New deep-branching environmental plastid genomes on the algal tree of life

Mahwash Jamy[1], Thomas Huber[2], Thibault Antoine[3], Hans-Joachim Ruscheweyh[4], Lucas Paoli[5], Eric Pelletier[3], Tom O. Delmont[3]*, Fabien Burki[2]*

[1]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, Sweden
[2]Department of Organismal Biology (Systematic Biology), Uppsala University, Uppsala, Sweden
[3]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France
[4]Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, ETH Zürich, 8093, Zürich, Switzerland
[5]Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland

*co-corresponding authors

Abstract

Marine algae support the entire ocean ecosystem and greatly impact planetary biology. The availability of algae in culture poorly represents their large environmental diversity, and we still have a limited understanding of their convoluted evolution by endosymbiosis. Here, we performed a phylogeny-guided plastid genome-resolved metagenomic survey of *Tara* Oceans expeditions. We present a manually curated resource of nearly 700 new non-redundant plastid genomes of environmental pelagic algae. This catalogue vastly expands the plastid genome diversity within major algal groups, often corresponding to algae without closely related reference genomes. We also recovered four genomes, including one near complete, forming a deep-branching plastid lineage of nano-size algae (<5 μm) that we informally named leptophytes. This group is globally distributed and generally rare, although it can reach relatively high abundance notably in the Arctic. Leptophytes encompass the enigmatic marine plastid group DPL2, one of the very few known plastid groups not clearly belonging to any major algal groups and for which only 16S amplicon data is available. Extensive phylogenetic analyses based on 93 plastid genes and gene content comparison firmly place leptophytes together with haptophytes and cryptophytes, although the exact relationships remain ambiguous. Collectively, our study demonstrates that metagenomics can reveal currently hidden diversity of organellar genomes, and shows the importance of including this diversity to improve models for plastid evolution.

Key words: Plastid evolution, endosymbiosis, plankton, *Tara* Oceans, genome-resolved metagenomics, ptMAGs, leptophytes, cryptophytes, haptophytes, red alga

Main text

Eukaryotic algae are essential to the success of life on earth, performing nearly 40% of primary production and forming the basis of many food webs[1]. Algae are incredibly diverse in forms and functions, ranging from some of the smallest known eukaryotes in the open ocean phytoplankton to undersea forests of giant coastal seaweed. The phylogenetic position of algae in the eukaryotic tree of life shows that they have evolved in multiple supergroups, owing their origins to a complex and largely unresolved history of endosymbioses. The first eukaryotic algae originated from the endosymbiotic integration of cyanobacteria into predatory host cells. This primary endosymbiosis led to the diversification of eukaryotes with photosynthetic organelles called primary plastids in the Archaeplastida supergroup, containing major algal groups such as red algae and green algae (including land plants)[2,3]. Subsequently, plastids were transferred from red and green algae to unrelated groups of eukaryotes by further endosymbioses, spreading photosynthetic capabilities throughout the tree of eukaryotes[4]. Major questions related to plastid evolution remain open, notably about the number of endosymbioses and partners involved to explain the observed phylogenetic distribution of plastids. The evolutionary history of complex plastids with red algal origin has been among the most difficult to resolve, in spite of the ecological importance of some of these algae in modern oceans such as diatoms, dinoflagellates, and coccolithophores[5,6]. Multiple hypotheses have been proposed, many involving successive endosymbioses between eukaryotes over vast evolutionary distances, but there is no consensus and very little empirical evidence to support one or another hypothesis[2,7–9].

A more complete catalogue of plastid genomes across the algal tree is needed to better understand the evolution of plastid endosymbiosis. In recent years, comparatively more data has been produced for nuclear genomes of algae—including many transcriptomes—than for the smaller and in principle easier to sequence plastid genomes[10]. There are currently several major groups of algae that have poor representation of plastid data. For example, only 17 plastid genomes are available for haptophyte species, one of the most abundant and bloom-forming marine algal groups containing as many as 1000 described species and a much larger environmental diversity[5]. Beyond known algae, a few so-called deep-branching plastid groups, *i.e.*, plastids that do not belong to any established algal groups, have been reported but currently lack any genomic or morphologic description. The discovery of deep-branching plastid groups is extremely rare, comparable to the discovery of new animal groups[11], owing to a good understanding of what the major algal groups are. The most notable cases of new deep-branching groups in the past 20 years include the rappemonads—but they have since been formally assigned to haptophytes into the new class Rappephyceae[12]—or the enigmatic environmental plastid lineages DPL1 and DPL2 for which only 16S rRNA amplicon sequences are available[13]. While DPL1 was weakly placed within haptophytes but outside of known classes, DPL2 branched deeper

possibly as sister to haptophytes but without statistical support[13]. Finally, the Picozoa (previously picobiliophytes), which were originally described as a new algal group with unknown affinities to eukaryotes, are now considered aplastidic members of Archaeplastida[14].

Here, we used 280 billion metagenomic reads covering a wide range of *Tara* Oceans planktonic size fractions[15,16] as well as additional resources[17,18] to assemble and manually curate 667 non-redundant environmental plastid genomes (ptMAGs). We show that genome-resolved metagenomics is a powerful approach for reconstructing organellar genomes and use this unprecedented plastid genomic resource to determine the phylogeny and geographical distribution of planktonic algae. Strikingly, we report a generally rare but widespread and deep-branching clade of plastid genomes encompassing the enigmatic DPL2 that we name leptophytes.

Results

A large resource of environmental plastid genomes from the sunlit ocean

To characterize the diversity of eukaryotic algae in the global sunlit ocean, we surveyed large *Tara* Oceans metagenomic co-assemblies (~12 million contigs >2.5 kb) covering plankton size fractions ranging from 0.8 µm to 2mm (Supplementary Table 1)[16,19]. These co-assemblies, each deriving from samples from a particular oceanic basin, have been shown to improve the recovery of genomes that are too scarce to be assembled from individual samples[16,20]. Briefly, we used a phylogeny-guided metagenomic approach based on the RNApolB gene, and complementary genomic and environmental information to characterize and manually curate an initial set of 1,448 plastid metagenome-resolved genomes (ptMAGs) with anvi'o[21,22] (see Methods). Two additional single-contig ptMAGs related to the focal clade of our study (see subsequent sections) were subsequently added from mOTUs-db[18]. After filtering for redundancy (average nucleotide identity <98%), our final plastid genome database derived from all oceanic regions consisted of 667 new ptMAGs, as well as 164 reference plastid genomes (Supplementary Fig. 1, Supplementary Table 1). The occurrence of 44 single copy plastid core genes indicated that the ptMAGs are 63.2% complete on average, with low redundancy (average: 2.5%) (Supplementary Fig. 2). 131 ptMAGs contained at least 40 of the 44 core plastid genes and were thus deemed highly complete. However, no ptMAG mapped as a circular sequence due to the absence of the ribosomal RNA operon which is often missing from MAGs. The ptMAGs had a median size of 66 kbp, with GC content ranging between 24.9% and 46.3%, and encoded up to 179 genes (Supplementary Table 1; Supplementary Figs. 3-5). Collectively, the ptMAGs were far more abundant in the sunlit oceans compared to selected culture representative genomes (84.2% of the total signal), in line with a previous comparison of nuclear genomes among the same metagenomes[16] (Supplementary Fig. 6).

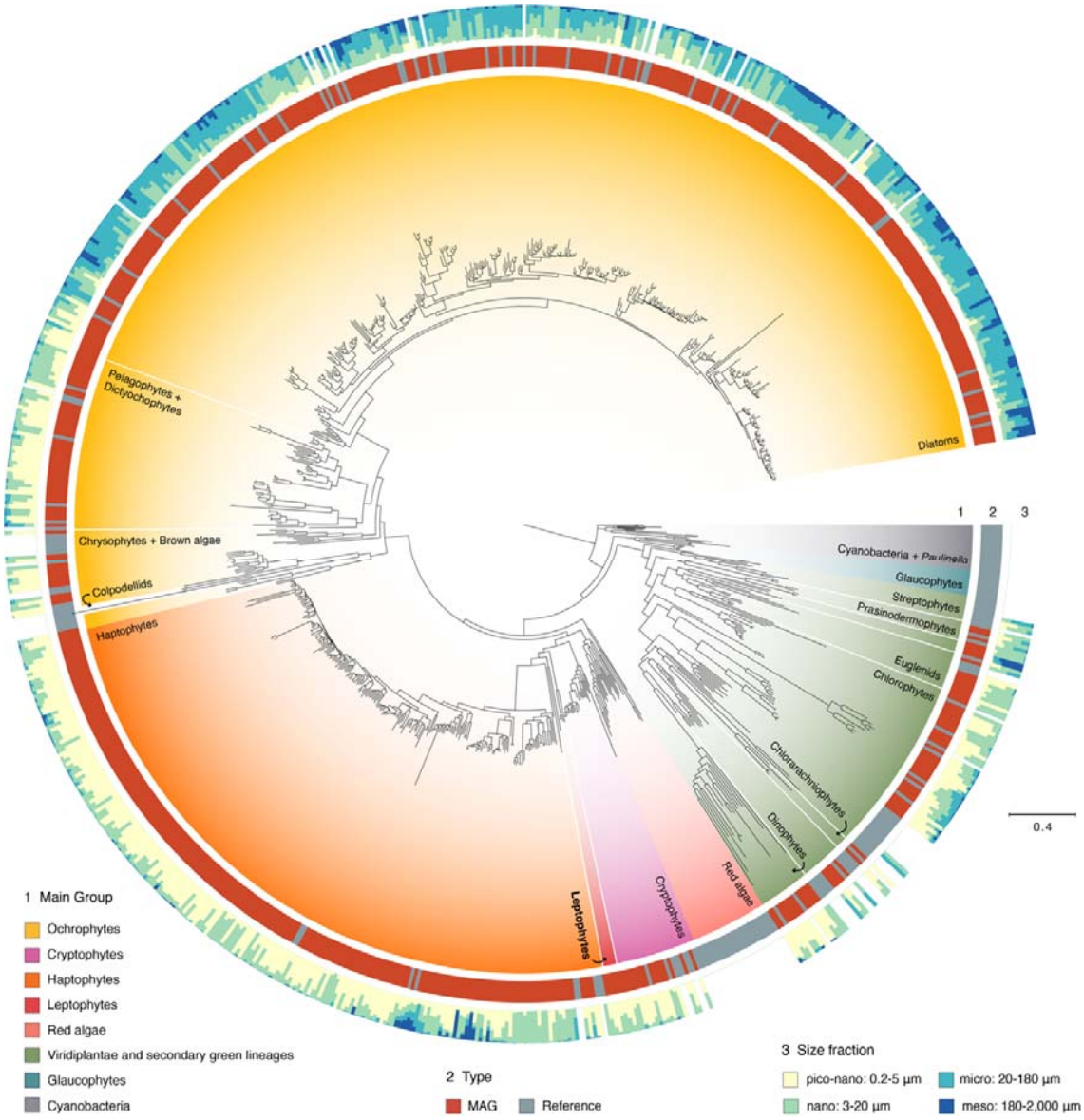


Fig. 1. Global phylogenetic analysis of plastid genomes from the sunlit ocean.
Maximum-Likelihood phylogeny inferred using 93 plastid-encoded genes and the LG+F+I+G4 model in IQ-TREE. The phylogeny contains 180 manually selected reference genomes and 667 plastid MAGs (ptMAGs) from various algal groups; cyanobacteria (including the *Paulinella* plastid) were set as outgroup. Ring 1 around the phylogeny indicates the algal groups; Ring 2 depicts whether the taxon is a reference sequence or a MAG; Ring 3 depicts the putative size of each marine taxon as determined by metagenomic mapping of 937 metagenomes of Tara Oceans of each genome (see Methods).

We reconstructed a broad multigene phylogeny of our plastid genome database to assess its taxonomic diversity. Given the large taxon sampling, we inferred a Maximum Likelihood phylogeny

using the relatively simple site-homogeneous LG+F+I+G4 model based on 93 plastid-encoded genes. This phylogeny recovered all major groups of algae and is overall in good agreement with previously published plastid trees (Fig. 1, Supplementary Fig. 7)[9,23]. The ptMAGs occurred in most, but not all, major groups of algae. We did not recover ptMAGs for red algae and glaucophytes, which are known to be rare or absent in the open ocean[24], nor for the divergent peridinin plastids of dinoflagellates, which are highly fragmented and lack the *rpoB* gene[25]. Conversely, the ptMAGs were affiliated to ochrophytes (n=377, including 305 diatoms displaying a very broad cellular size range), haptophytes (n=201, including *Phaeocystales* found in large size fractions likely due to their distinct colony-forming lifestyle[5]), cryptophytes (n=16, restrained to the small size fractions), and green algae (n=45, predominantly found in the smallest size fraction). In multiple cases, the ptMAGs represented novel genome diversity within established algal groups where cultured references are lacking. For instance, one ptMAG (REFM_CHLORO_00002) deeply branched as sister to Pavlovales (Fig. 1 and Supplementary Fig. 7). Eight chlorophyte ptMAGs recovered from different oceanic regions formed a monophyletic clade distantly related to a clade of Chloropicophyceae. Within ochrophytes, 15 ptMAGs with a broad geographic distribution were closely related but clearly different from the reference pelagophyte plastid genomes. Finally, among green algae, 13 ptMAGs formed a distinct clade related to pedinophyte endosymbionts of dinoflagellates, which include the undescribed MGD and TGD[26] as well as *Lepidodinium chlorophorum*[27] (Supplementary Fig. 8). This group might represent a larger diversity of endosymbionts than previously recognized. Taken together, these examples demonstrate the usefulness of metagenomics to recover a broad diversity of plastid genomes among multiple plastid major groups, in particular filling critical evolutionary gaps with environmental genomic diversity for clades lacking reference genomes.

Leptophytes: a new deep branching plastid group

Aside from the extended genomic diversity within clearly defined plastid groups, we also recovered a fully supported deep-branching clade formed by four red algal-derived ptMAGs that is related to, but distinct from all major groups with complex red plastids, *i.e.*, haptophytes, cryptophytes, ochrophytes, and myzozoans (Fig. 1). To facilitate discussion, we propose the informal name leptophytes for this new group of plastid genomes. The name refers to the putative small size (*lepto-* indicates something small) of the corresponding algal cells, which we infer to be below 5µm based on genomic size fraction distributions across the *Tara* Oceans metagenomes (Supplementary Fig. 9). Leptophytes were collectively detected in the surface waters of most oceanic basins (Supplementary Fig. 10). At the level of individual genomes, Lepto-01 and Lepto-02 were only detected in the Arctic Ocean while Lepto-03 and Lepto-04 showed a broader and non-polar distribution (Supplementary Fig. 11). Although leptophyte genomes were usually found in relatively low abundance (detected in 104 out of 147 *Tara* stations with a median cumulative coverage of 2.3x), Lepto-01 stood out as one of the most

detected plastid genomes in two Arctic stations (up to 124x mean coverage in the 0.8-2000µm size fraction), providing evidence that leptophytes can be abundant under certain conditions (Supplementary Table 2). However, efforts to identify potential hosts for Lepto-01 at stations where its signal was strongest were unsuccessful (see Methods). Interestingly, two 16S rRNA gene fragments were recovered in Lepto-01 and Lepto-04, respectively, which revealed that leptophytes are most similar to a few amplicon sequences of the enigmatic DPL2 plastid group[13]. The 16S rRNA of Lepto-04 was 100% similar to DPL2, while Lepto-01 shared 91% similarity to DPL2. The relatedness of DPL2 to leptophytes was confirmed by 16S rRNA phylogeny, which also showed a weak affinity of the group to haptophytes (Supplementary Fig. 12). The reported sub-tropical distribution and pico-size of DPL2[13] largely overlap with Lepto-03 and Lepto-04 but not with the Arctic preference of Lepto-01 and Lepto-02, thus altogether indicating that leptophytes form a more diverse and widespread group than the narrower DPL2 lineage.

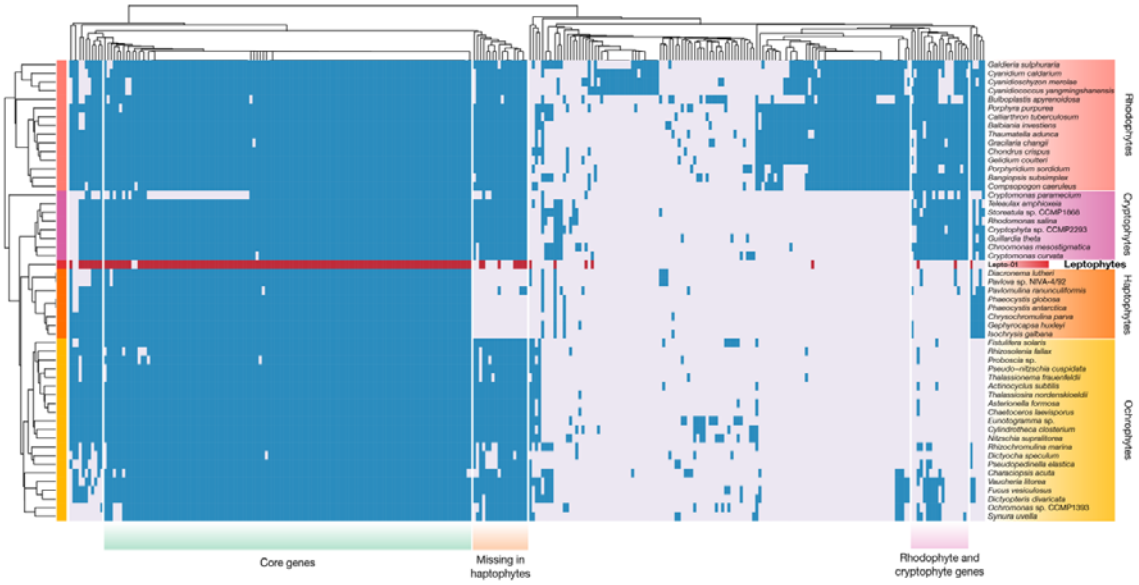


Fig. 2. Distribution patterns of plastid-encoded genes across algae with red plastids

A binary heatmap and UPGMA clustering based on the presence or absence of 355 plastid-encoded genes. Blue and grey boxes represent gene presence and absence, respectively. For leptophytes, present genes are highlighted in red for improved clarity. A larger version of this figure, including gene names, is presented in Supplementary Fig. 15.

Lepto-01 is the best representative genome for the group. This assembly had the highest coverage (detected in 20 stations with an average mean coverage of 6.3x) but it was also near complete and contiguous with a single contig of 104,203 bp. This ptMAG contained 144 genes with 118 proteins (no detectable introns) and 24 tRNAs (Supplementary Fig. 13). The functional repertoire of Lepto-01 shows that it is a photosynthetic plastid genome, retaining the core components of both photosystem I

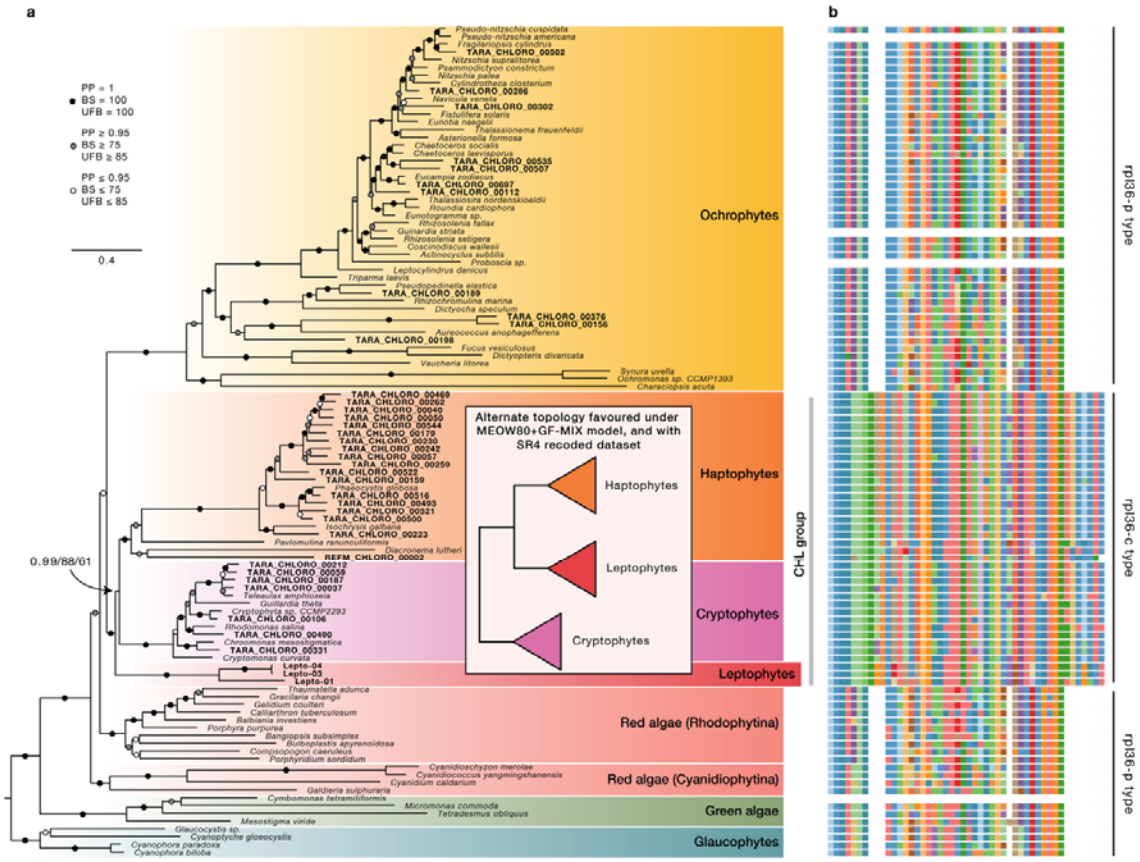
and II (psa- and psb- genes), carbon fixation (rbcL and rbcS), cytochrome b6/f complex (pet- genes), ATP synthase (atp- genes), and chlorophyll biosynthesis (chlI) (Fig. 2, Supplementary Fig. 13). While less complete and more fragmented, the three other leptophyte ptMAGs are largely syntenic with Lepto-01 and support these functional trends (Supplementary Fig. 14). The near complete assembly of Lepto-01 allowed us to compare its gene content with that of red algae and red algal-derived plastids. UPGMA-clustering of homologous gene occurrence showed that the gene content of leptophytes is most similar to that of haptophytes (Fig. 2). In particular, leptophytes are lacking 10 genes (including *dnaB*, *ftsH*, *petF*, *psbW*, *rpl4*, *rpl18*, *rpl29*, *rpl35*, *sufC*, *ycf33*) that are among the 18 genes almost completely absent in haptophytes, but mostly present in all other algal groups (Fig. 2, Supplementary Fig. 15).

Leptophytes are phylogenetically related to haptophytes and cryptophytes

To gain further insight into the position of leptophytes in the plastid tree, we reconstructed and evaluated phylogenies using several complex site-heterogeneous mixture models in Bayesian and ML implementations. These models, which consider site-specific amino-acid frequencies, have been shown to generally improve tree estimations but they come at heavier computational costs[28]. We thus produced a reduced dataset based on the same 93 genes as in the full phylogeny but containing 107 plastid genomes to speed up the analyses while maintaining the phylogenetic breadth for red algae and red algal plastid-containing lineages. Due to overall higher completeness, reference genomes were preferred over ptMAGs when possible, and the least complete leptophyte ptMAG (Lepto-02) was excluded. The battery of advanced phylogenetic analyses all unambiguously confirmed that leptophytes form a novel group of plastid genomes related to haptophytes and cryptophytes (Fig. 3a). This affinity is also supported by a shared rare bacterial gene replacement by horizontal transfer of the ribosomal gene rpl36 into the plastid genomes of all three groups (Fig. 3b). We will refer to this group as the Cryptophytes-Haptophytes-Leptophytes (CHL) group.

Leptophytes were recovered as sister to both haptophytes and cryptophytes (HC-sister topology) with the Bayesian model CAT+GTR+G4 (PP=0.99), as well as the ML models GTR+CAT-PMSF+G4 (BS=88) and LG+MEOW80+G4 (UFB=61) in which site profiles are estimated directly from the data[29,30]. The same position for leptophytes was recovered with the empirical mixture model cpREV+C60+G4 model and in an analysis that removed the 10 fastest-evolving taxa (Supplementary Fig. 16-17). Because of the relatively modest bootstrap support for the position of leptophytes in ML inferences, we compared the likelihood of the three possible topologies within the CHL group, enforced as topological constraints but letting other branches and parameters to be optimised by ML. Under the GTR+CAT-PMSF model, both alternative topologies were rejected by a likelihood χ^2 test with a Bonferroni correction for multiple comparisons[31], and the sister position to cryptophytes was

additionally rejected by an AU test (Fig 3c). The likelihood values under the LG+MEOW80+G4 model were less than four points difference across the three topologies, and none were statistically rejected (Fig. 3c).



c

Dataset	Untreated				Stationary trimmed
	ML analyses using fixed topology trees				
Model	GTR-CAT-PMSF	LG-MEOW80	LG-MEOW80 + GHOST (8 cat)	LG-MEOW80 + GF-MIX †	LG-MEOW80
Topology					
Leptophytes sister to haptophytes and cryptophytes	**-931909.236**	**-1056271.087**	**-1045458.696**	-1055620.569 *	**-592787.090**
Leptophytes sister to haptophytes	-931920.074 *	-1056274.766	-1045464.078 *	**-1055586.193**	-592793.164 *
Leptophytes sister to cryptophytes	-931924.826 **	-1056273.024	-1045463.152 *	-1055602.069 *	-592795.016 *

* Rejected by Bonferroni corrected χ2 test
** Rejected by Bonferroni corrected χ2 test and AU test
† AU test not possible under GF-MIX model

Fig. 3. Phylogenomic analyses based on a reduced taxon-sampling.
(a) Bayesian inference based on 20,292 amino acid sites across 107 taxa under the CAT+GTR+G4 model. Support values on branches show in that order: Bayesian posterior probabilities (PPs), non-parametric bootstrap support (BS; 100 replicates) under the GTR+CAT-PMSF model, and ultrafast bootstrap support (UFB; 1,000 replicates) under the LG+MEOW80+G4 model. The inset shows the alternative topology of the CHL clade. **(b)** Alignment of the *rpl-36* gene across corresponding taxa in the tree. **(c)** Maximum likelihood estimates under different models for alternate positions of leptophytes within the CHL clade. Likelihoods in bold text represent the highest scoring topology. Difference in log-likelihood scores between each topology and the best-scoring topology is shown by data bars in blue. A single asterisk indicates topologies rejected by the Bonferroni-

corrected chi-squared test, while two asterisks indicate topologies also rejected by an Approximately Unbiased (AU) test. See also Supplementary Table 3.

We then evaluated whether the position of leptophytes was influenced by two common types of systematic errors in phylogenetic reconstruction: heterotachy of evolutionary rates, and compositional heterogeneity across tree branches. To do this, we calculated the likelihoods of the three alternate phylogenies as fixed topologies under newly developed complex models as additional parameters to the LG+MEOW80+G4 model. Heterotachy was modeled with GHOST[32], while compositional heterogeneity was accounted for with the GFmix model[30,31]. The GHOST model continued to favour the HC-sister topology, but the best topology under the GFmix model changed to support the monophyly of leptophytes and haptophytes (H-sister topology) (Fig 3a and c). With both models, the alternative tested topologies were rejected by the Bonferroni-corrected likelihood χ^2 test (Fig 3c). Given the observation that compositional bias may have impacted our phylogenetic reconstructions, we attempted to minimize this artefact by recoding the 20 amino acids to a reduced 4-class alphabet (SR4 recoding)[33], and by removing the most heterogeneous sites (~25% of positions in the alignment) with a stationary-based trimmer[34]. In agreement with the GFmix model, the recoded dataset recovered the H-sister topology, albeit with low support (PP=0.75). However, the stationary-trimmed alignment recovered the HC-sister topology, also with low support (UFB=68) (Supplementary Fig. 18). Overall, these results indicate that resolving the topology of the CHL group in general and placement of leptophytes in particular is not trivial, as both HC-sister and H-sister topologies remain credible (Supplementary Note 1).

Discussion

The vast majority of genomic knowledge of plastids has been gained from cultured organisms. We demonstrate that organellar genomes can readily be recovered from large metagenomic assemblies using phylogeny-guided genome-resolved metagenomics. This study represents the first global-scale survey of environmental plastid genomes (ptMAGs), providing a culture-independent view of the diversity, functioning and evolutionary history of plastid genomes occurring in marine surface water. ptMAGs better represent the diversity and abundance of planktonic algae than reference genomes, spanning most major marine phytoplanktonic groups and often filling in evolutionary gaps in areas of the tree poorly represented by references. One very interesting new group of plastid genomes are the leptophytes, a globally distributed and generally rare deep-branching lineage formed by four ptMAGs in our data. This is remarkable because novel plastid diversity at this taxonomic depth is very rarely reported. Currently, only the environmental plastid lineages DPL1 and especially DPL2 are still

considered possible deep-branching phytoplankton lineages outside of eukaryotic supergroups[13]. Based on 16S rRNA data recovered from the ptMAGs and biogeographical comparison, we show not only that one of the leptophytes (Lepto-04) corresponds to DPL2, thus representing the first genomic data available for this enigmatic group, but also that leptophytes form a more diverse, widespread, and abundant group than previously known.

The plastids of leptophytes are evolutionarily related to cryptophytes and haptophytes, indicating that they possess red-algal derived plastids. This is firmly supported by phylogenetic analyses as well as gene content comparison and the exclusive presence of the c-type paralog of the ribosomal gene *rpl36* in cryptophytes, haptophytes and leptophytes (the CHL group). The exact phylogenetic position of leptophytes within the CHL group remains ambiguous in spite of the use of a dense taxon-sampling, 93 genes, and a range of sophisticated evolutionary models and data treatments. Taking into account this phylogenetic uncertainty, we propose two competing hypotheses for the evolutionary origin of leptophyte plastids: 1) the HC-sister topology (leptophytes are sister to both cryptophytes and haptophytes), which was favoured by all site-heterogeneous models, or 2) the H-sister topology (leptophytes are sister to haptophytes, to the exclusion of cryptophytes), which was favoured only after specifically taking into consideration amino acid compositional heterogeneity across taxa, and also by the 16S rRNA phylogeny. In contrast to the robust support for most bipartitions in the tree (Fig 3a), the ambiguity in the placement of leptophytes likely resulted from weak stochastic phylogenetic signals due to the ancient origin of the plastids and closely spaced speciation events early in the CHL group. Although a precise timeframe for the origin of the CHL plastids is unavailable, it has been suggested that both cryptophytes and haptophytes acquired their plastids more than 1 billion years ago[35]. Based on the phylogenetic position of leptophytes, it is reasonable to assume that this timeframe represents a minimum age estimate for the divergence of this group. This ancient age, combined with the extremely short internal branch uniting haptophytes and cryptophytes (Fig 3) and the limited information of plastid genomes, means that it is probably not possible with our current phylogenetic models and data to better resolve this part of the tree.

Despite the uncertainties around the position of leptophytes, the discovery of this group bears important implications on our understanding of the origin and spread of complex red plastids across the eukaryotic tree. To explain the patchy distribution of these plastids in the tree, the existing models of serial endosymbiosis all identify cryptophytes as the unique recipient of secondary red plastids (plastids derived directly from red algae), mostly because cryptophytes are the only known algae that have retained a red algal nucleomorph[36,37]. Most, but not all, models also specifically propose a plastid transfer between cryptophytes and haptophytes to take into account the rare rpl36 gene replacement into their plastid genomes[38] and affinities of both groups in several plastid phylogenies[23,39]. Our new

extensive plastid phylogeny is consistent with available data, but adding leptophytes allows us to make new predictions on the relative timing of events.

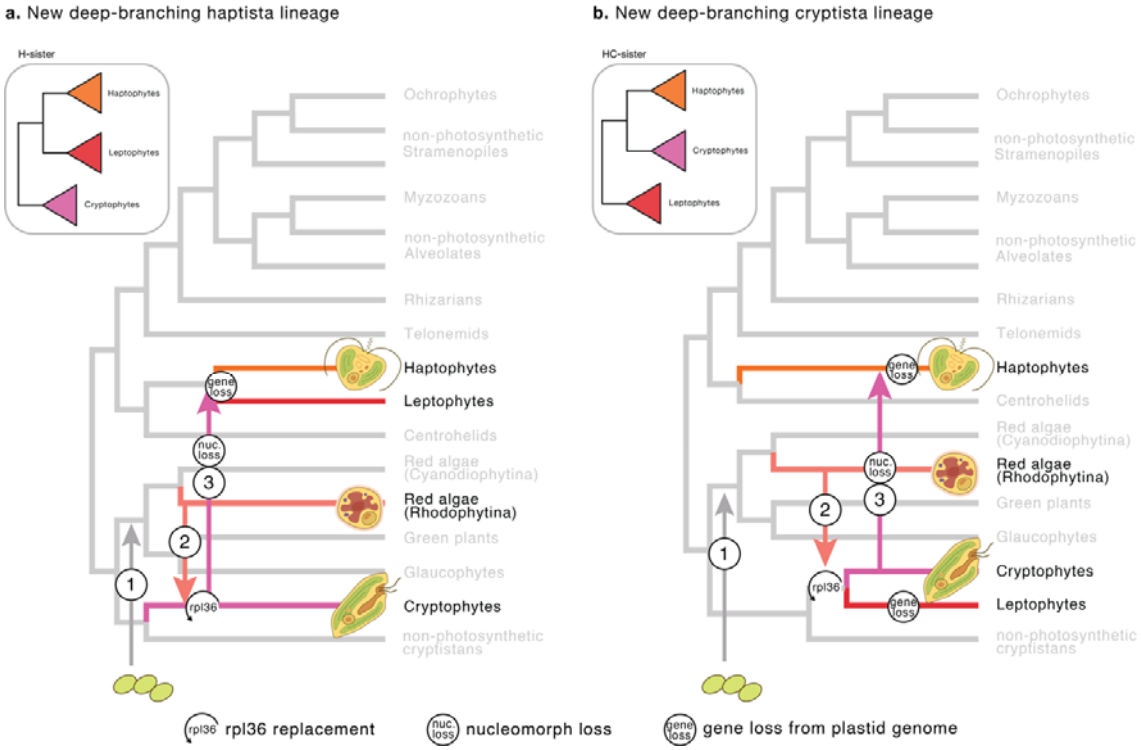


Fig. 4. Main proposed models for evolution of complex red plastids

The trees depict hypothetical host relationships with red algae, cryptophytes, haptophytes, and leptophytes highlighted. Numbers on the tree represent the level of endosymbiosis events. Inset phylogenies depict the corresponding plastid phylogenies. Plastid transfers to ochrophytes and myzozoans are not depicted for clarity. Illustrations of algae are from [40].

Reconciling our two hypotheses for the placement of leptophytes in the plastid phylogeny with the best current estimates of eukaryotic phylogeny[41], we propose two main models for the evolution of complex red plastids, focusing here on the CHL group for clarity (Fig 4). The first model assumes that leptophytes are sister to haptophytes (H-sister) also in the host phylogeny (Fig 4a). Leptophytes would have inherited their plastids, including the rpl36-c paralog, in a common ancestor with haptophytes by tertiary endosymbiosis from cryptophytes and would lack nucleomorph. Possible candidates for leptophytes in this scenario include early-branching environmental diversity of haptophytes[42,43], or even marine pico centrohelids. The second model reconciles the HC-sister plastid topology with a position of leptophytes as sister to cryptophytes in the host phylogeny, either as a new direct sister to photosynthetic cryptophytes or more deeply (Fig 4b). Deeper candidates in Cryptista exist, for example the environmental CRY3 group[44], or members of katablepharids some of which are known to

engage in tight photosymbiosis[45,46]. In this scenario, the secondary plastid from red algae was established in the common ancestor of leptophytes and cryptophytes, where the ancestral *rpl36*-p gene was replaced, before passing that plastid further to haptophytes from a stem cryptophyte. The model places the loss of nucleomorph concomitant to this tertiary transfer, opening up the possibility that leptophytes have also retained this organelle. To weigh in on both models, we searched for evidence of red algal nucleomorph in the same metagenomic data used to reconstruct the ptMAGs. While we found ample evidence of cryptophyte nucleomorphs elsewhere, we found no evidence of putative leptophyte nucleomorphs (see Methods). The H-sister model is also more parsimonious as it requires fewer parallel gene losses based on our plastid gene content comparison (7 vs 40) (Supplementary Fig. 19). Other models are at face value possible, including the possibility that leptophytes represent a deeper eukaryotic lineage with no close relationship to known supergroup (Supplementary Fig. 20). However, we deem these models less likely because they all require additional endosymbioses.

Ultimately, deciding between these scenarios will rest on characterizing the nuclear and putative nucleomorph genomes of leptophytes, ideally in the context of cultivation, in order to determine their specific position in the eukaryotic tree and investigate their cell biology. This will be challenging, as our data suggest that leptophytes are generally rare and it remains unclear if they occur in easily accessible coastal regions. Despite the fundamental importance of algae with complex red plastid as marine primary producers, we still understand little of how this great diversity of algae came to be. Our data show that players still hidden in the rare biosphere hold important clues that could bring new critical evidence. We demonstrated that organellar genome-resolved metagenomics allows access to this hidden diversity and should be considered in the future as a valuable approach to survey not only plastids but also mitochondria across biomes.

Material and Methods

***Tara* Oceans metagenomes.** The 937 metagenomes from *Tara* Oceans used in the study are publicly available at the EBI under project PRJEB402 (https://www.ebi.ac.uk/ena/browser/view/PRJEB402).

Phylogeny-guided genome-resolved metagenomics for plastids. We used two complementary datasets (metagenomic assemblies and evolutionary-informative proteins) to characterize ptMAGs. On the one side, we used 11 large metagenomic co-assemblies from *Tara* Oceans (~12 million contigs from 798 eukaryote-enriched metagenomes) which were organized into 2,550 metabins using constrained automatic binning and processed with anvi'o[21,22] for manual binning purposes[16]. On the other side, we used proteins corresponding to the DNA-dependent RNA polymerase B subunit (RNApolB) and found in those *Tara* Oceans metabins, in the context of a phylogeny of representative sequences (amino acid level with sequence similarity <90%)[47]. Until now, the two datasets had only

been used to characterize genomes corresponding to RNApolB clades of giant viruses and mirusviruses. Here, we focused on a large plastid RNApolB clade most closely related to that of Cyanobacteria and used this signal as guidance for genome-resolved metagenomics using the anvi'o interactive interface. Briefly, we mainly used sequence composition and differential coverage across metagenomes of the corresponding co-assembly to characterize and manually curate 1,448 ptMAGs in the corresponding metabins. The strong correlation between initial signal (number of plastid RNApolB genes) and outcome (number of characterized ptMAGs) across metabins (Supplementary Fig. 1) demonstrated the effectiveness of this method to survey the environmental genomics of plastids in complex marine metagenomes.

Creation of a non-redundant plastid genomic database. We collected all the manually curated Tara Oceans ptMAGs and included 164 manually selected reference plastid genomes from culture, as well as two contigs (one affiliated to the leptophytes and the other one corresponding to a deep-branching clade of haptophytes) characterized from Tara Oceans metagenomes as part of the mOTU global metagenomic resource[18]. We determined the average nucleotide identity (ANI) of each pair of genomes using skani[48] v0.2.1. Genomes were considered redundant when their ANI was >98% (minimum alignment of >25% of the smaller genome in each comparison). For genomes found to be redundant, reference chloroplast genomes were selected over ptMAGs, and in most cases where only MAGs were redundant, the longest one was selected. This analysis provided a non-redundant database containing the 164 reference plastid genomes and 667 ptMAGs. We estimated the completeness and redundancy of each genome based on the occurrence of 44 core-plastid genes[49] (Supplementary Table 4).

Biogeography of the GOEV database. We performed a mapping of 937 metagenomes from *Tara* Oceans to calculate the mean coverage (vertical coverage) and detection (horizontal coverage) of each genome in the non-redundant plastid genomic database. Briefly, we used BWA v0.7.15 (minimum identity of 95%) and a FASTA file containing all contigs from the database to recruit short reads from each metagenome[50]. We considered a genome to be detected in a metagenome when >25% of its length was covered by reads. The number of recruited reads below this cut-off was set to 0 before determining the mean coverage of genomes, as an effort to minimize non-specific read recruitments.

Genome annotation of ptMAGs. Plastid annotation was done with MFannot[51] v1.3.6 using the genetic code 11. The resulting ASN format files were converted to GenBank format using the *asn2gb* utility ([12]) from the NCBI toolkit. We then extracted all annotated proteins from the GenBank files using the script *gb_to_prot.py* (available at https://github.com/burki-lab/beta-Cyclocitral/blob/master/src/gb_to_prot.py). For the leptophyte ptMAGs, intron prediction was carried

out using RNAweasel (https://megasun.bch.umontreal.ca/apps/rnaweasel/), and maps were generated using OGDraw version 1.3.1 (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html) [52].

Taxonomic annotation of ptMAGs. We manually determined the taxonomy of ptMAGs based on the global phylogenetic analysis (see section "Phylogenomic analyses"), using guidance from the reference plastid genomes.

Phylogenomic dataset assembly

We used a set of 93 genes for phylogenetic analyses in this study (Supplementary Table 5). Briefly, we started with a gene set of 66 genes from [53], which was then expanded by including the gene sets used in [54] and [55] when they were also present in leptophytes.

We first built a dataset of 93 protein-coding genes from the 180 reference taxa. This was done as follows: (1) Homologous sequences from all reference taxa were retrieved with BLASTP[56] searches using sequences from *Fucus vesiculosus* (which has a well-annotated plastid genome; NC_016735) as queries (e-value: 1e-02), (2) For each gene, sequences were aligned with MAFFT[57] v.7.407 using the -auto option, and then trimmed with trimAL[58] v1.4.1 using a gap threshold of 0.8, (3) Single-gene trees were inferred with raxml-ng[59] v.1.2.0 using the LG4X model and 100 rapid bootstrap searches, (4) Gene trees were manually parsed to identify duplicates and extremely long branches corresponding to sequences from another gene, which were then removed from the dataset.

We then proceeded to construct a dataset initially including the 1448 ptMAGs identified (including redundant ptMAGs). Using reference sequences from the previous step as queries, we retrieved homologous sequences from the ptMAGs using BLASTP (e-value: 1e-02) and the gene annotation from MFannot. The 93 single-gene dataset was aligned individually using MAFFT-G-INS-I (using the --unalign 0.6 option to avoid over-alignment), and gently trimmed with trimAL (gap threshold: 0.9). Single-gene trees were inferred with IQTree[60] v.2.2.2.6 using the best fitting model as determined by ModelFinder Plus[61]. We manually parsed the single gene trees to check for putative chimeras and generated our final dataset comprised 93 genes and 847 taxa (667 ptMAGs and 180 references) with > 4% data.

The concatenated alignment was generated as follows: for each curated gene, we first used Prequal[62] v1.02 to filter sequence stretches with no clear homology using a posterior probability threshold of 0.95, then aligned with MAFFT-G-INS-I (--unalign 0.6), and subsequently trimmed with BMGE v1.12[34] (BLOSUM35 matrix, gap threshold: 0.8). After concatenating, we obtained a supermatrix (847-taxon dataset) with 18,759 aligned amino acid sites, which was used for broad-scale phylogenetic analyses. Finally, we also generated a smaller, 107-taxon dataset to enable analyses

under more complex and computationally demanding models. For this dataset, raw sequences for the 93 genes were again filtered, aligned, trimmed and concatenated as previously described, yielding a supermatrix with 20,292 amino acid sites.

Phylogenomic analyses

Site-homogeneous models

To assess the phylogenetic diversity of the ptMAGs, the full 847-taxon supermatrix was used to compute a maximum-likelihood tree using the best-fitting site-homogeneous model, LG+F+I+G4 in IQ-TREE v.2.2.2.6. Branch support was assessed with 1000 ultrafast bootstrap replicates.

Site-heterogeneous models

We used the smaller, 107-taxon dataset to perform analyses with the more complex, site-heterogeneous models in order to determine the position of leptophytes in the plastid tree. An initial tree was inferred using the best-fitting cpREV+C60+G4 model in IQ-TREE with 1000 ultrafast bootstrap replicates. This phylogeny was used as a guide tree to compute custom site-profiles directly from the input alignment using the programme MEOW[30] (-ri -p R -C 5 -f H -l), an approach which often outperforms generic mixture models (such as the C-series)[63]. MEOW (MAMMaL Extension On Whole alignments) extends the MAMMaL approach by using all variable sites in the alignment to estimate custom site-profile classes. Following Williamson et al[30], we estimated three sets of profile classes using varying proportions of high- and low rate classes respectively: (1) MEOW(80,0) indicates 80 classes from high rate sites and none from low rate sites, (2) MEOW(60,20) represents 60 classes from high rate sites and 20 classes from low rate sites, and (3) MEOW(40,40) corresponds to 40 classes each from the two partitions. We inferred a maximum likelihood phylogeny under the best fitting model based on the BIC criterion, LG+MEOW(60,20)+G4, with 1000 UFBOOT replicates, and for brevity, we refer to this model as LG+MEOW80+G throughout the manuscript.

Another Maximum-likelihood phylogeny was inferred with the 107-taxon dataset with the CAT-PMSF approach[29]. Briefly, a phylogeny inferred under the site-homogeneous LG+G model in IQ-TREE was used as a fixed tree for analysis with the CAT+GTR+G model in PhyloBayes-MPI v1.8[64]. Two Markov Chain Monte Carlo (MCMC) chains were run for more than 5,000 cycles until the effective sample sizes of nearly all parameters (seven out of eight) were above 100 (Supplementary Table 6). The posterior mean exchangeabilities and the custom site-frequency profiles were extracted for analysis in IQ-TREE and support was assessed with 100 non-parametric bootstrap replicates.

Finally, we also analysed the 107-taxon dataset in PhyloBayes under the CAT+GTR+G model via three independent MCMC chains. The chains were run for almost 9,000 cycles each. After discarding

the first 2,000 cycles as burn-in, convergence was assessed (with maxdiff reaching 0.047) and consensus trees were generated using the bpcomp command.

Derivatives of 107-taxon dataset

First, we tested the impact of fast-evolving taxa in our dataset by removing 10 taxa (all ochrophytes) with the largest root to tip distances, generating a 97-taxon dataset. Second, we dealt with possible compositional heterogeneity across taxa by using BMGE's stationary-based trimmer to remove sites based on Stuart's test of marginal homogeneity (BMGE -s FAST -h 0:1 -g 1; 4752 out of 20292 sites = 23.4% sites removed). Finally, we dealt with possible compositional heterogeneity and sequence saturation by recoding the dataset into SR4 categories[33]. Both the non-recoded datasets were analysed in IQ-TREE with the LG+MEOW80+G model with 1,000 ultrafast bootstraps, while the SR4 recoded dataset was used for phylogenetic inference under the CAT+GTR+G model in PhyloBayes (three MCMC chains with ~22,000 cycles, burn-in of 4,000, and maxdiff reaching 0.101).

Assessing alternative topologies

Site-heterogeneous models

To assess alternate positions of leptophytes, we performed constrained tree searches in IQ-TREE. We considered three possible positions in the algal tree: (1) sister to haptophytes and cryptophytes (HC-sister), (2) sister to haptophytes (H-sister), and (3) sister to cryptophytes (C-sister). For each of these three positions, we considered two topologies; one in which the complex plastids are monophyletic, and the other where the complex plastids are non-monophyletic (as recently recovered by [55]), bringing the total number of tested topologies to six. Constrained topologies were inferred with IQ-TREE under the GTR+CAT-PMSF+G and LG+MEOW80+G models for the 107-taxon dataset, and the xmC60SR4 model for the SR4 recoded dataset, with all parameters optimized freely. In the case of the LG+MEOW80+G model, ten independent runs in IQ-TREE were conducted for each constraint to avoid likelihood scores from searches stuck in local optima[65]. In all cases, the topologies were evaluated with the Approximately Unbiased (AU)[66] test implemented in IQ-TREE. Additionally, the best scoring topology under each model was compared with alternative topologies using the Bonferroni-corrected χ^2 test[30,31]. Briefly, the likelihood ratio test statistic (LRS; two times the difference between the maximised log-likelihood of the two topologies under consideration) is first calculated. The p-value for the χ^2 test (χ) is calculated as the probability that a chi-squared random variable would be larger than the LRS with the degrees of freedom defined as the number of branches collapsed in the strict consensus tree of the two trees. The Bonferroni-corrected p-value is calculated as $(1- \{1 - p\}^{A})$ where A is the set of trees compatible with the strict consensus tree.

GHOST and GF-mix models

We additionally tested the six alternative topologies under GHOST[32] and GFmix[30,31] as additional parameters to the LG+MEOW80 base model. For GHOST, the log-likelihood scores of the six topologies (inferred under the LG+MEOW80+G model) were calculated under the LG+MEOW80+H8 model in IQ-TREE, using the topologies as fixed trees. The LG+MEOW80+H8 indicates a model with eight linked heterotachy classes and was the best fit model using the BIC criterion when compared to models with four or six heterotachy classes. Topologies were tested using the AU test and the Bonferroni-corrected χ^2 test as described previously.

The GFmix model models compositional heterogeneity by modifying the vector of amino acid frequencies in a branch-specific manner to account for shifts in the relative frequencies of amino acids in different branches of the tree. While prior studies using this model have divided their alignment into multiple partitions to account for varying compositional heterogeneity in mitochondrial and nuclear encoded genes (e.g., [30,31]), we opted to use a single partition. This is because we do not expect substantial discrepancy in compositional heterogeneity between the only plastid-encoded genes considered in this study. With no *a priori* knowledge of how taxa might be compositionally biased, we first estimated the groups of amino acids that are depleted and enriched in different lineages from our data as done in [30]. Briefly, we used a custom R script (C. McCarthy, Dalhousie University, Canada) to carry out the following steps: (1) tabulate the counts of all 20 amino acids for each taxon, (2) perform a chi-squared test on these counts and use the Pearson's residuals to construct a UPGMA tree of taxa, (3) separate taxa into two groups based on the UPGMA tree, (4) conduct a binomial test for each amino acid and calculate a score reflecting the differences in counts between the two groups, and (5) classify amino acids into "enriched," "depleted," or "other" categories based on their scores (Supplementary Note 1). The enriched and depleted classes of amino acids were used by the GFmix programme to compute the log-likelihoods of the six alternative topologies, which were inferred using the LG+MEOW80+G model and provided as fixed input trees.

16S phylogeny

Barrnap v0.9 (https://github.com/tseemann/barrnap) was used to detect 859 bp and 808 bp fragments of the 16S gene in the Lepto-01 and Lepto-04 ptMAGs respectively. A BLAST search of the Lepto-01 16S sequence revealed that it was most similar to the DPL2[13] group with ~91% sequence similarity (accession numbers EF574856 and KX935025). However, the Lepto-04 16S sequence best BLAST hits included DPL2 as well as cyanobacteria. Manually BLASTing fragments of the Lepto-04 16S sequence against nt indicated that it was likely chimeric, with the first ~440 bp corresponding to DPL2 (100% similarity), and the latter half originating from cyanobacteria. We therefore trimmed the Lepto-04 16S sequence to only the first 440 bp. To confirm the association of leptophytes with DPL2, we constructed a 16S dataset including Lepto-01, Lepto-04, DPL2, and representatives of each eukaryotic algal clade. The 65-taxon dataset was aligned with MAFFT-L-INS-I, and trimmed with

trimAL (-gt 0.1) to get an alignment with 1496 sites. A Maximum-Likelihood phylogeny was inferred with raxml-ng with 20 ML searches and non-parametric bootstrapping until convergence using the GTR+G model.

Rpl36 alignment and phylogeny

To investigate whether leptophytes possessed the *rpl36*-p type or *rpl36*-c type gene [38], we extracted the rpl36 amino acid sequences from the 107 taxa in the subset phylogeny when available. The sequences were aligned with MAFFT-L-INS-I, trailing ends trimmed manually, and the alignment was visualised alongside the 93-gene PhyloBayes phylogeny using the R package ggtree v3.6.2[67].

Synteny analysis of leptophyte ptMAGs

Synteny analysis of the leptophyte ptMAGs was performed using the GenBank files generated during genome annotation using PyGenomeViz v1.0.0 via the PyGenomeViz Streamlit Web Application (https://pygenomeviz.streamlit.app/). Sequence similarity of protein coding genes was estimated using MMseqs2 for reciprocal best-hit CDS search[68].

Survey for the nuclear sequences of leptophytes

Survey for putative nucleomorphs of leptophytes

We searched for genomic signal corresponding to red-algal nucleomorphs associated with leptophytes across the 11 Tara Oceans metagenomic co-assemblies. Specifically, we surveyed the assembled RNApolB genes, following the protocol used to characterize ptMAGs but this time using publicly available nucleomorph RNApolB genes as a reference. With this approach, we did not identify any nucleomorph-affiliated RNApolB genes co-occurring with the ptMAGs of leptophytes.

Survey for leptophyte psbO sequences

We considered the psbO gene as a good candidate to try and find the nuclear genome of leptophytes, as it is encoded by all photosynthetic eukaryotes, usually, in a single copy, in their nuclear genome, and is absent from non-photosynthetic eukaryotes[69]. We searched for HMM psbO sequences in the Arctic metagenomic co-assembly (available at: https://www.genoscope.cns.fr/tara/; where Lepto-01 is abundant) using an HMM search based on the PFAM accession PF01716. A total of 51 hits were detected which were added to 346 reference psbO sequences from different algal groups. Sequences were aligned (MAFFT-L-INS-I), trimmed (trimAL, -gt 0.1), and then used to infer a tree with raxml-ng (20 searches, non-parametric bootstrapping until convergence) to determine the phylogenetic affiliation of the 51 psbO sequences. However, all psbO candidates fell within established algal groups, and no putative psbO sequence corresponding to leptophytes could be identified. Additionally, we focused on the two filter where Lepto-01 was among the most abundant eukaryotic plastids (Supplementary Table 1), and extracted the top 25 most abundant psbO sequences in these

two filters from the recently assembled psbO database[69]. We generated an alignment and inferred a phylogeny as before, but again, did not identify any strong candidates for leptophytes.

Survey for leptophyte 18S sequences

We used two approaches to search for leptophyte 18S sequences: (1) First, we coassembled the two samples where Lepto-01 was abundant using Megahit v1.2.9[70], and extracted 18S sequences from the coassembly using Barrnap (length cutoff: 0.2). These 18S sequences were BLASTed (e-value: 1e-10) against PR2 v5.0.0[71] and the nt database to identify them, however, all sequences were linked to well established clades and no candidate leptophyte 18S sequence could be identified. (2) Second, 18S-V9 metabarcoding ASV sequences from the 15 filters where Lepto-01 was most abundant were extracted (n=5,336). We attempted to match the ASVs to Lepto-01 based on correlation of abundances following [72], but did not succeed in finding suitable candidates.

Data availability. Data our study generated has been deposited in an online repository: https://figshare.scilifelab.se/articles/dataset/_/28212173. This link provides access to the individual FASTA files on each plastid genome used in our study (including the 667 non-redundant ptMAGs), individual gene alignments, concatenated and trimmed alignments, and maximum-likelihood and Bayesian tree files for the phylogenomic dataset.

Code availability. All scripts used for genome annotation and phylogenetic analyses are available on GitHub: https://github.com/burki-lab/beta-Cyclocitral.

Acknowledgements

We thank Shinichi Sunagawa for having facilitated the recovery of relevant data from the mOTU metagenomic database maintained by his research group at the Department of Biology at ETH Zürich. We thank A. Roger, H. Baños, and C. McCarthey for discussions, and for kindly providing custom scripts for running the phylogenetic models MEOW and GF-MIX. Our survey was made possible by the sampling and sequencing efforts of the *Tara* Oceans Project. *Tara* Oceans (which includes the *Tara* Oceans and *Tara* Oceans Polar Circle expeditions) would not exist without the leadership of the *Tara* Oceans Foundation and the continuous support of 23 institutes (https://oceans.taraexpeditions.org/). Phylogenetic analyses were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS 2024/5-197), partially funded by the Swedish Research Council through grant agreement no. 2022-06725. M.J. was supported by the Swedish Research Council (International Postdoc grant 2022-00351). FB's research is supported by grants from the European Research Council (ERC consolidator grant 101044505), the

Swedish Research Council VR (2021-04055), and Science for Life Laboratory. TD's research is supported by a grant from the l'Agence Nationale de la Recherche (ANR-23-CE02-0022).
We also thank the commitment of the CNRS and Genoscope/CEA. Some of the computations were performed using the platine, titane and curie HPC machine provided through GENCI grants (t2011076389, t2012076389, t2013036389, t2014036389, t2015036389 and t2016036389).

Author contributions

F.B. and T.O.D. conceived the project. T.O.D. characterized the ptMAGs. M.J., F.B. and T.H., performed phylogenetic analyses. T.A., E. P. and T.O.D. created the plastid genomic database and performed surveys for nucleomorphs. H.J.R. and L.P. retrieved relevant data from mOTUs. M.J. and T.H. annotated the ptMAGs, and E.P. performed mapping analyses. F.B., M.J. and T.O.D. wrote the manuscript with input from all the authors.

References

1. Andersen, R. A. Diversity of eukaryotic algae. *Biodivers. Conserv.* **1**, 267–292 (1992).
2. Archibald, J. M. The puzzle of plastid evolution. *Curr. Biol.* **19**, pR81-R88 (2009).
3. Irisarri, I., Strassert, J. F. H. & Burki, F. Phylogenomic Insights into the Origin of Primary Plastids. *Syst. Biol.* **71**, 105–120 (2022).
4. Keeling, P. J. The endosymbiotic origin, diversification and fate of plastids. *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 729–748 (2010).
5. Penot, M., Dacks, J. B., Read, B. & Dorrell, R. G. Genomic and meta-genomic insights into the functions, diversity and global distribution of haptophyte algae. *Appl. Phycol.* **3**, 340–359 (2022).
6. Malviya, S. *et al.* Insights into global diatom distribution and diversity in the world's ocean. *Proc. Natl. Acad. Sci.* **113**, E1516–E1525 (2016).
7. Dorrell, R. G. *et al.* Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. *eLife* **6**, e23717 (2017).
8. Stiller, J. W. *et al.* The evolution of photosynthesis in chromist algae through serial endosymbioses. *Nat. Commun.* **5**, 5764 (2014).

9. 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).
10. Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLOS Biol.* **12**, e1001889 (2014).
11. Moreira, D. & López-García, P. The rise and fall of Picobiliphytes: How assumed autotrophs turned out to be heterotrophs. *BioEssays* **36**, 468–474 (2014).
12. Kawachi, M. *et al.* Rappemonads are haptophyte phytoplankton. *Curr. Biol.* **31**, 2395–2403 (2021).
13. Choi, C. J. *et al.* Newly discovered deep-branching marine plastid lineages are numerically rare but globally distributed. *Curr. Biol.* **27**, R15–R16 (2017).
14. Schön, M. E. *et al.* Single cell genomics reveals plastid-lacking Picozoa are close relatives of red algae. *Nat. Commun.* **12**, 6651 (2021).
15. Sunagawa, S. *et al.* Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* **18**, 428–445 (2020).
16. Delmont, T. O. *et al.* Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics* **2**, 100123 (2022).
17. Ruscheweyh, H.-J. *et al.* Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. *Microbiome* **10**, 212 (2022).
18. Dmitrijeva, M. *et al.* The mOTUs online database provides web-accessible genomic context to taxonomic profiling of microbial communities. *Nucleic Acids Res.* **53**, D797–D805 (2025).
19. Song, W., Zhang, S. & Thomas, T. MarkerMAG: linking metagenome-assembled genomes (MAGs) with 16S rRNA marker genes using paired-end short reads. *Bioinformatics* **38**, 3684–3688 (2022).
20. Delmont, T. O. *et al.* Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat. Microbiol.* **3**, 804–813 (2018).

21. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
22. Eren, A. M. *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nat. Microbiol.* **6**, 3–6 (2021).
23. Kim, J. I. *et al.* Evolutionary Dynamics of Cryptophyte Plastid Genomes. *Genome Biol. Evol.* **9**, 1859–1872 (2017).
24. Vargas, C. de *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).
25. Muñoz-Gómez, S. A. & Slamovits, C. H. Plastid Genomes in the Myzozoa. in *Advances in Botanical Research* (eds. Chaw, S.-M. & Jansen, R. K.) vol. 85 55–94 (Academic Press, 2018).
26. Matsuo, E. *et al.* Comparative Plastid Genomics of Green-Colored Dinoflagellates Unveils Parallel Genome Compaction and RNA Editing. *Front. Plant Sci.* **13**, 918543 (2022).
27. Kamikawa, R. *et al.* Plastid Genome-Based Phylogeny Pinpointed the Origin of the Green-Colored Plastid in the Dinoflagellate *Lepidodinium chlorophorum*. *Genome Biol. Evol.* **7**, 1133–1140 (2015).
28. Baños, H., Susko, E. & Roger, A. J. Is Over-parameterization a Problem for Profile Mixture Models? *Syst. Biol.* **73**, 53–75 (2024).
29. Szánthó, L. L., Lartillot, N., Szöllősi, G. J. & Schrempf, D. Compositionally Constrained Sites Drive Long-Branch Attraction. *Syst. Biol.* **72**, 767–780 (2023).
30. Williamson, K. *et al.* A robustly rooted tree of eukaryotes reveals their excavate ancestry. 2024.09.04.611237 Preprint at https://doi.org/10.1101/2024.09.04.611237 (2024).
31. Muñoz-Gómez, S. A. *et al.* Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria. *Nat. Ecol. Evol.* **6**, 253–262 (2022).
32. Crotty, S. M. *et al.* GHOST: Recovering Historical Signal from Heterotachously Evolved Sequence Alignments. *Syst. Biol.* **69**, 249–264 (2020).
33. Susko, E. & Roger, A. J. On Reduced Amino Acid Alphabets for Phylogenetic Inference. *Mol. Biol. Evol.* **24**, 2139–2150 (2007).

34. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 1–21 (2010).
35. Strassert, J. F. H., Irisarri, I., Williams, T. A. & Burki, F. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nat. Commun.* **12**, 1879 (2021).
36. McFadden, G. I. The cryptomonad nucleomorph. *Protoplasma* **254**, 1903–1907 (2017).
37. Tanifuji, G. *et al.* Complete Nucleomorph Genome Sequence of the Nonphotosynthetic Alga *Cryptomonas paramecium* Reveals a Core Nucleomorph Gene Set. *Genome Biol. Evol.* **3**, 44–54 (2011).
38. Rice, D. W. & Palmer, J. D. An exceptional horizontal gene transfer in plastids: gene replacement by a distant bacterial paralog and evidence that haptophyte and cryptophyte plastids are sisters. *BMC Biol.* **4**, 31 (2006).
39. Khan, H. *et al.* Plastid Genome Sequence of the Cryptophyte Alga *Rhodomonas salina* CCMP1319: Lateral Transfer of Putative DNA Replication Machinery and a Test of Chromist Plastid Phylogeny. *Mol. Biol. Evol.* **24**, 1832–1842 (2007).
40. Keeling, P. J. & Eglit, Y. Openly available illustrations as tools to describe eukaryotic microbial diversity. *PLOS Biol.* **21**, e3002395 (2023).
41. Burki, F., Roger, A. J., Brown, M. W. & Simpson, A. G. B. The New Tree of Eukaryotes. *Trends Ecol. Evol.* **35**, 43–55 (2020).
42. Egge, E. S., Eikrem, W. & Edvardsen, B. Deep-branching Novel Lineages and High Diversity of Haptophytes in the Skagerrak (Norway) Uncovered by 454 Pyrosequencing. *J. Eukaryot. Microbiol.* **62**, 121–140 (2015).
43. Edvardsen, B., Egge, E. S. & Vaulot, D. Diversity and distribution of haptophytes revealed by environmental sequencing and metabarcoding – a review. *Perspect. Phycol.* **3**, 77–91 (2016).
44. Kim, E. & Archibald, J. M. Ultrastructure and Molecular Phylogeny of the Cryptomonad *Goniomonas avonlea* sp. nov. *Protist* **164**, 160–182 (2013).

45. Okamoto, N. & Inouye, I. *Hatena arenicola* gen. et sp. nov., a Katablepharid Undergoing Probable Plastid Acquisition. *Protist* **157**, 401–419 (2006).
46. Kwon, J. E. *et al.* Newly discovered role of the heterotrophic nanoflagellate *Katablepharis japonica*, a predator of toxic or harmful dinoflagellates and raphidophytes. *Harmful Algae* **68**, 224–239 (2017).
47. Gaïa, M. *et al.* Mirusviruses link herpesviruses to giant viruses. *Nature* **616**, 783–789 (2023).
48. Shaw, J. & Yu, Y. W. Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nat. Methods* **20**, 1661–1665 (2023).
49. Puerta, M. V. S., Bachvaroff, T. R. & Delwiche, C. F. The Complete Plastid Genome Sequence of the Haptophyte *Emiliania huxleyi*□: a Comparison to Other Plastid Genomes. *DNA Res.* **12**, 151–156 (2005).
50. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
51. Lang, B. F. *et al.* Mitochondrial genome annotation with MFannot: a critical analysis of gene identification and gene model prediction. *Front. Plant Sci.* **14**, 1222186 (2023).
52. Greiner, S., Lehwark, P. & Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* **47**, W59–W64 (2019).
53. 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.* **107**, 10949–10954 (2010).
54. Ponce-Toledo, R. I. *et al.* An Early-Branching Freshwater Cyanobacterium at the Origin of Plastids. *Curr. Biol. CB* **27**, 386–391 (2017).
55. Pietluch, F., Mackiewicz, P., Ludwig, K. & Gagat, P. A New Model and Dating for the Evolution of Complex Plastids of Red Alga Origin. *Genome Biol. Evol.* **16**, evae192 (2024).
56. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

57. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

58. 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).

59. Kozlov, A., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).

60. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

61. 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).

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

63. Susko, E., Lincker, L. & Roger, A. J. Accelerated Estimation of Frequency Classes in Site-Heterogeneous Profile Mixture Models. *Mol. Biol. Evol.* **35**, 1266–1283 (2018).

64. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.* **62**, 611–615 (2013).

65. Liu, C. *et al.* The Influence of the Number of Tree Searches on Maximum Likelihood Inference in Phylogenomics. *Syst. Biol.* **73**, 807–822 (2024).

66. Shimodaira, H. An Approximately Unbiased Test of Phylogenetic Tree Selection. *Syst. Biol.* **51**, 492–508 (2002).

67. Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinforma.* **69**, e96 (2020).

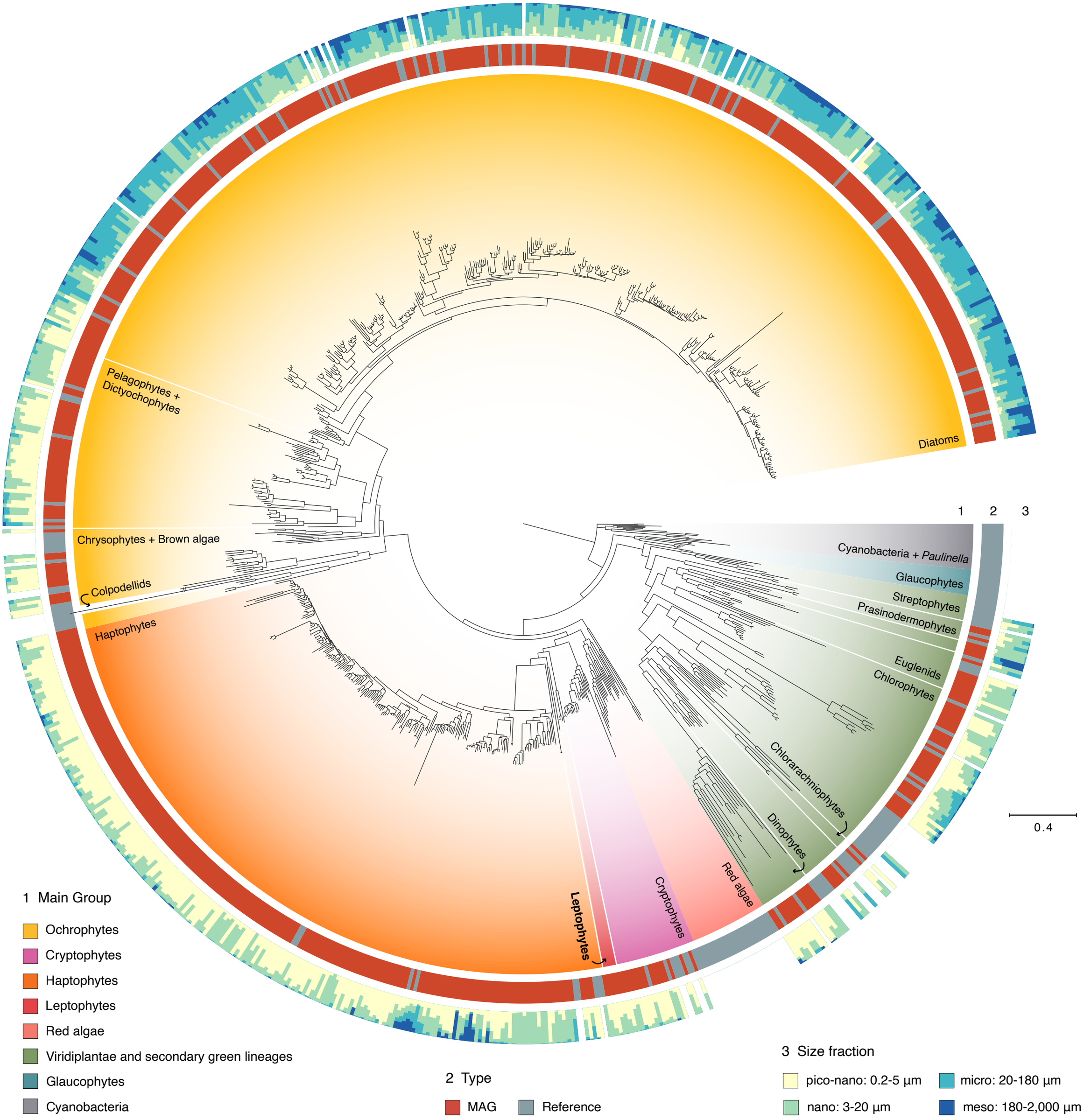
68. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

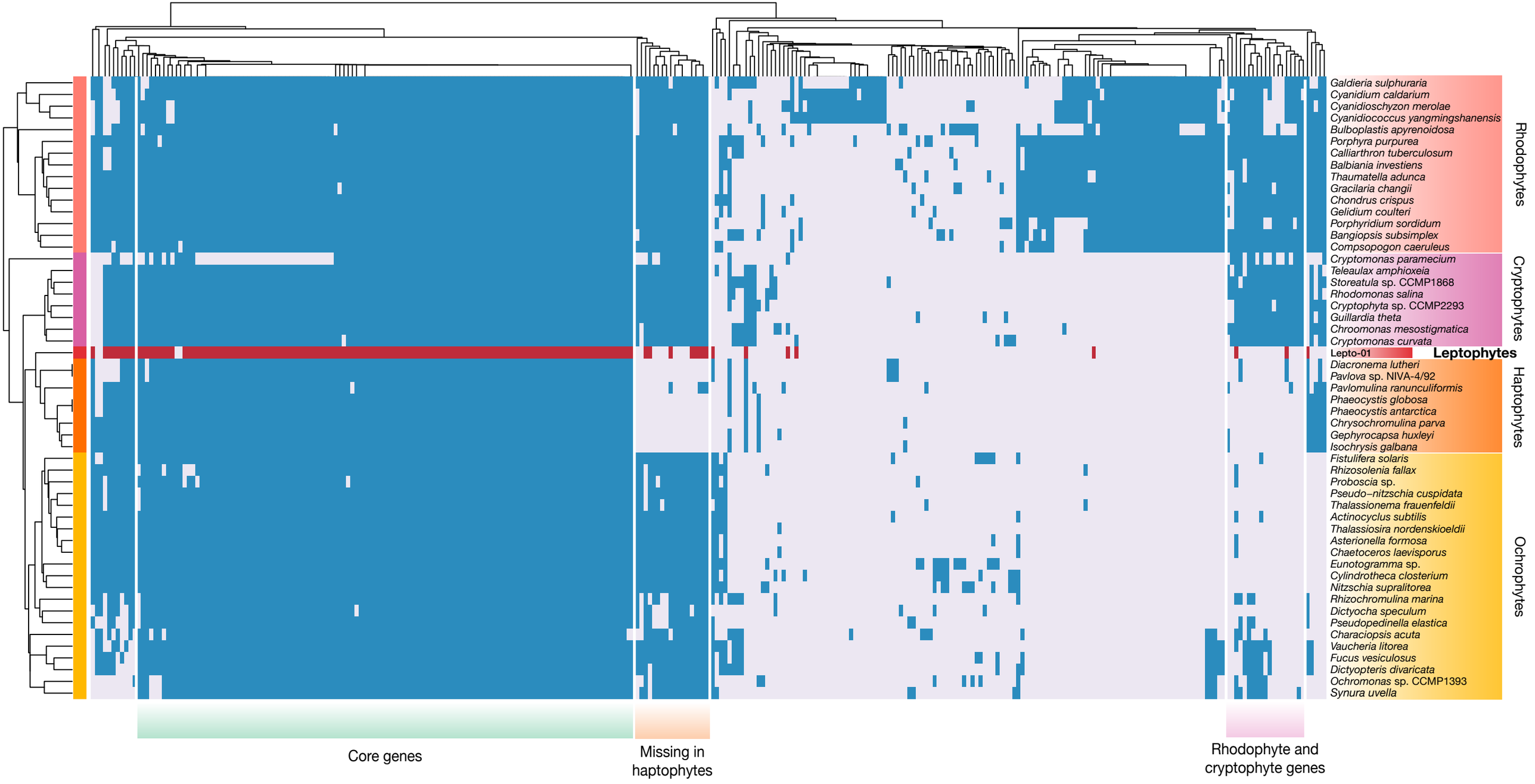
69. Pierella Karlusich, J. J. *et al.* A robust approach to estimate relative phytoplankton cell abundances from metagenomes. *Mol. Ecol. Resour.* **23**, 16–40.

70. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

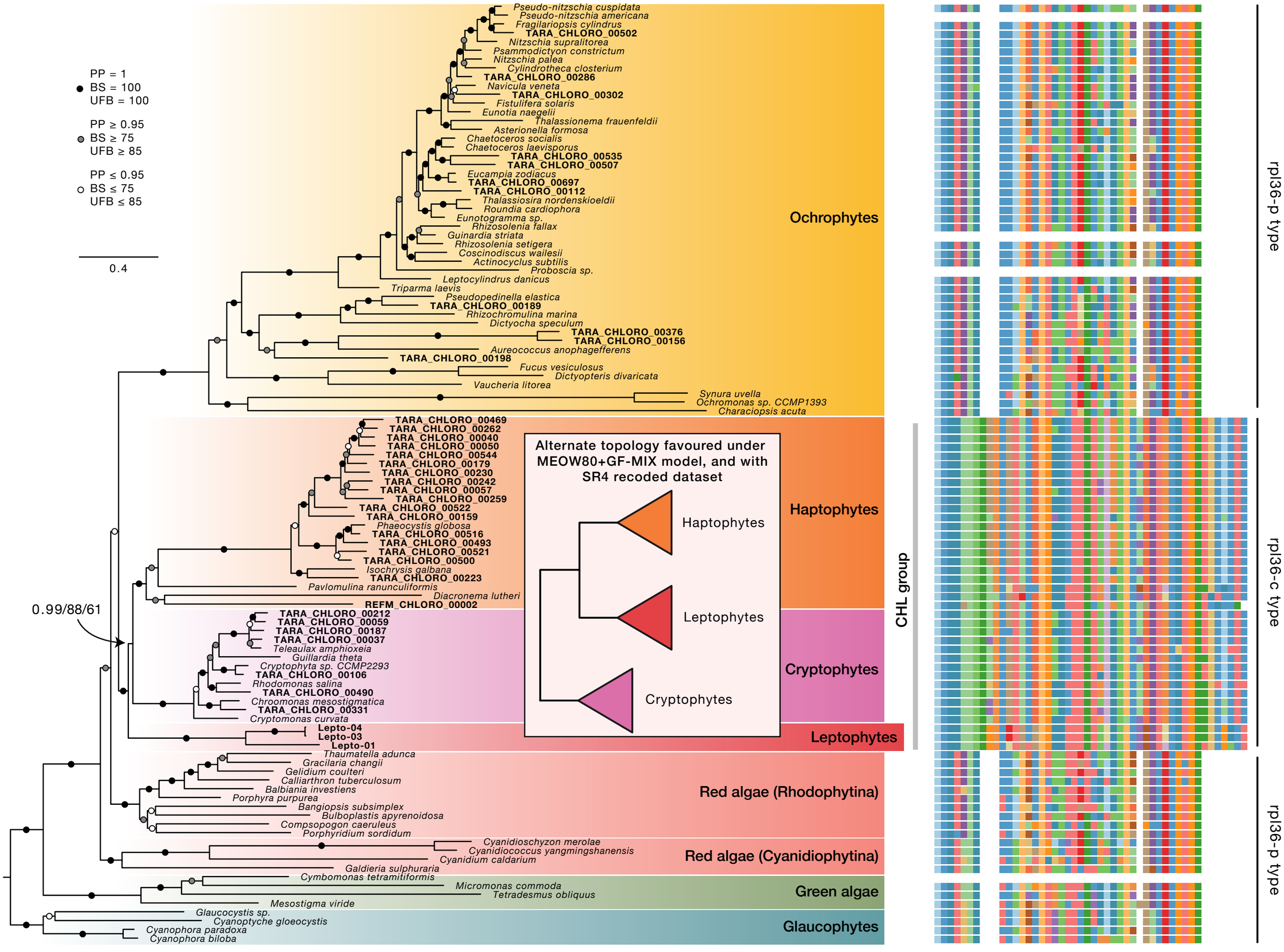
71. Guillou, L. *et al.* The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* **41**, D597–D604 (2013).

72. Zavadska, D., Henry, N., Auladell, A., Berney, C. & Richter, D. J. Diverse patterns of correspondence between protist metabarcodes and protist metagenome-assembled genomes. *PLOS ONE* **19**, e0303697 (2024).


Pelagophytes + Dictyochophytes
Diatoms
Chrysophytes + Brown algae
Colpodellids
Haptophytes
Leptophytes
Cryptophytes
Red algae
Dinophytes
Chlorarachniophytes
Chlorophytes
Euglenids
Prasinodermophytes
Streptophytes
Glaucophytes
Cyanobacteria + Paulinella
1
2
3
0.4
1 Main Group
Ochrophytes
Cryptophytes
Haptophytes
Leptophytes
Red algae
Viridiplantae and secondary green lineages
Glaucophytes
Cyanobacteria
2 Type
MAG
Reference
3 Size fraction
pico-nano: 0.2-5 µm
micro: 20-180 µm
nano: 3-20 µm
meso: 180-2,000 µm


Galdieria sulphuraria
Cyanidium caldarium
Cyanidioschyzon merolae
Cyanidiococcus yangmingshanensis
Bulboplastis apyrenoidosa
Porphyra purpurea
Calliarthron tuberculosum
Balbiania investiens
Thaumatella adunca
Gracilaria changii
Chondrus crispus
Gelidium coulteri
Porphyridium sordidum
Bangiopsis subsimplex
Compsopogon caeruleus
Rhodophytes
Cryptomonas paramecium
Teleaulax amphioxeia
Storeatula sp. CCMP1868
Rhodomonas salina
Cryptophyta sp. CCMP2293
Guillardia theta
Chroomonas mesostigmatica
Cryptomonas curvata
Cryptophytes
Lepto-01
Leptophytes
Diacronema lutheri
Pavlova sp. NIVA-4/92
Pavlomulina ranunculiformis
Phaeocystis globosa
Phaeocystis antarctica
Chrysochromulina parva
Gephyrocapsa huxleyi
Isochrysis galbana
Haptophytes
Fistulifera solaris
Rhizosolenia fallax
Proboscia sp.
Pseudo–nitzschia cuspidata
Thalassionema frauenfeldii
Actinocyclus subtilis
Thalassiosira nordenskioeldii
Asterionella formosa
Chaetoceros laevisporus
Eunotogramma sp.
Cylindrotheca closterium
Nitzschia supralitorea
Rhizochromulina marina
Dictyocha speculum
Pseudopedinella elastica
Characiopsis acuta
Vaucheria litorea
Fucus vesiculosus
Dictyopteris divaricata
Ochromonas sp. CCMP1393
Synura uvella
Ochrophytes
Core genes
Missing in haptophytes
Rhodophyte and cryptophyte genes

a



c

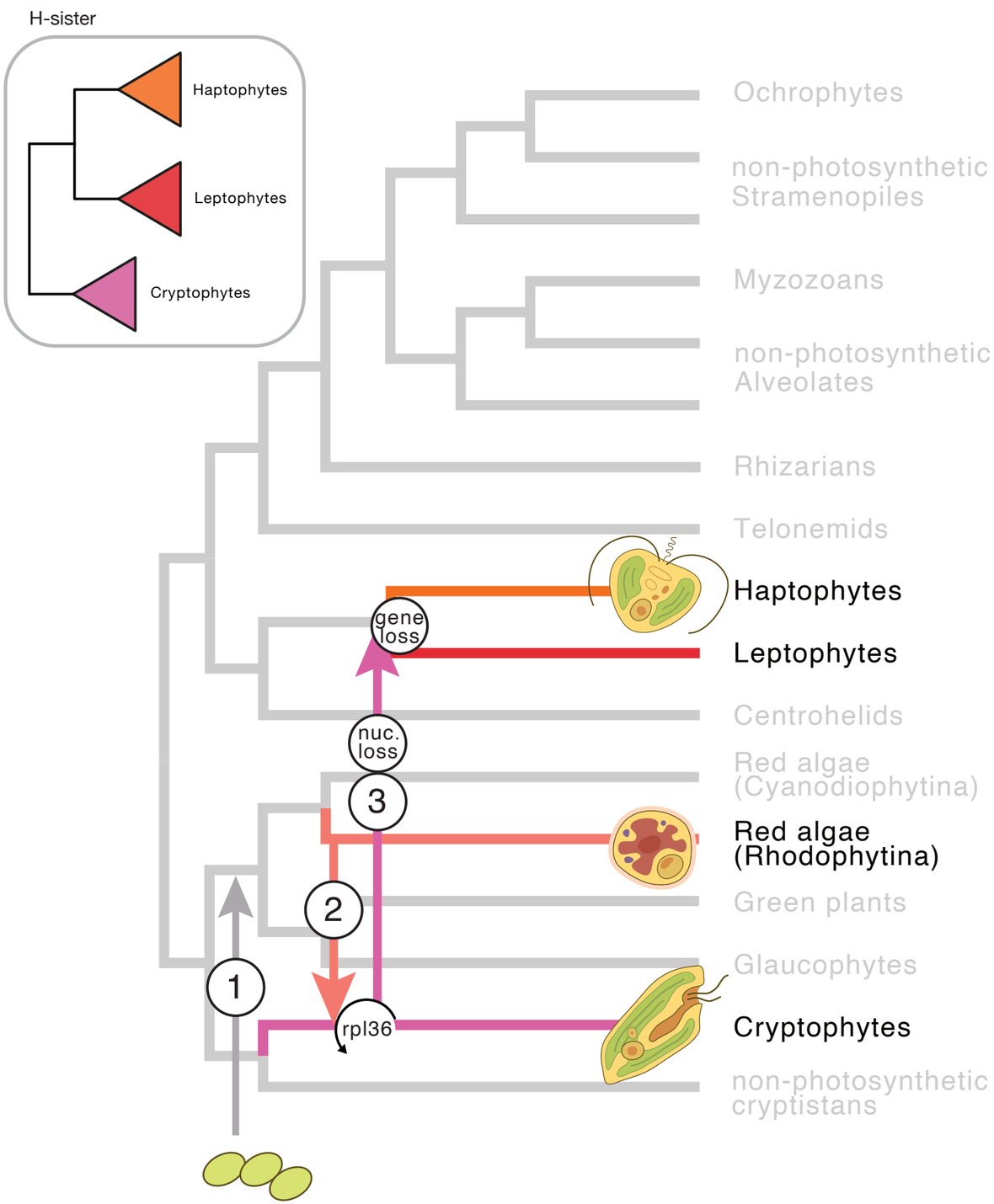
ML analyses using fixed topology trees					
Dataset	Untreated				Stationary trimmed
Model	GTR-CAT-PMSF	LG-MEOW80	LG-MEOW80 + GHOST (8 cat)	LG-MEOW80 + GF-MIX †	LG-MEOW80
Topology					
Leptophytes sister to haptophytes and cryptophytes	**-931909.236**	**-1056271.087**	**-1045458.696**	-1055620.569 *	**-592787.090**
Leptophytes sister to haptophytes	-931920.074 *	-1056274.766	-1045464.078 *	**-1055586.193**	-592793.164 *
Leptophytes sister to cryptophytes	-931924.826 **	-1056273.024	-1045463.152 *	-1055602.069 *	-592795.016 *

* Rejected by Bonferroni corrected χ2 test

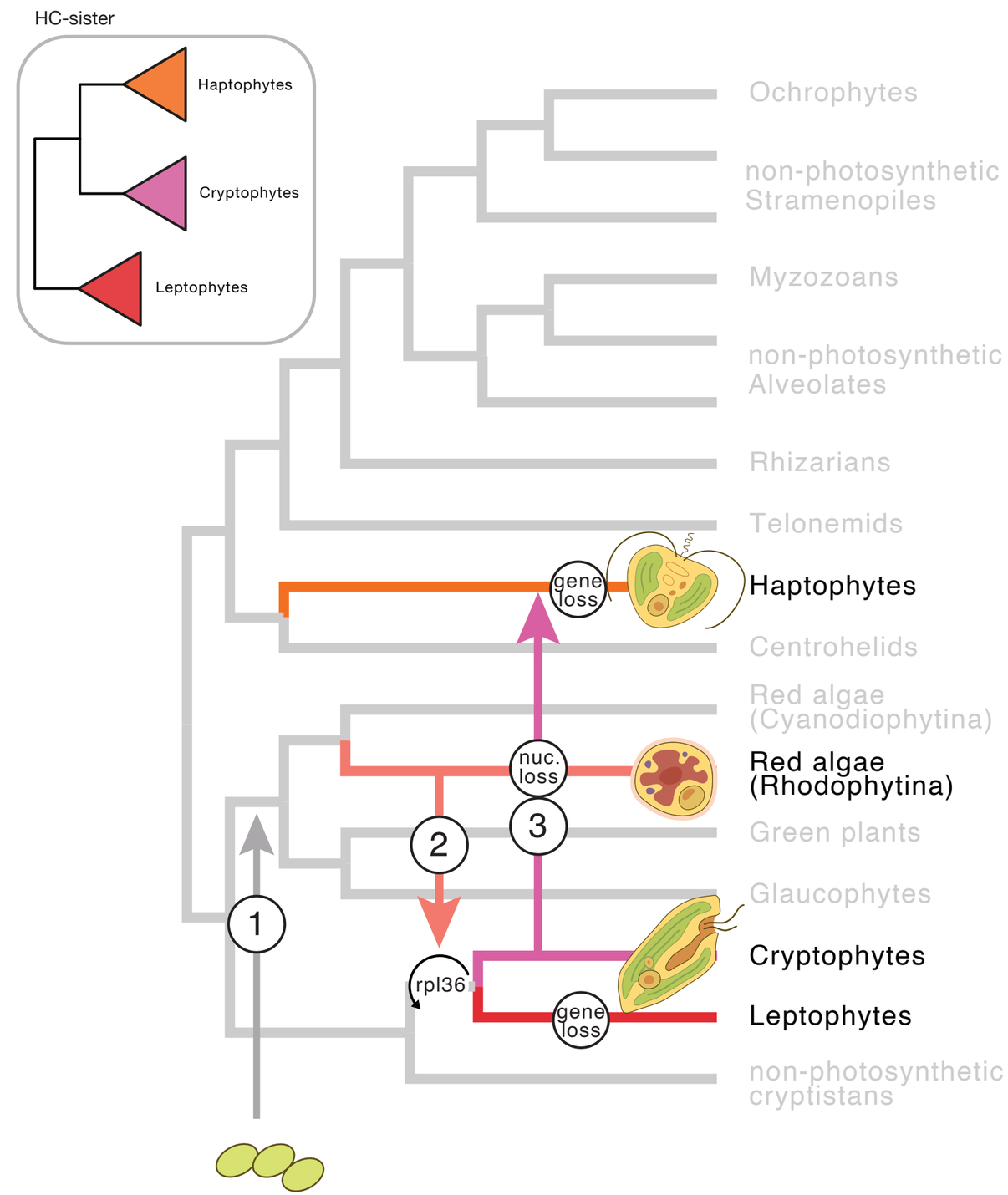
** Rejected by Bonferroni corrected χ2 test and AU test

† AU test not possible under GF-MIX model

a. New deep-branching haptista lineage



b. New deep-branching cryptista lineage



rpl36 rpl36 replacement

nuc. loss nucleomorph loss

gene loss gene loss from plastid genome