nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-26918-0 OPEN

Single cell genomics reveals plastid-lacking Picozoa are close relatives of red algae

Max E. Schön[1,2], Vasily V. Zlatogursky[1,10], Rohan P. Singh[3], Camille Poirier[4,5,11], Susanne Wilken[5,12], Varsha Mathur[6], Jürgen F. H. Strassert[1,13], 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]✉

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. All known archaeplastids still retain some form of primary plastids, which are widely assumed to have a single origin. Here, we use 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, the analysis of 43 single-cell genomes shows that Picozoa belong to Archaeplastida, specifically related to red algae and the phagotrophic rhodelphids. These picozoan genomes support the hypothesis that Picozoa lack a plastid, and further reveal no evidence of an early cryptic endosymbiosis with cyanobacteria. These findings change our understanding of plastid evolution as they either represent the first complete plastid loss in a free-living taxon, or indicate that red algae and rhodelphids obtained their plastids independently of other archaeplastids.

[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, Kiel, Germany. [5] Monterey Bay Aquarium Research Institute, Moss Landing, CA, USA. [6] Department of Botany, University of British Columbia, Vancouver, BC, Canada. [7] Centre for Ecology and Evolution in Microbial Model Systems—EEMiS, Linnaeus University, Kalmar, Sweden. [8] Laboratory of Microbiology, Wageningen University and Research, 6708 WE Wageningen, The Netherlands. [9] Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [10] Present address: Department of Invertebrate Zoology, Faculty of Biology, St. Petersburg State University, St. Petersburg, Russia. [11] Present address: Department of Zoology, University of Oxford, 11a Mansfield Road, Oxford OX1 3SZ, UK. [12] Present address: Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, The Netherlands. [13] Present address: Department of Ecosystem Research, Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany. ✉email: fabien.burki@ebc.uu.se

The origin of plastids by endosymbiosis between a eukaryotic host and cyanobacteria was a fundamental transition in eukaryotic evolution, giving rise to the first photosynthetic eukaryotes. These ancient primary plastids, estimated to have originated >1.8 billion years ago[1], are found in Rhodophyta (red algae), Chloroplastida (green algae, including land plants), and Glaucophyta (glaucophytes)—together forming the eukaryotic supergroup Archaeplastida[2]. Unravelling the 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[3,4]. Recently, two newly described phyla (Prasinodermophyta and Rhodelphidia) were found to branch as sister to green and red algae, respectively[5,6]. Most transformative was the discovery that rhodelphids are obligate phagotrophs that maintain cryptic non-photosynthetic plastids, implying that the ancestor of red algae was likely mixotrophic, a finding that greatly alters our perspectives on early archaeplastid evolution[5].

While there is substantial evidence that Archaeplastida is a group descended from a photosynthetic ancestor, non-photosynthetic and plastid-lacking lineages have been found to branch near the base or even within archaeplastids in phylogenomic trees. For example, Cryptista (which includes plastid-lacking and secondary plastid-containing species) have been inferred to be sister to either green algae and glaucophytes[7], or red algae[5,8], although other phylogenomic analyses have recovered the monophyly of Archaeplastida to the exclusion of the cryptists[5,9,10]. Another non-photosynthetic group that recently showed affinities to red algae based on phylogenomics is Picozoa[5,9,10]. 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)[11]. 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[12]. Based on 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[13]. However, the hypothesis that the cells were photosynthetic was challenged by the characterisation of SAG data from three picozoan cells isolated by fluorescence-activated cell sorting (FACS)[11]. The analysis of these SAGs revealed neither plastid DNA nor nuclear-encoded plastid-targeted proteins, but the scope of these conclusions is limited due to the small number of analysed cells and the highly fragmented and incomplete obtained data[11]. 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[14]. 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[14]. 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 Californian coast and from the Baltic Sea. Using a gene and taxon-rich phylogenomic dataset, these data allowed us to robustly infer Picozoa as a lineage of archaeplastids, branching with 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 PCR with Picozoa-specific primers[14] or 18S rRNA gene sequencing using general eukaryotic primers, followed by Illumina sequencing of the MDA products (see 'Methods'). The sequencing reads were assembled into genomic contigs, with a total assembly size ranging from 350 kbp to 66 Mbp (Fig. 1a and Supplementary Data 1). 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 (Supplementary Fig. 1). Based on this tree, we identified 6 groups representing 32 SAGs that possessed nearly identical 18S rRNA gene sequences within each group. 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 to 109 Mbp (Fig. 1a and Supplementary Data 1), an increase of 5—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[15], and (ii) a set of 317 conserved marker genes derived from a previous pan-eukaryote phylogenomic dataset[1] that we used here as starting point in downstream analyses (Fig. 1b). These comparisons showed that while most SAGs were highly incomplete (Fig. 1a, 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.[13], but SAG11 was placed outside of these groups. The deep-branching picozoan lineages identified by Moreira and López-García[16], as well as other possibly early-diverging lineages were not represented in our data (Fig. 1c). Interestingly, one CO-SAG (COSAG03) was closely related to the only described species, *Picomonas judraskeda*, for which no genomic data are available (18S rRNA gene 100% identical). 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[17]. 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, Supplementary Fig. 2). 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 seems 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.

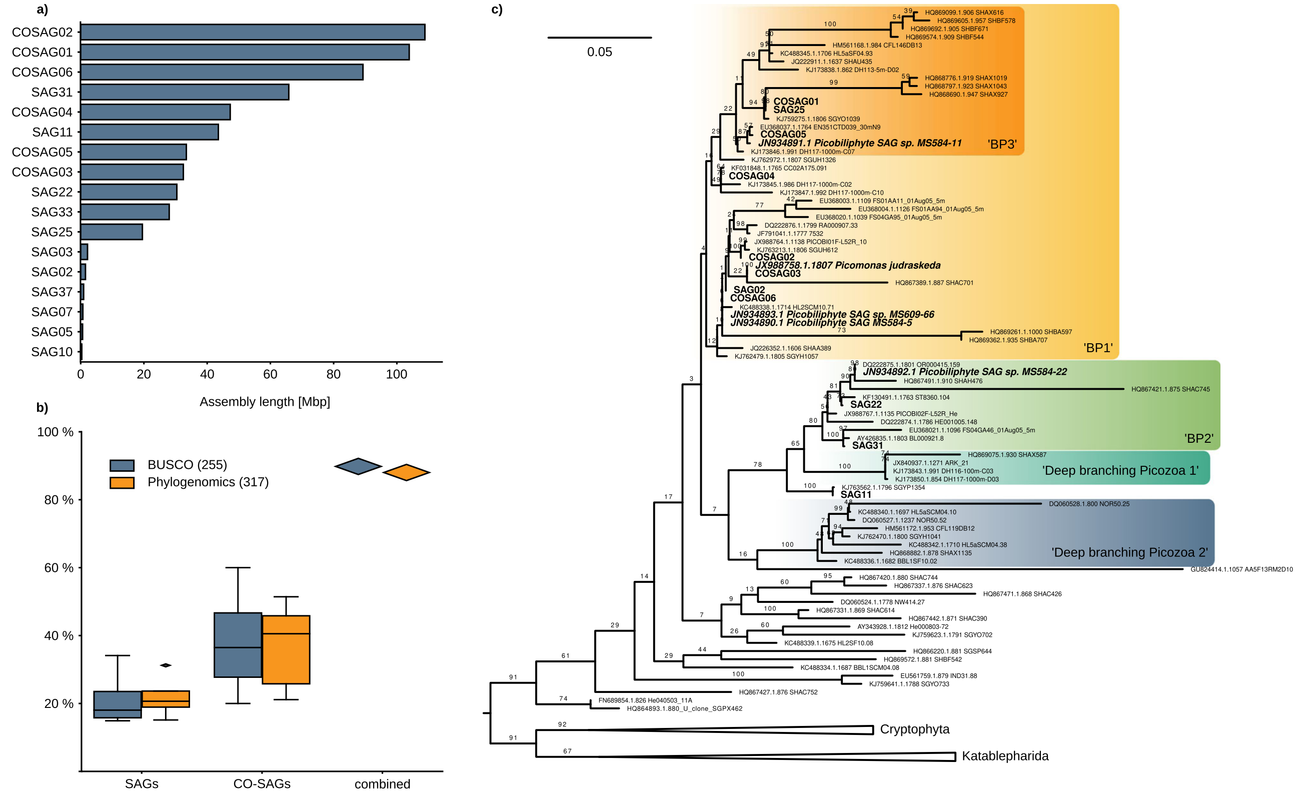

Fig. 1 Description and identification of single-cell genomes. **a** Assembly length in Mbp for 17 SAGs and CO-SAGs used for further analysis. Source data are provided in Supplementary Data 1. **b** Estimated completeness of the ten most complete SAGs (*n* = 4) and CO-SAGs (*n* = 6) as assessed using presence/absence of the BUSCO dataset of 255 eukaryotic markers and a dataset of 317 Phylogenomic marker genes. These ten assemblies were used for the phylogenomic inference. The boxes show the minimum and maximum (excluding outliers), first quartile and third quartile as well as the median. Source data are provided in a Source Data file. **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.[11] in bold italic. The group labels "BP1-3" are taken from Cuvelier et al.[13] and "deep-branching" lineages from Moreira and López-Garcia[16].

Phylogenomic dataset construction. To infer the evolutionary origin of Picozoa, we expanded on a phylogenomic dataset that contains a broad sampling of eukaryotes and a large number of genes that was recently used to study deep nodes in the eukaryotic tree[1]. Homologues from the SAGs and CO-SAGs as well as a number of newly sequenced key eukaryotes were added to each single gene (see Supplementary Table 1 for a list of taxa). After careful examination of the single genes for contamination and orthology based on individual phylogenies (see 'Methods'), we retained all six CO-SAGs and four individual SAGs together with the available SAG MS584-11 from a previous study[11]. The rest of the SAGs were excluded due to poor data coverage (less than five 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 794 taxa and 317 protein-coding genes, with orthologues from Picozoa included in 279 genes (88%) (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 COSAG01, from which we identified orthologues for 163 (51%) of the markers.

Picozoa group with 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 1000 replicates (Supplementary Fig. 3). 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 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 (Supplementary Data 2). 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 (BS = 100%; PP = 1). Consistent with the initial ML tree (Supplementary Fig. 3), red algae and rhodelphids branched together (BS = 95%; PP = 1), with Picozoa as their sister with full support (BS = 100%; PP = 1). This grouping was robust to fast-evolving sites removal analysis (Supplementary Fig. 4), trimming of the 25 and 50% compositionally most biased

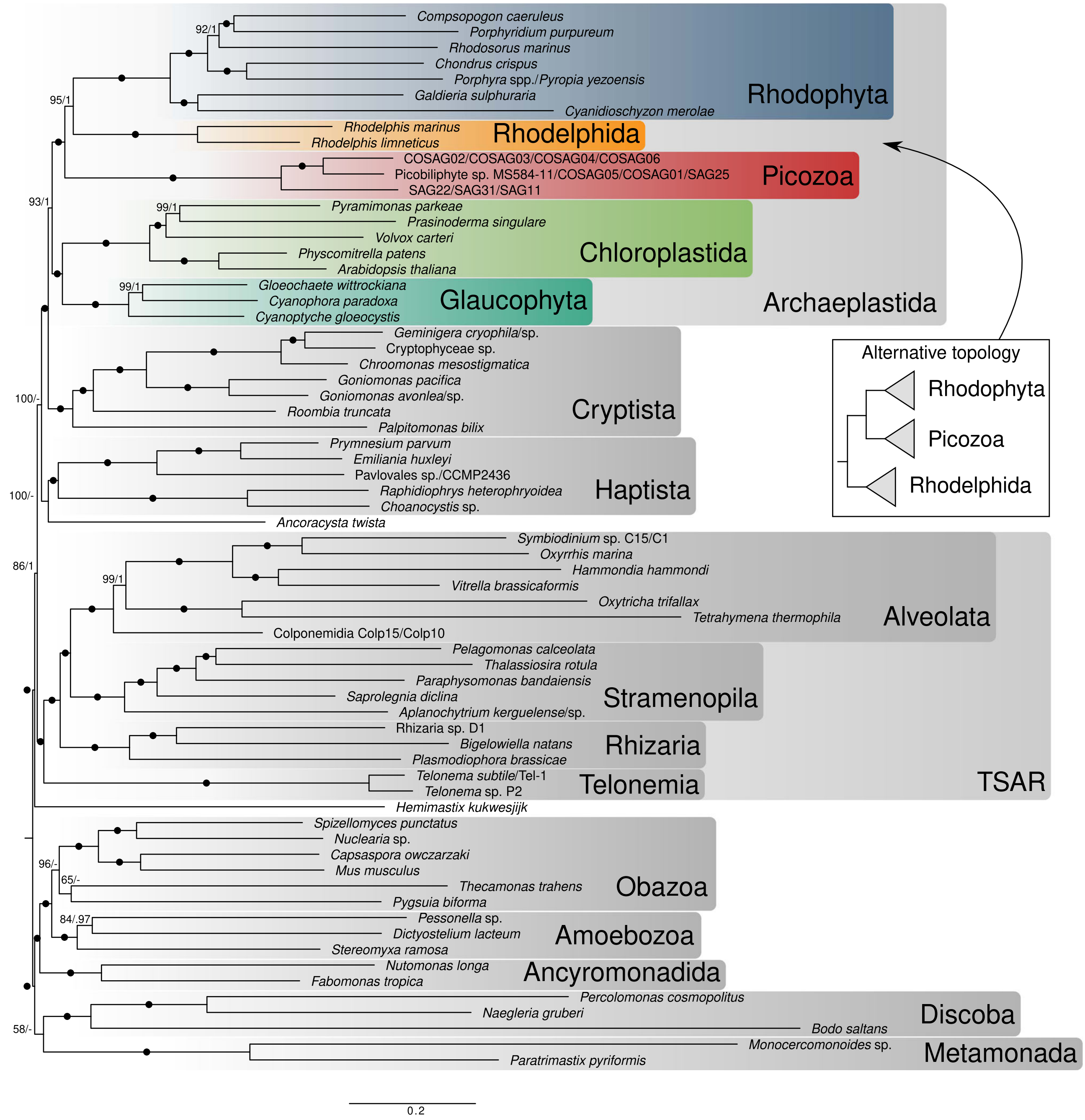

Fig. 2 Maximum likelihood tree of eukaryotic species showing the position of Picozoa. The tree is based on the concatenated alignment of 317 marker genes and was reconstructed using the site-heterogeneous model LG + C60 + F + G-PMSF. Support values correspond to 100 non-parametric bootstrap replicates/posterior probability values estimated using PhyloBayes CAT-GTR + G. Black circles denote full support (=100/1.0). Insert shows the only other topology not rejected in an AU topology test, which was also recovered when trimming the 50% most heterogeneous sites of the alignment.

sites (Supplementary Fig. 5), and was also recovered in a supertree method (ASTRAL-III) consistent with the multi-species coalescent model (Supplementary Fig. 6). Although this group is robust, we observed one variation in the branching order between Picozoa, rhodelphids and red algae when trimming the 50% most heterogeneous sites (Supplementary Fig. 7) and after removing genes with less than two picozoan sequences (Supplementary Fig. 8). In these analyses, Picozoa and red algae were most closely related, although this relationship was never significantly supported. An approximately unbiased (AU) test rejected all tested topologies except in the two cases where Picozoa branched as the closest sister to red algae ($p = 0.237$) and the topology of Fig. 2 ($p = 0.822$; Supplementary Table 2). Finally, we identified in Picozoa and rhodelphids a two amino acids replacement signature in the eukaryotic translation elongation factor 2 protein (SA instead of the ancestral GS residues, see Supplementary Data 3) that was previously shown to unite red and green algae (and land plants), haptophytes and some cryptists[18]. The presence of SA in Picozoa supports their affiliation with red algae and rhodelphids.

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. 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 (Supplementary Data 4). 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[19] (Supplementary Fig. 9). 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[20].

Second, we investigated the possibility that the plastid genome was lost while the organelle itself has been retained—as is the case for *Rhodelphis*[5]. 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[5,21,22]. 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 Supplementary Data S5). 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 predicted N-terminal plastid transit peptides. We also searched for picozoan homologues of all additional proteins ($n = 62$) that were predicted to be targeted to the cryptic plastid in rhodelphids[5]. This search resulted in one protein (Arogenate dehydrogenase, OG0000831) 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 or plastid transport machinery genes, which revealed no obvious candidates. In contrast, we readily identified mitochondrial genes (e.g. homologues of the mitochondrial import machinery from the TIM17/TIM22 family), which further strengthened our inference that the single-cell data are 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) (Supplementary Data 6). We built phylogenies for the 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 and 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 (Supplementary Fig. 10). 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.

Discussion

The 17 SAGs and CO-SAGs of Picozoa obtained in this study provide 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, most likely 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 refs. [23,24] and references therein for review), as well as plastid phylogenetics[25,26]. The phylogenetic support for Archaeplastida based on host (nuclear) data has been less certain[7,8,27], 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[1,9,10]. This position has important implications for our understanding of plastid origins because, in contrast to all other archaeplastids known to date, our results indicate that Picozoa lack plastids and plastid-associated EGTs. The lack of plastid in Picozoa was also inferred based on smaller initial SAG data[11] as well as ultrastructural observation of *P. judraskeda*[14]. 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 (and possibly *Rhodelphis*) arose from one or two 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[24]. 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[28–30], and has been argued before for primary plastids[31–34]. However, the current bulk of cell and molecular evidence suggests that multiple independent origins of primary plastids are unlikely, including several features of plastid biology that are not present in cyanobacteria (e.g., protein targeting

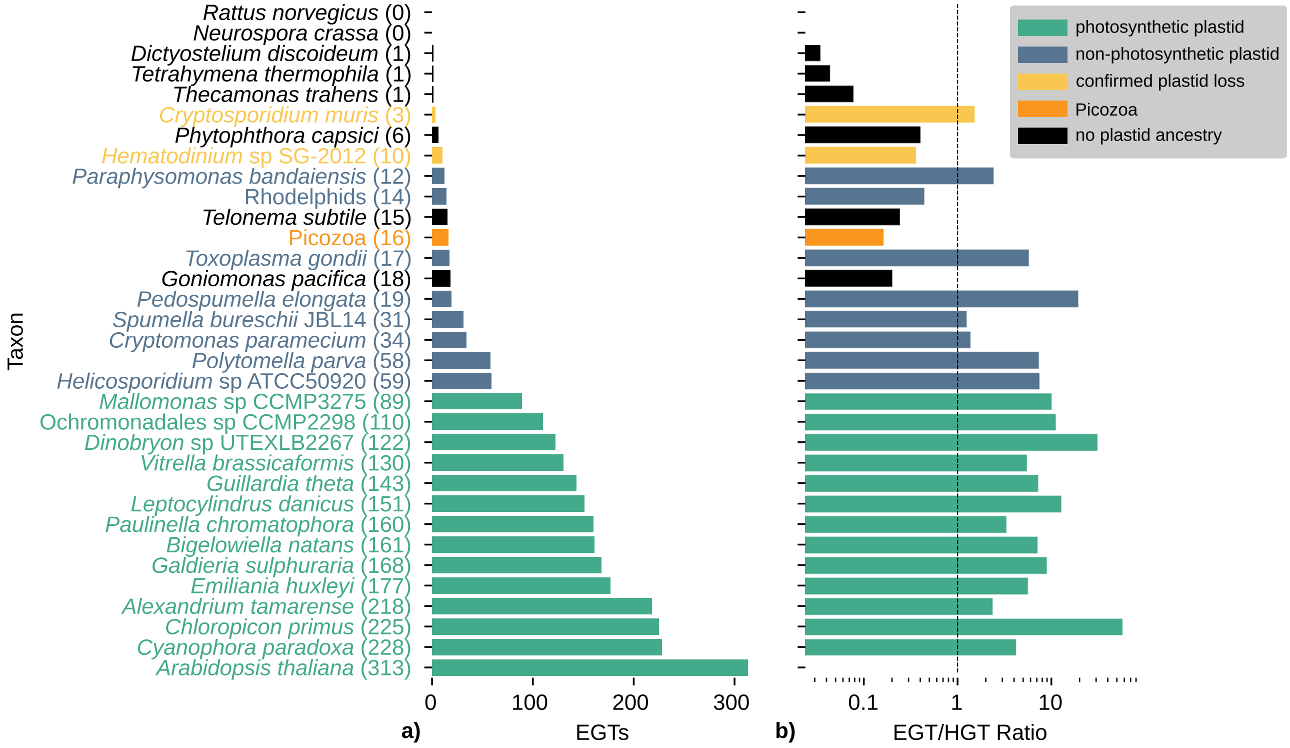

Fig. 3 Inferred endosymbiotic gene transfers (EGT) and horizontal gene transfers (HGT). a Number of inferred endosymbiotic gene transfers (EGT) across a selection of 33 species that represent groups with photosynthetic plastids (green), non-photosynthetic plastids (blue), confirmed plastid loss (yellow) 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* because there were no detectable HGT events. Source data are provided in Supplementary Table 3.

systems, light-harvesting complex proteins, or plastid genome architecture)[23,24,35]. A related explanation could involve a secondary endosymbiosis where the plastid in red algae, for example, was secondarily acquired from a green alga[36]. This latter scenario would be made unlikely 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*[37], certain gregarines[22,38], and the dinoflagellate *Hematodinium*[39]). 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[40,41]. 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[42–46]. 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*[47], or as few as nine genes in tertiary endosymbiosis in dinoflagellates[48]. 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 may have happen after their divergence from rhodelphids.

In this study, we used single-cell genomics to demonstrate that Picozoa are a plastid lacking major lineage of archaeplastids. To our knowledge, 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. In order to discriminate plastid loss from multiple plastid gains in the early archaeplastid evolution, 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[4]. Similar examples of integrated plastid endosymbionts but with apparently very few EGTs are known in dinoflagellates[48,49]. 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 archaeplastid evolution with the search and characterisation of novel diversity of lineages without plastids. The fact that this lineage has never been successfully maintained in culture, with just one study achieving transient culture[14], might indicate a lifestyle involving close association with other organisms (such as symbiosis) and further underscores the enigma of picozoan biology, the lack of information on which hinders our interpretation of their evolution.

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-fractionised. 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 min. 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× 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 (Summit v 6.3.1).

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 1× PBS 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 min 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 (primers PICOBI01F, 5′-CGGATTTTGGCATCACGC-3′, and P01ITS1R, 5′-CATCTCAATGTTCACGTGG-3′), as described in ref. [14] 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 ref. [50]. 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 ref. [50]. For initial screening, 18S rRNA gene amplicons were amplified from each well using the Illumina adapted TAReuk454FWD1 (5′-CCAGCASCYGCGGTAATTCC-3′) and TAReukREV3 (5′-ACTTTCGTTCTTGATYRA-3′) primers targeting the V4 hypervariable region. PCR reactions contained 10 ng of template DNA and 1× 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 s, 50°C for 60 s and 72°C for 90 s; 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 (https://github.com/najoshi/sickle). Paired-end reads were merged using USEARCH v.9.0.2132 when reads had $a \geq 40$ 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[51], 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 350 bp. For six samples, less than 100 ng was used (between 87 and 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 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with default parameter and assembled into genomic contigs with SPAdes v3.13.0[52] 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[53] and rRNA genes were predicted using barrnap v0.9 (https://github.com/tseemann/barrnap) for eukaryotes. All 18S rRNA gene sequences were, together with available reference sequences from the protist ribosomal reference database (PR2, https://pr2-database.org/), aligned with MAFFT E-INS-i v7.429[54] and trimmed with trimal[55] (gap threshold 0.01%). After performing a modeltest using ModelFinder[56] (best model: GTR + R6 + F), a phylogenetic tree was reconstructed in IQ-TREE v2.1.1[57] with 1000 ultrafast bootstrap replicates (see Supplementary Fig. 11 for a tree with extended taxon sampling). Additionally, we estimated the average nucleotide identity (ANI) for all pairs of SAGs using fastANI v1.2[58] (Supplementary Fig. 12). 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[15] with 255 markers for eukaryotes (Supplementary Fig. 13) as well as using the 320 marker phylogenomic dataset as described below. General genome characteristics were computed with QUAST v5.0.2[59]. 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.

For each of these 17 assemblies we estimated the amount of prokaryotic/viral contamination by comparing the predicted proteins against the NCBI nr database using DIAMOND in blastp mode[60]. If at least 60% of all proteins from a contig produced significant hits only to sequences annotated as prokaryotic or viral, we considered that contig to be a putative contamination. In general only a small fraction of each assembly was found to be such a contamination (Supplementary Fig. 14).

Phylogenomics. Existing untrimmed alignments for 320 genes and 763 taxa from ref. [1] were used to create HMM profiles in HMMER v3.2.1[61], 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 (Supplementary Table 1). Each single gene dataset was filtered using PREQUAL v1.02[62] to remove non-homologous residues prior to alignment, aligned using MAFFT E-INS-i, and filtered with Divvier -partial v1.0[63]. Alignments were then used to reconstruct gene trees with IQ-TREE (-mset LG, LG4X; 1000 ultrafast bootstraps with the BNNI optimisation). All trees were manually scrutinised to identify contamination and paralogues. 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 (Supplementary Data 2). 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, v1.12) and concatenated into two supermatrices. Model selection of mixture models was performed using ModelFinder[56] 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)[64] approximation of this mixture model in IQ-TREE and support was assessed with 100 non-parametric bootstraps (see Supplementary Fig. 15 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[65] 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 programme of PhyloBayes. Partial convergence was achieved between chains 1 and 2 with a maxdiff value of 0.26 (Supplementary Fig. 16). The third chain differed only in the position of haptists and *Ancoracysta twista*, but not in the relationships within Archaeplastida and the position of Picozoa (Supplementary Fig. 17).

In order to test the robustness of our results, we additionally performed a fast-site removal analysis[66], iteratively removing the 5000 fastest evolving sites (up to a total of 55,000 removed sites). For each of these 11 alignments, we reconstructed an ML tree using the model LG + C60 + G + F in IQ-TREE with ultrafast bootstraps (1000 UFBoots) and evaluated the support for the branching of Picozoa with rhodelphids and red algae as well as for other groupings (Supplementary Fig. 4). We also performed trimming of the 25 and 50% most heterogeneous sites based on the χ^2 metric[67] and performed tree reconstruction using the same model as above (Supplementary Figs. 5, 7). We also prepared a supermatrix alignment (BMGE trimmed) from 224 genes with at least two Picozoa sequences in the final dataset and performed similar tree reconstruction (model LG + C60 + G + F in IQ-TREE with 1000 ultrafast bootstraps, Supplementary Fig. 8).

Furthermore, we performed a supertree-based phylogenetic reconstruction using ASTRAL-III v5.7.3[68]. We reconstructed gene trees for each of the 317 alignments of the 67-taxa dataset using IQ-TREE (-m TEST -mset LG -mrate G,R4 -madd LG4X,LG4X + F,LG4M,LG4M + F, using 1000 ultrafast bootstraps) and performed multi-locus bootstrapping based on the bootstrap replicates (option -b in ASTRAL-III) (Supplementary Fig. 6).

Finally, we performed an approximately unbiased (AU) test in IQ-TREE of 15 topologies (see Supplementary Table 2), including previously recovered positions of Picozoa (as sister to red algae, cryptists, telonemids, Archaeplastida etc.).

Mitochondrial contig identification and annotation. Using the published picozoan mitochondrial genome (Picozoa sp. MS584-11: MG202007.1 from ref. [19]), BLAST searches were performed on a dedicated sequenceServer[69] to identify mitochondrial contigs in the 43 picozoan SAGs. Putative mitochondrial contigs were annotated using the MFannot server (https://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl). All contigs with predicted mitochondrial genes or whose top hits in the NCBI nr database was the published picozoan mitochondrial genome (MG202007.1, https://www.ncbi.nlm.nih.gov/nuccore/MG202007.1/) were considered to be bona fide mitochondrial contigs and retained (Supplementary Materials). Manual annotation was conducted as needed.

Plastid genes and EGT. GetOrganelle v1.7.1[70] 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[60] 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 (https://www.ncbi.nlm.nih.gov/nuccore/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 ref. [22]. 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' and reconstructed phylogenetic trees using IQ-TREE (-m LG4X; 1000 ultrafast bootstraps with the BNNI optimisation) 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 v0.7[71] to recruit reads and perform gene-centric assembly of these genes[72]. 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[73]. For Picozoa and a selection of 32 photosynthetic or heterotrophic lineages (Supplementary Table 3), 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 optimisation). We then identified trees where 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 *Phytophthora*, 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[17] 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[74] (--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.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data used for the analyses as well as results files such as contigs and single gene trees are available at figshare (https://doi.org/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 PRJNA747736. Source data are provided with this paper.

Code availability

All custom scripts used in this study are available at https://github.com/maxemil/picozoa-scripts (https://doi.org/10.5281/zenodo.5561108) under an MIT license.

Received: 8 May 2021; Accepted: 29 October 2021;
Published online: 17 November 2021

References

1. Strassert, J. F. H., Irisarri, I., Williams, T. A. & Burki, F. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nat. Commun.* **12**, 1879 (2021).
2. Burki, F., Roger, A. J., Brown, M. W. & Simpson, A. G. B. The new tree of eukaryotes. *Trends Ecol. Evol.* **35**, 43–55 (2019).

3. Marin, B., Nowack, E. C. M. & Melkonian, M. A plastid in the making: evidence for a second primary endosymbiosis. *Protist* **156**, 425–432 (2005).
4. Nowack, E. C. M. & Weber, A. P. M. Genomics-informed insights into endosymbiotic organelle evolution in photosynthetic eukaryotes. *Annu. Rev. Plant Biol.* **69**, 1–34 (2018).
5. Gawryluk, R. M. R. et al. Non-photosynthetic predators are sister to red algae. *Nature* **572**, 240–243 (2019).
6. Li, L. et al. The genome of Prasinoderma coloniale unveils the existence of a third phylum within green plants. *Nat. Ecol. Evol.* **4**, 1220–1231 (2020).
7. Burki, F. et al. 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 (2016).
8. Strassert, J. F. H., Jamy, M., Mylnikov, A. P., Tikhonenkov, D. V. & Burki, F. New phylogenomic analysis of the enigmatic phylum telonemia further resolves the eukaryote tree of life. *Mol. Biol. Evol.* **36**, 757–765 (2019).
9. Lax, G. et al. Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature* **564**, 410–414 (2018).
10. Irisarri, I., Strassert, J. F. H. & Burki, F. Phylogenomic insights into the origin of primary plastids. *Syst. Biol.* https://doi.org/10.1093/sysbio/syab036 (2021).
11. Yoon, H. S. et al. Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* **332**, 714–717 (2011).
12. Not, F. et al. Picobiliphytes: a marine picoplanktonic algal group with unknown affinities to other eukaryotes. *Science* **315**, 253–255 (2007).
13. Cuvelier, M. L. et al. Widespread distribution of a unique marine protistan lineage. *Environ. Microbiol.* **10**, 1621–1634 (2008).
14. Seenivasan, R., Sausen, N., Medlin, L. K. & Melkonian, M. Picomonas judraskeda Gen. Et Sp. Nov.: the first identified member of the Picozoa Phylum Nov., a widespread group of picoeukaryotes, formerly known as 'picobiliphytes'. *PLoS ONE* **8**, e59565 (2013).
15. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
16. Moreira, D. & López-García, P. The rise and fall of Picobiliphytes: how assumed autotrophs turned out to be heterotrophs. *Bioessays* **36**, 468–474 (2014).
17. Vargas, Cde et al. Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).
18. Kim, E. & Graham, L. E. EEF2 analysis challenges the monophyly of archaeplastida and chromalveolata. *PLoS ONE* **3**, e2621 (2008).
19. Janouškovec, J. et al. A new lineage of eukaryotes illuminates early mitochondrial genome reduction. *Curr. Biol.* **27**, 3717–3724.e5 (2017).
20. Wideman, J. G. et al. Unexpected mitochondrial genome diversity revealed by targeted single-cell genomics of heterotrophic flagellated protists. *Nat. Microbiol.* **5**, 154–165 (2020).
21. Dorrell, R. G. et al. Principles of plastid reductive evolution illuminated by nonphotosynthetic chrysophytes. *Proc. Natl Acad. Sci. USA*. https://doi.org/10.1073/pnas.1819976116 (2019).
22. Mathur, V. et al. Multiple independent origins of Apicomplexan-like parasites. *Curr. Biol.* **29**, 2936–2941.e5 (2019).
23. Reyes-Prieto, A., Weber, A. P. M. & Bhattacharya, D. The origin and establishment of the plastid in algae and plants. *Annu. Rev. Genet.* **41**, 147–168 (2007).
24. Gould, S. B., Waller, R. F. & McFadden, G. I. Plastid evolution. *Annu. Rev. Plant Biol.* **59**, 491–517 (2008).
25. Shih, P. M. et al. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc. Natl Acad. Sci. USA* **110**, 1053–1058 (2013).
26. Ponce-Toledo, R. I. et al. An early-branching freshwater cyanobacterium at the origin of plastids. *Curr. Biol.* **27**, 386–391 (2017).
27. Yabuki, A. et al. Palpitomonas bilix represents a basal cryptist lineage: insight into the character evolution in Cryptista. *Sci. Rep.* **4**, 4641 (2014).
28. Hehenberger, E., Gast, R. J. & Keeling, P. J. A kleptoplastidic dinoflagellate and the tipping point between transient and fully integrated plastid endosymbiosis. *Proc. Natl Acad. Sci. USA* **116**, 17934–17942 (2019).
29. Sarai, C. et al. Dinoflagellates with relic endosymbiont nuclei as models for elucidating organellogenesis. *Proc. Natl Acad. Sci. USA* **117**, 5364–5375 (2020).
30. Yamada, N., Sakai, H., Onuma, R., Kroth, P. G. & Horiguchi, T. Five non-motile dinotom dinoflagellates of the genus Dinothrix. *Front. Plant Sci.* **11**, 591050 (2020).
31. Stiller, J. W., Reel, D. C. & Johnson, J. C. A single origin of plastids revisited: convergent evolution in organellar genome content. *J. Phycol.* **39**, 95–105 (2003).
32. Larkum, A. W. D., Lockhart, P. J. & Howe, C. J. Shopping for plastids. *Trends Plant Sci.* **12**, 189–195 (2007).
33. Howe, C. J., Barbrook, A. C., Nisbet, R. E. R., Lockhart, P. J. & Larkum, A. W. D. The origin of plastids. *Philos. Trans. R. Soc. B Biol. Sci.* **363**, 2675–2685 (2008).
34. Stiller, J. W. Toward an empirical framework for interpreting plastid evolution. *J. Phycol.* **50**, 462–471 (2014).
35. Bhattacharya, D., Archibald, J. M., Weber, A. P. M. & Reyes-Prieto, A. How do endosymbionts become organelles? Understanding early events in plastid evolution. *Bioessays* **29**, 1239–1246 (2007).
36. Kim, E. & Maruyama, S. A contemplation on the secondary origin of green algal and plant plastids. *Acta Soc. Bot. Pol.* **83**, 331–336 (2014).
37. Zhu, G., Marchewka, M. J. & Keithly, J. S. Cryptosporidium parvum appears to lack a plastid genome. *Microbiology* **146**, 315–321 (2000).
38. Janouškovec, J. et al. Apicomplexan-like parasites are polyphyletic and widely but selectively dependent on cryptic plastid organelles. *Elife* **8**, e49662 (2019).
39. Gornik, S. G. et al. Endosymbiosis undone by stepwise elimination of the plastid in a parasitic dinoflagellate. *Proc. Natl Acad. Sci. USA* **112**, 5767–5772 (2015).
40. Timmis, J. N., Ayliffe, M. A., Huang, C. Y. & Martin, W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* **5**, 123–135 (2004).
41. Archibald, J. M. Genomic perspectives on the birth and spread of plastids. *Proc. Natl Acad. Sci. USA* **112**, 10147–10153 (2015).
42. Burki, F. et al. Re-evaluating the green versus red signal in eukaryotes with secondary plastid of red algal origin. *Genome Biol. Evol.* **4**, 626–635 (2012).
43. Deschamps, P. & Moreira, D. Reevaluating the green contribution to diatom genomes. *Genome Biol. Evol.* **4**, 683–688 (2012).
44. Qiu, H., Yoon, H. S. & Bhattacharya, D. Algal endosymbionts as vectors of horizontal gene transfer in photosynthetic eukaryotes. *Front. Plant Sci.* **4**, 1–8 (2013).
45. Morozov, A. A. & Galachyants, Y. P. Diatom genes originating from red and green algae: implications for the secondary endosymbiosis models. *Mar. Genom.* **45**, 72–78 (2019).
46. Sibbald, S. J. & Archibald, J. M. Genomic insights into plastid evolution. *Genome Biol. Evol.* **12**, evaa096 (2020).
47. Singer, A. et al. Massive protein import into the early-evolutionary-stage photosynthetic organelle of the Amoeba Paulinella chromatophora. *Curr. Biol.* **27**, 2763–2773.e5 (2017).
48. Burki, F. et al. Endosymbiotic gene transfer in tertiary plastid-containing dinoflagellates. *Eukaryot. Cell* **13**, 246–255 (2014).
49. Hehenberger, E., Burki, F., Kolisko, M. & Keeling, P. J. Functional relationship between a dinoflagellate host and its diatom endosymbiont. *Mol. Biol. Evol.* **33**, 2376–2390 (2016).
50. Needham, D. M. et al. A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proc. Natl Acad. Sci. USA* **116**, 20574–20583 (2019).
51. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *Embnet J.* **17**, 10–12 (2011).
52. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
53. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *Bmc Bioinform.* **11**, 119 (2010).
54. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
55. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
56. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Haeseler, Avon & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
57. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
58. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
59. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
60. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
61. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
62. Whelan, S., Irisarri, I. & Burki, F. PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences. *Bioinformatics* **34**, 3929–3930 (2018).
63. Ali, R. H., Bogusz, M. & Whelan, S. Identifying clusters of high confidence homologies in multiple sequence alignments. *Mol. Biol. Evol.* **36**, 2340–2351 (2019).
64. Wang, H.-C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Syst. Biol.* **67**, 216–235 (2018).

65. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.* **62**, 611–615 (2013).
66. Susko, E., Field, C., Blouin, C. & Roger, A. J. Estimation of rates-across-sites distributions in phylogenetic substitution models. *Syst. Biol.* **52**, 594–603 (2003).
67. Viklund, J., Ettema, T. J. G. & Andersson, S. G. E. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol. Biol. Evol.* **29**, 599–615 (2012).
68. Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *Bmc Bioinform.* **19**, 153 (2018).
69. Priyam, A. et al. Sequenceserver: a modern graphical user interface for custom BLAST databases. *Mol. Biol. Evol.* **36**, 2922–2924 (2019).
70. Jin, J.-J. et al. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* **21**, 241 (2020).
71. Schön, M. E., Eme, L. & Ettema, T. J. G. PhyloMagnet: fast and accurate screening of short-read meta-omics data using gene-centric phylogenetics. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz799 (2019).
72. Huson, D. H. et al. Fast and simple protein-alignment-guided assembly of orthologous gene families from microbiome sequencing reads. *Microbiome* **5**, 11 (2017).
73. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
74. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *Peerj* **4**, e2584 (2016).

Acknowledgements

This work was supported by a grant from Science for Life Laboratory available to F.B. and a scholarship from Carl Tryggers Stiftelse to V.V.Z. (PI: F.B.). T.J.G.E. 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 M.E.S. P.J.K. and V.M. were funded by an Investigator Grant from the Gordon and Betty Moore Foundation (https://doi.org/10.37807/GBMF9201). The Pacific Ocean work was supported by GBMF3788 to A.Z.W. 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 programme EcoChange to J.P. 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. Finally, we thank Eunsoo Kim and Sally D. Warring for sharing peptide models from *Palpitomonas bilix* and *Roombia truncata*.

Author contributions

F.B. and J.G.W. conceived the study. For the Baltic samples, V.V.Z. and J.P. performed sampling and cell sorting; V.V.Z. and J.F.H.S. performed genome amplification and sequencing preparation. For the Pacific samples, C.P., S.W. and A.Z.W. conceived, developed and implemented sort protocols; performed sampling, cell sorting and sequencing; C.P. and A.Z.W. performed initial sequence analyses and phylogenetics. M.E.S. under the supervision of T.J.G.E. and F.B. performed assembly of SAGs and Co-SAGs, phylogenomic analyses, and searched for plastid evidence and gene transfers with the help of V.M. and P.J.K. R.P.S. and J.G.W. assembled mitochondrial genomes. M.E.S., F.B., and J.G.W. drafted the manuscript. T.J.G.E., P.K.J., A.Z.W., J.P., V.V.Z. and J.F.H.S. contributed edits to the manuscript. All authors read and approved the final version.

Funding

Open access funding provided by Uppsala University.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-26918-0.

Correspondence and requests for materials should be addressed to Fabien Burki.

Peer review information *Nature Communications* thanks David Moreira and Hwan Su Yoon for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.