nature portfolio

Peer Review File

Global Metagenomics Reveals Plastid Diversity and Unexplored Algal Lineages

Corresponding Author: Dr Bikash Shrestha

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Shrestha et al., surveyed environmental metagenomes to retrieve novel plastid genomes that could shed light on algal diversity and evolution. They recovered a large number of plastid genomes, many of which are relatively novel. The authors also focused on cyanobacterial diversity, considerably expanding their known genome diversity in databases. They then inferred large-scale phylogenetic trees to reconstruct the origins and evolutionary history of plastids in eukaryotes.

Increasing phylogenetic diversity of prokaryotic lineages has become a common practice, but doing this for organelles genomes had not yet been done despite being the obvious next step to many of us for years. Shrestha et al., and another preprint by Jamy et al., do this for plastid genomes.

Compared to Jamy et al., the present study has some strengths, namely, the surveying of metagenomes from more diverse environments, and the re-examination of the phylogenetic origins of other plastids, such as secondary green plastids and the chromatophores of Paulinella chromatophore. In the process, the authors discovered some new ptDNAs most closely related to the secondary green ptDNAs of Lepidodinium chlorophorum. The authors may want to do a better job in highlighting these differences from the study of Jamy et al., which has its own strengths.

Perhaps the most glaring deficiency of this study is the use of simpler site-homogenous evolutionary models that do not properly model the complexities of protein evolution, especially at these vast phylogenetic distances. The reason for this is puzzling as many of these models are readily available in IQ-TREE and the authors surely have access to vast computational resources. I can only speculate that this was not done—but are probably and hopefully are currently running—because of pressure to get this manuscript out in light of the competing preprint. This is an important point, because the use of site-homogenous models considerably decreases the confidence in many of the phylogenetic conclusions made. Despite this, it is interesting to note that a considerable increase in taxon sampling improves phylogenetic inference in the absence of more complex and better-fitting models. Nonetheless, the use of these better-fitting models that account for compositional heterogeneity across sites is the standard for these deep phylogenetic questions and undoubtedly improves phylogenetic reconstruction.

For example, one of the major conclusions is that secondary red plastids have two separate origins from within red algae, the Cryptophyta and Haptophyta from the clade that unites the Eurhodophytina and Proteorhodophytina, and the Ochrophyta from the Cyanidiophytina. This conclusion cannot be trusted as the analyses were not conducted with best-fitting heterogeneous models. This pattern has been found before, but the use of better models often recovers secondary red plastids as monophyletic (e.g., https://doi.org/10.1016/j.cub.2017.04.054). Separate origins would also be at odds with the phenotypic features that unite secondary red (chromist) plastids: chlorophyll c and the SELMA complex as the protein translocase of the periplastid membrane: https://doi.org/10.1146/annurev-ecolsys-110411-160320

The fact that no ptMAGs were recovered for glaucophytes, for instance, suggests that metagenomics data is limited when it comes to some rare or low-abundance microorganisms. This thus leaves plenty of room for future discoveries.

The authors often report ranges of ptMAGs but many of these are likely incomplete. It is a better idea to report estimated ptDNA sizes. For this, an estimate of completeness of ptMAGs is needed. The authors can come up with sets of lineage-defining ptDNA markers that allow to assess for ptDNA completeness for a group.

Minor issues:

L55-56: Mention new additions to the Archaeplastida, the Rhodelphidia and Picozoa: https://doi.org/10.1038/s41586-019-1398-6 and https://doi.org/10.1038/s41467-021-26918-0
L70: Referring to dinoflagellates as a whole in imprecise here as dinoflagellates ancestrally have a peridinin plastid of red algal origin.
L74: Is Colopodellida used here in a way that it encompasses chromerids?
L78-79: The argument that multiple plastid losses are improbable does not rest on phylogenomic analyses, but mostly on the very few observations of complete plastid loss that have been reported.
L80: The Rhodoplex hypothesis is just one of many hypotheses that explain the distribution of chromist/chromalveolate plastids. For example: 10.1016/j.tree.2008.11.003 and 10.1111/j.1529-8817.2008.00559.x
L305-310: The name PCC-6307 is used for the most inclusive clade and another that excludes RSCF101 and WH5701. This is confusing and should be clarified.
L386: Which phylum or subphylum of the Rhodophyta does this species belong to?
L427-428: This statement is misleading. BIN141 is not sister to the red algae, but to a clade formed by Cryptophyta and Haptophyta.
L477-478: What is exactly meant by “the authors were unable to identify the host of leptophytes.”? I presume they refer to the inability of finding host nuclear genome sequences in the same samples.

Reviewer #2

(Remarks to the Author)
This is a beautiful and well-written study that not only expands the known plastid diversity and challenges the notion of a single origin for secondary red algal plastids, but also contributes a wealth of new plastid metagenome-assembled genomes (ptMAGs), including one that branch before the last common plastid ancestor of Cryptophyta and Haptophyta. The analyses are clearly presented, supported by high-quality figures and extensive supplementary material.

However, I would like to raise several points for clarification and improvement:

The manuscript frequently refers to the discovery of "novel" lineages (e.g., “reveals previously unidentified groups” or “likely represent undescribed algal lineages”). While the plastid genomes may indeed be newly sequenced, many of these lineages could correspond to already described taxa that simply lack plastid genome data. For instance, the authors the claim that six Streptophyta MAGs form a monophyletic clade sister to charophyte algae and represent “new members of class Zygnematophyceae”. By the way, this claim is not fully supported by the phylogeny in Fig. 2, as only three MAGs form a clade sister to Cylindrocystis and Zygnema. However, it is highly probable that this group in fact represents a known lineage with yet not sequened plastid genomes, such as Serritaeniales or other deep lineage within Zygnematales. It would be more accurate to describe these as “clades without reference plastomes” (as done appropriately on line 414), rather than asserting novelty.

While most phylogenetic relationships are congruent with previous studies, a few placements raise concerns and should be checked for potential misannotations or methodological artifacts:
- Interfilum terricola appears sister to Koliella, which is unexpected as it should fall within Streptophyta.
- Monoraphidium dybowskii is placed within Chlamydomonadales, though it is typically classified in Sphaeropleales.
- Trebouxiophyceae is not recovered as monophyletic, with four distinct lineages (Chlorellales, Helicosporidium, Stichococcus, and others). This is consistent with previous findings, but should be acknowledged in the text, especially in light of the statement on line 384 (“forming a distinct clade within Trebouxiophyceae”). Stichococcus might be worth highlighting due to its proximity to Chlorarachniophyceae.

Minor Comments and Corrections:

- Lines 50–65: The two primary endosymbioses are well described. However, why were other cases where cyanobacteria function as plastids not mentioned? For example, diazoplasts in Epithemia or nitroplasts in Braarudosphaera bigelowii could provide additional context for the diversity of plastid-like organelles.
- Line 70: Please revise to “and some dinoflagellates acquired secondary plastids.”
- Line 311: The reference to Fig. 1d should be corrected to Fig. 1c.
- Supplementary Figs. S6, S7: The label “Synurophyceae” should be corrected to “Chrysophyceae,” as Synurophyceae is nested within the latter.

Finally, I would like to mention that as a reviewer without direct experience in generating and filtering MAGs, I was unable to critically assess the methodology in this area. It would be helpful if the authors could provide a brief explanation or justification of their filtering criteria and quality thresholds for ptMAG inclusion.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have satisfactorily addressed my concerns.

I think that it is very important that the discussion about the evolution of SELMA and chlorophyll c is included in the final manuscript as this is a major conclusion of the study and a major point of debate in the field. (Note that even though I think the arguments made by the authors are valid, I think the discussion can be more precise and done more carefully.)

Reviewer #2

(Remarks to the Author)
I am very satisfied with the revision and recommend accepting the manuscript.

Reviewer #1 (Remarks to the Author):

Shrestha et al., surveyed environmental metagenomes to retrieve novel plastid genomes that could shed light on algal diversity and evolution. They recovered a large number of plastid genomes, many of which are relatively novel. The authors also focused on cyanobacterial diversity, considerably expanding their known genome diversity in databases. They then inferred large-scale phylogenetic trees to reconstruct the origins and evolutionary history of plastids in eukaryotes.

Increasing phylogenetic diversity of prokaryotic lineages has become a common practice, but doing this for organelles genomes had not yet been done despite being the obvious next step to many of us for years. Shrestha et al., and another preprint by Jamy et al., do this for plastid genomes.

Response: We appreciate your thoughtful comments and feedback, which have helped us improve the quality of our work. Your expertise and insights have been invaluable to us. Below, you will find our responses and resulting edits to the manuscript. Page and line numbers given below refer to the version of the edited manuscript **without** track changes annotated.

Compared to Jamy et al., the present study has some strengths, namely, the surveying of meta genomes from more diverse environments, and the re-examination of the phylogenetic origins of other plastids, such as secondary green plastids and the chromatophores of Paulinella chromatophore. In the process, the authors discovered some new ptDNAs most closely related to the secondary green ptDNAs of Lepidodinium chlorophorum. The authors may want to do a better job in highlighting these differences from the study of Jamy et al., which has its own strengths.

Response: We have now added the following text to the manuscript to better highlight differences between the two studies.

Original: “Our findings were further corroborated by a recent preprint by Jamy et al. (2025), which describes a new algal group based on plastid sequences referred to as “Leptophytes” and places them deep within the algal tree of life.”

Modified (Page 17, Lines 538-549): “Our results were further corroborated by findings described in a recent preprint by Jamy et al. (2025)[81], which identified a new algal group “Leptophytes” based on plastid sequences, placing it deep within the algal tree of life. However, Jamy et al. (2025) focused exclusively on the identification of plastid sequences from the oceanic pelagic zone.”

Perhaps the most glaring deficiency of this study is the use of simpler site-homogenous evolutionary models that do not properly model the complexities of protein evolution, especially at these vast phylogenetic distances. The reason for this is puzzling as many of these models are readily available in IQ-TREE and the authors surely have access to vast computational resources. I can only speculate that this was not done—but are probably and hopefully are currently running—because of pressure to get this manuscript out in light of the competing preprint. This is an important point, because the use of site-homogenous models considerably

decreases the confidence in many of the phylogenetic conclusions made. Despite this, it is interesting to note that a considerable increase in taxon sampling improves phylogenetic inference in the absence of more complex and better-fitting models. Nonetheless, the use of these better-fitting models that account for compositional heterogeneity across sites is the standard for these deep phylogenetic questions and undoubtedly improves phylogenetic reconstruction.

For example, one of the major conclusions is that secondary red plastids have two separate origins from within red algae, the Cryptophyta and Haptophyta from the clade that unites the Eurhodophytina and Proteorhodophytina, and the Ochrophyta from the Cyanidiophytina. This conclusion cannot be trusted as the analyses were not conducted with best-fitting heterogeneous models. This pattern has been found before, but the use of better models often recovers secondary red plastids as monophyletic (e.g.,https://doi.org/10.1016/j.cub.2017.04.054).

Response: The study by Munoz-Gomez et al. 2017, (https://doi.org/10.1016/j.cub.2017.04.054), did not mention whether a best-fit model analysis was carried out. However, we agree the most parameter-rich models (LG+C60+F+R6 and LG+PMSF+F+R6) used by Munoz-Gomez et al. 2017 did recover a single origin of secondary red plastids, with Haptophyta, Cryptophyta, and Ochrophyta forming a sister clade to red algae. Notably, the authors also recovered an alternate topology (Figs. S13-S18), in which Ochrophyta plastids shared ancestors with plastids from thermoacidophilic red algae (Cyanidiophytina), while Haptophyta and Cryptophyta plastids are sister to those of mesophilic red algae (subphyla Proteorhodophytina and Eurhodophytina) plastids, suggesting the possibility of two origins of secondary red plastids. In that study, this alternate topology was supported by several site-heterogeneous models, such as LG4X+F+R5, LG+C20+F+R5, LG+C40+F+R5 and including LG+C60+F+R5 and LG+PMSF1+F+R5, which differ only in the number of rate categories compared to the most parameter-rich models.

We acknowledge the limitation of site-homogenous evolutionary models in resolving deep phylogenetic relationships. We have now included in the manuscript additional analyses using a range of complex site-heterogeneous models implemented in IQ-TREE (please see the methods under “Estimating plastid phylogeny” section, **Page 8, Lines 284-297**). Briefly, we reduced the original 537-taxon dataset to 157-taxon subset for best-fit model selection. We then inferred maximum likelihood phylogenies using the best site-heterogenous model as well as several other site-heterogeneous models including two complex models (LG+PMSF+F+R6 and LG+C60+F+R6) used by Munoz-Gomez et al. 2017, (https://doi.org/10.1016/j.cub.2017.04.054).

Maximum likelihood trees inferred using all site-heterogenous models were generally congruent with the topology inferred through the site-homogeneous model, specifically supporting two separate origins of secondary red plastids: (i) plastids of Haptophyta and Cryptophyta derived from red algae uniting subphyla Proteorhodophytina and Eurhodophytina, and (ii) plastids of Ochrophyta derived from Cyanidiophytina red algae (Supplementary Figs. S6-S10). These

additional analyses yielded results consistent with our original conclusions, suggesting that the use of complex models doesn't substantially change the main phylogenetic relationships.

A major discrepancy among the different models used concerns the phylogenetic position of Glaucophyta plastids relative to plastids from other photosynthetic eukaryotes. We recovered strongly supported trees in which Glaucophyta were either (i) placed as sister to red algae and secondary red algae plastids (Supplementary Figs. S8-S10), or (ii) recovered as an early branching clade sister to plastids from all other photosynthetic eukaryotes (Supplementary Figs. S6 & S7). To reflect this, we have revised the Results section (**Page 11-12, Lines 384-391**) as follows:

"Under the site-homogeneous model, Glaucophyta plastids were strongly supported as a sister group to the red-algal plastids (UFBoot >95%; Fig. 2, Supplementary Figs. S4 & S5). In contrast, site-heterogeneous models yielded two alternative topologies: some placed Glaucophyta as the earliest branching lineage, sister to all other plastids (Supplementary Figs. S6 & S7), while others recovered it as sister to red-algal plastids (Supplementary Figs. S8-S10). This model-dependent pattern indicates that the position of Glaucophyta remains unresolved, consistent with the ambiguity present in previous phylogenomic studies[70]."

Separate origins would also be at odds with the phenotypic features that unite secondary red (chromist) plastids: chlorophyll c and the SELMA complex as the protein translocase of the periplastid membrane: https://doi.org/10.1146/annurev-ecolsys-110411-160320.

Response: Since protein import across the periplastidial membrane is considered to have derived from the preexisting symbiont-derived ER-associated protein degradation (ERAD) system (Stork et al, 2012, https://doi.org/10.1128/ec.00183-12), this raises the question of how many times this system was independently derived and reestablished in lineages with secondary red plastids. If we consider a single red algal endosymbiont and subsequent distribution of that endosymbiont through tertiary and quaternary endosymbiosis, then the SELMA complex would have to be reestablished (i.e., gene transfers from the endosymbiont to host) multiple times during/following these higher order endosymbioses. In our model, two independent adaptations and two reestablishments are required (in haptophytes and in apicomplexa), whereas under the single origin model, a single adaptation followed by multiple reestablishments are required, and both models are equally parsimonious, as suggested by Pietluch et al. 2024 (https://doi.org/10.1093/gbe/evae192).

Unexpectedly, given the widely accepted logic supported by phenotypic features that unite secondary red (chromist) plastids, molecular phylogenetics of the SELMA complex and chlorophyll *c* synthase have not provided definitive evidence with respect to the events that led to photosynthesis in the CASH lineages. Ponce-Toledo, et al. (https://doi.org/10.1093/molbev/msaf167) have recently published a comprehensive collection of single-gene trees for individual components of the SELMA complex. The authors predict mosaic evolutionary origins for components of SELMA, such that in some algal lineages a particular SELMA protein can be traced to red algal origins but in other lineages the analogous SELMA

protein appears to be the result of a duplication event involving the ancestral non-endosymbiont gene. Therefore, the SELMA phylogenies in Ponce-Toleda, et al. provide support that components of the SELMA complex have evolved independently in different lineages. Indeed, to reconcile the phylogenies with the hypothesis that SELMA evolved once and then was spread through endosymbiosis, the authors had to evoke scenarios where the red-alga-derived gene must have been lost and then replaced with a duplicated copy of the host's nuclear gene. Also, in the phylogenies presented in Ponce-Toleda, et al., the SELMA complex proteins from the subphylum Cyanidiophytina (of which there are only two taxa) are never recovered in a monophyletic clade, suggesting that additional analyses are required to interrogate evolutionary relationships with red algae. Nevertheless, Ponce-Toledo, et al. do find that SELMA proteins from haptophytes, stramenopiles and apicomplexa (but not cryptophytes) are typically monophyletic, suggesting evolution of SELMA components in a common ancestor, endosymbiotic gene transfer, lateral gene transfer in an ancestor, or convergent evolution.

Several groups have also performed phylogenetic analyses with the recently discovered chlorophyll *c* synthase gene *CHLC* (Jiang et al. (DOI: 10.1126/science.adg7921) and Jinkerson, et al. (DOI: 10.1016/j.cub.2023.12.068)), resulting in nearly identical reconstructions. Based on these analyses, the enzyme responsible for chlorophyll *c* synthesis has evolved independently at least two times, since entire algal lineages that are known to depend on chlorophyll *c* are missing *CHLC* (i.e., a second yet-to-be-discovered chlorophyll *c* synthase gene has evolved in those algae). Additionally, the phylogenies of CHLC cannot rule out the possibility of lateral gene transfer among ancestors or "cryptic" endosymbiotic gene transfer, with Jiang et al. speculating that the CHLC phylogeny supports an ochrophyte to cryptophyte transfer.

As such, the SELMA and chlorophyll *c* synthase phylogenies highlight the possibility of more complex evolutionary paths for the gain of these functions and do not rule out the possibility of ancient lateral gene transfer events or "cryptic" endosymbiotic gene transfer events that are harmonious with shared phenotypes and separate endosymbiotic events that resulted in the extant plastid genomes.

The fact that no ptMAGs were recovered for glaucophytes, for instance, suggests that metagenomics data is limited when it comes to some rare or low-abundance microorganisms. This thus leaves plenty of room for future discoveries.

Response: We absolutely agree. Indeed, this and other work highlight the need to discover "missing links" in rhodophyta and other lineages to better resolve the identity of the ancient endosymbiont(s). We have added the following comment to the manuscript:

Modified (Page 14, Lines 432-438): "We did not recover any ptMAGs representing Glaucophyta and Prasinodermaphyta, and both phyla were solely represented by RefSeq plastomes. We recovered a single new ptMAG from red algae, suggesting that the available metagenomics data is limited with respect to the plastids of rare or low-abundance algae. Additional deeper sequencing of diverse environmental samples and the identification of new

ptMAGs could help to resolve ancient relationships, such as the endosymbiotic events involving red agal endosymbiont(s)."

The authors often report ranges of ptMAGs but many of these are likely incomplete. It is a better idea to report estimated ptDNA sizes. For this, an estimate of completeness of ptMAGsis needed. The authors can come up with sets of lineage-defining ptDNA markers that allow to assess for ptDNA completeness for a group.

Response: To provide an estimated measure of ptMAGs completeness, we identified lineage specific plastid panorthologs. First, we performed OrthoFinder analysis using both ptMAGs and the NCBI reference plastomes. Next, we identified core proteins for each eukaryotic phylum that were present in 90% of reference taxa and then assessed the presence of these core proteins in each ptMAG. The completeness estimates and the ptMAG size are now provided in Supplementary Tables S4 & S5, and we have described the completeness in the main text. Additionally, the set of core proteins for both ptMAGs and reference species are presented in Supplementary Tables S8-S13. The corresponding methods are also now added to the manuscript as follows:

Modified (Page 8, Lines 303-310): "To assess plastid completeness, we conducted OrthoFinder v2.5.5[63] analysis using a dataset comprising dereplicated ptMAGs and NCBI reference plastomes that were also used for phylogenetic inference among photosynthetic eukaryotes. OrthoFinder results were parsed to identify core proteins for each phylum (e.g., Streptophyta, Chlorophyta, Rhodophyta, etc.), defined as orthologs present in 90% of the NCBI reference plastomes. For each ptMAG, we calculated the proportion of these core proteins detected for the corresponding phylum, thereby providing a lineage-specific measure of plastid completeness."

Minor issues:

L55-56: Mention new additions to the Archaeplastida, the Rhodelphidia and Picozoa: https://doi.org/10.1038/s41586-019-1398-6 and https://doi.org/10.1038/s41467-021-26918-0

Response: Thank you for pointing this out. We have modified the sentence as follows:

Original: "The first event, occurring approximately 1.5 billion years ago, involved a eukaryote and β-cyanobacterium endosymbiont, leading to the formation of plastids in Archaeplastida, a group that includes green algae and land plants (Viridiplantae), red algae (Rhodophyta), and glaucophytes (Glaucophyta)[1–5]."

Modified (Page 2, Lines 57-61): "The first event, occurring approximately 1.5 billion years ago, involved a eukaryote and β-cyanobacterium endosymbiont, leading to the formation of plastids in Archaeplastida. This group comprises green algae and land plants (Viridiplantae), glaucophytes (Glaucophyta), and red algae (Rhodophyta), as well as non-photosynthetic members such as Rhodelphia and Picozoa, which are closely related to Rhodophyta[3–9]."

L70: Referring to dinoflagellates as a whole is imprecise here as dinoflagellates ancestrally have a peridinin plastid of red algal origin.

Response: We have now changed the line as suggested:

Modified (Page 2, Line 76): "and **some** dinoflagellates acquired secondary plastids through the endosymbiosis of distinct green algae,"

L74: Is Colopodellida used here in a way that it encompasses chromerids?

Response: We have changed "Colpodellida" to "Chrompodellids (Chromerids + Colpodellids)" to avoid confusion. We have modified the sentence as follows:

Modified (Pages 2-3 Line 80-83): "However, the evolutionary origins and spread of red alga-derived plastids among algal groups, including Cryptophyta, Haptophyta, Ochrophyta (photosynthetic Stramenopila), and Alveolates (comprising **Chrompodellids**, Apicomplexa, and dinoflagellates), collectively known as CASH lineages, remain contentious."

L78-79: The argument that multiple plastid losses are improbable does not rest on phylogenomic analyses, but mostly on the very few observations of complete plastid loss that have been reported.

Response: Please see edits to this section in the following response.

L80: The Rhodoplex hypothesis is just one of many hypotheses that explain the distribution of chromist/chromalveolate plastids. For example: 10.1016/j.tree.2008.11.003 and 10.1111/j.1529-8817.2008.00559.x

Response: To avoid confusion and highlight the diversity of alternative hypotheses, we have made the following changes:

Original: "The Chromalveolate hypothesis, first proposed by Cavalier-Smith[16], suggests that the CASH lineages share a common algal ancestor, with multiple plastid losses to explain non-photosynthetic protist lineages. However, phylogenomic studies imply that multiple plastid losses would be highly improbable[17,18]. An alternative model, the Rhodoplex hypothesis, proposes that secondary/complex red algal plastids originated within one CASH lineage (likely Cryptophyta) and subsequently spread via tertiary or higher order endosymbiosis (serial endosymbiosis[18–21]). Despite these competing hypotheses, empirical evidence supporting them remains limited[22]."

Modified (Page 3, Lines 82-91): "The Chromalveolate hypothesis, first proposed by Cavalier-Smith[20], suggests that the CASH lineages share a common ancestor bearing secondary plastids, which would require multiple plastid losses to explain the existence of non-photosynthetic protist lineages. However, there is little empirical evidence to support such widespread plastid loss[21,22]. Several alternative hypotheses have been proposed to explain the distribution of plastids in CASH lineages. For example, the Rhodoplex hypothesis posits that secondary/complex red algal plastids originated within one CASH lineage (likely Cryptophyta)

and subsequently spread via tertiary or higher order endosymbiosis (serial endosymbiosis[22–26]). Despite these competing hypotheses, empirical evidence supporting them remains limited[27]."

L305-310: The name PCC-6307 is used for the most inclusive clade and another that excludes RSCF101 and WH5701. This is confusing and should be clarified.

Response: In the most recent GTDB taxonomic classifications, this order is called "PCC-6307". We now see how this is confusing to readers as this is also a strain designation. To address this, we have edited Figure 1c to distinguish the order PCC-6307 from the clade by renaming the clade "Remaining Cyanobiaceae", indicating that it represents the rest of the PCC-6307 species. We have also modified the sentence as follows:

Original: "However, while *Synechococcus* sp. WH5701 was the assumed closest extant relative[12], our results place the *Paulinella* plastid within the cyanobacterial order PCC-6307, after the diversification of *Synechococcus* sp. RSCF101 (Fig. 1c, Supplementary Fig. S3). The *Synechococcus* sp. RSCF101 shared ~79% nucleotide identity with *Synechococcus* sp. WH5701, and the two *Synechococcus* species belonged to distinct phylogenetic clades within the PCC-6307 (Fig. 1d, Supplemental Fig. S3)."

Modified (Page 11, Lines 358-368): "However, while *Synechococcus* sp. WH5701 was previously assumed to be the closest extant relative[16], our results place the *Paulinella* plastid after the diversification of *Synechococcus* sp. RSCF101 (Fig. 1c, Supplementary Fig. S3). The *Synechococcus* strains RSCF101 and WH5701 shared ~79% nucleotide identity but belonged to distinct phylogenetic subclades within the larger Synechococcales lineage (referred to as "order PCC-6307" in GTDB taxonomy; Fig. 1d, Supplemental Fig. S3)."

L386: Which phylum or subphylum of the Rhodophyta does this species belong to?

Response: We have modified the sentence:

Original: "A single novel rhodophyte ptMAG of size ~171 Kb was recovered as a sister to *Membranoptera platyphylla* (family Delesseriaceae) plastome…"

Modified (Page 15, Lines 452-454):" A single novel rhodophyte ptMAG (~171 Kb, 98% completeness) was recovered as a sister taxon to the *Membranoptera platyphylla* plastome, suggesting it belongs to the family Delesseriaceae, subphylum Eurhodophytina."

L427-428: This statement is misleading. BIN141 is not sister to the red algae, but to a clade formed by Cryptophyta and Haptophyta.

Response: We have modified the sentence to avoid confusion:

Original: "This ptMAG (IMGM3300009544_BIN141) is sister to mesophilic red algae and branches before the last common plastid ancestor of Cryptophyta and Haptophyta.""

Modified (Page 16, Lines 494-497): “This ptMAG (IMGM3300009544_BIN141) is phylogenetically affiliated with the clade comprising the plastids of Cryptophyta and Haptophyta, and together these three lineages form a monophyletic clade that is sister to mesophilic red algae.”

L477-478: What is exactly meant by “the authors were unable to identify the host of leptophytes.”? I presume they refer to the inability of finding host nuclear genome sequences in the same samples.

Response: Thank you for bringing to our attention this typo. Yes, we meant the nuclear genome of leptophytes, which would shed insight into the identity of this presumed new lineage of algae. We were unable to connect nuclear genomes identified for marine eukaryotes with the leptophyte plastid sequence, such as through sample co-occurrence. We have made the following changes to correct the typo:

Original: “However, an alternative topology where leptophyte plastids closer to haptophyte plastids than cryptophytes was also presented by authors creating uncertainty about their exact phylogenetic placement. Much like our ongoing efforts, the authors were unable to identify the host of leptophytes.”

Modified (Page 17, Line 544-548): “However, an alternative topology was also presented, suggesting leptophyte plastids branch closer to haptophyte plastids than cryptophytes, leading to uncertainty regarding their exact phylogenetic placement. In addition, the authors were unable to identify candidate nuclear genome sequences belonging to these algae.”

Reviewer #2 (Remarks to the Author):

This is a beautiful and well-written study that not only expands the known plastid diversity and challenges the notion of a single origin for secondary red algal plastids, but also contributes a wealth of new plastid metagenome-assembled genomes (ptMAGs), including one that branch before the last common plastid ancestor of Cryptophyta and Haptophyta. The analyses are clearly presented, supported by high-quality figures and extensive supplementary material.

Response: We appreciate the time and effort you dedicated to reviewing our manuscript and providing constructive comments and feedback. Your input has significantly enhanced the quality of our work, and we are grateful for your expertise and valuable insights. Below, we have outlined our responses and corresponding edits to the manuscript. Please note that the page and line numbers referenced below correspond to the edited manuscript **without** tracked changes.

However, I would like to raise several points for clarification and improvement:

The manuscript frequently refers to the discovery of "novel" lineages (e.g., “reveals previously unidentified groups” or “likely represent undescribed algal lineages”). While the plastid genomes may indeed be newly sequenced, many of these lineages could correspond to already

described taxa that simply lack plastid genome data. For instance, the authors the claim that six Streptophyta MAGs form a monophyletic clade sister to charophyte algae and represent “new members of class Zygnematophyceae”. By the way, this claim is not fully supported by the phylogeny in Fig. 2, as only three MAGs form a clade sister to Cylindrocystis and Zygnema. However, it is highly probable that this group in fact represents a known lineage with yet not sequened plastid genomes, such as Serritaeniales or other deep lineage within Zygnematales. It would be more accurate to describe these as “clades without reference plastomes” (as done appropriately on line 414), rather than asserting novelty.

Response: Yes, we agree that the only statements we can make are that many of the plastid sequences identified in our study are either identical, similar, or not similar to publicly available plastid sequences; we are also able to present predictions regarding the evolutionary relatedness between the ptMAGs identified here, sequenced cyanobacterial genomes, and sequenced plastid genomes isolated from known algae. As such, we should have been more considerate/conservative with the use of “novel” and related phrases. Given the historical shortage of algal genome sequencing, some of the ptMAGs could very well belong to described algae and described lineages that lack sequenced plastid genomes and would therefore be “invisible” in our analysis. We have now thoroughly revised the manuscript to remove insinuations about “previously unidentified groups” or “undescribed algal lineages”. When we use the phrase “novel ptMAG” we refer to sequences that have less than 95% average nucleotide identity with either align_fraction_ref ≥ 50% and align_fraction_query ≥ 50% (ptMAGs) or align_fraction_query ≥95% compared to all publicly available plastid sequences (see page 7, lines 250-259), and we are careful not to assume that the novel ptMAGs belong to yet-to-be described algae or algal lineages. The only time we keep this phrasing is if the phylogenies suggest the existence of an algal lineage that has been not described.

Regarding the ptMAGs within Zygnematophyceae, our dataset contains seven such ptMAGs, three ptMAGs sister to *Spirogyra maxima,* a ptMAG sister to *Cosmarium botrytis*, and three more ptMAGs sister to *Zygnema circumcarinatum* + *Cylindrocystis brebissonii*. While only three of these ptMAGs form a clade sister to *Zygnema* + *Cylindrocystis*, the remaining four are sister to other Zygnematophyceae members. We agree the original phrase may have caused confusion, so we have modified the sentence to avoid the confusion as follows:

Original: “Within the green lineages, assembly sizes for Streptophyta ptMAGs ranged from ~61 Kb to ~131 Kb, and ptMAGs (n=6) recovered from freshwater and deep subsurface formed a monophyletic clade sister to charophyte algae, suggesting they are new members of class Zygnematophyceae (Fig. 2).”

Modified (Page 14, Lines 441-445): “Within the green lineages, assembly sizes for Streptophyta ptMAGs (n=10) ranged from ~61 Kb to ~131 Kb, with completeness ranging from 39% to 96 % (90% median; Supplementary Table S4 & S5). Among these, seven ptMAGs from freshwater and deep subsurface habitats were associated with Zygnematophyceae algae, indicating they represent new plastid sequences belonging to this class (Fig. 2).”

While most phylogenetic relationships are congruent with previous studies, a few placements raise concerns and should be checked for potential misannotations or methodological artifacts:

- Interfilum terricola appears sister to Koliella, which is unexpected as it should fall within Streptophyta.

Response: Thank you for bringing this to our attention. We were able to trace the issue with this NCBI plastid sequence, and it seems to be a case of "Interfilum terricola" being used incorrectly by NCBI as a synonym for "Geminella terricola". The plastid genome of *Interfilum terricola* (NC_025542) used in our study was originally reported by Lemieux et al. (2014, doi:10.1186/s12862-014-0211-2), where it was referred as *Geminella terricola* (KM462881, Table 1). Both plastid genomes (NC_025542 and KM462881) are identical in size (187,843 bp) and pairwise alignment revealed 100% nucleotide identity. In Lemieux et al. (2014), phylogenomic analysis (Figure 1) placed *Geminella terricola* within the chlorophytes (Geminella clade), as an early-branching lineage sister to the rest of the core Trebouxiophyceae, which is consistent with the phylogenetic position of *Interfilum terricola* (NC_025542) in our study. Therefore, although NCBI lists this plastid genome as belonging to *Interfilum terricola*, the host is *Geminella terricola.* It is unknown why this *Geminella terricola* sequence was deposited or changed in NCBI with the name *Interfilum terricola*. Mikhailyuk, et al. 2008 (10.1111/j.1529-8817.2008.00606.x) did report that *Geminella terricola* strain SAG 2100 was mis-identified and is actually a streptophyte alga, renaming the isolate to *Interfilum terricola*. It appears, then, that there is some confusion that *Geminella terricola* and *Interfilum terricola* are synonyms, when only *Geminella terricola* strain SAG 2100 and *Interfilum terricola* are synonyms.
To avoid misleading readers, we have now used the name *Geminella terricola* in our figures and tables, since there is sufficient evidence (i.e., supported by Lemieux et al. 2014 and as we were able to trace the likely origin of the mis-labeling) to make the point that this plastome is mis-labeled in NCBI. We also added a note in the manuscript and in the Supplementary Table S4 that we renamed this accession.

We identified additional cases of potential mislabeling among NCBI plastomes, but do not necessarily have analogous breadcrumb trails. Therefore, as such and given the goals of this study and the conclusions that we are presenting, we saw it as impractical to manually check and re-curate all publicly available plastid genomes at this time (however, such "cleaning up" of the databases clearly needs to happen!). We did, however, ensure that potential/likely database misannotations did not lead to erroneous conclusions in the manuscript, and we did manually check publicly available reference plastome sequences used to infer a taxonomic inference for a metagenome-assembled plastid genome that we consider to be "novel".

In the revised manuscript, we have now added a paragraph that refers to database misannotation (new text pasted below). For example, a BLAST search of proteins from plastome NC_058752, assigned to *Grateloupia livida* (Rhodophyta, Florideophyceae), showed strong matches with proteins from NC_036804, which belong to *Dictyopteris divaricate* (Ochrophyta, Phaeophyceae); these two plastomes are 99.9% identical at the nucleotide level. Similarly, BLAST hits for plastome NC_058751, attributed to *Chorda filum* (Ochrophyta, Phaeophyceae), matched best with proteins from seed plants, specifically the Zosteraceae

species (marine monocots), and our phylogenetic analysis placed this plastome within Streptophyta. We believe these plastomes (and others) are mislabeled in the databases. These mis-labeled sequences did not affect our identification of novel plastome sequences and, as such, were not included in the phylogenetic reconstruction used to study the ptMAGs presented in Fig. 2.

We have added the following paragraph to the **Method section in the main text** (**Page 6 Lines 233-238)** as follows:
"During this process, we identified that the NCBI plastome for *Interfilum terricola* (NC_025542) is misclassified, with phylogenetic evidence supporting its reassignment to *Geminella terricola* (Chlorophyta). Accordingly, this plastome was renamed to *Geminella terricola* in our figures, and further details on this case as well as other potential plastome misclassifications are provided in the Supplementary Text."

And in the **Supplementary Text** (**Page 6-7, Lines 235-258**) as follows:

"Potentially misclassified plastomes in the NCBI database

In our study, we identified a few cases of potential mislabeling among plastomes in the NCBI database. Although, we note that these examples are not expected to be an exhaustive list of such database errors. The Interfilum terricola plastome (NC_025542) is classified in NCBI as belonging to a streptophyte alga, but our phylogenomic analysis placed this sequence as an early diverging taxon within Trebouxiphyceae, consistent with the position recovered by Lemieux et al. (2014)[28] for Geminella terricola (KM462881) within chlorophytes. The sequences for these two accessions share 100% nucleotide identity, and the mislabeling likely arose due to confusion caused by renaming of Geminella terricola strain SAG 2100 to Interfilum terricola by Mikhailyuk, et al. 2008[29]. Additional examples of likely mis-labeled plastomes include: (i) plastome NC_058752, attributed to Grateloupia livida (Rhodophyta, Florideophyceae), which shares 99.9% identity with plastome NC_036804 of Dictyopteris divaricate (Ochrophyta, Phaeophyceae), (ii) proteins of plastome NC_058751 of Chorda filum (Ochrophyta, Phaeophyceae), which matched best with proteins from seed plants, specifically Zosteraceae species (marine monocots), and our phylogenetic analysis placed this plastome within Streptophyta (results not shown), (iii) the Symbiochloris sp. SG-2018 plastome (NC_065389), listed as Trebouxiophyceae in NCBI, which was placed within Ulvophyceae in our preliminary analysis. This sequence is published as belonging to Symbiochlorum hainanensis (Ulvophyceae), with the NCBI accession ON645925 mis-labeled as Symbiochloris. Other than the Interfilum terricola plastome (NC_025542) that we corrected to Geminella terricola in figures and tables, these mis-labeled plastomes were not found to be sister to an identified ptMAG and were therefore excluded from the phylogeny presented in our study. These findings underscore the necessity for careful verification of taxonomic assignments when using publicly available plastome sequences."

- Monoraphidium dybowskii is placed within Chlamydomonadales, though it is typically classified in Sphaeropleales.
Response: Because of the unexpected placement of *Monoraphidium dybowskii* shown in Fig. S4 and because inclusion of this plastome did not provide information useful in the

determination of ptMAG relatedness to publicly available plastome sequenced during preliminary analyses, this plastome was removed before performing the MAG analysis represented in Fig. 2. Although this deposited chloroplast genome is labeled as belonging to Sphaeropleales in NCBI, we consistently observed that this plastid sequence is sister to *Haematococcus lacustris* within the Chlamydomonadales using the marker genes selected for this study. We recover this relationship whether using the plastids from the NCBI plastome database or a larger sampling of publicly available chlorophyte plastid genomes. Since this plastid genome sequence is not as-of-yet associated with a publication, we cannot confirm whether this inconsistency is because of a mis-attribution/labeling of the species name in the database or some other reason. To make Fig. S4 more consistent with the analysis in Fig. 2, we have now replaced Fig. S4 with the taxa used in Fig. 2. The only difference between Fig. 2 and the new Fig. S4 is the absence of plastid MAGs in the phylogenetic reconstruction in the new Fig. S4. Fig. S4 was included to test whether the inclusion/exclusion of MAGs influenced the predicted relationships between protist algae and rhodophytes.

- Trebouxiophyceae is not recovered as monophyletic, with four distinct lineages (Chlorellales, Helicosporidium, Stichococcus, and others). This is consistent with previous findings, but should be acknowledged in the text, especially in light of the statement on line 384 (“forming a distinct clade within Trebouxiophyceae”). Stichococcus might be worth highlighting due to its proximity to Chlorarachniophyceae.

Response: Thank you for pointing out this imprecise statement on line 384; we have revised this sentence to acknowledge the complex trebouxiophyte plastid phylogeny. We have now added the following text to the manuscript to acknowledge this observation and provide citations for previous findings:

Original: “Among these, potential novel lineages include ptMAGs (n=5) from marine ecosystems forming a monophyletic clade sister to Pedinophyceae, and ptMAGs (n=4) from marine and freshwater forming a distinct clade within Trebouxiophyceae.”

Modified (Page 14-15, Lines 447-452): “Particularly, five new ptMAGs (90% median completeness) from marine ecosystems formed a clade sister to Pedinophyceae, while four new ptMAGs (92% median completeness) from marine and freshwater habitats grouped in a distinct clade within the family Chlorellaceae (class Trebouxiophyceae). As observed previously[74,75], the plastomes from Trebouxiophyceae were not recovered as monophyletic (Fig. 2, Supplementary Fig. S4).”

Regarding Stichococcus, we do not consistently observe this plastid sequence near Chlorarachniophyceae in the various reconstructions. We have now added dashed magenta branches to Supplementary Fig. S4 that belong to those plastid genomes whose placement in the tree changes when plastid MAGs are added to the reconstruction and have added an explanation of that annotation in the figure legend.

Minor Comments and Corrections:

- Lines 50–65: The two primary endosymbioses are well described. However, why were other cases where cyanobacteria function as plastids not mentioned? For example, diazoplasts in

Epithemia or nitroplasts in Braarudosphaera bigelowii could provide additional context for the diversity of plastid-like organelles.

Response: We have edited Figure 1c to highlight the phylogenetic placement of the itroplast- and diazoplast-genomes, and added the following paragraph to incorporate your suggestion:

Original: "To date, only two confirmed instances of primary endosymbiosis exist."

Modified (Page 2, Lines 51-57): "To date, while only two confirmed cases of primary endosymbiosis-derived plastids have been documented, other intriguing examples of cyanobacteria evolving from endosymbionts into organelles in marine eukaryotes have been described recently. Notable examples include nitroplasts in *Braarudosphaera bigelowii*[1] and diazoplasts in *Epithemia clementina*[2], both enabling endosymbiotic nitrogen fixation and highlighting ongoing cyanobacterial integration in eukaryotes."

Line 70: Please revise to "and some dinoflagellates acquired secondary plastids."

Response: We have made the change as suggested.

- Line 311: The reference to Fig. 1d should be corrected to Fig. 1c.

Response: We have made the change as suggested.

- Supplementary Figs. S6, S7: The label "Synurophyceae" should be corrected to "Chrysophyceae," as Synurophyceae is nested within the latter.

Response: We have made the change as suggested.

Finally, I would like to mention that as a reviewer without direct experience in generating and filtering MAGs, I was unable to critically assess the methodology in this area. It would be helpful if the authors could provide a brief explanation or justification of their filtering criteria and quality thresholds for ptMAG inclusion.

Response: Below we provide a high-level summary and rationale for our workflow for generating and filtering ptMAGs:

To generate ptMAGs, we followed a well established workflow (similar to what's described in our previous study, https://www.nature.com/articles/s41587-020-0718-6) in which metagenomic sequences are grouped based on tetranucleotide frequencies and relative abundance into so called metagenome assembled genomes (MAGs), followed by QC with established tools such as GTDB-tk (https://doi.org/10.1093/nar/gkab776) and CheckM1/CheckM2 (10.1038/s41592-023-01940-w). This is a widely used approach in the field.

(i) Generation of ptMAGs:

Resulting MAGs were then screened to identify contigs containing 16S rRNA genes using the bacterial SSU covariance model, while accounting for both complete and partial 16S rRNA gene

sequences. The resulting contigs with the SSU rRNA sequences were taxonomically annotated against the SILVA and PR2 databases, and only those with a plastid sequence as the best hit were selected. A minimum contig length of ≥5 kb was required to ensure that the assemblies are large enough for confident taxonomic assignment.

(ii) Filtering of ptMAGs:

To retain only high-confidence plastid MAGs (ptMAGs), we annotated genes on each contigs and applied several additional filtering and quality control steps (which are complementary to the MAG QC procedure described above):

(a) *Minimum gene content:* ptMAGs containing less than 10% of a universal set of 56 single-copy marker genes (UNI56) were excluded, since incomplete assemblies would not be useful for phylogenetic inference.
(b) *Removal of non-cyanobacterial or non-plastid sequences:* All predicted proteins per contig were searched against the NCBI nr database. For each protein, we recorded the taxonomy and amino acid identity of the best hit. Contigs were assigned a consensus taxonomy based on the majority of the proteins. Contigs not assigned to Cyanobacteriota or Eukaryota were discarded, except for those associated with non-cyanobacterial taxa where the best protein hits had < 70% identity, as they were not reliably indicative of non-cyanobacterial origin. This protein-level taxonomic filtering removes likely mis-binned bacterial sequences of non-cyanobacterial origin or genuine lateral gene transfers from non-cyanobacteria, either of which would lead to artifacts in the phylogenetic reconstruction.
(c) *Exclusion of mitochondrial sequences:* Because eukaryotic contigs containing 16S rRNA can originate from mitochondria, we compared all contigs to RefSeq mitochondrial sequences. Any contigs that shared 95% nucleotide identity over >95% of their alignment length with RefSeq mitochondrial sequences were discarded. We found some ptMAGs belonging to Streptophyta sharing higher ANI (>90%) with mitochondrial sequences but these sequences were incorporated in our analyses as they share higher ANI with plastid sequences as well (please see Supplementary Text **Page 2, Lines 63-72**).

These thresholds and filtering steps were chosen to maximize the inclusion of genuine plastid sequences while removing bacterial or mitochondrial sequences.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have satisfactorily addressed my concerns.

I think that it is very important that the discussion about the evolution of SELMA and chlorophyll c is included in the final manuscript as this is a major conclusion of the study and a major point of debate in the field. (Note that even though I think the arguments made by the authors are valid, I think the discussion can be more precise and done more carefully.)

Response: Thank you for the positive assessment and for highlighting the importance of including discussion on SELMA and chlorophyll *c* evolution in the manuscript. In response to this comment, we have added a following section (Pages 7-8, Lines 252-303) that addresses the implications of our findings for the evolution of secondary red algal plastids, and we welcome any additional suggestions the reviewer may have regarding this discussion:

Modified (Pages 7-8, Lines 252-303):

"***Independent origins of secondary red algal plastids***

Interestingly, in all our phylogenetic reconstructions, the Rhodophyta is split (Fig. 2, Supplementary Figs. 4-10). The mesophilic red algae were monophyletic and strongly supported as a sister to Cryptophyta and Haptophyta. Whereas the early diverging red algal subphylum Cyanidiophytina was strongly supported as a sister to Ochrophyta (Fig. 2). This finding is consistent with the initial observation by Kim et al. (2015)[45], and a more recent study by Pietluch et al. (2024)[46], which presented a nearly identical topology, suggesting two independent origins of secondary red algal plastids and proposing a "multiple and serial endosymbiosis" model. This model extends the serial endosymbiosis hypothesis, which posits an initial secondary endosymbiosis of red algae plastids by cryptophytes, followed by acquisition of cryptophytes plastids by ochrophytes and haptophytes following tertiary or higher order endosymbiosis[23–25]. Our study, along with the findings of Pietluch et al. (2024), challenges the notion of a single origin for secondary red algal plastids and the vertical inheritance of the plastids in all lineages with red algal-derived secondary/complex plastids, presenting evidence for two distinct engulfment events involving a red alga.

Two separate events are potentially at odds with phenotypic features shared by algae in the CASH lineages, specifically the symbiont-specific ERAD-like machinery (SELMA) complex and occurrence of chlorophyll *c*[20,47], which are commonly cited as evidence for a single plastid origin. However, the evolutionary history of these features maybe more nuanced and complex. Since the SELMA complex, which facilitates protein import across the periplastidal membrane, is derived from preexisting symbiont-derived ER-

associated protein degradation (ERAD) machinery[48,49], the frequency of independent derivation and reestablishment of this system remains uncertain. Under our model, two independent adaptations (in cryptophytes and ochrophytes) and two establishment (in haptophytes and alveolates) would be required, whereas under a single origin model, a single adaptation followed by multiple reestablishments are required, making both scenarios comparably parsimonious as suggested by Pietluch et al. (2024)[46]. Phylogenetic analyses reveal that SELMA components show mosaic evolutionary origins, with only some SELMA proteins tracing back to red algae and displaying monophyly across CASH lineages[50]. Reconciling these patterns with the single origin and serial endosymbiosis model requires invoking lineage-specific duplications involving ancestral non-endosymbiont genes and replacement of SELMA components rather than simple vertical inheritance[50]. Similarly, while chlorophyll *c* synthase (CHLC) appears monophyletic across CASH lineage, phylogenetic analyses reveal several inconsistencies including polyphyly of ochrophytes, cryptophytes nested within ochrophytes rather than positioned as the basal lineage, and clustering of some ochrophytes and haptophytes members as a clade, potentially due to poor phylogenetic resolution or gene transfer between lineages[51,52]. Notably, some ochrophytes entirely lack CHLC homologs despite producing chlorophyll *c*, demonstrating that alternative chlorophyll *c* biosynthetic pathways have evolved independently at least once. While this absence has been attributed to gene loss and pathway remodeling[52], the independent evolution of chlorophyll *c* synthesis indicates multiple independent innovations and complicates interpretation of CHLC monophyly as definitive evidence for a single endosymbiotic origin."

Reviewer #2 (Remarks to the Author):

I am very satisfied with the revision and recommend accepting the manuscript.
Response: We thank the reviewer for their thorough evaluation and assessment of the revised manuscript.