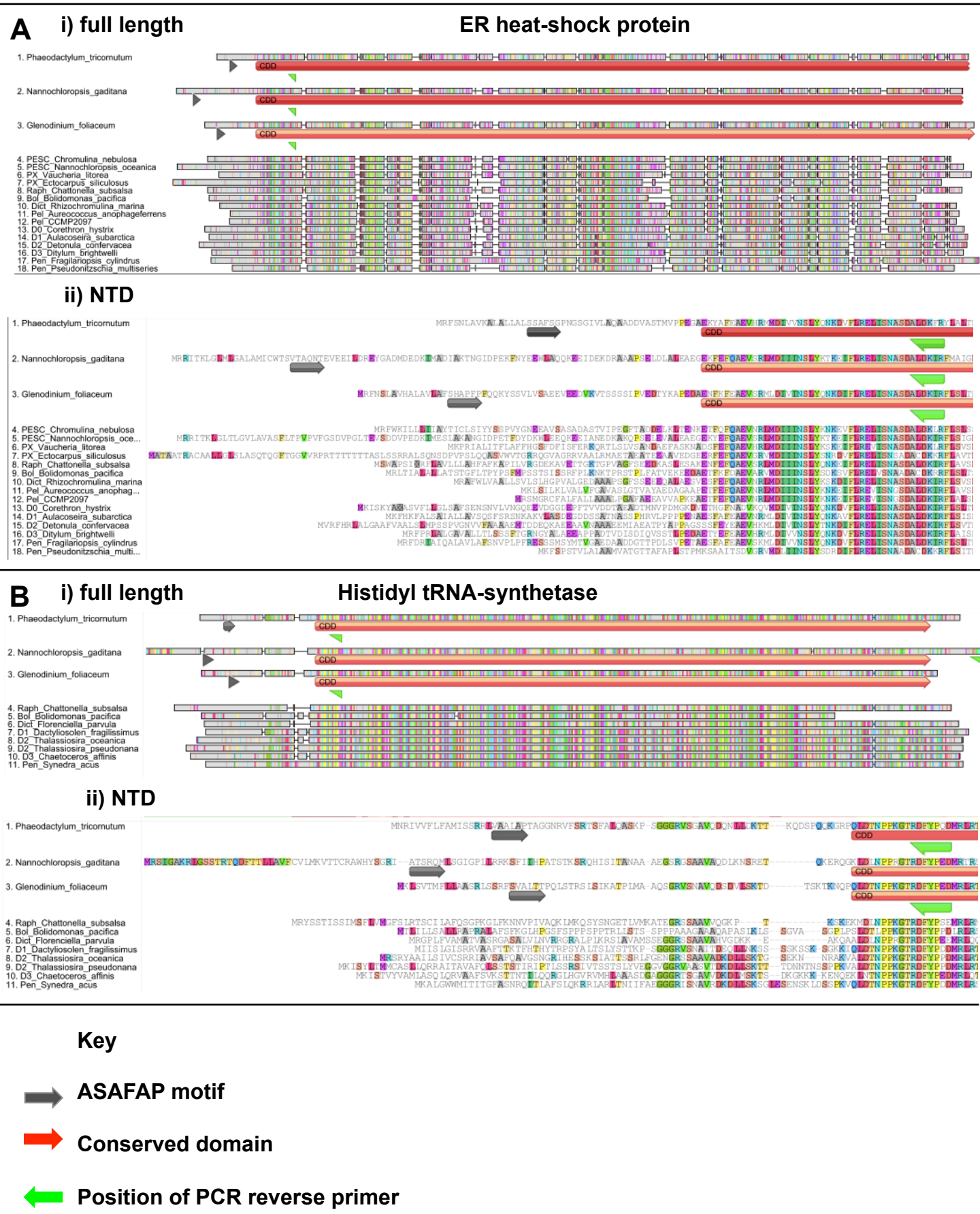


Fig. 1- figure supplement 1. Overview of eukaryotic diversity. This figure, adapted from a previous review[3], profiles the diversity of different eukaryotic nuclear lineages. Each grey ellipse corresponds to one major clade, or "supergroup" of eukaryotes. A brown ellipse within the stramenopile clade 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.



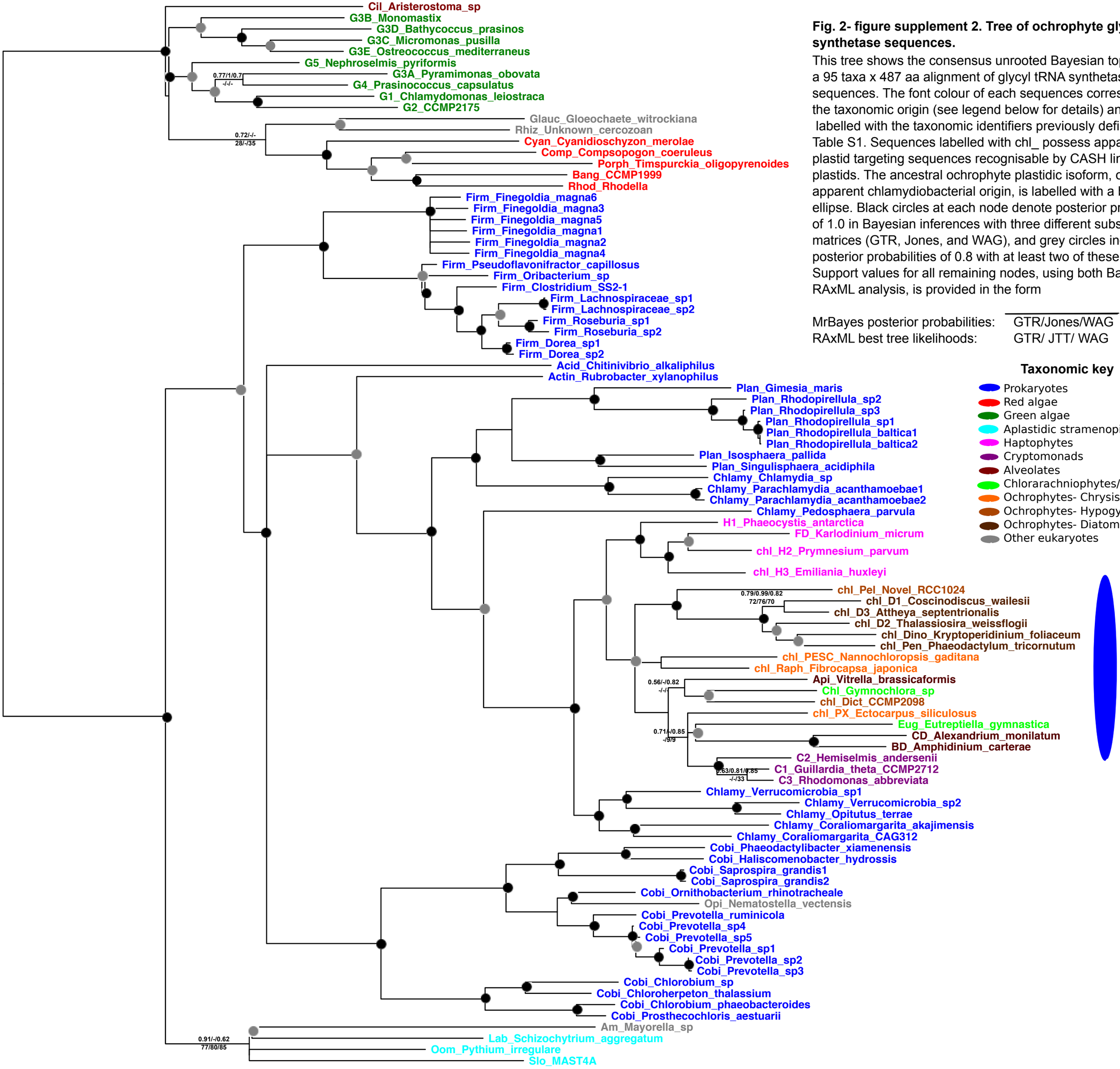


Fig. 2- figure supplement 2. Tree of ochrophyte glycyl-tRNA synthetase sequences.

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- Hypogyristea
- Ochrophytes- Diatoms
- Other eukaryotes

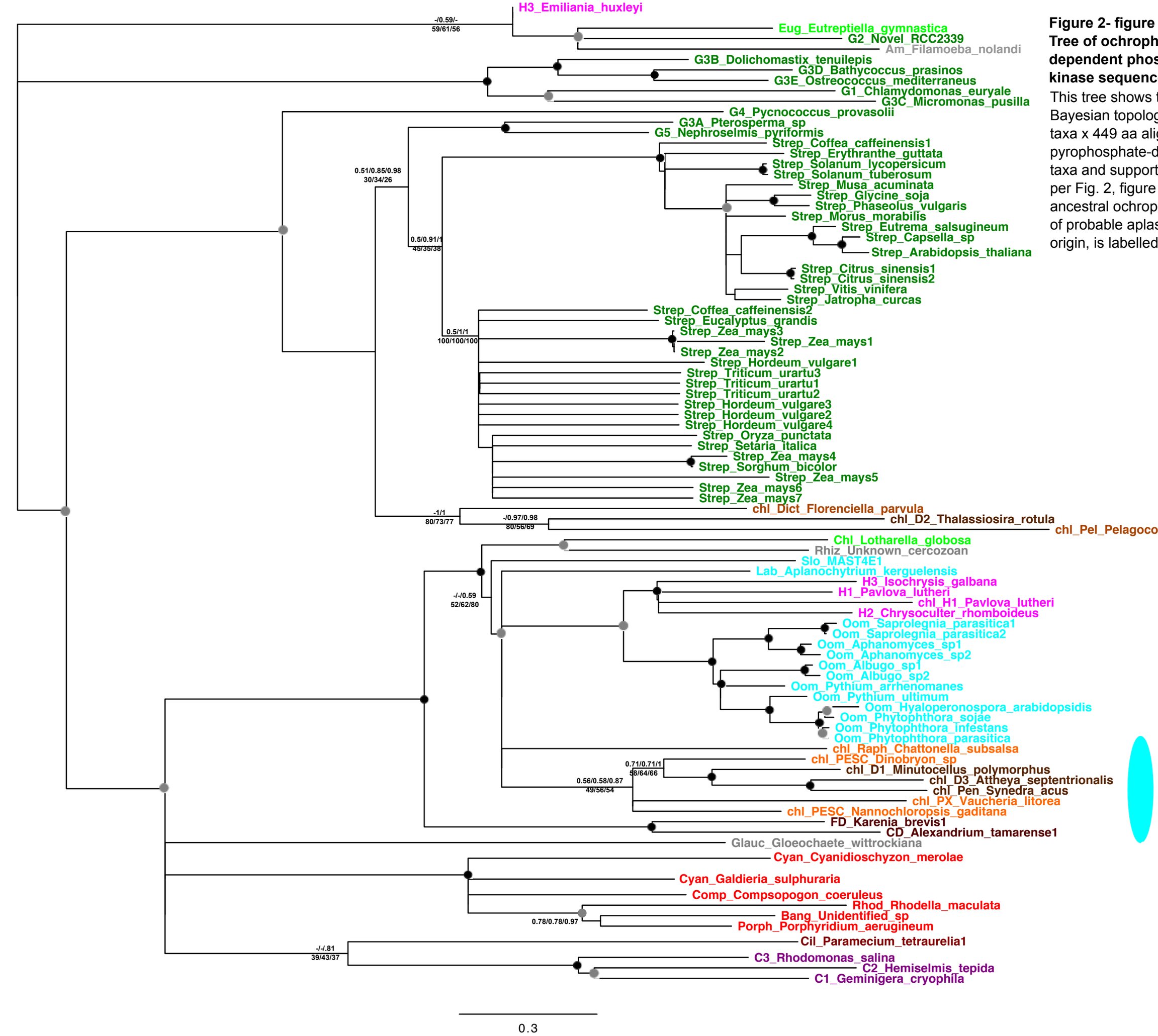


Figure 2- figure supplement 3. Tree of ochrophyte pyrophosphate dependent phosphofructo-1-kinase sequences.

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 aplastidic stramenopile origin, is labelled with a cyan ellipse.

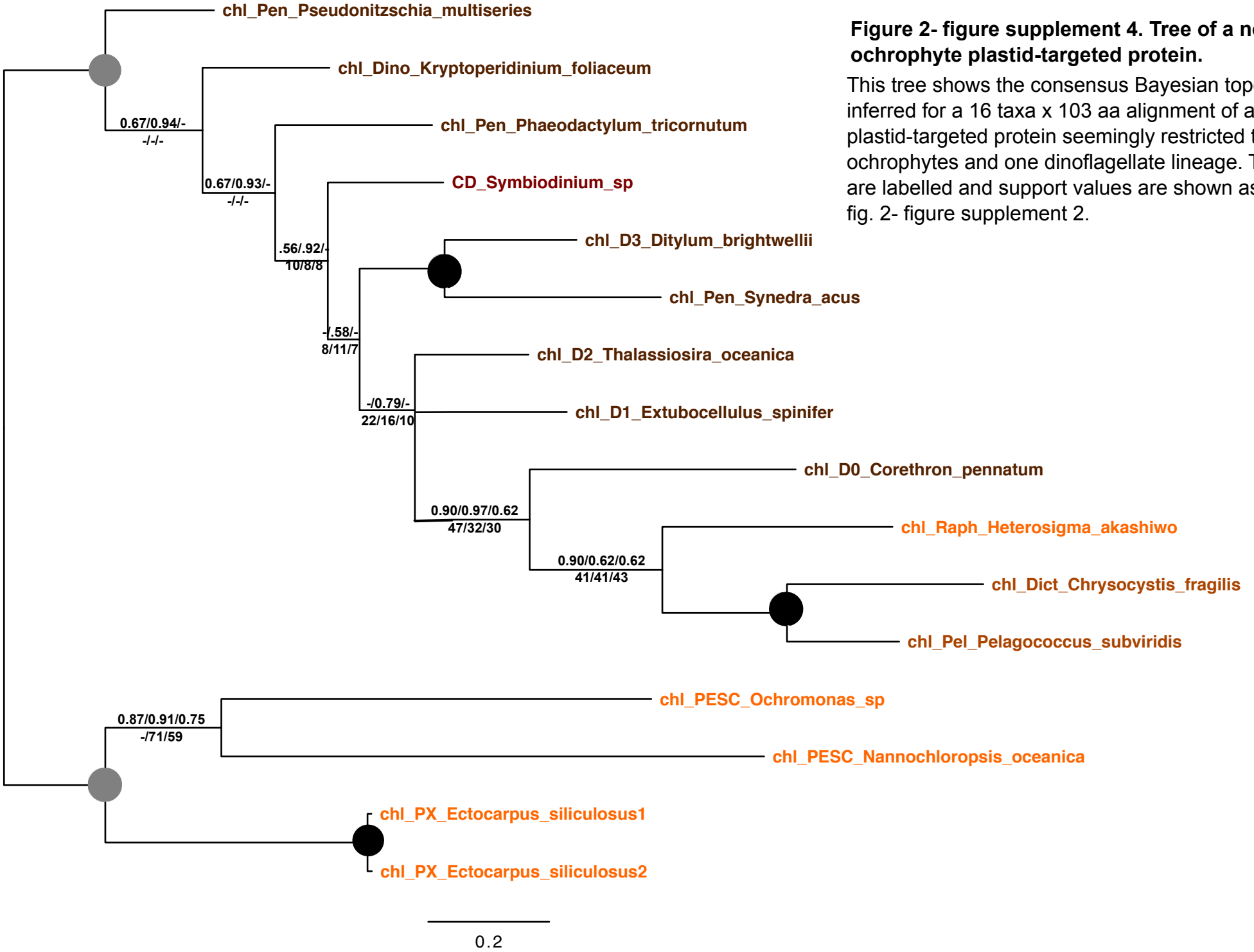


Figure 2- figure supplement 4. Tree of a novel ochrophyte plastid-targeted protein.

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).

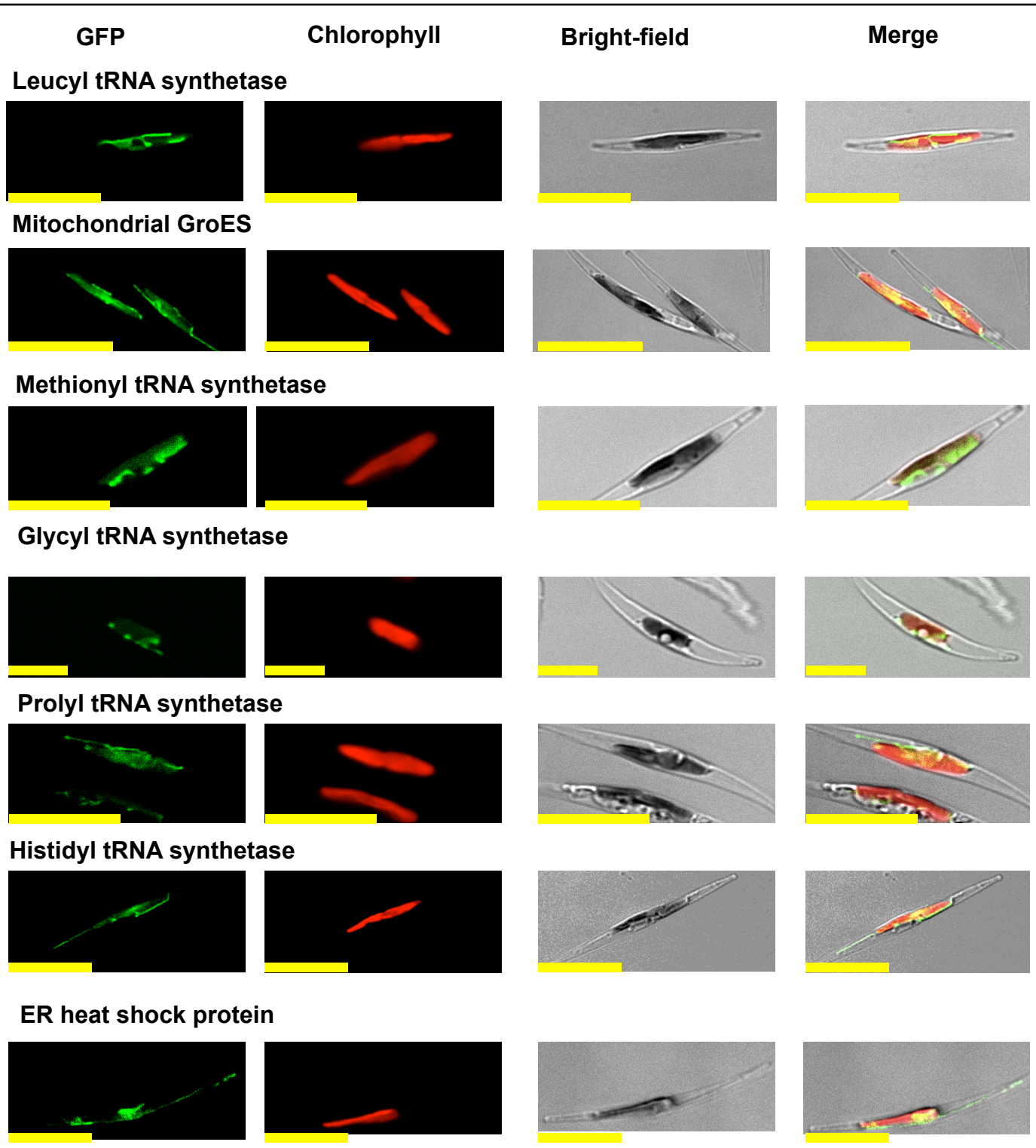


Fig. 2- figure supplement 6.Heterologous expression constructs of multipartite plastid-targeted proteins. This figure shows the localisation of GFP overexpression constructs for copies of two proteins from the dinotom *Glenodinium foliaceum* (**Panel A**) , and three proteins from the eustigmatophyte *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.

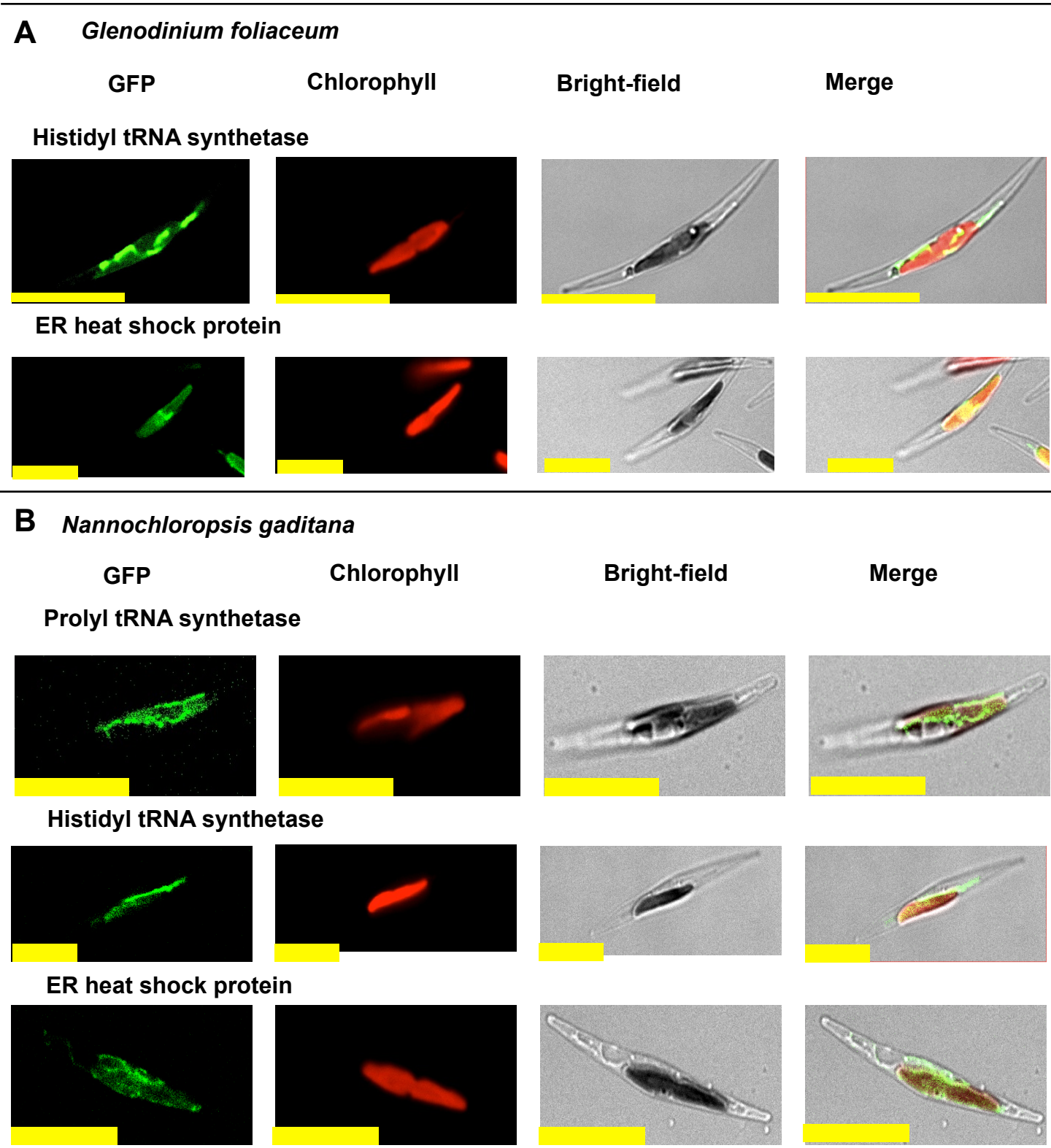


Fig. 2- figure supplement 7. Exemplar control images for confocal microscopy. This figure shows fluorescence patterns for wild-type *Phaeodactylum tricornutum* cells (**i**), and transformant *Phaeodactylum* cells expressing GFP that has not been fused to any N-terminal targeting sequence (**ii**), both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

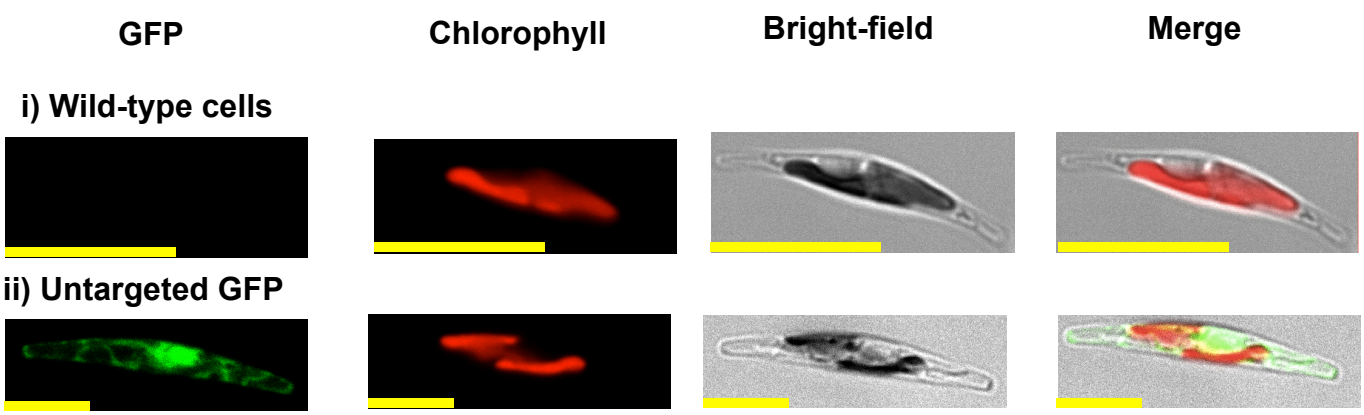


Fig. 4- figure supplement 1. Sampling richness associated with ancestral HPPGs of green algal origin. This figure shows the number of sub-different archaeplastid orthologues for ancestral HPPGs verified by combined BLAST top hit and single-gene tree analysis to be of either green algal origin (green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

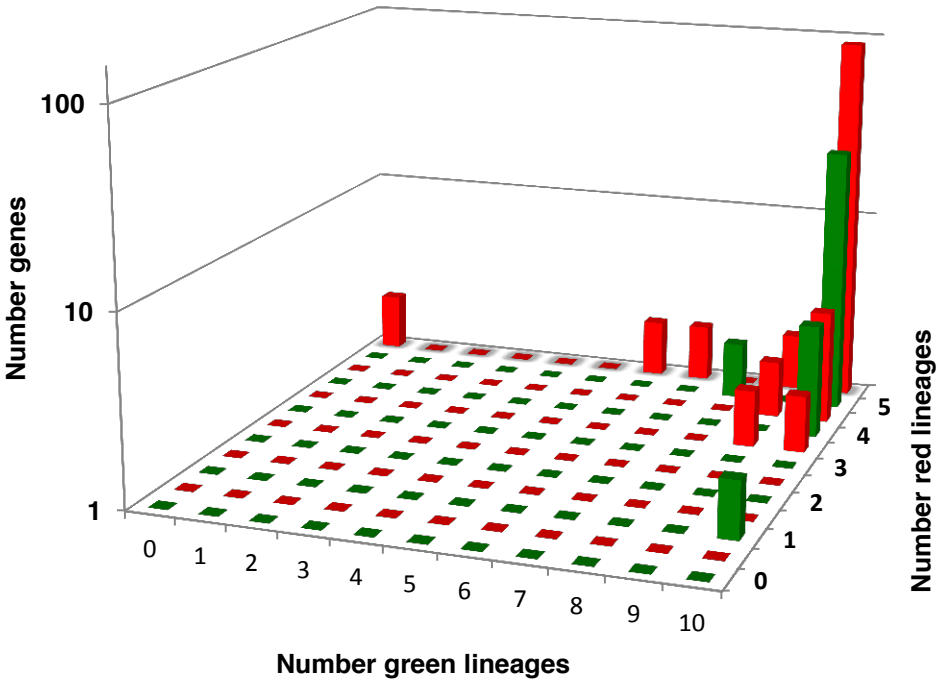
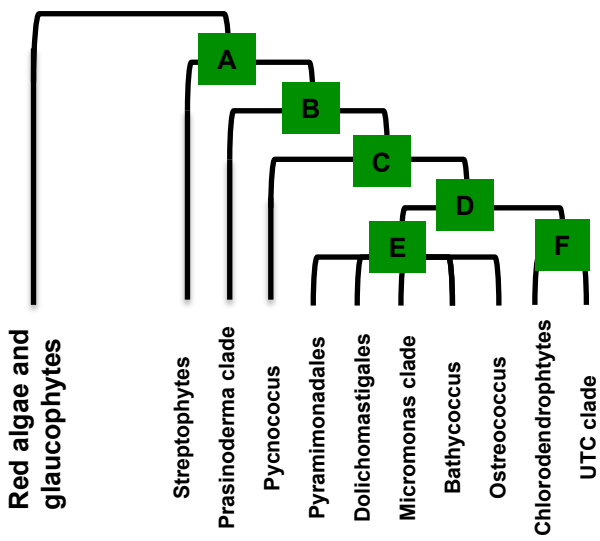


Fig.4- figure supplement 2. Heatmaps of nearest sister-groups of ancestral HPPGs of verified green origin. This figure shows the specific topologies of single gene trees for HPPGs verified to be of green origin by combined BLAST and phylogenetic analysis. **Panel A** shows a reference topology of evolutionary 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.



2bg (Chlorophyll A-B binding protein)
2fb (Hypothetical protein)
xiw (Magnesium-protoporphyrin IX methyltransferase)
xkc (GDT1-like membrane protein)
xkf (SAM-dependent methyltransferases)
xkm (Uroporphyrinogen decarboxylase)
xlb (Malonyl-CoA:ACP transacylase)
xsq (Thylakoid lumen 15kDa protein)
2gk (Hypothetical protein)
2ht (Putative Gun4 protein)
2ek (Hypothetical protein)
2fu (Hypothetical protein)
2ev (ATP-dependent RNA helicase)
xgl (coproporphyrinogen oxidase)
xqm (Ubiquinol oxidase)
xdz (Hypothetical protein)
2af (Metallophosphoesterase)
2da (Plastid lipid-associated protein/fibrillin)
2iq (Predicted unusual protein kinase)
2iu (Delta 6-fatty acid desaturase/delta-8 sphingolipid desaturase)
2mh (fructose-bisphosphate aldolase)
7ah (Cycloeucalenol cycloisomerase)
xdk (Ferredoxin-NADP oxidoreductase)
xie (Uncharacterized enzymes related to aldose 1-epimerase)
xik (Glucose/ribitol dehydrogenase)
xkt (Ribose-5-phosphate isomerase)
xps (Glutaredoxin)
2cz (Uncharacterized membrane protein, predicted efflux pump)
xmm (3-isopropylmalate dehydratase)
2he (Hypothetical protein)
2kl (Protochlorophyllide reductase A)
xcr (Hypothetical protein)
xen (6-phosphogluconolactonase)
2ml (Hypothetical protein)
xdd (Hypothetical protein)
xny (Hypothetical protein)
2cd (Hypothetical protein)
2cg (Glutaredoxin and related proteins)
2rd (Probable plastid-lipid-associated protein)
xoz (Hypothetical protein)
xom (Hypothetical protein)
2hv (Peptidase S1)
xqh (Hypothetical protein)
2cj (Glutathione S-transferase)
2ed (Hypothetical protein)
2ew (Predicted dehydrogenase)
2it (Hypothetical protein)
xlz (Defense-related protein containing SCP domain)
2ii (Hypothetical protein)
3aw (Hypothetical protein)
xke (Bacteriochlorophyll/chlorophyll synthetase)
xkn (Glutamate-1-semialdehyde 2,1-aminomutase (GSA))
3bb (Hypothetical protein)
xeq (Violaxanthin de-epoxidase)
2lu (Cobalamin synthesis protein)
xat (Hypothetical protein)
2ah (Hypothetical protein)
xmk (DAHP synthetase, class II)
2lo (Hypothetical protein)
2oi (Hypothetical protein)
xnb (Hypothetical protein)
2ct (Na+-independent Cl/HCO3 exchanger AE1 and related transporters (SLC4 family))
2ir (Peptide methionine sulfoxide reductase)
xbf (Alcohol dehydrogenase, class III)
xcv (Pheophytinase, Plastidic)
2jx (fucoxanthin chlorophyll a/c protein)
2ls (Predicted membrane protein)

A- All green lineages (16)

B- All chlorophytes (20)

C- All chlorophytes exc. *Prasinoderma* (3)

D- Mamiellophytes and core chlorophytes (6)

E- Mamiellophytes (3)

F- Core chlorophytes (2)

Specific chlorophyte sub-categories (18)

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.

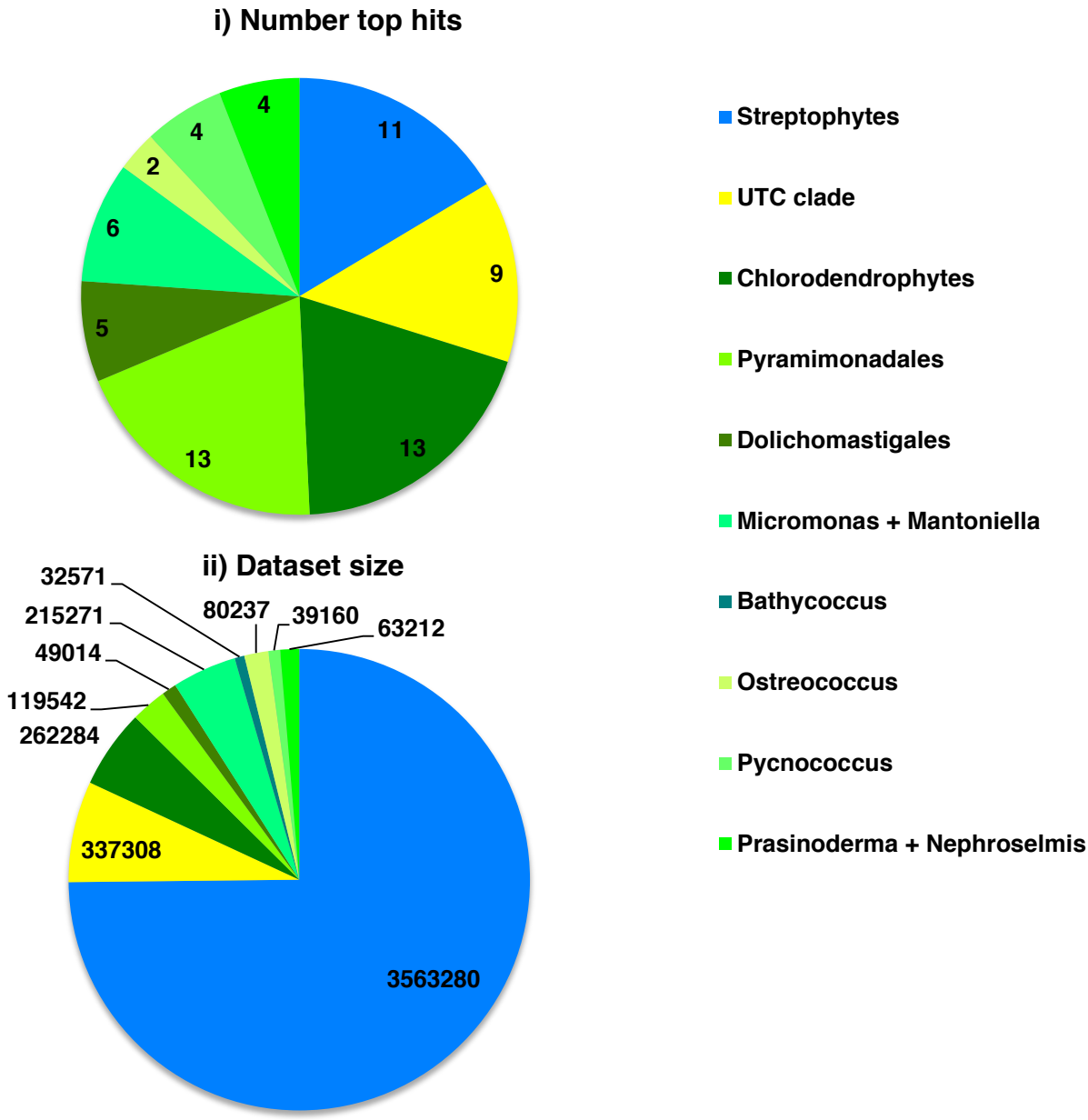


Fig. 4- figure supplement 4. Earliest evolutionary origins of shared plastid residues. This figure shows the number of residues in the concatenated alignment of HPPGs of verified green origin, which have been subsequently vertically inherited in all major photosynthetic eukaryoties that are present in 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 a 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*.

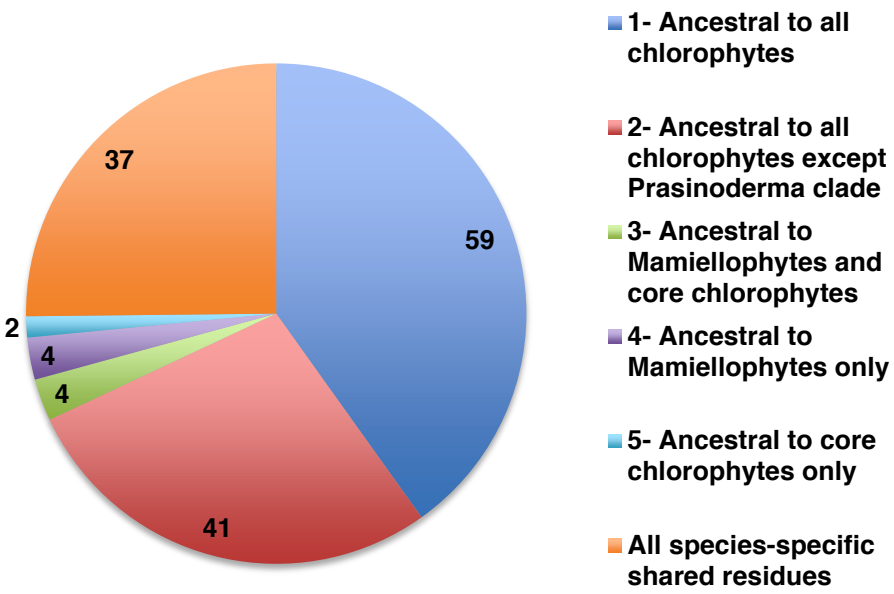


Fig. 4- figure supplement 5. Origins and HECTAR based targeting tests of proteins encoded by conserved ochrophyte gene clusters. Panel A shows the most probably evolutionary origin, identified using BLAST top hit analysis, for 7140 conserved gene clusters inferred to have been present in the last common ochrophyte ancestor. **Panel B** shows the number of these gene familieies 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.

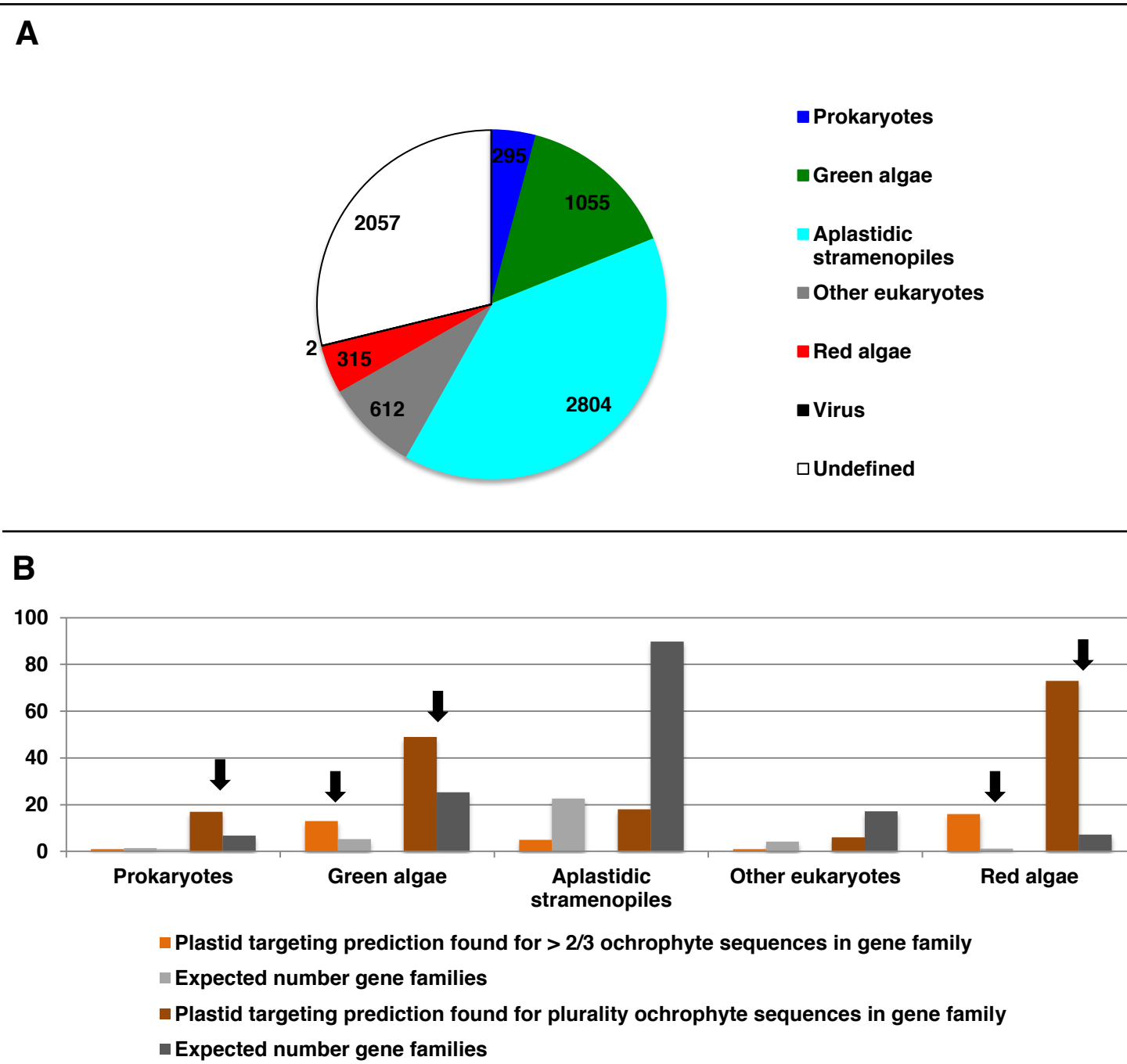


Fig. 5- figure supplement 1. Reconstructed metabolism pathways and core biological processes in the ancestral ochrophyte plastid.
This figure tabulates each of the ancestral ochrophyte HPPGs corresponding to 350 central plastid metabolism and other biological processes. The "origin" column shows the 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.

Key	Origin	Confidence: High	Confidence: Low
	Plastid-encoded		n/a
	Red algae		
	Green algae		
	Prokaryotes		
	Aplastidic stramenopiles		
	Other/ unresolved	n/a	

Cluster	Enzyme	Origin
1. Light harvesting		
2ka	Divergent li818-type	
xhu	High light inducible protein	
2kb	LhcA-type protein 1	
2kd	LhcA-type protein 2	
2kc	LhcA-type protein 3	
2ke	LhcA-type protein 4	
2jx	LhcF-type protein 1	
2jw	LhcF-type protein 2	
2jz	LhcF-type protein 3	
2kf	LhcR-type protein 1	
abj	LhcR-type protein 2	
2kh	LhcR-type protein 3	
abk	LhcR-type protein 4	
2bg	Li818-type	
2. Photosynthesis		
-	PsbA,B,C,D,E,F,H,I,J,K,L,N,T,V,W,X,Y,Z	
xmo	PsbU	
xhz	Psb27	
2ax	PsbP	
2dx	Psb31	
abn	Psb31	
2gd	PsbP	
2js	PsbM	
xkb	PsbO	
2jn	PsbQ	
2am	PsbW superfamily	
abo	PsbW superfamily	
2fd	Psb29	
-	PetA,B,D,G,L,M,N	
xcc	petC	
2ai	petJ/ cytochrome c6	
xmc	CPLD51 protein required for cyt b6 assembly	
-	PsaA,B,C,D,E,F,I,J,L,M	
xqz	PSI subunit 223993351	
2dv	Ferredoxin 1	
xlg	Ferredoxin 2	
xnv	Ferredoxin 3	
2lt	Ferredoxin rieske component	
xdk	Ferredoxin-NADP oxidoreductase 1	
2do	Ferredoxin-NADP oxidoreductase 2	
xln	Plastoquinol terminal oxidase	
xla	Photosystem II assembly factor Hcf136	
xqz	PsaO	
2bh	PGR5 protein	
xpd	PGR5-like protein	
-	atpA,B,D,E,F,G,H,I	
xkk	atpC	
3. Central carbon metabolism		
a) CBB cycle		
-	Rubisco large subunit	
-	Rubisco small subunit	
2eg	Rubisco small subunit N-methyltransferase I	
xms	3-phosphoglycerate kinase	
2kn	Glyceraldehyde 3-phosphate dehydrogenase	
xip	Triosephosphate isomerase	
2mh	Fructose-bisphosphate aldolase	
-	Sedoheptulose bisphosphatase	
xic	Fructose-1,6-bisphosphatase 1	
2jl	Fructose-1,6-bisphosphatase 2	
5aa	Fructose-1,6-bisphosphatase 3	
xji	Transketolase	
xkt	Ribose-5-phosphate isomerase	
2jj	D-ribulose-5-phosphate 3-epimerase	
acq	Phosphoribulokinase 1	
xfl	Phosphoribulokinase 2	
b) Glycolysis/ gluconeogenesis		
3bo	Phosphoglycerate mutase 1	
2je	Phosphoglycerate mutase 2	
2ko	Enolase	
xlr	Pyruvate kinase	
2dd	Pyruvate dehydrogenase	
xmt	Dihydrolipoamide dehydrogenase	
xan	Dihydrolipoamide acetyltransferase	
2hk	Pyruvate carboxylase	
xlj	Pyruvate phosphate dikinase	
2is	Pyrophosphate-dependent phosphofructo-1-kinase	
xnl	Phosphoglucomutase	
xdv	Beta-glucan synthase	
4ay	Glucan 1,3-beta-glucosidase	
xru	Hexose and triose phosphate transporter	

Cluster	Enzyme	Origin
4. Fatty acid biosynthesis		
a) Fatty acid synthesis		
xlv	Long-chain acyl-CoA transporter	
xjx	Long-chain acyl-CoA synthetase	
xpy	Acetyl-coA:carboxylase	
xlb	Malonyl-CoA:ACP transacylase	
xph	Beta-ketoacyl synthase	
abx	Beta-ketoacyl-ACP reductase	
xik	Enoyl: ACP reductase	
2ig	Long chain fatty acid elongase 1	
2qi	Long chain fatty acid elongase 2	
2ge	Fatty acid desaturase 1	
2iu	Fatty acid desaturase 2	
2jh	Fatty acid desaturase 3	
b) Glycerol metabolism		
2ky	Glycerol-3-phosphate dehydrogenase	
2kn	Glyceraldehyde 3-phosphate dehydrogenase	
5. Tetrapyrrole biosynthesis		
a) Common branch		
2ol	Glutamyl-tRNA synthetase	
aai	Glutamyl-tRNA reductase	
xkn	Glutamate-1-semialdehyde 2,1-aminomutase (GSA)	
xin	Delta-aminolevulinic acid dehydratase	
xjr	Porphobilinogen deaminase	
2be	Uroporphyrin III synthase	
2nc	Uroporphyrinogen decarboxylase 1	
xjh	Uroporphyrinogen decarboxylase 2	
xkm	Uroporphyrinogen decarboxylase 3	
xgl	Coproporphyrinogen III oxidase 1	
xls	Coproporphyrinogen III oxidase 2	
4gz	Protoporphyrinogen oxidase	
b) Chlorophyll branch		
xnp	Magnesium chelatase subunit D	
xks	Magnesium chelatase subunit H	
-	Magnesium chelatase subunit I	
xiw	Magnesium-PPIX methyltransferase	
-	Magnesium-PPIX methylmonoester cyclase	
2kl	Protochlorophyllide reductase A	
xmw	3,8-divinyl protochlorophyllide a 8-vinyl reductase	
xke	Chlorophyll synthetase	
c) Haem branch		
xjs	Ferrochelatase	
2kr	haem oxygenase 1	
2ks	haem oxygenase 2	
xjj	haem transporter	
d) Catabolism		
xcv	Pheophytinase	
2gw	Pheophorbide a oxidase	
6. Carotenoid biosynthesis		
2mc	Deoxyxylulose-5-phosphate synthase	
xif	1-deoxy-D-xylulose 5-phosphate reductoisomerase	
xlf	4-diphosphocytidyl-2C-methyl-D-erythritol synthase	
xob	4-diphosphocytidyl-2C-methyl-D-erythritol kinase	
xos	2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	
2bp	1-hydroxy-2-methyl-2-(E)-butenyl-4-diphosphate synthase	
xjf	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	
xjd	Geranylgeranyl pyrophosphate synthase 1	
xaf	Geranylgeranyl pyrophosphate synthase 2	
-	Isopentenyl diphosphate isomerase	
2oz	Phytoene synthase	
2cb	Phytoene desaturase 1	
xjy	Phytoene desaturase 2	
xhm	Zeta-carotene isomerase	
2kz	Zeta-carotene desaturase	
2dc	Beta-carotene isomerase	
2kx	Lycopene beta cyclase	
xeq	Violaxanthin de-epoxidase	
xjg	Zeaxanthin epoxidase	
7. Iron-sulphur cluster biosynthesis		
-	SufB	
-	SufC	
2hx	FeS assembly protein SufD	
xjm	Cysteine desulphurase NFS1	
xki	sulphite reductase (ferredoxin) 1	
xra	sulphite reductase (ferredoxin) 2	
xlo	Fe-S cluster biosynthesis protein ISA1	
xgg	Mitochondrial Fe-S cluster biosynthesis protein ISA2	
2dy	NifU protein 1	
2oq	NifU protein 2	

Cluster	Enzyme	Origin
8. Riboflavin biosynthesis		
2oe	4-dihydroxy-2-butanone 4-phosphate synthase	
2oe	GTP cyclohydrolase	
4hc	Riboflavin synthase	
9. Glutamate/ glutamine/ aspartate/ lysine biosynthesis		
a) Glutamine branch		
2bx	Pyruvate transporter	
2ct	Bicarbonate transporter	
2hk	Pyruvate carboxylase	
2mi	Aspartate aminotransferase	
2jb	Glutamine synthetase	
xso	Kynurenine aminotransferase	
2jk	Glutamate synthase	
b) Aspartate branch		
xlk	Aspartate kinase	
2cy	Aspartate-semialdehyde dehydrogenase 1	
xgf	Aspartate semialdehyde dehydrogenase 2	
2br	Dihydropicolinate synthase	
4ba	Homoserine dehydrogenase	
xis	Dihydrodipicolinate reductase	
xlw	Diaminopimelate aminotransferase	
xoi	Diaminopimelate epimerase	
xkz	Diaminopimelate decarboxylase	
10. Aromatic amino acid biosynthesis		
a) Chorismate branch		
xmk	DAHP synthetase	
-	3-dehydroquinate synthase	
xjv	3-dehydroquinate reductase/ Shikimate dehydroge	
-	Shikimate kinase	
xiq	EPSP synthase	
xsl	Chorismate synthase	
b) Phenylalanine/Tyrosine branch		
9af	Chorismate mutase	
2mi	Aspartate aminotransferase	
xso	Kynurenine/ phenylpyruvate aminotransferase	
xsg	Prephenate dehydrogenase	
c) Tryptophan branch		
xhr	Anthranilate synthase	
xld	Anthranilate phosphoribosyltransferase	
2qg	Phosphoribosylanthranilate isomerase/ indole 3-gly	
abs	Tryptophan synthase alpha	
xfu	Tryptophan synthase beta	
11. Branched chain amino acid biosynthesis		
a) Valine/ Isoleucine branch		
-	Acetolactate synthase	
xmy	Keto-acid reductoisomerase	
xkg	Dihydroxy-acid dehydratase	
xiz	Branched-chain amino acid aminotransferase II	
b) Leucine branch		
-	Isopropylmalate synthase	
xmm	3-isopropylmalate isomerase	
xkq	3-isopropylmalate dehydrogenase	
xiz	Branched-chain amino acid aminotransferase II	
12. Serine and cysteine biosynthesis		
a) Serine branch		
xhx	Serine hydroxymethyltransferase	
b) Cysteine branch		
xly	ATP sulphurylase (sulphate adenylyltransferase)	
-	Adenosine sulphate kinase	
xiv	Phosphoadenosine phosphosulphate reductase 1/	
xkd	Phosphoadenosine phosphosulphate reductase 2	
xki	sulphite reductase (ferredoxin) 1	
xra	sulphite reductase (ferredoxin) 2	
xct	Cysteine synthase	
xgx	Serine O-acetyltransferase	

Cluster	Enzyme	Origin
13. Aminoacyl tRNA synthetases		
xhe	Alanyl-tRNA synthetase	
2mg	Cysteinyl-tRNA synthetase 1	
2cw	Cysteinyl-tRNA synthetase 2	
2np	Aspartyl-tRNA synthetase	
2ol	Glutamyl-tRNA synthetase	
2mr	Phenylalanyl-tRNA synthetase 1	
2lx	Phenylyalanyl-tRNA synthetase 2	
2ja	Glycyl-tRNA synthetase	
xmg	Histidyl-tRNA synthetase	
2qe	Isoleucyl-tRNA synthetase	
9aa	Lysyl-tRNA synthetase	
2kq	Leucyl-tRNA synthetase	
xdl	Methionyl-tRNA synthetase	
xqu	Asparaginyl-tRNA synthetase	
xsa	Prolyl-tRNA synthetase	
xho	Glutaminyl-tRNA synthetase	
9ac	Arginyl-tRNA synthetase	
xhf	Seryl-tRNA synthetase	
xiu	Threonyl-tRNA synthetase	
xky	Valyl-tRNA synthetase	
xhg	Tryptophanyl-tRNA synthetase	
xfs	Tyrosyl-tRNA synthetase	
14. Nucleotide synthesis and import		
2cs	Adenylate kinase	
xnf	Guanylate kinase	
2ow	UMP-CMP kinase	
2hu	Nucleotide triposphate transporter 1	
15. Ribosome		
	rps2-14, rps16-20	
	rpl1-6, rpl11-16, rpl18-24, rpl27, rpl29, rpl31-36	
xjb	rps1A	
xih	rps1B	
-	rps15	
2mt	rpl9	
xix	rpl10	
xhv	rpl17	
xju	rpl28	
xmh	rps30A	
xid	rps21	
16. Translation initiation		
xjt	Translation initiation factor 1	
xjc	Translation initiation factor 2	
4ac	Translation initiation factor 3	
-	Translation elongation factor EF-Tu	
-	Translation elongation factor EF-Ts	
xqr	Translation elongation factor P	
xjl	Translation elongation factor G	
2qq	Ribosome release factor	
xqf	Ribosome recycling factor	
17. Plastid protein import		
-	secA, G, Y	
-	tatC	
2eq	TatA/B	
2qs	Signal peptidase complex subunit Srp12	
xdo	Signal peptidase complex subunit Srp22	
2fq	Signal peptidase coimplex subunit FtsY	
2jv	Signal recognition particle, subunit Srp54	
2ce	Tic20	
2ly	Tic21	
2iz	Tic110	
18. Plastid division		
xia	Cell division protein FtsH	
3ap	Cell division protein FtsZ	
xhc	Plastid division protein minD	
2dk	Plastid division protein minE	
19. Clp protease		
xit	Chaperone protein ClpA	
2qd	Chaperone protein ClpB	
2ms	Adaptor protein ClpS 1	
xtf	Adaptor protein ClpS 2	
xsp	Adaptor protein ClpS 3	
-	Proteolytic subunit ClpC	
xme	Proteolytic subunit ClpP 1	
2cx	Proteolytic subunit ClpP 2	
xlh	Proteolytic subunit ClpP 3	

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	Hypogyristea 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 green lineage	Tanaka and Tanaka 2007
Isopentenyl diphosphate isomerase	Carotenoid biosynthesis	Not found	Dispensible 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

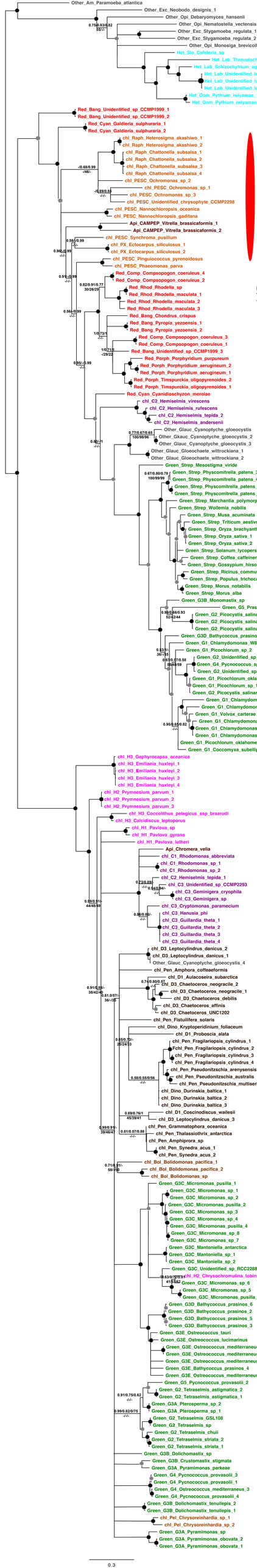


Fig 5 - figure supplement 3. Tree of ochrophyte sedoheptulose-7-bisphosphatase sequences.

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 hypogyristea and diatoms, of probable green algal origin- are shown respectively by red and green ellipses.

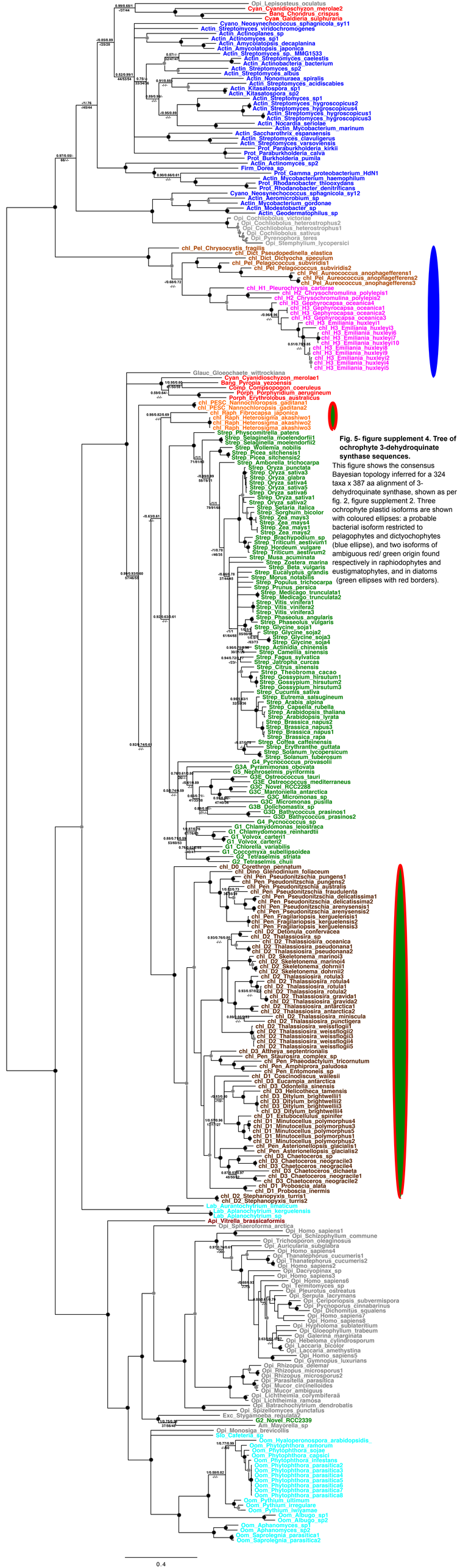


Fig. 5- figure supplement 4. Tree of ochrophyte 3-dehydroquinate synthase sequences.

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).

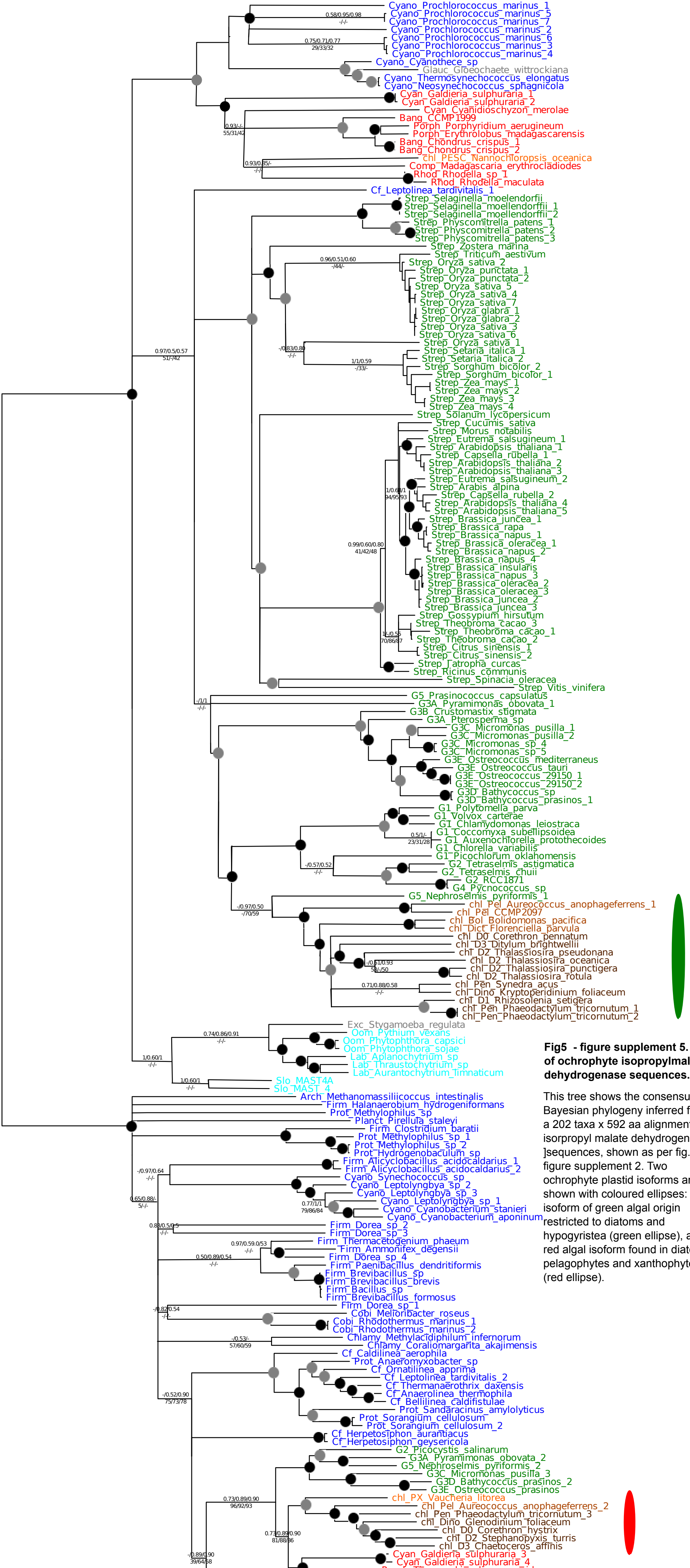


Fig5 - figure supplement 5. Tree of ochrophyte isopropylmalate dehydrogenase sequences.

This tree shows the consensus Bayesian phylogeny inferred for a 202 taxa x 592 aa alignment of isorpropyl 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 hypogyristea (green ellipse), and a red algal isoform found in diatoms, pelagophytes and xanthophytes (red ellipse).

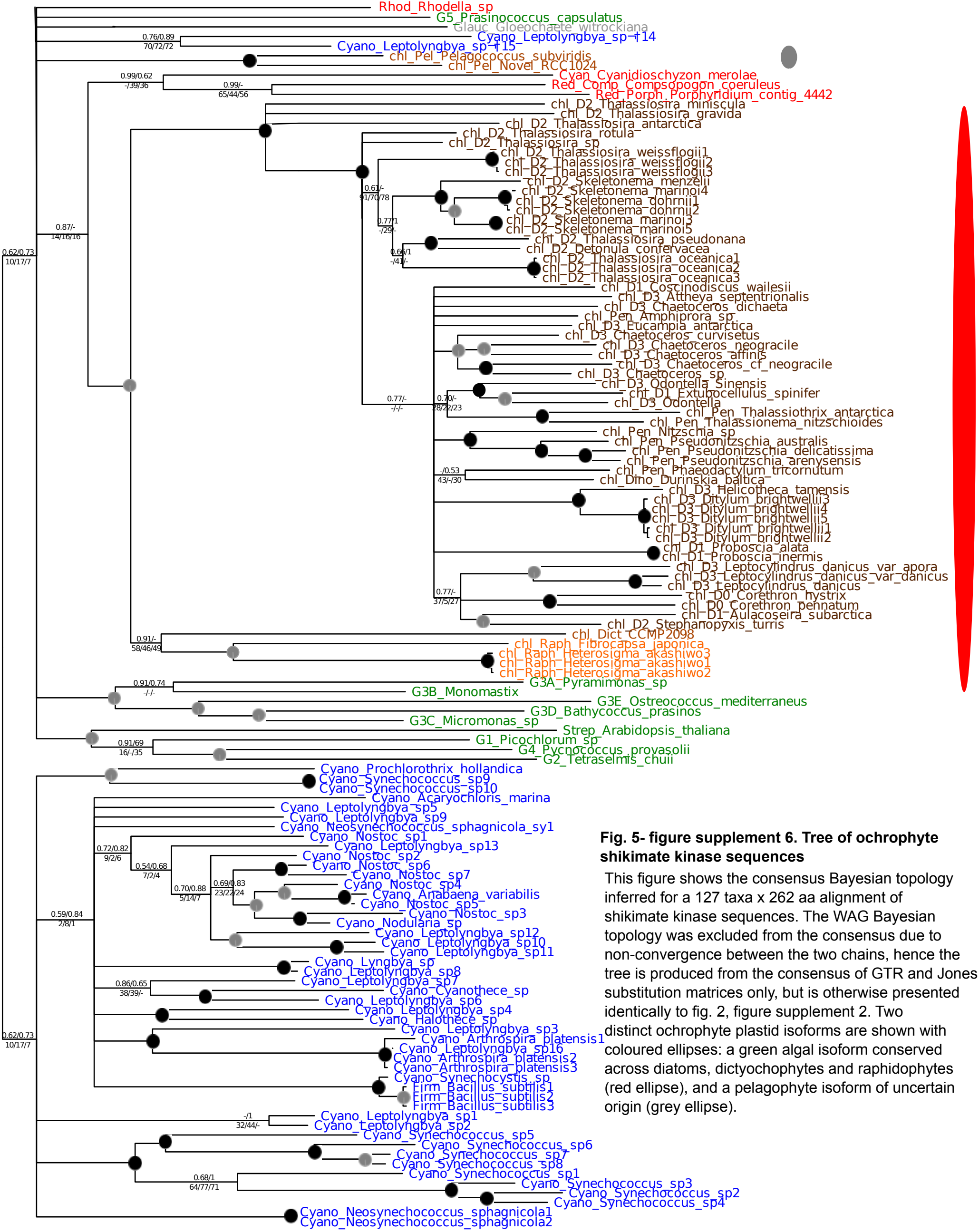


Fig. 5- figure supplement 6. Tree of ochrophyte shikimate kinase sequences

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.

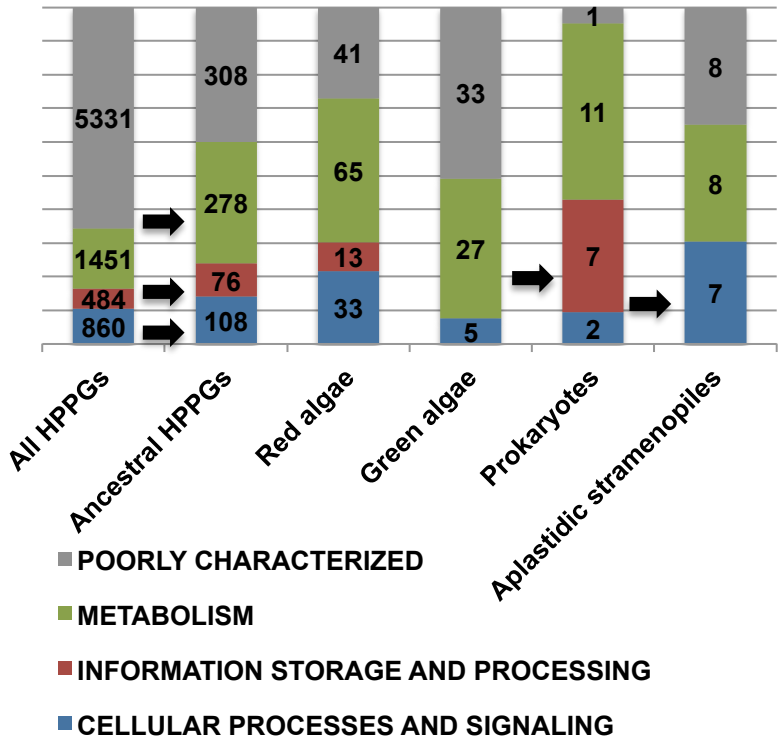
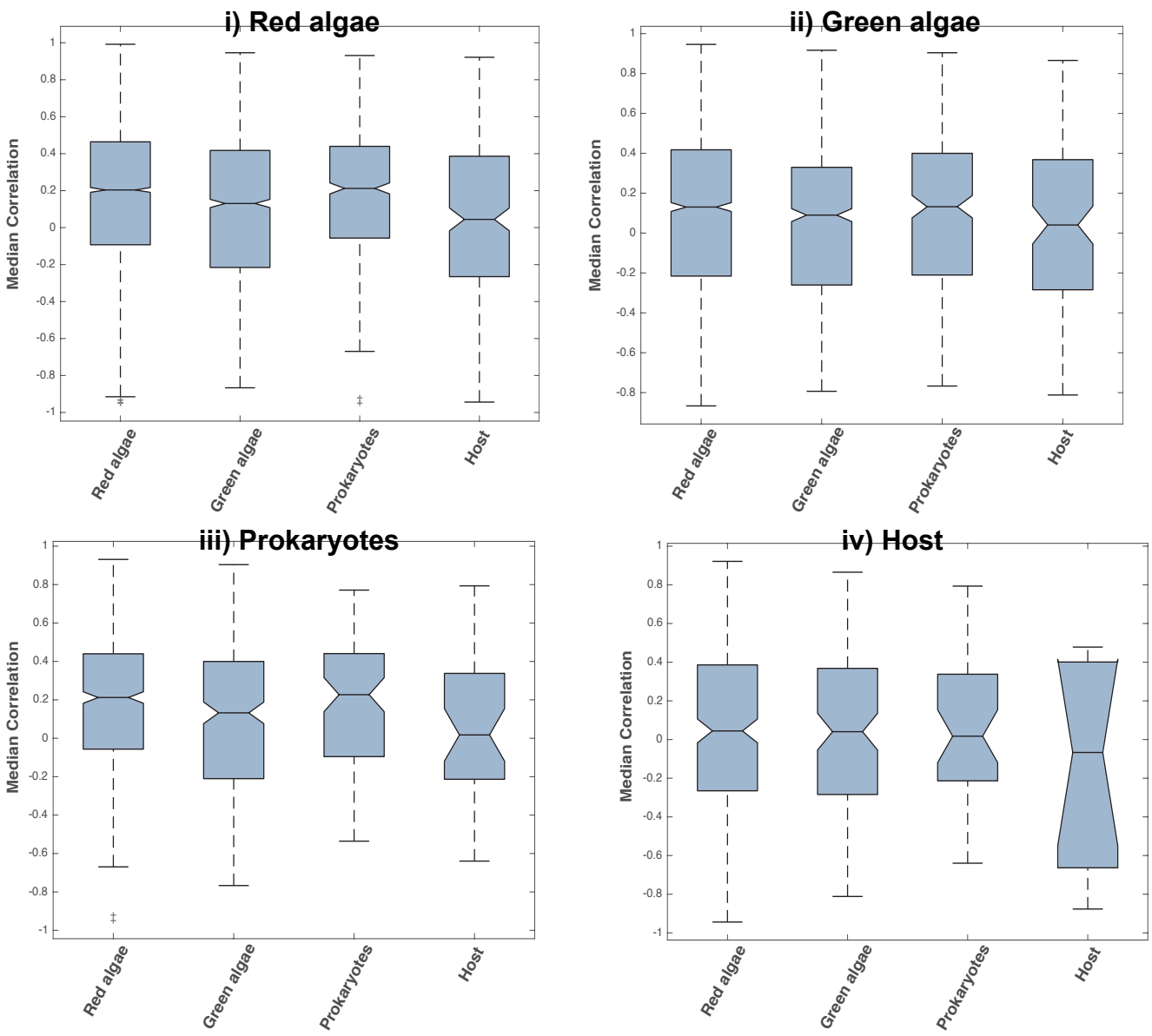


Fig. 5- figure supplement 8. Coregulation of genes incorporated into HPPGs of different origin in the model diatom *Phaeodactylum tricornutum*. Panel A shows boxplots of the correlation coefficients between the expression profiles of genes encoding members of ancestral HPPGs of red algal origin **(i)**, 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.

A

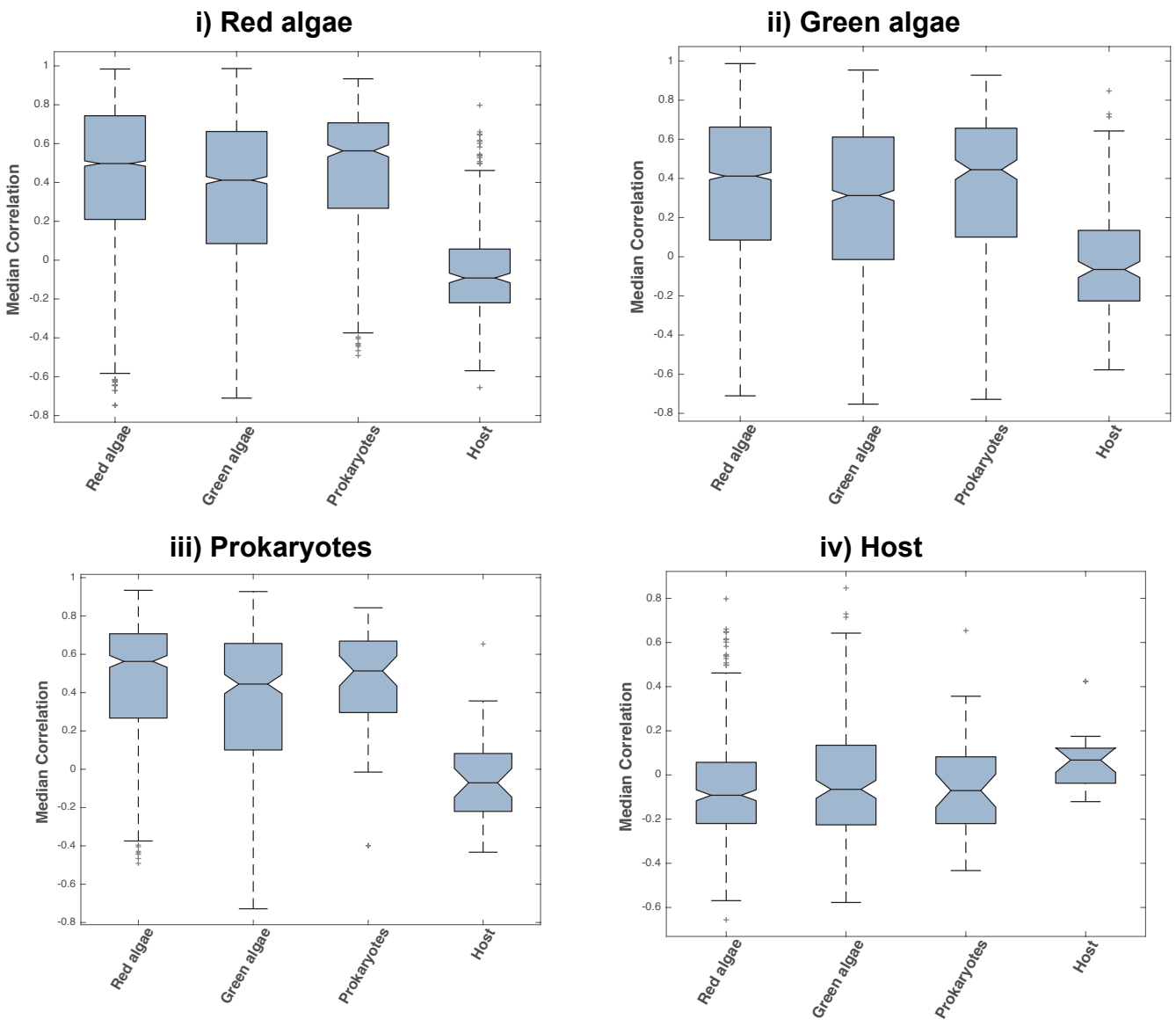


B

	Red	Green	Prokaryotic	Host
Red		0.393	-0.945	0.491
Green	-0.555		-0.780	0.905
Prokaryotic	-0.358	-0.432		0.564
Host	-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
Red		0.296	-0.833	0.005
Green	-0.376		-0.564	0.093
Prokaryotic	0.279	0.473		0.019
Host	-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.

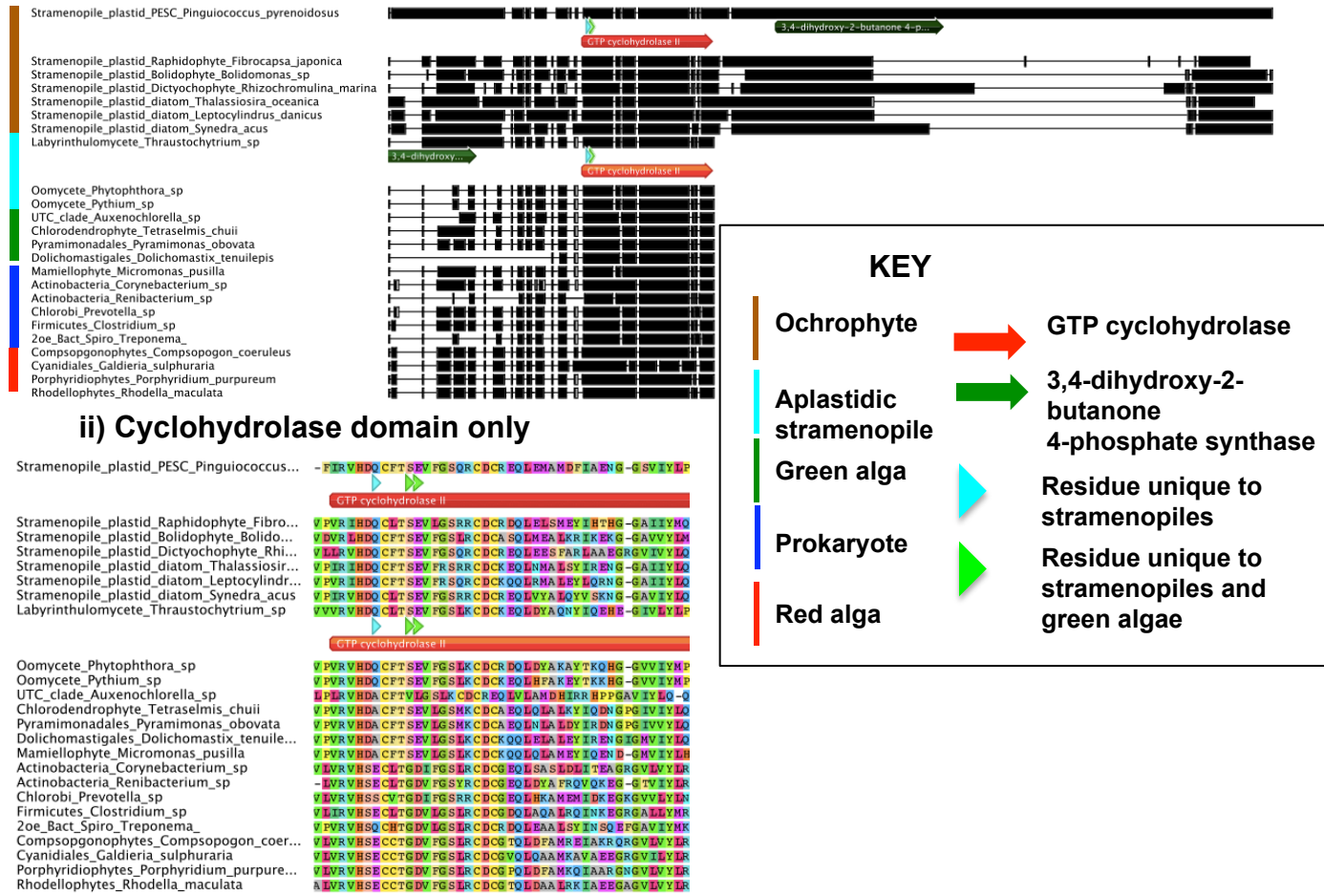


Fig. 6- figure supplement 2. Origins of ochrophyte plastid 3,4-dihydroxy-2-butanone 4-phosphate synthase. This figure shows the consensus Bayesian topology inferred for a 22 taxa x 206 aa alignment of 3,4-dihydroxy-2-butanone 4-phosphate synthase domains from different lineages, inferred using Jones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

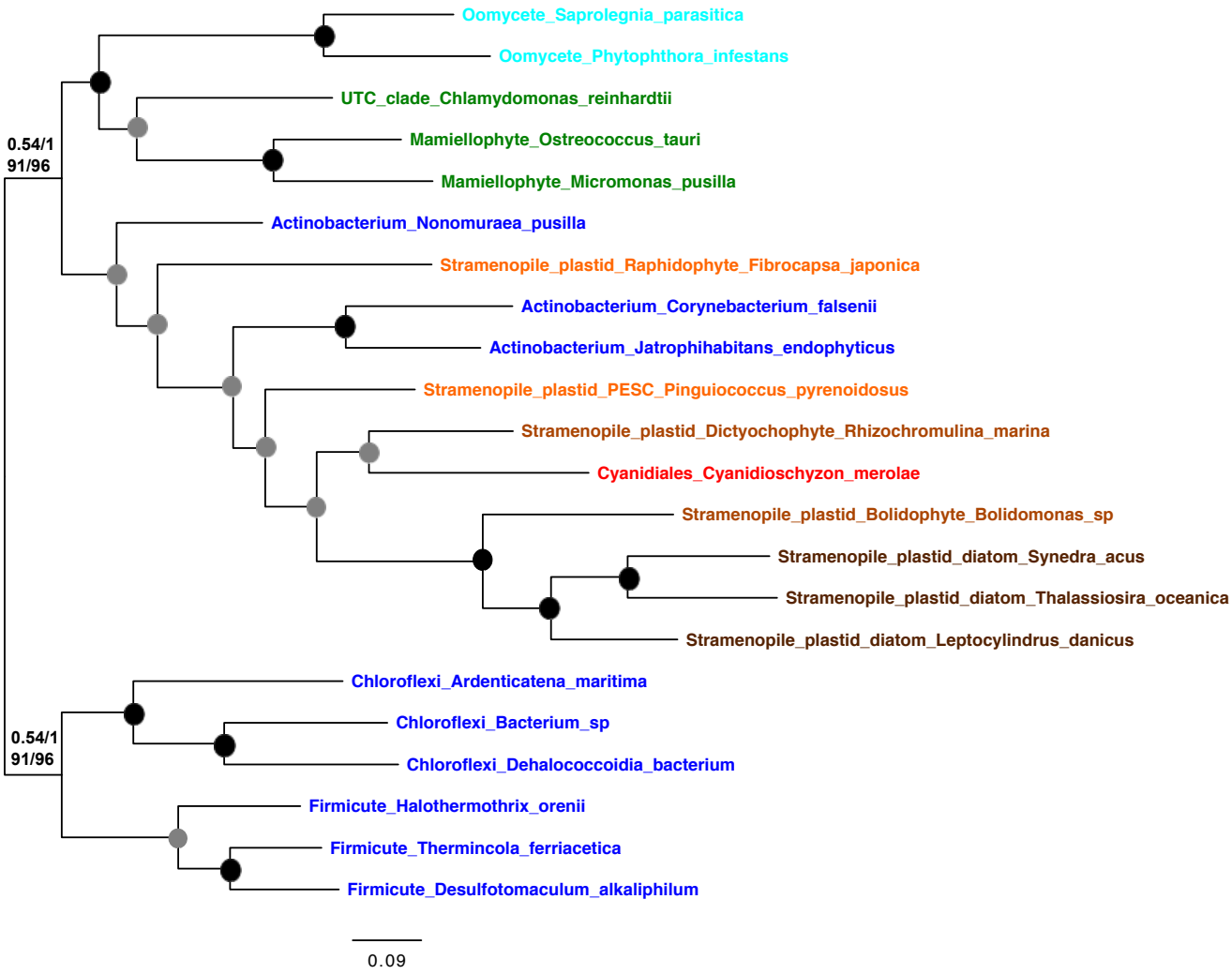


Fig. 6- figure supplement 3. An ochrophyte-specific Tic20 fusion protein. This figure shows alignments of the full length (**i**) and conserved region only (**ii**) of plastid Tic20 sequences, displayed as per figure supplement 9.

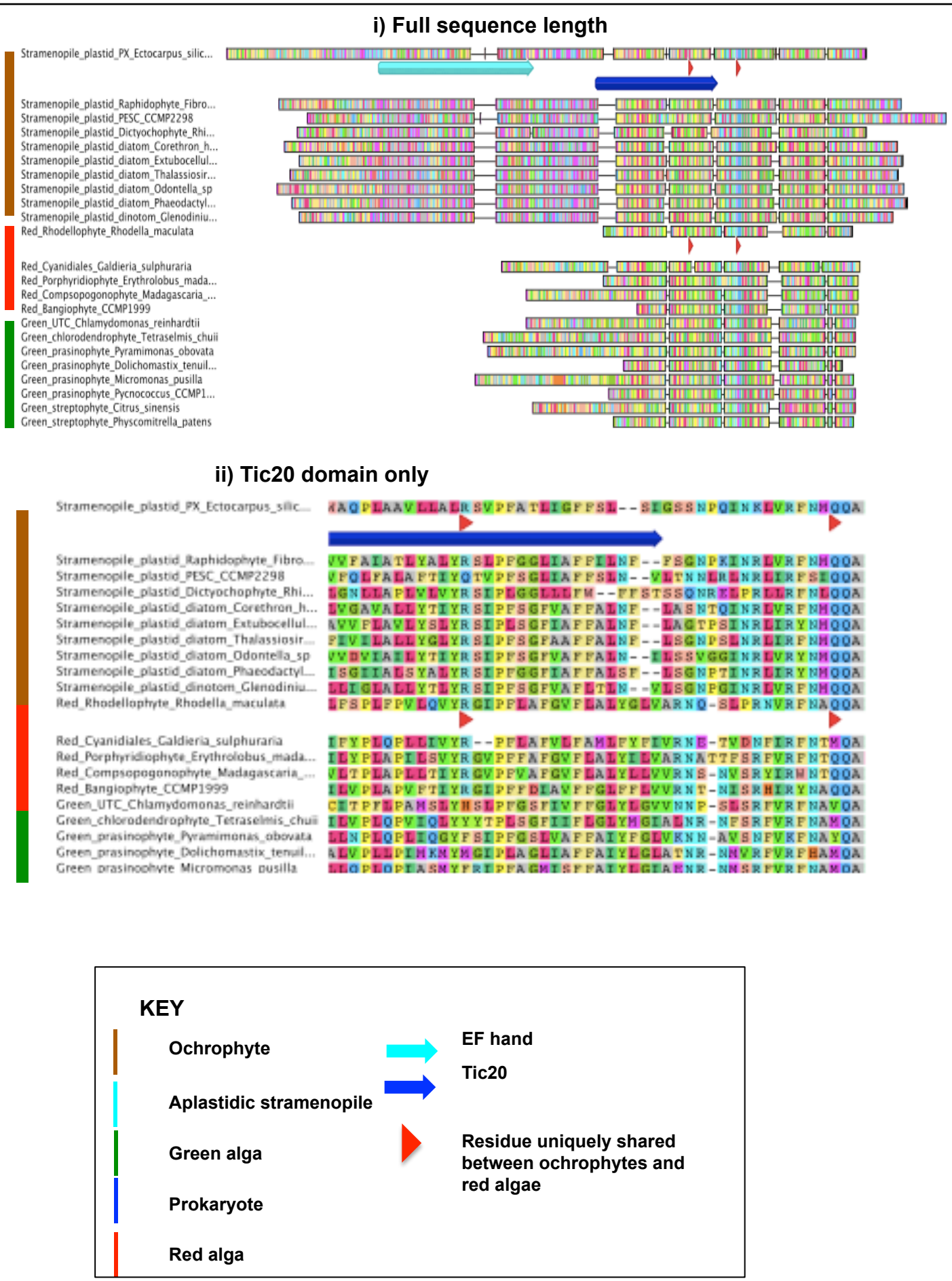
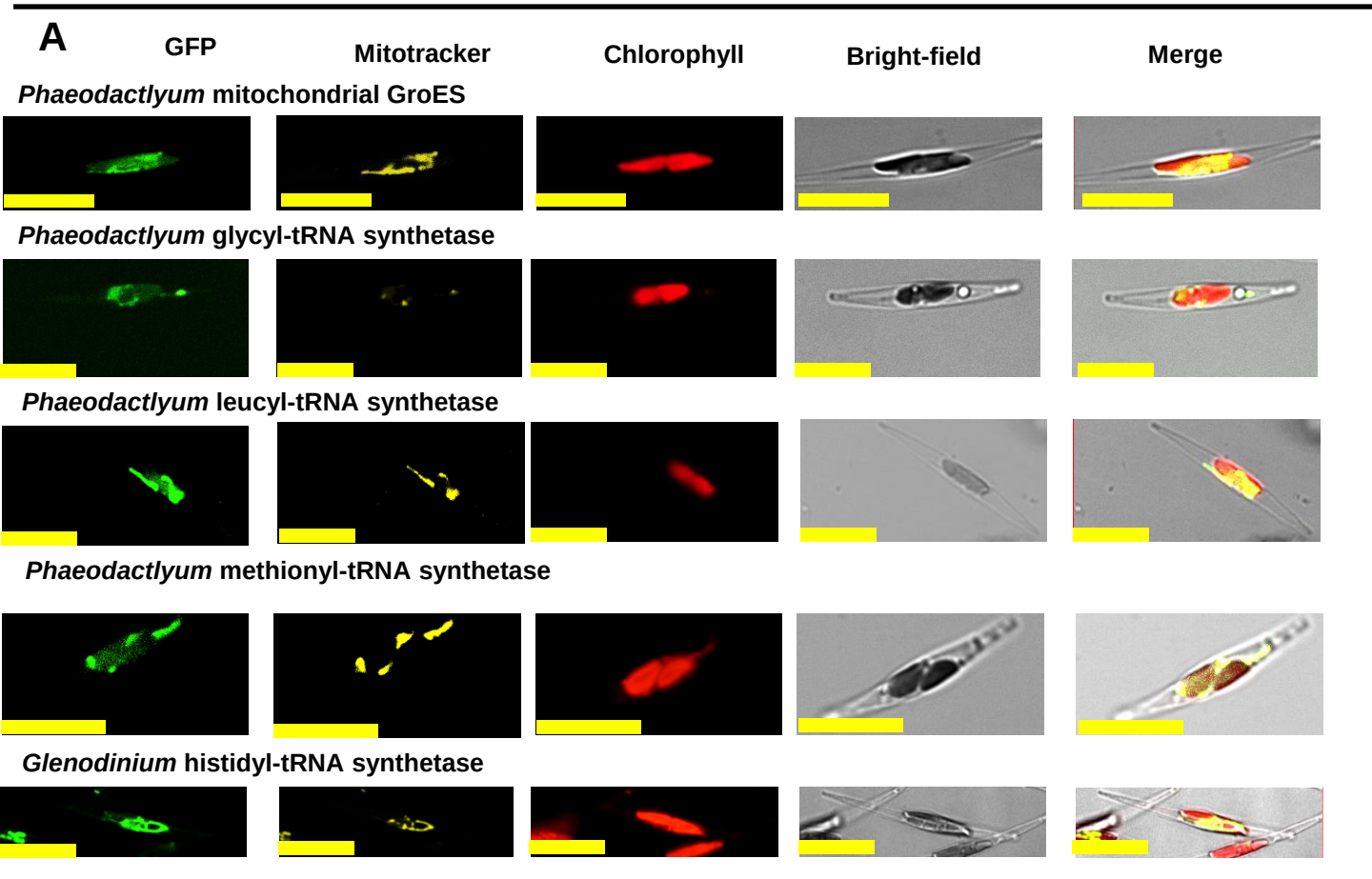


Fig. 7- figure supplement 1. Experimental verification of additional ochrophyte dual-targeted proteins. Panel A shows Mitotracker-orange stained *Phaeodactylum tricornutum* lines expressing four additional dual-targeted proteins (glycyl-, leucyl-, and methionyl-tRNA synthetases, and a predicted mitochondrial 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.



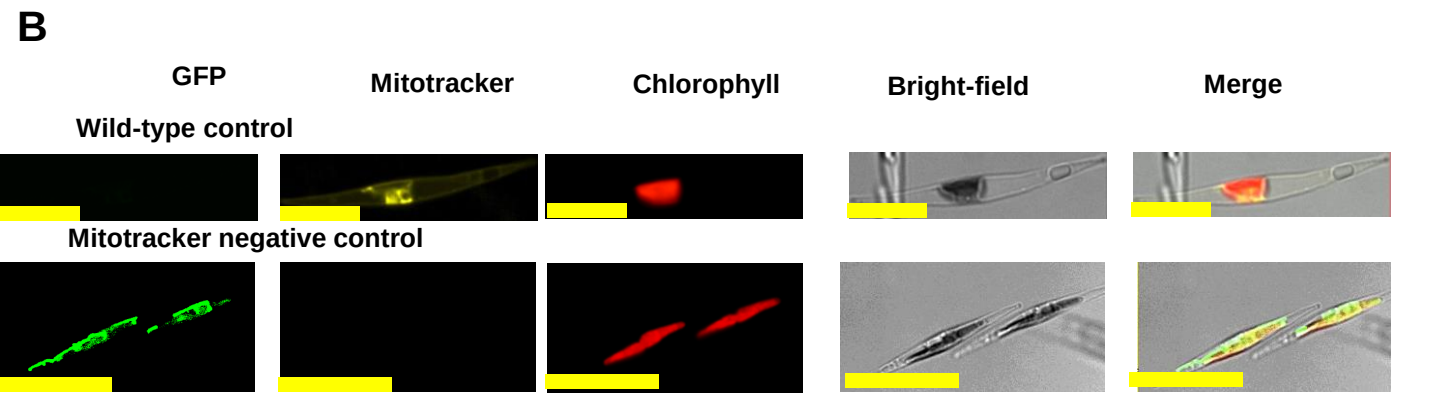
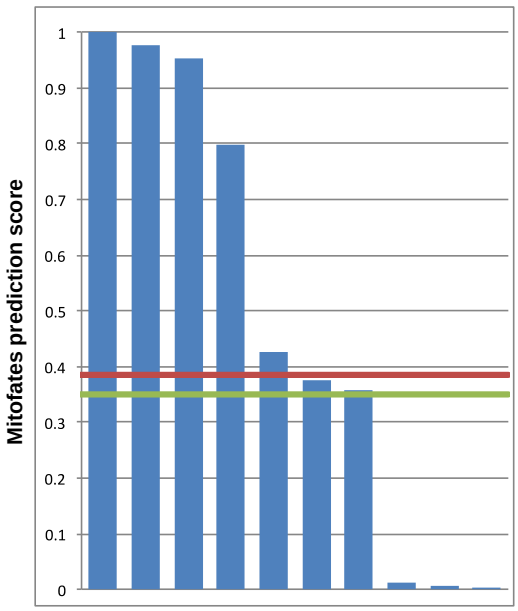
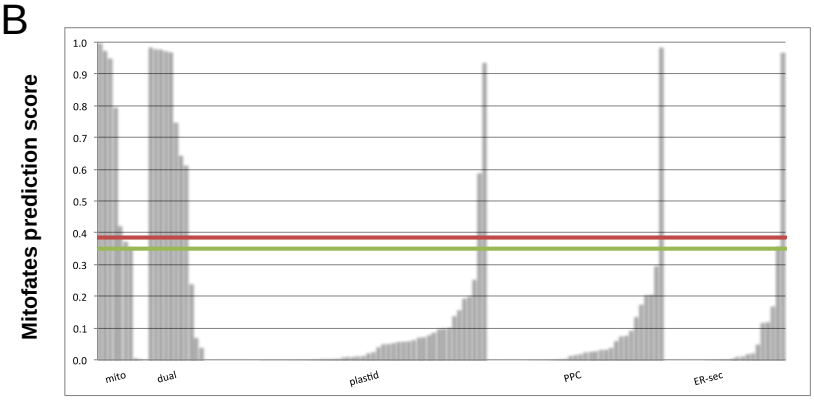


Fig. 7- figure supplement 2. Comparison of different in silico targeting prediction programmes for the identification of dual-targeted ochrophyte proteins. Panel A shows Mitofates scores for ochrophyte proteins verified experimentally to be dual targeted in this and a previous study[9]. **Panel B** shows 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

	% predicted mitochondrially targeted			
validated localization	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 hypogyristea); 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 hypogyristea.

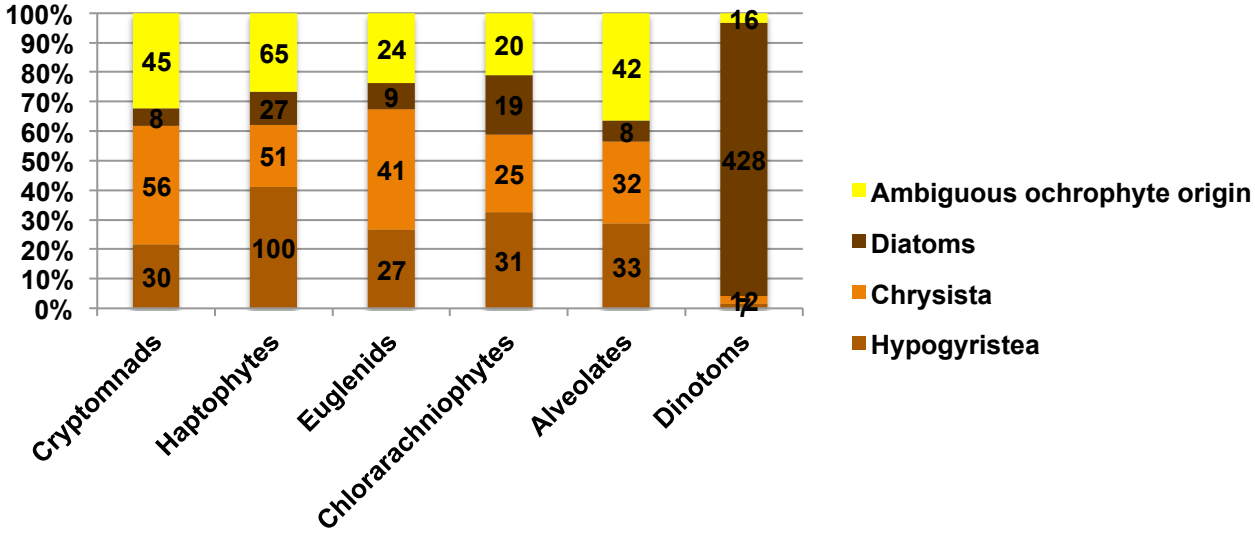


Fig.8- figure supplement 2. Heatmaps of nearest sister-groups to haptophytes in ancestral ochrophyte HPPG trees. This figure shows the specific ochrophyte lineages implicated in the origin of haptophyte plastid-targeted proteins, as inferred from the nearest ochrophyte sister-groups to haptophytes in 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.

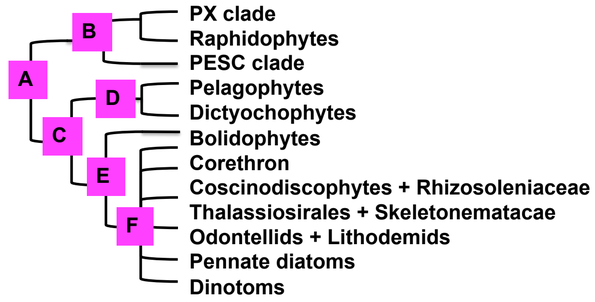


Fig. 8- figure supplement 3. Internal evolutionary affinities of haptophyte plastid-targeted proteins incorporated into ancestral ochrophyte HPPGs. This figure profiles the evolutionary origins of haptophyte plastid-targeted proteins incorporated into ancestral ochrophyte HPPGs by BLAST top hit analysis. 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.

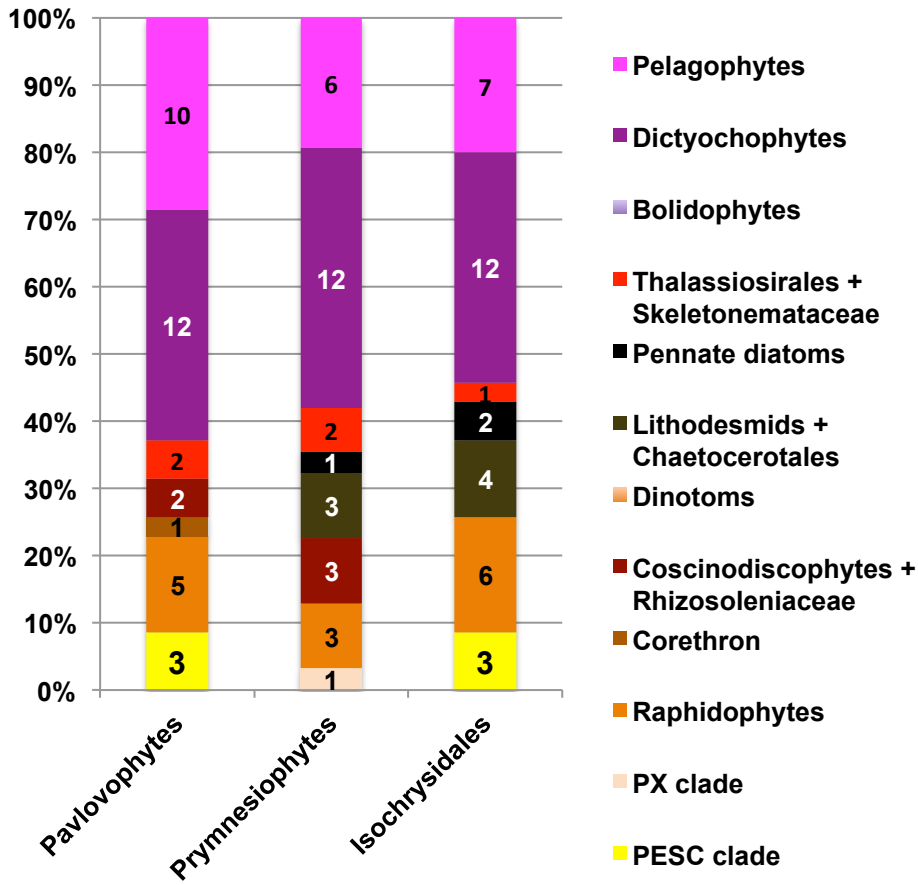


Fig. 8- figure supplement 4. Evidence for gene transfer from pelagophytes and dictyochyophytes into haptophytes. Panel A shows the next deepest sister groups identified for haptophyte proteins of hypogyristean origin in single-gene trees. The pie chart **(i)** compares the number of single-gene trees in which the combined clade of hatpohyte and hypogyristean 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/ hypogyristean 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 hypogyristea 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 hypogyristea, and subsequently transferred to haptophytes.

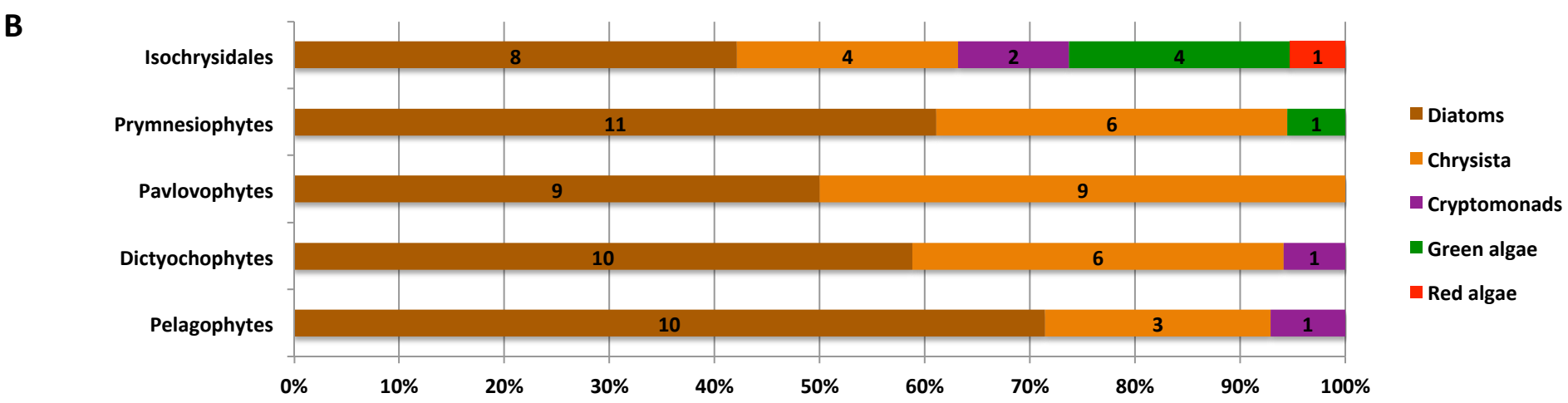


Fig. 8- figure supplement 5. Earliest possible origin points of uniquely conserved sites in haptophyte plastid-targeted proteins. This figure shows the total number of residues that are uniquely shared between a 37 proteins that have clearly been transferred between the ochrophytes and haptophytes, and 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 hypogyristea and diatoms, consistent with the gene transfer having occurred from an ancestor of the pelagophytes and dictyochophytes into the haptophytes, rather than the converse.

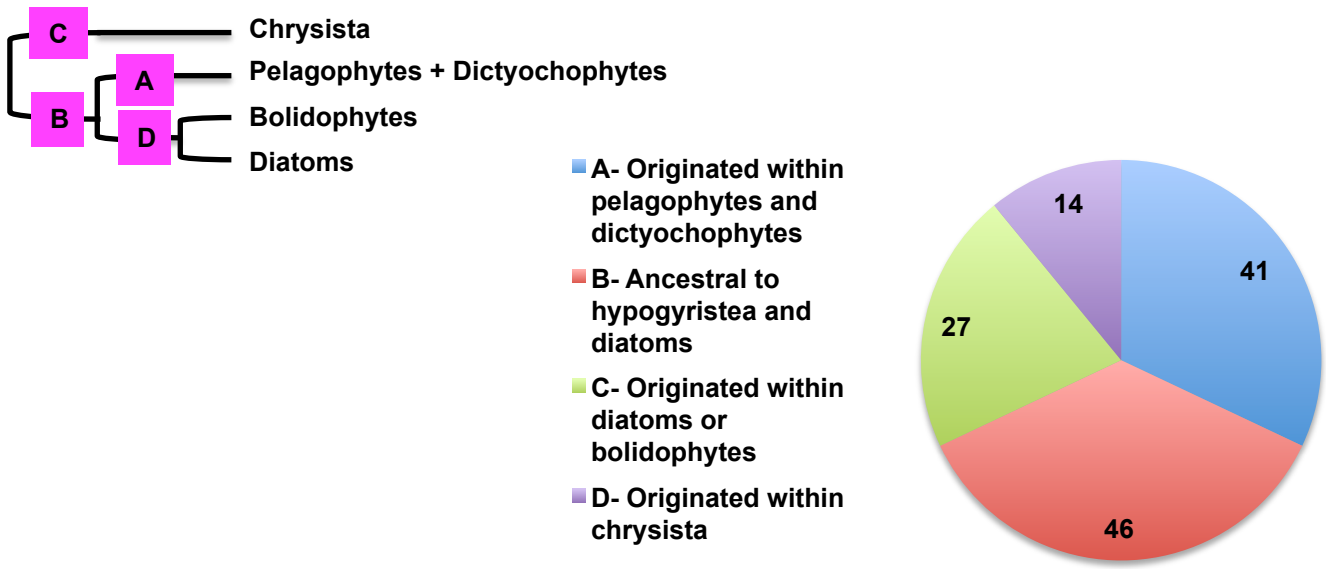


Fig. 8- figure supplement 6. Evolutionary origin of ancestral haptophyte genes. This figure shows the most likely evolutionary origin assigned by BLASt top hit analysis to the 12728 conserved gene families inferred to have been present in the last common haptophyte ancestor.

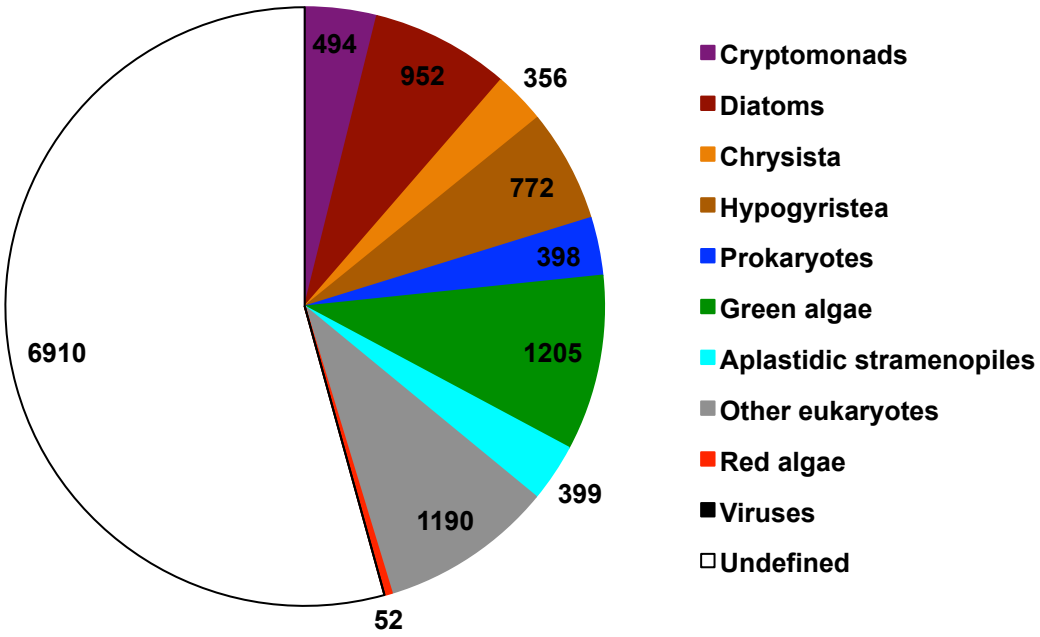
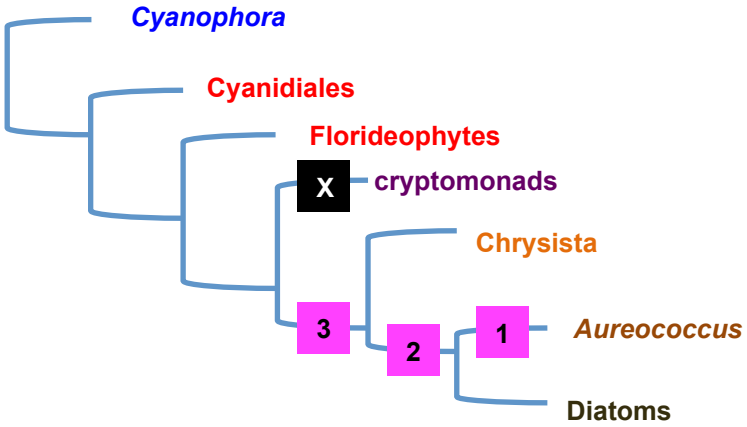


Fig. 9- figure supplement 1. Alternative topology tests of plastid genome trees. Tests were performed with the RAxML + JTT trees inferred for the gene-rich (**panel A**) and taxon-rich (**panel B**) plastid-encoded protein alignments. In each case, a schematic diagram of the tree topology obtained is 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.

A

i)

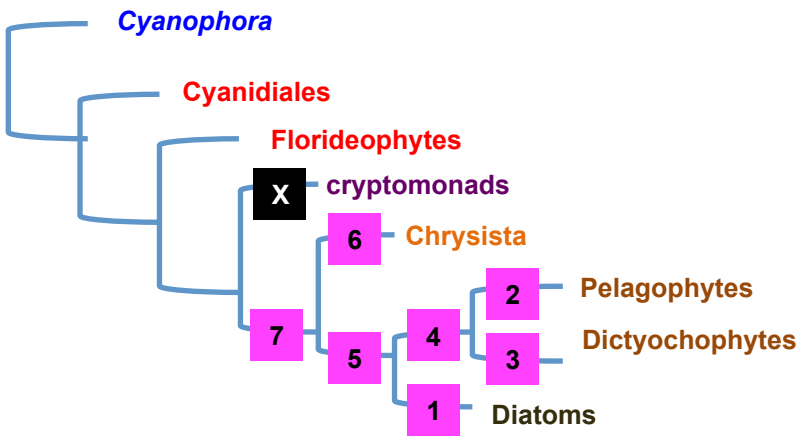


ii)

	haptophytes sister to	AU	NP	BP	PP	KH	SH	WKH	WSH
1	***Aureococcus***	0.004	8E-05	0	0	0	0	0	0
2	***Aureococcus* and diatoms**	0.001	6E-05	0	0	0	0	0	0
3	**All ochrophytes**	**0.057**	0.049	0.05	2E-23	**0.051**	**0.385**	**0.051**	**0.137**

B

i)



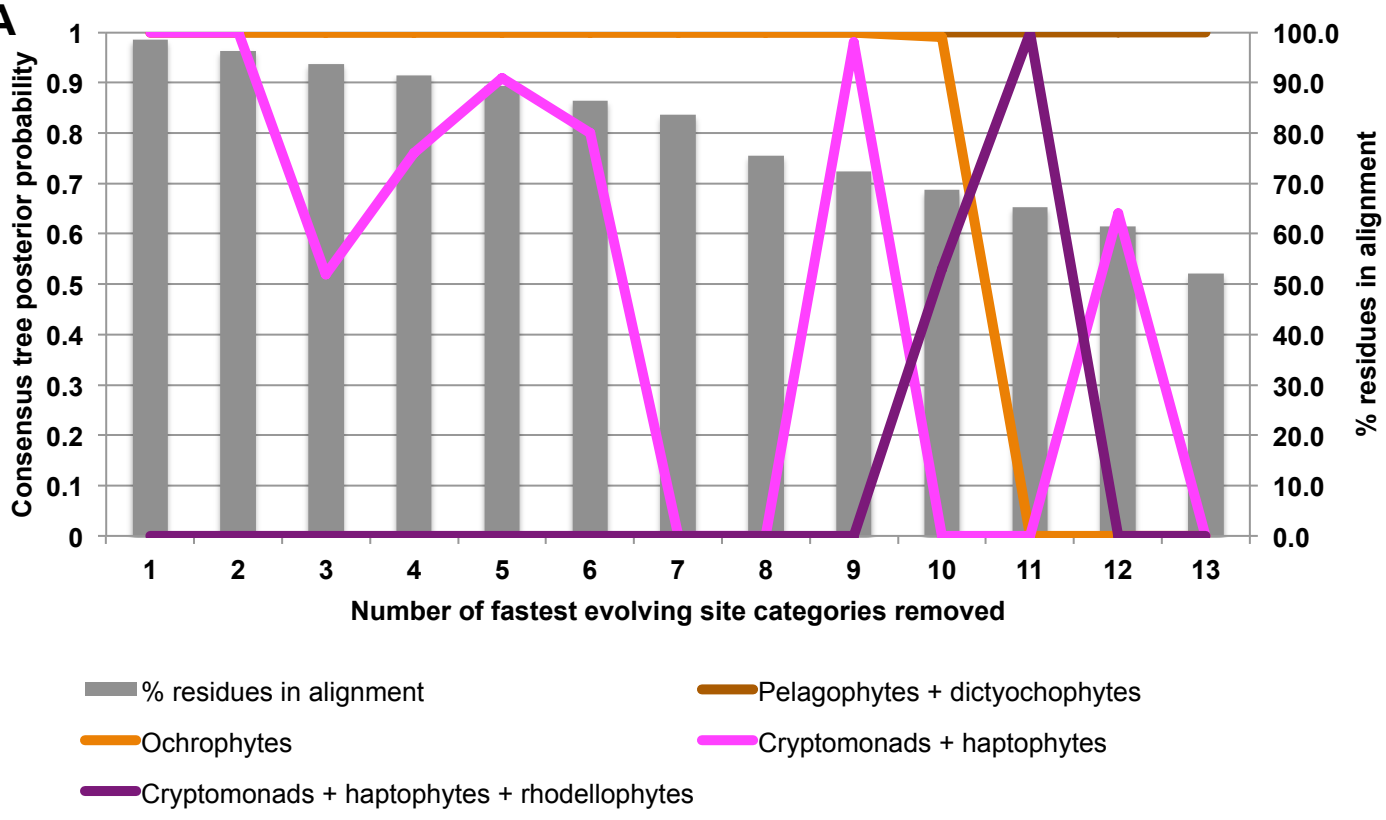
ii)

	haptophytes sister to	AU	NP	BP	PP	KH	SH	WKH	WSH
1	**Diatoms**	1E-95	3E-24	0	1E-94	0	0	0	0
2	**Pelagophytes**	5E-05	7E-06	0	1E-109	0	0	0	0
3	**Dictyochophytes**	6E-47	1E-16	0	5E-111	0	0	0	0
4	**Hypogyristea**	2E-06	2E-06	0	1E-106	0	0	0	0
5	**Hypogyristea + diatoms**	7E-75	2E-21	0	2E-99	9E-05	9E-05	0	0
6	**Chrysista**	2E-38	6E-15	0	1E-94	8E-05	8E-05	0	0
7	**All ochrophytes**	**0.423**	**0.418**	**0.419**	3E-04	**0.414**	**0.812**	**0.414**	**0.808**

- **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 hypogyristean 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.

A



B

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

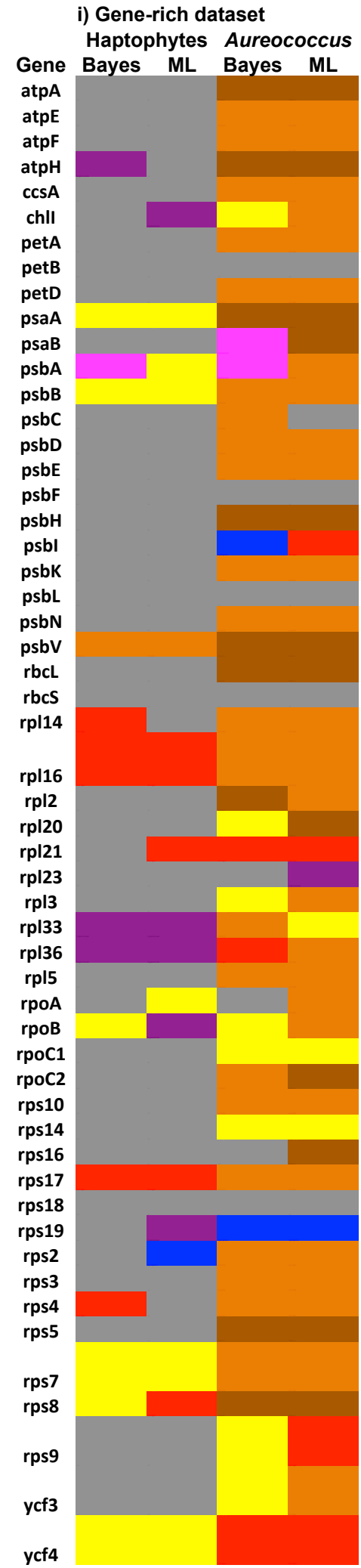
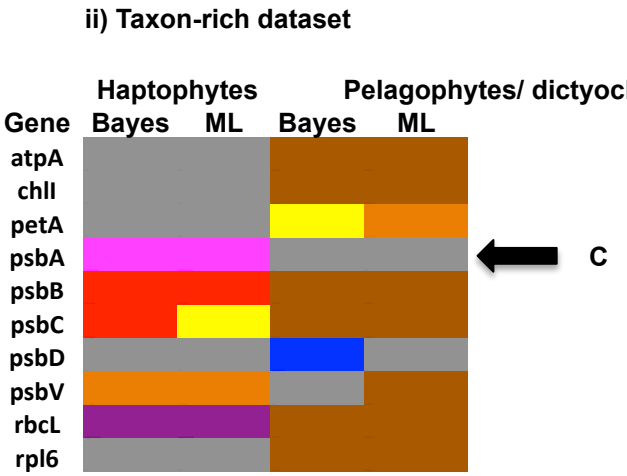


Fig. 9- figure supplement 3. Single-gene tree topologies associated with individual plastid-encoded genes. These heatmaps show the first sister-groups identified to haptophytes, and members of the pelagophyte/ dictyochophyte clade, in single-gene trees of component genes included in concatenated trees 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.



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 pelagophyes, 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 haptopthyte sequences. This might be consistent with a green algal contribution in the endosymbiotic ancestor of cryptomonad, haptophyte and ochrophyte plastids.

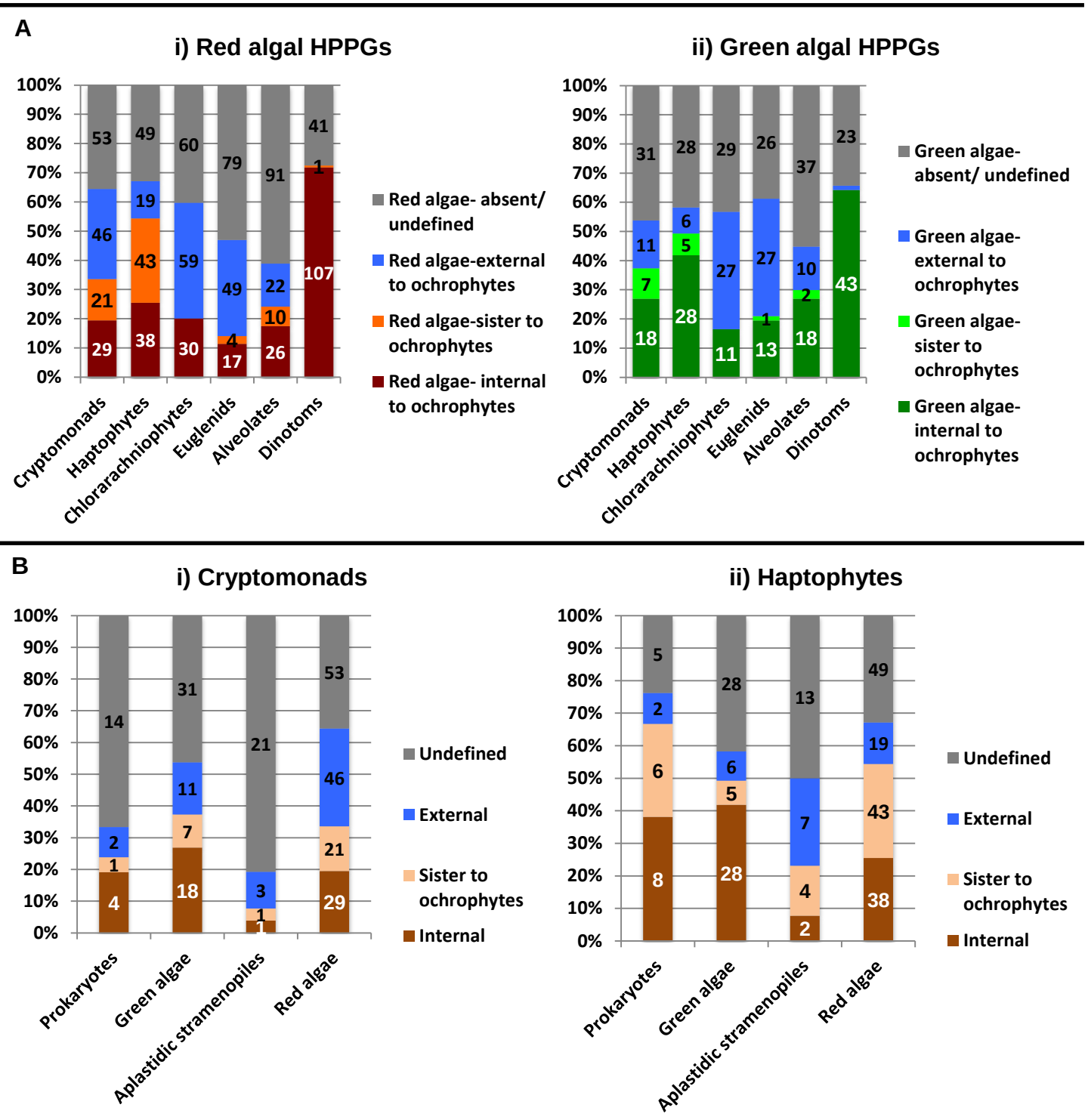


Fig. 10 –figure supplement 2. Different scenarios for the origins of haptophyte plastids. This schematic tree diagram shows different possibilities for the origins of the haptophyte plastid as predicted from the data within this study. No inference is made here regarding the ultimate origin of the ochrophyte 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 evoluitonarily distinct from the pelagophytes (**2B**). It is possible that the haptophyte plastid may have been acquired through the secondary endosymbiotis of a different lineage of red algae to the ochrophyte, either via a cryptomonad intermediate (**2C**) or directly (**2D**).