

Multiple parallel origins of parasitic Marine Alveolates

Received: 3 August 2023

Accepted: 20 October 2023

Published online: 03 November 2023

 Check for updates

Corey C. Holt ^{1,2,7} ✉, Elisabeth Hehenberger ^{1,3,7} ✉, Denis V. Tikhonenkov^{4,5}, Victoria K. L. Jacko-Reynolds ¹, Noriko Okamoto ^{1,2}, Elizabeth C. Cooney ^{1,2}, Nicholas A. T. Irwin ^{1,6} & Patrick J. Keeling ¹ ✉

Microbial eukaryotes are important components of marine ecosystems, and the Marine Alveolates (MALVs) are consistently both abundant and diverse in global environmental sequencing surveys. MALVs are dinoflagellates that are thought to be parasites of other protists and animals, but the lack of data beyond ribosomal RNA gene sequences from all but a few described species means much of their biology and evolution remain unknown. Using single-cell transcriptomes from several MALVs and their free-living relatives, we show that MALVs evolved independently from two distinct, free-living ancestors and that their parasitism evolved in parallel. Phylogenomics shows one subgroup (MALV-II and -IV, or Syndiniales) is related to a novel lineage of free-living, eukaryovorous predators, the eleutherids, while the other (MALV-I, or Ichthyodinida) is related to the free-living predator *Oxyrrhis* and retains proteins targeted to a non-photosynthetic plastid. Reconstructing the evolution of photosynthesis, plastids, and parasitism in early-diverging dinoflagellates shows a number of parallels with the evolution of their apicomplexan sisters. In both groups, similar forms of parasitism evolved multiple times and photosynthesis was lost many times. By contrast, complete loss of the plastid organelle is infrequent and, when this does happen, leaves no residual genes.

The marine alveolates (MALVs) are an elusive group of microbial eukaryotes first discovered through amplicon surveys of environmental rRNA^{1,2}, and since found to consistently dominate eukaryotic metabarcoding surveys in global oceans^{3–8}. Virtually all MALVs are known only from these rRNA SSU gene fragments, which have been phylogenetically linked to a handful of described parasites that branch deeply in the dinoflagellates^{1,9–11}. MALV sequences in marine environments are accordingly interpreted to represent a large and diverse population of dinospores that infect animal or protist hosts. Their abundance, diversity, and phylogenetic position all emphasize the evolutionary and ecological significance of MALVs but, since their

discovery, new insights beyond their SSU rRNA gene diversity have been limited by a lack of genomic data and the inability to culture all but a few species.

MALVs are typically subdivided into five groups based on rRNA (I–V) and are often considered to be a monophyletic clade comprising the order Syndiniales⁷. Phylogenetic analyses support the branching of MALVs at the base of the dinoflagellates, but concatenating SSU and LSU rRNA genes suggests MALVs may be paraphyletic, with MALV-II and -IV forming one group and MALV-I diverging earlier¹². Comprehensive genomic data are only available for two described MALVs, *Amoebophrya* and *Hematodinium*, which are in the MALV-II and IV

¹Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada. ²Hakai Institute, Heriot Bay, British Columbia, Canada. ³Institute of Parasitology, Biology Centre Czech Academy of Sciences, České Budějovice, Czech Republic. ⁴Papanin Institute for Biology of Inland Waters, Russian Academy of Sciences, Borok, Russia. ⁵AquaBioSafe Laboratory, University of Tyumen, Tyumen, Russia. ⁶Present address: Merton College, University of Oxford, Oxford, UK. ⁷These authors contributed equally: Corey C. Holt, Elisabeth Hehenberger. ✉ e-mail: corey.holt@ubc.ca; elisabeth.hehenberger@paru.cas.cz; pkeeling@mail.ubc.ca

groups, respectively^{13–15}. Both lack evidence of a plastid, despite the fact that plastids are found in all other closely-related dinoflagellate lineages^{14,15}. Losing photosynthesis is relatively common, but losing the plastid entirely is not^{14,16}, and this absence of a plastid has created some controversy about the evolution of the organelle in close relatives¹⁷. However, the lack of data beyond rRNA gene sequencing for most MALVs, and in particular the major MALV-I group (with the exception of uncultured Single-cell Assembled Genome (SAG) data), together with the lack of data from free-living relatives of these apparently obligate parasites, makes it difficult to untangle the evolution of parasitism and plastids.

Here, we show, using single-cell transcriptomes and phylogenomic analysis, that parasitism evolved multiple times in MALVs from two distinct, free-living lineages. MALV-II and MALV-IV (the Syndiniales) share a common ancestor with a new group of free-living heterotrophs, the eleftherids, whereas MALV-I shares a common ancestor with a different free-living lineage, *Oxyrrhis marina*. Transcriptomic data from the psammosids confirm their position as early-branching dinoflagellates, while new MALV-I data support the need to revise the genus *Euduboscquella* and suggest MALV-I (now the Ichthyodinida) retains a cryptic plastid and should no longer be considered part of the Syndiniales.

Results and discussion

MALV-II and -IV are closely related to the eleftherids, a new group of free-living heterotrophs

We isolated and cultured several strains collectively representing two new species of colorless, eukaryovorous flagellates, which we formally name *Eleftheros xomoi* (strains Colp-37 and Cur-11, isolated from the surface of a coral in the Caribbean Sea, Curaçao) and *Eleftheros karadeniz* (strain Colp-25, isolated from marine near-shore sediments in the Black Sea, Crimea) (see “Taxonomic summary” below for taxonomic diagnosis). Sequences sufficiently similar to eleftherid SSU rRNA genes were not found in any of the largest global planktonic environmental surveys^{18,19}, but we did recover small numbers of amplicons from marine sediment data, with a higher prevalence (albeit with often low read counts) found in deep sea samples²⁰ (Supplementary Fig. 1; Supplementary Data 1), suggesting that eleftherids are rare and benthic. *Eleftheros* are very small (~4 µm), bean-shaped or roundish, fast-swimming cells with two heterodynamic flagella, externally resembling the biflagellate dinospores of MALVs (Fig. 1, Supplementary Fig. 2 and Supplementary Movie 1). Many of their morphological features are reminiscent of dinoflagellates, including longitudinal and transverse grooves similar to cingulum and sulcus (Fig. 1a, c, f, j–m, o, q). Eleftherids possess large alveoli beneath the plasma membrane (Fig. 1n–r), flagellar basal bodies orientated predominantly at an acute angle to each other, a flagellar transition zone including an axosome with extended central microtubules and a transverse plate just below the cell surface (Fig. 1p and Supplementary Fig. 2a), enlarged perinuclear space containing tubular mastigonemes (Fig. 1n, q and Supplementary Fig. 2b), bowling pin-shaped trichocysts with square cross-sections scattered throughout the cytoplasm that are cross-striated after discharging (Fig. 1n, o, s and Supplementary Fig. 2c–e), a large convoluted mitochondrion with tubular cristae (Fig. 1n, o, q and Supplementary Fig. 2a, f), and storage compounds in the form of roundish granules (Fig. 1n and Supplementary Fig. 2a, g). Interestingly, we also observed large and distinctive vesicular compartments resembling the rhoptries of apicomplexan and perkinsid parasites, which are also present in the parasitic MALVs (Fig. 1o, r).

To determine their phylogenetic position, we generated transcriptomic data from both whole cultures and manually isolated eleftherid cells and built a concatenated 75-taxon/236 protein dataset (63,267 sites). Maximum-likelihood (ML) analyses consistently recovered eleftherids as sister to *Amoebophrya* and *Hematodinium* (MALV-II and -IV) with maximum statistical support (Fig. 2a). To examine the

possibility that long-branch attraction affected the position of eleftherids, we performed fast-evolving site removal analysis until 50% of sites were removed, which showed that the support for the eleftherids as sisters to MALV-II and -IV remained 100% (Fig. 2b).

MALV-I is closely related to *Oxyrrhis*, another free-living heterotroph

Currently, only rRNA gene fragments and a small number of SAGs are available from the other major MALV group, MALV-I, and none of this group are in culture. Accordingly, we sought infected hosts from marine samples and obtained one infected cell from a deep-water plankton tow near Quadra Island, British Columbia, and two cells from near-shore plankton tows at Piscadera Bay, Curaçao. The Quadra Island cell was identified from its morphology as the dinoflagellate *Polykrikos* sp. but was also observed to contain a large and distinctive circular “fried egg-like” inclusion measuring ~75 µm by 82 µm with a central nucleus and centripetal grooves on a cuticular disc (Fig. 2c; Supplementary Movie 2); a common morphological characteristic of a mature trophonts in MALV-I parasites, including others infecting dinoflagellates²¹. Of the two cells from Curaçao: one was an infected *Warnowia* sp. with a smaller, ovoid, brown inclusion measuring ~27 µm by 35 µm (Fig. 2d; Supplementary Movie 3), and the other was found with no discernable host. While the morphology of the former lacks sufficient resolution for definitive descriptions, it appears to contain a “brain-like” groove texture (Supplementary Movie 3) reminiscent of *Euduboscquella* sp. 4 (OP445725) and ciliate-infecting *Euduboscquella* species²¹. The remaining cell from Curaçao, on the other hand, showed clear centripetal grooves and ridges, measured ~68 µm by 70 µm, and resembled a late-stage trophont or perhaps a newly emerged tomont (Fig. 2e; Supplementary Movie 4). Although we could not determine the number of grooves (which can distinguish species of *Euduboscquella*) in each cell, like those described in Yoo et al.²¹, cells with this morphology appeared to be more similar (in terms of number of grooves) to *E. melo* (~30) compared to *E. nucleocola* (more than 100)²², both of which are dinoflagellate-infecting *Euduboscquella* spp.

All three cells were isolated, and single-cell transcriptomes were sequenced, which revealed two phylogenetically-distinct signals in two of the cells—one being the above-mentioned hosts and the other an uncharacterized MALV-I (Fig. 2a)—while no potential host sequences were found in the remaining cell. Phylogenetic analysis of full-length SSU rRNA gene sequences extracted from the *Polykrikos*-associated transcriptomic data showed the MALV-I parasite (herein referred to as Ichthyodinida sp. 1) clustering with full-support with a clade of unidentified environmental sequences in the MALV-I clade (Fig. 3). This fell within a larger group of parasites isolated from dinoflagellates, which were ascribed to *Euduboscquella* based on conserved morphological traits²¹. This group corresponds to MALV-I clade 3 elsewhere⁷. The MALV-I sequence derived from *Warnowia*-associated transcriptomic data (herein referred to as Ichthyodinida sp. 2), on the other hand, branched at the base of the ciliate-infecting *Euduboscquella* clade 4⁷, while the remaining, apparently host-less, MALV-I cell (Ichthyodinida sp. 3) clustered within a group of environmental sequences (MALV-I clade 5) noted to be one of the most commonly retrieved groups in environmental libraries⁷. *Euduboscquella* has been isolated from both ciliates and dinoflagellates^{21–25}; however, the SSU rRNA gene tree (Fig. 3; Supplementary Fig. 3) showed MALV sequences isolated from these two host types form two phylogenetically discrete groups. Indeed, MALV-I diversity seems to fall into a small number of highly host-specific clades, infecting ciliates, fish eggs, dinoflagellates, and radiolarians. However, host specificity is noted to vary between MALV lineages, with some infecting a single host species and others infecting multiple^{25–27}. That being said, the branch lengths of MALV-I lineages in phylogenomic analyses (Fig. 2a; Supplementary Fig. 4), together with the presence of multiple distinct *Euduboscquella* clades in SSU phylogenies (Fig. 3; Supplementary Fig. 5), all suggest that the genus

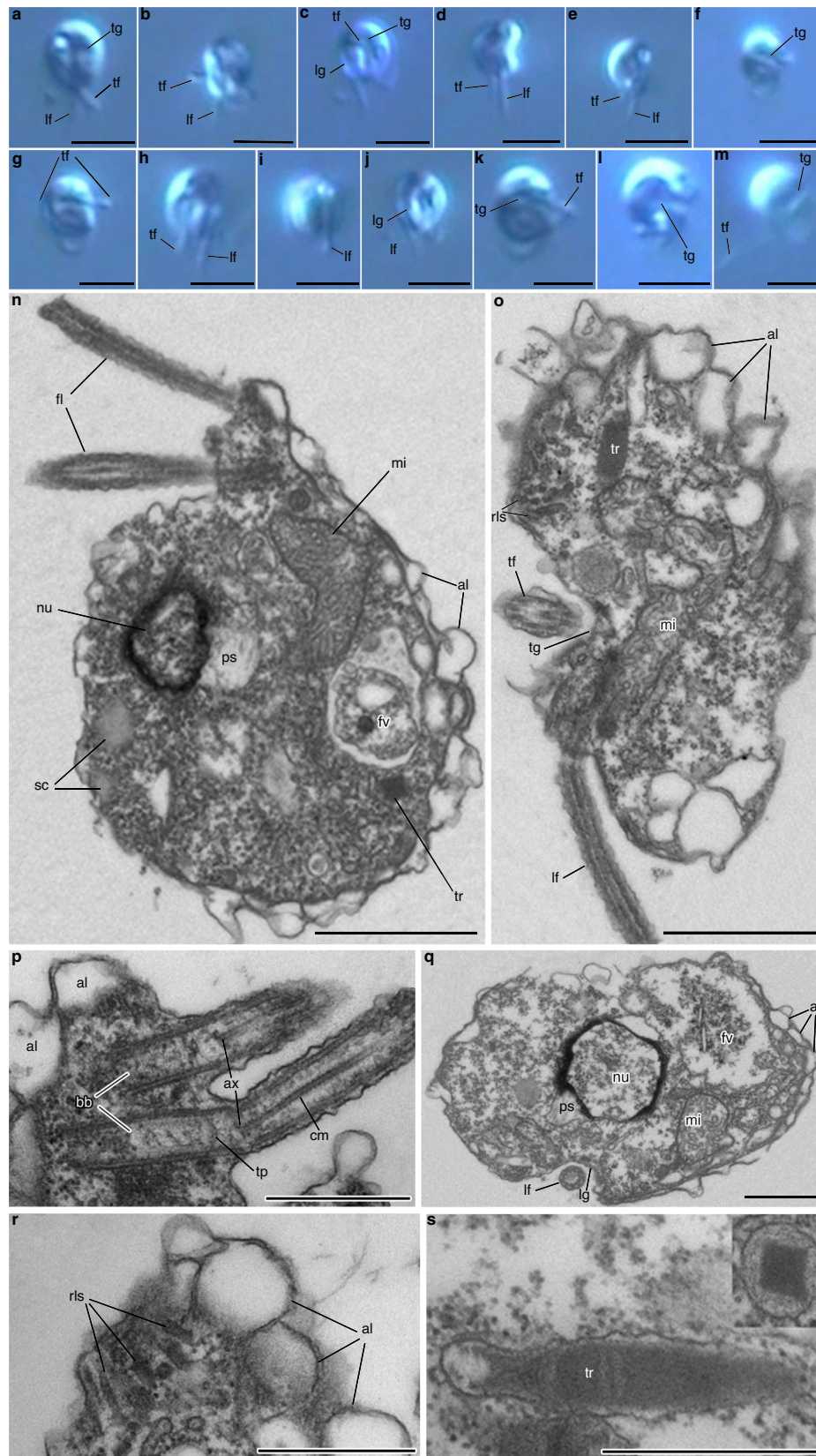


Fig. 1 | Cell morphology of *Eleftheros*. **a–j** Living cells of *E. xomoi* (strain Cur-11), visualized by light microscopy. **k–m** Living cells of *E. karadeniz*, visualized by light microscopy. **n–s** Cells of *E. karadeniz*, visualized by transmission electron microscopy. **n** Longitudinal section. **o** Oblique section showing the transverse groove. **p** Arrangement of basal bodies. **q** Transverse section showing the longitudinal groove. **r** Structure of the distal part with rhoptry-like structures and alveoli. **s** Longitudinal and cross section (inset) of trichocysts. al, alveoli; ax, axosome; bb, basal bodies; cm, central microtubules; fl, flagella; fv, food vacuole; lf, longitudinal flagellum; lg, longitudinal groove; mi, mitochondrion; nu, nucleus; ps, perinuclear space; rls, rhoptry-like structures; sc, storage compounds; tr, trichocyst; tf, transverse flagellum; tg, transverse groove; tp, transverse plate. Scale bars, 4 μm (**a–m**), 1 μm (**n, o, q**), and 0.5 μm (**p, r, s**). All observations were repeated at least 3 times with similar results.

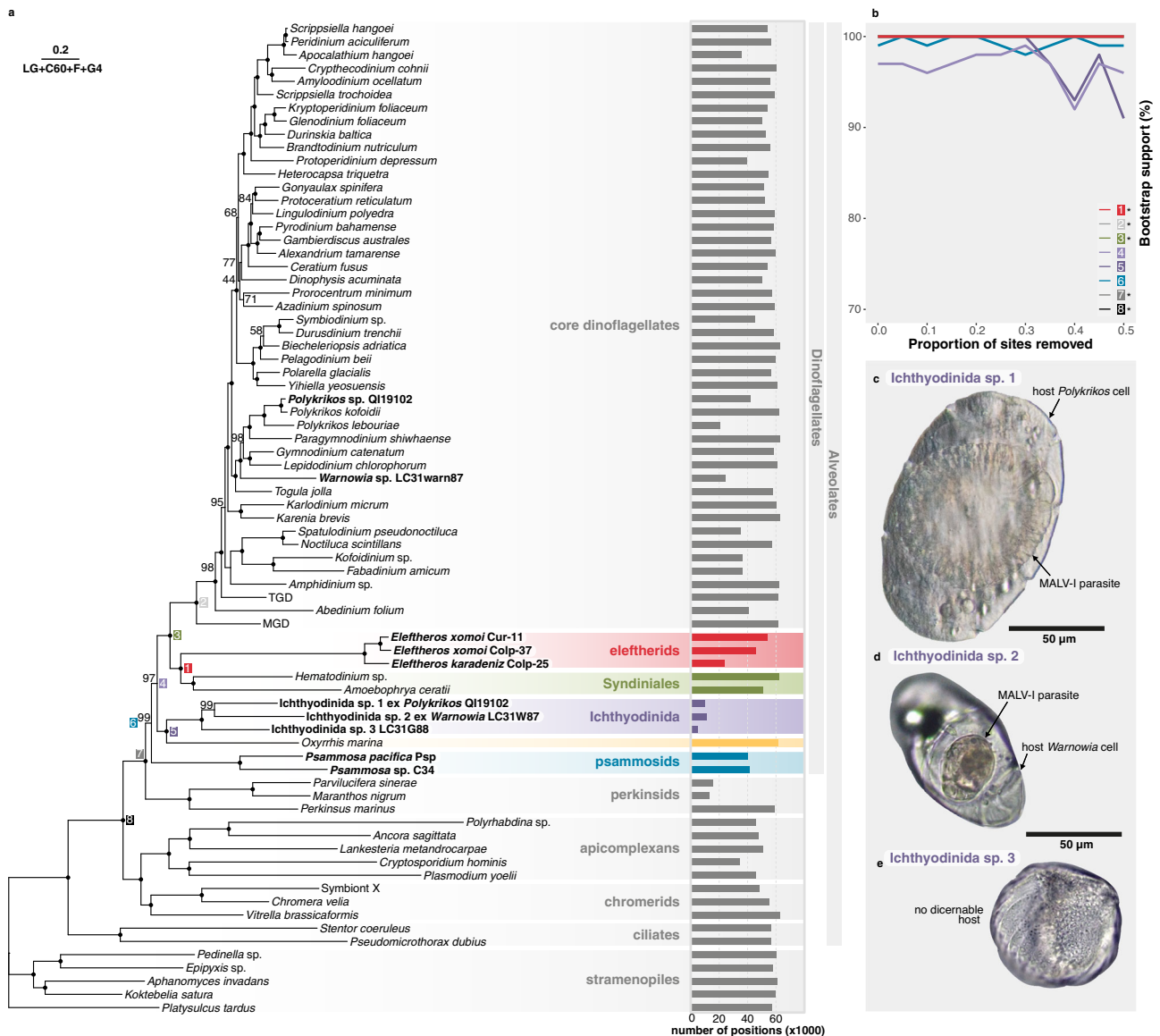


Fig. 2 | Phylogenomic reconstruction of the Marine Alveolates. **a** ML analysis of a 236 multi-protein alignment (LG + C60 + F + G4) separates MALV-I (purple) from the remaining Syndiniales (MALV-II and IV, green) to which the eleftherids (red) are the free-living sister group. Psammosids are shown in blue and *Oxyrrhis marina* in yellow. Black dots represent full statistical support (non-parametric bootstrap = 100%), values are shown for support below 100%. The percentage of amino acids present in the corresponding alignment for each taxon are shown as bars to the right of the tree. **b** UltraFast bootstrap support for given topologies, indicated by highlighted node labels in **(a)**, following removal of the fastest-evolving sites as

inferred using the LG + C 40 + F + G4 model. * indicates nodes that remained at 100%. Support for core dinoflagellate monophyly serves as a control for the presence of sufficient information for phylogenomic inference. **c** Living cell of *Polykrikos* sp. infected with a MALV-I-like parasite (Ichthyodinida sp. 1) showing characteristic centripetal grooves. **d** Living cell of *Warnowia* sp. infected with a MALV-I-like parasite (Ichthyodinida sp. 2). **e** MALV-I cell (Ichthyodinida sp. 3) showing no discernable host but characteristic centripetal grooves. Scale bar, 50 μm.

Euduboscquella requires revision, with two or potentially three additional genera needed to reflect the molecular diversity shown here.

Phylogenomic analyses including these new data reject the monophyly of MALV-I and MALV-II/IV, instead showing complete support for MALV-I branching specifically with another lineage of free-living heterotrophs, the common marine flagellate *Oxyrrhis* (Fig. 2a). This topology is further supported by the addition of published SAG data²⁸ (Supplementary Fig. 4). All topologies in which MALVs formed a monophyletic group were rejected by approximately unbiased (AU) tests (Supplementary Fig. 6). Indeed, MALVs, as currently defined, are neither monophyletic nor paraphyletic, since we can demonstrate complete support for separate lineages being sister groups to different free-living heterotrophs. Parasitic transitions in MALV-I and MALV-II therefore trace back to different free-living ancestors and must have

evolved their obligate parasitic life independently. We propose to retain the name Syndiniales for MALV-II/IV and to use the order Ichthyodinida proposed by Cavalier-Smith for MALV-I²⁹ (as opposed to the Coccidiales, which has been previously proposed but is based on a problematic type, *Coccidinium*^{23,30,31}). Whether MALV-III and the little-studied MALV-V can all be included in Ichthyodinida remains to be seen⁷ (Fig. 3, Supplementary Fig. 3).

Plastids and parasitism

The complete absence of evidence for a plastid in *Amoebophrya* and *Hematodinium*^{14,15} led to new hypotheses to account for the apparently complex distribution of plastids in early-diverging dinoflagellates and their sister clades, the apicomplexa, chrompodellids, and squirmids (or ACS clade)^{32–34}. However, it is difficult to evaluate any hypothesis

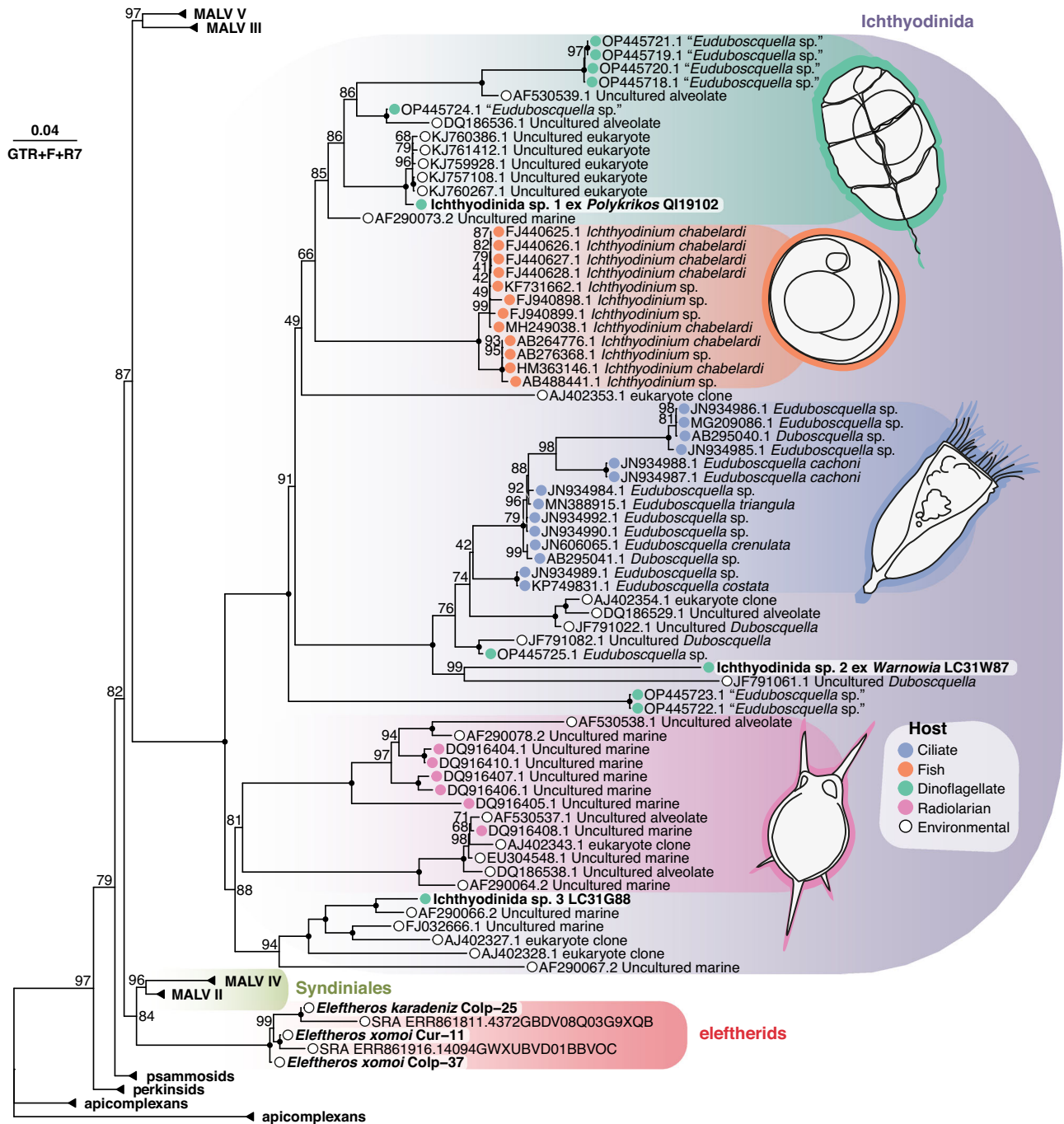


Fig. 3 | Contextualising new MALVs using SSU data. ML analysis (GTR + F + R7) showing MALV-I lineages (purple) cluster by host and those isolated in this study belong to distinct MALV-I clades. The eleftherids (red) again form a sister clade to the remaining Syndiniales (MALV-II and MALV-IV, green). Black dots at nodes represent full statistical support (UltraFast bootstrap = 100%), values are shown for

support below 100%. Colored circles accompanying MALV-I lineages reflect host identity: ciliates, blue; fish, orange; dinoflagellates, green; radiolarians, pink; and environmental, white. *Oxyrrhis* is not included due to its highly divergent rRNA SSU gene sequences.

about the origin of plastids in parasites without data from close free-living relatives. Now that such relatives are identified, we examined plastid function across the deeply-branching dinoflagellate groups, including another relevant lineage, *Psammosa*. *Psammosa* is a free-living heterotrophic dinoflagellate, but its exact position and whether it has a plastid are unclear in the absence of genomic data³⁵. *Psammosa pacifica* appears relatively rare in marine sediment datasets (Supplementary Data 1) but we re-established a culture of *P. pacifica* from the same location as the original description, alongside that of a novel lineage (with a partial SSU sequence 88.91% identical to *P. pacifica*),

and generated two transcriptomes from both strains. Phylogenomic analyses strongly support them forming an independent lineage at an important juncture in the tree, although exactly where they branch remains uncertain. In most analyses they branch basal to *Oxyrrhis*, Ichthyodinida, Syndiniales, and core dinoflagellates (Fig. 2a), but a position within the *Oxyrrhis*/Ichthyodinida clade itself is not rejected (Supplementary Fig. 6).

Eleftherids and (to a lesser extent) MALV-I and psammosids all possessed substantial numbers of nucleus-encoded genes for plastid-targeted proteins (as does *Oxyrrhis*), and these relate to metabolic

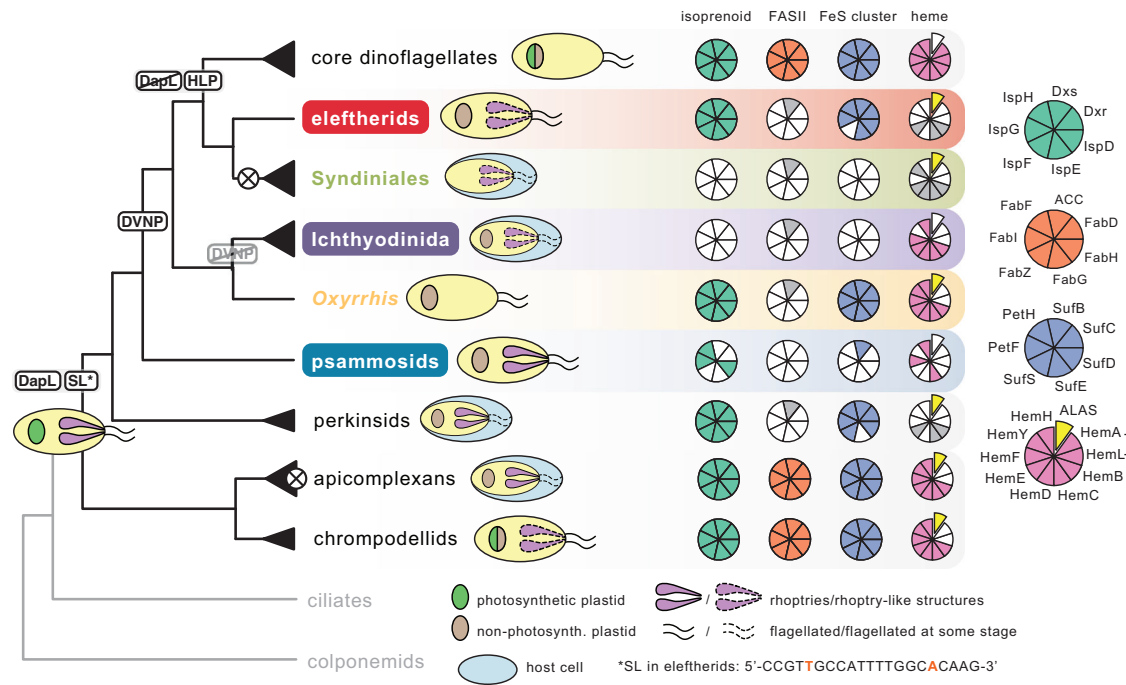


Fig. 4 | Character and plastid metabolic pathway evolution in dinoflagellates, apicomplexans, and related lineages. Schematic tree indicating abundance of and character evolution in dinoflagellates, apicomplexans, and related lineages. Black triangles/lines denote an estimation of the relative diversity of each lineage, while crossed-out circles indicate plastid loss (i.e., in the ancestor of the MALVs and within the apicomplexans). Taxa colored according to Fig. 2a: eleftherids, red; Syndiniales (MALV-II and -IV), green; Ichthyodinida (MALV-I), purple; *Oxyrrhis marina*, yellow; and psammosids, blue. Abbreviations on branches indicate the inferred origin of: L,L-diaminopimelate aminotransferase, DapL; SL, spliced leader; DVNPs, Dinoflagellate Viral NucleoProteins; HLPs, histone-like proteins. Next to each branch, Coulson plots depict selected plastidial metabolic pathways in the corresponding lineage. Grey segments indicate a cytosolic form of a given protein,

while an offset, yellow segment denotes the mitochondrial enzyme 5-aminolevulinic synthase (ALAS) that synthesizes 5-aminolevulinic in all lineages but the core dinoflagellates, where the same intermediate is synthesized by the plastidial enzymes glutamyl-tRNA reductase (HemA) and glutamate-1-semialdehyde 2,1-aminomutase (HemL). Red nucleotides in SL sequence indicate deviation from canonical SL. Pale blue ovals surrounding yellow cell depictions show host cells, indicating parasitism. Coulson plot headers refer to metabolic pathways: isoprenoid, methylerythritol phosphate (MEP) pathway for isoprenoid biosynthesis (green segments); FASII, type 2 fatty acid biosynthesis pathway (orange segments); FeS cluster, plastidial Fe-S cluster biosynthesis pathway including a ferredoxin system (navy blue segments); heme, heme biosynthesis (pink segments). Complete protein names are listed in Supplementary Data 2.

pathways typical for non-photosynthetic plastids³⁶. Specifically, in eleftherids and psammosids, we identified genes for the methylerythritol phosphate (MEP) pathway for isoprenoid biosynthesis and a plastidial SUF-type FeS cluster biosynthesis pathway including a ferredoxin redox system (Fig. 4 and Supplementary Fig. 7; phylogenetic trees available at <https://doi.org/10.6084/m9.figshare.19351700>). These are both widespread in non-photosynthetic dinoflagellates (non-photosynthetic core dinoflagellates and *Oxyrrhis*)^{36,37} and apicomplexans³⁴ but absent in Syndiniales¹⁴. The complete transcripts encode N-terminal extensions with one or more predicted transmembrane domains that are consistent with plastid targeting (IspF, SufC, SufD, SufE, PetF, PetH—in eleftherids), but do differ from canonical plastid-targeted sequences in core dinoflagellates³⁸ (alignments available at <https://doi.org/10.6084/m9.figshare.19351700>). No evidence for the plastidial type II fatty acid biosynthesis pathway was found, which is also absent in MALVs and all other dinoflagellates with non-photosynthetic plastids^{36,39} (Fig. 4 and Supplementary Fig. 7). Fatty acid biosynthesis in these groups is thought to occur via a cytosolic type I fatty acid synthase³⁶ (Supplementary Data 2). Heme is also synthesized in the plastids of some apicomplexans and dinoflagellates, but in eleftherids all these enzymes were mitochondrial (5-aminolevulinic synthase ALAS) or cytosolic (porphobilinogen synthase HemB, hydroxymethylbilane synthase HemC, uroporphyrinogen decarboxylase HemE), which is also the case in Syndiniales and *Perkinsus*. *Oxyrrhis*, by contrast, retains a mixed pathway with some enzymes inferred to be in the plastid, like dinoflagellates and apicomplexans (Supplementary Fig. 7; phylogenetic trees available at

<https://doi.org/10.6084/m9.figshare.19351700>). Like *Oxyrrhis*, we found MALV-I genomes encode HemB, HemC, HemE, HemF, and HemH which are all inferred to be targeted to the plastid: HemC exhibits a clear transmembrane region (TMR) which transports plastid proteins through the Golgi³⁸; HemB, HemH and HemF are truncated at the N-terminus but cluster with plastid-targeting dinoflagellates (like *Oxyrrhis*) in phylogenetic analyses; while HemE also contains N-terminal targeting information and falls within the so-called “eukaryotic” clade containing aplastidial organisms and some plastid-targeting dinoflagellates like *Oxyrrhis*. Notably HemE sequences from *Perkinsus*, *Hematodinium* and *Amoebophrya* do not have N-terminal extensions and are likely cytosolic. HemD (uroporphyrinogen-III synthase), the only enzyme encoded by MALV-II genomes that was proposed to have originated in the plastid¹⁴, is truncated in MALV-I. We did not find a HemD orthologue in eleftherids or psammosids (Fig. 4 and Supplementary Fig. 7), and it is also absent from many dinoflagellate transcriptomes, suggesting a low level of expression. However, in our phylogenetic reconstruction of HemD (Supplementary Fig. 8), photosynthetic eukaryotes and MALVs do not group with cyanobacteria, suggesting a bacterial origin for HemD independent from the plastid. The retention of plastidial versions of enzymes involved in heme biosynthesis in both MALV-I and *Oxyrrhis*, together with their absence in eleftherids and Syndiniales (Supplementary Fig. 7), indicates a functional redundancy that persisted long into the evolution of dinoflagellates and may have facilitated the repeated loss of plastid heme biosynthesis, and perhaps even the loss of the plastid itself in MALV-II.

The genome of *Hematodinium* (MALV-IV) contains another potential plastid pathway, the lysine biosynthesis enzyme L,L-diaminopimelate aminotransferase (DapL), which is only present in bacteria and a subset of plastids. Its phylogenetic origin in *Hematodinium* could not be resolved and support for its plastid origin was concluded to be ambiguous¹⁴. We identified DapL in eleutherids and psammosids (Fig. 4; protein absent in apicomplexans and most dinoflagellates) and phylogenetic reconstruction shows it to group with plastid homologs in cryptophytes, whereas the *Hematodinium* homolog appears to have originated independently from bacteria (Supplementary Fig. 9). While most eukaryotic DapL genes encode plastid-targeted sequences, eletherid and cryptophyte genes do not, indicating a likely cytosolic function (alignment available at <https://doi.org/10.6084/m9.figshare.19351700>). We could not identify DapL sequences in MALV-I.

We found no proteins related to photosynthesis or chlorophyll biosynthesis in eletherids, psammosids, MALV-I, or *Oxyrrhis* (as expected). We also identified no plastid import proteins; however, the import machinery is understudied in dinoflagellates and only a single component has been reported⁴⁰. There is no evidence for a plastid genome in any of the taxa investigated here, which is also expected since the plastid genome of photosynthetic dinoflagellates encodes only 12 proteins that are exclusively involved in photosynthesis⁴¹. We also investigated the 129 proteins predicted to localize to the apicoplast in the apicomplexan *Toxoplasma* by spatial proteomics⁴² and found no clear candidates for plastid-targeted homologs beyond the plastid pathways noted above. Taken together, the plastid in these non-photosynthetic lineages has been profoundly reduced.

Early character evolution in dinoflagellate-related lineages

Dinoflagellates evolved in many strange ways, beyond parasitism and plastids^{36,43,44}, and we now have a more accurate glimpse into these events as well. The nucleus, or “dinokaryon” is a good example: unlike other eukaryotes, dinoflagellates lack bulk nucleosomal chromatin and have huge genomes with permanently-condensed chromosomes structured by some combination of genes derived from horizontal gene transfer from viruses (Dinoflagellate Viral NucleoProtein, or DVNP)^{45,46} and bacteria (two Histone-Like Proteins, or HLPs). DVNP is present in all taxa analyzed here, with the notable exceptions of both psammosids and all MALV-I (Fig. 4 and Supplementary Data 2), whereas HLPs are present only in core dinoflagellates³⁶. DVNP is highly expressed and therefore generally easy to detect in dinoflagellate transcriptomes, so its complete absence from psammosid and MALV-I datasets raises the possibility that the gene might not be present. This is consistent with the phylogenetic position of *Psammosa*, but would require MALV-I to have lost this critical gene. Genomic data and/or biochemical evidence are needed to confirm this.

Dinoflagellate nuclei also perform trans-splicing of a highly-conserved spliced leader (SL) to the 5' end of many or most mRNAs⁴⁷. The same is also true of the perkinsids, however, the SL sequence is more variable⁴⁸. We identified complete SL sequences in a subset of transcripts from eleutherids, MALV-I, and psammosids, and found eletherids showed variations in sequence from the canonical SL sequence identified in core dinoflagellates and MALVs^{15,47} (Fig. 4).

The ultrastructure of eletherids also contains links to MALVs, which shed light on their diversification. Electron microscopy revealed the presence of rhoptyry-like structures (Fig. 10, r), similar to the electron-dense bodies found in the MALV-II *Amoebophrya*⁴⁹ and MALV-I *Ichthyodinium*¹⁰. Rhoptyries are components of the infection apparatus of apicomplexan parasites, their free-living relatives, as well as *Perkinsus* and *Psammosa*⁵⁰. In eletherids the presence of rhoptyry-like structures may indicate components of an apical complex were also retained with a role in feeding, as is the case in *Psammosa* and a handful of free-living lineages related to apicomplexans⁵¹. Altogether the distribution of rhoptyry-like structures and plastids supports a mixotrophic ancestor of dinoflagellates and apicomplexans (Fig. 4).

Parallelism in the evolution of parasitism

While we have focused mostly on the parasitic dinoflagellates, of course, the sisters to dinoflagellates are the much more famously parasitic apicomplexans (e.g., the malaria parasite *Plasmodium*). Due to their medical significance, apicomplexans are far better studied, and since it became clear they too evolved from algae and retained a cryptic plastid, their evolution has been thoroughly investigated^{52,53}. In the ACS clade, it is now clear that many of the major transitions seemingly associated with the origin of parasitism actually evolved multiple times in parallel. This includes the loss of photosynthesis, the loss of the plastid organelle, and even the origin of parasitism itself⁵⁴. Marine environmental surveys show MALVs dominate eukaryotic communities in terms of both diversity and abundance. Accordingly, they are likely to play an important role in marine ecology, yet we know shockingly little about their biology or evolution. It is now at least becoming clear that some of the same patterns emerging from apicomplexan research are also true of parasitic dinoflagellates, in particular parallel evolution in seemingly similar parasites.

Syndiniales (MALV-II and -IV) and Ichthyodinida (MALV-I) were previously thought to have evolved from a common parasitic ancestor^{1,2,7}, but phylogenomics now shows that they too evolved independently from different free-living ancestors, or in other words, that parasitism evolved twice in the Marine Alveolates. This is based on the same reasoning that is commonly applied to the ACS clade, namely that a transition from a free-living state to an obligately intracellular and parasitic state is more likely than a reversion from an obligate intracellular parasite to a free-living form. This is further supported by the fact that adaptation to intracellular parasitism is often reductive, and it is hard to reinvent complex traits (for example, feeding by phagocytosis) once lost. Moreover, because plastids are present in nearly every apicomplexan and dinoflagellate subgroup, and are demonstrably related⁵⁴, the ancestor of all these groups can be inferred to also have had a plastid. And because photosynthesis is still found in both dinoflagellates and ACS taxa (and they share a form of Rubisco unique among eukaryotes), this ancestor can also be inferred to have been photosynthetic. The prevalence of non-photosynthetic organisms branching at the base of the dinoflagellate tree has led to complex theories about the origin of their plastids¹⁷. However, if you also consider the ACS clade, a relatively straightforward model with multiple parallel losses of photosynthesis in early dinoflagellate evolution may not be that remarkable. While it is true that many cryptic plastids are found in early-branching dinoflagellates (MGD, *Abedinium*, TGD, *Eleftheros*, *Psammosa* and *Oxyrrhis*) and *Perkinsus*, they are all serviced by nucleus-encoded, plastid-targeted proteins that are phylogenetically related to homologs in peridinin plastids of photosynthetic dinoflagellates (<https://doi.org/10.6084/m9.figshare.19351700>), which very strongly argues they all evolved from the same ancestor. Based on the current tree, this would require multiple independent losses of photosynthesis, and it is these parallel events that have prompted more complex theories to explain plastid distribution¹⁷. However, if you count the number of losses in deep-branching taxa on the ACS side of the tree, you arrive at roughly the same number of independent losses of photosynthesis. Yet, multiple losses in the ACS clade attract no serious debate and are widely accepted to be the result of convergence. The key difference appears to be that the large, diverse, ecologically-successful apicomplexans are non-photosynthetic, so the retention of photosynthesis in a few “deep-branching” lineages (*Chromera* and *Vitrella*) leads to those lineages being treated as “oddballs”. On the dinoflagellate side of the tree, it is the large and successful group that (mostly) retained photosynthesis, so this is treated as somehow different when modeling ancient evolutionary events. We argue they should be treated equally: a loss is still a loss whether it preceded one modern species or a million of them.

The outright loss of the plastid organelle is very rare compared to the loss of photosynthesis, but this also happened in parallel in both

dinoflagellate and apicomplexan lineages. Plastids were lost multiple times in basal-branching apicomplexans³⁴ and at least once in dinoflagellates (Syndiniales)¹⁴. In all these cases, it is interesting to note how little evidence of the existence of a plastid is retained once the organelle has been lost. The idea that the endosymbiotic origin of plastids necessarily had an extensive and lasting genetic impact on their host genome has penetrated deeply into our thinking about organelle evolution, but the presence of a plastid in the ancestor of eleutherids and Syndiniales has left no impression on the genome of Syndiniales¹⁵, and indeed the only reason we know a plastid ever existed is due to their phylogenetic position, not because of any genetic “footprint” of the organelle. This is also true of apicomplexan lineages that have lost their plastid entirely^{34,53} and is further reinforced by the recent demonstration that the Picozoa are plastid-lacking members of the archaeplastids⁵⁵. Therefore, perhaps our baseline expectation should be that when an organelle is lost, it leaves no genetic trace, making it very difficult, without clear phylogenetic evidence, to distinguish a cell where an organelle was lost from one that never had the organelle in the first place.

Taxonomic summary

Taxonomy

Eukaryota

SAR Burki et al. 2008, emend. Sar Adl et al. 2012

Alveolata Cavalier-Smith 1991

Dinoflagellata Bütschli 1885, emend. Fensome et al. 1993, emend. Adl et al. 2005

Eleftheros n. gen. Tikhonenkov, Hehenberger, and Keeling

Etymology: Generic epithet means ‘free’. From Greek ‘Eleftheria’, freedom, liberty

Diagnosis: Naked and solitary eukaryovorous protist with two heterodynamic longitudinal and transverse flagella, large alveoli beneath the plasmalemma, bowling pin-shaped trichocysts, and rhoptry-like structures near the plasma membrane. Two basal bodies lie at an acute angle to each other or almost parallel. Large mitochondrion with tubular cristae.

Taxonomic remark: Free-living and fast-swimming cells outwardly resemble biflagellate dinospores of syndinian parasites (e.g. *Ichthyodinium*, *Amoebophrya*).

Zoobank registration: LSID for this publication:

urn:lsid:zoobank.org:pub:CBA7B765-7996-4F42-8AAF-42460CE7F77E. LSID for the new genus: urn:lsid:zoobank.org:act:DD5CBE62-6058-4ADA-BDBD-7B3A486447E1.

Type species: *Eleftheros xomoi*

Eleftheros xomoi n. sp. Tikhonenkov, Hehenberger, and Keeling

Diagnosis: Cells 3.6–4.8 µm long, 3.0–4.1 wide, bean-shaped or oval with wider anterior end. Longitudinal flagellum strait, lies in the longitudinal groove and directed backward, 1.5 times longer than the cell. Transverse flagellum about 2 times as long as cell, undulates in the transverse groove, which descends down the cell and widens. Coming out of the transverse groove below the cell, transverse flagellum makes an undulating loop around the cell, while the tip of the flagellum is directed downward.

Type figure: Fig. 1a.

Type material: The specimen shown in Fig. 1a is the holotype (see International Code of Zoological Nomenclature, Art. 72.5.6, Declaration 45).

Gene sequence: The SSU rRNA gene sequence has the GenBank Accession Number [OR427353](https://doi.org/10.26434/chemrxiv-2023-08).

Type locality: Coastal waters of Curacao

Etymology: Species epithet means ‘phantom’ in Aboriginal language of Curacao population.

Zoobank registration: LSID for this publication: urn:lsid:zoobank.org:pub:CBA7B765-7996-4F42-8AAF-42460CE7F77E. LSID for the

new species: urn:lsid:zoobank.org:act:1B42392D-9FE1-464F-B637-50761EFC7A08.

Eleftheros karadeniz n. sp. Tikhonenkov, Hehenberger, and Keeling

Diagnosis: Cells 4.2–5.5 µm long, roundish or wide-oval. Longitudinal flagellum strait, lies in the longitudinal groove and directed backward, 2 times longer than the cell. Transverse flagellum about 2.5 times as long as cell, undulates in the transverse groove.

Type figure: Fig. 1k.

Type material: The specimen shown in Fig. 1k is the holotype (see International Code of Zoological Nomenclature, Art. 72.5.6, Declaration 45).

Gene sequence: The SSU rRNA gene sequence has the GenBank Accession Number [OR427351](https://doi.org/10.26434/chemrxiv-2023-08).

Type locality: Near shore bottom sediments in the Black Sea.

Etymology: The species epithet means ‘Black Sea’ in Turkish and Crimean Tatar languages.

Zoobank registration: LSID for this publication: urn:lsid:zoobank.org:pub:CBA7B765-7996-4F42-8AAF-42460CE7F77E. LSID for the new species: urn:lsid:zoobank.org:act:6264A9FC-BBD1-4D8D-BOE6-EB05C540D628.

Methods

Data reporting

No statistical methods were used to predetermine the sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Cell isolation and culture establishment

Strains Colp-37 and Cur-11 (*Eleftheros xomoi* gen. et sp. nov.) were obtained from the surface of the brain coral *Colpophyllia natans* Houttuyn, 1772 in coastal waters of Curaçao (Caribbean Sea) in April 2016 and 2018, respectively. Strain Colp-25 (*Eleftheros karadeniz* gen. et sp. nov.) was isolated from the near shore bottom sediments in the Black Sea near T.I. Vyazemsky Karadag Scientific Station, Crimea, May 2015. The samples were enriched with a suspension of *Aeromonas sobria* bacteria and examined on the third, sixth, and ninth day of incubation (25 °C, darkness) in accordance with methods described previously⁵⁶. To obtain clonal cultures, the individual cells were transferred using a drawn-out glass micropipette into Petri dishes containing a clonal culture of the eukaryotic prey *Proccryptobia sorokini* (strain B-69), which were grown in marine Schmalz–Pratt’s medium at a final salinity of 20‰ using the bacterium *A. sobria* as food⁵⁷. Strains perished after several months of cultivation.

The MALV-I infected *Polykrikos* cell was collected with a 20 µm mesh net towed at Heriot Bay near the Hakai Quadra Island Ecological Observatory in British Columbia, Canada (October 2022). The MALV-I infected *Warnowia* cell and host-less MALV-I were collected from near-shore tows at Piscadera Bay, Curaçao (February 2023), using the same net. Cells were isolated with a drawn-out glass micropipette and washed with (0.2 µm) filtered water before imaging and lysis.

Psammusa pacifica Psp strain was isolated from Boundary Bay, British Columbia, Canada (the type location of the original description) and maintained according to Okamoto et al.³⁵. *Psammusa* sp. (strain Colp-34) was isolated from the near shore bottom sediments in the Kapsel Bay, Black Sea, Crimea (May 2016) and maintained following the abovementioned protocol.

Light and electron microscopy

To observe living cells, an AxioScope A1 light microscope (Carl Zeiss, Jena, Germany) with DIC water immersion objective 63× and an inverted microscope Leica DM IL LED with DIC objectives 40× and 63× were used. Images were captured with a Sony α7 R camera.

For transmission electron microscopy (TEM), cells were centrifuged and fixed at 1°C for 60 min in a cocktail of 0.6% glutaraldehyde and 2% OsO₄ (final concentration) prepared using a 0.1 M cacodylate buffer (pH 7.2). Fixed cells were dehydrated in alcohol and acetone series (30, 50, 70, 96, and 100%, 20 min in each step). Afterward, the cells were embedded in a mixture of Araldite and Epon (Fluka, 45345). Ultrathin sections (60 nm) were prepared with a Leica EM UC6 ultramicrotome (Leica Microsystems, Germany) and observed by using a JEM 1011 transmission electron microscope (JEOL, Japan).

Preparation of libraries and sequencing

RNA isolation and cDNA preparation. Cells from clonal culture were collected by centrifugation (1000 × g, room temperature) onto the 0.8 μm membrane of a Vivaclear mini column (Sartorius Stedim Biotech Gmng, VK01P042). Total RNA was then extracted using a RNeasy-Micro Kit (Invitrogen, AM1931) and reverse transcribed into cDNA using the Smart-Seq2 protocol⁵⁸, which uses poly-A selection to enrich mRNA. Additionally, cDNA of Cur-11 and Colp-37 were obtained from 20 single cells using the Smart-Seq2 protocol (cells were manually picked from culture using a drawn-out glass micropipette and transferred to a 0.2 mL thin-walled PCR tube containing 2 μL of cell lysis buffer –0.2% Triton X-100 and RNase inhibitor (Invitrogen)). Likewise, cDNA from the MALV-I associated cells and *Psammusa* spp. was prepared following the same protocol.

Sequencing dataset assembly and decontamination. The *Polykrikos-associated* MALV-I library was sequenced with an Illumina NextSeq 500 system using a Mid Output Flow Cell (2 × 150 bp reads), while the remaining two MALV-I associated cells were sequenced on the same system using a High Output Flow Cell (2 × 150 bp reads). *Eleftheros* and *Psammusa* libraries were sequenced on an Illumina MiSeq platform with read lengths of 2 × 300 bp (strains Colp-25, Colp-37, Psp, and Colp-34) and on an Illumina HiSeq 2500 machine, 2 × 125 bp reads (strain Cur-11). Sequence quality and adapter contamination of reads from transcriptomic datasets were assessed with FastQC v.0.10.1⁵⁹.

Reads of clone Colp-37 (culture and 20-cell preparations) and clone Colp-25 (culture) were merged with PEAR v.0.9.6⁶⁰ and resulting assembled as well as unassembled reads were separately trimmed with Trimmomatic⁶¹ as implemented in Trinity v.2.0.6⁶², removing Illumina adapters with ILLUMINACLIP, with a maximum of two mismatches, a palindrome clip threshold of 30 and a simple clip threshold of 10. Low-quality sequences were discarded, using a sliding window of 4 bp and a minimum trimmed length of 25 bp. Trimmed assembled and unassembled reads were combined into a single file and transcriptomes were assembled with Trinity, using the --single flag. Contaminating non-eukaryotic (bacterial, archaeal, and viral) and prey contigs were identified using BLASTn⁶³ queries of the NCBI nt database. Sequences that aligned with ≥100 nucleotides, had a query coverage of ≥80%, and were ≥90% identical to noneukaryotic entries were removed, while all sequences resulting in a kinetoplastid best hit were removed.

Reads of the later sampled clone Cur-11 (culture and 20-cell preparations) were trimmed and assembled without prior read merging in a single step, using the Trimmomatic plugin embedded in Trinity. Trimming parameters remained unchanged except for a simple clip threshold of 9. Due to results from initial analyses of clone Colp-37, the transcriptome was not subjected to any cleaning steps to retain putative bacterial horizontal gene transfers as well as sequences with bacterial best hits in nt that were identified as likely being eukaryotic in phylogenetic reconstructions in clone Colp-37. Instead, all genes of interest were subjected to phylogenetic analysis, as described below, to clarify the taxonomic identity of the gene.

Reads from the MALV-I-associated and psammosid libraries were trimmed using Cutadapt v3.2⁶⁴ before assembly with rnaSPAdes v3.15.1⁶⁵. Contaminating sequences were removed using BLASTx⁶⁶ and BLASTn⁶³ searches against the NCBI nt and UniProt databases (E-value

cut-off = 1 × 10⁻²⁵). Prey contigs, *Spumella elongata* for *Psammusa pacifica* and *Procryptobia sorokini* for *Psammusa* sp. were characterized using BLASTn⁶³ against prey transcriptome data and removed.

Prediction of protein coding regions in all assemblies was performed by TransDecoder v.5.0.2, including BLASTp⁶⁶ queries of the Swiss-Prot database (E-value cut-off = 1 × 10⁻⁵).

Phylogenomic dataset preparation and analysis

In addition to the taxa described above we added to or updated the following available transcriptome or genome data in an existing phylogenomic framework⁶⁷. We added the re-assembled transcriptomes of several core dinoflagellates generated by the MMETSP project⁶⁸, several transcriptomes from the EukProt database⁶⁹, the transcriptomes of the dinoflagellates *Lepidodinium chlorophorum* (<https://www.ncbi.nlm.nih.gov/bioproject/481676>), TGD and MGD⁷⁰, *Abedinium*⁷¹, *Amyloodinium ocellatum*⁷² and additional taxa from Cooney et al.⁷³, including *Spatulodinium*, *Kofooidinium* and *Fabadinium*. We updated the data for the MALVs *Amoebophrya ceratii*⁷⁴ and *Hematodinium* sp¹⁴., for *Oxyrrhis marina*⁶⁸, for the perkinsids *Perkinsus marinus* (https://protists.ensembl.org/Perkinsus_marinus_atcc_50983_gca_000006405/Info/Index), *Maranthos nigrum* and *Parvilucifera sinerae*⁴⁸, the apicomplexa *Ancora sagittate*⁵², and for the chromerids *Vitrella brassicaformis* (https://cryptodb.org/cryptodb/app/record/dataset/NCBITAXON_1169540) and Symbiont X⁵². Transcriptomic datasets were subjected to TransDecoder coding region prediction to extract peptides for downstream analyses. Peptide sequences from predicted MALV SAGs described in Delmont et al.²⁸ were also included.

All listed datasets were searched for 263 proteins to generate single-protein trees as described in Burki et al.⁶⁷. In brief: BLASTp was used to identify homologs to the 263 genes in the new datasets. After a parsing step (E-value ≤ 1 × 10⁻²⁰), a maximum of four non-redundant hits was added to the initial 263-protein set. The expanded gene sets were aligned with MAFFT L-ins-i v.7.222⁷⁵ and trimmed automatically with trimAl v.1.2⁷⁶, with a gap threshold of 80%. Single-protein ML phylogenies were reconstructed under the LG + G4 model using RAxML v.8.1.6⁷⁷ in combination with 100 rapid bootstraps and resulting trees were manually screened to flag paralogues and sequences derived from prey or other contamination. Cleaned protein sets were aligned and trimmed as above and taxa were selected upon concatenation with SCAFOs v.1.2.5⁷⁸ to select proteins sequences present in ≥60% of all taxa. To improve data presence, the concatenated sequences of clone Colp-37 derived from culture and 20-cell preparations were merged, as were the sequences for the two clone Cur-11 datasets. The final concatenated alignment included 75 taxa, 236 proteins and 63,267 amino acid sites. Clone Cur-11 was represented with 92% of proteins and 86% of sites, clone Colp-37 with 82% of proteins and 72% of sites and clone Colp-25 with 49% of proteins and 38% of sites. The host *Polykrikos* sp. was represented with 75% of proteins and 66% of sites and *Ichthyodinida* sp. 1 with 25% and 15% respectively. The host *Warnowia* sp. was represented with 61% of proteins and 39% of sites, and *Ichthyodinida* sp. 2 with 28% and 17%, respectively. *Ichthyodinida* sp. 3 was represented with 15% of proteins and 8% of sites *Psammusa pacifica* Psp was represented with 76% and 63% of sites, *Psammusa* sp. was represented with 75% of proteins and 65% of sites.

ML phylogenomic tree reconstruction was performed using IQ-TREE v. 1.6.5⁷⁹ using the C60 empirical mixture model in combination with the LG matrix, amino acid frequencies computed from the data, and four gamma categories for handling the rate heterogeneity across sites (LG + C60 + F + G4 model with 1000 UFBoot replicates⁸⁰). Phylogenetic trees were visualized in R⁸¹ with the ggtree⁸² and treeio⁸³ packages.

Fast-evolving sites were estimated with IQ-TREE using the -wsr option, using the LG + C60 + F + G4 model. Sites were removed in increments of 5% of the original alignment length, up to 50%, and for

each subsample with a reduced number of sites, trees were reconstructed with the same model. The support for the sister relationship of eleutherids and MALVs was assessed at each removal increment. The monophyly of core dinoflagellates was tested as a control.

Small subunit ribosomal RNA (SSU rRNA) phylogeny

The most complete sequences for SSU rRNA genes were identified from each eletherid transcriptome with BLASTn, using a known dinoflagellate SSU query, whereas barnnap⁸⁴ was used to extract MALV-I SSU rRNA sequences prior to identification with BLAST. An alignment of all near full-length MALV sequences deposited in GenBank was curated to include all five previously described MALV groups in Guillou et al.⁷. The resulting SSU collection was aligned using MAFFT with the E-INS-I algorithm and inspected for misaligned and chimeric sequences. Shorter sequences obtained from surveys of environmental data (see below) were aligned using the addfragments flag. The final alignment was trimmed with trimAl (-gt 0.1, -st 0.001) before generating a ML tree in IQ-TREE with 1000 ultrafast bootstrap replicates⁸⁰, using the GTR + F + R7 model (and GTR + R + R6 for Supplementary Fig. 3), selected with ModelFinder⁸⁵. *Oxyrrhis marina* and *Ellobiopsis chattonii* were omitted to minimize long branch attraction.

Identification of putative plastid-targeted proteins

Putative plastid-targeted proteins were identified combining a BLASTp-based similarity search and a hidden Markov models (HMMs) based screen. Known dinoflagellate³⁹ or MALV¹⁴ proteins involved in plastid metabolic pathways were used as queries in a BLASTp search against a comprehensive custom database containing representatives from most major eukaryotic groups (excluding the long-branching excavates and the data-poor group of Rhizaria) and RefSeq data from all bacterial phyla at NCBI (last accessed December 2017). The database was subjected to CD-HIT⁸⁶ clustering with a similarity threshold of 85% to reduce redundant sequences and paralogues, except for the data sets created in this study (clustered at 98%). The search results of the BLASTp step were parsed for hits with an E-value threshold $\leq 1 \times 10^{-25}$ and a query coverage of $\geq 30\%$ to reduce the possibility of paralogs and extremely short sequences and at the same time recover possibly fragmented homologs. The number of bacterial hits was restrained to 20 hits per phylum (for FCB group, most classes of Proteobacteria, PVC group, Spirochetes, Actinobacteria, Cyanobacteria, and Firmicutes) or 10 per phylum (remaining bacterial phyla) as defined by NCBI taxonomy. In some cases (HemD and DapL) these numbers were expanded to 20 and 40, respectively, for a more representative bacterial sampling. In addition, protein data were used to search the Pfam-A database release 33.0 with hmmscan (HMMER3.1; hmmer.org), employing the manually curated Pfam gathering threshold. The results were queried, using a keyword search, for Pfam domains present in plastid-associated proteins, including proteins involved in metabolic pathways in Hehenberger et al.³⁹, photosynthesis, plastid import (TIC/TOC components) and Calvin cycle (RuBisCO). Candidates with domains of interest were used as BLASTp queries as described above, and parsed hits (query coverage of $\geq 50\%$) were combined with the recovered hits from the known proteins. After a deduplication step, sequences were aligned with MAFFT using the -auto option, trimmed using trimAl (-gt 0.8) and ML tree reconstructions were performed with FastTree v.2.1.7⁸⁷ using the default options in a preliminary analysis. The resulting phylogenies and underlying alignments were inspected manually to remove contaminant, divergent, and/or low-quality sequences. The cleaned, unaligned sequences were then subjected to filtering with PREQUAL⁸⁸ using the default options, followed by alignment with MAFFT G-INS-I using the VSM option (-unalignedlevel 0.6). The alignments were subjected to Divvier⁸⁹ using the -mincol 4 and the -divvygap option before trimming with trimAl (-gt 0.01). Final trees were calculated with IQ-TREE, using the -mset option to restrict

model selection to LG for ModelFinder⁸⁵, while branch support was assessed with 1000 ultrafast bootstrap replicates⁸⁰.

We also searched for homologs to the 129 proteins predicted to localize to the apicoplast in *Toxoplasma*⁴². The predicted *Toxoplasma* proteins were used as queries in a BLASTp search against our custom database, using an initial E-value threshold of $\leq 1 \times 10^{-25}$ and a query coverage of $\geq 30\%$ for parsing. All proteins recovering putative homologs were used in a preliminary tree reconstruction analysis as described above. After manual inspection of the phylogenies, potential plastid-targeted candidates were further investigated by using the recovered homologs as additional BLASTp queries and combining the resulting hits (query coverage of $\geq 50\%$) with the initial BLASTp output as described above. The BLASTp search using the *Toxoplasma* proteins was repeated with relaxed parameters (E-value threshold $\leq 1 \times 10^{-5}$) to recover additional candidates. All candidates investigated were also submitted to a web BLASTp search against nr to exclude the possibility of contamination not present in our database.

N-terminal extensions of putative plastid-targeted sequences were investigated in the respective protein alignments, visualized with AliView⁹⁰. For easier recognition of such extensions, only the dinoflagellate sequences of the protein of interest plus the prokaryotic sequence with the highest sequence similarity were viewed. Sequences with N-terminal extensions relative to prokaryotic sequences were submitted to SignalP-3.0⁹¹ and TMHMM 2.0⁹² to predict putative signal peptides and transmembrane domains, respectively.

For proteins with known bacterial orthologs, bacterial naming conventions were applied, with the exception of the multifunctional enzyme Acetyl-CoA carboxylase (ACC) not present in bacteria and the enzyme 5-aminolevulinic synthase (ALAS), where eukaryotic conventions were applied.

Identification of molecular characteristics of dinoflagellates

Dinoflagellate Viral NucleoProteins (DVNPs) were identified using a BLASTp search against our custom database using all described *Hematodinium* DVNP proteins⁴⁵ as well as by performing an hmmsearch (HMMER3.1; hmmer.org) against our database using the DVNP profile HMM downloaded from pfam.xfam.org, employing default thresholds. The two approaches recovered the same set of eletherid DVNP candidates. No DVNP sequences were detected in psammosids or MALV-I. Only host DVNP sequences could be found in the *Polykrikos*/Ichthyodinida sp. 1 library.

We also searched for histone-like proteins (HLPs) using known representatives of the two known types of HLPs in dinoflagellates³⁶ as query, *Cryptocodinium cohnii* AAM97522.1 (HLP1; <https://www.ncbi.nlm.nih.gov/protein/AAM97522.1>) and *Noctiluca scintillans* ABV22345.1 (HLP2; <https://www.ncbi.nlm.nih.gov/protein/ABV22345.1>) in a BLASTp search against the predicted peptides (E-value threshold $\leq 1 \times 10^{-25}$). Additionally, we performed an hmmsearch using a profile HMM constructed from a curated alignment of dinoflagellate and prokaryotic HLPs.

Dinoflagellate spliced leader (SL) sequences were identified using BLASTn searches (using the option -task blastn to allow for short input queries and to identify short matches) against all transcriptomes, using the canonical 21-nucleotide dinoflagellate SL sequence⁴⁷ as a query. The recovered 21-nucleotide sequence in eletherids, differing in 2 nucleotides (5 A > T and 17 T > A) from the canonical sequence, was used as a search string to identify all eletherids transcripts containing a full-length SL. To avoid SL sequences in host sequences, MALV-I peptides identified in single-gene trees were aligned against host/MALV-I transcripts with tBLASTn to obtain MALV-I only transcripts.

Environmental distribution and abundance

To uncover the global distribution and abundance of eletherids, we searched for sequences similar to the eletherid SSU rRNA gene among published environmental SSU rRNA gene amplicon studies.

Full-length eleutherid SSU rRNA genes and their V4 and V9 hypervariable regions were used as queries in BLASTn searches (E-value threshold $\leq 1 \times 10^{-10}$) against the complete Tara Ocean database available on Ocean Gene Atlas¹⁸. We recovered no sequences with $\geq 95\%$ identity and a query cover of $\geq 90\%$, neither when using full-length nor hypervariable regions of eleutherid SSU rRNA gene queries.

Additionally, we searched a custom environmental sequence database containing data from 26 environmental amplicon studies with a focus on marine sediment, also including the large sequencing projects BioMarKs⁴ and Malaspina¹⁹. NCBI SRA files of amplicon data were downloaded using fastq-dump from sratoolkit 2.10.8. Raw sequence data were processed with MICrobial Community Analysis, using micca merge or mergepairs (-l 100 -d 30) for single-end or paired-end data respectively⁹³, and then concatenated and converted them into a BLAST database composed of 113,203,549 sequences with a total of 25,261,490,053 bases. Alternatively, paired-end data were processed in Qiime2⁹⁴ with the DADA2⁹⁵ denoising algorithm before creation of the BLAST database. This database was searched using full-length eleutherid SSU rRNA gene sequences as well as their V4 and V9 hypervariable regions in BLASTn searches (E-value threshold $\leq 1 \times 10^{-25}$). In addition to the environmental database, sequences similar to the eleutherid 18 S rRNA gene were searched using BLASTn against a V9 amplicon dataset collected from sandy beaches off the central coast of British Columbia (the European Nucleotide Archive project [PRJEB14727](https://doi.org/10.1038/s41467-023-42807-0)). A total of 73 unique sequences with $\geq 97\%$ identity to and $\geq 90\%$ coverage of the eleutherid V9 and V4 hypervariable regions resulted from both search approaches. Full-length and hypervariable regions V9 and V4 of *P. pacifica* were used as queries against the environmental database which resulted in a total of 4 unique sequences with $\geq 97\%$ identity to and $\geq 90\%$ coverage to psammosids. These sequences were then clustered at 99% identity using CD-HIT prior to being added to the SSU rRNA gene phylogeny to confirm their phylogenetic position as described above. Sequences less than 150 bp were removed from the final phylogeny. Sample coordinates corresponding to BLAST hits in the custom database were extracted and plotted in R⁸¹ with the rnaturalearth⁹⁶ package.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Raw transcriptome reads have been deposited in the GenBank Sequence Read Archive (SRA) database under the accessions SRR25604407–SRR25604420. SSU rRNA gene sequences retrieved from the transcriptomes have also been deposited in GenBank under the accessions OR427349–OR427355. All sequence data is linked to the BioProject accession code [PRJNA1003956](https://doi.org/10.1038/s41467-023-42807-0). Assembled transcriptomes, along with individual gene alignments, concatenated and trimmed alignments, and ML tree files for the phylogenomic dataset are available at Figshare (<https://doi.org/10.6084/m9.figshare.24293662>). The untrimmed and trimmed alignments, alignments depicting N-terminal extensions and tree files in nexus and pdf format for plastid-associated and other proteins of interest are available at Figshare (<https://doi.org/10.6084/m9.figshare.19351700>). The genus *Eleftheros* (urn:lsid:zoobank.org:act:DD5CBE62-6058-4ADA-BDBD-7B3A486447E1) and species *Eleftheros xomoi* (urn:lsid:zoobank.org:act:1B42392D-9FE1-464F-B637-50761EFC7A08) and *Eleftheros karadeniz* (urn:lsid:zoobank.org:act:6264A9FC-BBD1-4D8D-B0E6-EB05C540D628) have been registered with the Zoobank database (<http://zoobank.org/>). Zoobank Registration: LSID for this publication: urn:lsid:zoobank.org:pub:CBA7B765-7996-4F42-8AAF-42460CE7F77E.

Code availability

All unpublished code is available upon request from the corresponding authors.

References

1. Moon-van der Staay, S. Y., De Wachter, R. & Vaulot, D. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**, 607–610 (2001).
2. López-García, P., Rodríguez-Valera, F., Pedrós-Alió, C. & Moreira, D. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**, 603–607 (2001).
3. de Vargas, C. et al. Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).
4. Massana, R. et al. Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing: Protist diversity in European coastal areas. *Environ. Microbiol.* **17**, 4035–4049 (2015).
5. Clarke, L. J., Bestley, S., Bissett, A. & Deagle, B. E. A globally distributed Syndiniales parasite dominates the Southern Ocean micro-eukaryote community near the sea-ice edge. *ISME J.* **13**, 734–737 (2019).
6. Holt, C. C. et al. Microscopic marine invertebrates are reservoirs for cryptic and diverse protists and fungi. *Microbiome* **10**, 161 (2022).
7. Guillou, L. et al. Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). *Environ. Microbiol.* **10**, 3349–3365 (2008).
8. Rizos, I. et al. Beyond the limits of the unassigned protist microbiome: inferring large-scale spatio-temporal patterns of Syndiniales marine parasites. *ISME Commun.* **3**, 16 (2023).
9. Skovgaard, A., Massana, R., Balagué, V. & Saiz, E. Phylogenetic position of the copepod-infesting parasite *Syndinium turbo* (Dinoflagellata, Syndinea). *Protist* **156**, 413–423 (2005).
10. Gestal, C., Novoa, B., Posada, D., Figueras, A. & Azevedo, C. *Perkinsoidae chabelardi* n. gen., a protozoan parasite with an intermediate evolutionary position: possible cause of the decrease of sardine fisheries? *Environ. Microbiol.* **8**, 1105–1114 (2006).
11. Harada, A., Ohtsuka, S. & Horiguchi, T. Species of the parasitic genus *Duboscquella* are members of the enigmatic Marine Alveolate Group I. *Protist* **158**, 337–347 (2007).
12. Strasser, J. F. H. et al. Single cell genomics of uncultured marine alveolates shows paraphyly of basal dinoflagellates. *ISME J.* **12**, 304–308 (2018).
13. Jackson, C. J., Gornik, S. G. & Waller, R. F. The Mitochondrial genome and transcriptome of the basal dinoflagellate *Hematodinium* sp.: character evolution within the highly derived mitochondrial genomes of dinoflagellates. *Genome Biol. Evol.* **4**, 59–72 (2012).
14. Gornik, S. G. et al. Endosymbiosis undone by stepwise elimination of the plastid in a parasitic dinoflagellate. *Proc. Natl Acad. Sci. USA* **112**, 5767–5772 (2015).
15. Farhat, S. et al. Rapid protein evolution, organellar reductions, and invasive intronic elements in the marine aerobic parasite dinoflagellate *Amoebophrya* spp. *BMC Biol.* **19**, 1 (2021).
16. Xu, P. et al. The genome of *Cryptosporidium hominis*. *Nature* **431**, 1107–1112 (2004).
17. Waller, R. F. & Kořený, L. Plastid complexity in dinoflagellates: a picture of gains, losses, replacements and revisions. *Adv. Bot. Res.* **84**, 105–143 (2017).
18. Villar, E. et al. The Ocean Gene Atlas: exploring the biogeography of plankton genes online. *Nucleic Acids Res.* **46**, W289–W295 (2018).
19. Duarte, C. M. Seafaring in the 21st century: The Malaspina 2010 circumnavigation expedition. *Limnol. Oceanogr. Bull.* **24**, 11–14 (2015).
20. Cordier, T. et al. Patterns of eukaryotic diversity from the surface to the deep-ocean sediment. *Sci. Adv.* **8**, eabj9309 (2022).

21. Yoo, J., Coats, D. W. & Kim, S. Syndinean dinoflagellates of the genus *Euduboscquella* are paraphyletic. *J. Eukaryot. Microbiol.* **70**, e12953 (2023).
22. Cachon, J. Contribution à l'étude des péridiniens parasites. Cytologie, cycles évolutifs. *Annales des. Sci. Naturelles, Zoologie, Paris, Série 12*, 1–158 (1964).
23. Coats, D. W. & Bachvaroff, T. R. Parasites of Tintinnids. In *The biology and ecology of tintinnid ciliates: models for marine plankton* (eds. Dolan, J. R., Montagnes, D. J. S., Agatha, S., Coats, D. W. & Stoecker, D. K.), 145–170 (Wiley, 2012).
24. Jung, J.-H., Choi, J. M., Coats, D. W. & Kim, Y.-O. *Euduboscquella costata* n. sp. (Dinoflagellata, Syndinea), an intracellular parasite of the ciliate *Schmidingerella arcuata*: morphology, molecular phylogeny, life cycle, prevalence, and infection intensity. *J. Eukaryot. Microbiol.* **63**, 3–15 (2016).
25. Choi, J. M., Jung, J. H., Kim, K. H., Coats, D. W. & Kim, Y. O. A novel parasitic, syndinean dinoflagellate *Euduboscquella triangula* infecting the tintinnid *Helicostomella longa*. *Front Mar. Sci.* **8**, 720424 (2021).
26. Kim, S. Patterns in host range for two strains of *Amoebophrya* (dinophyta) infecting thecate dinoflagellates: *Amoebophrya* spp. ex *Alexandrium affine* and ex *Gonyaulax polygramma*. *J. Phycol.* **42**, 1170–1173 (2006).
27. Park, M. G., Kim, S., Shin, E.-Y., Yih, W. & Coats, D. W. Parasitism of harmful dinoflagellates in Korean coastal waters. *Harmful Algae* **30**, S62–S74 (2013).
28. Delmont, T. O. et al. Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genom.* **2**, 100123 (2022).
29. Cavalier-Smith, T. Kingdom Chromista and its eight phyla: a new synthesis emphasising periplastid protein targeting, cytoskeletal and periplastid evolution, and ancient divergences. *Protoplasma* **255**, 297–357 (2018).
30. Gómez, F. Diversity and classification of dinoflagellates. *Dinoflagellates. Classification, Evolution, Physiology, and Ecological Significance*, 1–38 (Nova Science Publishers, 2020).
31. Cachon, J. & Cachon, M. Parasitic dinoflagellates. In *The Biology of Dinoflagellates* (ed. Taylor, F. J. R.), 571–610 (Blackwell Scientific Publications, 1987).
32. Bodyt, A., Stiller, J. W. & Mackiewicz, P. Chromalveolate plastids: direct descent or multiple endosymbioses? *Trends Ecol. Evol.* **24**, 119–121 (2009).
33. Ševčíková, T. et al. Updating algal evolutionary relationships through plastid genome sequencing: did alveolate plastids emerge through endosymbiosis of an ochrophyte? *Sci. Rep.* **5**, 10134 (2015).
34. Mathur, V. et al. Reconstruction of plastid proteomes of apicomplexans and close relatives reveals the major evolutionary outcomes of cryptic plastids. *Mol. Biol. Evol.* **40**, msad002 (2023).
35. Okamoto, N., Horák, A. & Keeling, P. J. Description of two species of early branching dinoflagellates, *Psammosa pacifica* n. g., n. sp. and *P. atlantica* n. sp. *PLoS ONE* **7**, e34900 (2012).
36. Janouškovec, J. et al. Major transitions in dinoflagellate evolution unveiled by phylotranscriptomics. *Proc. Natl Acad. Sci. USA* **114**, E171–E180 (2017).
37. Sanchez-Puerta, M. V., Lippmeier, J. C., Apt, K. E. & Delwiche, C. F. Plastid genes in a non-photosynthetic dinoflagellate. *Protist* **158**, 105–117 (2007).
38. Patron, N. J., Waller, R. F., Archibald, J. M. & Keeling, P. J. Complex protein targeting to dinoflagellate plastids. *J. Mol. Biol.* **348**, 1015–1024 (2005).
39. Hehenberger, E., Gast, R. J. & Keeling, P. J. A kleptoplastidic dinoflagellate and the tipping point between transient and fully integrated plastid endosymbiosis. *Proc. Natl Acad. Sci. USA* **116**, 17934–17942 (2019).
40. Hehenberger, E., Imanian, B., Burki, F. & Keeling, P. J. Evidence for the retention of two evolutionary distinct plastids in dinoflagellates with diatom endosymbionts. *Genome Biol. Evol.* **6**, 2321–2334 (2014).
41. Dorrell, R. G. et al. Progressive and biased divergent evolution underpins the origin and diversification of peridinin dinoflagellate plastids. *Mol. Biol. Evol.* **34**, 361–379 (2017).
42. Barylyuk, K. et al. A comprehensive subcellular atlas of the *Toxoplasma* proteome via hyperLOPIT provides spatial context for protein functions. *Cell Host Microbe* **28**, 752–766.e9 (2020).
43. Gavelis, G. S. et al. Microbial arms race: Ballistic “nematocysts” in dinoflagellates represent a new extreme in organelle complexity. *Sci. Adv.* **3**, e1602552 (2017).
44. Hackett, J. D., Anderson, D. M., Erdner, D. L. & Bhattacharya, D. Dinoflagellates: a remarkable evolutionary experiment. *Am. J. Bot.* **91**, 1523–1534 (2004).
45. Gornik, S. G. et al. Loss of nucleosomal DNA condensation coincides with appearance of a novel nuclear protein in dinoflagellates. *Curr. Biol.* **22**, 2303–2312 (2012).
46. Irwin, N. A. T. et al. Viral proteins as a potential driver of histone depletion in dinoflagellates. *Nat. Commun.* **9**, 1535 (2018).
47. Zhang, H. et al. Spliced leader RNA trans-splicing in dinoflagellates. *Proc. Natl Acad. Sci. USA* **104**, 4618–4623 (2007).
48. Alacid, E. et al. A diversified and segregated mRNA spliced-leader system in the parasitic Perkinsozoa. *Open Biol.* **12**, 220126 (2022).
49. Miller, J. J., Delwiche, C. F. & Coats, D. W. Ultrastructure of *Amoebophrya* sp. and its changes during the course of infection. *Protist* **163**, 720–745 (2012).
50. Okamoto, N. & Keeling, P. J. The 3D structure of the apical complex and association with the flagellar apparatus revealed by serial TEM tomography in *Psammosa pacifica*, a distant relative of the Apicomplexa. *PLoS ONE* **9**, e84653 (2014).
51. Leander, B. S. & Keeling, P. J. Morphostasis in alveolate evolution. *Trends Ecol. Evol.* **18**, 395–402 (2003).
52. Janouškovec, J. et al. Apicomplexan-like parasites are polyphyletic and widely but selectively dependent on cryptic plastid organelles. *eLife* **8**, e49662 (2019).
53. Mathur, V. et al. Multiple independent origins of apicomplexan-like parasites. *Curr. Biol.* **29**, 2936–2941.e5 (2019).
54. Janouškovec, J., Horák, A., Oborník, M., Lukeš, J. & Keeling, P. J. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proc. Natl Acad. Sci. USA* **107**, 10949–10954 (2010).
55. Schön, M. E. et al. Single cell genomics reveals plastid-lacking Picozoa are close relatives of red algae. *Nat. Commun.* **12**, 6651 (2021).
56. Tikhonenkov, D. V., Mazej, Y. A. & Embulaeva, E. A. Degradation succession of heterotrophic flagellate communities in microcosms. *Zh. Obs Biol.* **69**, 57–64 (2008).
57. Tikhonenkov, D. V. et al. Description of *Colponema vietnamica* sp.n. and *Acavomonas peruviana* n. gen. n. sp., two new alveolate phyla (Colponemidia nom. nov. and Acavomonidia nom. nov.) and their contributions to reconstructing the ancestral state of alveolates and eukaryotes. *PLoS One* **9**, e95467 (2014).
58. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
59. Andrews, S. FastQC: A quality control tool for high throughput sequence data. version 0.10.1. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2010).
60. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
61. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

62. Grabherr, M. G. et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* **29**, 644–652 (2013).
63. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J. Comput Biol.* **7**, 203–214 (2000).
64. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
65. Bankevich, A. et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput Biol.* **19**, 455–477 (2012).
66. Camacho, C. et al. BLAST+: Architecture and applications. *BMC Bioinforma.* **10**, 1–9 (2009).
67. Burki, F. et al. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of centrohelida, haptophyta and cryptista. *Proc. R. Soc. B Biol. Sci.* **283**, 20152802 (2016).
68. Johnson, L. K., Alexander, H. & Brown, C. T. Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. *Gigascience* **8**, 1–12 (2019).
69. Richter, D. J. et al. EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. *Peer Community J.* **2**, e56 (2022).
70. Sarai, C. et al. Dinoflagellates with relic endosymbiont nuclei as models for elucidating organellogenesis. *Proc. Natl Acad. Sci. USA* **117**, 5364–5375 (2020).
71. Cooney, E. C. et al. Single-cell transcriptomics of *Abedinium* reveals a new earlybranching dinoflagellate lineage. *Genome Biol. Evol.* **12**, 2417–2428 (2020).
72. Byadgi, O. et al. Transcriptome analysis of *Amyloodinium ocellatum* tomtomts revealed basic information on the major potential virulence factors. *Genes* **11**, 1252 (2020).
73. Cooney, E. C., Leander, B. S. & Keeling, P. J. Phylogenomics shows unique traits in Noctilucales are derived rather than ancestral. *PNAS Nexus* **1**, pgac202 (2022).
74. John, U. et al. An aerobic eukaryotic parasite with functional mitochondria that likely lacks a mitochondrial genome. *Sci. Adv.* **5**, 1–12 (2019).
75. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
76. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
77. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
78. Roure, B., Rodríguez-Ezpeleta, N. & Philippe, H. SCaFoS: A tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol. Biol.* **7**, 1–12 (2007).
79. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
80. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
81. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/> (2021).
82. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
83. Wang, L.-G. et al. Treeio: An R package for phylogenetic tree input and output with richly annotated and associated data. *Mol. Biol. Evol.* **37**, 599–603 (2020).
84. Seemann T. barrnap 0.9: rapid ribosomal RNA prediction. <https://github.com/tseemann/barrnap>
85. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
86. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the nextgeneration sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
87. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
88. Whelan, S., Irisarri, I. & Burki, F. PREQUAL: Detecting non-homologous characters in sets of unaligned homologous sequences. *Bioinformatics* **34**, 3929–3930 (2018).
89. Ali, R. H., Bogusz, M., Whelan, S. & Tamura, K. Identifying clusters of high confidence homologies in multiple sequence alignments. *Mol. Biol. Evol.* **36**, 2340–2351 (2019).
90. >Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).
91. Dyrlov Bendtsen, J., Nielsen, H., Von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795 (2004).
92. Krogh, A., Larsson, B., Von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. Edited by F. Cohen. *J. Mol. Biol.* **305**, 567–580 (2001).
93. Albanese, D., Fontana, P., De Filippo, C., Cavalieri, D. & Donati, C. MICCA: A complete and accurate software for taxonomic profiling of metagenomic data. *Sci. Rep.* **5**, 1–7 (2015).
94. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
95. Callahan, B. J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
96. Massicotte P, South A. *rnaturalearth: World map data from natural earth*. R package version 0.3.2.9000, <https://docs.ropensci.org/rnaturalearth/><https://github.com/ropensci/rnaturalearth> (2023).

Acknowledgements

We thank A.P. Mylnikov and T.G. Simdyanov for help with sample collection, fixation, and interpretation of transmission electron microscopy images. This research was supported by grants from the Hakai Institute (P.J.K.), the Gordon and Betty Moore Foundation (<https://doi.org/10.37807/GBMF9201>, P.J.K.), the Natural Sciences and Engineering Research Council of Canada (Grant Number 2019-03994, P.J.K.), the Czech Academy of Sciences (Grant Number LQ200962204, E.H.) the Russian Foundation for Basic Research (Grant Number 20-04-00583, D.V.T.), Tyumen Oblast Government, as part of the West-Siberian Inter-regional Science and Education Center’s project No. 89-DON (2) and carried out within the framework of State Assignment no. 121051100102-2.

Author contributions

C.C.H., E.H., D.V.T., and P.J.K. designed the study. D.V.T. is isolated and cultured eleutherid cells. C.C.H. and E.C.C. isolated MALV-I. V.K.L.J.-R and N.O. isolated and cultured psammosid cells. C.C.H., E.H., D.V.T., V.K.L.J.-R., and E.C.C. generated material for sequencing. D.V.T. performed microscopy experiments. C.C.H., E.H., E.C.C., and N.A.T.I. performed phylogenomic analyses. C.C.H. and E.C.C. performed

phylogenetic analysis of the SSU rRNA genes. C.C.H. analyzed published SAG data. V.K.L.J.-R. performed the environmental distribution and abundance analysis. C.C.H., E.H., and V.K.L.J.-R. performed transcriptomic analyses and phylogenetic analysis of plastid and other proteins. C.C.H., E.H., and P.J.K. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-023-42807-0>.

Correspondence and requests for materials should be addressed to Corey C. Holt, Elisabeth Hehenberger or Patrick J. Keeling.

Peer review information *Nature Communications* thanks Marek Eliáš for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023