Nucleomorph phylogenomics suggests a deep and ancient origin of cryptophyte plastids within Rhodophyta

Lukas V. F. Novak[1,2]*, Sergio A. Muñoz-Gómez[1,3], Fabian van Beveren[1,4], Maria Ciobanu[1], Laura Eme[1], Purificación López-García[1], David Moreira[1]*

1) Unité d'Ecologie Systématique et Evolution, CNRS, AgroParisTech, Université Paris-Saclay, Gif-sur-Yvette, France.

2) Biologie et Ecologie des Ecosystèmes Marins Profonds (BEEP), CNRS, Université de Bretagne Occidentale, Institut Universitaire Européen de la Mer, Plouzané, France.

3) Department of Biological Sciences, Purdue University, IN, USA.

4) Laboratoire de Recherche en Sciences Végétales (LRSV), Université de Toulouse, CNRS, Castanet Tolosan, France.

* lukas.novak@univ-brest.fr (LVFN), david.moreira@universite-paris-saclay.fr (DM)

Abstract

The evolutionary origin of red algae-derived complex plastids remains obscure. Cryptophyta, one of four eukaryotic lineages harboring these plastids, still contain nucleomorphs, highly reduced remnants of red algal nuclei. The genes present on nucleomorph genomes can be used for phylogenomic reconstruction in order to unravel the evolutionary origin of all red complex plastids and provide data independent from previously analyzed plastid-encoded datasets. Here, we leverage these genes in a first attempt at pinpointing the position of cryptophyte nucleomorphs within a comprehensive diversity of Rhodophyta, including new sequence representatives from seven deep-branching red algae. Our analyses place cryptophyte nucleomorphs as sister to the extremophilic subphylum Cyanidiophytina. This conflicts with published analyses based on plastidial genes that placed red complex plastids closer to the mesophilic Rhodophytina. Our results reject a nucleomorph origin within any known subgroup of Rhodophyta, instead suggesting a deep and possibly freshwater origin of complex red plastids.

Introduction

The origin of complex plastids of red algal (Rhodophyta) ancestry found in photosynthetic members of Cryptista, Alveolata, Stramenopiles, and Haptista ("CASH") remains one of the great unresolved questions of eukaryotic evolution. Two competing hypotheses have tried to explain the presence of these organelles of apparently shared origin in these four not-so-closely-related lineages of hosts (Yoon et al. 2002; Burki et al. 2020). On the one hand, the chromalveolate hypothesis (Cavalier-Smith 1999) posits a single origin of these plastids from a single secondary endosymbiosis event in a common ancestor of CASH, and its subsequent vertical inheritance accompanied by multiple independent plastid losses in non-photosynthetic CASH species. On the other hand, the serial endosymbiosis hypothesis (Sanchez-Puerta and Delwiche 2008) postulates that secondary endosymbiosis occurred more recently in a single CASH lineage, after the four lineages had already diverged, and spread to the rest horizontally *via* tertiary, or even higher-degree, endosymbioses. This serial endosymbiosis hypothesis has been proposed in multiple versions positing different numbers and order of symbiotic events (Sanchez-Puerta and Delwiche 2008; Petersen et al. 2014; Stiller et al. 2014; Bodył 2018).

Regardless of which particular scenario is true, the identity of the red algal donor of complex plastids requires further investigation. Phylogenomic analyses of plastid-encoded genes (Yoon et al. 2002; Iida et al. 2007; Janouškovec et al. 2010; Stiller et al. 2014; Ševčíková et al. 2015; Kim et al. 2017) have consistently placed the CASH complex plastids in the vicinity of the Rhodophytina, the mesophilic branch of Rhodophyta, as opposed to the ancestrally extremophilic, freshwater Cyanidiophytina (Yoon et al. 2017). However, our insufficient understanding of phylogenetic relationships between red algal groups, as well as poor taxon sampling of rhodophyte genomes, have for a long time limited the usefulness of such studies

in precisely pinpointing the donor taxon. Recent large-scale phylogenomic analyses of plastid- and mitochondrion-encoded genes (Muñoz-Gómez et al. 2017; van Beveren et al. 2022) have greatly advanced our understanding of the internal phylogeny of the phylum Rhodophyta. These analyses support the basic division between Cyanidiophytina (class Cyanidiophyceae) and Rhodophytina and split the latter into two monophyletic subphyla: Proteorhodophytina (classes Compsopogonophyceae, Porphyridiophyceae, Rhodellophyceae, and Stylonematophyceae), comprising red algae with single-celled or simple multicellular body plans, and Eurhodophytina (classes Bangiophyceae and Florideophyceae), with complex multicellularity.

To attempt to phylogenetically place CASH plastids within the tree of Rhodophyta, we used an alternative source of genetic data – the nucleomorph genomes of Cryptophyta (Cryptista). Nucleomorphs (NMs) are highly reduced nuclei of archaeplastid algae that reside in the periplastidial space of certain complex plastids. There are at least three independent examples of NMs derived from green algae (Ludwig and Gibbs 1989; Sarai et al. 2020), but only one derived from red algae (Ludwig and Gibbs 1987; Douglas et al. 2001). These organelles are relatively well studied; phylogenetic analyses of their rRNAs have suggested a rhodophyte origin (Douglas et al. 1991; Maier et al. 1991), and they have even been included in exploration of rhodophyte genomic diversity (Wong et al. 2023). However, to our knowledge, there has not been any published phylogenomic analysis attempting to locate the position of cryptophyte NMs within Rhodophyta. Cryptophyte NMs contain several hundred open reading frames (ORFs) that can be divided into three categories: conserved ORFs with annotated homologues outside cryptophyte NMs, “nORFs” with annotated homologues only in other cryptophyte NMs, and “ORFans” without recognizable sequence similarity to any other known sequence (Moore et al. 2012). Only conserved ORFs are therefore useful for phylogenomic analyses aiming at placing NMs in a broader eukaryotic context. With this

goal in mind, we have improved the taxon sampling of red algae by sequencing seven novel nuclear genomes and performed a series of phylogenomic investigations into the position of cryptophyte NMs among Rhodophyta using a large dataset of 180 conserved ORFs.

Results and Discussion

To improve the taxon sampling of nucleus-encoded protein sequences of Rhodophyta available for phylogenomic analyses, we sequenced DNA of seven red algae: *Galdieria* sp. ACUF 613 (Cyanidiophyceae), *Madagascaria erythrocladioides* (Compsopogonophyceae), *Erythrolobus coxiae*, *Porphyridium aerugineum*, *Timspurckia oligopyrenoides* (Porphyridiophyceae), *Dixoniella grisea*, and *Rhodella violacea* (Rhodellophyceae). We assembled their nuclear genomes and predicted protein-coding genes. These new sequences, together with previously published genome- and transcriptome-derived predicted proteomes available in the EukProt database (Richter et al. 2022), cover all seven classes of Rhodophyta and, for the first time, allow analyzing the cryptophyte NMs phylogenetic position in relation to the full known diversity of red algae.

In order to maximize the independence of our phylogenomic matrix from plastid data, we started with a seed dataset of 230 NM-encoded proteins (markers) from the non-photosynthetic cryptophyte *Cryptomonas paramecium*, which were annotated as not being associated with the plastid (Tanifuji et al. 2010). Our final phylogenomic sequence supermatrix contained 51,610 amino acid positions belonging to 180 housekeeping protein markers (annotation in supplementary file S1); it included 54 species covering Viridiplantae, Glaucophyta, *Rhodelphis*, all seven classes of Rhodophyta including three of the four known orders of Cyanidiophyceae (Park et al. 2023), and NMs of four distantly related cryptophyte species: *C. paramecium*, *Guillardia theta*, *Chroomonas mesostigmatica*, and *Hemiselmis andersenii* (Hoef-Emden et al. 2002; Kim and Archibald 2013; Greenwold et al. 2023).

We carried out phylogenomic analyses of the sequence matrix by maximum likelihood (ML) with the LG+C60+G+F mixture model and by Bayesian inference with the CAT-GTR model. The resulting trees were outgroup-rooted on the non-rhodophyte archaeplastida. Both

analyses (figure 1) recovered the expected topology of Archaeplastida and Rhodophyta, including the monophyly of Proteorhodophytina with full support (1/100/100: Bayesian posterior probability/ultrafast bootstrap/nonparametric bootstrap). Surprisingly, the analyses recovered the cryptophyte NMs as sister to Cyanidiophytina with strong support (1/99/99). This result is conflicting with all previously published phylogenomic analyses of CASH plastids (including Cryptophyta) based on plastid-encoded genes, which unanimously placed CASH as sister to the other major branch of red algae, the Rhodophytina. To test the robustness of the recovered topology, we carried out a series of analyses (tests 1–8) designed to uncover potential influence of long-branch or compositional attractions on the ML results (figure 2).

We first inferred individual phylogenetic trees for each of the 180 markers, and then progressively removed from the supermatrix the markers producing the longest trees based on (1) the length of the NM stem branch (first common ancestor to last common ancestor of NMs; test 1), and (2) the average length of the NM branches, from the last common ancestor of NMs to the tips (test 2). We also progressively removed individual long-branching NM sequences based on their cumulative branch length from the root of NMs to the tip (test 3). For test 4, we used the estimated evolutionary rate of each position in the supermatrix to progressively remove the fastest-evolving sites. For test 5, we evaluated the amino acid composition of sequences from individual species in the supermatrix; we determined that cryptophyte NMs are significantly enriched in F, S, I, N, and K and depleted in A, P, V, D, and M amino acids, while Cyanidiophytina orthologs are moderately enriched in L and W amino acids (figure 2F). We thus progressively removed the most compositionally biased sites in the NMs from the supermatrix.

For all these tests, we inferred ML trees from supermatrices generated by gradually reducing divergent data in increments of 10%. Subsequently, we visualized the fluctuations in ultrafast bootstrap statistical support for the branches of interest. (figure 2A–E). Only in test 2 do we see the support for the monophyly of cryptophyte NMs and Cyanidiophytina substantially drop, and an alternative topology (NMs sister to all Rhodophyta) rose over 50% support in a single tree (with 60% of markers remaining). None of the other tests showed this alternative topology and the support for NMs sister to Cyanidiophytina remained above 90%.

For test 6, we recoded the supermatrix into a four-character state amino acid alphabet in order to minimize phylogenetic artefacts stemming from poorly modelled compositional heterogeneity (Susko and Roger 2007; Brown et al. 2013). The resulting ML tree presented the same topology as the original with 99% support for NMs as sister to Cyanidiophytina (figure 2G).

In test 7, we performed the Approximately Unbiased (AU) test (Shimodaira 2002) to compare the original tree with alternative topologies (figure 2H). The AU test rejected all topologies except for the original ML tree (figure 2G: topology 7.1); the topology consistent with the previous plastid-encoded proteins analyses i.e., NMs sister to Rhodophytina could not be rejected, although the p-value (0.0924) was on the cusp of the cut-off for significance (figure 2G: topology 7.3). Notably, the topology briefly supported in test 2 (figure 2G: topology 7.2) was rejected by the AU test (p-value=0.0154).

Finally, in test 8, we removed 7% of the most functionally divergent sites from the supermatrix, as estimated by FunDi (Gaston et al. 2011), in order to evaluate the effect of functional divergence in the protein sequences. This aimed to reduce potential artefacts from model misspecification since phylogenetic model currently available do not properly capture functional divergence in proteins (Inagaki et al. 2003; Muñoz-Gómez et al. 2022). ML

phylogenetic analysis of the resulting supermatrix again recovered the original topology with 99% support for the monophyly of NMs and Cyanidiophytina (figure 2G).

Taken together, our analyses show that the hypothesis of cryptophyte NMs branching sister to Cyanidiophytina is the one best supported by housekeeping genes still preserved in cryptophyte nucleomorph genomes. There is no indication whatsoever that this result was influenced by long-branch attraction or compositional bias. However, the previously suggested position of NMs as sister to Rhodophytina is not strongly rejected by the NM data and given its recovery in plastid genome phylogenetic analyses (Iida et al. 2007; Janouškovec et al. 2010; Ševčíková et al. 2015; Kim et al. 2017), we cannot at this stage favor one over the other.

Most importantly, our analyses neither show a close relationship between cryptophyte NMs and any of the six classes of Rhodophytina, nor their origin from within the known extant diversity of Cyanidiophytina. On the contrary, they support a deep and ancient origin of the cryptophyte NMs and, therefore, presumably also of all CASH plastids, within red algae. They likely emerged as sister to either Cyanidiophytina or Rhodophytina, two alternative positions placed just a single node away from each other.

Our results suggest that the secondary red endosymbiosis occurred either before the diversification of extant rhodophyte classes, or later, from a so-far-unknown deep branching lineage of red algae. This currently unknown lineage has either gone extinct (except for the complex plastids derived from it) or still exists, waiting to be discovered. Indeed, the discovery of a novel deep branching lineage in the phylogenetic vicinity of Rhodophyta would not be unprecedented (Gawryluk et al. 2019) and, as recent advances in primary endosymbiosis research indicate (Ponce-Toledo et al. 2017) , it may take a lucky find of a microbe closely related to our organelle of interest to resolve a long-standing disagreement

between evolutionary scenarios. Furthermore, if the sister position of cryptophyte NMs to Cyanidiophytina recovered in our analyses gets confirmed, it would support a freshwater and/or thermoacidophilic ancestry of CASH plastids, consistent with the notion that most of the early evolution of photosynthetic eukaryotes occurred in freshwater environments (Lewis 2017; Ponce-Toledo et al. 2017; Sánchez-Baracaldo et al. 2017).

Materials and Methods

For culturing, DNA extraction, and sequencing methods see (van Beveren et al. 2022). Raw genomic reads obtained by Illumina HiSeq 2500 and NovaSeq 6000 (for *D. grisea*, *E. coxiae*, *Galdieria* sp. ACUF 613, *M. erythrocladioides*, *P. aerugineum*, and *T. oligopyrenoides*) and by Illumina HiSeq 2500 and NovaSeq 6000 plus Oxford Nanopore MinION Mk 1B (for *R. violacea*) were assembled using SPAdes assembler, version 3.14.1 (Prjibelski et al. 2020). Nuclear contigs were separated from organellar and prokaryotic contamination-derived sequences using Whokaryote, version 1.1.2 (Pronk and Medema 2022). Raw reads were mapped to resulting contigs using the BWA-MEM algorithm (github.com/lh3/bwa) and BLASTn (Altschul et al. 1990) was used to search the contigs against the NCBI nt database, release 240. Results of the read mapping and BLASTn search were visualized in BlobTools2 (Challis et al. 2020) and used for additional manual decontamination, resulting in preliminary genomic assemblies (not shown).

Previously published transcriptomic reads for *M. erythrocladioides* (SRX554333), *T. oligopyrenoides* (SRX554192), and *R. violacea* (ERX2100208) were downloaded from the NCBI SRA database and mapped to their respective genomic assemblies using the BWA-MEM algorithm and visualized in the Integrative Genomics Viewer, version 2.9.4 (Robinson et al. 2011). Sets of >200 protein-coding gene models were manually predicted for each of the three assemblies with mapped transcripts and used as training sets for the online

AUGUSTUS gene predictor (Hoff and Stanke 2013). The resulting species parameters, as well as previously available *Galdieria sulphuraria* (Cyanidiophyceae) species parameters, were used in parallel for AUGUSTUS automatic gene prediction based on each of the seven genomic assemblies. The completeness of the individual sets of AUGUSTUS-predicted genes was evaluated using BUSCO, version 5.4.3, with the eukaryota_odb10 dataset (Manni et al. 2021) and the one with the highest estimated completeness was selected for each of the seven assemblies. These preliminary predicted proteomes were searched for protein sequences of interest for use in further analyses.

The 230 *C. paramecium* NM-encoded protein sequences from the "Conserved ORFs" dataset (Tanifuji et al. 2010) were searched against a custom protein database consisting of the seven novel rhodophyte preliminary predicted proteomes plus three cryptophyte NM predicted proteomes downloaded from the NCBI GenBank database, release 245, and multiple predicted proteomes representing a broad diversity of eukaryotes downloaded from the EukProt database, version 2 (Richter et al. 2022). Two best (lowest E-value) BLASTp hits were recovered per proteome for each query sequence. The sequences were aligned using MAFFT, version 7.455 (Katoh and Standley 2013) and resulting alignments were trimmed using BMGE, version 1.12 (Criscuolo and Gribaldo 2010) and then used to reconstruct initial single-gene trees using FastTree, version 2.1.8 with JTT+CAT model (Price et al. 2010). The trees (supplementary file S2) were manually inspected and used to: 1) select 180 proteins that produced trees where all the NM sequences branched together to the exclusion of all other sequences, 2) select 54 proteomes most relevant to the studied question, i.e. four from cryptophyte NMs and 50 from four lineages of Archaeplastida (Rhodophyta, *Rhodelphis*, Glaucophyta, and Viridiplantae), and 3) when paralogs were present, select a single paralog per tree – determined by a shorter branch. This selection resulted in the initial phylogenomic

dataset of 180 markers in 54 species. Details about the selected species are shown in supplementary file S3.

The selected protein sequences were gathered in full length into 180 single-gene multi-FASTA files. PREQUAL, version 1.01 (Whelan et al. 2018) was used to identify and mask positions with non-homologous characters and the masked sequences were aligned using MAFFT, version 7.455 (Katoh and Standley 2013). The alignments were then trimmed using DIVVIER (-partial -mincol 28), version 1.01 (Ali et al. 2019). The resulting alignments (supplementary file S4) were used to reconstruct single-marker phylogenetic trees as well as to assemble a concatenated supermatrix using phyutility, version 2.2 (Smith and Dunn 2008). Details about the matrix are shown in supplementary file S1.

The supermatrix (supplementary file S5) was analyzed with IQ-TREE, version 2.2.0 (Minh et al. 2020) using the substitution model LG+C60+G+F and both ultrafast bootstrap (-bb 1000) and nonparametric bootstrap (-bo 100) for the maximum likelihood (ML) phylogenetic analysis and with PhyloBayes-MPI, version 1.8c (Lartillot et al. 2013) with the CAT-GTR substitution model (Lartillot and Philippe 2004) for the Bayesian inference phylogenetic analysis (figure 1). The PhyloBayes-MPI analysis with two chains was run for more than 50,000 generations. The chains did not converge (maxdiff=1, meandiff=0.0190534, burn-in: 1000, sub-sampling every 10 trees), as is often the case with large phylogenomic datasets. Trace files were inspected with Tracer (Rambaut et al. 2018). Trees were visualized and edited using TreeViewer, version 2.1.0 (Bianchini and Sánchez-Baracaldo 2024).

Single-marker trees (supplementary file S6) were reconstructed using IQ-TREE (-bb 1000 -m LG+C60+G+F) version 2.2.0. PhyKIT, version 1.11.3 (Steenwyk et al. 2021) was used to extract branch lengths in the 180 single-marker trees and this information guided further tests of topology robustness by sequential removal of markers or individual sequences.

In test 1, all markers were sorted by the length of the root branch of cryptophyte NMs (first common ancestor to last common ancestor of NMs) from longest to shortest to generate additional supermatrices, by removing, in a stepwise fashion, 10% of the markers. In test 2, all markers were sorted by the average compound length of cryptophyte NM branches (first common ancestor of NMs to tips) from longest to shortest to generate 8 supermatrices, by removing, in a stepwise fashion, 10% of the markers. In test 3, only individual sequences were removed, not entire markers. All NM sequences were sorted by their compound branch length (first common ancestor of NMs to tips) from longest to shortest to generate 8 supermatrices, by removing, in a stepwise fashion, 10% of the NM sequences. In this test, 14 sequences (from NMs, Rhodophyta, and *Rhodelphis*), identified in the IQ-TREE-generated single-marker trees as disrupting NM monophyly, were removed together with the first 10% longest-branch sequences. Details about the removed markers and sequences and their sorting in tests 1–3 are shown in supplementary file S7.

In test 4, IQ-TREE, version 2.2.0 was used to calculate site-specific evolutionary rates (--rate -mlrate -m LG+C60+G+F) for each position in the original supermatrix (supplementary file S8) and SiteStripper, version 1.02 (github.com/hverbruggen/SiteStripper) was used to generate 8 supermatrices, each by removing, in a stepwise fashion, 10% of the positions with the highest evolutionary rates.

In test 5, amino acid compositional bias of the species in the supermatrix was determined according to the $\chi 2$ metric as described before (Viklund et al. 2012; Martijn et al. 2018; Muñoz-Gómez et al. 2019). Occurrences of all amino acids in all positions in the matrix were counted and the positions were sorted according to the ratio F+S+I+N+K / A+P+V+D+M in NMs divided by the ratio F+S+I+N+K / A+P+V+D+M in all other species. In order to avoid dividing by zero, an arbitrary value of 0.01 was added to all the amino acid occurrence values

before the calculation. Details about this calculation and sorting are given in supplementary file S9. A custom Python script was used to generate 8 supermatrices, each with 10% of the remaining positions removed starting with those with NMs most enriched in F+S+I+N+K and other species most depleted in A+P+V+D+M.

The modified supermatrices generated in tests 1–5 were used to reconstruct ML trees (IQ-TREE, -bb 1000 -m LG+C60+G+F) and the ultrafast bootstrap values supporting the original or alternative positions of NMs within the resulting topologies were plotted in figure 2.

To further evaluate the effects of compositional heterogeneity on phylogenetic inference, in test 6, we recoded the supermatrix into a four-character state amino acid alphabet that minimizes the compositional heterogeneity present in the data. The program minmax-chisq, which implements the methods of Susko and Roger (Susko and Roger 2007), was used to find the best recoding scheme for our supermatrix. The best recoding scheme (S4) was found to be NQFPV, RDCLKMST, AEGHIY, and W. Phylogenetic analyses on the recoded supermatrix were performed with IQ-TREE, version 2.2.2.7 and the across-site compositionally heterogeneous GTR+C60(S4)+F+R6 model. C60(S4) is an adaptation of the C60 mixture model to four-character states. It is obtained by adding the frequencies of the amino acids that belong to each bin in the dataset-specific four-character state scheme S4.

In test 7, we calculated ML trees (IQ-TREE, -bb 1000 -m LG+C60+G+F) with alternative topologies by constraining possible relationships under following scenarios: 1) unconstrained original tree, i.e. NMs+Cyanidiophytina, 2) NMs outside Rhodophyta, 3) NMs+Rhodophytina, 4) NMs+Eurhodophytina, and 5) NMs+Proteorhodophytina. AU-test (Shimodaira 2002) was used as implemented in IQ-TREE (-m LG+C60+G+F -n 0 -zb 1000 -au) to test the alternative hypotheses.

To evaluate the effect of functional divergence in the protein sequences, in test 8 we used the software FunDi, version 1.1 (Gaston et al. 2011) to estimate functionally divergent sites in the branch that separates nucleomorphs from all other nuclear genomes. About 7% of all amino acid sites were estimated to be functionally divergent with a posterior probability of > 0.5. These were subsequently removed using a custom Perl script. The resulting supermatrix consisted of 47,951 amino acids. Phylogenetic analyses on the supermatrix without functionally divergent sites were performed with IQ-TREE version 2.2.2.7 and the across-site compositionally heterogeneous LG+C60+G4+F mixture model. All alternative trees used in the tests can be found in supplementary file S10.

Data Availability

All new sequences used in the phylogenomic analysis were deposited in NCBI GenBank under accession numbers PP113013–PP114032.

Acknowledgments

This work was funded by the European Research Council Advanced Grants Protist World (No. 322669, P.L.-G.) and Plast-Evol (No. 787904, D.M.) and the ERC Starting grant MacroEpik (No. 803151, L.E.). We thank the UNICELL single-cell genomics platform (deemteam.fr/en/unicell) for help in DNA preparation and Nanopore sequencing.

References

Ali RH, Bogusz M, Whelan S. 2019. Identifying clusters of high confidence homologies in multiple sequence alignments. *Molecular Biology and Evolution* 36:2340.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215:403–410.

van Beveren F, Eme L, López-García P, Ciobanu M, Moreira D. 2022. Independent size expansions and intron proliferation in red algal plastid and mitochondrial genomes. *Genome Biology and Evolution* 14:evac037.

Bianchini G, Sánchez-Baracaldo P. 2024. TreeViewer: Flexible, modular software to visualise and manipulate phylogenetic trees. *Ecology and Evolution* 14:e10873.

Bodył A. 2018. Did some red alga-derived plastids evolve via kleptoplastidy? A hypothesis. *Biological Reviews* 93:201–222.

Brown MW, Sharpe SC, Silberman JD, Heiss AA, Lang BF, Simpson AGB, Roger AJ. 2013. Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. *Proceedings of the Royal Society B: Biological Sciences* 280:20131755.

Burki F, Roger AJ, Brown MW, Simpson AGB. 2020. The new tree of Eukaryotes. *Trends in Ecology & Evolution* 35:43–55.

Cavalier-Smith T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *Journal of Eukaryotic Microbiology* 46:347–366.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. 2020. BlobToolKit – Interactive quality assessment of genome assemblies. *G3 (Bethesda)* 10:1361–1374.

Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology* 10:210.

Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng L-T, Wu X, Reith M, Cavalier-Smith T, Maier U-G. 2001. The highly reduced genome of an enslaved algal nucleus. *Nature* 410:1091–1096.

Douglas SE, Murphy CA, Spencer DF, Gray MW. 1991. Cryptomonad algae are evolutionary chimaeras of two phylogenetically distinct unicellular eukaryotes. *Nature* 350:148–151.

Gaston D, Susko E, Roger AJ. 2011. A phylogenetic mixture model for the identification of functionally divergent protein residues. *Bioinformatics* 27:2655–2663.

Gawryluk RMR, Tikhonenkov DV, Hehenberger E, Husnik F, Mylnikov AP, Keeling PJ. 2019. Non-photosynthetic predators are sister to red algae. *Nature* 572:240–243.

Greenwold MJ, Merritt K, Richardson TL, Dudycha JL. 2023. A three-genome ultraconserved element phylogeny of cryptophytes. *Protist* 174:125994.

Hoef-Emden K, Marin B, Melkonian M. 2002. Nuclear and nucleomorph SSU rDNA phylogeny in the Cryptophyta and the evolution of cryptophyte diversity. *J Mol Evol* 55:161–179.

Hoff KJ, Stanke M. 2013. WebAUGUSTUS—A web service for training AUGUSTUS and predicting genes in eukaryotes. *Nucleic Acids Research* 41:W123–W128.

Iida K, Takishita K, Ohshima K, Inagaki Y. 2007. Assessing the monophyly of chlorophyll-c containing plastids by multi-gene phylogenies under the unlinked model conditions. *Molecular Phylogenetics and Evolution* 45:227–238.

Inagaki Y, Blouin C, Susko E, Roger AJ. 2003. Assessing functional divergence in EF□1α and its paralogs in eukaryotes and archaebacteria. *Nucleic Acids Research* 31:4227–4237.

Janouškovec J, Horák A, Oborník M, Lukeš J, Keeling PJ. 2010. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proceedings of the National Academy of Sciences* 107:10949–10954.

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

Kim E, Archibald JM. 2013. Ultrastructure and molecular phylogeny of the cryptomonad *Goniomonas avonlea* sp. nov. *Protist* 164:160–182.

Kim JI, Moore CE, Archibald JM, Bhattacharya D, Yi G, Yoon HS, Shin W. 2017. Evolutionary dynamics of cryptophyte plastid genomes. *Genome Biology and Evolution* 9:1859–1872.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution* 21:1095–1109.

Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Systematic Biology* 62:611–615.

Lewis LA. 2017. Hold the salt: Freshwater origin of primary plastids. *Proceedings of the National Academy of Sciences* 114:9759–9760.

Ludwig M, Gibbs SP. 1987. Are the nucleomorphs of cryptomonads and *Chlorarachnion* the vestigial nuclei of eukaryotic endosymbionts? *Annals of the New York Academy of Sciences* 503:198–211.

Ludwig M, Gibbs SP. 1989. Evidence that the nucleomorphs of *Chlorarachnion reptans* (Chlorarachniophyceae) are vestigial nuclei: Morphology, division and DNA-DAPI fluorescence. *Journal of Phycology* 25:385–394.

Maier U-G, Hofmann CJB, Eschbach S, Wolters J, Igloi GL. 1991. Demonstration of nucleomorph-encoded eukaryotic small subunit ribosomal RNA in cryptomonads. *Molec. Gen. Genet.* 230:155–160.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution* 38:4647–4654.

Martijn J, Vosseberg J, Guy L, Offre P, Ettema TJG. 2018. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* 557:101–105.

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.

Moore CE, Curtis B, Mills T, Tanifuji G, Archibald JM. 2012. Nucleomorph genome sequence of the cryptophyte alga *Chroomonas mesostigmatica* CCMP1168 reveals lineage-specific gene loss and genome complexity. *Genome Biol Evol* 4:1162–1175.

Muñoz-Gómez SA, Hess S, Burger G, Lang BF, Susko E, Slamovits CH, Roger AJ. 2019. An updated phylogeny of the Alphaproteobacteria reveals that the parasitic Rickettsiales and Holosporales have independent origins. *eLife* 8:e42535.

Muñoz-Gómez SA, Mejía-Franco FG, Durnin K, Colp M, Grisdale CJ, Archibald JM, Slamovits CH. 2017. The new red algal subphylum Proteorhodophytina comprises the largest and most divergent plastid genomes known. *Current Biology* 27:1677-1684.e4.

Muñoz-Gómez SA, Susko E, Williamson K, Eme L, Slamovits CH, Moreira D, López-García P, Roger AJ. 2022. Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria. *Nat Ecol Evol* 6:253–262.

Park SI, Cho CH, Ciniglia C, Huang T-Y, Liu S-L, Bustamante DE, Calderon MS, Mansilla A, McDermott T, Andersen RA, et al. 2023. Revised classification of the Cyanidiophyceae based on plastid genome data with descriptions of the Cavernulicolales ord. nov. and Galdieriales ord. nov. (Rhodophyta). *Journal of Phycology* 59:444–466.

Petersen J, Ludewig A-K, Michael V, Bunk B, Jarek M, Baurain D, Brinkmann H. 2014. *Chromera velia*, endosymbioses and the Rhodoplex hypothesis—Plastid evolution in cryptophytes, alveolates, stramenopiles, and haptophytes (CASH lineages). *Genome Biology and Evolution* 6:666–684.

Ponce-Toledo RI, Deschamps P, López-García P, Zivanovic Y, Benzerara K, Moreira D. 2017. An early-branching freshwater cyanobacterium at the origin of plastids. *Current Biology* 27:386–391.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLOS ONE* 5:e9490.

Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes de novo assembler. *Current Protocols in Bioinformatics* 70:e102.

Pronk LJU, Medema MH. 2022. Whokaryote: Distinguishing eukaryotic and prokaryotic contigs in metagenomes based on gene structure. *Microbial Genomics* 8:000823.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarization in bayesian phylogenetics using Tracer 1.7. *Syst Biol* 67:901–904.

Richter DJ, Berney C, Strassert JFH, Poh Y-P, Herman EK, Muñoz-Gómez SA, Wideman JG, Burki F, de Vargas C. 2022. EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. *Peer Community Journal* 2:e56.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative Genomics Viewer. *Nat Biotechnol* 29:24–26.

Sánchez-Baracaldo P, Raven JA, Pisani D, Knoll AH. 2017. Early photosynthetic eukaryotes inhabited low-salinity habitats. *Proc Natl Acad Sci U S A* 114:E7737–E7745.

Sanchez-Puerta MV, Delwiche CF. 2008. A hypothesis for plastid evolution in chromalveolates. *Journal of Phycology* 44:1097–1107.

Sarai C, Tanifuji G, Nakayama T, Kamikawa R, Takahashi K, Yazaki E, Matsuo E, Miyashita H, Ishida K, Iwataki M, et al. 2020. Dinoflagellates with relic endosymbiont nuclei as models for elucidating organellogenesis. *Proceedings of the National Academy of Sciences* 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.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. *Syst Biol* 51:492–508.

Smith SA, Dunn CW. 2008. Phyutility: A phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* 24:715–716.

Steenwyk JL, Buida TJ III, Labella AL, Li Y, Shen X-X, Rokas A. 2021. PhyKIT: A broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data. *Bioinformatics* 37:2325–2331.

Stiller JW, Schreiber J, Yue J, Guo H, Ding Q, Huang J. 2014. The evolution of photosynthesis in chromist algae through serial endosymbioses. *Nat Commun* 5:5764.

Susko E, Roger AJ. 2007. On reduced amino acid alphabets for phylogenetic inference. *Molecular Biology and Evolution* 24:2139–2150.

Tanifuji G, Onodera NT, Wheeler TJ, Dlutek M, Donaher N, Archibald JM. 2010. Complete nucleomorph genome sequence of the nonphotosynthetic alga *Cryptomonas paramecium* reveals a core nucleomorph gene set. *Genome Biol Evol* 3:44–54.

Viklund J, Ettema TJG, Andersson SGE. 2012. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol Biol Evol* 29:599–615.

Whelan S, Irisarri I, Burki F. 2018. PREQUAL: Detecting non-homologous characters in sets of unaligned homologous sequences. *Bioinformatics* 34:3929–3930.

Wong DK, Grisdale CJ, Slat VA, Rader SD, Fast NM. 2023. The evolution of pre-mRNA splicing and its machinery revealed by reduced extremophilic red algae. *Journal of Eukaryotic Microbiology* 70:e12927.

Yoon HS, Hackett JD, Pinto G, Bhattacharya D. 2002. The single, ancient origin of chromist plastids. *Proceedings of the National Academy of Sciences* 99:15507–15512.

Yoon HS, Nelson W, Lindstrom SC, Boo SM, Pueschel C, Qiu H, Bhattacharya D. 2017. Rhodophyta. In: Archibald JM, Simpson AGB, Slamovits CH, editors. Handbook of the Protists. Cham: Springer International Publishing. p. 89–133. Available from: https://doi.org/10.1007/978-3-319-28149-0_33

Figure legends

Figure 1. Phylogenetic tree placing cryptophyte NMs among Archaeplastida. (**A**) Phylogenetic tree constructed from a concatenation of 180 nucleus/nucleomorph-encoded proteins (51,610 positions, 54 taxa) with maximum likelihood (LG+C60+G+F). Branch support values are derived from Bayesian posterior probability (CAT-GTR) and maximum likelihood ultrafast and nonparametric bootstrap. (**B**) Completeness of the individual datasets shown as a schematic representation of the complete supermatrix alignment (gaps in white).

Figure 2. Results of the tests of topology robustness. (**A–E**) Changes in ultrafast bootstrap support for the sister relationship between cryptophyte NMs and Cyanidiophytina or alternative topologies as calculated based on reduced supermatrices in tests 1–5. (**F**) Principal component analysis of amino acid composition of the individual datasets used in the phylogenetic analyses. (**G**) Simplified schematic representation of the resulting phylogenetic tree with positions of alternative placement of cryptophyte NMs used in test 7 and ultrafast bootstrap support for the sister relationship between cryptophyte NMs and Cyanidiophytina as calculated based on modified supermatrices in tests 6 and 8. (**H**) Results of topology tests calculated with IQ-TREE in test 7.

Figures

Figure 1

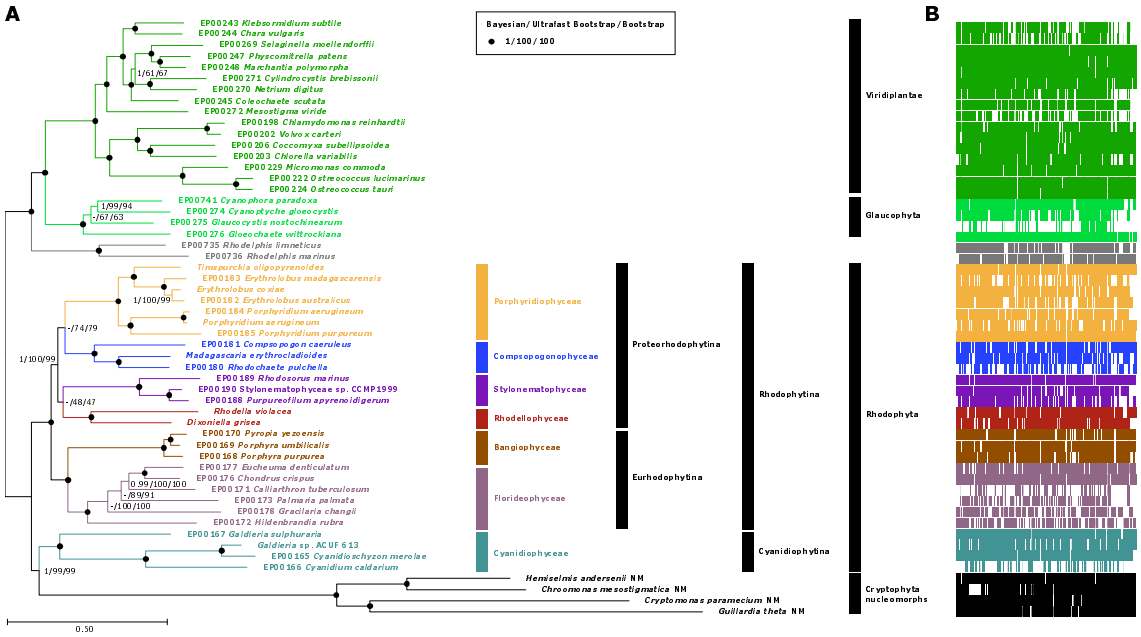
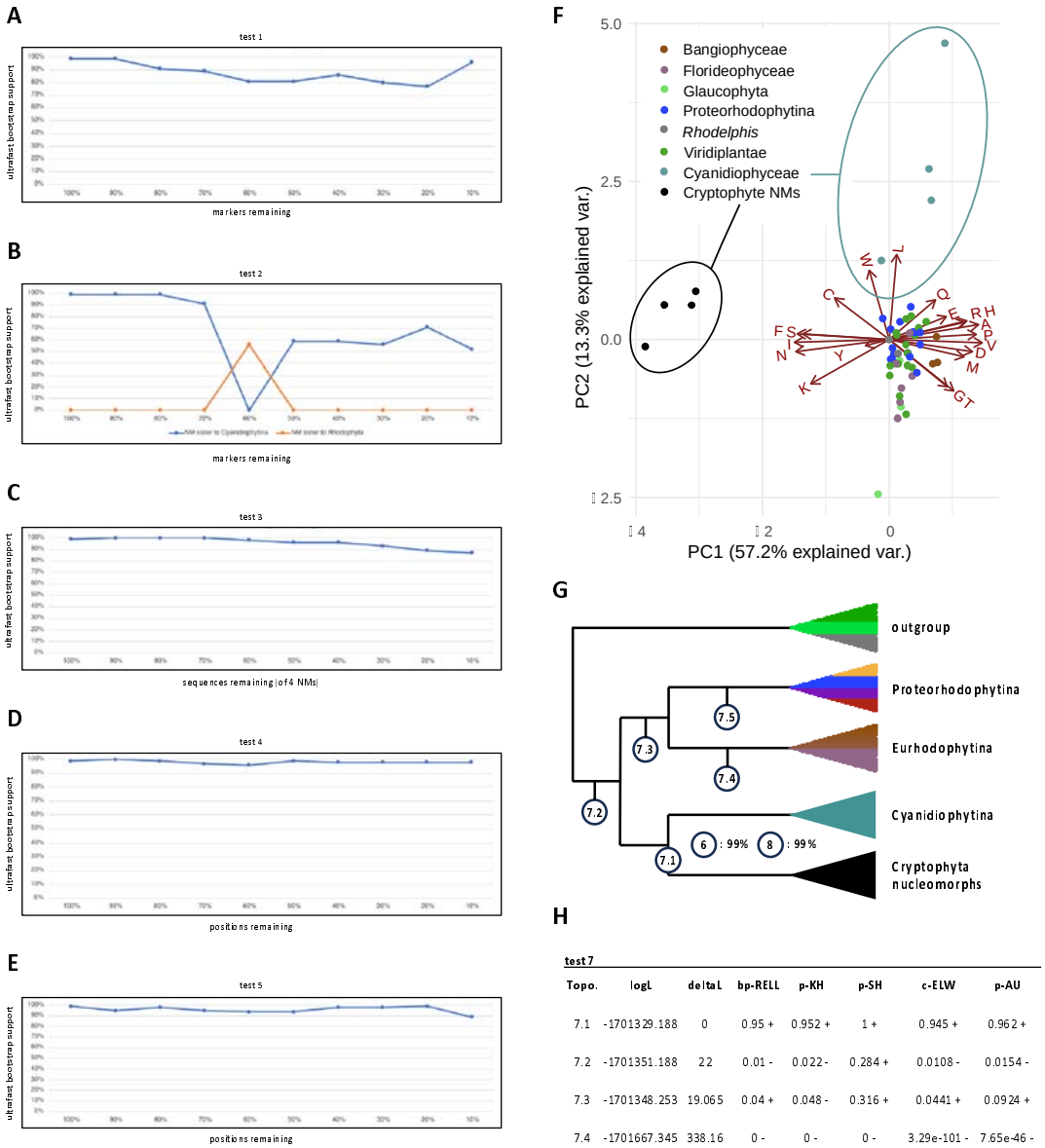


Figure 2



H

test 7

Topo.	logL	deltaL	bp-RELL	p-KH	p-SH	c-ELW	p-AU
7.1	-1701329.188	0	0.95 +	0.952 +	1 +	0.945 +	0.962 +
7.2	-1701351.188	22	0.01 -	0.022 -	0.284 +	0.0108 -	0.0154 -
7.3	-1701348.253	19.065	0.04 +	0.048 -	0.316 +	0.0441 +	0.0924 +
7.4	-1701667.345	338.16	0 -	0 -	0 -	3.29e-101 -	7.65e-46 -
7.5	-1701666.461	337.27	0 -	0 -	0 -	1.98e-99 -	4.11e-52 -