Article

TRANSPARENT PROCESS OPEN ACCESS CHECK FOR UPDATES

EMBO reports

New plastids, old proteins: repeated endosymbiotic acquisitions in kareniacean dinoflagellates

Anna MG Novák Vanclová[1,2,3]✉, Charlotte Nef[1,2], Zoltán Füssy[4], Adél Vancl[5], Fuhai Liu[1,6,7], Chris Bowler[1,2] & Richard G Dorrell[1,2,8]✉

Abstract

Dinoflagellates are a diverse group of ecologically significant micro-eukaryotes that can serve as a model system for plastid symbiogenesis due to their susceptibility to plastid loss and replacement via serial endosymbiosis. Kareniaceae harbor fucoxanthin-pigmented plastids instead of the ancestral peridinin-pigmented ones and support them with a diverse range of nucleus-encoded plastid-targeted proteins originating from the haptophyte endosymbiont, dinoflagellate host, and/or lateral gene transfers (LGT). Here, we present predicted plastid proteomes from seven distantly related kareniaceans in three genera (*Karenia*, *Karlodinium*, and *Takayama*) and analyze their evolutionary patterns using automated tree building and sorting. We project a relatively limited (~10%) haptophyte signal pointing towards a shared origin in the family Chrysochromulinaceae. Our data establish significant variations in the functional distributions of these signals, emphasizing the importance of micro-evolutionary processes in shaping the chimeric proteomes. Analysis of plastid genome sequences recontextualizes these results by a striking finding the extant kareniacean plastids are in fact not all of the same origin, as two of the studied species (*Karlodinium armiger*, *Takayama helix*) possess plastids from different haptophyte orders than the rest.

Keywords Myzozoa; Protists; Automated Tree Sorting; Post-Endosymbiotic Organelle Evolution; Shopping Bag Model
Subject Categories Evolution & Ecology; Microbiology, Virology & Host Pathogen Interaction; Organelles
https://doi.org/10.1038/s44319-024-00103-y
Received 24 November 2023; Revised 19 January 2024;
Accepted 6 February 2024
Published online: 18 March 2024

Introduction

The origin of oxygenic photosynthesis, and its adoption by eukaryotes, has fundamentally changed the evolutionary landscape of life on Earth. Originating in cyanobacteria more than 3 billion years ago, photosynthesis has spread across the eukaryotes through multiple endosymbiotic acquisitions of plastids. These began with the primary plastids (i.e., chloroplasts) of cyanobacterial origin that represent the defining synapomorphy of Archaeplastida (glaucophytes, red and green algae, and plants) and are surrounded by two membranes. Plastids subsequently spread into other groups (e.g., diatoms, haptophytes, dinoflagellates, and euglenids) through multiple endosymbiotic acquisitions of eukaryotic algae, forming "complex plastids" that possess additional membranes, often continuous with the endoplasmic reticulum. Across the tree of life, plastids provide their hosts with novel metabolic and biosynthetic capabilities and serve as important vectors of horizontal gene transfer (Howe et al, 2008; Ku et al, 2015; Novák Vanclová et al, 2020; Dorrell et al, 2021).

Plastids have their own genomes that typically encode core photosynthetic proteins, alongside some of their genetic housekeeping machinery and primary metabolic enzymes. Their size and coding capacity vary, typically possessing 20–250 genes, driven by gradual and somewhat convergent gene loss (Uthanumallian et al, 2021). Plastids, however, require a much larger inventory of proteins to function (e.g., >2000 in *Arabidopsis* and model eukaryotic algal groups; Kleffmann et al, 2004; Gruber et al, 2015; Novák Vanclová et al, 2020); and most of these are encoded in the nucleus and post-translationally imported based on cleavable N-terminal targeting signals. These plastid-targeted proteins comprise the expression products of plastid-derived genes that were transferred to the host nucleus, host-derived genes newly adapted for plastidial function, and genes originating from lateral gene transfer (LGT) from other sources, including other symbionts that the host has interacted with over its evolutionary history (Moustafa et al, 2008). This makes plastid proteomes innately chimeric entities, or "shopping bags", whose composition may reflect both general principles of organelle evolution and the unique evolutionary histories of individual lineages (Larkum et al, 2007; Suzuki and Miyagishima, 2010; Dorrell et al, 2017; Novák Vanclová et al, 2020). These processes are poorly understood and very problematic to model in systems with very ancient plastids, such as plants, due to weak phylogenetic signal. However, a very pertinent model in this respect can be found in dinoflagellates.

Dinoflagellates are a diverse group of unicellular algae that are ecologically and economically significant; including crucial

[1]Institut de Biologie de l'École Normale Supérieure (IBENS), École Normale Supérieure, CNRS, INSERM, Université PSL, Paris, France. [2]CNRS Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GOSEE, Paris, France. [3]Institute Jacques Monod, Paris, France. [4]Faculty of Science, Charles University, BIOCEV, Vestec, Czechia. [5]Faculty of Mathematics and Physics, Charles University, Prague, Czechia. [6]Centre de Recherches Interdisciplinaires, Paris, France. [7]Tsinghua-UC Berkeley Shenzhen Institute, Shenzhen, China. [8]CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative - UMR 7238, Sorbonne Université, Paris, France.
✉E-mail: anna.novak-vanclova@ijm.fr; richard.dorrell@sorbonne-universite.fr

symbionts of corals, and toxin-producing species that can cause major losses in fisheries and aquaculture industries as well as endangering public health (Hackett et al, 2004; Wang, 2008). Within the eukaryotic tree of life, they fall among the supergroup Alveolata, specifically Myzozoa, along with chrompodellids and apicomplexans (Ševčíková et al, 2015). Dinoflagellates are of further evolutionary interest given their unique and idiosyncratic cell biology. Their nuclear genomes are generally extremely large (up to several hundred Gbp) with permanently condensed chromosomes that lack histones, while their mitochondrial genomes are very streamlined (Lin, 2011; Dorrell and Howe, 2015). The evolutionarily ancestral dinoflagellate plastids are descended from red algae and bound by three membranes (Moog and Maier, 2017). The ancestral dinoflagellate plastid is further marked by multiple functional oddities. These include a unique carotenoid pigment, peridinin (Haxo et al, 1976), and peridinin-chlorophyll a-proteins (PCP) that together form membrane-extrinsic light-harvesting antennae (Ogata et al, 1994), while the chlorophyll-binding proteins are massively paralogized and synthesized as polyproteins (Boldt et al, 2012). Another striking characteristic of these plastids is a bacterial-like nucleus-encoded form II RuBisCO (Morse et al, 1995) in contrast to the partially plastid-encoded form I RuBisCO (consisting of homo-octamers of large and small subunits) found in all other plastid lineages except chromerids (Janouškovec et al, 2010). Finally, dinoflagellate mitochondrial and plastid genomes have generally extremely reduced coding contents (three and twelve protein-coding genes, respectively), with the plastid genome further fragmented into small unigenic elements termed “minicircles” (Zhang et al, 1999; Nash et al, 2007). The transcripts from dinoflagellate plastid minicircles undergo distinctive maturation events, including extensive substitutional editing and the addition of a 3’ poly(U) tail not found in most other plastid lineages (Zauner et al, 2004; Wang and Morse, 2006; Dorrell and Howe, 2015).

Dinoflagellates as a group are ecologically opportunistic: most plastid-bearing representatives are, in fact, mixotrophic (Cohen et al, 2021; Jeong et al, 2021), and in global meta-genomic analyses (e.g., *Tara* Oceans) they appear mostly as predators and parasites (Pierella Karlusich et al, 2022). Their plastids are frequently secondarily reduced to non-photosynthetic organelles, lost completely (Saldarriaga et al, 2001; Sanchez-Puerta et al, 2007; Gornik et al, 2015; Cooney et al, 2022), or replaced in new symbiogenetic events. These may result in fully integrated plastids (Tengs et al, 2000; Matsumoto et al, 2011), complex endosymbionts retaining plastids, mitochondria, and nuclei (Imanian and Keeling, 2007; Sarai et al, 2020), and kleptoplastids of various degrees of stability (Koike et al, 2005; Myung et al, 2006; Gast et al, 2007). In some cases, the coexistence of a new organelle or endosymbiont with a remnant of the ancestral plastid has been proposed (Hehenberger et al, 2019), and in one lineage, a complete repurposing of parts of the plastid and mitochondrion into an enigmatic, photosynthesis-unrelated optical structure termed an “ocelloid” has taken place (Gavelis et al, 2015).

Kareniaceae, typified by the genera *Karenia*, *Karlodinium*, and *Takayama*, are mixotrophic dinoflagellates whose plastids are related to those of haptophytes. Kareniaceae are the most abundant dinoflagellate lineage with non-peridinin plastids in *Tara* Oceans (Dataset EV1, based on de Vargas et al, 2015), and both *Karenia* and *Karlodinium* have important impacts as toxic components of harmful algal blooms. Instead of peridinin, plastids of the Kareniaceae contain the accessory pigment fucoxanthin typical of their haptophyte ancestors (Zapata et al, 2004, 2012). Kareniacean plastids are further assumed to be bound by four membranes, the outermost of which is continuous with the endoplasmic reticulum as per haptophytes, distinct from the three membrane-bound plastids found in peridinin dinoflagellates. Further ecological and evolutionary complexity is found within the Kareniaceae in the Antarctic native Ross Sea dinoflagellate (RSD) (Gast et al, 2007; Hehenberger et al, 2019), which uses a kleptoplastid from the haptophyte *Phaeocystis antarctica* acquired independently to that of other Kareniaceae, *Shimiella gracilenta* with cryptophyte kleptoplastids (Ok et al, 2021), and *Gertia stigmatica* which contains peridinin instead of fucoxanthin (Takahashi et al, 2019). The well-characterized identity of both the ancestral and current kareniacean plastids, alongside their ecological abundance, diversity, and relevance, renders them a particularly appealing system for understanding the fundamental events associated with serial endosymbiosis and plastid replacement. Previous studies have suggested that similarly complex endosymbiotic interactions occurred more broadly across the eukaryotic tree of life and may have been key to the resulting symbiogenetic events (Hannaert et al, 2003; Huang and Gogarten, 2007; Dorrell and Smith, 2011; once-Toledo et al, 2019).

Partial kareniacean plastid genomes and transcriptomes have been sequenced for the species *Karlodinium micrum* (syn. *veneficum*) and *Karenia mikimotoi*, indicating the retention of around 100 genes on a circular or linear chromosome (Gabrielsen et al, 2011; Dorrell et al, 2016), which is somewhat fewer than the ca. 140 genes associated with haptophyte plastids, alongside the presence of episomal minicircles (Richardson et al, 2014). Further studies, based on expressed sequence tags (ESTs) (Ishida and Green, 2002; Nosenko et al, 2006; Patron et al, 2006) and, more recently, transcriptomes of *Karenia brevis* and *Karlodinium micrum* realized through the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (Burki et al, 2014; Keeling et al, 2014; Hehenberger et al, 2019), have provided foundational insights into the nucleus-encoded proteome of the kareniacean plastid. These are defined by a bipartite targeting sequence, consisting of an N-terminal signal peptide followed by a hydrophilic transit peptide, similar to those of haptophytes, although with apparent differences in composition (Patron and Waller, 2007; Yokoyama et al, 2011). Phylogenetic analysis of the genes encoding plastid-targeted kareniacean proteins (Nosenko et al, 2006; Patron et al, 2006; Burki et al, 2014; Bentlage et al, 2016; Matsuo and Inagaki, 2018) indicate that some resolve with haptophytes (i.e., the endosymbiont ancestor), while others show alternative origins. These include key elements of the plastid gene expression machinery (e.g., RNA editing and poly(U) tail addition), which are likely to have been derived from the dinoflagellate host and potentially associated with the ancestral peridinin plastid (Dorrell and Howe, 2012; Jackson et al, 2013); and cofactor biosynthetic pathways (e.g., isopentenyl pyrophosphate and protoporphyrin/heme), which resolve principally either with dinoflagellates or sources suggesting horizontal acquisition (Matsuo and Inagaki, 2018). To our knowledge, a reconstruction of the fucoxanthin plastid proteome as a whole, either in silico or proteomics-based, has yet to be attempted, and its overall metabolic composition and specific evolutionary origin within the

Table 1. Assembly statistics and assessment of completeness for the five new transcriptomes.

	Karenia mikimotoi	*Karenia papilionacea*	*Karlodinium armiger*	*Karlodinium micrum* RCC3446	*Takayama helix*
Raw transcripts	373,333	241,077	265,372	202,000	662,715
N50	1221 bp	1168 bp	1231 bp	1332 bp	1326 bp
Avg contig length	752 bp	764.5 bp	791 bp	856 bp	799 bp
Total size	208.9 Mbp	155.4 Mbp	167.9 Mbp	145.1 Mbp	256.8 Mbp
Proteins in final dataset	167,296	134,380	139,785	112,599	196,617
Complete BUSCOs	71.4%	68.2%	71.0%	70.2%	77.3%
Fragmented BUSCOs	10.2%	13.7%	11.0%	11.0%	6.3%
Missing BUSCOs	18.4%	18.1%	18.0%	18.8%	16.4%

haptophytes remain unresolved (Yoon et al, 2002; Takahashi et al, 2019; Leblond et al, 2022).

Here, we explore kareniacean plastid evolution as a model for the endosymbiotic establishment of plastids and the spread of photosynthesis across eukaryotes. We leverage newly sequenced transcriptomic datasets for two additional strains of *Karenia* and *Karlodinium* and the first transcriptome for the genus *Takayama*, alongside high-throughput phylogenetic and environmental data to understand the diverse evolutionary trajectories of the kareniacean plastid following its acquisition. Our data point to a complex origin of the kareniacean plastid, whose probable nucleus-encoded plastid-targeted proteins appear to be of predominantly dinoflagellate origin, but whose composition is influenced by between-pathway, between-protein, and even between-species variations in phylogenetic origin. Considering the phylogenetic origins of both the nucleus-encoded and plastid-encoded proteins, we further reveal at least three independent acquisitions of the kareniacean plastid from different haptophyte lineages. This, however, does not reflect on the nucleus-encoded plastid protein inventory which exhibits a very similar and largely shared Chrysochromulinaceae-like phylogenetic signal suggesting monophyly of this part of the plastid proteome. Finally, we evaluate the different distributions of key lineages in the contemporary ocean. The repeated acquisition of the kareniacean plastid with different evolutionary consequences across different species can provide insights into the functional principles that constrain post-endosymbiotic organelle evolution, and the symbiogenetic establishment of photosynthesis across eukaryotes.

Results

Pan-transcriptomics reveals the core fucoxanthin plastid proteome

Five new transcriptomic datasets were produced for fucoxanthin dinoflagellates: *Karenia mikimotoi* and *Karenia papilionacea*, the RCC3446 strain of *Karlodinium micrum* and its distant relative *Karlodinium armiger*, and for *Takayama helix*, a predator of other dinoflagellates more closely related to *Karlodinium* than *Karenia* (Jeong et al, 2016). To our knowledge, this latter transcriptome represents the first one reported for the genus *Takayama*. These libraries, alongside four previously sequenced MMETSP libraries for *Karenia brevis*, one for *Karlodinium micrum* strain CCMP2283, and an independently sequenced transcriptome for the RSD (Data refs: Keeling et al, 2014; Ryan et al, 2014; Hehenberger et al, 2019), represent the pan-kareniacean library used for all downstream analyses.

The raw dataset sizes range from around 200 to 660 Mbp but contain large amounts of redundant or partially redundant transcripts, reflecting the extensive paralogization and pseudogenization associated with dinoflagellate nuclear genomes (Stern et al, 2010). To reduce the size and redundancy of our datasets for subsequent analysis, we selected the most complete paralogs of each studied protein by cd-hit. The completeness of the datasets as assessed by BUSCO is around 70% which is rather high in comparison to other dinoflagellate datasets (Table 1; Dataset EV2).

In silico prediction reveals within-lineage divergence in fucoxanthin plastid proteome contents

A modified version of ASAFind (Gruber et al, 2015), which has previously been used to reconstruct the haptophyte plastid proteome *in silico* (Dorrell et al, 2017), was constructed with a custom scoring matrix for fucoxanthin dinoflagellate plastids (see Materials and Methods) and used in combination with SignalP 5.0 in the final prediction pipeline. After preliminary automatic annotation of the retrieved *in silico* plastidial proteomes of all studied organisms by KAAS (https://www.genome.jp/kegg/kaas/; Moriya et al, 2007), we noticed that some of the integral plastidial proteins missing from these datasets could be captured by an alternative prediction approach combining PrediSI (Hiller et al, 2004) and ChloroP (Emanuelsson et al, 1999) employed previously for *Euglena gracilis* data (Ebenezer et al, 2019), and so we included these as well. The proportion of redundancy-treated translated transcripts predicted as plastid-targeted ranged from 7.5–14.5% in *Karlodinium micrum* and *Karenia brevis*, respectively (~9.5% being both the average and the modal value). Of these, most have a targeting sequence predictable by the modified ASAFind, while between 16–22% of the predicted plastid-targeted sequences represent additions based solely on the abovementioned alternative signal.

The putative phylogenetic origin of all predicted plastid proteins was investigated via a custom pipeline, integrating homology mining, single-gene tree construction and sorting based on topology and its statistical support (Appendix Fig. S1). Between 40 and 55% of each protein dataset had some homologs in our reference database containing representatives for most major eukaryotic groups, particularly enriched in haptophyte and dinoflagellate transcriptomes (see Materials and Methods). In about one-third of these, the identified

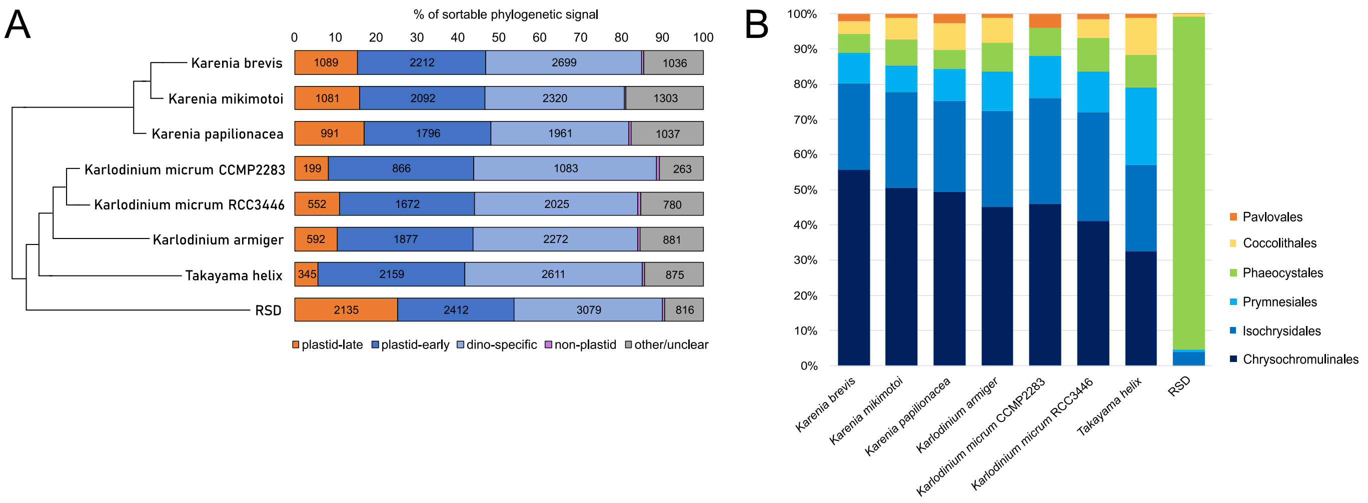


Figure 1. Ratios of phylogenetic signal of plastid proteins with detectable homologs in Kareniacean transcriptomes.

This plot shows:plastid-late (haptophyte-like; orange), plastid-early (dinoflagellate-like or dinoflagellate-specific; blue), ancestrally alveolate-like (purple), and other or unclear (gray) (**A**), and further breakdown of the plastid-late signal traceable to a concrete haptophyte family (Chrysochromulinaceae (dark blue), Isochrysidales (blue), Prymnesiales (light blue), Phaeocystales (green), Coccolithales (yellow), and Pavlovales (orange)) (**B**).

homologs were shared solely with other dinoflagellates and were therefore considered dinoflagellate-specific proteins without any further phylogenetic analysis. For the rest, single-gene trees (29,329 in total) were constructed and automatically sorted based on which evolutionary origin they support. The main two investigated categories were "plastid-early" (i.e., vertically inherited from the dinoflagellate host; color-coded as blue and purple dependent on whether the protein retrieved only dinoflagellate homologs or showed deeper homology to other alveolates or myzozoan lineages) and "plastid-late" (i.e., immediately clustering with the haptophytes; color-coded as orange in all graphical materials). In addition to this, potential cases of lateral gene transfer (LGT) from prokaryotes, green algae, and ochrophytes (as previously identified major sources of LGT within eukaryotic algal lineages, e.g., Moustafa et al, 2008; Matsuo and Inagaki, 2018; Novák Vanclová et al, 2020; Dorrell et al, 2021) were identified (examples shown in Appendix Figs. S6–S12).

Of all the proteins with homologs in the database, 65–80% were categorized as "plastid-early" in all studied organisms (Fig. 1A). The proportion of proteins categorized as "plastid-late" is generally in accord with the previously published estimates (Burki et al, 2014) but there was a noticeable difference between the genera, with ~15% noted in all three species of *Karenia*, 10% in *Karlodinium*, and 6% in *Takayama*. The proportion of proteins shared and clustering with ciliates, i.e., primarily heterotrophic alveolates (the "non-plastid signal"), is almost negligible in all studied organisms, as is the contribution of prokaryotic or "green" LGT (cases of which do not exceed 20 per organism). The amount of "brown" LGT is slightly higher than the amount of non-plastid signal (varying from 12–107 cases per organism). Between 10 and 20% of the signal in each organism was not resolvable into any of the investigated categories.

Phylogenomic evidence for different evolutionary histories of kareniacean plastidial and non-plastidial proteins

We used the whole transcriptomes as input for the PhyloFisher pipeline (Tice et al, 2021) to construct a matrix of 241 pan-eukaryotic nuclear-encoded genes for a reference phylogeny of the studied kareniacean nuclear genomes in the context of the eukaryotic tree of life. The topology reconstructed from the PhyloFisher matrix is consistent with recently published rRNA trees (Takahashi et al, 2019; Ok et al, 2021) with *Karenia* as a sister to both *Karlodinium* and *Takayama* (Fig. EV1, simplified in Fig. 1A).

Previous studies of the fast-evolving fucoxanthin plastid genome revealed a deep-branching position sister to all prymnesiophytes but no affinity towards individual orders (Choi et al, 2017; Klinger et al, 2018; Kawachi et al, 2021). To recover a more specific affinity, we enumerated the predicted plastid-targeted proteins clustering specifically with one of six haptophyte subgroups (Chrysochromulinaceae, Isochrysidales, Phaeocystales, Coccolithales, Pavlovales, and Prymnesiales; Fig. 1B), constituting approximately one-quarter of the total plastid-late signal. We also recorded how many trees showing these phylogenetic affiliations also recover Kareniaceae as monophyletic (1418 in total, 220 for Chrysochromulinaceae, 93 for Isochrysidales, 33 for Prymnesiales, and <20 for the rest; Appendix Fig. S2). For each of the seven studied fucoxanthin-containing kareniaceans the most abundant category (up to >50%) was Chrysochromulinaceae. In RSD, which possesses a kleptoplastid derived from *Phaeocystis antarctica*, the main source was Phaeocystales (>70%) and only a very small proportion (~2.5%) was shared with the other kareniaceans. We noted a strong secondary signal to the order Isochrysidales in all kareniacean species, including the RSD, even normalizing for dataset size (Appendix Fig. S3), although many were specific to individual genera (Appendix Figs. S4, S5).

We also sorted the single-gene trees based on the inner kareniacean topology they support (Fig. 2) and note that while the organismal topology (Fig. EV1) is predominantly recovered in trees defined as plastid-early (1505 out of 1922, 78%), the plastid-late trees often show different topologies, especially the one in which *Karenia* and *Karlodinium* are sisters to the exclusion of *Takayama* which forms a sister or

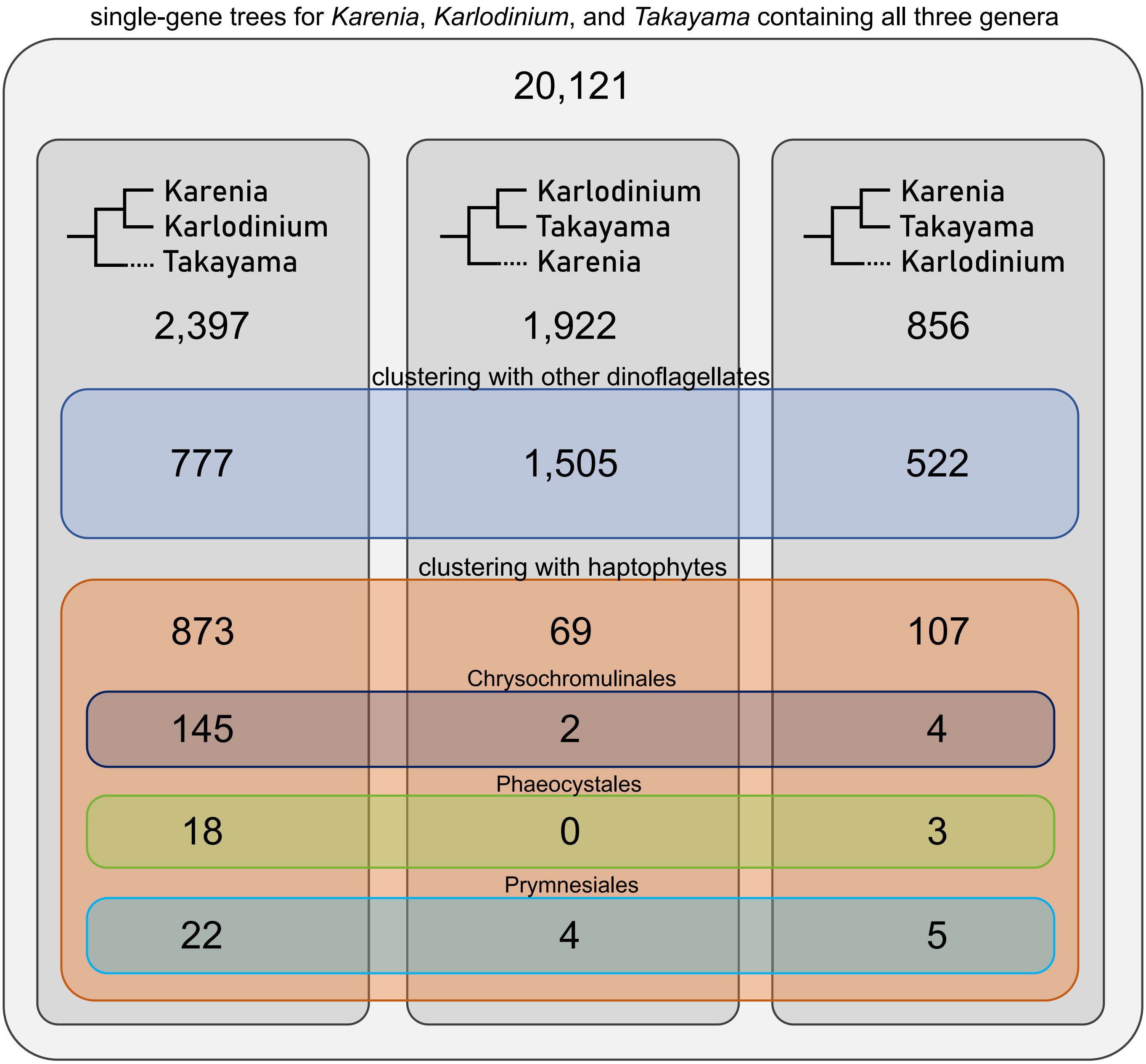


Figure 2. Proportions of single-gene trees supporting three different internal topologies of the three studied Kareniacean genera, grouped by broader phylogenetic origin.

The plastid-early genes clustering with other dinoflagellates primarily support the topology obtained from 18S trees and PhyloFisher trees (i.e., with *Karenia* sister to *Takayama/Karlodinium*). In contrast, the plastid-late genes clustering with haptophytes predominantly support *Takayama* as an outgroup of *Karenia* and *Karlodinium*. *Karlodinium* is less frequently recovered as an outgroup to a monophyletic clade of *Takayama/Karenia*.

separate branch (873 trees compared to 176 that support the two alternative topologies). We further investigated these relationships using a selected set of 23 nucleus-encoded plastid-late genes to build manually curated single-gene and partial-concatenation trees (Fig. EV2; Dataset EV10). Consistent with the above results, they largely supported the monophyly of *Karenia* and *Karlodinium* and exclusion of *Takayama*, and a close relationship to Chrysochromulinaceae.

A hybrid plastid proteome with species-specific innovations

The plastid proteins of each kareniacean were automatically annotated with enzymatic functions based mainly on KEGG, presented as a pivot table (Dataset EV3). Considering χ^2 enrichments of individual KEGG annotations (KO IDs) across all seven fucoxanthin-containing species, we noted a statistically

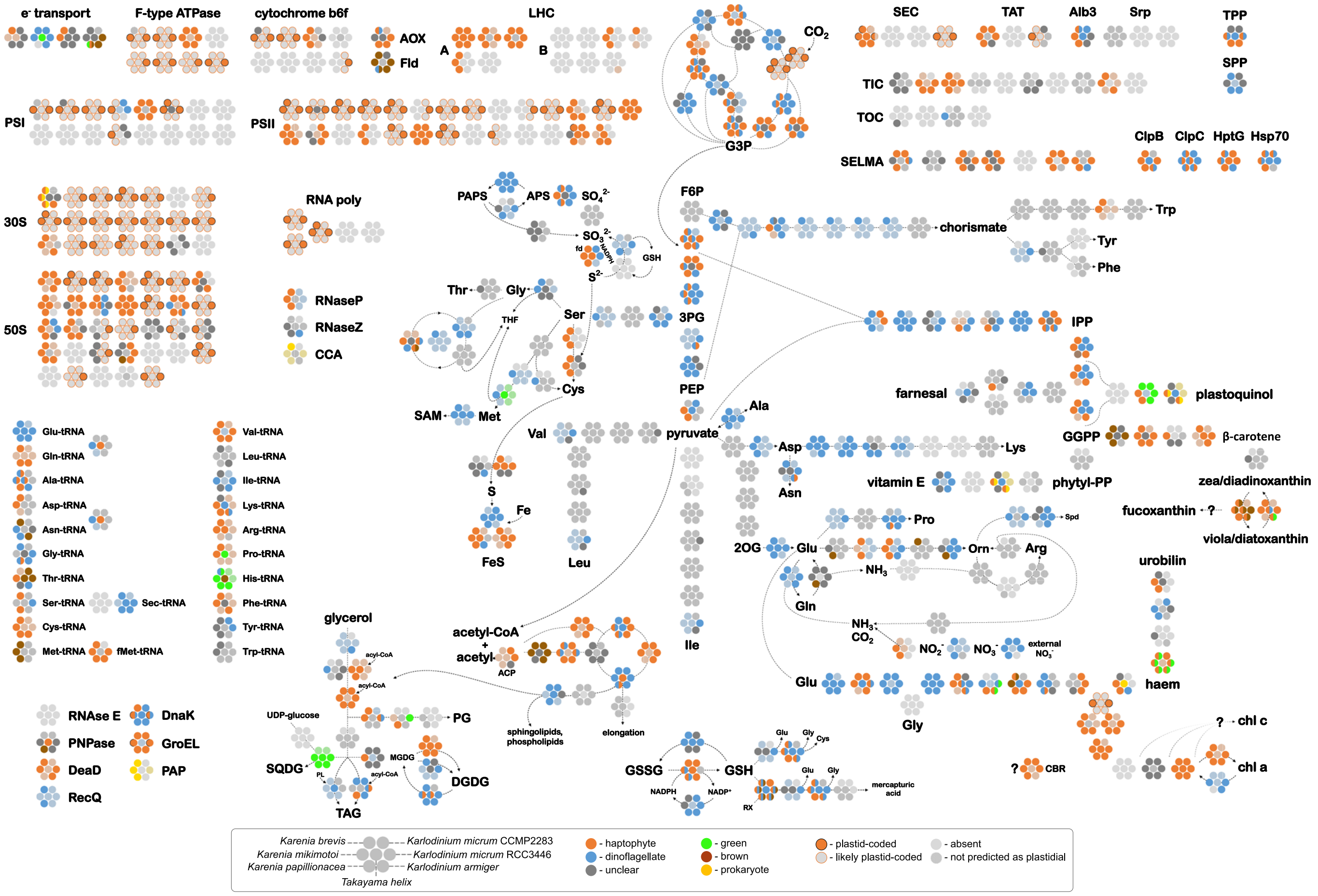


Figure 3. Reconstruction of major metabolic pathways of plastids of the seven kareniaceans, adapted from KEGG (Kanehisa and Goto, 2000).

Plastid proteins are arranged by major metabolic pathway or biological process, with each protein shown as rosettes. Each rosette (described in the legend) relates to the homologs of a specific nucleus-encoded and plastid-targeted protein, with each circle within each rosette corresponding to a different species, and different colors corresponding to different evolutionary affiliations. Proteins of plastid-late (haptophyte) origin, such as are concentrated in photosystem and ribosomal processes, are colored red; and proteins of plastid-early (dinoflagellate) origin, such as are concentrated in carbon and amino acid metabolism are colored blue. Additional phylogenetic origins are denoted by further examples in the legend, e.g., LGT shown in orange, green, and brown. Proteins encoded in plastid genomes (red with black border) are provided for the incomplete genome assembly for *Karlodinium micrum*, and partial plastid transcriptomes are identified for *Karenia mikimotoi* and *Karenia brevis*. In certain cases (shown as rosettes with multiple colors), homologs from different species have different evolutionary origins, e.g., *Karenia* possessing plastid-late and *Karlodinium/Takayama* plastid-early. A detailed breakdown of the individual sequences with their predicted origins used for this reconstruction are available in Dataset EV3, and an alternative graphic, including the RSD, is available as Appendix Fig. S13.

significant enrichment ($p < 0.01$) in the plastid-late signal in the BRITE category "Photosynthesis", the combined category "Translation", and the KEGG Pathway "Porphyrin and chlorophyll metabolism", while the plastid-early signal was enriched in the BRITE categories "Membrane trafficking", "Protein kinases, phosphatases, and associated proteins", and "Ion channels". Reflecting the probable plastid-early origin of the fucoxanthin plastid RNA processing machinery, categories relating to transcription and mRNA do not appear to be biased in their origin.

When analysed manually and qualitatively, additional evolutionary patterns were apparent in the overall metabolic map (Fig. 3), some of which were previously described in *Karenia brevis* and *Karlodinium micrum* (Matsuo and Inagaki, 2018). These include notable differences between the genera and species in some pathways or enzymes, and several general particularities such as unexpectedly missing, duplicated, or laterally gained proteins.

Tetrapyrrole metabolism

The tetrapyrrole pathway is clearly divided into three parts: a backbone leading from glutamine to protoporphyrin, the heme and bilin pathway, and the chlorophyll-synthesizing branch (Tanaka and Tanaka, 2006). While the former two are generally of mixed origin (including isolated cases of putative LGT), perhaps slightly skewed towards a plastid-early signal, the latter comprises almost exclusively plastid-late proteins. The only plastid-early enzyme associated specifically with chlorophyll is chlorophyllase, which has additional roles in phytol recycling to chlorophyll synthesis. Two evolutionarily distinct versions of the biliverdin-producing heme oxygenase seem to be present in the plastid of representatives of both *Karenia* and *Karlodinium*: one of plastid-late and one of green-like origin (each in one or two copies, Fig. EV3). This enzyme is often present in multiple copies that have been proposed to play slightly different roles in other reactions of bilin metabolism, especially in photosynthetic organisms where these molecules serve as chromophores of phytochromes (Dammeyer and Frankenberg-Dinkel, 2008).

Terpenoid metabolism

While the origin of the central MEP-DOXP pathway is mixed (slightly skewed towards plastid-early) and often not consistent between the genera or even species, this is not the case for other pathways branching from it. Carotenoid metabolism comprises mostly plastid-late proteins and does not contain any plastid-early proteins. Two enzymes (15-cis-phytoene synthase (Appendix Fig. S10) and zeaxanthin epoxidase) represent LGT from brown algae. MPBQ/MSBQ methyltransferase (VTE3) and homogentisate solanesyltransferase (HST; Appendix Fig. S11) likely represent LGTs from bacteria and green algae, respectively. However, neither the plastoquinol nor tocopherol synthesis pathway seems to be complete.

Fatty acid and plastid lipid biosynthesis

Retention of a plastid-derived type II FAS pathway has already been implied in fucoxanthin and in peridinin dinoflagellates (Janouskovec et al, 2017) and our data supports this with all steps of this pathway identified in at least some of the species. Most of these enzymes are of plastid-late origin or have both plastid-early and plastid-late copies. The origins are not always consistent between the genera (compare FabG and FabI in Fig. 3 or Dataset EV3). While one of the two enzymes for sulfoquinovosyldiacylglycerol (SQDG) synthesis, SQD2 is of green algal-like origin in all genera, representing a reliable case of conserved LGT (Appendix Fig. S9), the other, SQD1, was not identified in any of the seven transcriptomes. That said, SQD1 (and sometimes SQD2 as well) were not detectable in various algae in previous studies (Novák Vanclová et al, 2020, Riccio et al, 2020), possibly due to very low expression or presence of as yet undiscovered alternative enzyme.

Photosynthesis

The majority of photosystem and light-harvesting complex subunits and parts of the photosynthetic electron transport chain are of plastid-late origin. We detected no homologs of peridinin-chlorophyll-binding proteins (PCP) in any kareniacean transcriptome. PsaD is the only photosystem subunit that was vertically inherited from the ancestral plastid in *Karlodinium* and likely *Takayama* (where the protein was not predicted as plastid-targeted, likely due to N-terminal truncation), but it remains plastid-coded in *Karenia* (Dorrell et al, 2016). Another part of the photosynthetic apparatus that is not of plastid-late origin is plastidial ferredoxin (PetF) which shows a plastid-early origin in *Karenia* and *Karlodinium*, and is replaced by a green-like homolog in *Takayama*. PetF is also associated with non-photosynthetic metabolism (e.g., in leucoplasts of non-photosynthesizing plant tissues), and is plastid-encoded in the non-photosynthetic chrysophyte *Spumella* NIES-1486 (Dorrell et al, 2019), and its specific metabolic functions in Kareniaceae remain to be determined.

ATP synthase

Most subunits of the plastidial F-type ATP synthase are plastid-encoded in Kareniaceae and their sequences were not recovered in the transcriptomes. The gamma subunit (atpG, K02115) is encoded in the nucleus and is of plastid-late origin in all species. Surprisingly, the delta subunit (atpH, K02113) was not detected in any of the transcriptomes. AtpH is a conserved and essential subunit that is usually plastid-encoded (including in haptophytes) but was not retrieved as such in *Karlodinium micrum* and *Karenia mikimotoi* (Gabrielsen et al, 2011; Dorrell et al, 2016). Interestingly, an unrelated and structurally dissimilar but functionally analogous subunit of mitochondrial ATPase (ATP5D, K02134) seems to be duplicated, but it is unclear whether these two phenomena are linked.

Calvin cycle

The Calvin cycle in kareniaceans is a true evolutionary mixture, with most enzymes having multiple copies of both early and late origins. Sedoheptulose-bisphosphatase is the only protein that seems to be of purely plastid-late origin in all species, which may reflect that it is redox-regulated and has exclusive functions in photosynthesis (Gütle et al, 2016). In contrast, triosephosphate isomerase and transketolase are exclusively of plastid-early origin, while further enzymes that function reversibly in both glycolysis and pentose phosphate pathway (Kroth et al, 2008) possess both plastid-late and plastid-early copies. Notably, phosphoribulokinase is an exception from this overall trend as it is a photosynthetic carbon fixation-specific enzyme that possesses uniquely plastid-early origins.

Amino acid metabolism

While enzymes for the synthesis of most amino acids are present in the transcriptomes, only a small portion of them were predicted as plastid-targeted: the synthesis of alanine, aspartate, and asparagine from pyruvate and several following enzymes in the pathway towards lysine (up to 4-hydroxy-tetrahydrodipicolinate reductase), some of the enzymes converting glutamine to ornithine, the final step of proline synthesis (pyrroline-5-carboxylate reductase), some enzymes for serine and glycine synthesis from 3-phosphoglycerate, and the GS/GOGAT pathway. The shikimate pathway does not seem to be plastid-localized. The only exceptions to this are the AROM polypeptide of *Karenia brevis;* and 3-dehydroquinate synthase, which is plastid-targeted in five of the species and shows a plastid-early origin in *Karlodinium* and *Takayama* and a plastid-late origins in *Karenia*. Interestingly, the same evolutionary division can be observed for two of the enzymes of ornithine synthesis from glutamate, acetylglutamate kinase and *N*-acetyl-gamma-glutamyl-phosphate reductase.

Sulfur and nitrogen metabolism

Sulfur metabolism, including enzymes converting serine to cysteine, is evolutionarily mixed but comparatively richer in plastid-late signals. The iron-sulfur cluster assembly (SUF) is not uniform in its origin and comprises plastid-early SufS and SufB but plastid-late SufC, D, and E. Most of the nitrogen assimilation enzymes were not predicted as plastid-targeted, but the ones present are also evolutionarily non-uniform: the putative plastidial nitrate/nitrite transporter is pan-alveolate (shared with ciliates); nitrate reductase is plastid-early; while nitrite reductase shows a plastid-late origin.

Protein import

The kareniacean protein import machinery shows a predominant plastid-late origin. This is expected as the SELMA (Symbiont-specific ERAD-Like MAchinery for protein import) is present in haptophytes and kareniaceans but not peridinin dinoflagellates. Nonetheless, there are isolated cases of plastid-early SELMA components targeted to the plastid, although these may equally relate to host ERAD components with false positive predictions. Alb3 represents an interesting case of a protein strongly implicated in plastid biogenesis and photosynthesis with both plastid-early and plastid-late homologs in *Karenia brevis* and only a plastid-early homolog in *Karlodinium micrum* and *T. helix*.

Ribosomal proteins

Plastid-targeted ribosomal proteins show mostly plastid-late origins with only a few exceptions. These include L7/L12 in *Karlodinium* and *Takayama*, L17 of *Karlodinium*, and L15 in *Karenia mikimotoi* and *Karlodinium armiger*, which are of plastid-early origin, and might represent either genes retained from the ancestral plastid, or retargeted, originally mitochondrial proteins. All three species of *Karenia* contain an additional copy of ribosomal protein S1 that likely represents bacterial origin and is shared with only a few eukaryotic algae. Single gene trees of these proteins indicate the presence of further, non-plastid-targeted homologs in all *Karenia* transcriptomes, of as yet unclear function.

Aminoacyl-tRNA synthesis

The plastidial inventory of aminoacyl-tRNA synthetases is evolutionarily mixed, with relatively high concentrations of LGT (Figs. EV4 and EV5; Appendix Figs. S6–S8). There are two cases of LGT from green algae: one in *Takayama* only (PARS, proline-tRNA synthetase) and a second (HARS, histidine-tRNA synthetase) in all species except *Takayama* where, in turn, the protein seems to be brown in origin. Three other synthetases (TARS, threonyl-, NARS, asparaginyl-, and MARS, methionyl-) are of brown origin in at least some representatives. We did not detect clear evidence for plastid/mitochondrion dual-targeting in the identified plastid-targeted tRNA synthetases.

Targeting sequences of kareniacean plastid proteins vary with their evolutionary origin

Next, we wished to test if there are differences in the plastid-targeting pre-sequences of plastid-late and plastid-early proteins, by comparison to analogous protein regions from peridinin dinoflagellate and haptophyte references. The signal peptides of the predicted kareniacean plastid-targeted proteins were found to typically contain a central LACLAC motif and a terminal GHG motif directly preceding the cleavage site regardless of whether they were of plastid-late or plastid-early origin (Fig. 4A,B), whereas these motifs do not occur in haptophytes nor in peridinin dinoflagellates (Fig. 4C,D). Comparisons of the unique three-letter motifs associated with the plastid-targeting sequences of each group revealed particularly strong enrichments in the LACLAC and GHG motif (and its variants) in kareniacean proteins of plastid-late origin, compared to proteins of plastid-early origin and haptophyte and dinoflagellate equivalents (Fig. 4E), despite similar overall amino acid composition of signal peptides across all datasets (Appendix Fig. S14). In contrast, limited similarities were found in the transit peptides of the predicted kareniacean plastid-targeted proteins, apart from the double-arginine motif immediately following the cleavage site and conserved proline at +15 and arginine at +22 after the cleavage sites (Fig. 4F). No notable differences in the investigated pre-sequences were observed between proteins whose functions are consistent with either thylakoid or stromal localizations (Appendix Fig. S15).

Plastid phylogenetics reveals different origins of kareniacean plastid genomes and proteomes

To further investigate the aforementioned incongruence between plastidial and non-plastidial phylogenetic signals, we sequenced 6 plastid-encoded genes (*psbA*, *psbC*, *psbD*, *psaA*, *rbcL*, 16S) previously used as markers for the origins of the fucoxanthin plastid genome (Takishita et al, 1999; Dorrell and Howe, 2012). We chose to use m/rRNA sequences, sequences amplified by RT-PCR from *Karenia papillionacea*, *Karlodinium armiger*, and *T. helix* mRNA, and retrieved by reciprocal tBLASTn/ BLASTx searches of *Karenia brevis* MMETSP libraries. Considering a concatenated alignment of the five protein-coding genes, alongside equivalent sequences from 19 haptophyte plastid genomes and transcriptomes (Dorrell et al, 2017; Strassert et al, 2021), we obtained a tree topology in which *Karlodinium micrum* and all three species of *Karenia* form a monophylum sister to all haptophytes except Pavlovales and rappemonads, while *Takayama* and *Karlodinium armiger* fall outside this clade (Fig. 5). *Takayama* resolves inside Phaeocystales with 95% bootstrap support. The position of *Karlodinium armiger* is likewise distinct from *Karlodinium micrum* and *Karenia* (90% bootstrap support) and inside the Prymnesiales (86% bootstrap support). Much more densely-sampled 16S rDNA sequence trees broadly support these placements, albeit with weaker bootstrap support, and ambiguity over the monophyly versus polyphyly of the *Karenia*/ *Karlodinium veneficum* grouping (Dataset EV10).

Global distribution of fucoxanthin dinoflagellates reveals negative co-occurrence between Kareniaceae and relatives of their plastid donors

To provide biogeographical and ecological context for the interpretation of our transcriptomic and phylogenetic results, we investigated the distribution of *Karenia*, *Karlodinium*, and *Takayama* genera in the *Tara* Oceans database using the V9 18S metabarcoding data (Dataset EV4). Our results highlight different patterns: *Karenia* seems to contribute the most to dinoflagellate populations, although this may be biased by the different ribosomal DNA copy numbers in different genera (Galluzzi et al, 2010). Concerning distributions, *Karlodinium* clearly differs from the other two genera in its higher overall station occupancy and more cosmopolitan presence, which is noticeable especially in the Indian and Arctic Oceans (Fig. 6). Partial least square (PLS) analysis showed a negative correlation of phosphate, nitrate, and iron to the abundances of both *Karlodinium* and *Takayama*, whereas *Karenia* showed positive correlations to these variables, alongside temperature and oxygen (Appendix Fig. S16).

A second series of PLS analysis of the abundances of the three kareniacean genera to one another revealed a positive correlation of *Takayama* with both *Karenia* and *Karlodinium*, although no internal correlations between the latter two genera (Appendix Fig. S17). Strikingly, all three genera had strong negative correlations to the relative abundances of six haptophyte subgroups (Chrysochromulinaceae, Phaeocystales, Isochrysidales,

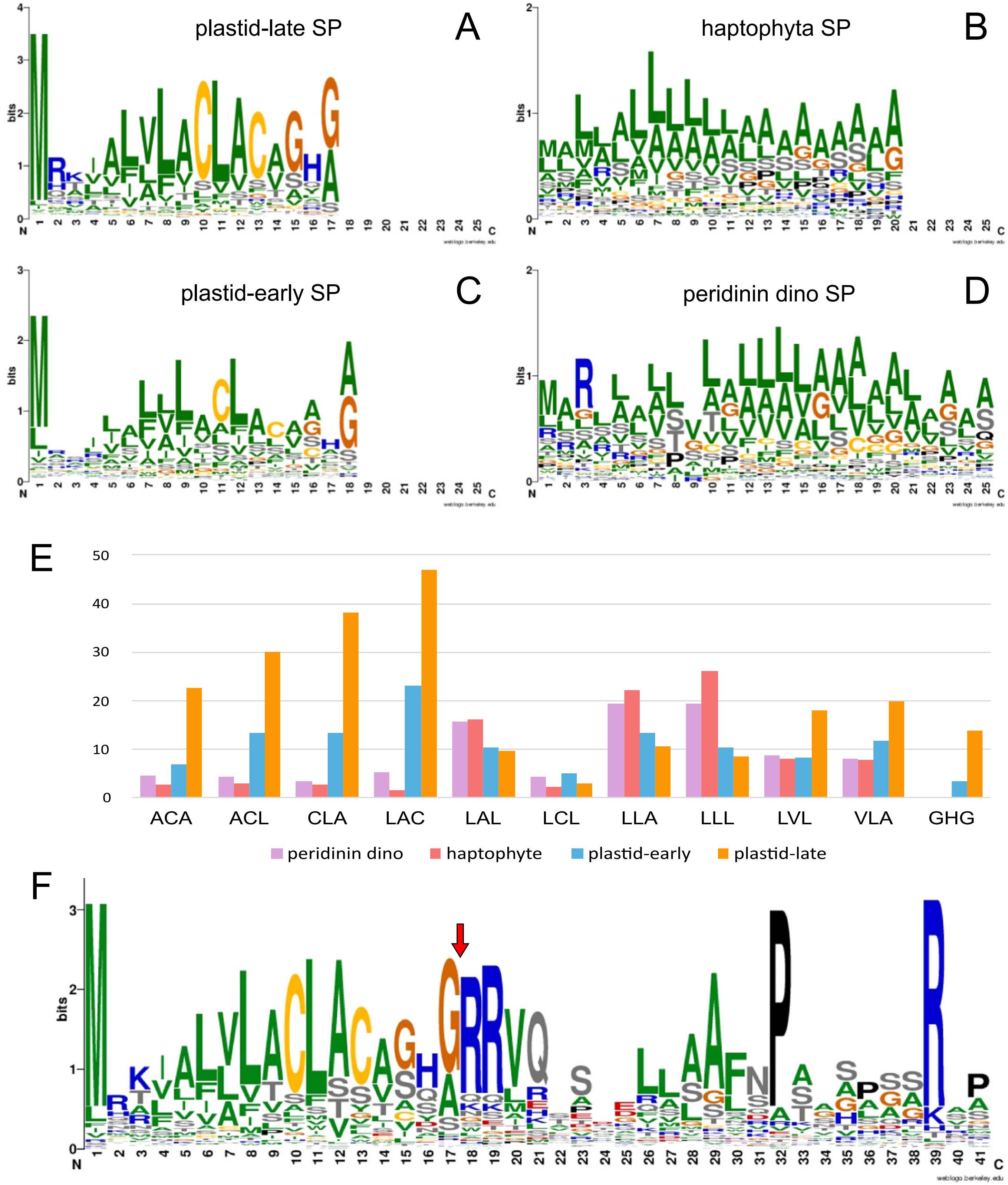


Figure 4. Sequence logos for aligned signal peptides associated with kareniacean plastid proteomes.

These plots show comparable signal peptides for four datasets of different evolutionary identity: haptophyte, peridinin dinoflagellate, plastid-late and plastid-early kareniacean signal peptides (**A–D**), occurrences of three-letter sequence motifs in signal peptides of each dataset, represented as a percentage of signal peptides in which at least one such motif occurs (**E**), and the sequence logo of the signal peptide and partial transit peptide for kareniacean plastidial proteins regardless of their phylogenetic origin with a red arrow denoting the signal peptide cleavage site (**F**).

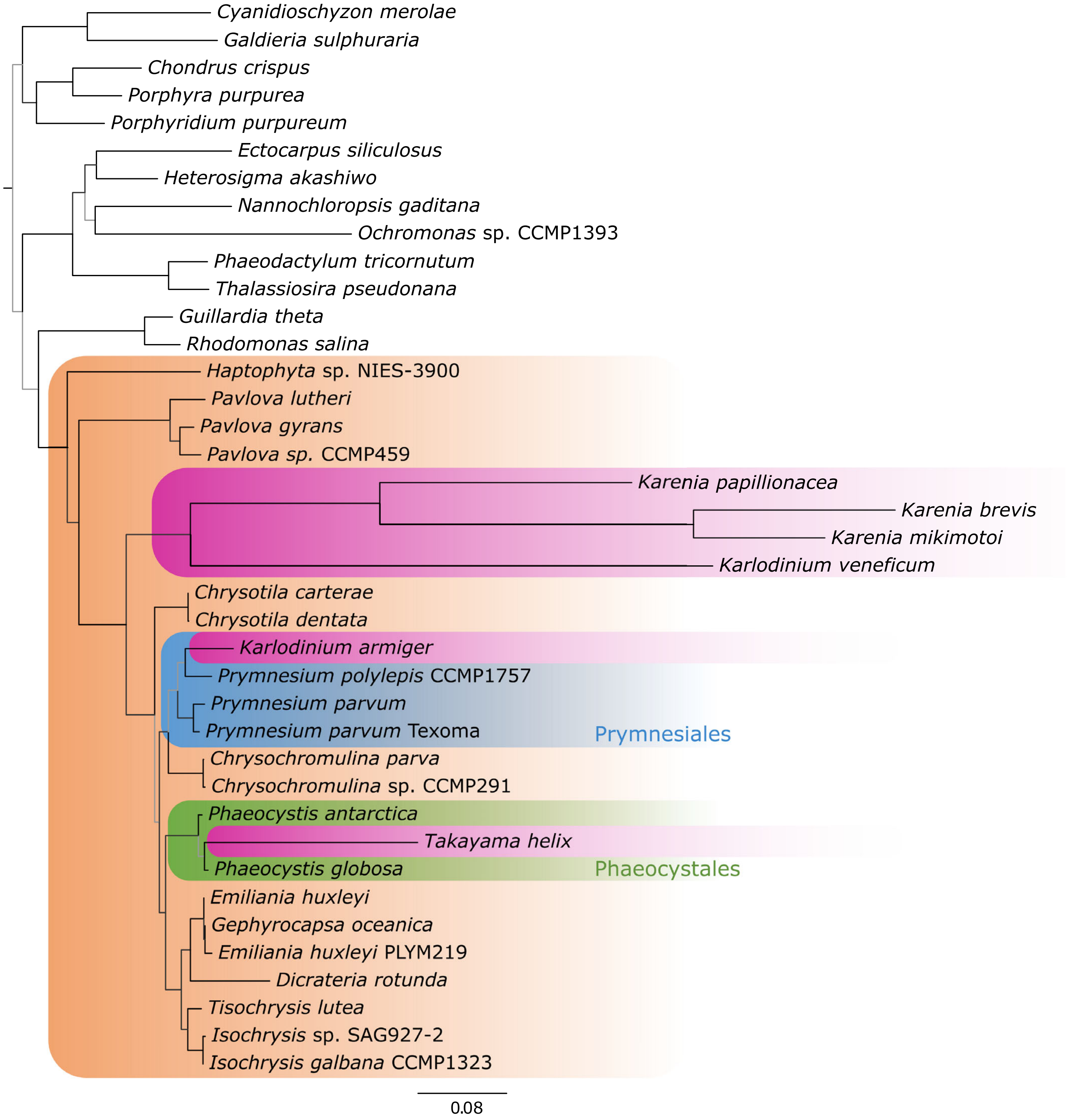


Figure 5. Polyphyletic origins of kareniacean plastid genomes.

Phylogenetic tree reconstructed by IQ-TREE based on five plastid-encoded proteins (PsaA, PsbA, PsbC, PsbD, and RbcL) of six kareniaceans and 20 haptophytes for which partial or complete plastid genomic datasets are available, alongside 13 further algae with primary and complex red plastids; bootstrap support is expressed by the branch color (black for ≥90%, dark gray for ≥75%, and light gray for <75%). Phylogeny based on plastid genome sequences does not support the monophyly of kareniacean plastids and their close relationship to Chrysochromulinaceae retrieved based on nucleus-encoded plastid proteins. Instead, *Takayama helix* and *Karlodinium armiger* split from the rest, placing within Phaeocystales and Prymnesiales, respectively, with high support.

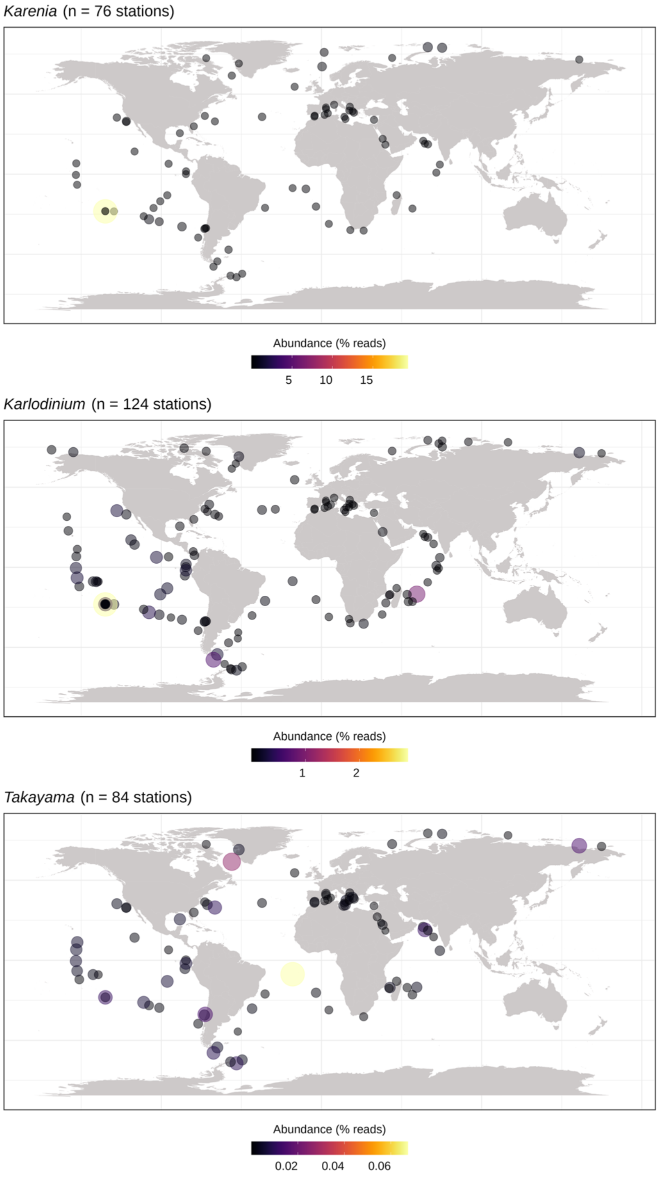

Karenia (n = 76 stations)
Abundance (% reads)
5
10
15
Karlodinium (n = 124 stations)
Abundance (% reads)
1
2
Takayama (n = 84 stations)
Abundance (% reads)
0.02
0.04
0.06

Figure 6. Relative abundances of the three studied kareniacean genera in *Tara* Oceans stations based on ribosomal 18S SSU V9 metabarcoding—*Karenia* exhibits higher abundance compared to both *Karlodinium* and *Takayama*, the latter displaying the lowest.

Karlodinium exhibits more cosmopolitan distributions in comparison to the others. A Correlation heatmap of each species against one another, showing a significant positive correlation between the relative abundances of *Karlodinium* and *Takayama* is provided in Appendix Fig. S17, while correlation circles against individual environmental variables are provided in Appendix Fig. S16.

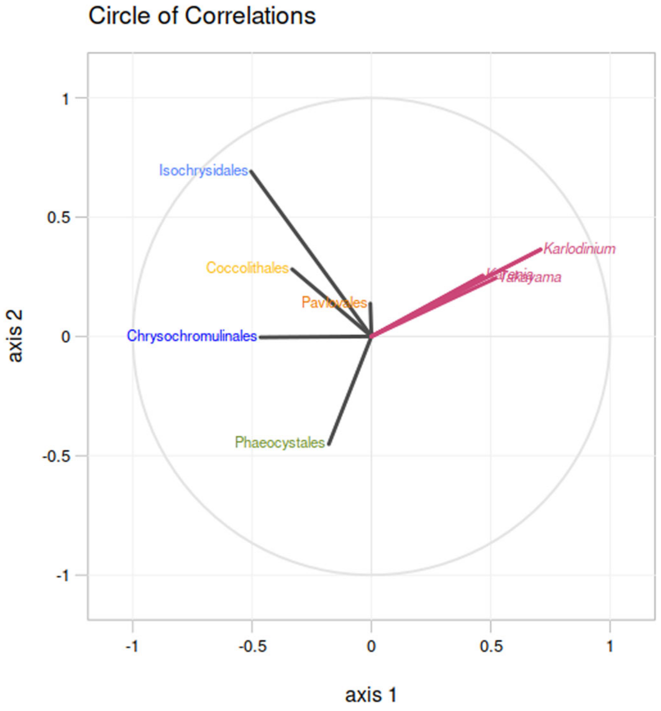


Figure 7. Non-coincidence of kareniaceae and haptophytes in global metagenome datasets.

Correlation circle based on the partial least square analysis of the *Tara* Oceans relative abundances (surface depth only) of the three studied kareniacean genera against those of different haptophyte families (color-coding is the same as in Fig. 1B) as observable variables; all three kareniaceans are negatively correlated to all families and most strongly to Chrysochromulinaceae.

Coccolithales, and Pavlovales), with the strongest negative correlations observed between *Karlodinium*, Chrysochromulinaceae, and Phaeocystales (Fig. 7).

Discussion

In this study, we harness novel transcriptomes, customized signal prediction, automated phylogenetic pipelines, and metabarcoding analyses from *Tara* Oceans to explore the evolution of the fucoxanthin-containing plastids of the kareniacean dinoflagellates, a system of special evolutionary importance for understanding endosymbiosis. Our data emphasize the relaxed nature of plastid acquisition and loss across Kareniaceae and their importance as a model system for studying the fundamental mechanistics and micro-evolutionary variation in plastid evolution.

The idea that individual genes encoding plastid-targeted proteins may exhibit evolutionary affiliation with other groups than the plastid donor, typifying the "shopping bag" model (Larkum et al, 2007), is well-established in many plastid lineages. Our data show that the dinoflagellate host was the principal contributor of nucleus-encoded proteins supporting the kareniacean plastid proteome (Fig. 1A). We do not detect higher numbers of proteins of Phaeocystales origin in *Takayama* or of Prymnesiales origin in *Karlodinium armiger*, in contrast to the origins of their current plastid genomes (see below). Both of these results typify the fact that all of these plastids represent replacements of previous organelles of the same type rather than complete innovations, and were therefore able to build upon a set of pre-existing genes and adaptations. This explains the observed "momentum" of the nucleus-encoded plastid proteome in comparison to the organelle itself.

Nonetheless, based on a large-scale sorting based on tree topology and manually curated phylogenies, we propose a specific haptophyte subgroup, the Chrysochromulinaceae, as a major contributor to the plastid-late part of kareniacean plastid proteomes. Combined with the frequent monophyly of genes of this origin (Figs. 2 and EV2), this strongly suggests that a common ancestor of the studied organisms either possessed a stable plastid or had a long-term symbiotic relationship (e.g., kleptoplastidic) with a haptophyte lineage related to the extant Chrysochromulinaceae.

The evolutionary patterns in the compiled metabolic pathways provide further insights into the specific roles of endosymbiotic gene transfer (EGT), LGT and host-derived proteins in the early evolution of the fucoxanthin plastid (Fig. 3). The preferential incorporation of plastid-late pathways into photosynthesis-associated and ribosomal pathways may reflect that these functions were not encoded in the kareniacean host and that the new plastid (or much less frequently LGT) was the only available source. It is also possible that there has been a direct niche competition between the peridinin and fucoxanthin plastid that may have coexisted in the same host for a period of time with possibly different selective pressure on the retention of individual imported proteins. For example, there may be a selective advantage to plastid-late photosystem and ribosomal proteins interacting with plastid-encoded proteins of the same origin.

In contrast, the retention of dinoflagellate-derived proteins in other pathways (e.g., isoprenoid, heme, and amino acid biosynthesis) implies that the kareniacean ancestor retained these functions. We note that these functions are also often associated with protist lineages that have secondarily lost photosynthesis but retain vestigial plastids (Hadariová et al, 2018; Záhonová et al, 2018; Dorrell et al, 2019). It has previously been suggested that the RSD retains a non-photosynthetic form of the peridinin plastid (Hehenberger et al, 2019), and a similar organelle might have been the donor of the plastid-early pathways. One interesting case illustrating this scenario is the plastid-early chlorophyllase, the only enzyme of this origin in the chlorophyll branch of tetrapyrrole metabolism, which might have been retained by a heterotrophic kareniacean ancestor for the purpose of phytol recycling from chlorophyll contained in its algal prey. Many of the plastid-early gene trees copy the organismal topology and may represent proteins vertically inherited from a cryptic peridinin plastid. Nonetheless, the presence (or eventual fate) of the peridinin plastid in fucoxanthin-containing lineages awaits structural confirmation via microscopy.

The apparent retention of ancestral RNA processing pathways (e.g., poly(U) tail addition and substitutional editing) in the fucoxanthin plastid is puzzling, given the predominantly plastid-late origins of plastid translation. If the Kareniacean ancestor possessed a non-photosynthetic plastid, it would be expected to lose its plastid genome and associated expression machinery as the peridinin plastid only encodes core photosystem subunits (Dorrell and Howe, 2012; Jackson et al, 2013; Cooney et al, 2022). That said, RNA editing is widespread in dinoflagellate mitochondrial and nuclear genomes (Lin et al, 2002; Liew et al, 2017), and is also present in the green plastids of *Lepidodinium* that also secondarily replaced the peridinin-containing ones (Matsuo et al, 2022). It remains to be determined if the fucoxanthin plastid RNA processing machinery is of peridinin plastid or non-plastidial host origin.

These general trends aside, the plastid metabolic map is generally divided between parts strongly enriched in plastid-late proteins and parts assembled from both sources, likely in an evolutionarily neutral, patchwork-like way (Fig. 3). This pattern does not only manifest at the level of one pathway being formed by enzymes of different evolutionary affiliations but often also at the level of one step in the pathway being carried out by homologs of different origin in different organisms. We propose that fucoxanthin plastid establishment may have been a relatively gradual process during which peridinin-like and haptophyte-like genes coexisted and competed for place, even after the divergence of individual species. This is also reflected in the partial plastid genomes assembled from *Karenia* and *Karlodinium*, which contain partially non-overlapping sets of genes that suggest independent post-endosymbiotic plastid genome reduction (Gabrielsen et al, 2011; Richardson et al, 2014; Dorrell et al, 2016). Our data establishes the importance of species-level variation as playing a substantial role in shaping the kareniacean plastid proteome.

Strikingly, our three phylogenomic analyses based on (a) plastid-unrelated conserved nuclear genes, (b) nucleus-encoded and haptophyte-like plastid genes, and (c) plastid genome-encoded genes, yield contradicting tree topologies (Figs. 5 and EV1,EV2). Kareniacean plastid genomes are split into three lineages, each associated with different haptophytes: Prymnesiales for *Karlodinium armiger*, Phaeocystales for *Takayama*, and a further unresolved but clearly distinct source for the genus *Karenia* and *Karlodinium micrum*. Remarkably, this phylogenetic diversity is in accord with differences in pigment composition between these lineages (Garcés et al, 2006; Zapata et al, 2012).

Furthermore, the phylogenetic signal from kareniacean plastid-targeted proteins, despite suggesting a common origin and evolutionary affinity (Fig. 1; Appendix Fig. S2), does not agree with the organismal topology and suggests a closer association of *Karlodinium* and *Karenia* in comparison to *Takayama* (Fig. 2). This result is further complicated by more recently described kareniacean species that do not possess a fucoxanthin plastid, i.e., RSD (Hehenberger et al, 2019), *Gertia* (Takahashi et al, 2019), and *Shimiella* (Ok et al, 2021) and their phylogenetic positions as closer relatives to *Karlodinium* and *Takayama* to the consistent exclusion of *Karenia*. While *Gertia* retains reduced but photosynthetically active peridinin plastid, RSD has been proposed to possess its relict form (Hehenberger et al, 2019).

In Fig. 8, we present three main hypothetical scenarios attempting to reconcile the seemingly contradicting evolutionary histories of kareniacean nuclei, plastids, and their nuclear-encoded protein inventories (top panels), further discussed below.

Given the shared haptophyte signal, often showing clear monophyly and phylogenetic association with Chrysochromulinaceae, in all organisms, we propose that they indeed possessed the same type of fucoxanthin plastid, already a replacement of the ancestral peridinin one, at some point of their history, which was later lost and replaced in *Karlodinium armiger* and *Takayama helix* but conserved in other *Karlodinium* and *Karenia*. Previous studies have noted the extensive capacity of both *Takayama helix* and *Karlodinium armiger* for mixotrophic metabolism (Berge et al, 2008; Berge and Hansen, 2016; Jeong et al, 2021), supporting the capacity for endosymbiotic replacement. *Karlodinium armiger* also produces significantly shorter branches than the others in plastid genome trees, which suggests a recent acquisition.

The scenario we find the most parsimonious and best reflecting the results of our phylogenetic analyses is the one where the fucoxanthin plastid was acquired by the common ancestor of

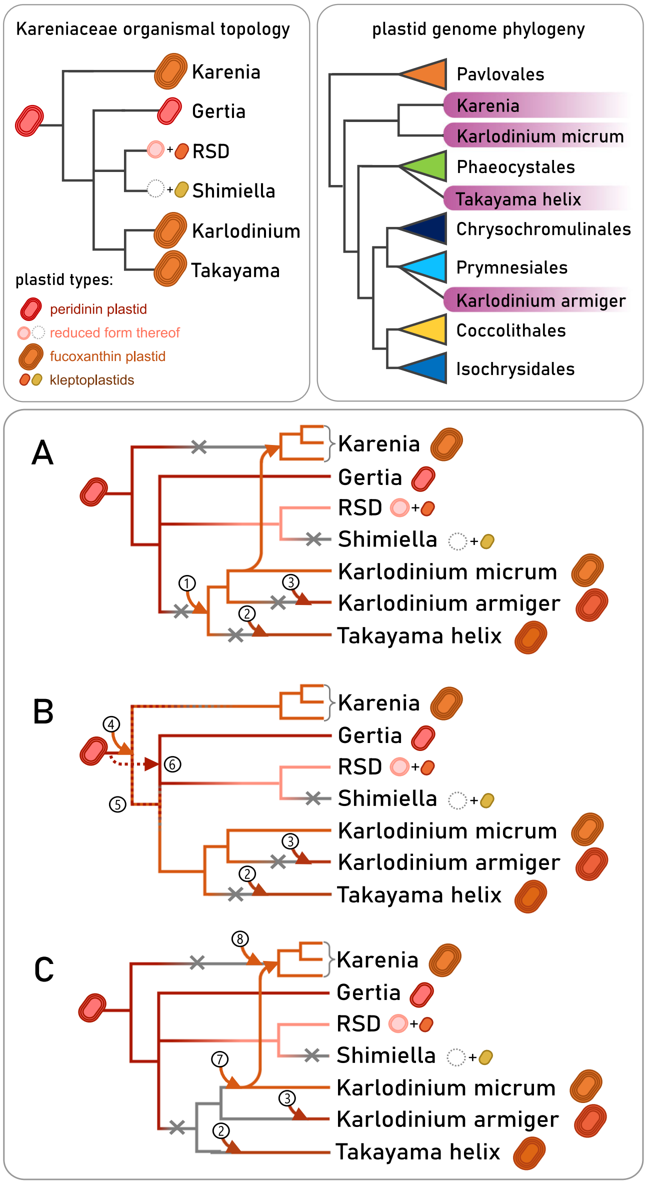


Figure 8. Proposed scenarios of plastid evolutionary history in Kareniaceae integrating previously published organismal topology with our data based on plastid protein and plastid genome phylogeny.

Scenario (**A**) proposes that the first fucoxanthin plastid was acquired by the common ancestor of *Karlodinium* and *Takayma* (1) and later transferred to *Karenia* and independently lost and replaced in *Takayama* and *Karlodinium armiger* (2, 3). This scenario best explains the shared Chrysochromulinaceae-like signal and tree topologies recovered by our phylogenetic analyses (Figs. 2 and EV2) but requires a higher number of plastid losses. Scenario (**B**) places the fucoxanthin plastid origin at the base of all Kareniaceae (4) and requires the peridinin plastid to either have coexisted with the fucoxanthin one for a period of time (5) or have been lost and re-acquired (6) to explain its presence in *Gertia* and RSD. It proposes less plastid losses and no additional endosymbiotic transfer between fucoxanthin species, but it explains the shared Chrysochromulinaceae-like signal only partially as it is congruent with its monophyly but not the inner relationships. Scenario (**C**) proposes an independent plastid origin for all four lineages (2,3,7,8) or for three lineages with subsequent transfer between *Karlodinium* and *Karenia* (2,3,7) or vice versa (only one direction is shown in the schematic). This scenario considers the Chrysochromulinaceae-like signal a phylogenetic artifact caused by the under-sampling of the plastid donors or sources of non-specific LGT.

Karlodinium and *Takayama* (Fig. 8A) and subsequently underwent an additional endosymbiotic transfer from *Karlodinium* to *Karenia*, explaining their sister position frequently recovered in our trees (Fig. 2). We note that the predator-prey relationships that may have facilitated such transfer are not uncommon even between toxin-producing dinoflagellates (Jeong et al, 2016). Conversely, the smaller size (de Salas et al, 2008) and cosmopolitan distribution of *Karlodinium* might make it a potential prey for *Karenia*, although this awaits direct observation. While it is true that we observe a small number of plastid-late genes shared between the RSD and other kareniaceae, which may be signals from a relict fucoxanthin plastid (cf., peridinin-derived signals in dinoflagellates that have completely lost a plastid; Gornik et al, 2015), this number is very small and may constitute phylogenetic noise.

In an alternative scenario, the fucoxanthin plastid was already present in the common ancestor of all three genera (Fig. 8B), i.e., the ancestor shared with non-fucoxanthin kareniaceans, all of which would have to have lost this plastid. This scenario would also have to involve either a period of coexistence of peridinin and fucoxanthin plastid or a loss and re-acquisition of the peridinin plastid to explain its presence in *Gertia*.

The third possibility considers independent acquisitions in all four lineages (Fig. 8C), with the shared phylogenetic signal representing a phylogenetic artifact stemming from the under-sampling of closely related plastid donors and sources of non-endosymbiotic LGT, or three independent acquisitions and a transfer from *Karlodinium* to *Karenia* or vice versa.

From the diversity of signal sequences targeting proteins into fucoxanthin plastids, we consider possible structural similarities that may hint at their evolutionary origin. The most striking feature of kareniacean signal peptides (both plastid-late and plastid-early proteins) is their high degree of conservation (Fig. 4A–D). At the same time, they are very dissimilar to those of peridinin dinoflagellates and haptophytes. If they were indeed inherited from one or the other, they have clearly evolved profoundly under very strong selection to the point of being more strictly conserved than their ancestors and are untraceable. The other possibility is that they arose *de novo* during fucoxanthin plastid establishment, and their highly similar sequence reflects their common and possibly relatively recent origin and spread, e.g., via gene duplications, domain swapping, or the activity of transposable elements (Burki et al, 2012). The recent assembly of a draft *Karenia brevis* nuclear genome (Nelson et al, 2021) may allow us to assess the genomic context of genes encoding plastid-targeted proteins.

Finally, the *Tara* Oceans distributions of each lineage raise questions about the broader ecological processes underpinning their plastid evolution. Despite their different distributions, all three kareniacean genera show a strong negative correlation with haptophytes. The strong negative association with Chrysochromulinaceae and the lack of negative correlation with Pavlovales are notable in regard to their relative genetic imprints on the kareniacean plastids (highest for the former, lowest for the latter; Cf., Figs. 1B, 7). These anti-correlations might reflect competitive exclusion or active predation avoidance (Lima-Mendez et al, 2015; Vincent & Bowler, 2020) or simply different niche preferences between each group. We note that the basic metabarcoding analysis of the biogeographical distribution performed in this study is preliminary, particularly as it relies on the analysis of entire haptophyte orders. It will therefore be worthwhile in the future to assess the distributions of other more recently developed marker genes (Penot et al, 2022; Pierella Karlusich et al, 2022). The specific ecological interactions between the progenitors of Kareniaceae, and the haptophyte ancestors of their fucoxanthin plastid, may be best inferred via ancestral niche reconstruction for each lineage.

Methods

Reference database preparation

An in-house protein database was used for homology searches throughout this project. This database consisted of 155 transcriptomic and genomic datasets fetched from NCBI (https://www.ncbi.nlm.nih.gov), 1KP (https://db.cngb.org/onekp/; Leebens-Mack et al, 2019), and decontaminated MMETSP (Keeling et al, 2014) as described and in Dorrell et al, 2017: 30 for eubacteria, 10 for archaea, and 115 for eukaryotes with sampling spanning most of the diversity with focus on plastid-bearing groups (100 datasets) and particularly enriched in dinoflagellate and haptophyte sequences (37 and 34 datasets, respectively). The source database can be accessed at https://osf.io/ykxes/, and an annotated list of datasets used in our study is available in Dataset EV2.

Data acquisition and de novo transcriptome sequencing

The previously published peptide sequences for *Karenia brevis*, *Karlodinium micrum* (syn. *veneficum*), and Ross Sea Dinoflagellate (RSD) were downloaded from their respective public repositories (Data refs: Keeling et al, 2014; Ryan et al, 2014, Hehenberger et al, 2019) and decontaminated as described above. In case of *Karenia brevis*, the four available assemblies from four clones were pooled together and treated as a single dataset with redundant sequences removed randomly.

Karenia mikimotoi RCC1513, *Karenia papilionacea* RCC6516 and *Karlodinium* sp. RCC3446 were obtained from Roscoff Culture Collection (https://roscoff-culture-collection.org), *Takayama helix* CCMP3082 was obtained from the Bigelow National Center for Marine Algae and Microbiota (https://ncma.bigelow.org), and *Karlodinium armiger* K0668 was obtained from the Norwegian Culture Collection of Algae (https://niva-cca.no). The cultures were grown in natural seawater (Horsey Mere seal colony, North Sea) supplemented with k/2 or enriched artificial seawater on a 12 h light (50μE): 12 h dark cycle at 19 °C for 4 weeks, to late-exponential phase (1–2 million cells/ml). Cells were concentrated by centrifugation at 4000 rpm for 10 min, washed three times in sterile marine PBS (mPBS), and snap-frozen in liquid nitrogen. Total cellular RNA was extracted as previously described (Dorrell and Howe, 2012) using Trizol (Invivogen) and chloroform phase extraction, isopropanol precipitation, and resuspended in nuclease-free water (Qiagen). RNA integrity was confirmed by electrophoresis on ethidium bromide-stained agarose gels and quantified using a NanoDrop photospectrometer (Thermo Fisher). Two μg RNA was subsequently treated with 2U RNA-free DNAse (Promega) following the manufacturers' instructions, precipitated with isopropanol, resuspended in nuclease-free water, and quality-checked and quantified again.

For each species, 1 μg DNase-treated RNA was polyA-tail selected, and strand-specific libraries were prepared with TruSeq Stranded mRNA Library Prep kit (Illumina). The resulting libraries were sequenced using a Novaseq flow cell (Fasteris) with 100 bp paired reads, followed by in-house base-calling and basic Q30 enumeration (https://www.fasteris.com).

The obtained paired-end reads were further quality controlled using the RNA-QC-chain toolkit (Zhou et al, 2018) by filtering based on phred score with the threshold of 30; and GC content with the allowed range set to 42–60% based on the expected values for kareniacean dinoflagellates (Lidie et al, 2005), and by removing reads potentially representing rRNA contamination (default settings were used for this step). Filtered reads were then assembled by Trinity (Grabherr et al, 2011) with --NO_SEQTK and --trimmomatic options, and the resulting assembly was selected for the longest isoforms by the respective script from the Trinity toolkit. The quality and completeness of the transcriptomic datasets were assessed by TrinityStats and BUSCO (eukaryota_odb10, n:255) (Simão et al, 2015). The datasets were translated to proteins by TransDecoder (with default settings; https://github.com/TransDecoder/TransDecoder) and subsequently decontaminated by homology search using LAST (Kiełbasa et al, 2011) against the in-house database and removing sequences homologous to eubacterial, archaeal, or metazoan sequences with e-value of 1E-10 or lower. The translated datasets were used for further analyses alongside the previously published transcriptomes in their translated forms and are available at https://figshare.com/articles/dataset/full_protein_datasets_tar/21647771. The sequence data were deposited at NCBI under BioProject ID PRJNA788777.

Plastid signal prediction

First, the relative effectiveness of two SignalP versions (3.0 and 5.0) were determined for discriminating plastid-targeted proteins using a model dataset (Appendix Fig. S20; Dataset EV6). SignalP 5.0 (Almagro Armenteros et al, 2019) was determined to possess greater specificity for the identification of plastid-targeted proteins (22/728 non-plastid control proteins predicted to possess signal peptides, compared to 34 for SignalP 3.0) and used for all subsequent analysis.

Both original sequences and sequences cut to the first methionine were analysed to accommodate for potential translated UTRs or spliced leaders, and the results were pooled afterward prioritizing the positive signal peptide prediction in case of different results. The obtained SignalP results were converted into a tabular format compatible with SignalP versions 3.0–4.1 for downstream applications with a custom script (Computer code EV1). For the second step predicting the rest of the bipartite targeting signal, a modified version of ASAFind (Gruber et al, 2015; Füssy et al, 2019) was used with a custom scoring matrix (Computer code EV1; Dataset EV5) that reflects the differences in sequence surrounding the signal peptide cleavage site of kareniaceans (for instance, instead of the phenylalanine motif, multiple arginine residues are typically present). The matrix was prepared based on sequence logos prepared by Seq2Logo (Thomsen and Nielsen, 2012; Appendix Fig. S18) from model datasets of *Karenia brevis* and *Karlodinium micrum* proteins determined as having exclusively plastidial, mitochondrial, nuclear, or endomembrane functions, e.g., photosynthetic core machinery proteins for plastidial functions, and identified from reciprocal BLAST against well-curated proteomic datasets from other organisms (Butterfield et al, 2013; Klinger et al, 2013; Dorrell et al, 2017; Bannerman et al, 2018; Beauchemin and Morse, 2018), the queries and retrieved model sequences are available in Dataset EV5. The modified script and custom scoring matrix were subsequently tested and optimized for sensitivity and specificity, and in the final version represented an improvement of 56.4 and 0.5 percentage points, respectively, over unmodified ASAFind. The detection thresholds used with SignalP were also tested, with no apparent improvement in specificity/ sensitivity tradeoffs over the default settings; as well as prediction by HECTAR and a combination of PrediSI and ChloroP (Gschloessl et al, 2008; Dorrell et al, 2019; Ebenezer et al, 2019; Appendix Figs. S19, S20; Dataset EV6).

Additionally, we employed an alternative prediction combining PrediSI (Hiller et al, 2004) and ChloroP (Emanuelsson et al, 1999) with the intention to capture potential proteins with peridinin plastid-like signal domains similar to those of *Euglena* on which it was previously tested (Patron and Waller, 2007; Ebenezer et al, 2019). Proteins judged as not coding for a plastid-targeting sequences by the modified ASAFind but identified as plastid-targeted by this approach were included in the predicted plastid proteomes as well, albeit with prefix “SPTP-” denoting the presence of an alternative signal.

The presence of large numbers of highly similar sequences caused by very recent paralog duplication and/or alternative splicing is common in dinoflagellates. To reduce this complexity for phylogenetic analyses, all datasets were additionally clustered by cd-hit with an identity threshold of 0.9. This was performed after signal prediction to avoid accidentally discarding an isoform/ paralog with a predictable targeting signal in favor of one without it.

Phylogenetic analysis

Each of the predicted plastid proteomes was processed independently in order not to omit any variant signals in individual proteins. First, homologs were mined by LAST (Kiełbasa et al, 2011) search with threshold e-value 1E-10 against the in-house database plus all the fucoxanthin dinoflagellate transcriptomes. The best hit from all organisms in the in-house dataset were retained alongside the best five hits from each other fucoxanthin dinoflagellate library to account for potential paralogs in fucoxanthin dinoflagellate transcriptomes. Proteins with no hits whatsoever or with no homologs in any taxa other than dinoflagellates (Kareniaceae or other) were annotated as “lineage-specific” or “dinoflagellate-specific” respectively, and their phylogeny was not investigated further. The remaining homolog sets were aligned by MAFFT (version 7.310, Mar 17 2017; Katoh and Standley, 2013) with default settings and trimmed by TrimAl (version 1.4, Dec 17 2013; Capella-Gutierrez et al, 2009) with -gappyout parameter. The trimmed alignments were then filtered by a custom python script that discarded sequences comprising of more than 75% gaps, and then rejected alignments shorter than 100 positions or containing fewer than 10 taxa. The overall statistics of these alignments (distributions of their total lengths, numbers of sequences, and percentages of gaps) are provided in Appendix Figs. S21–S23. The satisfactory alignments were then used for maximum-likelihood tree construction by IQ-TREE (multicore version 2.0.5, May 15 2020; Minh et al, 2020) with automatic model selection and 1000 ultrafast bootstraps. For a graphical summary of this pipeline, see Appendix Fig. S1. All trees are available at https://figshare.com/articles/dataset/all-automatically-built-trees_tar/21647768; in-house scripts used in the pipeline are available among supplementary files (Computer code EV1); as well as the final predicted plastid protein datasets with their phylogenetic annotations (Dataset EV9).

TreeSorter, a freeware program written in Python (https://github.com/vanclcode/treesorter/), builds abstract tree structures from a tree file in Nexus format. It then takes arguments defining sets of taxa, and criteria based on those sets, that determine what constitutes a valid subtree. Each tree is traversed and the highest bootstrap value of an edge (bipartition) that produces a subtree containing the taxon of interest and matching the defined criteria is reported. Criteria allow for further quantification of taxon sets, such that the minimum and/or maximum number or proportion of taxa from a particular set per subtree. TreeSorter was used for large-scale analysis of the tree topologies and sorting and annotating the seed plastid proteins by their probable evolutionary origin into custom categories: haptophyte (plastid-late), dinoflagellate (plastid-early), alveolate (non-plastid). Further signals considered included potential LGTs/EGTs from green (i.e., green algal, plant, or chlorarachniophyte), brown (i.e., ochrophyte), or prokaryotic sources as sequences of these evolutionary origins were noticed in these organisms in previous studies (Nosenko et al, 2006; Waller et al, 2006); all remaining trees for which one of the above annotations could not be identified (i.e., trees with no clear phylogenetic signal) were classed as other/unresolved. In order to be assigned to one of the categories, the seed protein had to be either sister to or nested within a clade comprising exclusively members of said category and of at least 2 taxa. The bootstrap value of the highest-supported branch dividing such clade from the rest of the tree was then saved as the numerical score of this sorting result. In case the protein can be assigned to more than one category, the one with a higher bootstrap was selected, and if they had equal scores, the protein was considered unresolved. In the case of nested categories (such as dinoflagellates and alveolates), the wider category is selected. Bootstrap scores of 75 were generally considered as the lower threshold in interpreting the evolutionary origins, however under specific circumstances (e.g., when the same protein was assigned the same origin with score ≥75 in other studied species), lower scoring results are reported as well.

The predicted plastid proteomes were automatically annotated by KAAS and major metabolic pathways were reconstructed using KEGG Mapper (https://www.genome.jp/kegg/mapper/reconstruct.html). Additional functions and missing enzymes were manually identified by targeted HMMER, based on their conserved domains as detected by PfamScan (Finn et al, 2014).

Phylogenomic analysis

First, a phylogenetic matrix for non-plastid proteins (72,162 positions) was prepared using the standard workflow of the PhyloFisher toolkit (Tice et al, 2021), with the manual step of paralog annotation by parasorter, and removal of sequences containing >66% of gaps; the matrix constructor statistics, input metadata, and the matrix itself are available as supplementary files (Dataset EV10).

Manually curated phylogenetic matrices for plastid-targeted proteins were prepared from genes that passed our criteria of (a) tree-based evidence of plastid-late origins, (b) occurrence in at least 5 of the 7 organisms, and (c) manually verified functional annotations and plastid-associated functions: 1-acyl-sn-glycerol-3-phosphate acyltransferase, 3-hydroxyacyl-ACP dehydratase, 3-oxoacyl-ACP synthase, ACP, cysteinyl-tRNA synthetase, digalactosyldiacylglycerol synthase, glutaminyl-tRNA synthetase, heme oxygenase, chlorophyll a synthase, chlorophyllide b reductase, LHCA1, lycopene beta cyclase, magnesium chelatase subunit H, magnesium-protoporphyrin *O*-methyltransferase, protochlorophyllide reductase, PsbO, PsbP, PsbU, RP-L1, RP-L13, SecA, SufC, and SufD. The alignments, single-gene and concatenated trees are available in supplementary files (Dataset EV10).

Next, a phylogenetic matrix for plastid-encoded proteins (2404 positions) was prepared using five genes (*PsaA*, *PsbA*, *PsbC*, *PsbD*, and *RbcL*) obtained from previously published plastid genomes and sequences encoded in MMETSP transcriptomes and *de novo* sequencing of PCR products (Dorrell et al, 2017), and inspected with IQ-TREE as above. Preliminary single-gene trees for the plastid-targeted proteins were manually checked to remove paralogs, very long branches, and likely cases of LGT, if present; these trees with the removed taxa highlighted, as well as all original and trimmed datasets, and both final matrices are available as supplementary files (Dataset EV10). Finally, trees based on a single gene for 16S rRNA with up to 644 haptophyte sequences, including environmental data (Choi et al, 2017), were constructed using various alignment, trimming, and models (details in the supplementary files).

Concatenated trees were then constructed by IQ-TREE with the LG + C60 + F model for the plastid matrices and posterior mean site frequency (PMSF) model (LG + C60 + F + G with a guide tree constructed with C20) for the PhyloFisher matrix (Wang et al, 2018).

Signal peptide analysis

Four experimental datasets for this analysis consisted of two sets of plastid-targeted plastid-early (264) and plastid-late (437) proteins with assigned KEGG annotation with uniquely plastidial function and two analogous sets selected from MMETSP transcriptomes for five peridinin dinoflagellates (358) and five haptophytes (272) (details in Dataset EV7) based on presence of (a) detectable signal peptide and (b) the same KEGG annotations to produce comparable datasets without functional bias. Signal peptide regions of these proteins were determined by SignalP 5.0 (the region of ~25 positions immediately following the predicted cleavage site was considered as putative partial transit peptide region). The regions were aligned by ClustalW, manually curated to keep their conserved part and used to create sequence logos for each dataset using WebLogo (https://weblogo.berkeley.edu/logo.cgi; Crooks et al, 2004). The overall amino acid composition and number of occurrences of three-letter sequence motifs suggested by the logos in each dataset was also determined.

Metabarcoding distributions analysis

To explore the distribution of the three Kareniaceae genera, the V9 18S rDNA metabarcoding data from *Tara* Oceans corresponding to the genera *Karenia*, *Karlodinium* and *Takayama* were extracted from (Ibarbalz et al, 2019) (available at https://zenodo.org/record/3768510#.Ye-xqpHMJzp). Overall, a total of 69 barcodes were extracted for *Karenia*, 76 for *Karlodinium*, and 131 for *Takayama*. Their respective distributions (at least equal to three reads) in the eukaryotic-enriched size fractions (Pierella Karlusich et al, 2019) (0.8–5 μm; 5–20 μm; 20–180 μm, and 180–2000 μm) and two depths (surface and deep-chlorophyll maximum (DCM)) were pooled together and normalized against the total reads assigned to dinoflagellates, extracted as above, within each sample to account for their respective contribution to dinoflagellate populations. Results were plotted using the R package “mapdata” v2.3.0 (http://cran.nexr.com/web/packages/mapdata/index.html). The dinoflagellate occupancy was estimated as the percentage of stations in which their respective barcodes were identified compared with the total stations investigated.

A correlation analysis was conducted on the pooled Kareniaceae size fractions and depths, including a total of 76 stations in which at least one of the organisms was retrieved, with the R package “corrplot” (Wei and Simko, 2021). Partial least square analyses were implemented with the R package “plsdepot” (https://github.com/gastonstat/plsdepot) to explore the correlations between each kareniacean genus and a diverse set of environmental variables (range-transformed median values of nutrient concentrations of iron, nitrate, silicate, phosphate; and oxygen, salinity, temperature, all extracted from the PANGAEA database; Ardyna et al, 2017; Guidi et al, 2017a, 2017b). A second set of partial least square analyses were performed to explore the correlations of each kareniacean genus to each other, and to the total relate abundance of V9 metabarcode data for five haptophyte orders (Chrysochromulinaceae, Coccolithales, Isochrysidales, Pavlovophyceae, and Phaeocystales) as described in Penot et al, 2022. Abundances were all Hellinger-transformed. Abundances and environmental data considered for the partial least square analyses were computed only for the surface layer with all four size fraction abundances pooled together.

Plastid transcript amplification and sequencing

Nucleotide sequences corresponding to plastidial transcripts from MMETSP libraries of *Karenia brevis* and *Karlodinium micrum* were identified by tBLASTn searches using a composite sequence library of predicted plastid-encoded proteins from three haptophyte species (*Emiliania huxleyi*, *Phaeocystis antarctica*, *Pavlova lutheri*) and one kareniacean dinoflagellate (*Karenia mikimotoi*) with threshold e-value 1E-05, following previous studies (Dorrell et al, 2016; Dorrell et al, 2017) (Dataset EV8). Retrieved transcripts were first validated by BLASTx search against the entire NCBI database (accessed December 2017) to exclude bacterial, mitochondrial, chimeric, and artifactual BLAST hits, and then assembled by species and genes into contigs using the in-built sequence assembly function in Geneious vR10.0.9 (Kearse et al, 2012). Transcript sequences were trimmed to remove non-conserved nucleotides and inferred frameshifts prior to conceptual translation with a standard plastid translation table. Identified transcript sequences are provided in Dataset EV8.

cDNA was synthesized from 200 ng DNase-treated *Karenia papillionacea*, *Karlodinium armiger*, and *Takayama helix* RNA using Maxima First Strand cDNA Synthesis Kit for RT-qPCR (Thermo Fisher) following the manufacturer’s instructions. cDNAs corresponding to six plastid-encoded marker genes (*psaA*, *psbA*, *psbC*, *psbD*, *rbcL*, and 16 S rDNA/ *rrnS*) sequenced in previous studies of *Karenia mikimotoi* and *Karlodinium micrum* (Dorrell and Howe, 2012; Richardson et al, 2014) were amplified from each species using DreamTaq PCR MasterMix (Thermo Fisher), following the manufacturer's instructions. Consensus primers (typically: four forward and three reverse primers, with an additional oligo-d(A) reverse primer that anneals to transcript poly(U) tails) were designed from alignments of *Karenia mikimotoi*, *Karenia brevis*, and *Karlodinium micrum* RNA sequences as inferred above (Dorrell et al, 2016), and are provided in Dataset EV8. PCR conditions were typically: 10 min at 95 °C, followed by 35 cycles of 45 s at 95 °C, 45 s at 55 °C, and 2 min at 72 °C, with a terminal 5-min elongation at 72 °C. Different combinations of forward and reverse primer were tested for each gene, and for specific genes, the annealing phase temperature was modified on a gradient between 48 °C and 60 °C.

Products of the expected size, as inferred from agarose gel electrophoresis, were purified using a DNA column spin kit (Macherey-Nagel). In certain cases, products were cloned into pGEM-T Easy vector (Promega) and transformed into competent *E. coli* cells, with inserts amplified using standard T7 and SP6 primers and purified as before. Purified products were Sanger sequenced using each PCR primer under standard conditions (Eurofins genomics). 18S rDNA and ITS1 sequences were amplified and sequenced above, using consensus primers as provided in Gachon et al, 2013 (Dataset EV8). Novel sequences associated with this project are publicly available from GenBank under sequence IDs OQ679836-OQ679852 and OQ680141-OQ680143.

Data availability

The datasets and computer code produced in this study are available in the following databases: • translated transcriptomic datasets: https://figshare.com/articles/dataset/full_protein_datasets_tar/21647771. • raw reads and nucleotide assemblies: NCBI SRA BioProject ID PRJNA788777 (https://www.ncbi.nlm.nih.gov/bioproject/788777). • trees produced by the automatic pipeline: https://figshare.com/articles/dataset/all-automatically-built-trees_tar/21647768. • alignments produced and used by the automatic pipeline: https://figshare.com/articles/dataset/all-automatically-generated-alignments_rar/24347032. • predicted plastid protein datasets with phylogenetic annotations: https://figshare.com/ articles/dataset/phylo-annotated_proteins_sets/24407914. • genomes and transcriptomes used for homolog mining: https://osf.io/ykxes/. • automatic tree sorting script: https://github.com/vanclcode/treesorter/.

Expanded view data, supplementary information, appendices are available for this paper at https://doi.org/10.1038/s44319-024-00103-y.

Peer review information

A peer review file is available at https://doi.org/10.1038/s44319-024-00103-y

References

Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol 37:420–423

Ardyna, M, D'Ovidio, F, Speich, S, Leconte, J, Chaffron, S, Audic, S, Garczarek, L, Pesant, S, Tara Oceans Consortium, C., & Tara Oceans Expedition, P. (2017). Environmental context of all samples from the Tara Oceans Expedition (2009-2013), about mesoscale features at the sampling location. In: Tara Oceans Consortium, Coordinators; Tara Oceans Expedition, Participants (2017): Registry of all samples from the Tara Oceans Expedition (2009-2013).

Bannerman BP, Kramer S, Dorrell RG, Carrington M (2018) Multispecies reconstructions uncover widespread conservation, and lineage-specific elaborations in eukaryotic mRNA metabolism. PLoS ONE 13:e0192633

Beauchemin M, Morse D (2018) A proteomic portrait of dinoflagellate chromatin reveals abundant RNA-binding proteins. Chromosoma 127:29–43

Bentlage B, Rogers TS, Bachvaroff TR, Delwiche CF (2016) Complex ancestries of isoprenoid synthesis in dinoflagellates. J Eukaryot Microbiol 63:123–137

Berge T, Hansen PJ (2016) Role of the chloroplasts in the predatory dinoflagellate *Karlodinium armiger*. Mar Ecol Prog Ser 549:41–54

Berge T, Hansen PJ, Moestrup Ø (2008) Feeding mechanism, prey specificity and growth in light and dark of the plastidic dinoflagellate *Karlodinium armiger*. Aquat Microb Ecol 50:279–288

Boldt L, Yellowlees D, Leggat W (2012) Hyperdiversity of genes encoding integral light-harvesting proteins in the dinoflagellate symbiodinium sp. PLoS ONE 7:1–13

Burki F, Hirakawa Y, Keeling PJ (2012) Intragenomic spread of plastid-targeting presequences in the coccolithophore *Emiliania huxleyi*. Mol Biol Evol 29(9):2109–2112

Burki F, Imanian B, Hehenberger E, Hirakawa Y, Maruyama S, Keeling PJ (2014) Endosymbiotic gene transfer in tertiary plastid-containing dinoflagellates. Eukaryot Cell 13:246–255

Butterfield ER, Howe CJ, Nisbet RER (2013) An analysis of dinoflagellate metabolism using EST data. Protist 164:218–236

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973

Choi CJ, Bachy C, Jaeger GS, Poirier C, Sudek L, Sarma VVSS, Mahadevan A, Giovannoni SJ, Worden AZ (2017) Newly discovered deep-branching marine plastid lineages are numerically rare but globally distributed. Curr Biol 27:R15–R16

Cohen NR, McIlvin MR, Moran DM, Held NA, Saunders JK, Hawco NJ, Brosnahan M, DiTullio GR, Lamborg C, McCrow JP et al (2021) Dinoflagellates alter their carbon and nutrient metabolic strategies across environmental gradients in the central Pacific Ocean. Nat Microbiol 6:173–186

Cooney EC, Leander BS, Keeling PJ (2022) Phylogenomics shows unique traits in Noctilucales are derived rather than ancestral. PNAS Nexus 1:1–11

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14:1188–1190

Dammeyer T, Frankenberg-Dinkel N (2008) Function and distribution of bilin biosynthesis enzymes in photosynthetic organisms. Photochem Photobiol Sci 7:1121–1130

de Salas MF, Laza-Martínez A, Hallegraeff GM (2008) Novel unarmored dinoflagellates from the toxigenic family Kareniaceae (Gymnodiniales): five new species of Karlodinium and one new Takayama from the Australian sector of the Southern Ocean. J Phycol 44:241–257

de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, Lara E, Berney C, Le Bescot N, Probert I et al (2015) Eukaryotic plankton diversity in the sunlit ocean. Science 348:1261605–1/11

Dorrell RG, Azuma T, Nomura M, de Kerdrel GA, Paoli L, Yang S, Bowler C, Ishii K, Miyashita H, Gile GH et al (2019) Principles of plastid reductive evolution illuminated by nonphotosynthetic chrysophytes. Proc Natl Acad Sci USA 116:6914–6923

Dorrell RG, Gile G, McCallum G, Méheust R, Bapteste EP, Klinger CM, Brillet-Guéguen L, Freeman KD, Richter DJ, Bowler C (2017) Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. Elife 6:e23717

Dorrell RG, Hinksman GA, Howe CJ (2016) Diversity of transcripts and transcript processing forms in plastids of the dinoflagellate alga Karenia mikimotoi. Plant Mol Biol 90:233–247

Dorrell RG, Howe CJ (2012) Functional remodeling of RNA processing in replacement chloroplasts by pathways retained from their predecessors. Proc Natl Acad Sci USA 109:18879–18884

Dorrell RG, Howe CJ (2015) Integration of plastids with their hosts: Lessons learned from dinoflagellates. Proc Natl Acad Sci USA 112:10247–10254

Dorrell RG, Smith AG (2011) Do red and green make brown?: perspectives on plastid acquisitions within chromalveolates. Eukaryot Cell 10:856–868

Dorrell RG, Villain A, Perez-Lamarque B, Kerdrel GA, de, McCallum G, Watson AK, Ait-Mohamed O, Alberti A, Corre E, Frischkorn KR et al (2021) Phylogenomic fingerprinting of tempo and functions of horizontal gene transfer within ochrophytes. Proc Natl Acad Sci USA 118:e2009974118

Ebenezer TE, Zoltner M, Burrell A, Nenarokova A, Novák Vanclová AMG, Prasad B, Soukal P, Santana-Molina C, O'Neill E, Nankissoor NN et al (2019) Transcriptome, proteome and draft genome of Euglena gracilis. BMC Biol 17:11

Emanuelsson O, Nielsen H, von Heijne G (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci 8:978–984

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J et al (2014) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230

Füssy Z, Faitová T, Oborník M (2019) Subcellular compartments interplay for carbon and nitrogen allocation in *Chromera velia* and *Vitrella brassicaformis*. Genome Biol Evol 11:1765–1779

Gabrielsen TM, Minge MA, Espelund M, Tooming-Klunderud A, Patil V, Nederbragt AJ, Otis C, Turmel M, Shalchian-Tabrizi K, Lemieux C et al (2011) Genome evolution of a tertiary dinoflagellate plastid. PLoS ONE 6:e19132

Gachon CMM, Heesch S, Kuepper FC, Achilles-Day UEM, Brennan D, Campbell CN, Clarke A, Dorrell RG, Field J, Gontarek S, Menendez CR, Saxon, A RJ et al (2013) The CCAP KnowledgeBase: linking protistan and cyanobacterial biological resources with taxonomic and molecular data. Syst Biodivers 11:407–413

Galluzzi L, Bertozzini E, Penna A, Perini F, Garcés E, Magnani M (2010) Analysis of rRNA gene content in the Mediterranean dinoflagellate *Alexandrium catenella* and *Alexandrium taylori*: implications for the quantitative real-time PCR-based monitoring methods. J Appl Phycol 22:1–9

Garcés E, Fernandez M, Penna A, Van Lenning K, Gutierrez A, Camp J, Zapata M (2006) Characterization of nw Mediterranean Karlodinium spp. (Dinophyceae) strains using morphological, molecular, chemical, and physiological methodologies. J Phycol 42:1096–1112

Gast RJ, Moran DM, Dennett MR, Caron DA (2007) Kleptoplasty in an Antarctic dinoflagellate: caught in evolutionary transition? Environ Microbiol 9:39–45

Gavelis GS, Hayakawa S, White RA, Gojobori T, Suttle CA, Keeling PJ, Leander BS (2015) Eye-like ocelloids are built from different endosymbiotically acquired components. Nature 523:204–207

Gornik SG, Febrimarsa, Cassin AM, MacRae JI, Ramaprasad A, Rchiad Z, McConville MJ, Bacic A, McFadden GI, Pain A et al (2015) Endosymbiosis undone by stepwise elimination of the plastid in a parasitic dinoflagellate. Proc Natl Acad Sci USA 112:5767–5772

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652

Gruber A, Rocap G, Kroth PG, Armbrust EV, Mock T (2015) Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. Plant J 81:519–528

Gschloessl B, Guermeur Y, Cock JM (2008) HECTAR: a method to predict subcellular targeting in heterokonts. BMC Bioinform 9:1–13

Guidi, L, Morin, P, Coppola, L, Tremblay, J-É, Pesant, S, Tara Oceans Consortium, C., Tara Oceans Expedition, P. (2017a). Environmental context of all samples from the Tara Oceans Expedition (2009-2013), about nutrients in the targeted environmental feature. In: Tara Oceans Consortium, Coordinators; Tara Oceans Expedition, Participants (2017): Registry of all samples from the Tara Oceans Expedition (2009–2013). PANGAEA

Guidi, L, Picheral, M, Pesant, S, Tara Oceans Consortium, C., Tara Oceans Expedition, P. (2017b). Environmental context of all samples from the Tara Oceans Expedition (2009-2013), about sensor data in the targeted environmental feature. In: Tara Oceans Consortium, Coordinators; Tara Oceans Expedition, Participants (2017): Registry of all samples from the Tara Oceans Expedition (2009–2013). PANGAEA

Gütle DD, Roret T, Müller SJ, Couturier J, Lemaire SD, Hecker A, Dhalleine T, Buchanan BB, Reski R, Einsle O, Jacquot JP (2016) Chloroplast FBPase and SBPase are thioredoxin-linked enzymes with similar architecture but different evolutionary histories. Proc Natl Acad Sci USA 113:6779–6784

Hackett JD, Anderson DM, Erdner DL, Bhattacharya D (2004) Dinoflagellates: a remarkable evolutionary experiment. Am J Bot 91:1523–1534

Hadariová L, Vesteg M, Hampl V, Krajčovič J (2018) Reductive evolution of chloroplasts in non-photosynthetic plants, algae and protists. Curr Genet 64:365–387

Hannaert V, Saavedra E, Duffieux F, Szikora J-P, Rigden DJ, Michels PAM, Opperdoes FR (2003) Plant-like traits associated with metabolism of Trypanosoma parasites. Proc Natl Acad Sci USA 100:1067–1071

Haxo FT, Kycia JH, Somers GF, Benneit A, Siegelman HW (1976) Peridinin-chlorophyll a proteins of the dinoflagellate *Amphidinium carterae* (Plymouth 450). Plant Physiol 57:297–303

Hehenberger E, Gast RJ, Keeling PJ (2019) A kleptoplastidic dinoflagellate and the tipping point between transient and fully integrated plastid endosymbiosis. Proc Natl Acad Sci USA 116:17934–17942

Hiller K, Grote A, Scheer M, Münch R, Jahn D (2004) PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Res 32:375–379

Howe CJ, Barbrook AC, Nisbet RER, Lockhart PJ, Larkum AWD (2008) The origin of plastids. Philos Trans R Soc Lond B Biol Sci 363:2675–2685

Huang J, Gogarten J (2007) Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biol 8:R99

Ibarbalz FM, Henry N, Brandão MC, Martini S, Busseni G, Byrne H, Coelho LP, Endo H, Gasol JM, Gregory AC et al (2019) Global trends in marine plankton diversity across kingdoms of life. Cell 179:1084–1097

Imanian B, Keeling PJ (2007) The dinoflagellates *Durinskia baltica* and *Kryptoperidinium foliaceum* retain functionally overlapping mitochondria from two evolutionarily distinct lineages. BMC Evol Biol 7:1–11

Ishida K-I, Green BR (2002) Second-and third-hand chloroplasts in dinoflagellates: phylogeny of oxygen-evolving enhancer 1 (PsbO) protein reveals replacement of a nuclear-encoded plastid gene by that of a haptophyte tertiary endosymbiont. Proc Natl Acad Sci USA 99:9294–9299

Jackson CJ, Gornik SG, Waller RF (2013) A tertiary plastid gains RNA editing in its new host. Mol Biol Evol 30:788–792

Janouskovec J, Gavelis GS, Burki F, Dinh D, Bachvaroff TR, Gornik SG, Bright KJ, Imanian B, Strom SL, Delwiche CF et al (2017) Major transitions in dinoflagellate evolution unveiled by phylotranscriptomics. Proc Natl Acad Sci USA 114:E171–E180

Janouškovec J, Horák A, Oborník M, Lukeš J, Keeling PJ (2010) A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. Proc Natl Acad Sci USA 10710949-10954

Jeong HJ, Kang HC, Lim AS, Jang SH, Lee K, Lee SY, Ok JH, You JH, Kim JH, Lee KH et al (2021) Feeding diverse prey as an excellent strategy of mixotrophic dinoflagellates for global dominance. Sci Adv 7:eabe4214

Jeong HJ, Ok JH, Lim AS, Kwon JE, Kim SJ, Lee SY (2016) Mixotrophy in the phototrophic dinoflagellate Takayama helix (family Kareniaceae): Predator of diverse toxic and harmful dinoflagellates. Harmful Algae 60:92–106

Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780

Kawachi M, Nakayama T, Kayama M, Nomura M, Miyashita H, Bojo O, Rhodes L, Sym S, Pienaar RN, Probert I et al (2021) Rappemonads are haptophyte phytoplankton. Curr Biol 31:2395–2403

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C et al (2012) Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ et al (2014) The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol 12:e1001889

Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC (2011) Adaptive seeds tame genomic sequence comparison. Genome Res 21:487–493

Kleffmann T, Russenberger D, von Zychlinski A, Christopher W, Sjölander K, Gruissem W, Baginsky S (2004) The *Arabidopsis thaliana* chloroplast proteome reveals pathway abundance and novel protein functions. Curr Biol 14:354–362

Klinger CM, Klute MJ, Dacks JB (2013) Comparative genomic analysis of multi-subunit tethering complexes demonstrates an ancient pan-eukaryotic complement and sculpting in apicomplexa. PLoS ONE 8:e76278

Klinger CM, Paoli L, Newby RJ, Wang MYW, Carroll HD, Leblond JD, Howe CJ, Dacks JB, Bowler C, Cahoon AB et al (2018) Plastid transcript editing across dinoflagellate lineages shows lineage-specific application but conserved trends. Genome Biol Evol 10:1019–1038

Koike K, Sekiguchi H, Kobiyama A, Takishita K, Kawachi M, Koike K, Ogata T (2005) A novel type of kleptoplastidy in Dinophysis (Dinophyceae): presence of haptophyte-type plastid in *Dinophysis mitra*. Protist 156:225–237

Kroth PG, Chiovitti A, Gruber A, Martin-Jezequel V, Mock T, Parker MS, Stanley MS, Kaplan A, Caron L, Weber T et al (2008) A model for carbohydrate metabolism in the diatom *Phaeodactylum tricornutum* deduced from comparative whole genome analysis. PLoS ONE 3:e1426

Ku C, Nelson-Sathi S, Roettger M, Sousa FL, Lockhart PJ, Bryant D, Hazkani-Covo E, McInerney JO, Landan G, Martin WF (2015) Endosymbiotic origin and differential loss of eukaryotic genes. Nature 524:427–432

Larkum AWD, Lockhart PJ, Howe CJ (2007) Shopping for plastids. Trends Plant Sci 12:189–195

Leblond JD, Sabir K, Graeff JE (2022) Galactolipid composition of Gertia stigmatica (Dinophyceae), an atypical member of the Kareniaceae with a peridinin-containing, non-haptophyte-derived plastid. Phycol Res 70:137–141

Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, Grosse I, Li Z, Melkonian M, Mirarab S (2019) One thousand plant transcriptomes and the phylogenomics of green plants. Nature 574:679–685

Lidie KB, Ryan JC, Barbier M, van Dolah FM (2005) Gene expression in Florida red tide dinoflagellate *Karenia brevis*: analysis of an expressed sequence tag library and development of DNA microarray. Mar Biotechnol 7:481–493

Liew YJ, Li Y, Baumgarten S, Voolstra CR, Aranda M (2017) Condition-specific RNA editing in the coral symbiont Symbiodinium microadriaticum. PLoS Genet 13:e1006619

Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F, Chaffron S, Ignacio-Espinosa JC, Roux S, Vincent F et al (2015) Determinants of community structure in the global plankton interactome. Science 348:1262073

Lin S (2011) Genomic understanding of dinoflagellates. Res Microbiol 162:551–569

Lin S, Zhang H, Spencer DF, Norman JE, Gray MW (2002) Widespread and extensive editing of mitochondrial mRNAS in dinoflagellates. J Mol Biol 320:727–739

Matsumoto T, Shinozaki F, Chikuni T, Yabuki A, Takishita K, Kawachi M, Nakayama T, Inouye I, Hashimoto T, Inagaki Y (2011) Green-colored plastids in the dinoflagellate genus *Lepidodinium* are of core chlorophyte origin. Protist 162:268–276

Matsuo E, Inagaki Y (2018) Patterns in evolutionary origins of heme, chlorophyll a and isopentenyl diphosphate biosynthetic pathways suggest non-photosynthetic periods prior to plastid replacements in dinoflagellates. PeerJ 6:e5345

Matsuo E, Morita K, Nakayama T, Yazaki E, Sarai C, Takahashi K, Iwataki M, Inagaki Y (2022) Comparative plastid genomics of green-colored dinoflagellates unveils parallel genome compaction and RNA editing. Front. Plant Sci 13:918543

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534

Moog D, Maier UG (2017) Cellular compartmentation follows rules: the Schnepf theorem, its consequences and exceptions. BioEssays 39:1700030

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35:W182–W185

Morse D, Salois P, Markovic P, Hastings JW (1995) A nuclear-encoded form II RuBisCO in dinoflagellates. Science 268:1622–1624

Moustafa A, Reyes-Prieto A, Bhattacharya D (2008) Chlamydiae has contributed at least 55 genes to plantae with predominantly plastid functions. PLoS ONE 3:e2205

Myung GP, Kim S, Hyung SK, Myung G, Yi GK, Yih W (2006) First successful culture of the marine dinoflagellate *Dinophysis acuminata*. Aquat Microb Ecol 45:101–106

Nash EA, Barbrook AC, Edwards-Stuart RK, Bernhardt K, Howe CJ, Nisbet RER (2007) Organization of the mitochondrial genome in the dinoflagellate *Amphidinium carterae*. Mol Biol Evol 24:1528–1536

Nelson DR, Hazzouri KM, Lauersen KJ, Jaiswal A, Chaiboonchoe A, Mystikou A, Fu W, Daakour S, Dohai B, Alzahmi A et al (2021) Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution. Cell Host Microbe 29:250–266

Nosenko T, Lidie KL, van Dolah FM, Lindquist E, Cheng JF, Bhattacharya D (2006) Chimeric plastid proteome in the Florida "red tide" dinoflagellate *Karenia brevis*. Mol Biol Evol 23:2026–2038

Novák Vanclová AMG, Zoltner M, Kelly S, Soukal P, Záhonová K, Füssy Z, Ebenezer TE, Lacová Dobáková E, Eliáš M, Lukeš J et al (2020) Metabolic quirks and the colourful history of the *Euglena gracilis* secondary plastid. New Phytol 225:1578–1592

Ogata T, Kodama M, Nomura S, Kobayashi M, Nozawa T, Katoh T, Mimuro M (1994) A novel peridinin—chlorophyll a protein (PCP) from the marine dinoflagellate *Alexandrium cohorticula*: a high pigment content and plural spectral forms of peridinin and chlorophyll a. FEBS Lett 356:367–371

Ok JH, Jeong HJ, Lee SY, Park SA, Noh JH (2021) Shimiella gen. nov. and Shimiella gracilenta sp. nov. (Dinophyceae, Kareniaceae), a kleptoplastidic dinoflagellate from Korean waters and its survival under starvation. J Phycol 57:70–91

Patron NJ, Waller RF (2007) Transit peptide diversity and divergence: a global analysis of plastid targeting signals. Bioessays 29:1048–1058

Patron NJ, Waller RF, Keeling PJ (2006) A tertiary plastid uses genes from two endosymbionts. J Mol Biol 357:1373–1382

Penot M, Dacks JB, Read B, Dorrell RG (2022) Genomic and meta-genomic insights into the functions, diversity and global distribution of haptophyte algae. Appl Phycol 3:340–359

Pierella Karlusich JJ, Ibarbalz FM, Bowler C (2019) Phytoplankton in the Tara Ocean. Ann Rev Mar Sci 12:233–265

Pierella Karlusch, JJ, Nef, C, Bowler, C, Dorrell, RG (2022) Biogeographical patterns and genomes of aquatic photoautotrophs. In: Maberly SC and Gontero B (eds) Blue planet, red and green photosynthesis: productivity and carbon cycling in aquatic ecosystems. Wiley-ISTE, p 43–80

Ponce-Toledo RI, López-García P, Moreira D (2019) Horizontal and endosymbiotic gene transfer in early plastid evolution. New Phytol 224:618–624

Riccio G, de Luca D, Lauritano C (2020) Monogalactosyldiacylglycerol and sulfolipid synthesis in microalgae. Mar Drugs 18:237

Richardson E, Dorrell RG, Howe CJ (2014) Genome-wide transcript profiling reveals the coevolution of plastid gene sequences and transcript processing pathways in the fucoxanthin dinoflagellate *Karlodinium veneficum*. Mol Biol Evol 31:2376–2386

Ryan DE, Pepper AE, Campbell L (2014) De novo assembly and characterization of the transcriptome of the toxic dinoflagellate Karenia brevis. BMC Genomics 15:1–11

Ryan, DE, Pepper, AE, Campbell, L (2014). NCBI BioProject PRJNA214017 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA214017) [DATASET]

Saldarriaga JF, Taylor FJR, Keeling PJ, Cavalier-Smith T (2001) Dinoflagellate nuclear SSU rRNA phylogeny suggests multiple plastid losses and replacements. J Mol Evol 53:204–213

Sanchez-Puerta MV, Lippmeier JC, Apt KE, Delwiche CF (2007) Plastid genes in a non-photosynthetic dinoflagellate. Protist 158:105–117

Sarai C, Tanifuji G, Nakayama T, Kamikawa R, Takahashi K, Yazaki E, Matsuo E, Miyashita H, Ishida K-I, Iwataki M et al (2020) Dinoflagellates with relic endosymbiont nuclei as models for elucidating organellogenesis. Proc Natl Acad Sci USA 117:5364–5375

Ševčíková T, Horák A, Klimeš V, Zbránková V, Demir-Hilton E, Sudek S, Jenkins J, Schmutz J, Přibyl P, Fousek J et al (2015) Updating algal evolutionary relationships through plastid genome sequencing: did alveolate plastids emerge through endosymbiosis of an ochrophyte? Sci Rep 5:10134

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212

Stern RF, Horak A, Andrew RL, Coffroth M-A, Andersen RA, Küpper FC, Jameson I, Hoppenrath M, Véron B, Kasai F et al (2010) Environmental barcoding reveals massive dinoflagellate diversity in marine environments. PLoS ONE 5:e13991

Strassert JFH, Irisarri I, Williams TA, Burki F (2021) A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. Nat Commun 12:1879

Suzuki K, Miyagishima SY (2010) Eukaryotic and eubacterial contributions to the establishment of plastid proteome estimated by large-scale phylogenetic analyses. Mol Biol Evol 27:581–590

Takahashi K, Benico G, Lum WM, Iwataki M (2019) Gertia stigmatica gen. et sp. nov. (Kareniaceae, Dinophyceae), a new marine unarmored dinoflagellate possessing the peridinin-type chloroplast with an eyespot. Protist 170:125680

Takishita K, Nakano K, Uchida A (1999) Preliminary phylogenetic analysis of plastid-encoded genes from an anomalously pigmented dinoflagellate *Gymnodinium mikimotoi* (Gymnodiniales, Dinophyta). Phycol Res 47:257–262

Tanaka A, Tanaka R (2006) Chlorophyll metabolism. Curr Opin Plant Biol 9:248–255

Tengs T, Dahlberg OJ, Shalchian-Tabrizi K, Klaveness D, Rudi K, Delwiche CF, Jakobsen KS (2000) Phylogenetic analyses indicate flint the 19′hexanoyloxy-fucoxanthin- containing dinoflagellates have tertiary plastids of haptophyte origin. Mol Biol Evol 17:718–729

Thomsen MCF, Nielsen M (2012) Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. Nucleic Acids Res 40:W281–W287

Tice AK, Žihala D, Pánek T, Jones RE, Salomaki ED, Nenarokov S, Burki F, Eliáš M, Eme L, Roger AJ et al (2021) PhyloFisher: a phylogenomic package for resolving eukaryotic relationships. PLoS Biol 19:e3001365

Uthanumallian K, Iha C, Repetti SI, Chan CX, Bhattacharya D, Duchene S, Verbruggen H (2021) Tightly constrained genome reduction and relaxation of purifying selection during secondary plastid endosymbiosis. Mol Biol Evol 39:msab295

Vincent F, Bowler C (2020) Diatoms are selective segregators in global ocean planktonic communities. mSystems 5:e00444-19

Waller RF, Slamovits CH, Keeling PJ (2006) Lateral gene transfer of a multigene region from cyanobacteria to dinoflagellates resulting in a novel plastid-targeted fusion protein. Mol Biol Evol 23:1437–1443

Wang D-Z (2008) Neurotoxins from marine dinoflagellates: a brief review. Mar Drugs 6:349–371

Wang HC, Minh BQ, Susko E, Roger AJ (2018) Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. Syst Biol 67:216–235

Wang Y, Morse D (2006) Rampant polyuridylylation of plastid gene transcripts in the dinoflagellate Lingulodinium. Nucleic Acids Res 34:613–619

Wei T, Simko V (2021). R package “corrplot”: visualization of a correlation matrix. Available at: https://github.com/taiyun/corrplot

Yokoyama A, Takahashi F, Kataoka H, Hara Y, Nozaki H (2011) Evolutionary analyses of the nuclear-encoded photosynthetic gene psbo from tertiary plastid-containing algae in dinophyta. J Phycol 47:407–414

Yoon HS, Hackett JD, Bhattacharya D (2002) A single origin of the peridinin- and fucoxanthin-containing plastids in dinoflagellates through tertiary endosymbiosis. Proc Natl Acad Sci USA 99:11724–11729

Záhonová K, Füssy Z, Birčák E, Novák Vanclová AMG, Klimeš V, Vesteg M, Krajčovič J, Oborník M, Eliáš M (2018) Peculiar features of the plastids of the colourless alga Euglena longa and photosynthetic euglenophytes unveiled by transcriptome analyses. Sci Rep 8:17012

Zapata M, Fraga S, Rodríguez F, Garrido JL (2012) Pigment-based chloroplast types in dinoflagellates. Mar Ecol Prog Ser 465:33–52

Zapata, M, Jeffrey, SW, Wright, SW, Rodríguez, F, Garrido, JL, Clementson, L (2004). Photosynthetic pigments in 37 species (65 strains) of *Haptophyta*: implications for oceanography and chemotaxonomy. Mar Ecol Prog Ser 270:83–102.

Zauner S, Greilinger D, Laatsch T, Kowallik KV, Maier U-G (2004) Substitutional editing of transcripts from genes of cyanobacterial origin in the dinoflagellate *Ceratium horridum*. FEBS Lett 577:535–538

Zhang Z, Green BR, Cavalier-Smith T (1999) Single gene circles in dinoflagellate chloroplast genomes. Nature 400:155–159

Zhou Q, Su X, Jing G, Chen S, Ning K (2018) RNA-QC-chain: comprehensive and fast quality control for RNA-Seq data. BMC Genomics 19:144

Acknowledgements

RGD and ANV acknowledge support from a CNRS Momentum Fellowship awarded to RGD, 2019-2021, and an ANR JCJC ("PanArctica", ANR-21-CE02-0014) awarded to RGD, 2021-2022. RGD further acknowledges an ERC Starting Grant ("ChloroMosaic", project number 101039760), awarded 2023-2027. ANV acknowledges support from ANR ("ArchaeArf", ANR-20-CE13-0007), awarded 2020-2024. CB acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Diatomic; grant agreement No. 835067), the French Government "Investissements d'Avenir" programs MEMO LIFE (ANR-10-LABX-54) and PSL* Research University (ANR-1253 11-IDEX-0001-02), and from the Agence Nationale de la Recherche (BrownCut; grant agreement ANR-19-CE20-0020). Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. The authors thank Elisabeth Hehenberger for sharing translated and decontaminated RSD data and consultation regarding their analysis, Lukáš Novák for consultation regarding phylogenomics, and Emmanuel Alastra for aid in media substrate preparation. This article is a contribution #151 to *Tara* Oceans.

Author contributions

Anna MG Novák Vanclová: Conceptualization; Data curation; Software; Formal analysis; Investigation; Visualization; Methodology; Writing—original draft; Writing—review and editing. **Charlotte Nef**: Data curation; Formal analysis; Investigation; Methodology; Writing—original draft. **Zoltan Fussy**: Data curation; Formal analysis; Investigation; Methodology. **Adél Vancl**: Software. **Fuhai Liu**: Data curation; Software; Methodology. **Chris Bowler**: Supervision; Funding acquisition; Validation; Writing—review and editing. **Richard G Dorrell**: Conceptualization; Data curation; Formal analysis; Supervision; Funding acquisition; Validation; Investigation; Methodology; Writing—original draft; Project administration; Writing—review and editing.

Disclosure and competing interests statement

The authors declare no competing interests.

Expanded View Figures

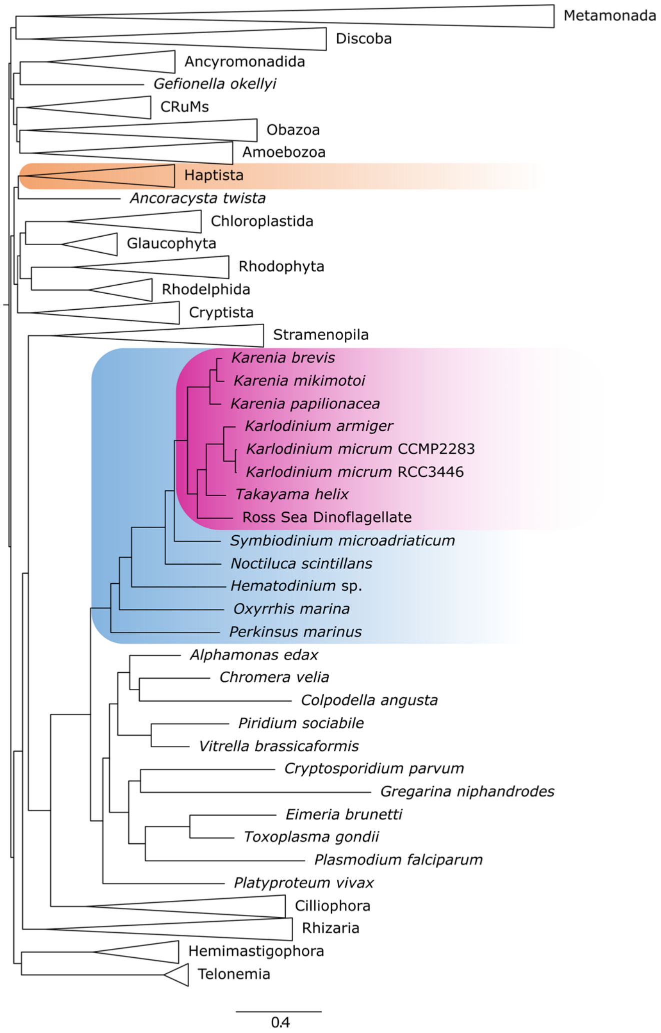
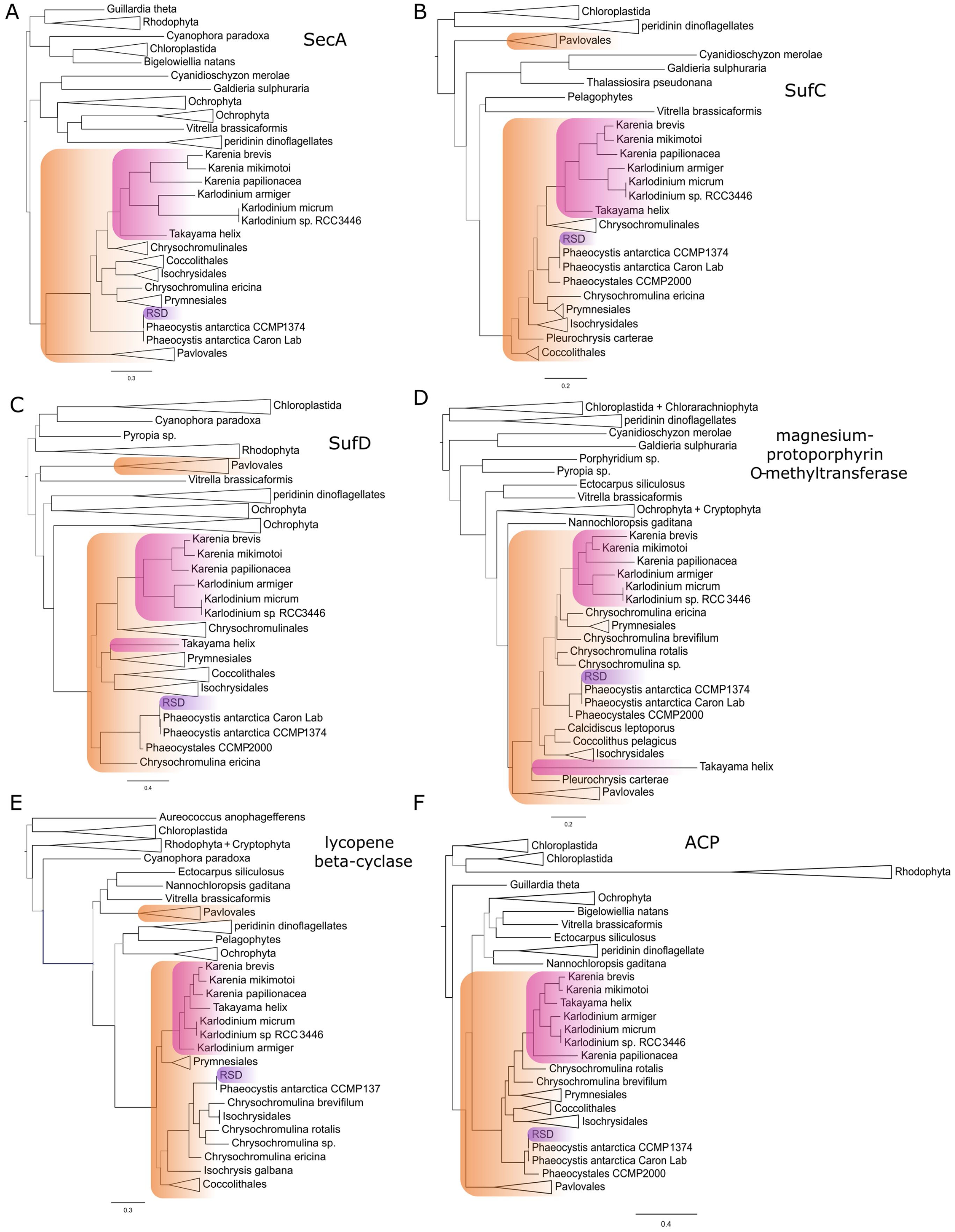


Figure EV1. The eight kareniaceans in the pan-eukaryotic phylogenetic context as reconstructed by IQ-TREE based on a matrix of 241 genes prepared using PhyloFisher toolkit.

The inner relationships between the studied kareniaceans are resolved with maximum support and confirm the results of previous phylogenetic analyses based on ribosomal subunits (Takahashi et al, 2019; Ok et al, 2021).


A
SecA
Guillardia theta
Rhodophyta
Cyanophora paradoxa
Chloroplastida
Bigelowiellia natans
Cyanidioschyzon merolae
Galdieria sulphuraria
Ochrophyta
Ochrophyta
Vitrella brassicaformis
peridinin dinoflagellates
Karenia brevis
Karenia mikimotoi
Karenia papilionacea
Karlodinium armiger
Karlodinium micrum
Karlodinium sp. RCC3446
Takayama helix
Chrysochromulinales
Coccolithales
Isochrysidales
Chrysochromulina ericina
Prymnesiales
RSD
Phaeocystis antarctica CCMP1374
Phaeocystis antarctica Caron Lab
Pavlovales
0.3
B
SufC
Chloroplastida
peridinin dinoflagellates
Pavlovales
Cyanidioschyzon merolae
Galdieria sulphuraria
Thalassiosira pseudonana
Pelagophytes
Vitrella brassicaformis
Karenia brevis
Karenia mikimotoi
Karenia papilionacea
Karlodinium armiger
Karlodinium micrum
Karlodinium sp. RCC3446
Takayama helix
Chrysochromulinales
RSD
Phaeocystis antarctica CCMP1374
Phaeocystis antarctica Caron Lab
Phaeocystales CCMP2000
Chrysochromulina ericina
Prymnesiales
Isochrysidales
Pleurochrysis carterae
Coccolithales
0.2
C
SufD
Chloroplastida
Cyanophora paradoxa
Pyropia sp.
Rhodophyta
Pavlovales
Vitrella brassicaformis
peridinin dinoflagellates
Ochrophyta
Ochrophyta
Karenia brevis
Karenia mikimotoi
Karenia papilionacea
Karlodinium armiger
Karlodinium micrum
Karlodinium sp RCC3446
Chrysochromulinales
Takayama helix
Prymnesiales
Coccolithales
Isochrysidales
RSD
Phaeocystis antarctica Caron Lab
Phaeocystis antarctica CCMP1374
Phaeocystales CCMP2000
Chrysochromulina ericina
0.4
D
magnesium-protoporphyrin O-methyltransferase
Chloroplastida + Chlorarachniophyta
peridinin dinoflagellates
Cyanidioschyzon merolae
Galdieria sulphuraria
Porphyridium sp.
Pyropia sp.
Ectocarpus siliculosus
Vitrella brassicaformis
Ochrophyta + Cryptophyta
Nannochloropsis gaditana
Karenia brevis
Karenia mikimotoi
Karenia papilionacea
Karlodinium armiger
Karlodinium micrum
Karlodinium sp. RCC3446
Chrysochromulina ericina
Prymnesiales
Chrysochromulina brevifilum
Chrysochromulina rotalis
Chrysochromulina sp.
RSD
Phaeocystis antarctica CCMP1374
Phaeocystis antarctica Caron Lab
Phaeocystales CCMP2000
Calcidiscus leptoporus
Coccolithus pelagicus
Isochrysidales
Takayama helix
Pleurochrysis carterae
Pavlovales
0.2
E
lycopene beta-cyclase
Aureococcus anophagefferens
Chloroplastida
Rhodophyta + Cryptophyta
Cyanophora paradoxa
Ectocarpus siliculosus
Nannochloropsis gaditana
Vitrella brassicaformis
Pavlovales
peridinin dinoflagellates
Pelagophytes
Ochrophyta
Karenia brevis
Karenia mikimotoi
Karenia papilionacea
Takayama helix
Karlodinium micrum
Karlodinium sp RCC3446
Karlodinium armiger
Prymnesiales
RSD
Phaeocystis antarctica CCMP137
Chrysochromulina brevifilum
Isochrysidales
Chrysochromulina rotalis
Chrysochromulina sp.
Chrysochromulina ericina
Isochrysis galbana
Coccolithales
0.3
F
ACP
Chloroplastida
Chloroplastida
Rhodophyta
Guillardia theta
Ochrophyta
Bigelowiellia natans
Vitrella brassicaformis
Ectocarpus siliculosus
peridinin dinoflagellate
Nannochloropsis gaditana
Karenia brevis
Karenia mikimotoi
Takayama helix
Karlodinium armiger
Karlodinium micrum
Karlodinium sp. RCC3446
Karenia papilionacea
Chrysochromulina rotalis
Chrysochromulina brevifilum
Prymnesiales
Coccolithales
Isochrysidales
RSD
Phaeocystis antarctica CCMP1374
Phaeocystis antarctica Caron Lab
Phaeocystales CCMP2000
Pavlovales
0.4

◀ **Figure EV2. Example single-gene trees for proteins of plastid-late origin, with specifically plastidial function and present in all studied organisms.**

The non-RSD kareniacean sequences typically resolve adjacent to Chrysochromulinales or a mixed Chrysochromulinales/Prymnesiales clade. Their inner topology varies and typically differs from the organismal one (Figure EV1), with *Takayama* often branching as sister to the rest (**A**, **B**) and sometimes even separately (**C**, **D**), albeit with low support, or inside the *Karenia/Karlodinium* clade but sister to *Karenia* rather than *Karlodinium* (**E**, **F**). All species of *Karenia* and *Karlodinium* are, however, consistently retrieved as monophyletic, suggesting a common chrysochromulinalean-like origin of their nucleus-encoded plastid protein inventory, which is for the most part, also shared with *Takayama*.

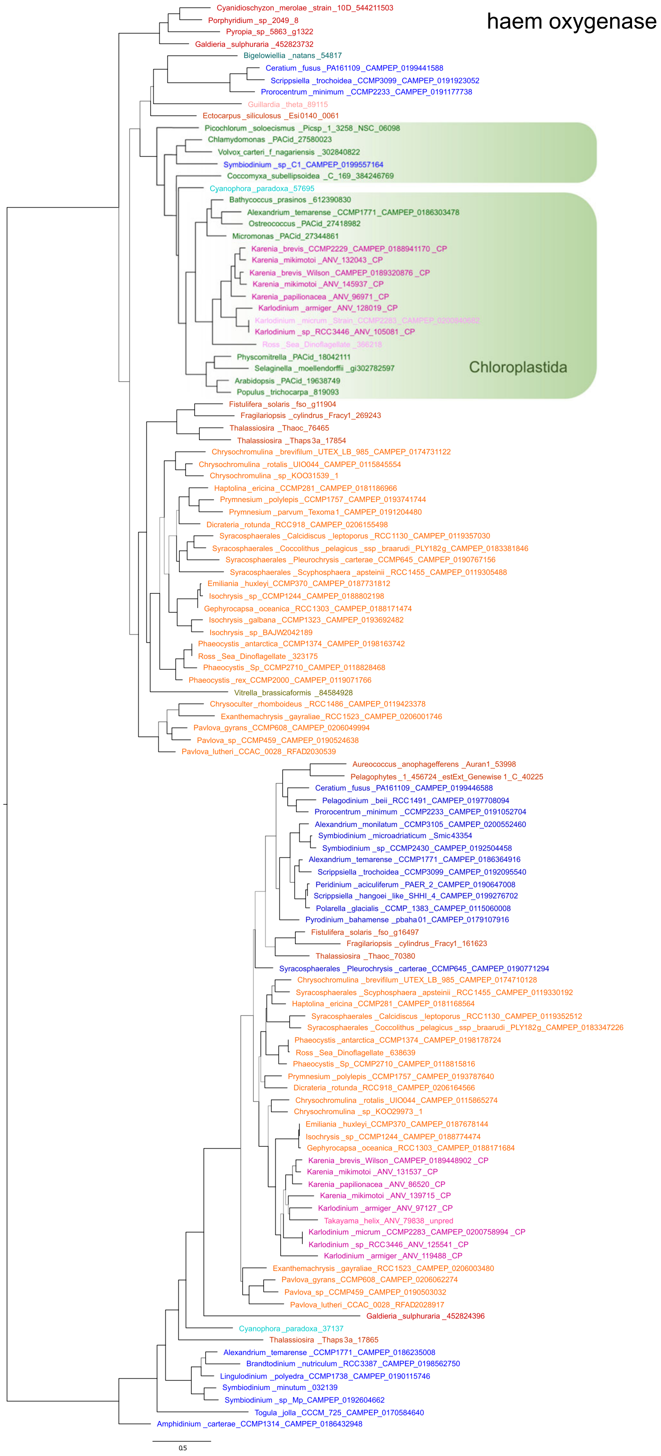

haem oxygenase
Cyanidioschyzon _merolae _strain _10D_544211503
Porphyridium _sp _2049 _8
Pyropia_sp _5863 _g1322
Galdieria _sulphuraria _452823732
Bigelowiellia _natans _54817
Ceratium _fusus _PA161109 _CAMPEP_0199441588
Scrippsiella _trochoidea _CCMP3099 _CAMPEP_0191923052
Prorocentrum _minimum _CCMP2233 _CAMPEP_0191177738
Guillardia _theta _89115
Ectocarpus _siliculosus _Esi0140 _0061
Picochlorum _soloecismus _Picsp _1_3258 _NSC_06098
Chlamydomonas _PACid_27580023
Volvox_carteri _f_nagariensis _302840822
Symbiodinium _sp _C1_CAMPEP_0199557164
Coccomyxa_subellipsoidea _C_169 _384246769
Cyanophora _paradoxa _57695
Bathycoccus _prasinos _612390830
Alexandrium _temarense _CCMP1771 _CAMPEP_0186303478
Ostreococcus _PACid_27418982
Micromonas _PACid_27344861
Karenia _brevis _CCMP2229 _CAMPEP_0188941170 _CP
Karenia _mikimotoi _ANV_132043 _CP
Karenia _brevis _Wilson _CAMPEP_0189320876 _CP
Karenia _mikimotoi _ANV_145937 _CP
Karenia _papilionacea _ANV_96971 _CP
Karlodinium _armiger _ANV_128019 _CP
Karlodinium _micrum _Strain _CCMP2283 _CAMPEP_0200840682
Karlodinium _sp _RCC3446 _ANV_105081 _CP
Ross _Sea _Dinoflagellate _366218
Physcomitrella _PACid_18042111
Selaginella _moellendorffii _gi302782597
Arabidopsis _PACid_19638749
Populus _trichocarpa _819093
Chloroplastida
Fistulifera _solaris _fso _g11904
Fragilariopsis _cylindrus _Fracy1 _269243
Thalassiosira _Thaoc _76465
Thalassiosira _Thaps 3a _17854
Chrysochromulina _brevifilum _UTEX_LB_985 _CAMPEP_0174731122
Chrysochromulina _rotalis _UIO044 _CAMPEP_0115845554
Chrysochromulina _sp _KOO31539 _1
Haptolina _ericina _CCMP281 _CAMPEP_0181186966
Prymnesium _polylepis _CCMP1757 _CAMPEP_0193741744
Prymnesium _parvum _Texoma 1 _CAMPEP_0191204480
Dicrateria _rotunda _RCC918 _CAMPEP_0206155498
Syracosphaerales _Calcidiscus _leptoporus _RCC1130 _CAMPEP_0119357030
Syracosphaerales _Coccolithus _pelagicus _ssp _braarudi _PLY182g _CAMPEP_0183381846
Syracosphaerales _Pleurochrysis _carterae _CCMP645 _CAMPEP_0190767156
Syracosphaerales _Scyphosphaera _apsteinii _RCC1455 _CAMPEP_0119305488
Emiliania _huxleyi _CCMP370 _CAMPEP_0187731812
Isochrysis _sp _CCMP1244 _CAMPEP_0188802198
Gephyrocapsa _oceanica _RCC1303 _CAMPEP_0188171474
Isochrysis _galbana _CCMP1323 _CAMPEP_0193692482
Isochrysis _sp _BAJW2042189
Phaeocystis _antarctica _CCMP1374 _CAMPEP_0198163742
Ross _Sea _Dinoflagellate _323175
Phaeocystis _Sp _CCMP2710 _CAMPEP_0118828468
Phaeocystis _rex _CCMP2000 _CAMPEP_0119071766
Vitrella _brassicaformis _84584928
Chrysoculter _rhomboideus _RCC1486 _CAMPEP_0119423378
Exanthemachrysis _gayraliae _RCC1523 _CAMPEP_0206001746
Pavlova _gyrans _CCMP608 _CAMPEP_0206049994
Pavlova _sp _CCMP459 _CAMPEP_0190524638
Pavlova _lutheri _CCAC_0028 _RFAD2030539
Aureococcus _anophagefferens _Auran1 _53998
Pelagophytes _1_456724 _estExt _Genewise 1_C_40225
Ceratium _fusus _PA161109 _CAMPEP_0199446588
Pelagodinium _beii _RCC1491 _CAMPEP_0197708094
Prorocentrum _minimum _CCMP2233 _CAMPEP_0191052704
Alexandrium _monilatum _CCMP3105 _CAMPEP_0200552460
Symbiodinium _microadriaticum _Smic43354
Symbiodinium _sp _CCMP2430 _CAMPEP_0192504458
Alexandrium _temarense _CCMP1771 _CAMPEP_0186364916
Scrippsiella _trochoidea _CCMP3099 _CAMPEP_0192095540
Peridinium _aciculiferum _PAER_2 _CAMPEP_0190647008
Scrippsiella _hangoei _like _SHHI_4 _CAMPEP_0199276702
Polarella _glacialis _CCMP_1383 _CAMPEP_0115060008
Pyrodinium _bahamense _pbaha01 _CAMPEP_0179107916
Fistulifera _solaris _fso _g16497
Fragilariopsis _cylindrus _Fracy1 _161623
Thalassiosira _Thaoc _70380
Syracosphaerales _Pleurochrysis _carterae _CCMP645 _CAMPEP_0190771294
Chrysochromulina _brevifilum _UTEX_LB_985 _CAMPEP_0174710128
Syracosphaerales _Scyphosphaera _apsteinii _RCC1455 _CAMPEP_0119330192
Haptolina _ericina _CCMP281 _CAMPEP_0181168564
Syracosphaerales _Calcidiscus _leptoporus _RCC1130 _CAMPEP_0119352512
Syracosphaerales _Coccolithus _pelagicus _ssp _braarudi _PLY182g _CAMPEP_0183347226
Phaeocystis _antarctica _CCMP1374 _CAMPEP_0198178724
Ross _Sea _Dinoflagellate _638639
Phaeocystis _Sp _CCMP2710 _CAMPEP_0118815816
Prymnesium _polylepis _CCMP1757 _CAMPEP_0193787640
Dicrateria _rotunda _RCC918 _CAMPEP_0206164566
Chrysochromulina _rotalis _UIO044 _CAMPEP_0115865274
Chrysochromulina _sp _KOO29973 _1
Emiliania _huxleyi _CCMP370 _CAMPEP_0187678144
Isochrysis _sp _CCMP1244 _CAMPEP_0188774474
Gephyrocapsa _oceanica _RCC1303 _CAMPEP_0188171684
Karenia _brevis _Wilson _CAMPEP_0189448902 _CP
Karenia _mikimotoi _ANV_131537 _CP
Karenia _papilionacea _ANV_86520 _CP
Karenia _mikimotoi _ANV_139715 _CP
Karlodinium _armiger _ANV_97127 _CP
Takayama _helix _ANV_79838 _unpred
Karlodinium _micrum _CCMP2283 _CAMPEP_0200758994 _CP
Karlodinium _sp _RCC3446 _ANV_125541 _CP
Karlodinium _armiger _ANV_119488 _CP
Exanthemachrysis _gayraliae _RCC1523 _CAMPEP_0206003480
Pavlova _gyrans _CCMP608 _CAMPEP_0206062274
Pavlova _sp _CCMP459 _CAMPEP_0190503032
Pavlova _lutheri _CCAC_0028 _RFAD2028917
Galdieria _sulphuraria _452824396
Cyanophora _paradoxa _37137
Thalassiosira _Thaps 3a _17865
Alexandrium _temarense _CCMP1771 _CAMPEP_0186235008
Brandtodinium _nutriculum _RCC3387 _CAMPEP_0198562750
Lingulodinium _polyedra _CCMP1738 _CAMPEP_0190115746
Symbiodinium _minutum _032139
Symbiodinium _sp _Mp _CAMPEP_0192604662
Togula _jolla _CCCM_725 _CAMPEP_0170584640
Amphidinium _carterae _CCMP1314 _CAMPEP_0186432948
0.5

Figure EV3. Automatically generated phylogenetic tree showing the green origin of one form of heme oxygenase in *Karenia*, *Karlodinium*, and the RSD (no *Takayama* homolog was retrieved); the second form is of plastid-late origin in all genera.

Bootstrap support is expressed by the branch color (black for ≥90%, dark gray for ≥75%, and light gray for <75%). Kareniacean sequences are colored pink with proteins with predicted plastid-targeting signal in darker shades and annotated "CP" suffix; dinoflagellates are coded blue; haptophytes are coded orange; green algae (Chloroplastida) are coded green.

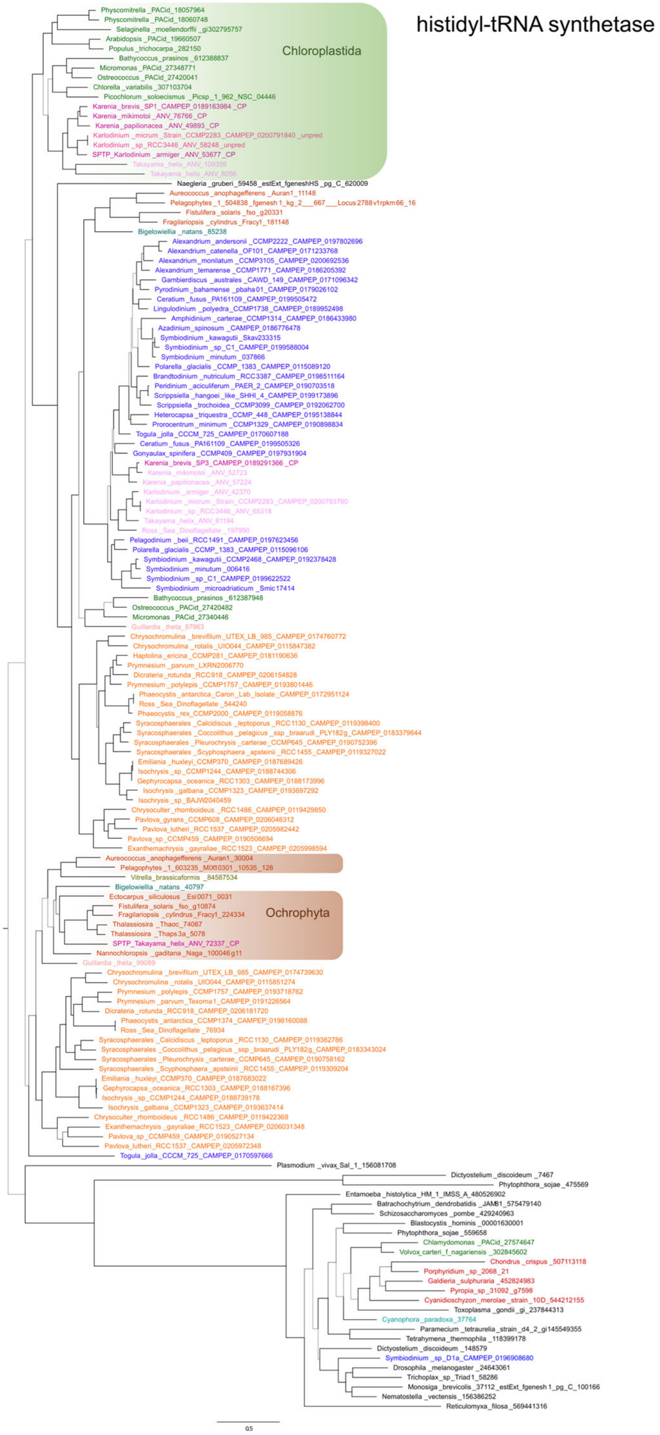

histidyl-tRNA synthetase
Chloroplastida
Physcomitrella _PACid_18057964
Physcomitrella _PACid_18060748
Selaginella _moellendorffii _gi302795757
Arabidopsis _PACid_19660507
Populus _trichocarpa _282150
Bathycoccus_prasinos _612388837
Micromonas _PACid_27348771
Ostreococcus _PACid_27420041
Chlorella _variabilis _307103704
Picochlorum _soloecismus _Picsp_1_962_NSC_04446
Karenia_brevis_SP1_CAMPEP_0189163984 _CP
Karenia _mikimotoi _ANV_76766 _CP
Karenia _papilionacea _ANV_49893 _CP
Karlodinium _micrum _Strain_CCMP2283 _CAMPEP_0200791840 _unpred
Karlodinium _sp_RCC3446_ANV_58248 _unpred
SPTP_Karlodinium _armiger _ANV_53677 _CP
Takayama _helix_ANV_109359
Takayama _helix_ANV_8056
Naegleria _gruberi _59458 _estExt_fgeneshHS _pg_C_620009
Aureococcus _anophagefferens _Auran1_11148
Pelagophytes _1_504838 _fgenesh 1_kg_2___667___Locus 2788 v1rpkm 66_16
Fistulifera _solaris _fso_g20331
Fragilariopsis _cylindrus _Fracy1_181148
Bigelowiellia _natans _85238
Alexandrium _andersonii _CCMP2222_CAMPEP_0197802696
Alexandrium _catenella _OF101_CAMPEP_0171233768
Alexandrium _monilatum _CCMP3105_CAMPEP_0200692536
Alexandrium _temarense _CCMP1771_CAMPEP_0186205392
Gambierdiscus _australes _CAWD_149_CAMPEP_0171096342
Pyrodinium _bahamense _pbaha 01_CAMPEP_0179026102
Ceratium _fusus _PA161109 _CAMPEP_0199505472
Lingulodinium _polyedra_CCMP1738 _CAMPEP_0189952498
Amphidinium _carterae_CCMP1314 _CAMPEP_0186433980
Azadinium _spinosum _CAMPEP_0186776478
Symbiodinium _kawagutii _Skav233315
Symbiodinium _sp_C1_CAMPEP_0199588004
Symbiodinium _minutum _037866
Polarella _glacialis _CCMP_1383 _CAMPEP_0115089120
Brandtodinium _nutriculum _RCC3387 _CAMPEP_0198511164
Peridinium _aciculiferum _PAER_2_CAMPEP_0190703518
Scrippsiella _hangoei _like_SHHI_4_CAMPEP_0199173896
Scrippsiella _trochoidea _CCMP3099 _CAMPEP_0192062700
Heterocapsa _triquestra _CCMP_448 _CAMPEP_0195138844
Prorocentrum _minimum _CCMP1329 _CAMPEP_0190898834
Togula _jolla _CCCM_725_CAMPEP_0170607188
Ceratium _fusus _PA161109 _CAMPEP_0199505326
Gonyaulax_spinifera _CCMP409 _CAMPEP_0197931904
Karenia _brevis_SP3_CAMPEP_0189291366 _CP
Karenia _mikimotoi _ANV_52723
Karenia _papilionacea _ANV_57224
Karlodinium _armiger _ANV_42370
Karlodinium _micrum _Strain _CCMP2283 _CAMPEP_0200793760
Karlodinium _sp_RCC3446_ANV_65318
Takayama _helix_ANV_81194
Ross _Sea _Dinoflagellate _197990
Pelagodinium _beii _RCC1491 _CAMPEP_0197623456
Polarella _glacialis _CCMP_1383 _CAMPEP_0115096106
Symbiodinium _kawagutii _CCMP2468 _CAMPEP_0192378428
Symbiodinium _minutum _006416
Symbiodinium _sp_C1_CAMPEP_0199622522
Symbiodinium _microadriaticum _Smic17414
Bathycoccus_prasinos _612387948
Ostreococcus _PACid_27420482
Micromonas _PACid_27340446
Guillardia _theta_87963
Chrysochromulina _brevifilum _UTEX_LB_985_CAMPEP_0174760772
Chrysochromulina _rotalis _UIO044_CAMPEP_0115847382
Haptolina _ericina _CCMP281_CAMPEP_0181190636
Prymnesium _parvum _LXRN2006770
Dicrateria _rotunda _RCC918 _CAMPEP_0206154828
Prymnesium _polylepis _CCMP1757 _CAMPEP_0193801446
Phaeocystis _antarctica _Caron _Lab _Isolate _CAMPEP_0172951124
Ross _Sea _Dinoflagellate _544240
Phaeocystis _rex_CCMP2000_CAMPEP_0119058876
Syracosphaerales _Calcidiscus _leptoporus _RCC1130 _CAMPEP_0119398400
Syracosphaerales _Coccolithus _pelagicus _ssp _braarudi _PLY182g _CAMPEP_0183379644
Syracosphaerales _Pleurochrysis _carterae _CCMP645_CAMPEP_0190752396
Syracosphaerales _Scyphosphaera _apsteinii _RCC1455 _CAMPEP_0119327022
Emiliania _huxleyi _CCMP370 _CAMPEP_0187689426
Isochrysis _sp_CCMP1244 _CAMPEP_0188744306
Gephyrocapsa _oceanica _RCC1303 _CAMPEP_0188173996
Isochrysis _galbana _CCMP1323 _CAMPEP_0193697292
Isochrysis _sp_BAJW2040459
Chrysoculter _rhomboideus _RCC1486 _CAMPEP_0119429850
Pavlova_gyrans _CCMP608 _CAMPEP_0206048312
Pavlova _lutheri _RCC1537 _CAMPEP_0205982442
Pavlova_sp_CCMP459 _CAMPEP_0190506694
Exanthemachrysis _gayraliae _RCC1523 _CAMPEP_0205998594
Aureococcus _anophagefferens _Auran1_30004
Pelagophytes _1_603235 _MIX10301 _10535 _128
Vitrella _brassicaformis _84587534
Bigelowiellia _natans _40797
Ochrophyta
Ectocarpus _siliculosus _Esi0071_0031
Fistulifera _solaris _fso_g10874
Fragilariopsis _cylindrus _Fracy1_224334
Thalassiosira _Thaoc_74067
Thalassiosira _Thaps3a_5078
SPTP_Takayama _helix_ANV_72337 _CP
Nannochloropsis _gaditana _Naga _100046 g11
Guillardia _theta _99089
Chrysochromulina _brevifilum _UTEX_LB_985_CAMPEP_0174739630
Chrysochromulina _rotalis _UIO044_CAMPEP_0115851274
Prymnesium _polylepis _CCMP1757 _CAMPEP_0193718762
Prymnesium _parvum _Texoma1_CAMPEP_0191226564
Dicrateria _rotunda _RCC918 _CAMPEP_0206181720
Phaeocystis _antarctica _CCMP1374 _CAMPEP_0198160088
Ross _Sea _Dinoflagellate _76934
Syracosphaerales _Calcidiscus _leptoporus _RCC1130 _CAMPEP_0119362786
Syracosphaerales _Coccolithus _pelagicus _ssp _braarudi _PLY182g _CAMPEP_0183343024
Syracosphaerales _Pleurochrysis _carterae _CCMP645 _CAMPEP_0190758162
Syracosphaerales _Scyphosphaera _apsteinii _RCC1455 _CAMPEP_0119309204
Emiliania _huxleyi _CCMP370 _CAMPEP_0187683022
Gephyrocapsa _oceanica _RCC1303 _CAMPEP_0188167396
Isochrysis _sp_CCMP1244 _CAMPEP_0188739178
Isochrysis _galbana _CCMP1323 _CAMPEP_0193637414
Chrysoculter _rhomboideus _RCC1486 _CAMPEP_0119422368
Exanthemachrysis _gayraliae _RCC1523 _CAMPEP_0206031348
Pavlova _sp_CCMP459 _CAMPEP_0190527134
Pavlova _lutheri _RCC1537 _CAMPEP_0205972348
Togula _jolla _CCCM_725_CAMPEP_0170597666
Plasmodium _vivax_Sal_1_156081708
Dictyostelium _discoideum _7467
Phytophthora _sojae _475569
Entamoeba _histolytica _HM_1_IMSS_A_480526902
Batrachochytrium _dendrobatidis _JAM81_575479140
Schizosaccharomyces _pombe _429240963
Blastocystis _hominis _00001630001
Phytophthora _sojae _559658
Chlamydomonas _PACid_27574647
Volvox_carteri_f_nagariensis _302845602
Chondrus _crispus _507113118
Porphyridium _sp_2068_21
Galdieria _sulphuraria _452824983
Pyropia_sp_31092_g7598
Cyanidioschyzon _merolae _strain _10D_544212155
Toxoplasma _gondii _gi_237844313
Cyanophora _paradoxa _37764
Paramecium _tetraurelia _strain _d4_2_gi145549355
Tetrahymena _thermophila _118399178
Dictyostelium _discoideum _148579
Symbiodinium _sp_D1a_CAMPEP_0196908680
Drosophila _melanogaster _24643061
Trichoplax _sp_Triad1_58286
Monosiga _brevicolis _37112 _estExt_fgenesh1_pg_C_100166
Nematostella _vectensis _156386252
Reticulomyxa _filosa _569441316
0.5

Figure EV4. Automatically generated phylogenetic tree showing the green origin of plastidial histidyl-tRNA synthetases of *Karenia* and *Karlodinium* and brown origin of the plastidial homolog in *Takayama*.

A green-like homolog present in *Takayama* is not predicted as plastidial. The *Karenia brevis* SP3 homolog clustering with the dinoflagellates was likely falsely predicted as plastid-targeted due to an artificial N-terminal extension of the transcript. Bootstrap support is expressed by the branch color (black for ≥90%, dark gray for ≥75%, light gray for <75%). Kareniacean sequences are colored pink with proteins with predicted plastid-targeting signal in darker shades and annotated "CP" suffix; dinoflagellates are coded blue; haptophytes are coded orange; green algae (Chloroplastida) are coded green; and brown algae (Ochrophyta) are coded brown.

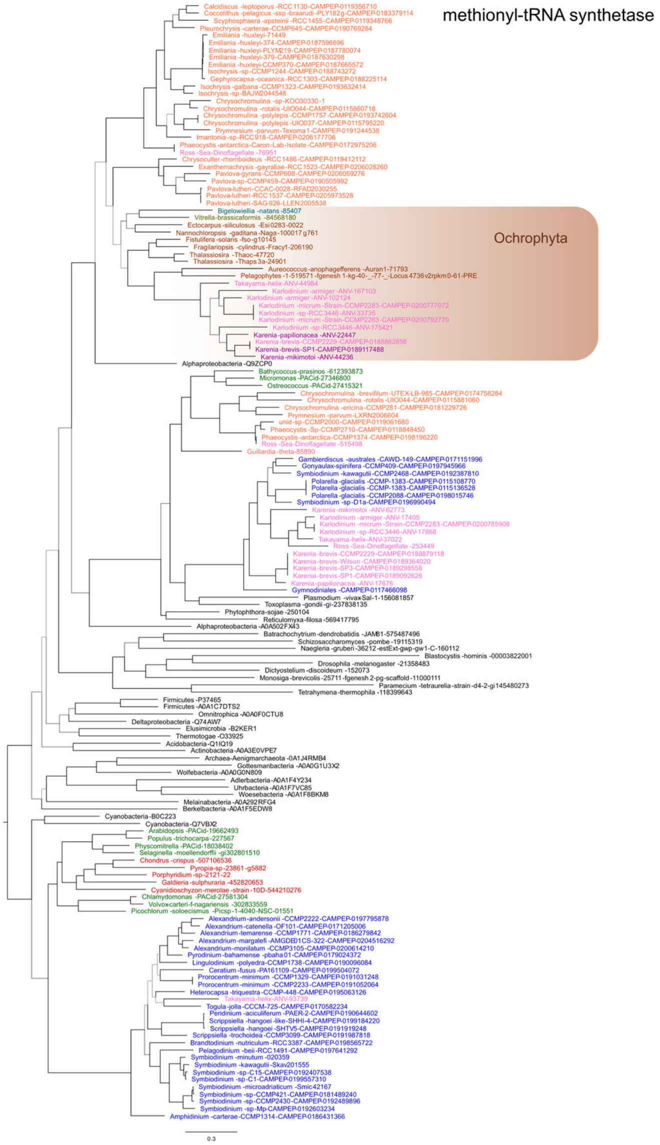

methionyl-tRNA synthetase
Calcidiscus -leptoporus -RCC1130-CAMPEP-0119356710
Coccolithus -pelagicus -ssp -braarudi -PLY182g-CAMPEP-0183379114
Scyphosphaera -apsteinii -RCC1455-CAMPEP-0119348766
Pleurochrysis -carterae-CCMP645-CAMPEP-0190769284
Emiliania -huxleyi-71449
Emiliania -huxleyi-374-CAMPEP-0187596696
Emiliania -huxleyi-PLYM219-CAMPEP-0187780074
Emiliania -huxleyi-379-CAMPEP-0187630298
Emiliania -huxleyi-CCMP370-CAMPEP-0187665572
Isochrysis -sp-CCMP1244-CAMPEP-0188743272
Gephyrocapsa -oceanica -RCC1303-CAMPEP-0188225114
Isochrysis -galbana -CCMP1323-CAMPEP-0193632414
Isochrysis -sp-BAJW2044548
Chrysochromulina -sp-KOO30330-1
Chrysochromulina -rotalis -UIO044-CAMPEP-0115860718
Chrysochromulina -polylepis -CCMP1757-CAMPEP-0193742604
Chrysochromulina -polylepis -UIO037-CAMPEP-0115795220
Prymnesium -parvum-Texoma1-CAMPEP-0191244538
Imantonia -sp-RCC918-CAMPEP-0206177706
Phaeocystis -antarctica-Caron-Lab-Isolate-CAMPEP-0172975206
Ross -Sea-Dinoflagellate -76951
Chrysoculter-rhomboideus -RCC1486-CAMPEP-0119412112
Exanthemachrysis -gayraliae-RCC1523-CAMPEP-0206028260
Pavlova-gyrans-CCMP608-CAMPEP-0206059276
Pavlova-sp-CCMP459-CAMPEP-0190505992
Pavlova-lutheri-CCAC-0028-RFAD2030255
Pavlova-lutheri-RCC1537-CAMPEP-0205973528
Pavlova-lutheri-SAG-926-LLEN2005538
Ochrophyta
Bigelowiellia -natans-85407
Vitrella-brassicaformis -84568180
Ectocarpus-siliculosus -Esi0283-0022
Nannochloropsis -gaditana-Naga-100017g761
Fistulifera-solaris -fso-g10145
Fragilariopsis -cylindrus-Fracy1-206190
Thalassiosira -Thaoc-47720
Thalassiosira -Thaps3a-24901
Aureococcus-anophagefferens -Auran1-71793
Pelagophytes-1-519571-fgenesh1-kg-40-_-77-_-Locus4736v2rpkm0-61-PRE
Takayama-helix-ANV-44984
Karlodinium -armiger -ANV-167103
Karlodinium -armiger -ANV-102124
Karlodinium -micrum -Strain-CCMP2283-CAMPEP-0200777072
Karlodinium -sp-RCC3446-ANV-33735
Karlodinium -micrum -Strain-CCMP2283-CAMPEP-0200792770
Karlodinium -sp-RCC3446-ANV-175421
Karenia-papilionacea -ANV-22447
Karenia-brevis-CCMP2229-CAMPEP-0188862856
Karenia-brevis-SP1-CAMPEP-0189117488
Karenia-mikimotoi -ANV-44236
Alphaproteobacteria -Q9ZCP0
Bathycoccus-prasinos -612393873
Micromonas -PACid-27346800
Ostreococcus -PACid-27415321
Chrysochromulina -brevifilum-UTEX-LB-985-CAMPEP-0174758284
Chrysochromulina -rotalis-UIO044-CAMPEP-0115881060
Chrysochromulina -ericina-CCMP281-CAMPEP-0181229726
Prymnesium -parvum-LXRN2006604
unid-sp-CCMP2000-CAMPEP-0119061680
Phaeocystis-Sp-CCMP2710-CAMPEP-0118848450
Phaeocystis-antarctica-CCMP1374-CAMPEP-0198196220
Ross -Sea-Dinoflagellate -515498
Guillardia-theta-85890
Gambierdiscus -australes -CAWD-149-CAMPEP-0171151996
Gonyaulax-spinifera -CCMP409-CAMPEP-0197945966
Symbiodinium -kawagutii-CCMP2468-CAMPEP-0192387810
Polarella-glacialis -CCMP-1383-CAMPEP-0115108770
Polarella-glacialis -CCMP-1383-CAMPEP-0115136528
Polarella-glacialis -CCMP2088-CAMPEP-0198015746
Symbiodinium -sp-D1a-CAMPEP-0196990494
Karenia-mikimotoi -ANV-62773
Karlodinium -armiger -ANV-17405
Karlodinium -micrum -Strain-CCMP2283-CAMPEP-0200785908
Karlodinium -sp-RCC3446-ANV-17868
Takayama-helix-ANV-37022
Ross -Sea-Dinoflagellate -253449
Karenia-brevis-CCMP2229-CAMPEP-0188879118
Karenia-brevis-Wilson-CAMPEP-0189364020
Karenia-brevis-SP3-CAMPEP-0189298558
Karenia-brevis-SP1-CAMPEP-0189092626
Karenia-papilionacea -ANV-17676
Gymnodiniales -CAMPEP-0117466098
Plasmodium -vivax-Sal-1-156081857
Toxoplasma -gondii-gi-237838135
Phytophthora-sojae-250104
Reticulomyxa-filosa-569417795
Alphaproteobacteria -A0A502FX43
Batrachochytrium-dendrobatidis -JAM81-575487496
Schizosaccharomyces -pombe-19115319
Naegleria -gruberi-36212-estExt-gwp-gw1-C-160112
Blastocystis -hominis -00003822001
Drosophila -melanogaster -21358483
Dictyostelium -discoideum -152073
Monosiga-brevicolis-25711-fgenesh2-pg-scaffold-11000111
Paramecium -tetraurelia-strain-d4-2-gi145480273
Tetrahymena-thermophila-118399643
Firmicutes -P37465
Firmicutes -A0A1C7DTS2
Omnitrophica -A0A0F0CTU8
Deltaproteobacteria -Q74AW7
Elusimicrobia -B2KER1
Thermotogae -O33925
Acidobacteria-Q1IQ19
Actinobacteria-A0A3E0VPE7
Archaea-Aenigmarchaeota -0A1J4RMB4
Gottesmanbacteria -A0A0G1U3X2
Wolfebacteria-A0A0G0N809
Adlerbacteria-A0A1F4Y234
Uhrbacteria-A0A1F7VC85
Woesebacteria -A0A1F8BKM8
Melainabacteria-A0A292RFG4
Berkelbacteria-A0A1F5EDW8
Cyanobacteria-B0C223
Cyanobacteria-Q7VBX2
Arabidopsis -PACid-19662493
Populus -trichocarpa-227567
Physcomitrella -PACid-18038402
Selaginella -moellendorffii -gi302801510
Chondrus -crispus -507106536
Pyropia-sp-23861-g5882
Porphyridium -sp-2121-22
Galdieria -sulphuraria -452820653
Cyanidioschyzon-merolae-strain-10D-544210276
Chlamydomonas -PACid-27581304
Volvox-carteri-f-nagariensis -302833559
Picochlorum -soloecismus -Picsp-1-4040-NSC-01551
Alexandrium-andersonii -CCMP2222-CAMPEP-0197795878
Alexandrium-catenella-OF101-CAMPEP-0171205006
Alexandrium-temarense -CCMP1771-CAMPEP-0186279842
Alexandrium-margalefi -AMGDE01CS-322-CAMPEP-0204516292
Alexandrium-monilatum -CCMP3105-CAMPEP-0200614210
Pyrodinium-bahamense -pbaha01-CAMPEP-0179024372
Lingulodinium -polyedra-CCMP1738-CAMPEP-0190096084
Ceratium-fusus-PA161109-CAMPEP-0199504072
Prorocentrum-minimum -CCMP1329-CAMPEP-0191031248
Prorocentrum-minimum -CCMP2233-CAMPEP-0191052064
Heterocapsa -triquestra-CCMP-448-CAMPEP-0195063126
Takayama-helix-ANV-93739
Togula-jolla-CCCM-725-CAMPEP-0170582234
Peridinium -aciculiferum -PAER-2-CAMPEP-0190644602
Scrippsiella -hangoei -like-SHHI-4-CAMPEP-0199184220
Scrippsiella -hangoei -SHTV5-CAMPEP-0191919248
Scrippsiella -trochoidea-CCMP3099-CAMPEP-0191987818
Brandtodinium -nutriculum-RCC3387-CAMPEP-0198565722
Pelagodinium -beii-RCC1491-CAMPEP-0197641292
Symbiodinium -minutum -020359
Symbiodinium -kawagutii-Skav201555
Symbiodinium -sp-C15-CAMPEP-0192407538
Symbiodinium -sp-C1-CAMPEP-0199557310
Symbiodinium -microadriaticum -Smic42167
Symbiodinium -sp-CCMP421-CAMPEP-0181489240
Symbiodinium -sp-CCMP2430-CAMPEP-0192489896
Symbiodinium -sp-Mp-CAMPEP-0192603234
Amphidinium -carterae-CCMP1314-CAMPEP-0186431366
0.3

Figure EV5. Automatically generated phylogenetic tree showing the brown origin of plastidial methionyl-tRNA synthetase in all three genera.

Bootstrap support is expressed by the branch color (black for ≥90%, dark gray for ≥75%, and light gray for <75%). Kareniacean sequences are colored pink with proteins with predicted plastid-targeting signal in darker shades and annotated "CP" suffix; dinoflagellates are coded blue; haptophytes are coded orange; and brown algae (Ochrophyta) are coded brown.