Phylogenetic dissection provides insights into the incongruity in the tree of Archaeplastida between the analyses of nucleus- and plastid-encoded proteins.

Ryu Isogai[1], Ryo Harada[1,2], Takuro Nakayama[3], Yuji Inagaki[3]

[1] Graduate School of Science and Technology, Univ. of Tsukuba, Ibaraki, Japan
[2] Faculty of Medicine, Dalhousie University, Halifax, Nova Scotia, Canada
[3] Center for Computational Sciences, Univ. of Tsukuba, Ibaraki, Japan

To whom correspondence should be addressed: YI, yuji@ccs.tsukuba.ac.jp

Abstract

Archaeplastida is defined as a taxonomic assemblage comprising three sub-clades, namely Chloroplastida, Glaucophyta, and Rhodophyta plus non-photosynthetic lineages sister to Rhodophyta (collectively termed "Rhodozoa" here). The members of Archaeplastida are the descendants of the eukaryote that took up and transformed a cyanobacterial endosymbiont into a primary plastid. Archaeplastida has been well studied as one of the major eukaryotic assemblages in science, as this assemblage initiated the diversity and evolution of eukaryotic phototrophs, and contains primary producers that are significant for the Earth's environments. Recent phylogenetic analyses of multiple proteins (phylogenomic analyses) stably recovered the monophyly of Archaeplastida, but uncertainty remains in the relationship among the three sub-clades in this assemblage. The phylogenomic analyses of nucleus-encoded proteins (nuc-proteins) put Rhodozoa at the base of the Archaeplastida clade, albeit Glaucophyta was often inferred as the basal lineage in the phylogenomic analyses of plastid-encoded proteins (pld-proteins). In this study, we challenged the previously recognized but as-yet-explicitly addressed issue in the tree of Archaeplastida (ToA). The phylogenomic analyses of the nuc-protein and pld-protein supermatrices processed by the removal of fast-evolving, random, and heterotachious positions revealed that taxon sampling can invoke different types of phylogenetic artifacts into the inferences from both supermatrices examined here. In the end, we propose the basal branching of Rhodozoa as the working hypothesis for the ToA, redefine the issue in the internal relationship among the three sub-clades, and provide future perspectives toward resolving the ToA.

1. Introduction

The plastids surrounded by double membranes have been found in three photosynthetic lineages, namely Chloroplastida, Glaucophyta, and Rhodophyta (Adl et al. 2005). This type of plastid, so-called "primary plastids," was most likely established through "primary endosymbiosis" between a cyanobacterium (endosymbiont) and a heterotrophic eukaryote (host) (Archibald 2009, Keeling 2013, Sibbald and Archibald 2020). It is generally believed that transforming a cyanobacterial endosymbiont into a plastid governed by the host eukaryotic cell is rare in the Tree of Life (ToL). Thus, Chloroplastida, Glaucophyta, and Rhodophyta have been anticipated to form a

monophyletic assemblage (i.e., a supergroup Archaeplastida), evolving from a common ancestor, the first eukaryotic phototroph bearing a primary plastid (Archibald 2009, Keeling 2013, Sibbald & Archibald 2020). Nevertheless, recent studies forced us to update the organismal diversity of Archaeplastida, as well as the rarity of primary endosymbiosis in the ToL. First, two non-photosynthetic groups, Rhodelphida and Picozoa, were reported to belong to Archaeplastida (Gawryluk et al. 2019, Schön *et al.* 2021). Phylogenetic analyses based on multiple proteins (phylogenomic analyses) placed both of them at the base of Rhodophyta (note that we tentatively name the clade of Rhodophyta, Rhodelphida, and Picozoa as "Rhodozoa" in this study). Second, testate amoeba belonging to Rhizaria (*Paulinella* spp.) appeared to bear photosynthetic organelles that are evolutionarily distinct from but equivalent to primary plastids in Archaeplastida (Macorano and Nowack 2021). The host and endosymbiont phylogenies of the eukaryotes bearing cyanobacterium-derived photosynthetic organelles with double membranes claim that primary endosymbiosis occurred (at least) two separate branches of the ToL (Stephens et al. 2021).

Primary plastids can be considered as a convincing plesiomorphy uniting Chloroplastida, Glaucophyta, and Rhodophyta (Rodríguez-Ezpeleta et al. 2005), albeit Rhodelphida and Picozoa lost it secondarily (Gawryluk et al. 2019, Schön *et al.* 2021). Nevertheless, the host monophyly of Chloroplastida, Glaucophyta, and Rhodophyta had been difficult to recover in phylogenomic analyses until recently (Mackiewicz and Gagat 2014). In phylogenomic analyses of nucleus-encoded proteins (nuc-proteins), the monophyly of Archaeplastida was often hindered by Cryptista comprising cryptophyte algae and their non-photosynthetic relatives (goniomonads, kathablepharids, and *Palpitomonas bilix*) (Burki et al. 2016, Gawryluk et al. 2019, Strassert et al. 2019). Fortunately, Archaeplastida was reconstructed as a clade with confidence after carefully assessing the proteins and taxa to be included in phylogenomic analyses (Yazaki et al. 2022, Irisarri et al. 2022). In particular, Yazaki et al. (2022) demonstrated that the taxa branched at the base of Rhodophyta (e.g., *Rhodelphis* spp.), and those positioned at the base of and sister to Cryptista (e.g., *Palpitomonas bilix* and *Microheliella maris*) were critical to suppress the putative false phylogenetic affinity between Rhodophyta and Cryptista and led to the recovery of the monophyly of Archaeplastida. The phylogenomic analyses, in which the suspicious affinity between Rhodophyta and Cryptophyta was

suppressed efficiently, further nominated the assemblage of *M. maris* and cryptists (collectively designated Pancryptista) as the closest lineage of Archaeplastida (Yazaki et al. 2022).

We are now comfortable about the monophyly of Archaeplastida, but there is still an unsettled issue in the tree of Archaeplastida (ToA)—the phylogenetic relationship among Chloroplastida, Glaucophyta, and Rhodozoa remains controversial. In theory, three types of the ToA, depending on which sub-clade occupies the basal position, are possible. Previously published nuc-protein-based phylogenomics often put Rhodozoa/Rhodophyta at the base of the ToA (Rhodo-basal tree) with high statistical support (Tice et al. 2021; Tikhonenkov et al. 2022; Eglit et al. 2024). We can also examine the internal branching patterns in the ToA by analyzing plastid genome-encoded proteins (pld-proteins) if the plastids in the extant members of Archaeplastida have been inherited vertically from their common ancestor. To our knowledge, there has been little ground to doubt the vertical inheritance of plastids in Archaeplastida, and thus the ToA inferred from pld-proteins is supposed to agree with, or at least does not conflict fundamentally with, that from nuc-proteins. In reality, the basal position in the ToA inferred from multiple pld-proteins was typically occupied by Glaucophyta (Glauco-basal tree) (Rodríguez-Ezpeleta et al. 2005, Figueroa-Martinez et al. 2019). Unfortunately, the uncertainty in the phylogenetic relationship among the three subclades of Archaeplastida has not been explicitly addressed in any past study.

We here conducted phylogenomic analyses of an alignment comprising 276 nuc-proteins (nuc276) and that comprising 54 pld-proteins (pld54) to understand the reasons why the ToA inferred from the two types of multi-protein alignments (supermatrices) conflicted with each other. The detailed and systematic assessments revealed that taxon sampling severely affected the phylogenetic inferences from both nuc276 and pld54. Depending on taxon sampling, fast-evolving positions and heterotachious positions appeared to overrate a particular type of the ToA in the nuc276 and pld54 analyses, respectively. Finally, we redefine the issue in the phylogenetic relationship among Chloroplastida, Glaucophyta, and Rhodozoa/Rhodophyta and discuss further perspectives.

2. Materials & Methods

2.1. Nuclear protein-based phylogenomic analyses

2.1.1. Preparation of supermatrices

We modified 351 single-protein alignments considered in a previous phylogenomic analysis (Harada et al. 2024). The single-protein alignments were updated by adding the sequences of 25 species belonging to Archaeplastida, two species of Anaeramoebidea (Stairs et al. 2021), and three operational taxonomic units (OTUs) for Picozoa (see below in this section; Schön et al. 2021). Overall, the above-mentioned archaeplastid species represent most of the classes to orders in Archaeplastida proposed in Adl et al. (2019). Separately, we updated the sequence data of the green alga *Mesostigma viride* in this study, as the coverage of this species was poor at approximately 17.56% in the original supermatrix. For 20 archaeplastid species, the amino acid sequence data were retrieved from the OneKP database (Leebens-Mack *e*t al. 2019; Carpenter et al. 2019). For six archaeplastid species (including *M. viride*) that are not available in the OneKP database, their transcriptome data were obtained from the NCBI Sequence Read Archive (SRA). The nucleotide sequence data were assembled into contigs using Trinity ver. 2.13.2 (Grabherr et al. 2011) after quality control by fastp ver. 0.23.2 (Chen 2023) with options '-q 20 -u 80.' The contig (nucleotide) sequences were translated into amino acid sequences using TransDecoder.LongOrfs ver. 5.5.0 (https://github.com/TransDecoder/TransDecoder) with an option '-m 50.' The amino acid sequence data of the two *Anaeramoeba* species were provided as a part of the supplementary data in Stairs et al. (2021). Schön et al. (2021) generated the single amplified genomes (SAGs) and COSAGs from the picozoan cells and created three OTUs. We followed the procedure taken by Schön et al. (2021) to generate the three picozoan OTUs by merging (i) the amino acids data of SAG22, SAG31, and SAG11, (ii) those of COSAG02, COSAG03, COSAG04, and COSAG06, and (iii) those of COSAG05, COSAG01, and SAG25. The amino acid sequence data considered additionally or updated for assessing the ToA (31 species/OTUs in total; see described above) were subjected to CD-HIT ver. 4.8.1 (Fu *et al.* 2012; Li and Godzik 2006) with an option '-c 0.98' to reduce the sequence redundancy and subsequently formatted into the databases for blastp (see below).

The 351 nuc-proteins, which comprised the supermatrix analyzed in Harada et al. (2024), were searched for individually in the 31 blastp databases (see above). We ran blastp ver. 2.12.0+ (Camacho et al. 2009) against each database with all the sequences included in the previous single-protein alignment as queries. The blastp options were set as '-evalue 1e-30 -max_target_seqs 5 -max_hsps 1.' The candidate sequences retrieved by blastp were aligned with the corresponding alignment by MAFFT v.7.490 (Katoh and Standley 2013), followed by trimming the positions occupied by a large number of gaps and those aligned ambiguously by trimAL v. 1.4.rev15 (Capella-Gutiérrez et al. 2009). Then, the updated alignments were subjected to the maximum likelihood (ML) phylogenetic analyses by using IQ-TREE v. 2.1.2 (Minh et al. 2020). We individually inspected the resultant ML trees to identify and discard potential paralogous and horizontally transferred sequences, as well as those evolved rapidly (long-branch sequences). The remaining sequences after the operations described above were realigned, and then the alignment positions were reselected for the phylogenetic analyses described below. Based on the results from the ML analyses of the refined alignments, at this point, internal parts of the picozoan sequences, in which amino acids were highly diverged compared to other orthologous sequences, were identified. Such highly diverged sequence stretches in the picozoan sequences were replaced with "X," representing an uncertain amino acid identity.

Before the concatenation of the selected single-protein alignments into the final supermatrix, we conducted the third series of aligning and trimming of the sequences by MAFFT with the L-INS-i algorithm and the combination of trimAL with '-gt 0.1' option and BMGE ver. 2.0 (Criscuolo and Gribaldo 2010) with '-g 0.3 -e 0:0.6 -w 5' options. We concatenated the 345 single-protein alignments into a preliminary supermatrix. The first supermatrix was refined by discarding the species and/or single-protein alignments that were poor in the coverage of alignment positions, and we obtained a supermatrix comprising 317 single-proteins from 133 OTUs (nuc317; 117,125 amino acid positions in total). nuc317 was used to infer the global eukaryotic phylogeny. Next, as we were primarily interested in the ToA, a smaller supermatrix, "nuc276," in which 276 single-proteins (94,907 amino acid positions in total) were sampled from 57 OTUs, was generated from nuc317. nuc276 comprised 25 species of Chloroplastida, four species of Glaucophyta, 16 species of Rhodozoa (11 species of rhodophytes, two rhodelphids, and

three picozoan OTUs), and 12 species of other eukaryotes as the outgroup. The outgroup in nuc276 included six species in Pancryptista, which was nominated as the closest related lineage of Archaeplastida (Yazaki et al. 2022), but any photosynthetic member (e.g., cryptophyte algae) was omitted to avoid biasing phylogenetic inferences due to genes potentially transferred from the photosynthetic endosymbionts that gave rise to the current plastids. The single-protein alignments, of which more than half the number of the species in each of the three subclades of Archaeplastida—(i) Chloroplastida, (ii) Glaucophyta, and (iii) Rhodozoa, were retained in nuc276. In every step preparing nuc276 from the preliminary supermatrix (see above), the single-protein alignments were individually realigned by MAFFT with L-INS-i algorithm and then subjected to a two-step trimming of ambiguously aligned positions and those occupied by gaps, initially by trimAl with '-gt 0.1' option and then by BMGE with '-g 0.2 -e 0:0.5 -w 5' options.

The supermatrices comprising 276 nuc-proteins described above were summarized in Table 1.

2.1.2. Phylogenomic analyses

We used IQ-TREE for phylogenetic analyses of nuc317 and nuc276 using the ML method (see below). The LG+C60+F+G model, accounting for both among-site rate variation approximated by gamma distributions (Yang and Kumer 1996) and the heterogeneity in amino acid frequency among alignment positions (Si Quang et al. 2008), was applied to tree searches and UFBoot2 (Hoang et al. 2018) that calculated ultrafast bootstrap values (UFBPs; 1,000 replicates). In non-parametric bootstrap analyses (100 replicates), the heterogeneity in amino acid frequency among alignment positions was approximated by the posterior mean site frequency model (Wang et al. 2018) estimated over the ML tree described above, coupled with the frequency vectors in the LG+C60+F+G model. nuc276 was also analyzed by Bayesian method by using PhyloBayes-MPI ver. 1.9 (Lartillot et al. 2013). We ran two Markov Chain Monte Carlo (MCMC) runs for 11,000 cycles with the GTR+CAT+G model (Lartillot and Philippe 2004) and stored trees every cycle during the MCMC runs. Although the two MCMC runs did not converge (max-diff = 1), we discarded the trees from the initial 3,000 cycles, and the remaining trees were used to calculate the consensus tree and Bayesian posterior probabilities (BPPs). nuc276 and supermatrices derived from nuc276 (see below) were

used for an approximately unbiased (AU) test (Shimodaira 2002), in which the number of RELL resampling was set to 10,000, to examine the relationship among Chloroplastida, Gaucophyta, and Rhodozoa. In the ML tree, Rhodozoa was placed at the base of the Archaeplastida clade in the ML tree (Rhodo-basal tree; see section 3.1.1). For the AU tests, we prepared two alternative trees of Archaeplastida, namely "Glauco-basal tree" and "Chloro-basal tree," in which Glaucophyta and Chloroplastida were forced to be basal, respectively. The two alternative trees were generated from the ML tree (i.e., Rhodo-basal tree) by interchanging the three sub-clades manually.

Three modified versions of nuc276, "nuc276ΔPR," "nuc276ΔR," and "nuc276ΔP," were prepared by excluding (i) two rhodelphids and three picozoan OTUs, (ii) the two rhodelphids, or (iii) the three picozoan OTUs from the original supermatrix, respectively (Table 1). Both ML and Bayesian analyses described above were repeated on nuc276ΔPR (the details of the analyses are the same as described above; in Bayesian analysis, the two MCMC runs did not converge as the max-diff stayed 1 after 11,000 cycles). In addition, the three supermatrices derived from nuc276 were subjected to the AU tests exploring the three candidate trees of Archaeplastida (see above).

We evaluated the impact of fast-evolving positions in the original and modified versions of nuc276 on the ToA. nuc276 and its corresponding ML tree were subjected to IQ-TREE with the '--rate' option to estimate the substitution rates for individual alignment positions (site-rates) using Bayesian approach. Then, the fastest-evolving positions were progressively removed from each of nuc276 and its derivatives at 20% intervals (fast-evolving position removal or FPR). We also removed random positions from the supermatrices described above at 20% intervals (random position removal or RPR). At each data point, 20 sets of alignment positions were chosen randomly and excluded from each of nuc276 and its derivatives to prepare 20 distinct RPR-processed supermatrices. This procedure allowed us to compare the phylogenetic inferences from a pair of nuc276ΔR and nuc276ΔP, from which the exact same set of randomly selected alignment positions was excluded (see section 3.1.3). The ML tree search, in conjunction with UFBoot2, was conducted on each of the supermatrices processed by both FPR and RPR.

The analyses of nuc276ΔR and nuc276ΔP revealed that rhodelphids and picozoan OTUs had different degrees of impact on the ToA (see section 3.1.3). We hypothesized

that such a difference stemmed from the difference in coverage between the rhodelphid and picozoan sequences in the alignment. The average percent of missing data/gap positions in nuc276 was 4.85% in the rhodelphid sequences and 52.88% in the picozoan sequences. Thus, we artificially increased the proportions of missing data/gap positions in the rhodelphid sequences to the compatible level of those in the picozoan sequences (see below). First, we extracted 50,140 "empty positions" occupied by gaps in more than two picozoan sequences from nuc276. Likewise, we defined 44,767 of the positions in which gaps occurred less than two picozoan sequences in nuc276 as "filled positions" (note that "empty" and "filled" positions are complementary to each other in the picozoan sequences in nuc276). Then, we replaced any amino acid characters of the rhodelphid sequences at the positions corresponding to the "empty positions" in the picozoan sequences with gaps. This manipulation reduced the average percent value of missing/gap positions in the rhodelphid sequences to 54.84%, which is compatible with that in the picozoan sequences in the original nuc276 (Table 1). We then excluded the three picozoan OTUs from the modified nuc276 to generate "nuc276ΔP*". We repeated the procedure described above, except for considering "filled positions" instead of "empty positions," and generated "nuc276ΔP**" (the averaged percent value of missing/gap positions in the rhodelphid sequences to 50.01%; Table 1). Both nuc276ΔP* and nuc276ΔP** were used for the AU test assessing the three candidate trees of Archaeplastida as described above.

2.2. Plastid protein-based phylogenomic analyses

2.2.1. Preparation of supermatrices

We prepared a supermatrix based on pld-proteins by referring to that analyzed in Ponce-Toledo et al. (2017) to investigate the relationship among Chloroplastida, Glaucophyta, and Rhodophyta. From the NCBI RefSeq or GenBank database, we retrieved the protein (amino acid) sequences encoded in the plastid genomes of 42 archaeplastid species that were also considered in nuc276 (see above) plus six complex plastids that are the remnants of red/green algal endosymbionts, and nine cyanobacterial genomes that were chosen by referring to the cyanobacteria phylogeny in Strunecký et al. (2023) (57 OTUs in total). The 97 proteins encoded in the plastid genome of a glaucophyte *Cyanophora paradoxa* were set as queries, and their homologues were searched for in the protein sets encoded in the plastid and cyanobacterial genomes

described above by blastp (options were set as '-evalue 1e-4 -max_target_seqs 5 -max_hsps 1'). In each blastp analysis, only the first hit was retained, as the proteins encoded in plastid genomes are usually single-copied or, even if duplicated, the copies are identical to each other at the amino acid level. We omitted 37 proteins that bore poor coverages or were absent from the plastid genomes of the chloroplastid species considered, leaving 57 proteins for the analyses described below. The preliminary versions of 57 single-protein alignments were generated by MAFFT combined with trimAl for alignment trimming and then subjected to ML analyses using IQ-TREE for the final inspection. We omitted three alignments in which phylogenetic signals potentially bias the phylogenetic inference from a supermatrix incorporating them. The above-mentioned process yielded the final selection of 54 pld-proteins for the phylogenomic analyses described in the next section. These sequences were realigned by MAFFT with the L-INS-i algorithm, followed by a two-step trimming procedure composed of trimAl with '-gt 0.1' option for the first step and BMGE with '-g 0.06 -e 0:0.5 -w 5' options for the second step. The final versions of the 54 single-protein alignments were then concatenated into a "pld54" alignment (57 OTUs; 12,787 amino acid positions in total). We modified pld54 by excluding (i) the six complex plastids, (ii) two green alga-derived complex plastids, and (iii) four red alga-derived complex plastids to generate "pld54Δg*r*," "pld54Δg*," and "pld54Δr*," respectively.

The supermatrices comprising 54 pld-proteins described above were summarized in Table 1.

2.2.2. Phylogenomic analyses

Both pld54 and pld54Δg*r* were subjected to a series of analyses using the ML methods, namely tree search, UFBoot2 (1,000 replicates), and non-parametric bootstrap analysis (100 replicates). Among the substitution models considering the heterogeneity in amino acid frequency among alignment positions by C60, the cpREV+C60+F+I+G model was selected by ModelFinder (Kalyaanamoorthy et al. 2017) implemented in IQ-TREE ver. 2.2.2.7. In non-parametric bootstrap analyses, the heterogeneity in amino acid frequency among alignment positions was approximated by the posterior mean site frequency model estimated over the ML tree described above, coupled with the frequency vectors in the cpREV+C60+F+I+G model. The two supermatrices were phylogenetically

analyzed with Bayesian method using PhyloBayes-MPI. We ran two MCMC runs for 120,000 cycles with the GTR+CAT+G model and stored trees every cycle during the MCMC runs. The MCMC runs converged in the analysis of pld54 with a max-diff of 0.0903 and in the analysis of pld54Δg*r* with a max-diff of 0.0265. After discarding the trees from the initial 30,000 cycles as "burn-in," the remaining trees were used to calculate the consensus tree and BPPs.

We conducted the pld54-based AU test comparing Chloro-basal, Glauco-basal, and Rhodo-basal trees. Two alternative trees were prepared by modifying the ML tree inferred from the corresponding supermatrix (note that either Glauco-basal or Chloro-basal tree was recovered as the ML estimate, depending on the supermatrix analyzed and the phylogenetic method applied; see section 3.2.1). The site-lnLs were calculated with the cpREV+C60+F+I+G model over the ML and two alternative trees and used for the AU test (the number of RELL replicates was set to 10,000). We repeated the AU test considering the modified version of pld54 (i.e., pld54Δg*r*, pld54Δr*, and pld54Δg*; Table 1), together with the three test trees (see above).

As done with the original and modified versions of nuc276 (see above), we subjected pld54 and pld54Δg*r* to both FPR and RPR, followed by the ML tree search and UFBoot2. The details were the same as described above, except for (i) the cpREV+C60+F+I+G model being applied to the phylogenetic analyses and (ii) removing alignment positions being repeated at 10% intervals.

We noticed the overall difference in branch length between the six complex plastids and others (i.e., primary plastids and cyanobacteria) in the pld54 phylogeny (see section 3.2.1). Thus, we focused on the alignment positions of which the substitution rates were changed with and without the complex plastids and explored their impact on the phylogenetic inference of pld54. The substitution rate at each position (site-rate) based on pld54 was divided by the corresponding site-rate based on pld54Δg*r*. The alignment positions associated with site-rate ratios that departed largely from 1 potentially underwent "heterotachious" substitution modes. A site-rate ratio >1 indicates that the site-rate up-shifts in the presence of the complex plastids relative to their absence. In the case of the site-rate down-shifting in the presence of the complex plastids relative to their absence, the ratio is supposed to be <1. We sorted the alignment positions in descending order of the site-rate ratio and progressively excluded the alignment positions associated

with the 5% largest site-rate ratios and then 10-90% (at 10% intervals) from pld54. The resultant alignments were subjected to the ML tree search coupled with UFBoot2. We also excluded the alignment positions associated with the 5% smallest site-rate ratios, and then 10-90% progressively, followed by the ML tree search and UFBoot2 (see above). Lastly, we modified pld54 by excluding the alignment positions associated with both the 5% largest and 5% smallest site-rate ratios (Table 1) and subjected it to the AU test examining the three types of the ToA.

3. Results and Discussion

3.1. Nuclear protein-based phylogenomic analyses

First, we phylogenetically analyzed nuc317 to examine whether the overall ToE was changed before and after updating the samplings of taxa and proteins in a supermatrix. The ML phylogeny inferred from nuc317 (Fig. S1) verified the monophyly of Archaeplastida, as well as the sister relationship between Archaeplastida and Pancryptista (Yazaki et al. 2022). Other major clades of eukaryotes, such as TSAR, Haptista, Amorphea, CRuMs, Discoba, and Metamonada, were also recovered. The above-mentioned clades received UFBPs ranging from 88-100% and non-parametric ML bootstrap percent values (MLBPs) of 99-100% (Fig. S1). nuc276, a more compact alignment than nuc317, was generated for assessing the internal relationship among the three sub-clades in Archaeplastida. nuc276 comprised 45 archaeplastid species and the outgroup taxa (i.e., six pancryptists and six other eukaryotes). Any phototroph was omitted from the outgroup because such eukaryotes may have endosymbiotically transferred genes that can bias the ToA.

3.1.1 Internal relationship among the three sub-clades in Archaeplastida: With and without Rhodelphida and Picozoa

In the ML phylogeny of nuc276 (Fig. 1A), the monophyly of Archaeplastida was recovered. In the Archaeplastida clade, Rhodozoa served as the basal lineage in the Archaeplastida clade. The three bipartitions described above received full support from UFBoot2, ML bootstrap, and Bayesian method (Fig. 1A; see also Fig. S2 for the details). We also examined the alternative tree topologies for the ToA by the AU test. In this test, the Rhodo-basal tree was the ML estimate, and both Glauco-basal and Chloro-basal trees

were rejected at the 1% level (Table 2).

The phylogenomic analyses presented in Yazaki et al. (2022) demonstrated that Rhodelphida and *M. maris*, which split the branch leading to the Rhodophyta clade and that leading to the Cryptista clade, respectively, were critical to ameliorating the false affinity between Rhodophyta and Cryptista and helped to recover the Archaeplastida clade. When the analyses presented in Yazaki et al. (2022) were conducted, the transcriptome data of Picozoa reported in Schön et al. (2021) was not publicly available. Consequently, Yazaki *et al.* (2022) did not evaluate the impact of Picozoa on the ToA or its intimate phylogenetic affinity between Archaeplastida and Pancryptista. In this study, we explored the significance of Rhodelphida and Picozoa to the relationship among the three sub-clades in Archaeplastida by analyzing the modified nuc276 excluding either or both of the non-photosynthetic relatives of Rhodophyta (i.e., nuc276ΔRP, nuc276ΔP, and nuc276ΔR; Table 1). In addition, we anticipated that the analyses of nuc276ΔRP provide clues for why the internal relationship in the ToA was incongruent between the phylogenomic trees based on nuc-proteins and pld-proteins, as no pld-protein data are available for the currently known members of Rhodelphida or Picozoa that most likely lost their plastid genomes. Intriguingly, the Glauco-basal tree was recovered from the ML analyses of nuc276ΔRP (Fig. 1B; see also Fig. S3 for the details). The Archaeplastida clade was recovered with UFBP of 99%, MLBP of 100%, and BPP of 1.0, and Chloroplastida and Rhodophyta grouped together excluding Glaucophyta with UFBP of 89%, MLBP of 100%, and BPP of 1.0 (Fig. 1B). The inference from nuc276ΔRP appeared to be not conclusive, as the AU test failed to reject the Rhodo-basal tree with a *p*-value of 0.163 (Table 2).

3.1.2 Analyses of the alignments excluding fast-evolving and random positions

It is important to pursue why the exclusion of both Rhodelphida and Picozoa switched the ML tree topology from Rhodo-basal to Glauco-basal tree (Fig. 1). As Rhodelphida and Picozoa split the somewhat long branch leading to the Rhodophyta clade (Fig. 1A), we wondered whether their absence triggered a phylogenetic artifact resulting in the position of Rhodophyta in the Archaeplastida clade instable. In general, fast-evolving positions in alignments are regarded as one of the typical sources of phylogenetic artifacts. Pioneering works have evaluated the degree of the contribution

from fast-evolving positions to the ML tree by comparing the tree inferred from the original alignment and those from the alignments processed by FPR (Yazaki et al. 2022; Tikhonenkov et al. 2022; Eglit et al. 2024). We here subjected nuc276 and nuc276ΔRP to both FPR and RPR to examine the contribution of fast-evolving positions to the ML estimates from the two supermatrices.

The analyses of nuc276 processed by FPR and RPR suggested little contribution of fast-evolving positions in nuc276 to the dominance of the Rhodo-basal tree over others. In Figs. 2A-D, the results from FPR and RPR were presented as line graphs and box plots, respectively. The Rhodo-basal tree was constantly recovered from the analyses after the removal of the top 20-60% fastest-evolving positions from nuc276 (Fig. 2B). After the top 80% fastest-evolving positions were removed, the remaining alignment positions were inappropriate to infer the relationship among the three sub-clades in the Archaeplastida clade, as Picozoa was tied to Glaucophyta instead of Rhodophyta and Rhodelphida. In the analyses of RPR-processed alignments, the Rhodo-basal tree was recovered at 20 out of the 20 trials (20-60% removals) and 14 out of the 20 trials (80% removal) (Fig. S4A). As the Rhodo-basal tree was recovered in the analyses of both FPR and RPR-processed supermatrices, we conclude that the contribution of fast-evolving positions in nuc276 to the ML estimate (i.e., Rhodo-basal tree) is negligible.

We yielded the opposite conclusion from the analyses of nuc276ΔRP processed by FPR and RPR (Figs. 2E-H). The basal lineage in the Archaeplastida clade switched from Glaucophyta to Rhodophyta after removing the top 20% fastest-evolving positions from nuc276ΔRP (see line graphs in Figs. 2F & G). The Rhodo-basal tree appeared dominant over the Glauco-basal tree until removing the top 80% fastest-evolving positions (Figs. 2F & G). In analyses excluding the top 80% fastest-evolving positions, the ML estimate was still the Rhodo-basal tree, albeit the Glauco-basal tree contended with the Rhodo-basal tree according to UFBPs (the Rhodo-basal and Glauco-basal trees received UFBPs of 53% and 26%, respectively; see line graphs in Figs. 2F & G). In sharp contrast to the analyses of FPR-processed alignments described above, no switch of the ML tree topology from the Glauco-basal tree to the Rhodo-basal tree was promoted by the exclusion of random positions in nuc276ΔRP (see box plots in Figs. 2F & G). The Glauco-basal tree was recovered in 12-19 out of the 20 trials in the analyses after removing 20-60% of random positions (Fig. S4B). We observed that the dominance of

the Glauco-basal tree over the Rhodo-basal tree was diminished by further removal of random positions (Fig. S4B). The results from the analyses described here suggest that the Glauco-basal tree relies heavily on the top 20% fastest-evolving positions, while the rest of the alignment positions favor the Rhodo-basal tree rather than the Glauco-basal tree.

3.1.3 Rhodelphida and Picozoa as the keys for inferring the ToA

In the analyses of nuc276, we found no apparent evidence for the phylogenetic inference, in which Rhodozoa was placed at the base of the Archaeplastida clade, being severely biased. Noteworthy, the results above reinforced the importance of Rhodelphida and Picozoa for inferring the ToA, as the absence of the two lineages introduced the phylogenetic artifact stemming from fast-evolving positions. Yazaki *et al.* (2022) demonstrated that including Rhodelphida (and *M. maris*) in phylogenomic analyses was critical to preventing the artifactual union of Rhodophyta and Cryptista, which was principally driven by cryptophytes, and recovering the monophyly of Archaeplastida. Note that the analyses of nuc276 and its derivatives contained no cryptophytes and, thus, were primarily immune to this phylogenetic artifact. Combining the results presented here with those reported in Yazaki *et al.* (2022), the non-photosynthetic lineages splitting from the branch leading to the Rhodophyta clade can be regarded as keys to inferring the early evolution of Archaeplastida.

Rhodelphida and Picozoa were shown to suppress the phylogenetic artifact from fast-evolving positions. Although not described above, either of the two lineages alone suppressed the Glauco-basal tree in the presence of the top 20% fastest-evolving positions. The Rhodozoa branched at the base of the Archaeplastida clade with UFBPs of 80-99% in the ML analyses of nuc276 modified by the reciprocal exclusion of Picozoa and Rhodelphida (i.e., nuc276ΔP and nuc276ΔR; Figs. 1A and S5). In the tests assessing the three types of the ToA based on nuc276ΔP and nuc276ΔR, the Rhodo-basal tree received greater likelihoods than the Glauco-basal or Chloro-basal tree (Table 2). Nevertheless, the exclusion of Rhodelphida alone and that of Picozoa alone appeared to bear different magnitudes in suppressing the Glauco-basal tree. The AU test considering Rhodelphida alone (i.e., nuc276ΔP) rejected the Glauco-basal tree at the 5% level, while the Glauco-basal tree failed to be rejected in the test considering Picozoa alone (i.e., nuc276ΔR)

(Table 2). The results from the analyses of RPR-processed nuc276ΔP and nuc276ΔR are shown in Supplementary Figs. S6A-D. The UFBPs for the Rhodo-basal tree tend to be lower in the analyses considering Picozoa alone than in those considering Rhodelphida as the sole basal lineage of Rhodophyta (ΔP and ΔR in Fig. S6B). In contrast, the UFBPs for the Glauco-basal tree appeared to be elevated in the analyses considering Picozoa alone, compared to those considering Rhodelphida alone (ΔP and ΔR in Fig. S6C).

We hypothesized that the difference in the proportion of gaps between the rhodelphida and picozoan sequences in nuc276 reflects the magnitude of suppressing the Glauco-basal tree described above. To examine the hypothesis, nuc276ΔP* and nuc276ΔP** were prepared from nuc276 by excluding Picozoa and then increasing the gap proportion in the rhodelphid sequences to a similar degree to the picozoan sequences (Table 1). Importantly, the two distinct patterns of introducing gaps into the rhodelphid sequences produced nuc276ΔP* and nuc276ΔP**. Subsequently, the two supermatrices were subjected to the AU test considering the three types of the ToA and FPR/RPR followed by the ML tree search coupled with UFBoot2. The AU test based on either of the two supermatrices failed to reject the Glauco-basal tree (Table 2). As the test based on nuc276ΔP, for which rhodelphid sequences have all of the original positions, rejected the Glauco-basal tree at the 1% level (Table 2), the increment of gap proportion in the rhodelphid sequences, regardless of how gaps being introduced, appeared to negatively affect the dominance of the Rhodo-basal tree over the Glauco-basal tree. A similar trend was observed in the analyses of RPR-processed nuc276ΔP* and nuc276ΔP** (Figs. S6A-D). Compared to the analyses of the RPR-processed nuc276ΔP, (i) the UFBPs for the basal branching of Rhodozoa in the Archaeplastida clade appeared to be lowered, while (ii) the UFBPs for the basal branching of Glaucophyta appeared to be elevated, in the analyses of the RPR-processed nuc276ΔP* and nuc276ΔP** (ΔP* and ΔP** in Figs. S6B & C). Regarding the ML tree topology, higher frequencies of recovering the Glauco-basal tree were observed after the increment of gap proportion in the rhodelphid sequences (compare Figs. S5C & E and S4C & F). Thus, we conclude that the difference in the magnitude of suppressing the Glauco-basal tree between Rhodelphida and Picozoa stemmed from the large proportion of gaps in the latter.

3.2. Plastid protein-based phylogenomic analyses

3.2.1 Internal relationship among the three sub-clades in Archaeplastida: With and without complex plastids

Previously published phylogenetic studies of the supermatrices comprising multiple pld-proteins often suggested the plastids in glaucophytes (henceforth designated as "glauco plastids") as the most ancestral in the plastid clade (i.e., Glauco-basal tree) (Ponce-Toledo et al. 2017; Figueroa-Martinez et al. 2019). We also recovered the Glauco-basal tree from pld54 by the ML method with UFBP of 80% and MLBP of 67% (Fig. 3A; see Fig. S7 for the details). On the other hand, Bayesian analysis of pld54 nominated the plastids in chloroplastids (green plastids) as the most ancestral in the plastid clade with a BPP of 1.0 (i.e., Chloro-basal tree; Fig. S8). In both ML and Bayesian analyses, the monophyly of plastids was recovered with UFBP of 100%, MLBP of 100%, and BPP of 1.0 (Fig. 3A). Corresponding with the uncertainties in the standard ML and Bayesian phylogenetic analyses (see above), the AU test comparing the Glauco-basal tree (ML) with (i) the Chloro-basal tree and (ii) the tree suggesting the plastids in rhodophytes (red plastids) as the most ancestral (i.e., Rhodo-basal tree) based on pld54 failed to reject the Chloro-basal tree (Table 3).

pld54 includes six complex plastids—two are the remnants of green algae engulfed by the ancestors of chlorarachniophytes and euglenids, and four represent the plastids, which can be traced back to red algal endosymbiont(s) directly or indirectly, in stramenopiles, a haptophyte alga, and a cryptophyte alga. The branches of the six complex plastids were longer than others in the plastid clade (Fig. S7), indicating the increment of substitution rates during the reductive process that transformed algal endosymbionts into plastids in the host eukaryotic cells. We regard the six complex plastids in pld54 as dispensable for assessing the relationship among the three sub-clades in the plastid clade, as (i) they are essentially red or green plastids and (ii) their rapidly evolving natures may bias the phylogenetic inferences (e.g., potential sources of long-branch attraction artifacts). The ML and Bayesian analyses recovered the Chloro-basal tree from pld54 after excluding the six complex plastids (pld54Δg*r*) with UFBP of 89%, MLBP of 83%, and BPP of 0.99 (Fig. 3B), but the AU test remained the possibility for the Glauco-basal tree (Table 3). Curiously, the ML analyses of pld54 excluding either green or red complex plastids (i.e., pld54Δg* or pld54Δr*) yielded the Chloro-basal tree as well (Figs. S10A &

B). The results described above clearly demonstrated that the recovery of the Glauco-basal tree demands a pair of green and red complex plastids in a supermatrix.

3.2.2 Analyses of the alignments excluding fast-evolving and random positions

The phylogenetic relationship among the three sub-clades in the plastid clade inferred from pld54 appeared to be different with and without the six complex plastids (see the previous section). Thus, we subjected pld54 and pld54Δg*r* to FPR and RPR and subjected them to the ML tree search and UFBoot2 (Figs. 4A-H), as done on nuc276 and its derivatives. In the analyses of FPR-processed pld54, the monophyly of plastids was recovered until the removal of the top 80% fastest-evolving positions (see the line graph in Fig. 4A). The Glauco-basal tree was dominant over the other two trees until the top 70% fastest-evolving positions were removed (see the line graphs in Figs. 4B-D), suggesting that fast-evolving positions are unlikely to be responsible for the basal position of glauco plastids in the plastid clade. In contrast, the Glauco-basal and Chloro-basal trees appeared to compete with each other in the analyses of RPR-processed pld54 (see the box plots in Figs. 4C & D). Regardless of the proportion of random positions removed, the UFBPs for the Glauco-basal tree ranged widely from 0.1 to 99.7% and those for the Chloro-basal tree from 0 to 99.9% (Figs. 4C & D). Similarly, the two trees were recovered from the RPR-processed pld54 at similar frequencies (Fig. S11A).

The analyses of FPR-processed pld54Δg*r* recovered the monophyly of plastids until the top 80% fastest-evolving positions were removed (see the line graph in Fig. 4E), as observed in the analyses of the FPR-processed pld54 (see above). Nevertheless, the basal branch in the plastid clade appeared to be either glauco or green plastids, depending on how far we removed fast-evolving positions from pld54Δg*r* (see the line graphs in Figs. 4G & H). The Glauco-basal tree was preferred over the Chloro-basal tree in the analyses after removing the top 10, 40, and 50% of the fastest-evolving positions. In contrast, the analyses of pld54Δg*r*, from which the top 20, 30, and 60% of the fastest-evolving positions were removed, recovered the Chloro-basal tree rather than the Glauco-basal tree. The contention between the Glauco-basal and Chloro-basal trees was reproduced in the analyses of RPR-processed pld54Δg*r*. As observed for RPR-processed pld54, the UFBPs for the two trees were distributed nearly from 0 to 100%, regardless of the proportion of alignment positions removed (see the box plots in Figs.

4G & H). Likewise, the two trees competed in terms of the frequency that occurred in the 20 trials at each of 10-70% removals of random positions (Fig. S11B).

The results described above provide us with two insights into the phylogenetic inferences from pld54 and pld54Δg*r*, namely (i) the lack of the phylogenetic "power" to resolve the ToA (see below) and (ii) a type of systematic bias introduced by the six complex plastids (see the next paragraph). First, the analyses of the RPR-processed pld54 demonstrated that the "balance" between the Glauco-basal and the Chloro-basal trees could be altered by choosing (even a slight amount of) alignment positions. For instance, we cannot be sure whether glauco or green plastids are the most ancestral, even after removing 10% of random positions from supermatrices (see the box plots in Fig. 4G). A similar "balance" between the Glauco-basal and Chloro-basal trees was observed in the RPR-processed pld54Δg*r*. It is noteworthy that the analyses after the removal of 10-50% of random positions from pld54 tend to provide more support for the Glauco-basal tree than the Chloro-basal tree, as the UFBPs for the Glauco-basal tree appeared to skew upward relative to the corresponding values from the analyses of RPR-processed pld54Δg*r* (Figs. 4C & G). Likewise, the Glauco-basal tree was recovered more frequently in the analyses of RPR-processed pld54 than those of the RPR-processed pld54Δg*r* (Figs. S11 A & B). Despite the skew of the UFBPs for the Glauco-basal tree described above hinting at a certain degree of the "signal" for the particular tree in the complex plastids, the analyses of the two supermatrices examined here are unlikely to clarify whether glauco or green plastids are the basal branch in the plastid clade.

Second, we here claim that the Glauco-basal tree was most likely overrated by the complex plastids in the phylogenetic inferences from pld54. The ML estimate from pld54 was the Glauco-basal tree (Figs. 3A & S7), albeit the same analysis excluding both or either green and red complex plastids yielded the Chloro-basal tree (Figs. 3B & S9 & S10), suggesting that the union of green and red plastids relies primarily on the presence of both types of complex plastids in the supermatrix. Considering the evolutionary trajectories of green and red complex plastids, they are dispensable to infer the relationship among green, glauco, and red plastids. Furthermore, from the general point of view in phylogenetic methodology, their rapidly evolving natures, which can be the source of various phylogenetic artifacts, may be detrimental. Curiously, the putative overcredibility to the Glauco-basal tree introduced by the two types of complex plastids

has nothing to do with fast-evolving positions in pld54, as the analyses of the FPR-processed pld54 constantly recovered the Glauco-basal tree (Fig. 4C). Thus, we examine another aspect in sequence evolution, namely heterotachy (Lopez et al. 2002), for the reason why the Glauco-basal tree is recovered from the phylogenetic analyses of pld54 in the next section.

3.2.3 Heterotachy is responsible for the union of red and green plastids in the analyses of plastid protein-supermatrices

The alignment positions with the shifts in site-rate, which were introduced by the presence of the complex plastids, were identified based on the ratios of site-rates based on pld54 and those based on pld54Δg*r*. We analyzed pld54 processed in two distinct ways of removing the alignment positions with shifts in site-rates with and without the complex plastids. The results of the analyses of pld54 processed by the progressive removal of the alignment positions with the largest site-rate ratios (i.e., the site-rates calculated with the complex plastids were greater than those without) are shown in Fig. 5A. Regardless of the depth of the removal of the alignment positions, the basal branching of glauco plastids stayed the ML estimate with UFBPs of 87-100% (Fig. 5A). Thus, we conclude that the alignment positions, in which the site-rates were accelerated by incorporating the complex plastids, are highly unlikely to be responsible for the dominance of the Glauco-basal tree over the Chloro-basal (or Rhodo-basal) tree.

The removal of the alignment positions with the smallest site-rate ratios (i.e., the site-rates calculated with the complex plastids were smaller than those without) gave a drastic change in the ML estimate (Fig. 5B). The removal of the positions with the 5% smallest site-rate ratios switched the ML estimate from the Glauco-basal tree to the Chloro-basal tree (Fig. 5B). Throughout this set of the removal of the positions based on the site-rate ratio, the Chloro-basal tree was constantly recovered as the ML estimates with UFBPs of 94.2-100%. In sharp contrast, the UFBPs for the Glauco-basal tree were negligible (0-5.8%). The results described above claim that the “signal” for the Glauco-basal tree is concentrated in the alignment positions with the 5% smallest site-rate ratios. In other words, the Glauco-basal tree (i.e., the basal position of glauco plastids in the plastid clade) principally depends on the heterotachy in pld54. Importantly, our claim above can reconcile the results from the analyses of the FPR-processed pld54 (Figs. 4C).

Even after the removal of the top 60% fastest-evolving positions from pld54, almost two-thirds of the alignment positions with the 5% smallest site-rate ratios remained for tree search and UFBoot2 (Fig. 5C), resulting in the recovery of the Glauco-basal tree as the ML estimate. In sum, we conclude that the basal position of glauco plastids in the plastid clade heavily depends on the heterotachy in pld54.

We reconstructed the ToA from pld54 after excluding the alignment positions with heterotachy. To be fair, the ML analysis was conducted on pld54 after excluding the alignment positions associated with the largest 5% site-rate ratios, of which exclusion yielded no change in the ML estimate (Fig. 5A), as well as those with the smallest 5% site-rate ratios. As anticipated from the analyses of pld54 processed by removing the alignment positions with the smallest 5% site-rate ratios (Fig. 5B), the Chloro-basal tree was recovered as the ML estimate with a UFBP of 90% (Fig. S12), albeit the AU test failed to reject the Glauco-basal tree (p = 0.231; Table 3). The results from the tests in the presence and absence of the heterotachy in pld54 appeared not to be incongruent with one another, leaving the possibility of both glauco and green plastids as the basal branch in the plastid clade (Table 3). Combined with the discussion derived from the analyses of the RPR-processed pld54 and pldΔg*r* (see the previous section), we again conclude that pld54 retains no sufficient signal to clarify the basal branch in the plastid clade.

3.3. Redefining the issue in the ToA

In pioneering phylogenomic studies, the analyses based on nuc-proteins inferred the Rhodo-basal tree of Archaeplastida (Irisarri et al. 2021; Yazaki et al. 2022), while those based on pld-proteins recovered the Glauco-basal tree (Ponce-Toledo *et al.* 2017; Figueroa-Martinez et al. 2019). This uncertainty in the ToA has been known but not addressed explicitly prior to the current study. We here addressed the question of why different trees of Archaeplastida were inferred from supermatrices of nuc-proteins and those of pld-proteins (henceforth abbreviated as “nuc-supermatrices” and “pld-supermatrices,” respectively). The systematic assessments on nuc-supermatrices (nuc276 and its derivatives) and pld-supermatrices (pld54 and its derivatives) demonstrated that taxon sampling in supermatrices had a large impact on the relationship among Chloroplastida, Glaucophyta, and Rhodozoa/Rhodophyta (and their plastids). The nuc276 analyses revealed Rhodelphida and Picozoa were critical to recovering the Rhodo-basal

tree by suppressing the phylogenetic artifact stemming from fast-evolving positions in the supermatrix (see section 3.1.2). On the other hand, the pld54 analyses are most likely insufficient to conclude whether the Glauco-basal or Chloro-basal tree is more appropriate than the other. Considering the analyses conducted in this study, we here re-address the issue in the ToA. The nuc276 analyses constantly suggested the Rhodo-basal ToA and excluded the possibility of the other two trees (Table 2). In sharp contrast, regardless of the presence or absence of the complex plastids that potentially bias the phylogenetic inference, the pld54 analyses excluded the Rhodo-basal tree (Table 3). Combined, we should redefine the issue in the ToA around the Rhodo-basal tree as the center, namely, why analyses of pld-supermatrices disfavor the Rhodo-basal tree severely, albeit those of nuc-supermatrices favor the same tree over the alternatives.

Currently, there is no practical procedure to pursue the redefined issue in the ToA, except for applying more complex substitution models than those used in the analyses mentioned above (i.e., LG+C60+F+G and cpREV+C60+F+I+G) (Table 4). We additionally examined three complex models, namely a model in which among-site rate heterogeneity was approximated by a FreeRate model (Soubrier et al. 2012), and GHOST models (Crotty et al. 2020) accounting for heterotachy in a supermatrix (LG+C60+F+R9, LG+C60+F+H4, and LG+C60+F+H8), in the AU tests comparing the Chloro-basal, Glauco-basal, and Rhodo-basal trees over nuc276. nuc276 was chosen because no apparent phylogenetic bias was detected (see above). Due to the severe computational burdens of the FreeRate and GHOST models, these models could not be used in the ML tree search, UFBoot2, or non-parametric bootstrap. Importantly, none of the three complex models brought any substantial change in the results of the AU tests, dominating the Rhodo-basal tree over the alternatives (Table 4).

For the pld-supermatrix for the AU tests with a FreeRate model (cpREV+C60+F+I+R7) and GHOST models (cpREV+C60+F+I+H4 and cpREV+C60+F+I+H8), we selected pld54. As discussed above, pld54 is most likely short of the phylogenetic power to resolve the ToA. Rather, this series of the AU test was conducted to examine whether the GHOST model can overcome the recovery of the Glauco-basal tree depending solely on the heterotachious positions in the supermatrix. The overall results (i.e., the rejection of the Rhodo-basal tree) were unchanged by switching the approximation of among-site rate heterogeneity from the gamma to the

FreeRate model (Table 4). Likewise, the test applying the cpREV+C60+F+I+H4 model gave the highest log-likelihood to the Glauco-basal tree and rejected the Rhodo-basal tree at the 1% α-level (Table 4). Intriguingly, a drastic change occurred in the result from the test incorporating a more parameter-rich model (cpREV+C60+F+I+H8); the Rhodo-basal tree received a higher log-likelihood than the Glauco-basal or Chloro-basal tree (Table 4). It is attractive to reconcile that not the H4 but the H8 model ameliorated sufficiently the heterotachy in pld54, inferring the Rhodo-basal tree from pld54. There is no evident sign of overparameterization in the H8 model. So far, the H8 model was judged to be more appropriate than the other models examined here under the Bayesian information criterion (Table 4). Although the result described above is encouraging, we need to ensure whether the GHOST model took into account the heterotachy in pld54 properly. It will also be intriguing which type of the ToA is recovered in future analyses of pld54, incorporating new substitution models, preferentially those accounting for heterotachy better than the GHOST model. In sum, the incongruity in the ToA between the analyses of nuc- and pld-supermatrices has yet to be resolved, and this matter needs to be pursued further in future studies.

4. Future perspectives toward resolving the ToA

We detected no apparent sign of methodological bias in the phylogenetic inferences from nuc276 as long as Rhodelphida and/or Picozoa are present in the supermatrix. On the other hand, the analyses of pld54 demonstrated that this supermatrix is fundamentally short to resolve the relationship among red, glauco, and green plastids, but bears the heterotachy introduced by the complex plastids, which overrates the phylogenetic affinity between red and green plastids. Thus, based on the results from the phylogenetic analyses of nuc276 and pld54, we favor the Rhodo-basal tree as the working hypothesis on the ToA for future examinations. Apparently, this conclusion is tentative and should be examined rigorously in future studies. In this section, we evaluate potential approaches toward resolving the ToA in the future.

One can argue that the issue of insufficient resolution in pld54 can be overcome by preparing and analyzing new pld-supermatrices larger than pld54 in size. There is a large difference in size between nuc276 and pld54 (i.e., 94,907 and 12,787 amino acid positions), and such a size difference likely resulted in the uncertainty remaining in the

phylogenetic inference from pld54. In theory, we can generate a larger supermatrix by combining the proteins encoded in plastid genomes and nucleus-encoded plastid proteins in archaeplastid species. We do not exclude such ideas, but need to be extremely cautious about which proteins to include, as they must have been inherited vertically from the cyanobacterial endosymbiont without any replacement by a distantly related homologue.

Besides phylogenetic analyses of pld-proteins and nuc-proteins, the ToA can be inferred from the proteins encoded in mitochondrial genomes (mt-proteins). In reality, alignments comprising multiple mt-proteins (mt-supermatrices) may suffer from the issue of size, as seen in pld-supermatrices (see above). Because of the number of mt-proteins being restricted, the size of a mt-supermatrix cannot be compatible with that of nuc276 or even smaller than pld54. For instance, some of us prepared and analyzed an alignment of 11 mt-proteins (3,533 amino acid positions in total) to assess whether *Palpitomonas bilix*, a kathablepharid, goniomonads, and cryptophyte algae formed a clade (Nishimura et al. 2020). Both ML and Bayesian trees recovered the monophyly of the species of interest (i.e., cryptists) but failed to group Chloroplastida, Glaucophyta, and Rhodophyta into a single clade. The phylogenetic inference from a similar mt-supermatrix (14 mt-proteins; 3,267 amino acid positions) recovered the monophyly of Archaeplastida, but the statistical support from the ML method was poor (Jackson and Reyes-Prieto 2014). Thus, mt-supermatrices are highly unlikely to be appropriate for inferring the ToA with confidence.

In the end, we contemplate that novel members of Archaeplastida, if they exist, may play a pivotal role in inferring the ToA. Such novel species are not totally speculative, as both Rhodelphida and Picozoa were documented as the early-branching species in Archaeplastida only recently (Gawryluk et al. 2019; Schön et al. 2021). Thus, novel members, which are photosynthetic and branch at the base of Chloroplastida, Glaucophyta, or Rhodophyta, may have been overlooked in natural environments. For instance, suppose that the Rhodo-basal tree of Archaeplastida is genuine, and the analyses of pld-supermatrices conducted to date falsely disfavored the ‘true’ tree. The pld-protein data of photosynthetic species currently unknown to science may change the relationship among red, glauco, and green plastids drastically or at least help the Rhodo-basal tree compete with the other two trees in the analyses of pld-supermatrices. On the contrary, there still remains the possibility of the analyses of nuc-supermatrices being biased by

unknown reasons, and thus, the Glauco-basal or Chloro-basal tree can be the genuine ToA. If so, as-yet-unknown species, particularly those branched at the base of Chloroplastida or Glaucophyta, may suppress the dominance of the Rhodo-basal tree in the analyses of the current nuc-supermatrices. In conclusion, the incongruity between the analyses of nuc- and pld-supermatrices can be ameliorated or even resolved if the novel, early branching members of Archaeplastida are identified in the future.

5. Acknowledgements

Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics. This work was supported by grants from the Japan Society for the Promotion of Science (JSPS) (23K27226 and BPI06050 awarded to Y. I., and 24K09587 to T. N.). R. H. was supported by the JSPS Overseas Research Fellowship.

6. Data availability

Supplementary data are available at https://docs.google.com/document/d/1B9cu68vf0mTt37_XcUYuFHjWzxsL4dmQQzRAzFTw7OI/edit?usp=sharing.

7. References

Adl S.M., Simpson A.G.B., Farmer M.A., Andersen R.A., Anderson O.R., Barta J.R., Bowser S.S., Brugerolle G., Fensome R.A., Fredericq S., James T.Y., Karpov S., Kugrens P., Krug J., Lane C.E., Lewis L.A., Lodge J., Lynn D.H., Mann D.G., Mccourt R.M., Mendoza L., Moestrup Ø., Mozley-Standridge S.E., Nerad T.A., Shearer C.A., Smirnov A.V., Spiegel F.W., Taylor M.F.J.R. 2005. The New Higher Level Classification of Eukaryotes with Emphasis on the Taxonomy of Protists. J. Eukaryot. Microbiol. 52:399–451.

Adl S.M., Bass D., Lane C.E., Lukeš J., Schoch C.L., Smirnov A., Agatha S., Berney C., Brown M.W., Burki F., Cárdenas P., Čepička I., Chistyakova L., del Campo J., Dunthorn M., Edvardsen B., Eglit Y., Guillou L., Hampl V., Heiss A.A., Hoppenrath M., James T.Y., Karnkowska A., Karpov S., Kim E., Kolisko M., Kudryavtsev A., Lahr D.J.G., Lara E., Le Gall L., Lynn D.H., Mann D.G., Massana R., Mitchell E.A.D., Morrow C., Park J.S., Pawlowski J.W., Powell M.J., Richter D.J., Rueckert S., Shadwick L., Shimano S., Spiegel F.W., Torruella G., Youssef N., Zlatogursky V., Zhang Q. 2019. Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. J. Eukaryot. Microbiol. 66:4–119.

Archibald, J.M. 2009. The puzzle of plastid evolution. Curr. Biol. 19:R18–R88. 10.1016/j.cub.2008.11.067

Burki F., Kaplan M., Tikhonenkov D.V., Zlatogursky V., Minh B.Q., Radaykina L.V.,

Smirnov A., Mylnikov A.P., Keeling P.J. 2016. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. Proc. R. Soc. B Biol. Sci. 283:20152802.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., Madden T.L. 2009. BLAST+: architecture and applications. BMC Bioinformat. 10:421.

Capella-Gutiérrez S., Silla-Martínez J.M., Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

Carpenter E.J., Matasci N., Ayyampalayam S., Wu S., Sun J., Yu J., Jimenez Vieira F.R., Bowler C., Dorrell R.G., Gitzendanner M.A., Li L., Du W., K. Ullrich K., Wickett N.J., Barkmann T.J., Barker M.S., Leebens-Mack J.H., Wong G.K.-S. 2019. Access to RNA-sequencing data from 1,173 plant species: The 1000 Plant transcriptomes initiative (1KP). GigaScience 8:giz126.

Chen S. 2023. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta. 2:e107.

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

Crotty S.M., Minh B.Q., Bean N.G., Holland B.R., Tuke J., Jermiin L.S., Haeseler A.V. 2020. GHOST: Recovering Historical Signal from Heterotachously Evolved Sequence Alignments. Syst. Biol. 69:249–264.

Eglit Y., Shiratori T., Jerlström-Hultqvist J., Williamson K., Roger A.J., Ishida K.-I., Simpson A.G.B. 2024. *Meteora sporadica*, a protist with incredible cell architecture, is related to Hemimastigophora. Curr. Biol. 34:451-459.

Figueroa-Martinez F., Jackson C., Reyes-Prieto A. 2019. Plastid Genomes from Diverse Glaucophyte Genera Reveal a Largely Conserved Gene Content and Limited Architectural Diversity. Genome Biol. Evol. 11:174–188.

Fu L., Niu B., Zhu Z., Wu S., Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152.

Gawryluk R.M.R., Tikhonenkov D.V., Hehenberger E., Husnik F., Mylnikov A.P., Keeling P.J. 2019. Non-photosynthetic predators are sister to red algae. Nature 572:240–243.

Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., di Palma F., Birren B.W., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A. 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat. Biotechnol. 29:644–652.

Haas B. TransDecoder. Available from https://github.com/TransDecoder/TransDecoder.

Harada R., Hirakawa Y., Yabuki A., Kim E., Yazaki E., Kamikawa R., Nakano K., Eliáš M., Inagaki Y. 2024. Encyclopedia of Family A DNA Polymerases Localized in Organelles: Evolutionary Contribution of Bacteria Including the Proto-Mitochondrion. Mol. Biol. Evol. 41:msae014.

Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Vinh L.S. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. 35:518–522.

Irisarri I., Strassert J.F.H., Burki F. 2021. Phylogenomic Insights into the Origin of Primary Plastids. Syst. Biol. 71:105–120. 10.1093/sysbio/syab036

Jackson C.J., Reyes-Prieto A. 2014. The mitochondrial genomes of the glaucophytes

Gloeochaete wittrockiana and *Cyanoptyche gloeocystis*: Multilocus phylogenetics suggests a monophyletic Archaeplastida. Genome Biol. Evol. 6:2774–2785. 10.1093/gbe/evu218

Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14:587–589.

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

Keeling P.J. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. 2013. Ann. Rev. Plant Biol. 64:583–607. 10.1146/annurev-arplant-050312-120144

Lartillot N., Philippe H. 2004. A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. Mol. Biol. Evol. 21:1095–1109.

Lartillot N., Rodrigue N., Stubbs D., Richer J. 2013. PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. Syst. Biol. 62:611–615.

Leebens-Mack J.H., Barker M.S., Carpenter E.J., Deyholos M.K., Gitzendanner M.A., Graham S.W., Grosse I., Li Z., Melkonian M., Mirarab S., Porsch M., Quint M., Rensing S.A., Soltis D.E., Soltis P.S., Stevenson D.W., Ullrich K.K., Wickett N.J., DeGironimo L., Edger P.P., Jordon-Thaden I.E., Joya S., Liu T., Melkonian B., Miles N.W., Pokorny L., Quigley C., Thomas P., Villarreal J.C., Augustin M.M., Barrett M.D., Baucom R.S., Beerling D.J., Benstein R.M., Biffin E., Brockington S.F., Burge D.O., Burris J.N., Burris K.P., Burtet-Sarramegna V., Caicedo A.L., Cannon S.B., Çebi Z., Chang Y., Chater C., Cheeseman J.M., Chen T., Clarke N.D., Clayton H., Covshoff S., Crandall-Stotler B.J., Cross H., dePamphilis C.W., Der J.P., Determann R., Dickson R.C., Di Stilio V.S., Ellis S., Fast E., Feja N., Field K.J., Filatov D.A., Finnegan P.M., Floyd S.K., Fogliani B., García N., Gâteblé G., Godden G.T., Goh F. (Qi Y., Greiner S., Harkess A., Heaney J.M., Helliwell K.E., Heyduk K., Hibberd J.M., Hodel R.G.J., Hollingsworth P.M., Johnson M.T.J., Jost R., Joyce B., Kapralov M.V., Kazamia E., Kellogg E.A., Koch M.A., Von Konrat M., Könyves K., Kutchan T.M., Lam V., Larsson A., Leitch A.R., Lentz R., Li F.-W., Lowe A.J., Ludwig M., Manos P.S., Mavrodiev E., McCormick M.K., McKain M., McLellan T., McNeal J.R., Miller R.E., Nelson M.N., Peng Y., Ralph P., Real D., Riggins C.W., Ruhsam M., Sage R.F., Sakai A.K., Scascitella M., Schilling E.E., Schlösser E.-M., Sederoff H., Servick S., Sessa E.B., Shaw A.J., Shaw S.W., Sigel E.M., Skema C., Smith A.G., Smithson A., Stewart C.N., Stinchcombe J.R., Szövényi P., Tate J.A., Tiebel H., Trapnell D., Villegente M., Wang C.-N., Weller S.G., Wenzel M., Weststrand S., Westwood J.H., Whigham D.F., Wu S., Wulff A.S., Yang Y., Zhu D., Zhuang C., Zuidof J., Chase M.W., Pires J.C., Rothfels C.J., Yu J., Chen C., Chen L., Cheng S., Li J., Li R., Li X., Lu H., Ou Y., Sun X., Tan X., Tang J., Tian Z., Wang F., Wang J., Wei X., Xu X., Yan Z., Yang F., Zhong X., Zhou F., Zhu Y., Zhang Y., Ayyampalayam S., Barkman T.J., Nguyen N., Matasci N., Nelson D.R., Sayyari E., Wafula E.K., Walls R.L., Warnow T., An H., Arrigo N., Baniaga A.E., Galuska S., Jorgensen S.A., Kidder T.I., Kong H., Lu-Irving P., Marx H.E., Qi X., Reardon C.R., Sutherland B.L., Tiley G.P., Welles S.R., Yu R., Zhan S., Gramzow

L., Theißen G., Wong G.K.-S., One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. Nature 574:679–685.

Li W., Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22:1658–1659.

Lopez P., Casane D., Philippe H. 2002. Heterotachy, an Important Process of Protein Evolution. Mol. Biol. Evol. 19:1–7.

Mackiewicz P., Gagat P. 2014. Monophyly of Archaeplastida supergroup and relationships among its lineages in the light of phylogenetic and phylogenomic studies. Are we close to a consensus? Acta Soc. Bot. Pol. 83:263–280.

Macorano L, Nowack E.C.M. 2021. *Paulinella chromatophora*. Curr. Biol. 31:R1024–R1026. 10.1016/j.cub.2021.07.028

Marin B., M. Nowack E.C., Melkonian M. 2005. A Plastid in the Making: Evidence for a Second Primary Endosymbiosis. Protist 156:425–432.

Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., von Haeseler A., Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37:1530–1534.

Nishimura Y., Kume K., Sonehara K., Tanifuji G., Shiratori T., Ishida K., Hashimoto T., Inagaki Y., Ohkuma M. 2020. Mitochondrial Genomes of *Hemiarma marina* and *Leucocryptos marina* Revised the Evolution of Cytochrome *c* Maturation in Cryptista. Front. Ecol. Evol. 8:140.

Ponce-Toledo R.I., Deschamps P., López-García P., Zivanovic Y., Benzerara K., Moreira D. 2017. An Early-Branching Freshwater Cyanobacterium at the Origin of Plastids. Curr. Biol. 27:386–391.

Rodríguez-Ezpeleta N., Brinkmann H.,Burey S.C., Roure B**.,** Burger G., Löffelhardt W., Bohnert H.J., Philippe H., Lang B.F. 2005.Monophyly of Primary Photosynthetic Eukaryotes: Green Plants, Red Algae, and Glaucophytes. Curr. Biol. 15:1325–1330. 10.1016/j.cub.2005.06.040

Schön M.E., Zlatogursky V.V., Singh R.P., Poirier C., Wilken S., Mathur V., Strassert J.F.H., Pinhassi J., Worden A.Z., Keeling P.J., Ettema T.J.G., Wideman J.G., Burki F. 2021. Single cell genomics reveals plastid-lacking Picozoa are close relatives of red algae. Nat. Commun. 12:6651.

Sibbald, S.J., Archibald, J.M. 2020. Genomic insights into plastid evolution. Genome Biol. Evol. 12:978–990. 10.1093/gbe/evaa096

Shimodaira H. 2002. An Approximately Unbiased Test of Phylogenetic Tree Selection. Syst. Biol. 51:492–508.

Si Quang L., Gascuel O., Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24:2317–2323.

Soubrier J., Steel M., Lee M.S.Y., Der Sarkissian C., Guindon S., Ho S.Y.W., Cooper A. 2012. The influence of rate heterogeneity among sites on the time dependence of molecular rates. Mol. Biol. Evol. 29:3345–3358. 10.1093/molbev/mss140

Stairs C.W., Táborský P., Salomaki E.D., Kolisko M., Pánek T., Eme L., Hradilová M., Vlček Č., Jerlström-Hultqvist J., Roger A.J., Čepička I. 2021. Anaeramoebae are a divergent lineage of eukaryotes that shed light on the transition from anaerobic mitochondria to hydrogenosomes. Curr. Biol. 31:5605-5612.e5.

Strassert J.F.H., Jamy M., Mylnikov A.P., Tikhonenkov D.V., Burki F. 2019. New Phylogenomic Analysis of the Enigmatic Phylum Telonemia Further Resolves the

Eukaryote Tree of Life. Mol. Biol. Evol. 36:757–765.
Stephens T.G., Gabr A., Calatrava V., Grossman A.R., Bhattacharya D. 2021 Why is primary endosymbiosis so rare? New Phytol. 231:1693–1699. 10.1016/j.cub.2021.07.028
Strunecký O., Ivanova A.P., Mareš J. 2023. An updated classification of cyanobacterial orders and families based on phylogenomic and polyphasic analysis. J. Phycol. 59:12–51.
Tice A.K., Žihala D., Pánek T., Jones R.E., Salomaki E.D., Nenarokov S., Burki F., Eliáš M., Eme L., Roger A.J., Rokas A., Shen X.-X., Strassert J.F.H., Kolísko M., Brown M.W. 2021. PhyloFisher: A phylogenomic package for resolving eukaryotic relationships. PLOS Biol. 19:e3001365.
Tikhonenkov D.V., Mikhailov K.V., Gawryluk R.M.R., Belyaev A.O., Mathur V., Karpov S.A., Zagumyonnyi D.G., Borodina A.S., Prokina K.I., Mylnikov A.P., Aleoshin V.V., Keeling P.J. 2022. Microbial predators form a new supergroup of eukaryotes. Nature 612:714–719.
Wang H.-C., Minh B.Q., Susko E., Roger A.J. 2018. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. Syst. Biol. 67:216–235.
Yang Z, Kumar S. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. 13:650–659. 10.1093/oxfordjournals.molbev.a025625
Yazaki E., Yabuki A., Imaizumi A., Kume K., Hashimoto T., Inagaki Y. 2022. The closest lineage of Archaeplastida is revealed by phylogenomics analyses that include *Microheliella maris*. Open Biol. 12:210376.

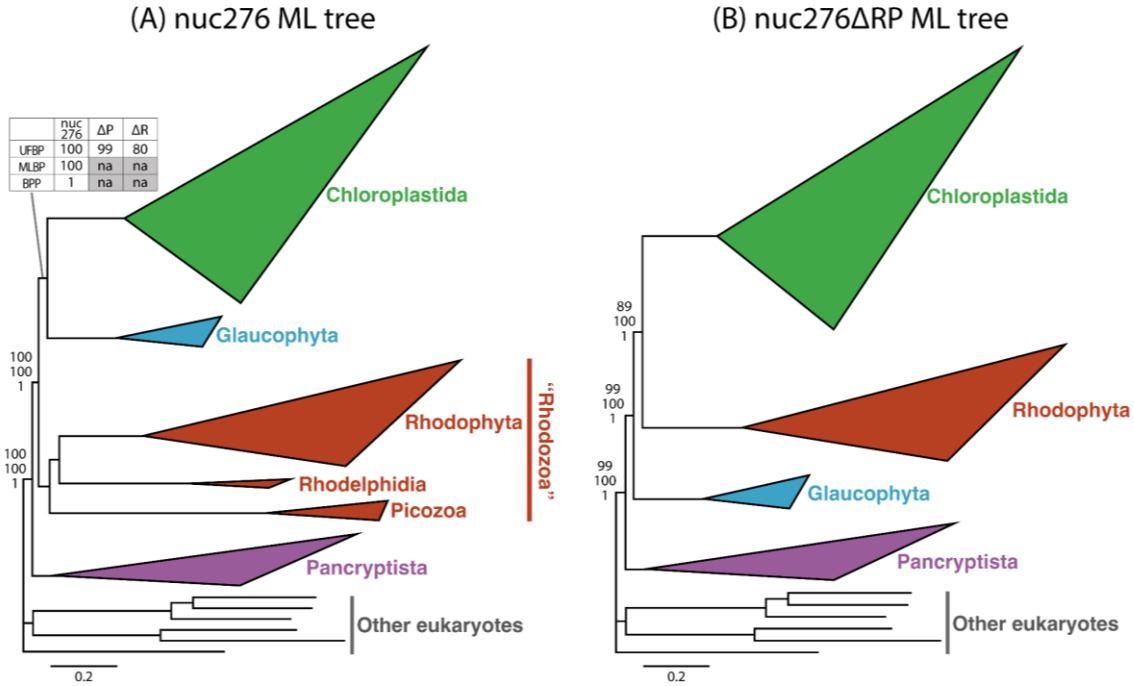

Figure 1. The trees of Archaeplastida inferred from nuc276 and nuc276ΔRP. The phylogenetic trees were recovered by the maximum likelihood (ML) method with the LG+C60+F+G model. The tree inferred from nuc276 and that from nuc276ΔRP are shown in (A) and (B), respectively. The sub-clades in Archaeplastida and Pancryptista were simplified into triangles. The size of each triangle is scaled to the number of the OTUs comprising the clade, and the longest and shortest distances from the ancestral to terminal nodes. All OTU names were omitted. We also omitted the support values for all bipartitions, except the values for (i) the CAM clade, (ii) the monophyly of Archaeplastida, and (iii) the basal branching of Rhodozoa or Glaucophyta (see Figs. S2 and S3 for the details). The support values shown on the top, middle, and bottom are ultrafast bootstrap percent values (UFBPs), ML bootstrap values (MLBPs), and Bayesian posterior probabilities (BPPs), respectively. The overall ML tree topologies inferred from nuc276ΔP and nuc276ΔR are essentially the same as that from nuc276 (see Fig. S4 for the details). The UFBPs for the bipartition uniting Chloroplastida and Glaucophyta calculated from nuc276ΔP and nuc276ΔR are listed in a table in (A).

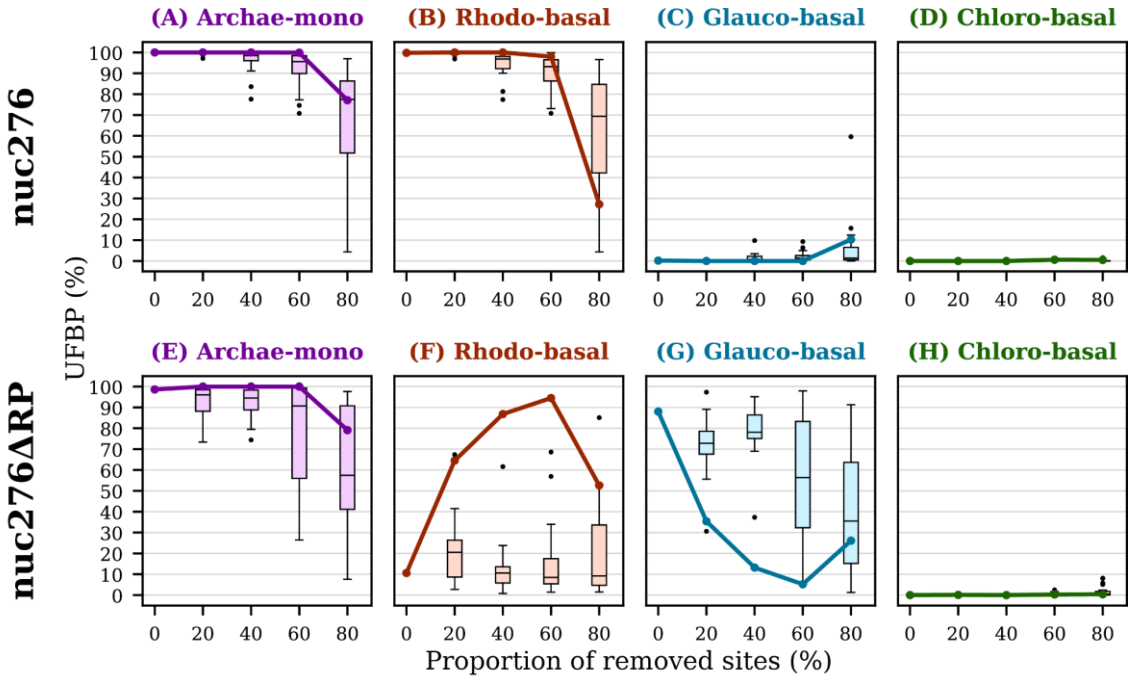

Figure 2. Analyses of nuc276 and nuc276ΔRP processed by the fast-evolving position removal (FPR) and random position removal (RPR) analyses. We removed the top 20%, 40%, 60%, and 80% of the fastest-evolving positions from nuc276 and nuc276ΔRP. The FPR-processed nuc276 and nuc276ΔRP were subjected to the ML tree search and UFBoot2 (1,000 replications). The ultrafast bootstrap percent values (UFBPs) for the monophyly of Archaeplastida (Archae-mono), the basal branching of Rhodozoa/Rhodophyta (Rhodo-basal), the basal branching of Glaucophyta (Glauco-basal), and the basal branching of Chloroplastida (Chloro-basal) are plotted as line graphs in (A/E), (B/F), (C/G), and (D/H), respectively. In the ultrafast bootstrap analyses described here, we defined "Rhodo-basal" trees satisfying both the monophyly of Archaeplastida and the separation of Rhodozoa/Rhodophyta from Glaucophyta and Chloroplastida. Likewise, "Glauco-basal" trees satisfied the monophyly of Archaeplastida and the separation of Glaucophyta from Rhodozoa/Rhodophyta and Chloroplastida. Finally, we selected the trees satisfying both the monophyly of Archaeplastida and the separation of Chloroplastida from Glaucophyta and Rhodozoa/Rhodophyta as "Chloro-basal" trees. The box plots represent the UFBPs of "Archae-mono" (A/E), "Rhodo-basal" (B/F), "Glauco-basal" (C/G), and "Chloro-basal" (D/H) that were calculated from the RPR-processed nuc276 and nuc276ΔRP.

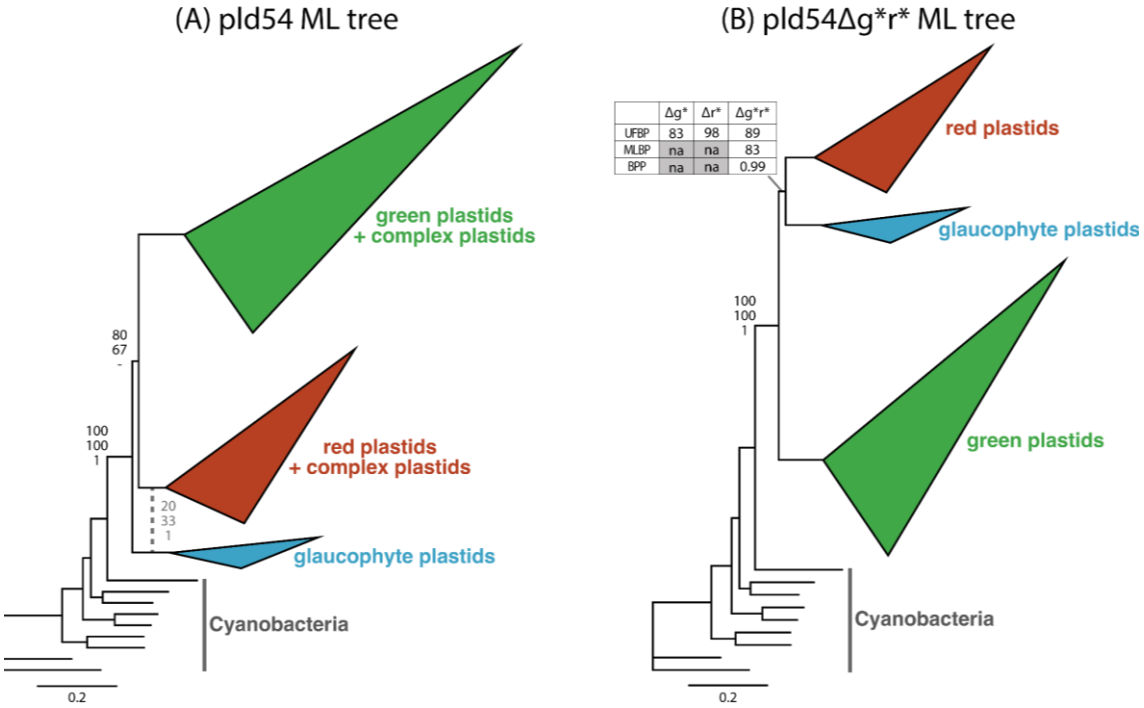

Figure 3. The trees of plastids inferred from pld54 and pldΔg*r*. The phylogenetic trees were recovered by the maximum likelihood (ML) method with the cpREV+C60+F+I+G model. The tree inferred from pld54 and that from pld54Δg*r* are shown in (A) and (B), respectively. The three sub-clades in the plastid clade were simplified into triangles, as done in Fig. 1. We also omitted the support values for all bipartitions, except the support values for the monophyly of plastids and the basal branching of glaucophyte or green plastids. The support values shown on the top, middle, and bottom are ultrafast bootstrap percent values (UFBPs), ML bootstrap values (MLBP), and Bayesian posterior probabilities (BPPs), respectively. A dotted line in (A) represents the union of the clade of red plastids and red alga-derived complex plastids (red* plastids) and the clade of glauco plastids in Bayesian analyses of pld54. The UFBP, MLBP, and BPP for the bipartition represented by the dotted line are shown in grey. The overall ML tree topologies inferred from pld54Δg* and pld54Δr* are essentially the same as that from pld54Δg*r* (see Figs. S10A & B). Thus, the UFBPs for the bipartition uniting red and glaucophyte plastids calculated from pld54Δg* and pld54Δr* are listed in a table in (B).

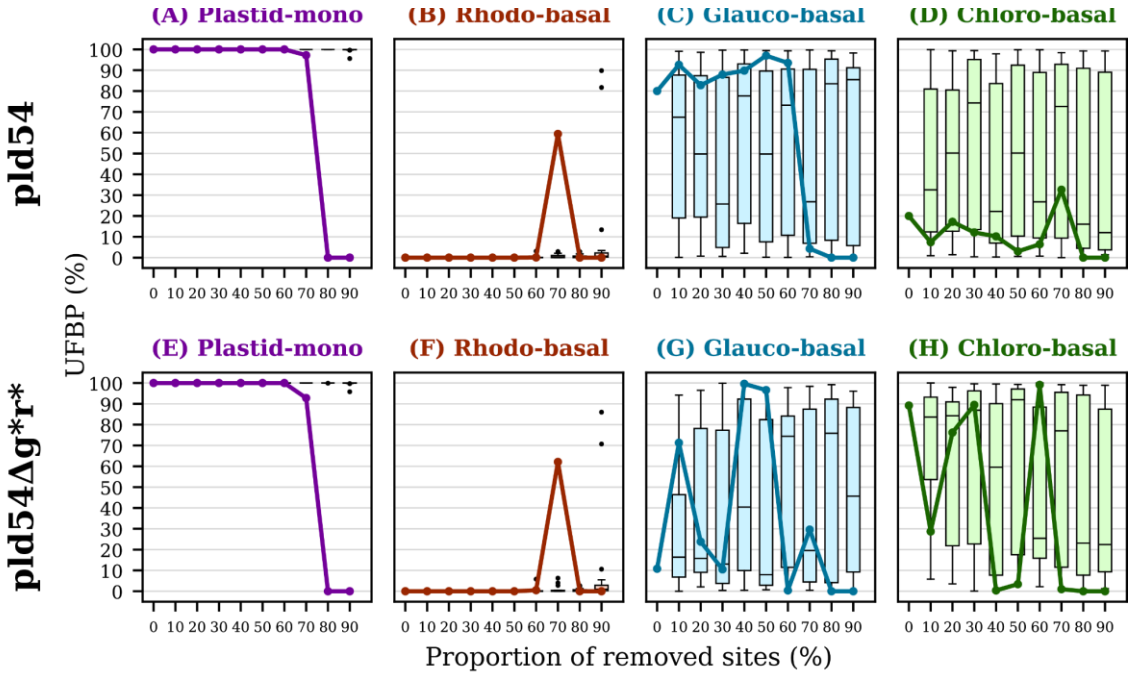

Figure 4. Analyses of pld54 and pld54Δg*r* processed by the fast-evolving position removal (FPR) and random position removal (RPR). We removed the top 10-90% of the fastest-evolving positions from pld54 and pld54Δg*r* at 10% intervals. The FPR-processed pld54 and pld54Δg*r* were subjected to the ML tree search and UFBoot2 (1,000 replications). The ultrafast bootstrap percent values (UFBPs) for the monophyly of plastids (Plastid-mono), the basal branching of red plastids with or without red alga-derived complex plastids (Rhodo-basal), the basal branching of glauco plastids (Glauco-basal), and the basal branching of green plastids with or without green alga-derived complex plastids (Chloro-basal) are plotted as line graphs in (A/E), (B/F), (C/G), and (D/H), respectively. In the ultrafast bootstrap analyses described here, we defined "Rhodo-basal" trees satisfying both the monophyly of plastids and the separation of red plastids with or without red alga-derived complex plastids from the other plastids considered. Likewise, "Glauco-basal" trees satisfied the monophyly of plastids and the separation of glauco plastids from the other plastids considered. Finally, we selected the trees satisfying both the monophyly of plastids and the separation of green plastids with or without green alga-derived complex plastids from the other plastids considered as "Chloro-basal" trees. The box plots represent the UFBPs of "Plastid-mono" (A/E), "Rhodo-basal" (B/F), "Glauco-basal" (C/G), and "Chloro-basal" (D/H) that were calculated from the RPR-processed pld54 and pld54Δg*r*.

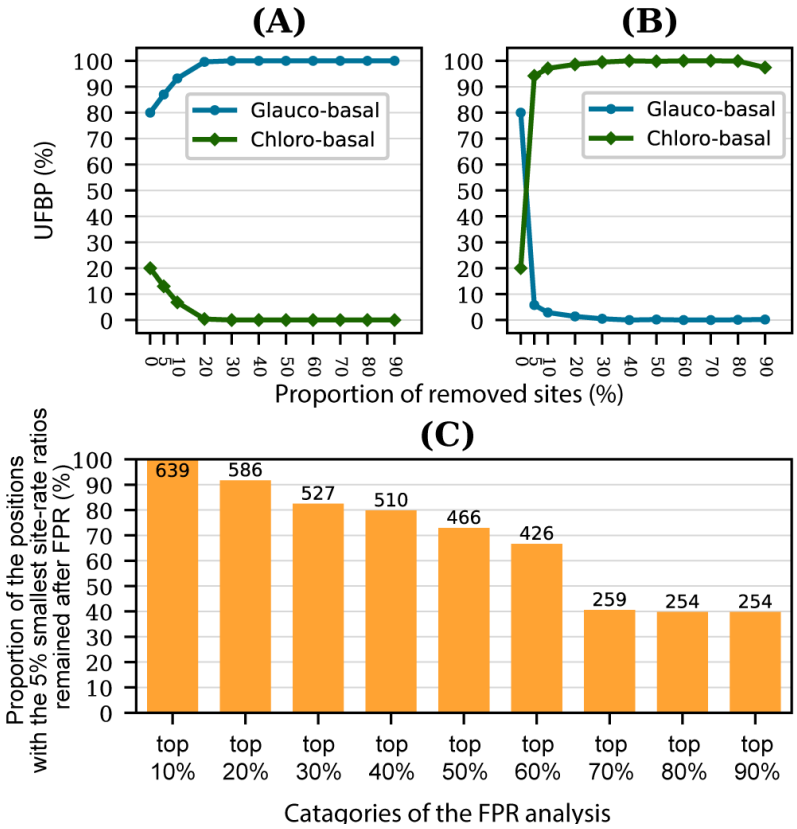

Figure 5. Analyses of pld54 processed by the progressive removal of the alignment positions based on site-rate ratios. The site-rates calculated based on pld54 were divided by those based on pld54Δg*r* to obtain site-rate ratios. In (A), we progressively removed the alignment positions with the 5-90% largest site-rate ratios from pld54, subjected each of the modified pld54 to the ML tree search coupled with UFBoot2, and plotted the ultrafast bootstrap values (UFBPs) for the Glauco-basal and Chloro-basal trees. We presented the results from the same analyses, but the removal of the alignment positions according to the 5-90% smallest site-rate ratios, in (B). The bar graph shown in (C) indicates the percent proportion of the alignment positions with the 5% smallest site-rate ratios remaining in the FPR-processed alignments. The figures associated with the bars indicate the actual numbers of the 5% largest site-rate ratios remaining in the FPR-processed pld54. For instance, the alignment positions after the exclusions of the top 10 and 60% fastest-evolving positions contained all (i.e., 639) and 426 positions of the 5% smallest site-rate ratios, respectively.

Table 1. Supermatrices that were used for phylogenomic analyses in this study.

Supermatrix	Number of OTUs	Number of proteins[1]	Notes
nuc317	133	317 nuc-proteins	A supermatrix to infer the tree of global eukaryotes
nuc276	57	276 nuc-proteins	A supermatrix to infer the tree of Archaeplastida
nuc276ΔRP	52	276 nuc-proteins	nuc276 without Picozoa or Rhodelphidia
nuc276ΔP	54	276 nuc-proteins	nuc276 without Picozoa
nuc276ΔR	55	276 nuc-proteins	nuc276 without Rhodelphidia
nuc276ΔP*	54	276 nuc-proteins	Same as nuc276ΔP, but the average coverage of the rhodelphid sequences was lowered to 45.16%
nuc276ΔP**	54	276 nuc-proteins	Same as nuc276ΔP, but the average coverage of the rhodelphid sequences was lowered to 49.99%
pld54	57	54 pld-proteins	A supermatrix to infer the tree of plastids
pld54Δg*r*	51	54 pld-proteins	pld54 without six complex plastids
pld54Δr*	53	54 pld-proteins	pld54 without red alga-derived complex plastids
pld54Δg*	55	54 pld-proteins	pld54 without green alga-derived complex plastids

[1] nuc-proteins, nucleus-encoded proteins; pld-proteins, plastid-encoded proteins.

Table 2. Comparisons of the three possible trees of Archaeplastida by AU tests based on nuc276 and its derivatives.

Supermatrix	Basal branch in the Archaeplastida clade	Difference in log likelihood[1]	*p* value[2]
nuc276	**Rhodozoa**	ML (-4429234.956)	-
	Glaucophyta	131.18	0.00133**
	Chloroplastida	222.63	2.17e-11**
nuc276ΔRP	Rhodophyta	39.842	0.163
	Glaucophyta	ML (-4098432.614)	-
	Chloroplastida	128.71	2.13e-05**
nuc276ΔP	**Rhodozoa**	ML (-4279550.183)	-
	Glaucophyta	72.172	0.0303*
	Chloroplastida	138.42	2.92e-05**
nuc276ΔR	**Rhodozoa**	ML (-4248544.063)	-
	Glaucophyta	32.038	0.23
	Chloroplastida	178.29	7.08e-99**
nuc276ΔP*	**Rhodozoa**	ML (-4180767.531)	-
	Glaucophyta	22.552	0.288
	Chloroplastida	120.11	0.000375**
nuc276ΔP**	**Rhodozoa**	ML (-4197252.482)	-
	Glaucophyta	20.888	0.309
	Chloroplastida	110.61	0.000114**

[1] For each ML tree, the corresponding log-likelihood value was given in parentheses.
[2] Double and single asterisks indicate that the AU test rejected the null hypothesis of no difference for the ML tree with $p < 0.01$ and < 0.05, respectively.

Table 3. Comparisons of the three possible trees of Archaeplastida by AU tests based on pld54 and its derivatives.

Supermatrix	Basal branch in the plastid clade[1]	Difference in log likelihood[2]	*p* value[3]
pld54	red plastids+	36.703	0.000365**
	glauco plastids	ML (-553602.8714)	-
	green plastids+	0.70825	0.486
pld54Δg*r*	red plastids	38.506	3.09e-07**
	glauco plastids	3.631	0.424
	green plastids	ML (-489684.9682)	-
pld54Δg*	red plastids+	36.413	1.2e-05**
	glauco plastids	0.85283	0.477
	green plastids	ML (-529919.7079)	-
pld54Δr*	red plastids	36.382	8.52e-06**
	glauco plastids	1.8319	0.461
	green plastids+	ML (-513445.9583)	-
pld54 after excluding the potential heterotachious positions	red plastids+	29.303	0.00351**
	glauco plastids	10.944	0.231
	green plastids+	ML (-513181.3828)	-

[1] Plus signs indicate red or green alga-derived complex plastids were included.
[2] For each ML tree, the corresponding log-likelihood value is given in parentheses.
[3] Double and single asterisks indicate that the AU test rejected the null hypothesis of no difference for the ML tree with $p < 0.01$ and < 0.05, respectively.

able 4. Comparisons of the three possible trees of Archaeplastida by AU tests with the FreeRate and GHOST models.

Alignment	Substitution model (BIC[4])	Basal branch in the Archaeplastida / plastid clade[1]	Difference in log likelihood[2]	*p* value[3]
nuc276	LG+C60+F+G (8,860,658.897)	**Rhodozoa**	ML (-4429234.956)	-
		Glaucophyta	131.18	0.00133**
		Chloroplastida	222.63	2.17e-11**
	LG+C60+F+R9 (8,843,760.388)	**Rhodozoa**	ML (-4420699.747)	-
		Glaucophyta	113.18	0.0026**
		Chloroplastida	196.28	1.29e-87**
	LG+C60+F+H4 (8,850,693.831)	**Rhodozoa**	ML (-4422332.764)	-
		Glaucophyta	140.01	0.000224**
		Chloroplastida	226.71	2.55e-06**
	LG+C60+F+H8 (8,823,860.076)	**Rhodozoa**	ML (-4406348.7)	-
		Glaucophyta	89.708	0.0318*
		Chloroplastida	175.88	1.64e-50**
pld54	cpREV+C60+F+I+G (1,109,021.330)	red plastids+	36.703	0.000365**
		glauco plastids	ML (-553602.8714)	-
		green plastids+	0.70825	0.486
	cpREV+C60+F+I+R7 (1,106,108.846)	red plastids+	37.529	2.45e-05**
		glauco plastids	ML (-552094.6204)	-
		green plastids	1.3258	0.47
	cpREV+C60+F+I+H4 (1,105,792.557)	red plastids+	37.785	1.35e-05**
		glauco plastids	ML (-550404.574)	-
		green plastids+	3.48	0.439
	cpREV+C60+F+I+H8 (1,102,983.706)	**red plastids+**	ML (-546860.8501)	-
		glauco plastids	21.113	0.231
		green plastids+	60.204	0.0123*

[1]Plus signs indicate the inclusion of red or green complex plastids.
[2]For each ML tree, the corresponding log-likelihood value is given in parentheses.
[3]Double and single asterisks indicate that the AU test rejected the null hypothesis of no difference for the ML tree with p <0.01 and <0.05, respectively.
[4]Bayesian information criterion values were calculated from the log-likelihoods of the Rhodo-basal tree inferred from nuc276, or that of the Glauco-basal tree inferred from pld54.