Picozoa are archaeplastids without plastid

Max E. Schön[1,2], Vasily V. Zlatogursky[1], Rohan P. Singh[3], Camille Poirier[4,5], Susanne Wilken[5], Varsha Mathur[6], Jürgen F. H. Strassert[1], Jarone Pinhassi[7], Alexandra Z. Worden[4,5], Patrick J. Keeling[6], Thijs J. G. Ettema[8], Jeremy G. Wideman[3], Fabien Burki[1,9]*

[1]Department of Organismal Biology, Program in Systematic Biology, Uppsala University, Uppsala, Sweden

[2]Department of Cell and Molecular Biology, Program in Molecular Evolution, Uppsala University, Uppsala, Sweden

[3]Biodesign Center for Mechanisms of Evolution, School of Life Sciences, Arizona State University, Tempe, AZ, USA

[4]Ocean EcoSystems Biology, RD3, GEOMAR Helmholtz Centre for Ocean Research Kiel, Germany

[5]Monterey Bay Aquarium Research Institute, Moss Landing, CA, USA

[6]Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada

[7]Centre for Ecology and Evolution in Microbial Model Systems – EEMiS, Linnaeus University, Kalmar, Sweden

[8]Laboratory of Microbiology, Wageningen University and Research, NL-6708 WE Wageningen, The Netherlands

[9]Science for Life Laboratory, Uppsala University, Uppsala, Sweden

Present address (VVZ): Department of Invertebrate Zoology, Faculty of Biology, St. Petersburg State University, Russia

Present address (JFHS): Department of Ecosystem Research, Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany

*Corresponding author: fabien.burki@ebc.uu.se

Abstract

The endosymbiotic origin of plastids from cyanobacteria gave eukaryotes photosynthetic capabilities and launched the diversification of countless forms of algae. These primary plastids are found in members of the eukaryotic supergroup Archaeplastida, and are widely assumed to have a single origin. Here, we used single-cell genomics from natural samples combined with phylogenomics to infer the evolutionary origin of the phylum Picozoa, a globally distributed but seemingly rare group of marine microbial heterotrophic eukaryotes. Strikingly, we find based on the analysis of 43 single-cell genomes that Picozoa belong to Archaeplastida as a robust sister group to the clade containing red algae and the phagotrophic rhodelphids. Our analyses of this extensive data support the hypothesis that Picozoa lack a plastid, and further show no evidence for an early cryptic endosymbiosis with cyanobacteria. The position of Picozoa in the eukaryotic tree represents the first known case of a plastid-lacking lineage closely related to one of the main archaeplastid branches. The implications of these findings for our understanding of plastid evolution are unprecedented, and can either be interpreted as the first report of complete plastid loss in a free-living taxon, or as an indication that red algae and rhodelphids obtained their plastids independently of other archaeplastids.

Introduction

The origin of plastids by endosymbiosis between a eukaryotic host and cyanobacteria is a cornerstone in eukaryotic evolution, giving rise to the first photosynthetic eukaryotes. These ancient primary plastids, estimated to have originated >1.8 billion years ago (Strassert et al. 2021), are found in Rhodophyta (red algae), Chloroplastida (green algae, including land plants), and Glaucophyta (glaucophytes)—together forming the eukaryotic supergroup Archaeplastida (Burki et al. 2020). Unravelling the complex sequence of events leading to the establishment of the cyanobacterial endosymbiont in Archaeplastida is complicated by antiquity, and by the current lack of modern descendants of early-diverging relatives of the main archaeplastidan groups in culture collections or sequence databases. Indeed, the only other known example of primary endosymbiosis are the chromatophores in one unrelated genus of amoeba (*Paulinella*), which originated about a billion years later (Marin, Nowack, and Melkonian 2005; Nowack and Weber 2018). Recently, two newly described major phyla (Prasinodermophyta and Rhodelphidia) were found to branch as sister to green and red algae, respectively (Gawryluk et al. 2019; Li et al. 2020). Most transformative was the discovery that rhodelphids are obligate phagotrophs that maintain cryptic non-photosynthetic plastids, implying that the ancestor of red algae was mixotrophic, a finding that greatly alters our perspectives on early archaeplastid evolution (Gawryluk et al. 2019).

While it has been widely assumed that Archaeplastida is a group derived from a photosynthetic ancestor, non-photosynthetic and plastid-lacking lineages have branched near the base or within archaeplastids in phylogenomic trees. For example, Cryptista (which includes plastid-lacking and secondary plastid-containing species), was inferred as sister to green algae and glaucophytes (Burki et al. 2016), or instead branched as sister to red algae (Lax et al. 2018; Gawryluk et al. 2019; Strassert et al. 2019). These relationships, however, are uncertain and recent phylogenomic analyses have instead recovered the monophyly of Archaeplastida to the exclusion of the cryptists (Lax et al. 2018; Irisarri, Strassert, and Burki 2020; Strassert et al. 2021). Another non-photosynthetic group that recently showed affinities to red algae based on phylogenomics is Picozoa (Lax et al. 2018; Gawryluk et al. 2019; Irisarri, Strassert, and Burki 2020). But as for cryptists, the position of Picozoa has lacked consistent support, mostly because there is no member of Picozoa available in continuous culture, and genomic data are currently restricted to a few incomplete single amplified genomes (SAGs) (Yoon et al. 2011). Thus, the origin of Picozoa remains unclear.

Picozoa (previously known as picobiliphytes) were first described in 2007 in marine environmental clone libraries of the 18S ribosomal RNA (rRNA) gene and observed by epifluorescence microscopy in temperate waters (Not et al. 2007). Due to orange autofluorescence reminiscent of the photosynthetic pigment phycobiliprotein and emanating from an organelle-like structure, picozoans were initially described as likely containing a plastid. Orange fluorescence was also observed in association with these uncultured cells in subtropical waters (Cuvelier et al. 2008). However, the hypothesis that the cells were putatively photosynthetic was later challenged by the investigation of SAG data from three picozoan cells isolated by fluorescence-activated cell sorting (FACS) (Yoon et al. 2011). The analysis of these SAGs revealed no plastid DNA or signs of nuclear-encoded plastid-targeted proteins, but the scope of these conclusions is

limited due to the low number of analyzed cells and highly fragmented and incomplete genomes (Yoon et al. 2011). Most interestingly, a transient culture was later established, enabling the formal description of the first (and so far only) picozoan species–*Picomonas judraskeda*–as well as ultrastructural observations with electron microscopy (Seenivasan et al. 2013). These observations revealed an unusual structural feature in two body parts, a feeding strategy by endocytosis of nano-sized colloid particles, and confirmed the absence of plastids (Seenivasan et al. 2013). Only the 18S rRNA gene sequence of *P. judraskeda* is available as the transient culture was lost before genomic data could be generated.

Here, we present an analysis of genomic data from 43 picozoan single-cell genomes sorted with FACS from the Pacific Ocean off the California coast and from the Baltic Sea. Using a gene and taxon-rich phylogenomic dataset, these new data allowed us to robustly infer Picozoa as a lineage of archaeplastids, branching sister to red algae and rhodelphids. With this expanded genomic dataset, we confirm Picozoa as the first archaeplastid lineage lacking a plastid. We discuss the important implications that these results have on our understanding of the origin of plastids.

Results

Single cell assembled genomes representative of Picozoa diversity

We isolated 43 picozoan cells (40 from the eastern North Pacific off the coast of California, 3 from the Baltic Sea) using FACS and performed whole genome amplification by multiple displacement amplification (MDA). The taxonomic affiliation of the SAGs was determined either by Pico-PCR (Seenivasan et al. 2013) or 18S rRNA gene sequencing (see methods), followed by Illumina sequencing of the MDA products. The sequencing reads were assembled into genomic contigs, ranging in length from 350 kbp to 66 Mbp (Fig 1a, Tab S1). From these contigs, the 18S rRNA gene was found in 37 out of the 43 SAGs, which we used to build a phylogenetic tree with reference sequences from the protist ribosomal reference PR2 database (Fig S1). Based on this tree, we identified 6 groups corresponding to 32 SAGs that possessed nearly identical 18S rRNA gene sequences. These SAGs with identical ribotype were reassembled by pooling all reads in order to obtain longer, more complete co-assemblies (CO-SAGs). The genome size of the CO-SAGs ranged from 32 Mbp to 109 Mbp (Fig 1a, Tab S1), an increase of 5% to 45% over individual SAGs. The genome completeness of the SAGs and CO-SAGs was estimated based on two datasets: (i) a set of 255 eukaryotic marker genes available in BUSCO (Simão et al. 2015), and (ii) a set of 317 conserved marker genes derived from a previous pan-eukaryote phylogenomic dataset (Strassert et al. 2021) that we used here as starting point (Fig 1b). These comparisons showed that while most SAGs were highly incomplete (Fig 1a and b), the CO-SAGs were generally more complete (up to 60%). When taken together, 90% of the BUSCO markers and 88% of the phylogenomic markers were present in at least one assembly, suggesting that while the single-cell genome assemblies are fragmentary, they together represent a much more complete Picozoa meta-assembly.

The final 17 assemblies (11 SAGs and 6 CO-SAGs) were mainly placed within the three proposed groups of Picozoa BP1-3 (Fig 1c), sensu (Cuvelier et al. 2008). One SAG (SAG11) was placed outside of these groups. The deep-branching lineages identified by (Moreira and López-García 2014), as well as other possibly early-diverging lineages were not represented in our data (Fig 1c). Interestingly, one CO-SAG (COSAG03) was closely related (18S rRNA gene 100% identical) to the only described species, *Picomonas judraskeda*, for which no genomic data is available. Using our assemblies and reference sequences from PR2 as queries, we identified by sequence identity 362 OTUs related to Picozoa (≥ 90 %) in the data provided by the *Tara* Oceans project (Vargas et al. 2015). Picozoa were found in all major oceanic regions, but had generally low relative abundance in V9 18S rRNA gene amplicon data (less than 1% of the eukaryotic fractions in most cases, Fig S2). An exception was the Southern Ocean between South America and Antarctica, where the Picozoa-related OTUs in one sample represented up to 30% of the V9 18S rRNA gene amplicons. Thus, Picozoa are widespread in the oceans but generally low in abundance based on available sampling, although they can reach higher relative abundances in at least circumpolar waters.

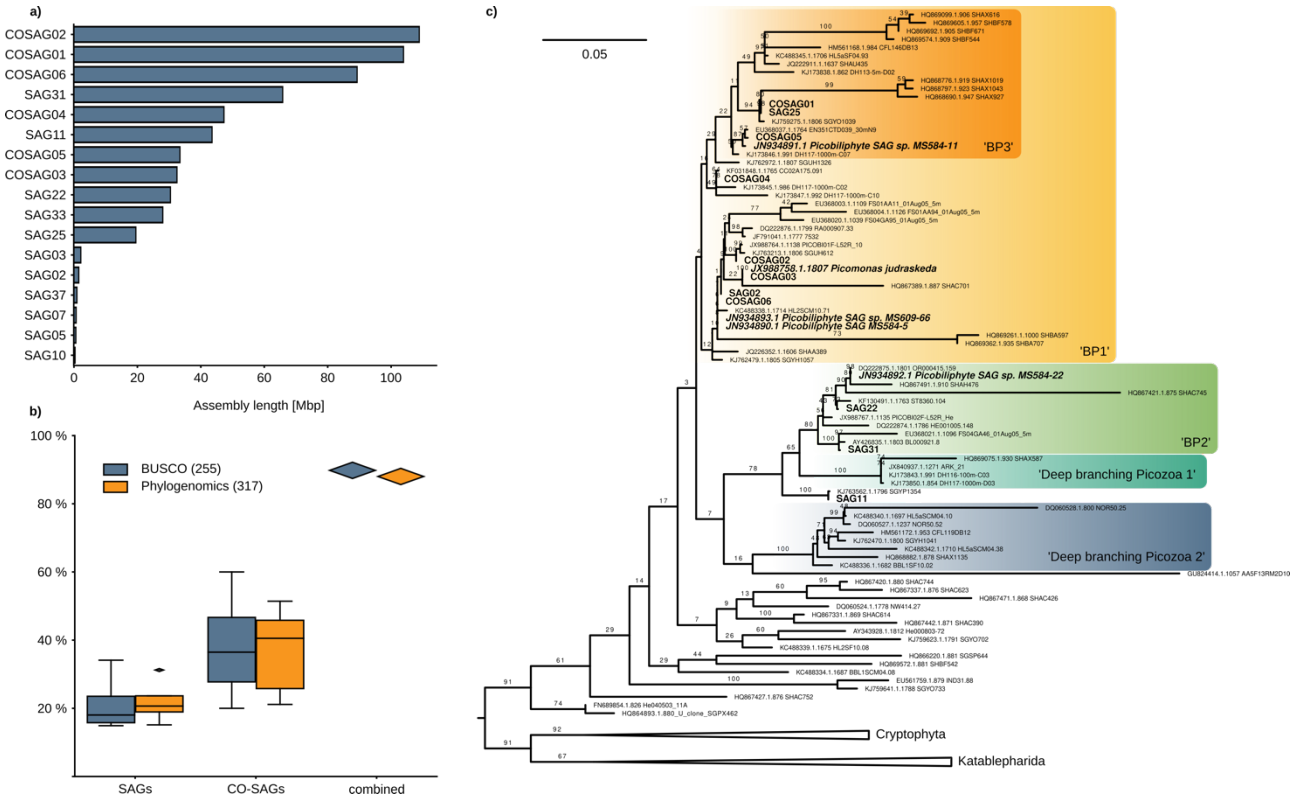

Figure 1. a) Assembly length in Mbp for 17 SAGs and CO-SAGs used for further analysis. **b)** Estimated completeness of the 10 most complete SAGs and CO-SAGs as assessed using presence/absence of the BUSCO dataset of 255 eukaryotic markers and a dataset of 317 Phylogenomic marker genes. These 10 assemblies were used for the phylogenomic inference. **c)** Maximum likelihood tree of the 18S rRNA gene, reconstructed using the model GTR+R4+F while support was estimated with 100 non-parametric bootstrap replicates in IQ-TREE. Picozoa CO-SAGs and SAGs are written in bold, the sequences of *Picomonas judraskeda* and the SAGs from Yoon et al. 2011 in bold italic. Groupings within Picozoa BP1-3 as per (Cuvelier et al. 2008) and deep branching lineages as defined by (Moreira and López-García 2014).

Phylogenomic dataset construction

To infer the evolutionary origin of Picozoa, we expanded on a phylogenomic dataset that contained a broad sampling of eukaryotes and a large number of genes that was recently used to study deep nodes in the eukaryotic tree (Strassert et al. 2021). Homologues from the SAGs and CO-SAGs as well as a number of newly sequenced key eukaryotes were added to each single gene (see Table S2 for a list of taxa). After careful examination of the single genes for contamination and orthology based on individual phylogenies (see material and methods), we retained all six CO-SAGs and four individual SAGs together with the previously available SAG MS584-11 from (Yoon et al. 2011). The rest of the SAGs were excluded due to poor data coverage (less than 5 markers present) and, in one case (SAG33), because it was heavily contaminated with sequences from a cryptophyte (see Data availability for access to the gene trees). In total, our phylogenomic dataset contained 317 protein-coding genes, with orthologues from Picozoa included in 279 genes (88%), and 794 taxa (Fig 1b). This represents an increase in gene coverage from 18% to 88% compared to the previously available genomic data for Picozoa. The most complete assembly was CO-SAG01, from which we identified orthologues for 163 (51%) of the markers.

Picozoa are sister to Rhodophyta and Rhodelphidia

Concatenated protein alignments of the curated 317 genes were used to infer the phylogenetic placement of Picozoa in the eukaryotic Tree of Life. Initially, a maximum likelihood (ML) tree was reconstructed from the complete 794-taxa dataset using the site-homogeneous model LG+F+G and ultrafast bootstrap support with 1,000 replicates (Fig S3). This analysis placed Picozoa together with a clade comprising red algae and rhodelphids with strong support (100% UFBoot2), but the monophyly of Archaeplastida was not recovered due to the internal placement of the cryptists. To further investigate the position of Picozoa, we applied better-fitting site-heterogeneous models to a reduced dataset of 67 taxa, since these models are computationally much more demanding. The process of taxon reduction was driven by the requirement of

maintaining representation from all major groups, while focusing sampling on the part of the tree where Picozoa most likely belong to, i.e. Archaeplastida, TSAR, Haptista and Cryptista. We also merged several closely related lineages into OTUs based on the initial ML tree in order to reduce missing data (Table S6). This 67 taxa dataset was used in ML and Bayesian analyses with the best-fitting site-heterogeneous models LG+C60+F+G+PMSF (with non-parametric bootstrapping) and CAT+GTR+G, respectively. Both ML and Bayesian analyses produced highly similar trees, and received maximal support for the majority of relationships, including deep divergences (Fig 2). Most interestingly, both analyses recovered the monophyly of Archaeplastida (BS=93%; PP=1), with cryptists as sister lineage. Consistent with the initial ML tree (Fig S3), red algae and rhodelphids branched together (BS=95%; PP=1), and Picozoa received strong support (BS=100%; PP=1) for a sister relationship to this group. An approximately unbiased (AU) test rejected the topology where Picozoa branched with rhodelphids (p=0.0362), but the position of Picozoa as the closest sister to red algae could not be rejected (p=0.236).

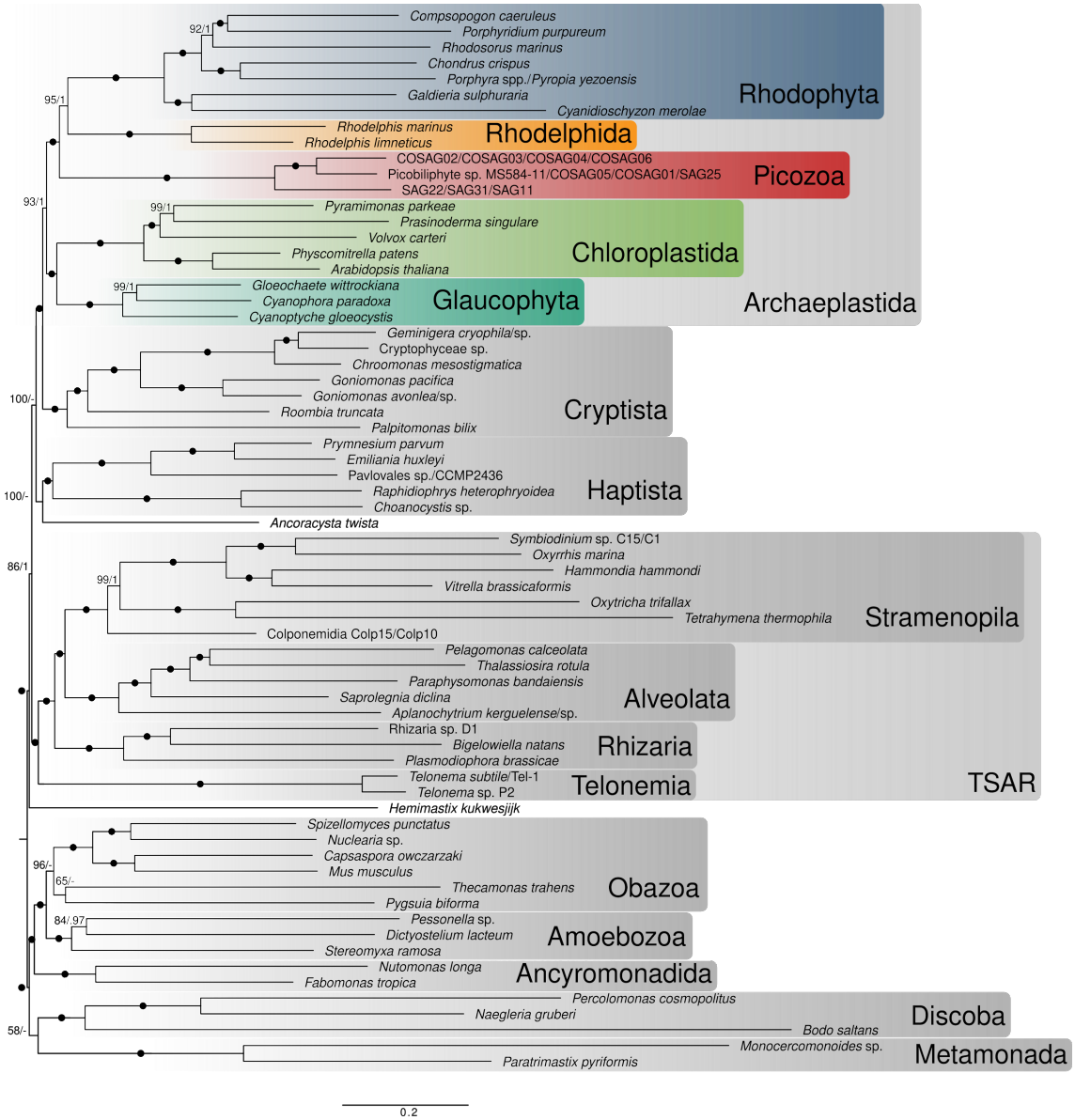

Figure 2. Maximum likelihood tree of eukaryotic species showing the position of Picozoa. The tree is based on the concatenated alignment of 317 marker genes (filtered with BMGE) and was reconstructed using the PMSF approximation of the site-heterogeneous model LG+C60+F+G. Support values correspond to 100 non-parametric bootstrap replicates/ posterior probability values estimated using PhyloBayes CAT-GTR. Black circles denote full support.

Picozoa SAGs show no evidence of a plastid

Since there have been conflicting conclusions about the occurrence of plastids in picozoans, we extensively searched our genomic data for evidence of cryptic plastids. First, we searched the SAG and CO-SAG assemblies for plastidial contigs as evidence of a plastid genome using the tool GetOrganelle (Jin et al.

2020). While there were some contigs that initially showed similarities to reference plastid genomes, these were all rejected as bacterial (non-cyanobacterial) contamination upon closer inspection. In contrast, mitochondrial contigs were readily identified in 26 of 43 SAGs (Table S3). Although mitochondrial contigs remained fragmented in most SAGs, four complete or near-complete mitochondrial genomes were recovered with coding content near-identical to the published mitochondrial genome from picozoa MS5584-11 (Janouškovec et al. 2017) (Fig S4). The ability to assemble complete mitochondrial genomes from the SAGs suggests that the partial nature of the data does not specifically hinder organelle genome recovery if present, at least in the case of mitochondria (Wideman et al. 2020).

Second, we investigated the possibility that the plastid genome was lost while the organelle itself has been retained—as is the case for *Rhodelphis* (Gawryluk et al. 2019). For this, we reconstructed phylogenetic trees for several essential nuclear-encoded biochemical plastid pathways derived by endosymbiotic gene transfer (EGT) that were shown to be at least partially retained even in cryptic plastids (Dorrell et al. 2019; Mathur et al. 2019; Gawryluk et al. 2019). These included genes involved in the biosynthesis of isoprenoids (ispD,E,F,G,H, dxr, dxs), fatty acids (fabD,F,G,H,I,Z, ACC), heme (hemB,D,E,F,H,Y, ALAS), and iron-sulfur clusters (sufB,C,D,E,S, NifU, iscA; see also Table S4). In all cases, the picozoan homologues grouped either with bacteria—but not cyanobacteria, suggesting contamination—or the mitochondrial/nuclear copies of host origin. Furthermore, none of the picozoan homologues contained N-terminal plastid transit peptides, as predicted by TargetP (Almagro Armenteros et al. 2019). We also searched for picozoan homologues of all additional proteins (n=62) that were predicted to be targeted to the cryptic plastid in rhodelphids (Gawryluk et al. 2019). This search resulted in one protein (Arogenate dehydrogenase) with picozoan homologues that were closely related to red algae and belonged to a larger clade with host-derived plastid targeted plant sequences, but neither the picozoan nor the red algal sequences displayed predicted transit peptides. Finally, to eliminate the possibility of missing sequences because of errors during the assembly and gene prediction, we additionally searched the raw read sequences for the same plastid-targeted genes, as well as several genes from the mitochondrial transport machinery using the tool PhyloMagnet (Schön, Eme, and Ettema 2020). This revealed no obvious candidates for plastid-targeted or plastid transport machinery genes in the raw data, but readily identified mitochondrial homologues or several additional bacterial contaminations. For example, we identified potential homologues of the mitochondrial import machinery from the TIM17/TIM22 family, which further strengthened our inference that the single-cell data is in principle adequate to identify organellar components, when they are present.

The lack of cryptic plastids in diverse modern-day picozoans does not preclude photosynthetic ancestry if the plastid was lost early in the evolution of the group. To assess this possibility, we searched more widely for evidence of a cyanobacterial footprint on the nuclear genome that would rise above a background of horizontal gene transfers for proteins functioning in cellular compartments other than the plastids. The presence of a significant number of such proteins may be evidence for a plastid-bearing ancestor. We clustered proteins from 419 genomes, including all major eukaryotic groups as well as a selection of bacteria into orthologous groups (OGs) (Table S5). We built phylogenies for OGs that contained at least cyanobacterial and algal sequences, as well as a sequence from one of 33 focal taxa, including Picozoa, a range of photosynthetic taxa, but also non-photosynthetic plastid-containing, and plastid-lacking taxa to be used as controls. Putative gene transfers from cyanobacteria (EGT) were identified as a group of plastid-bearing eukaryotes that included sequences from the focal taxa and branched sister to a clade of cyanobacteria. We allowed up to 10% of sequences from groups with no plastid ancestry. This approach identified 16 putative EGTs for Picozoa where at least 2 different SAGs/CO-SAGs grouped together, compared to between 89–313 EGTs for photosynthetic species, and up to 59 EGTs for species with non-photosynthetic plastids (Fig 3a). At the other end of the spectrum for species with non-photosynthetic plastids, we observed that the number of inferred cyanobacterial genes for e.g. rhodelphids (14) or *Paraphysomonas* (12) was comparable to Picozoa (16) or other, plastid-lacking taxa such as *Telonema* (15) or *Goniomonas* (18). In order to differentiate these putative endosymbiotic transfers from a background of bacterial transfers (or bacterial contamination), we next attempted to normalise the EGT signal by estimating an extended bacterial signal (indicative of putative HGT: horizontal gene transfers) using the same tree sorting procedure (Fig S5). When comparing the number of inferred EGT with that of inferred HGT, we found a marked difference between plastid-containing (including non-photosynthetic) and plastid-lacking lineages. While all plastid-containing taxa—with the notable exception of *Rhodelphis*—showed a ratio of EGT to HGT above 1, all species without plastid ancestry and *Hematodinium*, one of the few taxa with reported plastid loss, as well as *Rhodelphis* and Picozoa showed a much higher number of inferred HGT than EGT.

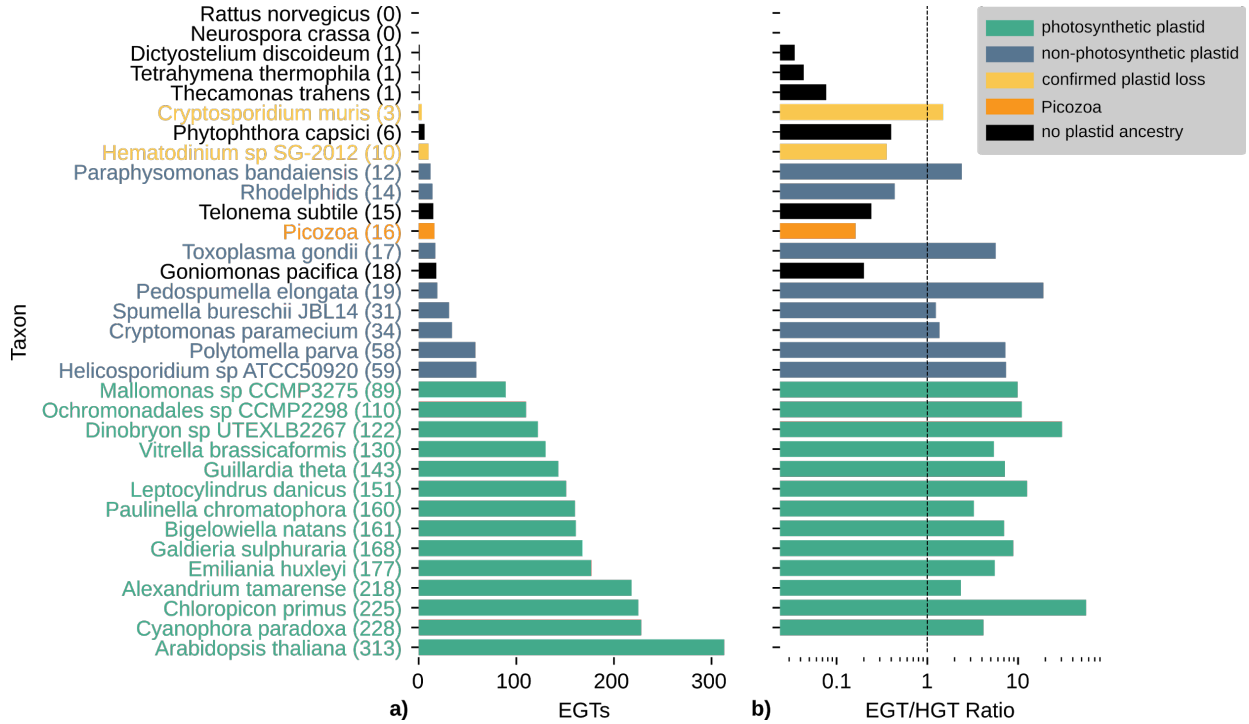

Figure 3. a) Number of inferred endosymbiotic gene transfers (EGT) across a selection of 33 species that represent groups with a photosynthetic plastid (green), a non-photosynthetic plastid (blue), confirmed plastid loss (yellow) and no known plastid ancestry (black). These species serve as a comparison to Picozoa (orange). **b)** The number of EGTs from a) is related to the number of inferred HGT across the same 33 selected species. A number below 1 indicates more HGT than EGT, while numbers above 1 indicate more EGT than HGT. No ratio could be calculated for *Arabidopsis* since there were no detectable HGT events.

Discussion

The 17 SAGs and CO-SAGs of Picozoa obtained in this study provide the first robust data for phylogenomic analyses of this important phylum of eukaryotes. With this data, we are able to firmly place Picozoa within the supergroup Archaeplastida, more specifically as a sister lineage to red algae and rhodelphids. Archaeplastids contain all known lineages with primary plastids (with the exception of *Paulinella*), which are widely viewed to be derived from a single primary endosymbiosis with a cyanobacterium. This notion of a common origin of primary plastids is supported by cellular and genomic data (see (Bhattacharya et al. 2007; Gould, Waller, and McFadden 2008) and references therein for review), as well as plastid phylogenetics (Shih et al. 2013; Ponce-Toledo et al. 2017). The phylogenetic support for Archaeplastida based on host (nuclear) data has been less certain (Yabuki et al. 2014; Burki et al. 2016; Strassert et al. 2019), but our analysis is consistent with recent reports that have also recovered a monophyletic origin—here including Picozoa—when using gene and taxon-rich phylogenomic datasets (Lax et al. 2018; Irisarri, Strassert, and Burki 2020; Strassert et al. 2021). This position has important implications for our understanding of plastid origins because, in contrast to all other archaeplastids known to date, our results strongly indicate that Picozoa lack plastids and plastid-associated EGTs. The lack of plastid in Picozoa was also inferred based on smaller initial SAG data (Yoon et al. 2011) as well as ultrastructural observation of *P. judraskeda* (Seenivasan et al. 2013). Two main possible hypotheses exist to explain the lack of plastids in Picozoa: that this group was never photosynthetic, or complete plastid loss occurred early in their evolution.

To suggest that Picozoa was never photosynthetic requires that the current distribution of primary plastids is due to multiple independent endosymbioses, specifically that red algae arose from a separate primary endosymbiosis from that leading to green algae and glaucophytes. This scenario would have involved the endosymbioses of closely related cyanobacterial lineages in closely related hosts to explain the many similarities between primary plastids (Gould, Waller, and McFadden 2008). Although this may sound unlikely, there is accumulating evidence that similar plastids were derived independently from similar endosymbionts in closely related hosts in dinoflagellates with tertiary plastids (Hehenberger, Gast, and Keeling 2019; Sarai et al. 2020; Yamada et al. 2020), and has been argued before for primary plastids (Stiller, Reel, and Johnson 2003; Howe et al. 2008; Stiller 2014). However, the current bulk of cell and

molecular evidence suggests that multiple independent origins of primary plastids is unlikely, including several features of plastid biology that are not present in cyanobacteria (e.g., protein targeting systems, light-harvesting complex proteins, or plastid genome architecture) (Gould, Waller, and McFadden 2008). A related explanation could involve a secondary endosymbiosis where the plastid in red algae, for example, was secondarily acquired from a green alga (Kim and Maruyama 2014). The latter scenario would be effectively refuted by the identification of host-derived plastid components shared between all archaeplastid lineages.

The second hypothesis implies that a common ancestor of Picozoa entirely lost its primary plastid. The possibility of plastid loss in a free-living lineage like Picozoa would be unprecedented because to date, the only known unambiguous cases of total plastid loss all come from parasitic lineages (all in myzozoan alveolates: in *Cryptosporidium* (Zhu, Marchewka, and Keithly 2000), certain gregarines (Mathur et al. 2019; Janouškovec et al. 2019), and the dinoflagellate *Hematodinium* (Gornik et al. 2015)). To evaluate this possibility, we searched our data for a cyanobacterial footprint in the nuclear genome that would result from an ancestral endosymbiosis. The transfer of genes from endosymbiont to host nucleus via EGT, and the targeting of the product of some or all of these genes back to the plastids, are recognised as a hallmark of organelle integration (Timmis et al. 2004; Archibald 2015). EGT has occurred in all algae, although its impact on nuclear genomes can vary and the inference of EGT versus other horizontally acquired genes (HGT) can be difficult to decipher for ancient endosymbioses (Burki et al. 2012; Deschamps and Moreira 2012; Qiu, Yoon, and Bhattacharya 2013; Morozov and Galachyants 2019; Sibbald and Archibald 2020). Our analysis of the normalised cyanobacterial signal in Picozoa, which we used as a proxy for quantifying EGT, provides no clear evidence for the existence of a plastid-bearing ancestor. However, it should be noted that evaluating the possibility of plastid loss in groups where a photosynthetic ancestry is not confirmed—such as Picozoa—is complicated because there is no baseline for the surviving footprint of endosymbiosis following plastid loss. Notably, we found no significant difference in the number of inferred EGTs in Picozoa compared to lineages with demonstrated plastid loss (e.g. *Hematodinium* with 10 inferred EGT), lineages with non-photosynthetic plastids (e.g. *Rhodelphis*: 14 inferred EGT), or with no photosynthetic ancestry (e.g. *Telonema*: 15 inferred EGT).

The lack of a genomic baseline to assess plastid loss in Picozoa is further complicated by limitations of our data and methods. The partial nature of eukaryotic SAGs makes it possible that EGTs are absent from our data, even with >90% of inferred genomic completeness. Additionally, the possibility exists that the number of EGT might have always been low during the evolution of the group, even if a plastid was once present. Recent endosymbioses where EGT can be pinpointed with precision showed a relatively low frequency. For example, they represent at most a few percent of the chromatophore proteome in *Paulinella* (Singer et al. 2017), or as few as 9 genes in tertiary endosymbiosis in dinoflagellates (Burki et al. 2014). Thus, it is possible that the much higher number of EGT inferred in red algae (e.g. 168 in *Galdieria*) occurred after the divergence of Picozoa, and that Picozoa quickly lost its plastid before more EGT occurred. An observation that supports this hypothesis is the low number of putative EGTs found in *Rhodelphis* (14), suggesting that the bulk of endosymbiotic transfers in red algae happened after their divergence from rhodelphids.

Conclusion

In this study, we used single-cell genomics to demonstrate that Picozoa are a plastid lacking major lineage of archaeplastids. This is the first example of an archaeplastid lineage without plastids, which can be interpreted as either plastid loss, or evidence of independent endosymbiosis in the ancestor of red algae and rhodelphids. Under the most widely accepted scenario of a single plastid origin in Archaeplastida, Picozoa would represent the first known case of plastid loss in this group, but also more generally in any free-living species. However, the fact that this lineage has never been successfully maintained in culture, with just one study achieving transient culture (Seenivasan et al. 2013), might indicate a lifestyle involving close association with other organisms (such as symbiosis) and further underscores the enigma of picozoan biology, which if elucidated might shed further light on their evolution. In order to discriminate plastid loss from multiple plastid gains in the early evolution of the Archaeplastida, and more generally during the evolution of secondary or tertiary plastids, a better understanding of the early steps of plastid integration is required. In the recently evolved primary plastid-like chromatophores of *Paulinella*, the transfer of endosymbiotic genes at the onset of the integration was shown to be minimal (Nowack and Weber 2018). Similar examples of integrated plastid endosymbionts but with apparently very few EGTs are known in dinoflagellates (Burki et al. 2014; Hehenberger et al. 2016). Therefore, new important clues to decipher the origin of plastids will

likely come from a better understanding of the role of the host in driving these endosymbioses, and crucially the establishment of a more complete framework for Archaeplastida evolution with the search and characterization of novel diversity of lineages without plastids.

Material and Methods

Cell Isolation, Identification, and genome amplification

Baltic Sea. Surface (depth: up to 2 m) marine water was collected from the Linnaeus microbial Observatory (LMO) in the Baltic Sea located at 56°N 55.85' and 17°E 03.64' on two occasions: 2 May 2018 (6.1°C and 6.8 ppt salinity) and 3 April 2018 (2.4°C and 6.7 ppt salinity). The samples were transported to the laboratory and filter-fractionized (see Table 1 for detailed sample information). The size fractions larger than 2 µm were discarded whereas the fraction collected on 0.2 µm filters was resuspended in 2 mL of the filtrate. The obtained samples were used for fluorescence-activated cell sorting (FACS). Aliquots of 4 µL of 1 mM Mitotracker Green FM (ThermoFisher) stock solution were added to the samples and were kept in the dark at 15°C for 15–20 minutes. Then the cells were sorted into empty 96-well plates using MoFlo Astrios EQ cell sorter (Beckman Coulter). Gates were set mainly based on Mitotracker intensity and the dye was detected by a 488 nm and 640 nm laser for excitation, 100 µm nozzle, sheath pressure of 25 psi and 0.1 µm sterile filtered 1 x PBS as sheath fluid. The region with the highest green fluorescence and forward scatter contained the target group and was thereafter used alongside with exclusion of red autofluorescence.

The SAGs were generated in each well with REPLI-g® Single Cell kit (Qiagen) following the manufacturer's recommendations but scaled down to 5 µl reactions. Since the cells were sorted in dry plates, 400 nl of 1xPBS was added prior to 300 nl of lysis buffer D2 for 10 min at 65°C and 10 min on ice, followed by 300 nl stop solution. The PBS, reagent D2, stop solution, water, and reagent tubes were UV-treated at 2 Joules before use. A final concentration of 0.5 µM SYTO 13 (Invitrogen) was added to the MDA mastermix. The reaction was run at 30°C for 6 h followed by inactivation at 65°C for 5 min and was monitored by detection of SYTO13 fluorescence every 15 minutes using a FLUOstar® Omega plate reader (BMG Labtech, Germany). The single amplified genome (SAG) DNA was stored at -20°C until further PCR screening. The obtained products were PCR-screened using Pico-PCR approach, as described in (Seenivasan et al. 2013) and the wells showing signal for Picozoa were selected for sequencing.

Eastern North Pacific. Seawater was collected and sorted using a BD InFlux Fluorescently Activated Cell Sorter (FACS) on three independent cruises in the eastern North Pacific. The instrument was equipped with a 100 mW 488 nm laser and a 100 mW 355 nm laser and run using sterile nuclease-free 1× PBS as sheath fluid. The stations where sorting occurred were located at 36.748°N, 122.013°W (Station M1; 20 m, 2 April 2014 and 10 m, 5 May 2014); 36.695°N, 122.357°W (Station M2, 10 m, 5 May 2014); and 36.126°N, 123.49°W (Station 67-70, 20 m 15 October 2013). Water was collected using Niskin bottles mounted on a CTD rosette. Prior to sorting samples were concentrated by gravity over a 0.8 µm Supor filter. Two different stains were used: LysoSensor (2 April 2014, M1) and LysoTracker (5 May 2014, M1; 15 October 2013, 67-70), or both together (5 May 2014, M2). Selection of eukaryotic cells stained with LysoTracker Green DND-26 (Life Technologies; final concentration, 25 nM) was based on scatter parameters, positive green fluorescence (520/35 nm bandpass), as compared to unstained samples, and exclusion of known phytoplankton populations, as discriminated by their forward angle light scatter and red (chlorophyll-derived) autofluorescence (i.e., 692/40 nm bandpass) under 488 nm excitation, similar to methods in (Needham et al. 2019). Likewise, selection of cells stained with LysoSensor Blue DND-167 (Life Technologies; final concentration, 1 µM), a ratiometric probe sensitive to intracellular *pH* levels, e.g. in lysosomes, was based on scatter parameters, positive blue fluorescence (435/40 nm bandpass), as compared to unstained samples, and exclusion of known phytoplankton populations, as discriminated by their forward angle light scatter and red (chlorophyll-derived) autofluorescence (i.e., 692/40 nm bandpass filter) under 355 nm excitation. For sorts using both stains all of the above criteria, and excitation with both lasers (with emissions collected through different pinholes and filter sets), were applied to select cells. Before each sort was initiated, the respective plate was illuminated with UV irradiation for 2 min. Cells were sorted into 96- or 384-well plates using the Single-Cell sorting mode from the BD FACS Software v1.0.0.650. A subset of wells was left empty or received 20 cells for negative and positive controls, respectively. After sorting, the plates were covered with sterile, nuclease free foil and frozen at -80 °C immediately after completion.

Whole genome amplification of individual sorted cells followed methods outlined in (Needham et al. 2019). For initial screening, 18S rRNA gene amplicons were amplified from each well using the Illumina

adapted TAReuk454FWD1 and TAReukREV3 primers targeting the V4 hypervariable region. PCR reactions contained 10 ng of template DNA and 1X 5PRIME HotMasterMix (Quanta Biosciences) as well as 0.4 mg ml^{-1} BSA (NEB) and 0.4 µM of each primer. PCR reactions entailed: 94 °C for 3 min; and 30 cycles at 94 °C for 45 sec, 50 °C for 60 sec and 72 °C for 90 sec; with a final extension at 72 °C for 10 min. Triplicate reactions per cell were pooled prior to Paired-end (PE) library sequencing (2 × 300 bp) and the resulting 18S V4 rRNA gene amplicons were trimmed at Phred quality (Q) of 25 using a 10 bp running window using Sickle 1.33 (Joshi and Fass 2011). Paired-end reads were merged using USEARCH v.9.0.2132 when reads had a ≥40 bp overlap with max 5% mismatch. Merged reads were filtered to remove reads with maximum error rate >0.001 or <200 bp length. Sequences with exact match to both primers were retained, primer sequences were trimmed using Cutadapt v.1.13 (Martin 2011), and the remaining sequences were *de novo* clustered at 99% sequence similarity by UCLUST forming operational taxonomic units (OTUs). Each of the cells further sequenced had a single abundant OTU that was taxonomically identified using BLASTn in GenBank's nr database.

Sequencing

Sequencing libraries were prepared from 100 ng DNA using the TruSeq Nano DNA sample preparation kit (cat# 20015964/5, Illumina Inc.) targeting an insert size of 350bp. For six samples, less than 100ng was used (between 87 ng-97 ng). The library preparations were performed by SNP&SEQ Technology Platform at Uppsala University according to the manufacturers' instructions. All samples were then multiplexed on one lane of an Illumina HiSeqX instrument with 150 cycles paired-end sequencing using the v2.5 sequencing chemistry, producing between 10,000 and 30,000,000 read pairs.

Genome Assembly and 18S rRNA gene analysis

The 43 Illumina datasets were trimmed using *Trim Galore* v0.6.1 (Krueger 2021) with default parameter and assembled into genomic contigs with SPAdes v3.13.0 (Bankevich et al. 2012) in single-cell mode (--sc --careful -k 21,33,55,99). Open reading frames (ORFs) were identified and translated using Prodigal v2.6.3 in 'anonymous' mode (Hyatt et al. 2010) and rRNA genes were predicted using barrnap v0.9 (Seemann [2013] 2021) for eukaryotes. All 18S rRNA gene sequences were, together with available reference sequences from the protist ribosomal reference database (PR2 (Guillou et al. 2013)), aligned with MAFFT E-INS-i v7.429 (Katoh and Standley 2013) and trimmed with trimal (gap threshold 0.01%, (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009)). After performing a modeltest using ModelFinder (best model: GTR+R6+F), a phylogenetic tree was reconstructed in IQ-TREE v2.1.1 (Kalyaanamoorthy et al. 2017; Minh et al. 2020) with 1000 ultrafast bootstrap replicates (see Fig S6 for a tree with extended taxon sampling). Additionally, we estimated the average nucleotide identity (ANI) for all pairs of SAGs using fastANI v1.2 (Jain et al. 2018) (Fig S7). Based on the 18S rRNA gene tree and the ANI value, groups of closely related SAGs with almost identical 18S rRNA gene sequences (sequence similarity above 99%) were identified for co-assembly. Co-assemblies were generated in the same way as described above for single assemblies, pooling sequencing libraries from closely related single cells. ORFs and rRNA genes were similarly extracted from the co-assemblies. The completeness of the SAGs and CO-SAGs was then assessed using BUSCO v4.1.3 (Simão et al. 2015) with 255 markers for eukaryotes as well as using the 320 marker phylogenomic dataset as described below. General genome characteristics were computed with QUAST v5.0.2 (Gurevich et al. 2013). Alignments were reconstructed for the 18S rRNA genes from the co-assemblies and those SAGs not included in any CO-SAG together with PR2 references for cryptists and katablepharids (the closest groups to Picozoa in 18S rRNA gene phylogenies) in the same way as described above. The tree was reconstructed using GTR+R4+F after model selection and support was assessed with 100 non-parametric bootstraps. The six CO-SAGs and the 11 individual SAGs were used in all subsequent analyses.

Phylogenomics

Existing untrimmed alignments for 320 genes and 763 taxa from (Strassert et al. 2021) were used to create HMM profiles in HMMER v3.2.1 (Eddy 2011), which were then used to identify homologous sequences in the protein sequences predicted from the Picozoa assemblies (or co-assemblies) as well as in 20 additional, recently sequenced eukaryotic genomes and transcriptomes (Table S2). Each single gene dataset was filtered using PREQUAL v1.02 (Whelan, Irisarri, and Burki 2018) to remove non-homologous residues prior to alignment, aligned using MAFFT E-INS-i, and filtered with Divvier -partial v1.0 (Ali et al. 2019). Alignments were then used to reconstruct gene trees with IQ-TREE (-mset LG, LG4X; 1000 ultrafast bootstraps with the BNNI optimization). All trees were manually scrutinized to identify contamination and paralogs. These steps were repeated at least two times, until no further contaminations or paralogs could be

detected. We excluded three genes that showed ambiguous groupings of Picozoa or rhodelphids in different parts of the trees. From this full dataset of 317 genes and 794 taxa, we created a concatenated supermatrix alignment using the cleaned alignments described above. This supermatrix was used to reconstruct a tree in IQ-TREE with the model LG+G+F and ultrafast bootstraps (1000 UFBoots) estimation with the BNNI improvement.

We then prepared a reduced dataset with a more focused taxon sampling of 67 taxa, covering all major eukaryotic lineages but focussing on the groups for which an affiliation to Picozoa had been reported previously. For this dataset, closely related species were merged into OTUs in some cases in order to decrease the amount of missing data per taxon (Table S6). The 317 single gene datasets were re-aligned using MAFFT E-INS-i, filtered using both Divvier -partial and BMGE (-g 0.2 -b 10 -m BLOSUM75) and concatenated into two supermatrices. Model selection of mixture models was performed using ModelFinder (Kalyaanamoorthy et al. 2017) for both datasets, and in both cases LG+C60+G+F was selected as the best-fitting model. Trees for both datasets were reconstructed using the Posterior Mean Site Frequency (PMSF, (Wang et al. 2017)) approximation of this mixture model in IQ-TREE and support was assessed with 100 non-parametric bootstraps (see Fig S8 for the Divvier derived tree)

In addition, we reconstructed a phylogenetic tree using the supermatrix alignment based on BMGE trimming in PhyloBayes MPI v1.8 (Lartillot et al. 2013) using the CAT+GTR+G model. We ran three independent chains for 3600 cycles, with the initial 1500 cycles being removed as burnin from each chain. We then generated a consensus tree using the bpcomp program of PhyloBayes. Partial convergence was achieved between chains 1 and 2 with a maxdiff value of 0.26 (Fig S9). The third chain differed only in the position of haptists and *Ancoracysta twista* (Janouškovec et al. 2017), but not in the relationships within Archaeplastida and the position of Picozoa (Fig S10).

Mitochondrial contig identification and annotation

Using the published picozoan mitochondrial genome (Picozoa sp. MS584-11: MG202007.1, (Janouškovec et al. 2017)), BLAST searches were performed on a dedicated sequenceServer to identify mitochondrial contigs in the 43 picozoan SAGs (Altschul et al. 1997; Priyam et al. 2019). Putative mitochondrial contigs were annotated using the MFannot server (Beck and Lang 2020). All contigs with predicted mitochondrial genes or whose top hits in the NCBI nr database was the published picozoan mitochondrial genome (MG202007.1) were considered to be *bona fide* mitochondrial contigs and retained (Supplementary materials). Manual annotation was conducted as needed.

Plastid Genes & EGT

GetOrganelle v1.7.1 (Jin et al. 2020) was used to identify organellar genomes. We searched the assemblies for putative plastid contigs with the subcommand 'get_organelle_from_assembly.py -F embplant_pt,other_pt', while we attempted to assemble such a genome directly using the command 'get_organelle_from_reads.py -R 30 -k 21,45,65,85,105 -F embplant_pt,other_pt'. We additionally searched the predicted proteins against available plastid protein sequences from ncbi using diamond v2.0.6 (Buchfink, Xie, and Huson 2015) in blastp mode (--more-sensitive). Contigs that were identified as putatively coming from a plastid genome were then checked manually by doing BLAST searches against NT, and contigs that showed similarity only to bacterial genomes or to the picozoa mitochondrial assembly MG202007.1 were rejected.

To search for known plastid pathways, we prepared Hidden Markov model (HMM) profiles for 32 gene alignments that were shown to be retained in lineages with non-photosynthetic plastids and included a wide diversity of plastid-bearing eukaryotes following a similar approach as in (Mathur et al. 2019). Using these profiles, we identified homologues in the Picozoa SAGs, and aligned them together with the initial sequences used to create the profiles using MAFFT E-INS-i. We trimmed the alignments using trimAl v1.4.rev15 '-gt 0.05' (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009) and reconstructed phylogenetic trees using IQ-TREE (-m LG4X; 1000 ultrafast bootstraps with the BNNI optimization) from these alignments. We then manually inspected the trees to assess whether picozoan sequences grouped with known plastid-bearing lineages. We additionally used the sequences from these core plastid genes to search the raw sequencing reads for any signs of homologues that could have been missed in the assemblies. We used the tool PhyloMagnet (Schön, Eme, and Ettema 2020) to recruit reads and perform gene-centric assembly of these genes (Huson et al. 2017). The assembled genes were then compared to the NR database using diamond in blastp mode (--more-sensitive --top 10).

To identify putative EGT, we prepared orthologous clusters for 419 species (128 bacteria and 291 eukaryotes) with a focus on plastid-bearing eukaryotes and cyanobacteria, but also including other eukaryotes and bacteria, using OrthoFinder v2.4.0 (Emms and Kelly 2019). For Picozoa and a selection of 32 photosynthetic or heterotrophic lineages (Table S7), we inferred trees for 2626 clusters that contained the species under consideration, at least one cyanobacterial sequence, and at least one archaeplastid sequence of red algae, green algae or plants. Alignments for these clusters were generated with MAFFT E-INS-i, filtered using trimAl '-gt 0.01' and phylogenetic trees were reconstructed using IQ-TREE (-m LG4X; 1000 ultrafast bootstraps with the BNNI optimization). We then identified trees where the the target species grouped with other plastid-bearing lineages (allowing up to 10% non-plastid sequences) and sister at least two cyanobacterial sequences. For Picozoa, we added the condition that sequences from at least two SAG/COSAG assemblies must be monophyletic. For species with no known plastid ancestry such as *Rattus* or *Phytophtora*, these putative EGTs can be interpreted as false positives due to contamination, poor tree resolution or other mechanisms, since we expect no EGTs from cyanobacteria to be present at all in these species. This rough estimate of the expected false positive rate for this approach can give us a baseline of false positives that can be expected for picozoa as well.

To put the number of putative EGTs into relation to the overall amount of gene transfers, we applied a very similar approach to the one described above for detecting putative HGT events. We prepared additional trees (in the same way as described for the detection of EGTs) for clusters that contained the taxon of interest and non-cyanobacterial bacteria and identified clades of the taxon under consideration (including a larger taxonomic group, e.g. Streptophyta for *Arabidopsis* or Metazoa for *Rattus*) that branched sister to a bacterial clade.

Distribution of Picozoa in Tara *Oceans*

We screened available OTUs that were obtained from V9 18S rRNA gene eukaryotic amplicon data generated by *Tara* Oceans (Vargas et al. 2015; De Vargas et al. 2017) for sequences related to Picozoa. Using the V9 region of the 18S rRNA gene sequences from the 17 Picozoa assemblies as well as from the picozoan PR2 references used to reconstruct the 18S rRNA gene tree described above, we applied VSEARCH v2.15.1 (Rognes et al. 2016) (--usearch_global -iddef 1 --id 0.90) to find all OTUs with at least 90 % similar V9 regions to any of these reference picozoan sequences. Using the relative abundance information available for each *Tara* Oceans sampling location, we then computed the sum for all identified Picozoa OTUs per station and plotted the relative abundance on a world map.

Acknowledgments

This work was supported by a grant from Science for Life Laboratory available to FB and a scholarship from Carl Tryggers Stiftelse to VZ (PI: FB). TJGE thanks the European Research Council (ERC consolidator grant 817834); the Dutch Research Council (NWO-VICI grant VI.C.192.016); Moore–Simons Project on the Origin of the Eukaryotic Cell (Simons Foundation 735925LPI, https://doi.org/10.46714/735925LPI); and the Marie Skłodowska-Curie ITN project SINGEK (H2020-MSCA-ITN-2015-675752) which provided funding for MES. The Pacific Ocean work was supported by GBMF 3788 to AZW. Sampling at the LMO station in the Baltic Sea was carried out by support from the Swedish Research Council VR and the marine strategic research program EcoChange to JP. Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala, part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation. Cell sorting and whole genome amplification was performed at the Microbial Single Cell Genomics Facility (MSCG) at SciLifeLab. Computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Projects SNIC 2019/3-305, SNIC 2020/15-58, SNIC 2021/5-50, Uppstore2018069.

Code & data availability

All custom scripts used in this study are available at https://github.com/maxemil/picozoa-scripts under a GPLv3 license. All data used for the analyses as well as results files such as contigs and single gene trees are available at figshare (10.6084/m9.figshare.c.5388176). A sequenceServer BLAST server was set up for the

SAG assemblies: http://evocellbio.com/SAGdb/burki/. Raw sequencing reads were deposited in the Sequence Read Archive (SRA) at NCBI under accession ### and will be available upon acceptance.

References

Ali, Raja Hashim, Marcin Bogusz, Simon Whelan, and Koichiro Tamura. 2019. 'Identifying Clusters of High Confidence Homologies in Multiple Sequence Alignments'. *Molecular Biology and Evolution* 36 (10): 2340–51. https://doi.org/10.1093/molbev/msz142.

Almagro Armenteros, Jose Juan, Marco Salvatore, Olof Emanuelsson, Ole Winther, Gunnar von Heijne, Arne Elofsson, and Henrik Nielsen. 2019. 'Detecting Sequence Signals in Targeting Peptides Using Deep Learning'. *Life Science Alliance* 2 (5). https://doi.org/10.26508/lsa.201900429.

Altschul, S F, T L Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, and D J Lipman. 1997. 'Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs.' *Nucleic Acids Research* 25 (17): 3389–3402.

Archibald, John M. 2015. 'Genomic Perspectives on the Birth and Spread of Plastids'. *Proceedings of the National Academy of Sciences* 112 (33): 10147–53. https://doi.org/10.1073/pnas.1421374112.

Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. 'SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing'. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 19 (5): 455–77. https://doi.org/10.1089/cmb.2012.0021.

Beck, N, and FB Lang. 2020. *MFannot, Organelle Genome Annotation Websever.* http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl.

Bhattacharya, Debashish, John M. Archibald, Andreas P. M. Weber, and Adrian Reyes-Prieto. 2007. 'How Do Endosymbionts Become Organelles? Understanding Early Events in Plastid Evolution'. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 29 (12): 1239–46. https://doi.org/10.1002/bies.20671.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson. 2015. 'Fast and Sensitive Protein Alignment Using DIAMOND.' *Nature Methods* 12 (1): 59–60.

Burki, Fabien, Pavel Flegontov, Miroslav Oborník, Jaromír Cihlář, Arnab Pain, Julius Lukeš, and Patrick J. Keeling. 2012. 'Re-Evaluating the Green versus Red Signal in Eukaryotes with Secondary Plastid of Red Algal Origin'. *Genome Biology and Evolution* 4 (6): 626–35. https://doi.org/10.1093/gbe/evs049.

Burki, Fabien, Behzad Imanian, Elisabeth Hehenberger, Yoshihisa Hirakawa, Shinichiro Maruyama, and Patrick J. Keeling. 2014. 'Endosymbiotic Gene Transfer in Tertiary Plastid-Containing Dinoflagellates'. *Eukaryotic Cell* 13 (2): 246–55. https://doi.org/10.1128/EC.00299-13.

Burki, Fabien, Maia Kaplan, Denis V. Tikhonenkov, Vasily Zlatogursky, Bui Quang Minh, Liudmila V. Radaykina, Alexey Smirnov, Alexander P. Mylnikov, and Patrick J. Keeling. 2016. 'Untangling the Early Diversification of Eukaryotes: A Phylogenomic Study of the Evolutionary Origins of Centrohelida, Haptophyta and Cryptista'. *Proceedings of the Royal Society B: Biological Sciences* 283 (1823): 20152802. https://doi.org/10.1098/rspb.2015.2802.

Burki, Fabien, Andrew J. Roger, Matthew W. Brown, and Alastair G.B. Simpson. 2020. 'The New Tree of Eukaryotes'. *Trends in Ecology and Evolution* 35 (1): 43–55. https://doi.org/10.1016/j.tree.2019.08.008.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. 'TrimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses'. *Bioinformatics* 25 (15): 1972–73. https://doi.org/10.1093/bioinformatics/btp348.

Cuvelier, Marie L, Alejandra Ortiz, Eunsoo Kim, Heike Moehlig, David E Richardson, John F Heidelberg, John M Archibald, and Alexandra Z Worden. 2008. 'Widespread Distribution of a Unique Marine Protistan Lineage' 10: 1621–34. https://doi.org/10.1111/j.1462-2920.2008.01580.x.

De Vargas, Colomban, Stephane Audic, Coordinators Tara Oceans Consortium, and Participants Tara Oceans Expedition. 2017. 'Total V9 RDNA Information Organized at the OTU Level for the Tara Oceans Expedition (2009-2012)'. PANGAEA. https://doi.org/10.1594/PANGAEA.873275.

Deschamps, Philippe, and David Moreira. 2012. 'Reevaluating the Green Contribution to Diatom Genomes'. *Genome Biology and Evolution* 4 (7): 683–88. https://doi.org/10.1093/gbe/evs053.

Dorrell, Richard G, Tomonori Azuma, Mami Nomura, Guillemette Audren, De Kerdrel, and Lucas Paoli. 2019. 'Principles of Plastid Reductive Evolution Illuminated by Nonphotosynthetic Chrysophytes'

116 (14). https://doi.org/10.1073/pnas.1819976116.
Eddy, Sean R. 2011. 'Accelerated Profile HMM Searches'. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1002195.
Emms, David M., and Steven Kelly. 2019. 'OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics'. *Genome Biology* 20 (1): 238. https://doi.org/10.1186/s13059-019-1832-y.
Gawryluk, Ryan M.R., Denis V. Tikhonenkov, Elisabeth Hehenberger, Filip Husnik, Alexander P. Mylnikov, and Patrick J. Keeling. 2019. 'Non-Photosynthetic Predators Are Sister to Red Algae'. *Nature* 572 (7768): 240–43. https://doi.org/10.1038/s41586-019-1398-6.
Gornik, Sebastian G., Febrimarsa, Andrew M. Cassin, James I. MacRae, Abhinay Ramaprasad, Zineb Rchiad, Malcolm J. McConville, et al. 2015. 'Endosymbiosis Undone by Stepwise Elimination of the Plastid in a Parasitic Dinoflagellate'. *Proceedings of the National Academy of Sciences of the United States of America* 112 (18): 5767–72. https://doi.org/10.1073/pnas.1423400112.
Gould, Sven B., Ross F. Waller, and Geoffrey I. McFadden. 2008. 'Plastid Evolution'. *Annual Review of Plant Biology* 59 (1): 491–517. https://doi.org/10.1146/annurev.arplant.59.032607.092915.
Guillou, Laure, Dipankar Bachar, Stéphane Audic, David Bass, Cédric Berney, Lucie Bittner, Christophe Boutte, et al. 2013. 'The Protist Ribosomal Reference Database (PR2): A Catalog of Unicellular Eukaryote Small Sub-Unit RRNA Sequences with Curated Taxonomy'. *Nucleic Acids Research* 41 (D1): D597–604. https://doi.org/10.1093/nar/gks1160.
Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. 'QUAST: Quality Assessment Tool for Genome Assemblies'. *Bioinformatics* 29 (8): 1072–75. https://doi.org/10.1093/bioinformatics/btt086.
Hehenberger, Elisabeth, Fabien Burki, Martin Kolisko, and Patrick J. Keeling. 2016. 'Functional Relationship between a Dinoflagellate Host and Its Diatom Endosymbiont'. *Molecular Biology and Evolution* 33 (9): 2376–90. https://doi.org/10.1093/molbev/msw109.
Hehenberger, Elisabeth, Rebecca J. Gast, and Patrick J. Keeling. 2019. 'A Kleptoplastidic Dinoflagellate and the Tipping Point between Transient and Fully Integrated Plastid Endosymbiosis'. *Proceedings of the National Academy of Sciences* 116 (36): 17934–42. https://doi.org/10.1073/pnas.1910121116.
Howe, C.j, A.c Barbrook, R.e.r Nisbet, P.j Lockhart, and A.w.d Larkum. 2008. 'The Origin of Plastids'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363 (1504): 2675–85. https://doi.org/10.1098/rstb.2008.0050.
Huson, Daniel H., Rewati Tappu, Adam L Bazinet, Chao Xie, Michael P. Cummings, Kay Nieselt, and Rohan Williams. 2017. 'Fast and Simple Protein-Alignment-Guided Assembly of Orthologous Gene Families from Microbiome Sequencing Reads'. *Microbiome* 5 (1): 11. https://doi.org/10.1186/s40168-017-0233-2.
Hyatt, Doug, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. 2010. 'Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification.' *BMC Bioinformatics* 11 (1): 119. https://doi.org/10.1186/1471-2105-11-119.
Irisarri, Iker, Jürgen F H Strassert, and Fabien Burki. 2020. 'Phylogenomic Insights into the Origin of Primary Plastids'. *BioRxiv*, 1–36.
Jain, Chirag, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. 2018. 'High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries'. *Nature Communications* 9 (1): 5114. https://doi.org/10.1038/s41467-018-07641-9.
Janouškovec, Jan, Gita G Paskerova, Tatiana S Miroliubova, Kirill V Mikhailov, Thomas Birley, Vladimir V Aleoshin, and Timur G Simdyanov. 2019. 'Apicomplexan-like Parasites Are Polyphyletic and Widely but Selectively Dependent on Cryptic Plastid Organelles'. Edited by John McCutcheon, Detlef Weigel, Christopher Howe, and Geoff McFadden. *ELife* 8 (August): e49662. https://doi.org/10.7554/eLife.49662.
Janouškovec, Jan, Denis V. Tikhonenkov, Fabien Burki, Alexis T. Howe, Forest L. Rohwer, Alexander P. Mylnikov, and Patrick J. Keeling. 2017. 'A New Lineage of Eukaryotes Illuminates Early Mitochondrial Genome Reduction'. *Current Biology* 27 (23): 3717-3724.e5. https://doi.org/10.1016/j.cub.2017.10.051.
Jin, Jian-Jun, Wen-Bin Yu, Jun-Bo Yang, Yu Song, Claude W. dePamphilis, Ting-Shuang Yi, and De-Zhu Li. 2020. 'GetOrganelle: A Fast and Versatile Toolkit for Accurate de Novo Assembly of Organelle Genomes'. *Genome Biology* 21 (1): 241. https://doi.org/10.1186/s13059-020-02154-5.
Joshi, NA, and JN Fass. 2011. *Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for FastQ Files* (version 1.33). https://github.com/najoshi/sickle.
Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas K.F. Wong, Arndt Von Haeseler, and Lars S. Jermiin.

2017. 'ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates'. *Nature Methods* 14 (6): 587–89. https://doi.org/10.1038/nmeth.4285.
Katoh, Kazutaka, and Daron M. Standley. 2013. 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability'. *Molecular Biology and Evolution* 30 (4): 772–80. https://doi.org/10.1093/molbev/mst010.
Kim, Eunsoo, and Shinichiro Maruyama. 2014. 'A Contemplation on the Secondary Origin of Green Algal and Plant Plastids'. *Acta Societatis Botanicorum Poloniae* 83 (4): 331–36. https://doi.org/10.5586/asbp.2014.040.
Krueger, Felix. 2021. 'Babraham Bioinformatics - Trim Galore!' 2021. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
Lartillot, Nicolas, Nicolas Rodrigue, Daniel Stubbs, and Jacques Richer. 2013. 'Phylobayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment'. *Systematic Biology* 62 (4): 611–15. https://doi.org/10.1093/sysbio/syt022.
Lax, Gordon, Yana Eglit, Laura Eme, Erin M. Bertrand, Andrew J. Roger, and Alastair G.B. Simpson. 2018. 'Hemimastigophora Is a Novel Supra-Kingdom-Level Lineage of Eukaryotes'. *Nature* 564 (7736): 410–14. https://doi.org/10.1038/s41586-018-0708-8.
Li, Linzhou, Sibo Wang, Hongli Wang, Sunil Kumar Sahu, Birger Marin, Haoyuan Li, Yan Xu, et al. 2020. 'The Genome of Prasinoderma Coloniale Unveils the Existence of a Third Phylum within Green Plants'. *Nature Ecology & Evolution* 4 (9): 1220–31. https://doi.org/10.1038/s41559-020-1221-7.
Marin, Birger, Eva C M Nowack, and Michael Melkonian. 2005. 'A Plastid in the Making: Evidence for a Second Primary Endosymbiosis'. *Protist* 156 (4): 425–32. https://doi.org/10.1016/j.protis.2005.09.001.
Martin, Marcel. 2011. 'Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads'. *EMBnet.Journal* 17 (1): 10–12. https://doi.org/10.14806/ej.17.1.200.
Mathur, Varsha, Martin Kolísko, Elisabeth Hehenberger, Nicholas A.T. Irwin, Brian S. Leander, Árni Kristmundsson, Mark A. Freeman, and Patrick J. Keeling. 2019. 'Multiple Independent Origins of Apicomplexan-Like Parasites'. *Current Biology* 29 (17): 2936-2941.e5. https://doi.org/10.1016/j.cub.2019.07.019.
Minh, Bui Quang, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era'. *Molecular Biology and Evolution* 37 (5): 1530–34. https://doi.org/10.1093/molbev/msaa015.
Moreira, David, and Purificación López-García. 2014. 'The Rise and Fall of Picobiliphytes: How Assumed Autotrophs Turned out to Be Heterotrophs'. *BioEssays* 36 (5): 468–74. https://doi.org/10.1002/bies.201300176.
Morozov, A. A., and Yuri P. Galachyants. 2019. 'Diatom Genes Originating from Red and Green Algae: Implications for the Secondary Endosymbiosis Models'. *Marine Genomics* 45 (June): 72–78. https://doi.org/10.1016/j.margen.2019.02.003.
Needham, David M., Susumu Yoshizawa, Toshiaki Hosaka, Camille Poirier, Chang Jae Choi, Elisabeth Hehenberger, Nicholas A. T. Irwin, et al. 2019. 'A Distinct Lineage of Giant Viruses Brings a Rhodopsin Photosystem to Unicellular Marine Predators'. *Proceedings of the National Academy of Sciences* 116 (41): 20574–83. https://doi.org/10.1073/pnas.1907517116.
Not, Fabrice, Klaus Valentin, Khadidja Romari, Connie Lovejoy, Ramon Massana, Kerstin Töbe, Daniel Vaulot, and Linda K Medlin. 2007. 'Picobiliphytes : A Marine Picoplanktonic Algal Group with Unknown Affinities to Other Eukaryotes', no. January.
Nowack, Eva C.M., and Andreas P.M. Weber. 2018. 'Genomics-Informed Insights into Endosymbiotic Organelle Evolution in Photosynthetic Eukaryotes'. *Annual Review of Plant Biology* 69 (1): 51–84. https://doi.org/10.1146/annurev-arplant-042817-040209.
Ponce-Toledo, Rafael I., Philippe Deschamps, Purificación López-García, Yvan Zivanovic, Karim Benzerara, and David Moreira. 2017. 'An Early-Branching Freshwater Cyanobacterium at the Origin of Plastids'. *Current Biology* 27 (3): 1–6. https://doi.org/10.1016/j.cub.2016.11.056.
Priyam, Anurag, Ben J Woodcroft, Vivek Rai, Ismail Moghul, Alekhya Munagala, Filip Ter, Hiten Chowdhary, et al. 2019. 'Sequenceserver: A Modern Graphical User Interface for Custom BLAST Databases'. *Molecular Biology and Evolution* 36 (12): 2922–24. https://doi.org/10.1093/molbev/msz185.
Qiu, Huan, Hwan Su Yoon, and Debashish Bhattacharya. 2013. 'Algal Endosymbionts as Vectors of Horizontal Gene Transfer in Photosynthetic Eukaryotes'. *Frontiers in Plant Science* 4.

https://doi.org/10.3389/fpls.2013.00366.
Rognes, Torbjørn, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. 2016. 'VSEARCH: A Versatile Open Source Tool for Metagenomics'. *PeerJ* 4 (October): e2584. https://doi.org/10.7717/peerj.2584.
Sarai, Chihiro, Goro Tanifuji, Takuro Nakayama, Ryoma Kamikawa, Kazuya Takahashi, and Euki Yazaki. 2020. 'Dinoflagellates with Relic Endosymbiont Nuclei as Models for Elucidating Organellogenesis' 117 (10). https://doi.org/10.1073/pnas.1911884117.
Schön, Max E, Laura Eme, and Thijs J G Ettema. 2020. 'PhyloMagnet: Fast and Accurate Screening of Short-Read Meta-Omics Data Using Gene-Centric Phylogenetics'. *Bioinformatics* 36 (6): 1718–24. https://doi.org/10.1093/bioinformatics/btz799.
Seemann, Torsten. (2013) 2021. *Tseemann/Barrnap*. Perl. https://github.com/tseemann/barrnap.
Seenivasan, Ramkumar, Nicole Sausen, Linda K Medlin, and Michael Melkonian. 2013. 'Picomonas Judraskeda Gen . Et Sp . Nov .: The First Identified Member of the Picozoa Phylum Nov ., a Widespread Group of Picoeukaryotes , Formerly Known as ' Picobiliphytes '' 8 (3). https://doi.org/10.1371/journal.pone.0059565.
Shih, P M, D Wu, A Latifi, S D Axen, D P Fewer, E Talla, A Calteau, et al. 2013. 'Improving the Coverage of the Cyanobacterial Phylum Using Diversity-Driven Genome Sequencing'. *Proc Natl Acad Sci U S A* 110 (3): 1053–58. https://doi.org/10.1073/pnas.1217107110.
Sibbald, Shannon J, and John M Archibald. 2020. 'Genomic Insights into Plastid Evolution'. *Genome Biology and Evolution* 12 (7): 978–90. https://doi.org/10.1093/gbe/evaa096.
Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. 'BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs'. *Bioinformatics* 31 (19): 3210–12. https://doi.org/10.1093/bioinformatics/btv351.
Singer, Anna, Gereon Poschmann, Stefan A Rensing, and Eva C M Nowack. 2017. 'Massive Protein Import into the Early-Evolutionary- Stage Photosynthetic Organelle of the Amoeba Paulinella Chromatophora Massive Protein Import into the Early-Evolutionary-Stage Photosynthetic Organelle of the Amoeba Paulinella Chromatophora', 2763–73. https://doi.org/10.1016/j.cub.2017.08.010.
Stiller, John W. 2014. 'Toward an Empirical Framework for Interpreting Plastid Evolution'. *Journal of Phycology* 50 (3): 462–71. https://doi.org/10.1111/jpy.12178.
Stiller, John W., DeEtte C. Reel, and Jeffrey C. Johnson. 2003. 'A Single Origin of Plastids Revisited: Convergent Evolution in Organellar Genome Content1'. *Journal of Phycology* 39 (1): 95–105. https://doi.org/10.1046/j.1529-8817.2003.02070.x.
Strassert, Jürgen F. H., Iker Irisarri, Tom A. Williams, and Fabien Burki. 2021. 'A Molecular Timescale for Eukaryote Evolution with Implications for the Origin of Red Algal-Derived Plastids'. *Nature Communications* 12 (1): 1879. https://doi.org/10.1038/s41467-021-22044-z.
Strassert, Jürgen F.H., Mahwash Jamy, Alexander P. Mylnikov, Denis V. Tikhonenkov, and Fabien Burki. 2019. 'New Phylogenomic Analysis of the Enigmatic Phylum Telonemia Further Resolves the Eukaryote Tree of Life'. *Molecular Biology and Evolution* 36 (4): 757–65. https://doi.org/10.1093/molbev/msz012.
Timmis, Jeremy N., Michael A. Ayliffe, Chun Y. Huang, and William Martin. 2004. 'Endosymbiotic Gene Transfer: Organelle Genomes Forge Eukaryotic Chromosomes'. *Nature Reviews Genetics* 5 (2): 123–35. https://doi.org/10.1038/nrg1271.
Vargas, Colomban de, Stéphane Audic, Nicolas Henry, Johan Decelle, Frédéric Mahé, Ramiro Logares, Enrique Lara, et al. 2015. 'Eukaryotic Plankton Diversity in the Sunlit Ocean'. *Science* 348 (6237). https://doi.org/10.1126/science.1261605.
Wang, Huai-Chun, Bui Quang Minh, Edward Susko, Andrew J Roger, and Bui Quang. 2017. 'Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation'. *Syst. Biol* 0 (0): 1–19. https://doi.org/10.1093/sysbio/syx068.
Whelan, Simon, Iker Irisarri, and Fabien Burki. 2018. 'PREQUAL: Detecting Non-Homologous Characters in Sets of Unaligned Homologous Sequences'. Edited by John Hancock. *Bioinformatics* 34 (22): 3929–30. https://doi.org/10.1093/bioinformatics/bty448.
Wideman, Jeremy G., Adam Monier, Raquel Rodríguez-Martínez, Guy Leonard, Emily Cook, Camille Poirier, Finlay Maguire, et al. 2020. 'Unexpected Mitochondrial Genome Diversity Revealed by Targeted Single-Cell Genomics of Heterotrophic Flagellated Protists'. *Nature Microbiology* 5 (1): 154–65. https://doi.org/10.1038/s41564-019-0605-4.
Yabuki, Akinori, Ryoma Kamikawa, Sohta A. Ishikawa, Martin Kolisko, Eunsoo Kim, Akifumi S. Tanabe, Keitaro Kume, Ken Ichiro Ishida, and Yuji Inagki. 2014. 'Palpitomonas Bilix Represents a Basal

Cryptist Lineage: Insight into the Character Evolution in Cryptista'. *Scientific Reports* 4: 1–6. https://doi.org/10.1038/srep04641.

Yamada, Norico, Hiroto Sakai, Ryo Onuma, Peter G. Kroth, and Takeo Horiguchi. 2020. 'Five Non-Motile Dinotom Dinoflagellates of the Genus Dinothrix'. *Frontiers in Plant Science* 11. https://doi.org/10.3389/fpls.2020.591050.

Yoon, Hwan Su, Dana C. Price, Ramunas Stepanauskas, Veeran D. Rajah, Michael E. Sieracki, William H. Wilson, Eun Chan Yang, Siobain Duffy, and Debashish Bhattacharya. 2011. 'Single-Cell Genomics Reveals Organismal Interactions in Uncultivated Marine Protists'. *Science* 332 (6030): 714–17. https://doi.org/10.1126/science.1203163.

Zhu, Guan, Mary J. Marchewka, and Janet S. YR 2000 Keithly. 2000. 'Cryptosporidium Parvum Appears to Lack a Plastid Genome'. *Microbiology* 146 (2): 315–21. https://doi.org/10.1099/00221287-146-2-315.