ARTICLES
https://doi.org/10.1038/s41564-019-0605-4
nature microbiology

Unexpected mitochondrial genome diversity revealed by targeted single-cell genomics of heterotrophic flagellated protists

Jeremy G. Wideman[1,2,3,4,11*], Adam Monier[1,11], Raquel Rodríguez-Martínez[1,5,11], Guy Leonard[1], Emily Cook[1], Camille Poirier[6,7], Finlay Maguire[1,8], David S. Milner[1], Nicholas A. T. Irwin[9], Karen Moore[1], Alyson E. Santoro[10], Patrick J. Keeling[8], Alexandra Z. Worden[6,7] and Thomas A. Richards[1*]

Most eukaryotic microbial diversity is uncultivated, under-studied and lacks nuclear genome data. Mitochondrial genome sampling is more comprehensive, but many phylogenetically important groups remain unsampled. Here, using a single-cell sorting approach combining tubulin-specific labelling with photopigment exclusion, we sorted flagellated heterotrophic unicellular eukaryotes from Pacific Ocean samples. We recovered 206 single amplified genomes, predominantly from underrepresented branches on the tree of life. Seventy single amplified genomes contained unique mitochondrial contigs, including 21 complete or near-complete mitochondrial genomes from formerly under-sampled phylogenetic branches, including telonemids, katablepharids, cercozoans and marine stramenopiles, effectively doubling the number of available samples of heterotrophic flagellate mitochondrial genomes. Collectively, these data identify a dynamic history of mitochondrial genome evolution including intron gain and loss, extensive patterns of genetic code variation and complex patterns of gene loss. Surprisingly, we found that stramenopile mitochondrial content is highly plastic, resembling patterns of variation previously observed only in plants.

Mitochondria originate from an alphaproteobacteria-like endosymbiont[1], often contain their own genomes, and make ATP via oxidative phosphorylation. Most of the 900–1,100 different mitochondrial proteins are encoded by nuclear DNA[2]. The genome of the progenitor endosymbiont encoded many more genes than extant mitochondrial genomes, many of which have been lost or transferred to the nucleus[3]. Mitochondria-encoded genes vary, but they include those essential for mitochondrial transcription and translation and the electron transport chain (ETC)[4]. Understanding the dynamics of mitochondrial gene loss and gene transfer to the nucleus is, however, limited by poor sampling from diverse lineages, especially heterotrophic flagellates[5].

Microbial eukaryotes, including heterotrophic flagellates, are important constituents of trophic networks and global biogeochemical cycles[5], but most remain uncultured. In the absence of cultures, researchers have used single-cell or targeted metagenome approaches to acquire genomic samples. Three studies have analysed partial nuclear or plastid genomes from photosynthetic marine cells[6,7], and considerable information exists for cultured phytoplankton[8]. Recent studies have tried to fill gaps, relying on hand-picking cells of interest[9,10] or fluorescence-activated cell sorting (FACS). Among the studies using FACS, a few have attempted genome sequencing and assembly[11–15], while others have analysed small subunit ribosomal RNA genes from PCR amplicons[16,17]. These FACS-based studies have used LysoTracker to stain acidic compartments such as food vacuoles[18], or the permissive DNA stain SYBR Green, in preserved cells[15,17], combined with chlorophyll exclusion to enrich for putatively phagotrophic cells. Where genome sequencing has been attempted, it has provided insight into the genome sequences of a few eukaryotes; however, the highly fragmented incomplete nature of single amplified genomes (SAGs) has restricted their use for comparative genomics.

Here, we hypothesize that mitochondrial DNAs (mtDNAs) will be sampled in SAGs at a tractable frequency, enabling comparative analysis. We developed a cell-sorting pipeline to select for the presence of tubulin, combined with chlorophyll exclusion to target heterotrophic flagellates for single-cell isolation. Using these samples, we performed whole-genome amplification and sequencing, recovering numerous and diverse mtDNAs. Using these data, we investigated evolution of mitochondrial gene content, confirming a dynamic pattern of gene loss and patterns of genetic code variation and intron acquisition.

Results

Single-cell sampling of marine flagellates. Heterotrophic protists have diverse lifestyles, which are important for ecosystem function. Since heterotrophs are poorly sampled, diverse methods are needed to recover the diverse forms. Many feed by phagotrophy, using

[1]Living Systems Institute, University of Exeter, Exeter, UK. [2]Wissenschaftskolleg zu Berlin, Berlin, Germany. [3]Department of Biochemistry & Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada. [4]Center for Mechanisms of Evolution, Biodesign Institute, School of Life Sciences, Arizona State University, Tempe, AZ, USA. [5]Laboratorio de Complejidad Microbiana y Ecología Funcional, Instituto Antofagasta, Universidad de Antofagasta, Antofagasta, Chile. [6]Monterey Bay Aquarium Research Institute, Moss Landing, CA, USA. [7]Ocean EcoSystems Biology Unit, Division of Marine Ecology, GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany. [8]Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada. [9]Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada. [10]Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, CA, USA. [11]These authors contributed equally: Jeremy G. Wideman, Adam Monier, Raquel Rodríguez-Martínez. *e-mail: Jeremy.Wideman@asu.edu; T.A.Richards@exeter.ac.uk

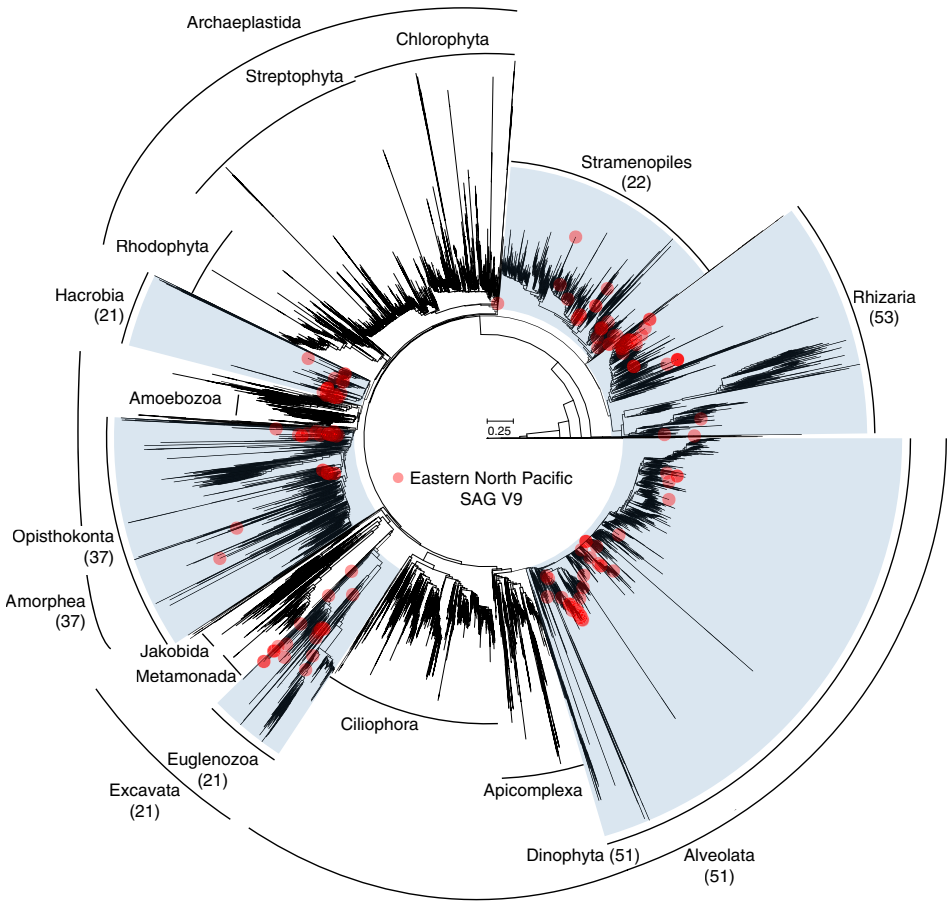


Fig. 1 | V9-nSSU phylogenetic mapping of Monterey Bay SAGs. Maximum-likelihood phylogenetic tree of reference nSSU sequences, retrieved and curated from the PR2 reference database onto which SAG nSSU V9 sequences (from 206 single cells from eastern North Pacific waters sorted by flagellum-targeted flow cytometry) were phylogenetically mapped (red circles). The maximum-likelihood tree was inferred under the GTR-CAT model, on the basis of multiple sequence alignment of 20,939 PR2 representative sequences, at a total of 1,750 sites. Major eukaryotic clades are labelled (See Fig. 2). Groups with representative SAGs are shaded in blue. Numbers in brackets next to taxon names indicate the number of SAGs that were obtained from each taxonomic group.

acidic vacuoles to digest engulfed prey. Previous studies have used FACS combined with LysoTracker, which stains acidic vacuoles, to target actively feeding cells for genomic investigation. However, many heterotrophic flagellates (for example, obligate osmotrophs[19]) do not phagocytose; furthermore, acidic vacuoles can be deployed for other cellular processes[20,21]. Therefore, such approaches can yield false positives[22]. To develop alternative ways of recovering heterotrophic flagellates while limiting the recovery of false positives (for example, prokaryotic cells and detrital particles), we developed an approach combining flow cytometry with tubulin-specific fluorescence staining, following the logic that many protists, especially in the marine water column, use their flagella to find food, hunt prey and, in some cases, infect hosts.

We sorted small tubulin-positive photopigment-lacking cells from the subsurface chlorophyll maximum (SCM, at a depth of 30 m) from the eastern North Pacific Ocean, isolated DNA and performed multiple displacement amplification (MDA)[23]. V9 PCR combined with Sanger sequencing identified 206 SAGs containing eukaryotic nuclear small subunit (nSSU) rRNA genes. Our strategy did not include subcloning of the SSU rDNA amplified template. The Sanger chromatographs did not show evidence of mixed amplicons, suggesting that the recovered V9 sequences were the predominant rDNA signal from each SAG. These were mapped to a universal eukaryotic reference tree, revealing a diversity of nSSU sequences that cluster with heterotrophic flagellates (Figs. 1 and 2). Of these, 189 (92%) branched closely to marine heterotrophic flagellates, demonstrating the efficacy of our approach (Figs. 1 and 2 and Supplementary Table 1). Six nSSU sequences grouped with taxa containing photosynthetic or heterotrophic forms (for example, haptophytes and ochrophytes), and 11 were derived from non-flagellated fungi previously sampled from marine environments[24] (Fig. 2).

A recent TARA Oceans-related project presented a broad diversity of heterotrophic nSSUs from sorted cryopreserved SAGs using SYBR green and chlorophyll exclusion, but enriched for different taxa compared with our analysis[17]. The majority of the heterotrophic flagellates recovered were marine stramenopiles (MASTs) (362, 71%), whereas our protocol recovered some MASTs (12, 6%), but we predominantly recovered cercozoans (53, 26%), marine alveolates (MALVs) and dinoflagellates (51, 25%), choanoflagellates (22, 11%), telonemids (13, 6%) and euglenozoans (20, 10%). Although the samples were obtained from different geographic sites, the differences in taxon diversity recovered highlights the importance of developing approaches that target different cellular attributes.

A rank abundance analysis on nSSU V9 diversity tag sequences was performed using DNA isolated from parallel seawater samples from the same depth and 10 m above. We searched these community profiles for representation of the 206 SAGs and found that our

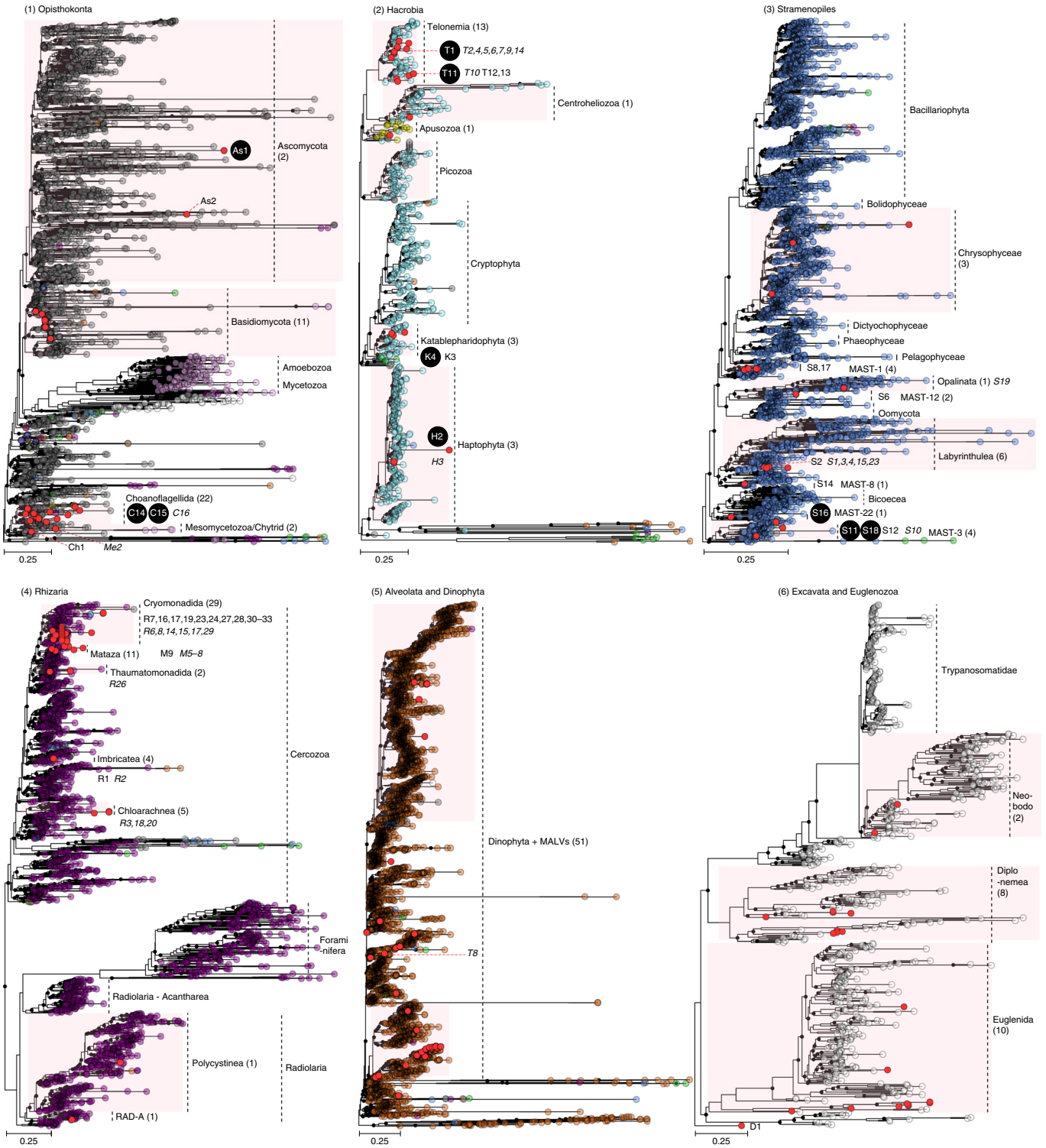


Fig. 2 | Clade-specific maximum-likelihood subtrees showing subsections of the eukaryotic diversity sampled. Six distinct eukaryotic clades from which numerous SAGs were recovered are shown. The SAG nSSU V9 sequences mapped to full-length reference tree that incorporates PR2 reference sequences. For each subtree, specific lineages that attracted SAG V9 sequences are highlighted in pink frames with the lineage name provided. Numbers in parentheses indicate the number of SAG V9 sequences that mapped onto the lineage. SAGs with mitochondrial contigs present are labelled as follows: complete mtDNAs, white font on black circle; near complete, bold; partial genome, italics. SH-like local node supports > 0.9 are shown (black circles). Taxonomic colour legend: Alveolata, orange; Apusozoa, yellow; Euglenozoa, white; Opisthokonta, grey; Stramenopiles, blue; Amoebozoa, pink; Archaeplastida, green; Hacrobia, turquoise; Rhizaria, purple. Scale bars represent the number of estimated substitutions per site.

SAGs were among the rarer taxa identified (Extended Data Fig. 1 and Supplementary Table 2). This was expected as the vast majority of eukaryotes at the SCM are photosynthetic. These data also indicate that many abundant heterotrophs were not recovered in our cell sampling. This could be a product of bias arising from size exclusion or due to the limitation of sampling hundreds of cells from

a community of millions. However, based on the taxonomic diversity recovered, we conclude that our sorting method was effective in targeting heterotrophic flagellates, while excluding phototrophs and non-target cells or particles, and can be applied to various environments (for example, freshwater and potentially, with modification, in soils).

Genome sequencing of single-cell samples. On the basis of the phylogenetic affiliation of the 206 SAGs, we chose 99 cells from under-sampled lineages for DNA sequencing (Fig. 2). We generated 204 Gbp of sequence with a mean (median) sequencing depth of approximately 1.61 (1.35) Gbp per SAG. The resulting reads were assembled, generating a mean assembly size of 14.5 ± 13.8 Mbp (mean ± s.d.) and N50 of 3.4 ± 2.4 kbp per SAG. Full-length nSSU rRNAs recovered from these assemblies were used to confirm the V9 phylogenetic position of the SAGs sampled by BLAST[25] discussed above. In all cases only a single full-length eukaryotic SSU type was recovered from each SAG, suggesting that there was minimal co-sampling of multiple eukaryotic cells. After database curation, three of the nSSU V9 types previously mapped to a tree were determined to be artefactual. Specifically, nSSU V9 sequences recovered from As1 and As2, which mapped as ascomycete fungi (probably due to long-branch attraction), were actually shown to represent a picozoan and a rhizarian, respectively. Furthermore, the T8-SAG assembly contained a complete telonemid SSU; thus, the V9 amplicon sequence was judged to have mapped erroneously as a dinoflagellate (Supplementary Table 3).

To estimate genome completion, we implemented the core eukaryotic genes mapping approach (CEGMA)[13,14], demonstrating recovery of 0.81–48% of CEGMA genes (mean 11.4%, median 6.5%) (Supplementary Tables 3 and 4), comparable to 2–45% recovery seen in other studies[11,13,14]. However, this approach to estimation of genome completion is subject to a range of artefacts stemming from: (1) sampling wells occupied by more than one cell and (2) underestimated completeness due to biases in the CEGMA reference taxa. In some cases, we know that our assemblies are derived from a mixture of eukaryotic, prokaryotic and viral signatures (Supplementary Data 1 https://doi.org/10.6084/m9.figshare.8859014); however, the lack of multiple SSUs in individual SAGs suggests that eukaryote–eukaryote contamination was minimal.

Biased recovery of mtDNAs from SAGs. Mitochondrial genome contigs were recovered in 70 of 99 SAGs (Supplementary Table 4). In the 53 SAGs that demonstrate more than 50% predicted mitochondrial completion, the relative coverage of mtDNAs was higher and more variable (17.0 ± 17.2 times (mean ± s.d.)) than the SAG assemblies (4.9 ± 2.5 times) (Supplementary Table 5), consistent with their derivation from organellar genomes that are often present in higher copy numbers than nuclear genomes. Notably, we observe three distinct groups of SAGs (Fig. 3), those with high nuclear CEGMA completion, those with high mitochondrial coverage, and those with both low and intermediate nuclear completion and mitochondrial coverage (Hotelling's T^2 test[26], $P = 9.07 \times 10^{-13}$), but no SAGs with high recovery of both nuclear DNA and mtDNA (Fig. 3). The mutually exclusive recovery of mtDNAs or higher CEGMA score could be due to several factors: mtDNA could be abundant in some cells, mtDNAs could be preferentially amplified by the SAG methodology (as a product of biased MDA of circular or AT-rich genomes) or alternatively, nuclear DNA sampling and amplification may be retarded relative to mtDNAs due to chromatin wrapping or the complex secondary structures of nuclear DNA. Regardless of the explanation, our data demonstrate that when mitochondrial DNA is preferentially recovered from SAG genomes, nuclear gene sampling is limited. The differences between mitochondrial and nuclear genome coverage, the lack of intervening stop codons in open reading frames and the absence of bordering nuclear sequence

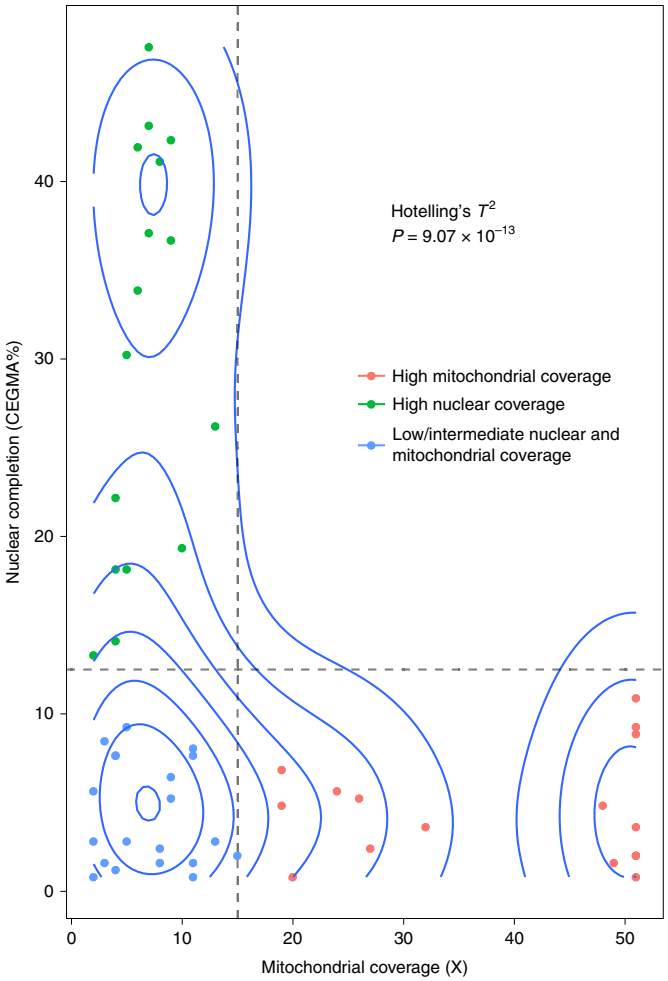


Fig. 3 | Distribution and groupings of mitochondrial sequence coverage relative to estimated nuclear genome completeness. Sequenced genomes showed either high nuclear completion (percentage of CEGMA) (green, $n = 17$ biologically independent mitochondrial genomes), high mitochondrial coverage (X-fold coverage of > 80% of mtDNA) (red, $n = 16$ biologically independent mitochondrial genomes) or simultaneously low nuclear and mitochondrial coverage (blue). The blue density contours were plotted using ggplot2. X-fold coverage of > 80% of each sequenced mtDNA was calculated using BamQC in BAMtools for each SAG with > 50% estimated completion of the mtDNA. This score was plotted against the estimated CEGMA completion scores (%). The result of a Hotelling's T^2 test to assess whether the SAGs with high mitochondrial coverage and those with high nuclear coverage support the rejection of the null hypothesis, which states that these are sampled from the same population. The rejection of the null hypothesis suggests that the there is a fundamental difference between these two SAG subpopulations. The dashed lines partition the samples detected; the empty quadrant lacks sample representation.

in mitochondrial contigs all suggest that we have sequenced bona fide mitochondrial genomes and not mitochondrial insertions into nuclear genomes.

A total of ten unique, complete circular-mapping mtDNAs were assembled from individual SAGs. These include: two telonemids (T1, (GenBank accession no.) MK188946 and T12, MN082145), a katablepharid (K4, MK188945), an unknown alveolate (As1, MK188935 (see below)), two MAST3s (S11, MK188941 and S18, MK188943), a MAST1 (S17, MK188942), a haptophyte (H2, MK188944) and two choanoflagellates (C14, MK188937 and C15, MK188938) (Fig. 4). Two cercozoan mtDNAs were assembled,

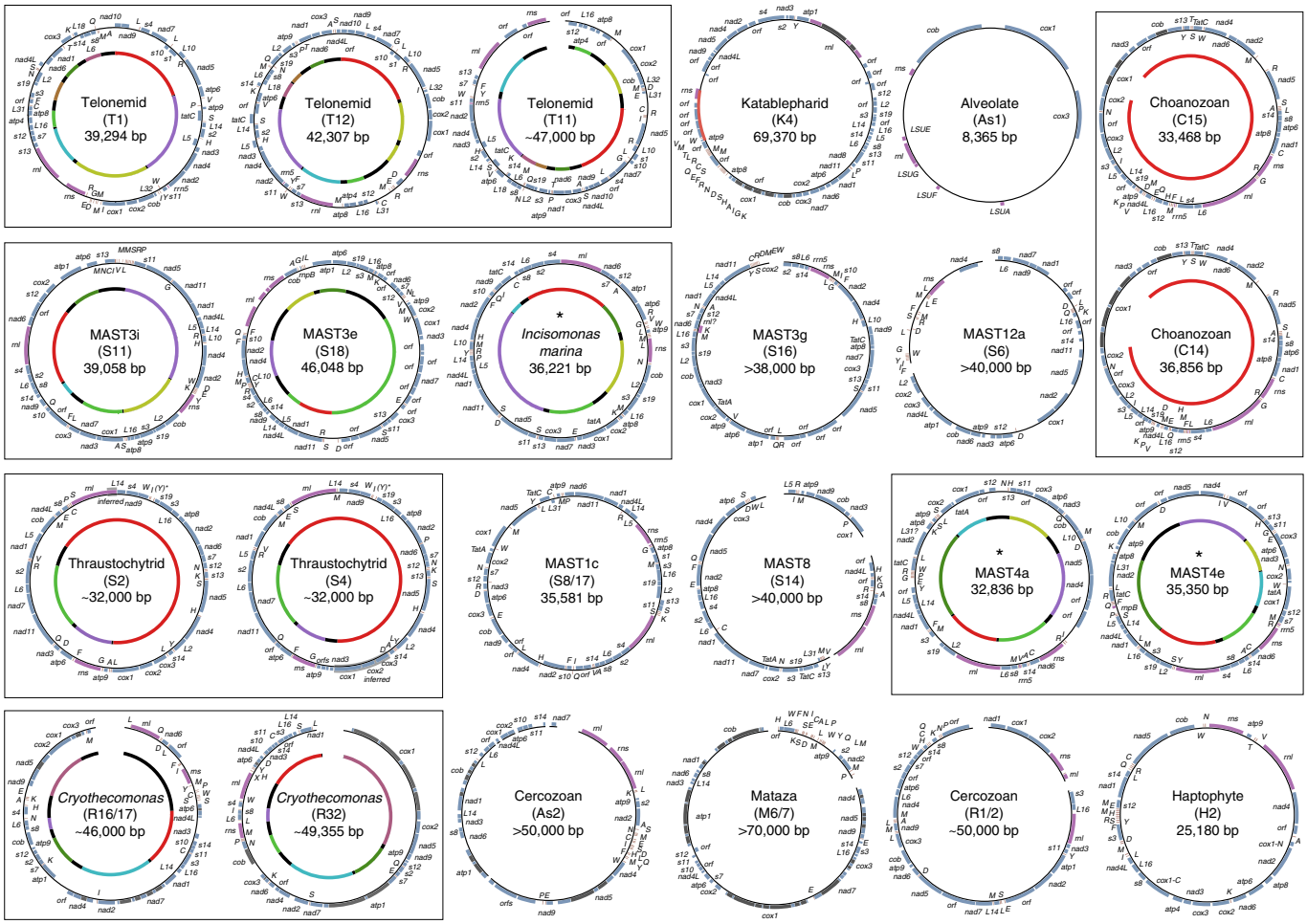


Fig. 4 | Uncharacterized mtDNAs from underrepresented eukaryotic groups. Complete and near-complete mtDNAs assembled from heterotrophic marine flagellate SAGs. Mitochondrial contigs were annotated using mfannot with manual corrections as needed (http://megasun.bch.umontreal.ca/RNAweasel/). MtDNAs are represented as circular diagrams or broken circles if contigs could not be joined. Complete genomes assembled herein using publicly available metagenomes and previously published SAG datasets are marked with an asterisk in the centre of the genome map. Genomes from *Cryothecomonas*-like cells did not map as circular. Where present, coloured central circles correspond to syntenic regions shared between closely related genomes (within boxes). Some mtDNAs were inferred from multiple cells with identical nSSU sequences containing nearly identical stretches of mtDNA sequences that could be stitched together (see Methods). Colour-coded genes: blue, protein coding; purple, rRNA; red, tRNA; dark grey, putative introns.

judged to be linear and are probably complete, on the basis of protein repertoires (R17, MK188936 and R32 MN082144 (Fig. 4)). A further nine unique near-complete (~75–95% complete, see Methods) mtDNAs were identified but could not be completed by additional assembly approaches or by PCR. In some cases, these incomplete mtDNAs provide additional samples validating the provenance of the mitochondrial sampling (Fig. 4). From publicly available datasets[14,27], we assembled three additional complete mtDNAs: *Incisomonas marina* (MAST3), a MAST4a and a MAST4e (Fig. 4, asterisks). Additionally, we identified a probably complete MAST4a mtDNA (EU795181.1), which was wrongly annotated as a bacterial fosmid in the NCBI database. A near-complete mtDNA from a MAST4d SAG[13] was also assembled (Fig. 5). In total, this effort provided 26 complete or near-complete unique mtDNAs from poorly sampled eukaryotic lineages.

To confirm that the mtDNAs belong to the expected taxa, we used our complete and near-complete mitochondrial assemblies as BLAST queries in the NCBI non-redundant database (Supplementary Table 6). The choanoflagellate (C14 and C15), katablepharid (K4) and haptophyte (H2) mtDNAs matched related mtDNAs (Supplementary Table 6). Surprisingly, the top hits for As1 were all alveolate dinoflagellates, indicating conflict between the mitochondrial and nuclear signal (see below). All mtDNAs from stramenopiles (S2, S4, S6, S11, S14, S16 and S18) except S17 retrieved other stramenopiles as best hits. Since the S17 mtDNA did not retrieve sequenced stramenopiles, the Cox1 protein sequence was extracted and used as a BLAST query; this retrieved only stramenopile sequences (Supplementary Table 6). Unexpectedly, the cercozoan mtDNAs and translated Cox1 sequences retrieved stramenopiles and other eukaryotes as top hits, but not sequenced cercozoans (Supplementary Table 6). We therefore reconstructed a multigene phylogeny using stramenopile and cercozoan mtDNAs (Fig. 6). Our cercozoans bifurcated with *Bigelowiella* and *Paracercomonas* and not stramenopiles with full support, confirming their probable identity as rhizarians. These results lead us to conclude that all assembled mtDNAs with the exception of As1 have the same taxonomic affiliation as the nSSUs present in each respective sample.

The As1 SAG contained a single assembled nSSU sequence 94% identical to the nSSU from picozoan MS5584–11 (ref. [11]) and a single circular mtDNA. The mtDNA encodes no transfer RNAs (tRNAs) and only five putative genes, including barely identifiable,

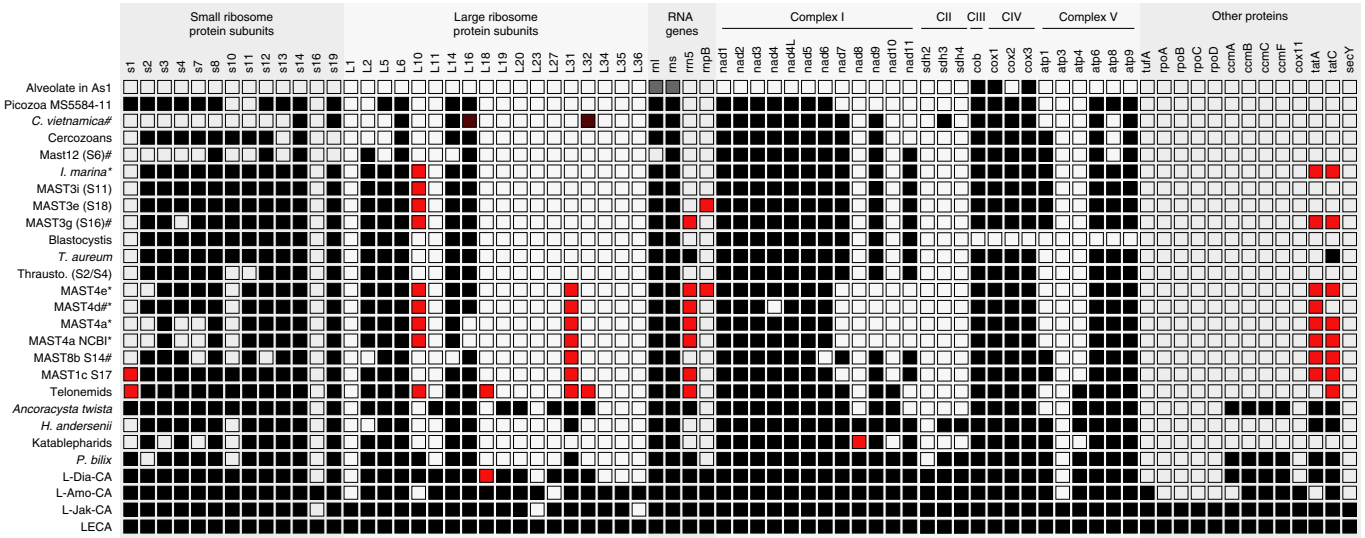


Fig. 5 | Comparison between mtDNA gene repertoires. Mitochondrial genomes newly assembled in this study, previously sequenced mtDNAs and ancestral reconstructions. L-Dia-CA, last Diaphoretickes common ancestor; L-Amo-CA, last Amorphean common ancestor (including malawimonads and collodictyonids); L-Jak-CA, last Jakobid common ancestor; LECA, last Eukaryote common ancestor. Black square, present; empty square, absent; red square, rare protein present. Hashtags indicate incomplete mtDNA and asterisks indicate genomes assembled from publicly available datasets. *C. vietnamica*, *Colponema vietnamica*; *H. andersenii*, *Hemiselmis andersenii*; Thrausto., Thaustrochytrid-like cells.

fragmented, mitochondrial small and large ribosomal RNA genes, *cob*, *cox1* and an unidentified open reading frame that—on the basis of the predicted transmembrane architecture of the protein—is probably a divergent *cox3* (ref. [28]) (Fig. 4). This repertoire is the same as that found in myzozoan alveolates, which differs considerably from picozoan MS5584–11 mtDNA[11,29]. Consistent with the BLAST results reported above, phylogenetic reconstruction using Cox1 demonstrated that the As1-derived protein branches within myzozoans (Extended Data Fig. 2). By contrast, the MS5584–11 Cox1 protein did not branch strongly within any eukaryotic group, as expected for orphan lineages. Given the phylogenetic position of Cox1 and the myzozoan-like coding content and ribosomal fragmentation, we conclude that the mtDNA assembled from As1 is derived from a myzozoan and not a picozoan, a result potentially arising from sampling a cryptic cell–cell interaction (predator–prey or host–parasite).

Evolutionary diversity of mitochondrial gene repertoires. The data reported here enabled us to sample a wide diversity of eukaryotic lineages and compare repertoires of mitochondrial genes (Figs. 4 and 5). Several gene families thought to be encoded in a small subset of eukaryotic mtDNAs were shown to be discontinuously distributed across a diversity of lineages (Fig. 5, red squares). For example, the telonemids possess 40 mitochondrial genes, including *rps1*, *rpl10*, *rpl18* (Extended Data Fig. 3), *rpl31*, *rpl32* and *tatC*, which are thought to be rare; whereas the katablepharids contain a single discontinuously distributed gene (*nad8*). Within the katablepharid mtDNAs, we also identified thirteen additional open reading frames with no similarity to ancestral mitochondrial proteins (Fig. 4). Some of these genes are similar to LAGLIDADG and GIY–YIG homing endonucleases; others may represent undescribed selfish elements or mitochondrial proteins with lineage-specific functions requiring further investigation. MAST mtDNAs encode additional discontinuously distributed gene families, including *tatA* and *tatC* in MAST1c, MAST3g, *I. marina*, MAST4 and MAST8; however, they are absent in other closely related lineages (MAST3i and MAST3e). Previous studies have noted the presence of *tatC* in labyrinthulomycete mtDNAs (KU183024.1 and AF288091.2 (refs. [30,31])), which is absent in our thraustochytrid-related cells (S2 MK188939 and S4 MK188940). We also identified the RNA component (*rnpB*) of RNase P, encoded by MAST3e and MAST4e; *rps1*, encoded by MAST1c; and *rpl31*, encoded by MAST1c, MAST4 and MAST8. The variable nature of stramenopile mtDNA repertoires reveals unexpected dynamics of gene loss and endosymbiotic transfer within this lineage.

Introns in diverse protist mtDNAs. In addition to the standard bacterial-derived mitochondrial gene repertoire, mtDNAs sporadically contain group I and group II self-splicing introns[32]. Using mfannot (http://megasun.bch.umontreal.ca/cgi- bin/mfannot/mfannotInterface.pl), we identified introns in cercozoan, choanoflagellate and katablepharid mtDNAs (Fig. 4, dark grey lines). Interestingly, the two recovered choanoflagellate mtDNAs have 97% identity, but contain different numbers of introns in the *cox1* gene (C14, 4; C15, 2; *Monosiga brevicollis*, 3) (Fig. 4). The two homing endonucleases encoded in C15 are similar to two found in C14 (89% and 98% amino acid identity), but none are similar to *M. brevicollis cox1* (AF538053.1), suggesting a complex pattern of replacement or rapid intron diversification[33].

Similarly, in the cercozoan mtDNAs, whereas no introns can be detected in the R1 and R2 mtDNAs, the cercozoan mtDNAs M9, As2, R32 and R16/17 contain 23, 8, 9 and 8 introns, respectively. Even among mtDNAs from closely related cercozoans (for example, R17 and R32, with 97% nSSU rRNA nucleotide identity (Fig. 4)), the differences between the number of introns and their positions (for example, R32 has four large introns in *cox1*, whereas R17 has no introns in *cox1*) suggests that most of the introns have been acquired recently or that the sampled genomes have undergone repeated invasion by related introns coupled with differential loss of intron variants (for example, in ref. [33]).

Whereas the *Palpitomonas bilix* and some cryptophyte mtDNAs contain no, or very few introns[34–36], the katablepharid K4 mtDNA contains seven introns (Fig. 4, dark grey). The published *Leucocryptos marina* partial mtDNA sequence contains group I introns encoding homing endonucleases in the *cob* and *cox1* genes in identical locations to introns identified in the katablepharid mtDNAs sampled here (49% and 73% amino acid identity, respectively). Our data confirm that multiple mitochondrial evolutionary

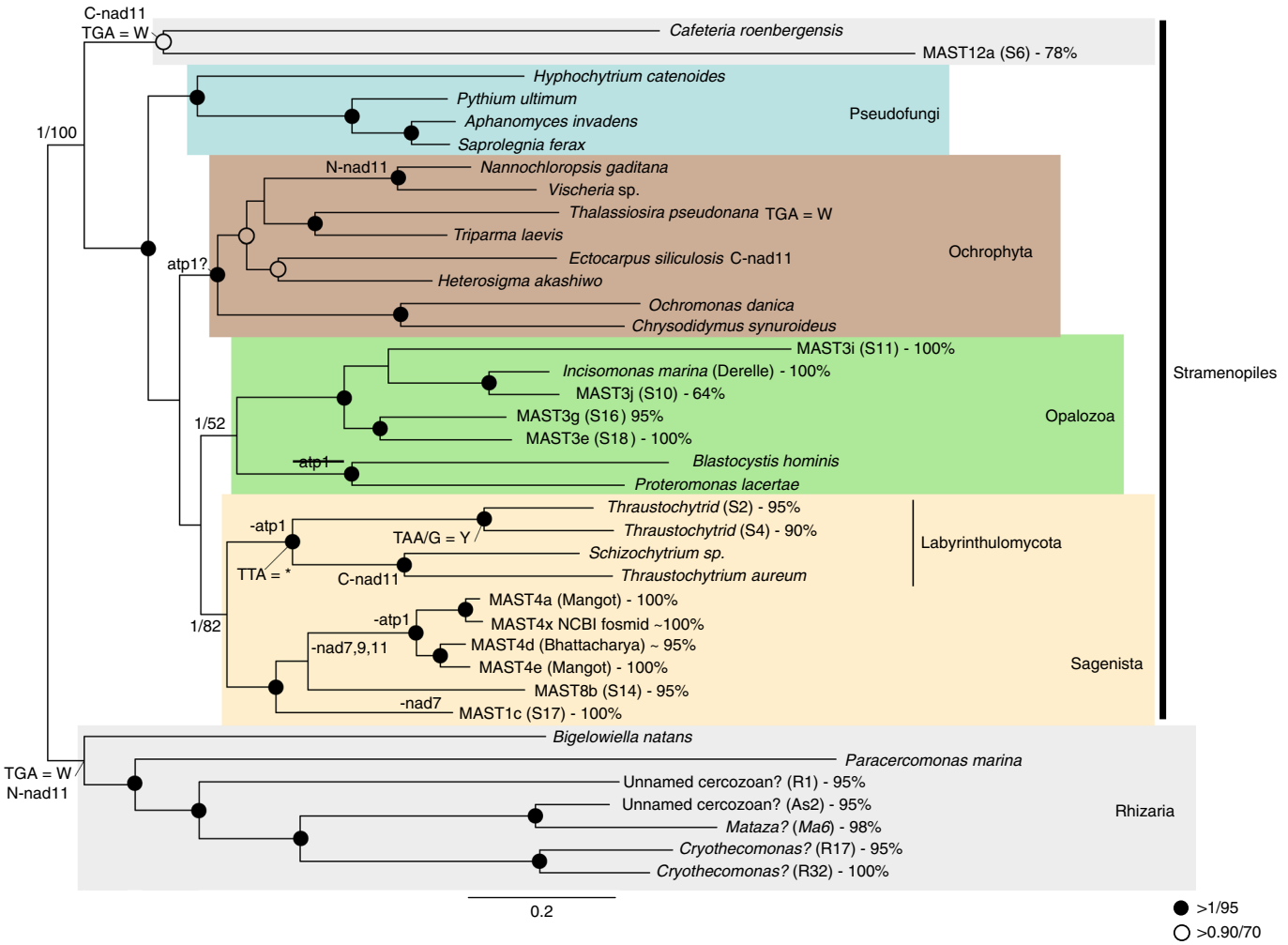


Fig. 6 | Phylogenetic reconstruction of representative stramenopiles using concatenated conserved mitochondria-encoded electron transport chain proteins. Electron transport chain proteins encoded in publicly available mtDNAs and our newly sequenced mtDNAs of stramenopiles and rhizarians were collected, aligned, masked and concatenated, resulting in a 16-protein 4,442-site alignment. We excluded alveolate mtDNAs from this analysis because most of these datasets encode very few (for example, dinoflagellates and apicomplexans) and/or highly divergent proteins (for example, ciliates). Phylogenies were reconstructed and node-support values were calculated using MrBayes v.3.2.6 for posterior probability[82] and RAxML v.8.2.10 for maximum likelihood[83], and are presented at bottom right of the figure (MrBayes/RAxML). The MrBayes tree topology is shown. Changes in genetic code are mapped to nodes as indicated. Genes encoding electron transport chain components (*atp1*, *nad7*, *nad9* and *nad11*) that have putatively moved to the nucleus are mapped to nodes as indicated. The *atp1* gene has been lost within the opalozoans and is indicated with a strikethrough. N-nad11 indicates that the N-terminal domain of *nad11* is encoded in the nucleus, while C-nad11 indicates the C-terminal domain of *nad11* is encoded in the nucleus. Percentages indicate the estimated completeness of each mtDNA presented in this study. 'TGA = W' indicates recoding of TGA from a stop codon to tryptophan; 'TTA = *' indicates TTA is used as a stop codon; 'TAA/G = Y' indicates TAX codons have been recoded to tyrosine.

lineages undergo a high turnover of self-splicing introns, whereas other lineages appear to be free from intron colonisation.

Stramenopile mitochondrial phylogeny identifies organelle-to-nucleus transfers and variations in the mitochondrial genetic code. Using our MAST mtDNAs and sampling from public databases, we sought to calculate a stramenopile mtDNA phylogeny. Using sixteen conserved ETC proteins, we reconstructed a 4,442-site concatenated protein phylogeny using members of cercozoans as an outgroup (Fig. 6). The recovered phylogeny previously established phylogenetic groups including Ochrophyta, Labyrinthulomycota and Pseudofungi[27,37]. Similar to other mitochondrial phylogenies[38], and in contrast to phylogenies based on nuclear proteins, we could not recover Ochrophyta–Pseudofungi sisterhood[27,39], suggesting that there is either a conflicting phylogenetic signal in mtDNA compared with nuclear markers, or that a systematic phylogenetic artefact is present, as discussed previously[38]. Our phylogeny provided some support for the placement of MAST clades previously proposed from nSSU rRNA phylogenies[40] and partially corroborated in a recent multi-gene phylogeny of nuclear-encoded genes[27]. These relationships include an opalozoan group that includes diverse MAST3s (although *Cafeteria roebergensis* and MAST12 fall outside this group) and a sagenistan group containing MAST4s, MAST8, MAST1c (unexpectedly) and labyrinthulomycetes (Fig. 6). Given previous evidence of contradictory relationships identified in stramenopile mitochondrial and nuclear gene phylogenies[38], the branching order presented here should be treated with caution. Additional sampling of stramenopile lineages is required to understand the conflict observed between mitochondrial and nuclear phylogenies.

Using the mitochondrial phylogeny, we sought to polarize mitochondrial traits onto the stramenopile tree. So far, recent and

frequent functional mitochondria-to-nuclear gene transfers have been reported only in Archaeplastida[41] (in particular, green plants). Identification of closely related lineages containing different mitochondrial genes (that is, MAST4s, MAST1 and MAST8) suggests that genes have been transferred relatively recently to the nucleus in stramenopile lineages. Indeed, there are numerous transfers of *atp1* and also partial transfers of *nad11* in multiple stramenopile lineages (Fig. 6 and refs. [39,42]). The mtDNA of MAST1c lacks *nad7* and MAST12 encodes only the N-terminal half of *nad11*, whereas MAST4s lack *nad7*, *nad9* and *nad11*, which are encoded in mtDNAs of most other stramenopiles. We therefore searched for nuclear-encoded versions of these genes in the MAST1c, MAST12 and MAST4 assemblies[14]. In MAST1c, we identified a short contig encoding the C-terminal region of *nad7* adjacent to sequence with no similarity to known proteins or genomic DNA. In MAST12, we identified a contig with a C-terminal domain of *nad11*, which appears to contain spliceosomal introns. Finally, we also identified a contig in a MAST4 assembly encoding *nad9* adjacent to the U4/U6 small nuclear ribonucleoprotein Prp4 along with a number of unidentified proteins (complex I contigs: https://doi.org/10.6084/m9.figshare.7314692). These results suggest that these essential genes have been relocated to the nucleus in these lineages.

Our results demonstrate that stramenopile mtDNA repertoires are extremely diverse compared to other major lineages such as animals and fungi and more closely resemble the dynamic repertoires in the plant lineage[41]. Interestingly, the patterns of variation identified (Fig. 5) generally correspond to a complex pattern of losses previously proposed as 'predictable', in which 'non-core' components of complexes (for example, complex I components *nad7–11*) are more readily transferred to the nucleus than core (defined as energetically central) components (for example, complex I components *nad1–6*)[43]. These results further support the hypothesis that the evolutionary diversification of the mitochondrial lineage, deep within the eukaryotic radiation, was typified by a pattern of early conservation of a wider gene repertoire, followed by numerous independent gene losses[29].

Lastly, our stramenopile and cercozoan mtDNA sequences have enabled us to trace the evolutionary history of three genetic code changes. Several mitochondrial code changes have been documented[44], the most common being the recoding of TGA from a stop codon to tryptophan. This simple change has occurred independently in several lineages, including holozoans, fungi, haptophytes, some diatoms, *C. roebergensis*, cercozoans, picozoan MS584–11, ciliates and some red and green algae (for example, in ref. [4]). We show that the TGA–tryptophan genetic code change observed in *C. roebergensis* is shared with MAST12 and can be traced to their common ancestor. Likewise, since all cercozoans—including sequences presented here—encode TGA as tryptophan, it is likely that the code change occurred very early in this lineage. More notably, we identified a genetic code present in our thraustochytrid mtDNAs (two near-complete (S2 and S4) and three fragmented (S1, S3 and S15)). In these mtDNAs, TGA and TTA (which usually encode leucine) serve as the only termination codons, and TAG and TAA (usually termination codons) encode tyrosine (Extended Data Fig. 4). This finding is supported by the identification of a UUA anticodon tRNA in the SAG mtDNAs (Extended Data Fig. 5). TTA was recoded as a stop codon in *Thraustochytrium aureum* (AF288091.2)[30]; therefore, we can trace stepwise changes in the mtDNA code in this lineage (Fig. 6). These data demonstrate a complex pattern of genetic code variation across stramenopile mitochondria.

Discussion

In this study, we demonstrate that mtDNAs are readily recovered from heterotrophic flagellates using tubulin-targeted single-cell sorting with chlorophyll exclusion followed by whole-genome amplification and sequencing. This represents a method for recovering mtDNAs from diverse uncultured eukaryotes that can be applied, with minor protocol variations, to investigate a range of environments. Such an approach will enable higher resolution studies of protist population structures and effective sampling of multiple genes with different rates of sequence variation that are useful for phylogenetic analyses. The data reported here have substantially increased the available heterotrophic flagellate mtDNA sequences. NCBI reports 9,520 complete mtDNAs, 8,685 from animals, 406 from photosynthetic algae and plants, 334 from fungi, and 50 from animal and plant parasites (apicomplexans and oomycetes). Of the remaining 44 genomes of heterotrophic protists, only 17 are heterotrophic flagellates, spread across the eukaryotic tree of life. Our data more than doubles this representation, adding complete or near-complete genomes from 5 unrepresented or underrepresented groups (Telonemida, Katablepharida, heterotrophic flagellated stramenopiles, Rhizaria and Choanozoa). Further investigation in diverse environments will expand our sampling of heterotrophic protist mtDNAs from across the eukaryotic tree.

Methods

Sample collection and preparation. Seawater was collected in Monterey Bay at 36.6893° N, 122.384° W (Monterey Bay Aquarium Research Institute time series station M2, 56 km from shore) on 7 October 2014 using a Niskin rosette. Water was collected at depths of 20 m and 30 m (SCM as determined by in vivo chlorophyll fluorescence). For general community diversity analyses, 500 ml of water was filtered using a 0.2 µm pore Supor filter (Pall catalogue no. 60301) and extracted using a modification of the DNeasy kit (Qiagen) including the addition of a mechanical lysis by bead-beating[45]. For single-cell sorting, the 30 m water sample was pre-filtered through a 30 µm mesh, then concentrated by gravity 70–100 times onto a 0.8 µm filter and stained with paclitaxel–Oregon Green 488 Conjugate (ThermoFisher, 100 µg ml^{-1} stock made in DMSO) at 10 µM (targeting tubulin from the cytoskeleton). Cells were washed twice with sterile artificial seawater to remove unbound dye, then stained with Hoechst 33342 (targeting DNA) at 2 µg ml^{-1}. Stained samples were diluted in sterile artificial seawater in preparation for flow cytometry.

Cell sorting of marine heterotrophic flagellates. Cells were analysed and sorted on a BD Influx flow cytometer equipped with 488 nm and 355 nm lasers and using sterile nuclease-free PBS pH 7.4 as sheath fluid (ThermoFisher catalogue no. AM9625). A combination of sort windows was applied to select cells that showed green and blue fluorescence (captured by a 520/35 nm and a 460/50 nm bandpass filter for Oregon Green (tubulin) and Hoechst 33342 (DNA), respectively) compared with unstained control samples, and baseline red fluorescence (692/40 nm bandpass filter), indicating the absence of chlorophyll, enabling us to exclude the majority of photosynthetic cells (see Extended Data Fig. 6). Eighteen SAGs with recovered mitochondrial genomes were obtained using this strategy, originating from sort 34 and sort 36 (Supplementary Table 3). A majority of SAGs (52) were recovered from sort 35, in which cells were targeted on the basis of Oregon Green fluorescence only, regardless of Hoechst fluorescence. However, sort windows were refined using the forward-angle light scatter (used as a proxy for cell size) to select cells larger than cyanobacterial cells present in the sample (that is, *Synechococcus*, recognizable by the orange fluorescence of the phycoerythrin in the cells detected with a 572/27 nm bandpass filter).

Targeted cells were sorted into 96-well plates so that all wells received one individual cell (single-cell sorting mode was implemented in BD FACS Sortware v.1.0.0.650), except for the outer column of wells, which was left empty for negative controls. Duplicate plates were obtained for sort 34 and 36 and triplicate plates were obtained for sort 35. The plates were illuminated by UV radiation inside the sort chamber for 2 min before the sort, covered with foil and placed at −80 °C immediately after the sort. The sort quality and correct drop delay were regularly checked by sorting a known number of polystyrene beads (Polysciences, catalogue no. 17153–10) on a slide and counting them on an epifluorescence microscope.

Single-cell genome amplification and sequencing. Samples (sorted cells and negative controls) were lysed for 10 min at 65 °C using alkaline solution from the Repli-g Single Cell Kit (Qiagen) according to the manufacturer's instructions for amplification of genomic DNA from single cells. After neutralization, samples were amplified using the Repli-g reagents to obtain a final volume of 50 µl. The MDA reactions were run in a thermal cycler for 8 h at 30 °C. All materials used during MDA procedures were UV-treated in a HL-2000 HybriLinker UV Crosslinker (UVP) for 30 to 90 min. Single-cell MDA products were screened using Sanger sequencing of the V9 region of the nuclear small subunit (nSSU) rRNA gene amplicons derived from each MDA product. An aliquot of each MDA product was diluted 100-fold in water and 2 µl of this dilution served as the template for each PCR reaction in 25 µl final volume. PCR amplification was carried out using

the primers forward 1389F (5′-TTGTACACACCGCCC-3′) and reverse 1510R (5′-CCTTCYGCAGGTTCACCTAC-3′) as in ref. [46]. PCR products were run on 1% agarose gels stained with GelGreen. Bands were cut using a Visi-Blue Plate (in a UVP transilluminator) to ensure that DNA was not damaged. Amplicons were purified with GeneJet gel extraction kit (Thermo Scientific), quantified with a Qubit fluorometer using the dsDNA BR kit (Invitrogen) and sent for Sanger sequencing (Eurofins).

For Illumina library preparation, an aliquot of each selected MDA sample (including six negative controls) was purified with AMPureXP magnetic beads (Beckmann) following the manufacturer's instructions, quantified with a Qubit and diluted in 10 mM TrisCl (pH 8.0) to a final volume of 130 µl and a concentration of 7.7 ng µl^{-1}. DNA was fragmented using focused acoustic waves (Covaris E220) and concentrated, and libraries were made with Nextflex Rapid DNA library preparation kit and indexes (BIOO Scientific) without PCR amplification. For a subset of samples, 3 µl of each was pooled and concentrated for 450–650 bp size selection using a Blue Pippin 1.5% agarose cassette with R2 marker. The average size of the recovered libraries was 420 bp (with 295 bp inserts). For a second subset, libraries were prepared similarly but used bead-based size selection (420–620 bp), rather than Blue Pippin, quantified by quantitative PCR and pooled in equimolar amounts (213 nM). Library pools were denatured, diluted and 250-paired-end sequenced across two lanes on a HiSeq 2500 using Rapid Run SBS v2 reagents (Illumina). Nine repeated samples were sequenced more deeply on an additional HiSeq 2500 lane to obtain better coverage of these genomes (Supplementary Table 4).

For environmental census of nSSU amplicon libraries, 10 ng environmental DNA was amplified in a two-step protocol following the Illumina amplicon library preparation strategy. Sequencing primers comprised Illumina Nextera pad sequence, a 12-base unique molecular identifier, a spacer sequence and the 1389F or 1510R sequences described above. Two cycles of PCR were performed using these primers in four 25 µl PCR reactions with 2.5 ng DNA in each. Reactions were pooled and purified using AmpureXP beads before adding NexteraXT indexes in a second PCR reaction (21 cycles) to complete the library preparations. Triplicate samples were prepared, pooled in equimolar amounts and quantified by quantitative PCR before 125 bp PE Illumina sequencing.

Single-cell genomic assembly. All SAG sample libraries were assembled using the automatic workflow available at https://doi.org/10.5281/zenodo.192677 or https://github.com/guyleonard/single_cell_workflow. All Illumina read library samples were uploaded to an Amazon EC2 instance (m4.10xlarge) of Ubuntu Linux. The 150 bp PE read libraries were then overlapped using the program PEAR[47] (v.0.9.8) to create 'long' reads; the resulting long reads and the pairs that did not overlap were subsequently quality- and adaptor-trimmed using the program Trim Galore! (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). The resulting libraries were then assembled with SPAdes v.3.7.1 (ref. [48]) using single-cell mode, the careful option and with a combination of k-mers (21, 33 and 55). Quality assessment of the resulting scaffolds was computed with the analysis software QUAST[49] (v.5.0.2) and completeness profiles were made using CEGMA v.2 (ref. [50]). A set of Blobtools (v.1.0) charts were also made with a combination of scaffolds, read mapping and megaBLAST hits to the NCBI nucelotide database[51]. Additional analyses, including KRONA taxonomy charts and Qualimap (v.2.2.1) reports of mapping were computed from these data[52].

Universal nSSU tree, V9 mapping and taxon identification from rDNA assemblies. For the SAG taxonomic classification, nSSU V9 sequences (primers 1389F–1510R[53], a single sequence from each sample) corresponding to each of our 206 SAGs were phylogenetically mapped onto an nSSU reference phylogenetic tree (see Supplementary Table 1) reconstructed from a processed version of the nSSU Protist Ribosomal Reference (PR2) database v.4.4[54] built from GenBank release 203. We first processed the PR2 database by removing short sequences (<400 bp) and/or sequences not spanning the V9 region. In addition, sequences from metazoan organisms (based on PR2 or GenBank taxonomic data) were also discarded. To remove sequence redundancy, the PR2 database was then clustered using CD-HIT v.4.6 (ref. [55]) at 90% sequence identity for sequences classified as Opisthokonta (resulting in 2,694 clusters) and at 98% for non-Opisthokonta sequences (18,245 clusters). This final processed PR2 database, used for subsequent phylogenetic analysis, was composed of 20,939 nSSU clusters, representing a total of 132,235 nSSU sequences.

Cluster representatives, along with SAG V9 sequences, were then aligned with PyNAST v.1.2 (ref. [56]) using the nSSU seed alignment from Silva release v.123 (ref. [57]) as a template alignment. The resulting alignment was then edited and trimmed using Trimal v.1.4 (ref. [58]) to remove sites with gaps in more than 25% of the sequences, but conserving at least half of the original alignment (that is, - gt 0.25 -cons 50 parameters); the final alignment was composed of 1,750 sites. Aligned SAG V9 sequences were removed from the alignment and the PR2-based maximum-likelihood tree was reconstructed using RAxML v.8.2 (multithreaded version; PTHREADS-SSE3)[59] under the GTR model with CAT approximation. SAG V9 sequences were mapped onto the PR2 reference maximum-likelihood tree using the RAxML evolutionary placement algorithm (EPA[60]) under GTR-CAT. To evaluate local node supports, a Shimodaira–Hasegawa-like test[61] was run using FastTree v.2.1 (double precision build[62] in 'accurate' mode (-mlacc 2 -slownni parameters)) and under GTR-CAT. Subsequently to the phylogenetic mapping, and for tree display purposes, taxa with long branches were pruned from the phylogenetic tree; specifically, branches were pruned if the length of the inner node's parent branch was longer than 0.2 substitutions per site or if the terminal branch (that is, linking a leaf to a node) was longer than 3 substitutions per site. These long branches were identified and removed using the Newick utilities package[63]. Note that no SAG V9 sequences were mapped onto these long branches. The figures corresponding to the full, circular PR2 phylogenetic tree with SAG V9 mapping (Fig. 1) and clade-specific trees (Fig. 2) were rendered using the R package ggtree[64].

Contigs from assemblies containing rRNA gene sequences were extracted and used as queries in BLAST searches to confirm V9 mapping results (Supplementary Table 2). Out of 99 sequenced SAGs, 96 V9 placements corresponded closely with the respective assembled nSSU BLAST hits, whereas 3 did not corroborate the V9 mapping results, including both sequences that mapped to ascomycetes and one sequence that mapped to dinophyte. In these cases, the nSSU assembly data clearly indicate that the V9 regions were misplaced during mapping, the first two due to long-branch attraction, and the third due to poor V9 sequence quality. The negative controls contained predominantly very small fragments of contigs most similar to bacterial SSU sequences, possibly due to contamination. However, two of the six total negative control samples subjected to sequencing contained low-coverage contigs most similar to the nSSU sequence of *Cryothecomonas aestivalis* (97–99% identity). Since these controls were taken from different 96-well plates than our samples related to *C. aestivalis*, it is extremely unlikely that these control wells were contaminated either biologically or during library preparation. Instead, it is much more likely that the large signal from the 25 SAG samples that contained contigs with extremely high coverage (sometimes in the thousands) most similar (97–99% identity) to nSSU sequences of *C. aestivalis* interfered with the detector during the sequencing run. The abundance and overrepresentation of these sequences in our SAG samples is a plausible source of the apparent technical contamination (that is, instrument-derived) of these two negative controls, as well as some other samples (see Supplementary Table 1).

Monterey Bay V9 tag sequencing diversity census of whole-seawater samples. Primers and other technical sequences were trimmed from demultiplexed paired-end reads using cutadapt v.1.14 (ref. [65]). To identify artefactual sequences, reads were searched against a V9 reference database (a V9-trimmed version of PR2, clustered at 80% sequence identity using CD-HIT) using BLASTn;[66] reads with no significant hit (E-value < 1×10^{-5}) against the reference database were discarded. Reads were then processed using DADA2 v.1.4 (ref. [67]). On the basis of quality profiles, forward reads were truncated at 150 bp, reverse reads were truncated at 100 bp and reads with more than two expected errors were filtered out. Forward and reverse reads were then independently corrected using run-specific error-rate modelling and dereplicated. Amplicon sequence variants (ASVs; that is, unique sequences) were inferred from these merged reads. Chimeric ASVs were identified and discarded from the datasets. Next, ASVs were assigned a taxonomy using the Ribosomal Database Project naïve Bayesian classifier[68] as implemented in DADA2 and using PR2 as a reference database. ASVs classified as bacteria, archaea, organelle, metazoa or with no eukaryotic supergroup classification (that is, classified only as 'Eukaryota') were discarded. The final Monterey Bay V9 census dataset comprised 1,073 ASVs representing a total of 89,376 quality controlled, merged sequences (Supplementary Table 6). Comparisons between V9 sequences from Monterey Bay SAGs and environmental census, in terms of sequence identity (Supplementary Table 5), were conducted using EMBOSS Water pairwise sequence alignment[69]. Subsequent V9 analyses were conducted using the R package Phyloseq v.1.20 (ref. [70]).

Mitochondrial genome contig identification, reassembly, annotation and confirmation. In 70 of 99 (70%) SAG assemblies, contigs encoding multiple mitochondrial-like genes were identified from the assembly. To ensure that no contaminating DNAs were included in our analysis we removed any contigs showing more than 90% identity to known bacterial, chloroplast or contaminating (for example, fungal) mitochondrial DNAs. To obtain better mitochondrial genome assemblies, reads mapping to each of the identified mitochondrial scaffolds for each SAG were extracted (using BWA (v0.7.17)[71], SAMtools v.1.9 (ref. [72]) and BAMtools v.2.4.0 (ref. [73])) and reassembled with SPAdes 3.7.1 (ref. [48]) in assembly-only mode. The best assemblies were chosen for further analysis, manual adjustment and annotation. Mitochondrial genes, including introns, were annotated using mfannot (http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl) with manual correction as needed. Myzozoan ribosomal fragments in the As1 mitochondrial genome were identified by nhmmer[74] searches with HMMER v.3.1 using hidden Markov models generated from alignments of known fragments[75] (E-value < 1×10^{-5}). Complete or near-complete contigs (see below) were used as queries to identify shorter (that is, encoding only single mitochondrial proteins or RNA genes) bona fide mitochondrial contigs in assemblies from closely related cells. Mitochondrial genome completion percentages were estimated by comparing incomplete mitochondrial genomes to complete (100% circular) or near complete genomes (arbitrarily designated at 95% when no coding sequence or nearly no coding sequence is missing, based on comparisons with closely related taxa).

Samples T1, T12, K4, As1, H2, C14, C15, S11, S17 and S18, as well as *I. marina* and two MAST4 mitochondrial genomes from previous studies were assembled into complete circular genomes. T11 could be assembled into a single contig but could not be circularized. R32 reassembled into a single linear contig with repeats at 5′ and 3′ ends, and was used to identify contigs in similar SAGs. R16 and R17 have identical nSSU rRNA and nearly identical mitochondrial sequences (>99.9%) and were used to infer a probably complete linear mitochondrial genome molecule with repeats at both 5′ and 3′ ends similar to R32. R1 and R2 also have identical SSU and nearly identical mitochondrial sequences (>99.5%). Overlapping contigs from R1 and R2 were joined to form two large contigs that could not be confidently joined further. As2 (mapped on V9-SSU rDNA phylogenetic trees near ascomycetes but was actually a rhizarian cell) contained a *Mataza*-like SSU and assembled into seven contigs that could not be joined but contained nearly all of the predicted genes present in R17 and R32. M6 and M7 mitochondrial genomes were nearly identical (>99.6%) and were used to infer a near-complete *Mataza* mitochondrial genome consisting of two non-overlapping contigs. Two SAGs related to thraustochytrids, S2 and S4, contained single large mitochondrial genome contigs that could not be circularized by PCR. However, on the basis of synteny, the missing stretches of DNA could be inferred, since the missing sequences were present in the reciprocal SAG (shaded and labelled 'inferred' in Fig. 4). S16 (MAST3g) assembled into a single contig and appears to be complete in terms of coding content; however, a repeat region was assembled in the 3′ region of the contig, which appears to contain fragments of *cox2*, which could indicate the presence of an inverted repeat. We could not verify this, as we did not recover any other MAST3g SAGs. Similarly, S6 (MAST12) and S14 (MAST8b) were assembled into two and three contigs, respectively. S14 appears complete with respect to coding content, although the contigs could not be joined. S6 was incomplete, but when compared with the coding content of its closest sequenced relative *C. roebergensis* (which also contains a TGA–W code change), it lacked only 7 of 32 genes and therefore was estimated at 78% complete. Complete and near-compete mitochondrial genomes were visualized using the CGview server[76] and manually edited for figure construction. Closely related mitochondrial genome molecules were manually examined for synteny (Fig. 4, inner coloured circles within boxed mitochondrial genomes).

Since mitochondrial genomes were well represented in SAG assemblies, we calculated the relative coverage of mitochondrial genomes compared with the total SAG assembly. We defined relative coverage as the minimum read coverage over 80% of the representative genome as defined by BamQC in BAMtools output reports (Supplementary Tables 3 and 4). The maximum coverage in the output of this tool was 51×. The relationship between relative mitochondrial genome coverage was compared with that of the nuclear coverage (as estimated by CEGMA%) using the ggplot2 (v.2.2.1)[77] and DescTools (v.0.99.23)[78] packages in R (v.3.4.3)[79]. A two-sided Hotelling's T^2 test[26] (df1 = 2, df2 = 30, T.2 = 44.942, $P = 9.07 \times 10^{-13}$) was used to test whether the groupings of SAGs showing high mitochondrial coverage ($n = 17$) and those with high nuclear coverage ($n = 16$) were sampled from populations showing distinct template profiles. This was performed under the assumption that they were independently sampled from multivariate normal distributions with approximately equal covariance matrices.

Identification of an alternative genetic code in thraustochytrids. The recovered thraustochytrid mitochondrial genomes (S1–S4 and S15) use TTA as a stop codon and contain in-frame TAG and TAA codons that align with conserved tyrosine residues when compared with homologues in other thraustochytrids (Extended Data Fig. 4), suggesting that these stop codons have been reassigned to code for tyrosine. *Cob* genes with internal stop codons were identified in mitochondrial contigs from each SAG and translated using the standard genetic code. These genes were aligned using MUSCLE[80] (https://www.ebi.ac.uk/Tools/msa/muscle/) with publicly available *cob* genes from thraustochytrid mitochondrial genomes (KU183024.1 and AF288091.2) (Extended Data Fig. 4). The lack of a tRNA containing the UAA anticodon and the presence of a tRNA with an AAU anticodon corroborates this hypothesis (Extended Data Fig. 5). Since *T. aureum* is known to have reassigned TTA to a stop codon (GenBank: AF288091.2), these findings support the sister relationship of thraustochytrids and the phylogenetically related SAGs sampled here (Fig. 6).

Phylogenetic analysis of representative stramenopiles from concatenated mitochondria-encoded ETC proteins. Since mitochondrial ribosomes and ribosomal proteins are fast-evolving and have a greater propensity to be lost or relocated to the nucleus, we chose to reconstruct a phylogeny of the stramenopiles using 16 conserved mitochondria-encoded ETC proteins. These included Nad1–Nad7, Nad4L, Nad9, Cob, Cox1–Cox3, Atp6, Atp8 and Atp9. After alignment and manual trimming using Mesquite v.2.75, this resulted in a concatenated alignment with 4,442 sites. IQ-Tree[81] was used for model testing, resulting in LG as the highest scoring model by BIC. Phylogenetic tree reconstructions were performed using MrBayes v.3.2.6 for Bayesian analysis[82]. MrBayes analyses were run with the following parameters: prset aamodelpr = fixed (WAG); mcmcngen = 1,000,000; samplefreq = 1000; nchains = 4; startingtree = random; sumt burnin = 250. Split frequencies were checked to ensure convergence. Maximum-likelihood bootstrap values (100 pseudoreplicates) were obtained using RAxML v.8.2.10 (ref. [83]) under the LG model[84].

Phylogenetic analysis of Cox1 proteins from diverse eukaryotes. Cox1 proteins were collected from representative eukaryote groups from the NCBI non-redundant protein database using BLAST[25]. Resulting sequences were aligned using MUSCLE[80], and manually trimmed to a resulting 402 sites. A phylogenetic reconstruction was conducted using RAxML v.8.2.10 (ref. [83]) (100 bootstrap pseudoreplicates) under the LG model[84].

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

Data availability

Complete mtDNA sequences assembled from this study are available at GenBank under the accession numbers MK188935 to MK188947, MN082144 and MN082145. Sequencing data are available under NCBI BioProject PRJNA379597. Reads have been deposited at NCBI Sequence Read Archive with accession number SRP102236. Partial mtDNA contigs and other important contigs mentioned in the text are available from Figshare at https://doi.org/10.6084/m9.figshare.7314728. Nuclear SAG assemblies are available from Figshare at https://doi.org/10.6084/m9.figshare.7352966. A protocol is available from protocols.io at: https://doi.org/10.17504/protocols.io.ywpfxdn.

Code availability

The bioinformatic workflow is available at https://doi.org/10.5281/zenodo.192677; additional statistical analysis code is available at https://doi.org/10.6084/m9.figshare.9884309.

Received: 19 November 2018; Accepted: 8 October 2019;
Published online: 25 November 2019

References

1. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).
2. Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. The origin and diversification of mitochondria. *Curr. Biol.* **27**, R1177–R1192 (2017).
3. Martin, W. & Herrmann, R. G. Gene transfer from organelles to the nucleus: how much, what happens, and why? *Plant Physiol.* **118**, 9–17 (1998).
4. Gray, M. W. et al. Genome structure and gene content in protist mitochondrial DNAs. *Nucleic Acids Res.* **26**, 865–878 (1998).
5. Worden, A. Z. et al. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. *Science.* **347**, 1257594 (2015).
6. Cuvelier, M. L. et al. Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proc. Natl Acad. Sci. USA* **107**, 14679–14684 (2010).
7. Worden, A. Z. et al. Global distribution of a wild alga revealed by targeted metagenomics. *Curr. Biol.* **22**, R675–R677 (2012).
8. Keeling, P. J. et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.* **12**, e1001889 (2014).
9. Gawryluk, R. M. R. et al. Morphological identification and single-cell genomics of marine diplonemids. *Curr. Biol.* **26**, 3053–3059 (2016).
10. Strassert, J. F. H. et al. Single cell genomics of uncultured marine alveolates shows paraphyly of basal dinoflagellates. *ISME J.* **12**, 304–308 (2018).
11. Yoon, H. S. et al. Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* **332**, 714–717 (2011).
12. Bhattacharya, D. et al. Single cell genome analysis supports a link between phagotrophy and primary plastid endosymbiosis. *Sci. Rep.* **2**, 356 (2012).
13. Roy, R. S. et al. Single cell genome analysis of an uncultured heterotrophic stramenopile. *Sci. Rep.* **4**, 4780 (2014).
14. Mangot, J.-F. et al. Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Sci. Rep.* **7**, 41498 (2017).
15. Seeleuthner, Y. et al. Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. *Nat. Commun.* **9**, 310 (2018).
16. Martinez-Garcia, M. et al. Unveiling in situ interactions between marine protists and bacteria through single cell sequencing. *ISME J.* **6**, 703–707 (2012).
17. Sieracki, M. E. et al. Single cell genomics yields a wide diversity of small planktonic protists across major ocean ecosystems. *Sci. Rep.* **9**, 6025 (2019).
18. Rose, J., Caron, D., Sieracki, M. & Poulton, N. Counting heterotrophic nanoplanktonic protists in cultures and aquatic communities by flow cytometry. *Aquat. Microb. Ecol.* **34**, 263–277 (2004).
19. Richards, T. A. & Talbot, N. J. Horizontal gene transfer in osmotrophs: playing with public goods. *Nat. Rev. Microbiol.* **11**, 720–727 (2013).
20. Vrieling, E. G., Gieskes, W. W. C. & Beelen, T. P. M. Silicon deposition in diatoms: control by the pH inside the silicon deposition vesicle. *J. Phycol.* **35**, 548–559 (1999).

21. Kawai, A., Uchiyama, H., Takano, S., Nakamura, N. & Ohkuma, S. Autophagosome–lysosome fusion depends on the pH in acidic compartments in CHO cells. *Autophagy* **3**, 154–157 (2007).
22. Wilken, S. et al. The need to account for cell biology in characterizing predatory mixotrophs in aquatic environments. *Philos. T. R. Soc. B* **374**, 20190090 (2019).
23. Dean, F. B. et al. Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl Acad. Sci. USA* **99**, 5261–5266 (2002).
24. Richards, T. A., Jones, M. D. M., Leonard, G. & Bass, D. Marine fungi: their ecology and molecular diversity. *Annu. Rev. Mar. Sci.* **4**, 495–522 (2012).
25. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
26. Hotelling, H. The generalization of Student's ratio. *Ann. Math. Stat.* **2**, 360–378 (1931).
27. Derelle, R., López-García, P., Timpano, H. & Moreira, D. A phylogenomic framework to study the diversity and evolution of stramenopiles (=heterokonts). *Mol. Biol. Evol.* **33**, 2890–2898 (2016).
28. Flegontov, P. et al. Divergent mitochondrial respiratory chains in phototrophic relatives of apicomplexan parasites. *Mol. Biol. Evol.* **32**, 1115–1131 (2015).
29. Janouškovec, J. et al. A new lineage of eukaryotes illuminates early mitochondrial genome reduction. *Curr. Biol.* **27**, 3717–3724 (2017).
30. Gray, M. W., Lang, B. F. & Burger, G. Mitochondria of protists. *Annu. Rev. Genet.* **38**, 477–524 (2004).
31. Wang, Z. et al. Complete mitochondrial genome of a DHA-rich protist *Schizochytrium* sp. TIO1101. *Mitochondrial DNA B* **1**, 126–127 (2016).
32. Saldanha, R., Mohr, G., Belfort, M. & Lambowitz, A. M. Group I and group II introns. *FASEB J.* **7**, 15–24 (1993).
33. Goddard, M. R. & Burt, A. Recurrent invasion and extinction of a selfish gene. *Proc. Natl Acad. Sci. USA* **96**, 13880–13885 (1999).
34. Hauth, A. M., Maier, U. G., Lang, B. F. & Burger, G. The *Rhodomonas salina* mitochondrial genome: bacteria-like operons, compact gene arrangement and complex repeat region. *Nucleic Acids Res.* **33**, 4433–4442 (2005).
35. Kim, E. et al. Complete sequence and analysis of the mitochondrial genome of *Hemiselmis andersenii* CCMP644 (cryptophyceae). *BMC Genomics* **9**, 215 (2008).
36. Nishimura, Y. et al. Mitochondrial genome of *Palpitomonas bilix*: derived genome structure and ancestral system for cytochrome *c* maturation. *Genome Biol. Evol.* **8**, 3090–3098 (2016).
37. Riisberg, I. et al. Seven gene phylogeny of heterokonts. *Protist* **160**, 191–204 (2009).
38. Oudot-Le Secq, M.-P., Loiseaux-de Goër, S., Stam, W. T. & Olsen, J. L. Complete mitochondrial genomes of the three brown algae (heterokonta: Phaeophyceae) *Dictyota dichotoma*, *Fucus vesiculosus* and *Desmarestia viridis*. *Curr. Genet.* **49**, 47–58 (2006).
39. Leonard, G. et al. Comparative genomic analysis of the 'pseudofungus' *Hyphochytrium catenoides*. *Open Biol.* **8**, 170184 (2018).
40. Massana, R., del Campo, J., Sieracki, M. E., Audic, S. & Logares, R. Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. *ISME J.* **8**, 854–866 (2014).
41. Kannan, S., Rogozin, I. B. & Koonin, E. V. MitoCOGs: clusters of orthologous genes from mitochondria and implications for the evolution of eukaryotes. *BMC Evol. Biol.* **14**, 237 (2014).
42. Ševčíková, T. et al. A comparative analysis of mitochondrial genomes in eustigmatophyte algae. *Genome Biol. Evol.* **8**, 705–722 (2016).
43. Johnston, I. G. & Williams, B. P. Evolutionary Inference across eukaryotes identifies specific pressures favoring mitochondrial gene retention. *Cell Syst.* **2**, 101–111 (2016).
44. Keeling, P. J. Genomics: evolution of the genetic code. *Curr. Biol.* **26**, R851–R853 (2016).
45. Demir-Hilton, E. et al. Global distribution patterns of distinct clades of the photosynthetic picoeukaryote *Ostreococcus*. *ISME J.* **5**, 1095–1107 (2011).
46. Logares, R. et al. Patterns of rare and abundant marine microbial eukaryotes. *Curr. Biol.* **24**, 813–821 (2014).
47. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
48. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single- cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
49. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
50. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
51. Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M. & Blaxter, M. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front. Genet.* **4**, 237 (2013).
52. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, btv566 (2015).
53. Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W. & Huse, S. M. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS ONE* **4**, e6372 (2009).
54. Guillou, L. et al. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* **41**, D597–D604 (2013).
55. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
56. Caporaso, J. G. et al. PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**, 266–267 (2010).
57. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2012).
58. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
59. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
60. Berger, S. A., Krompass, D. & Stamatakis, A. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst. Biol.* **60**, 291–302 (2011).
61. Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114–1116 (1999).
62. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
63. Junier, T. & Zdobnov, E. M. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* **26**, 1669–1670 (2010).
64. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
65. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
66. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
67. Callahan, B. J. et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
68. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microb.* **73**, 5261–5267 (2007).
69. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
70. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
71. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
72. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
73. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Stromberg, M. P. & Marth, G. T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691–1692 (2011).
74. Wheeler, T. J. & Eddy, S. R. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**, 2487–2489 (2013).
75. Jackson, C. J. et al. Broad genomic and transcriptional analysis reveals a highly derived genome in dinoflagellate mitochondria. *BMC Biol.* **5**, 41 (2007).
76. Grant, J. R. & Stothard, P. The CGView server: a comparative genomics tool for circular genomes. *Nucleic Acids Res.* **36**, W181–W184 (2008).
77. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer–Verlag, 2009).
78. Signorell, A. *DescTools: tools for descriptive statistics* R package v.0.99.23 (2017).
79. R Core Team. *R: a Language and Environment for Statistical Computing* http://www.r-project.org/ (R Foundation for Statistical Computing, 2013).
80. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
81. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
82. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
83. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
84. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).

Acknowledgements

We thank F. Lang and N. Beck for annotation assistance and access to an unreleased version of mfannot, D. Price for assistance with picozoan SAG data, and C. Dunn for

discussions and encouragement. This project was supported by a Gordon and Betty Moore foundation grant (GBMF3307) to T.A.R., A.E.S., A.Z.W. and P.J.K. and a Philip Leverhulme Award (PLP-2014–147) to T.A.R.. Field sampling was supported by the David and Lucile Packard Foundation and GBMF3788 to A.Z.W., T.A.R. and A.M. are supported by Royal Society University Research Fellowships. J.G.W. was supported by the European Molecular Biology Organization Long-term Fellowship (ALTF 761–2014) co-funded by the European Commission (EMBOCOFUND2012, GA-2012–600394) support from Marie Curie Actions and a College for Life Sciences Fellowship at the Wissenschaftskolleg zu Berlin. R.R.-M. is supported by CONICYT FONDECYT 11170748. F.M. is supported by Genome Canada.

Author contributions

J.G.W. performed bioinformatic and phylogenetic analyses and wrote the manuscript. R.R.-M. performed molecular biological analyses. A.M. performed bioinformatic and phylogenetic analyses and G.L. performed bioinformatic analyses. E.C. and C.P. collected the samples and performed flow cytometry. F.M. performed statistical and bioinformatic analyses. D.M. performed molecular biological experiments and generated biochemical reagents. K.M. performed genome sequencing. N.A.T.I. analysed genomic data. T.A.R. devised the project. J.G.W., A.E.S., P.J.K., A.Z.W. and T.A.R. supervised the project and wrote the manuscript. All authors contributed to the editing of the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

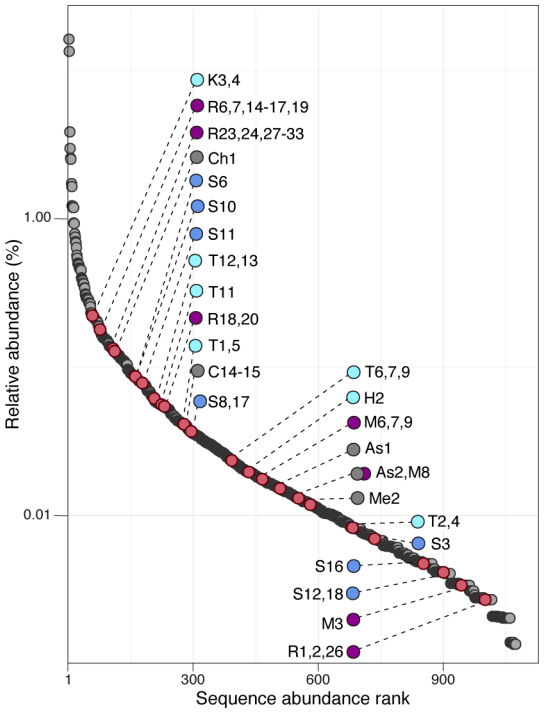
Extended data is available for this paper at https://doi.org/10.1038/s41564-019-0605-4.

Supplementary information is available for this paper at https://doi.org/10.1038/s41564-019-0605-4.

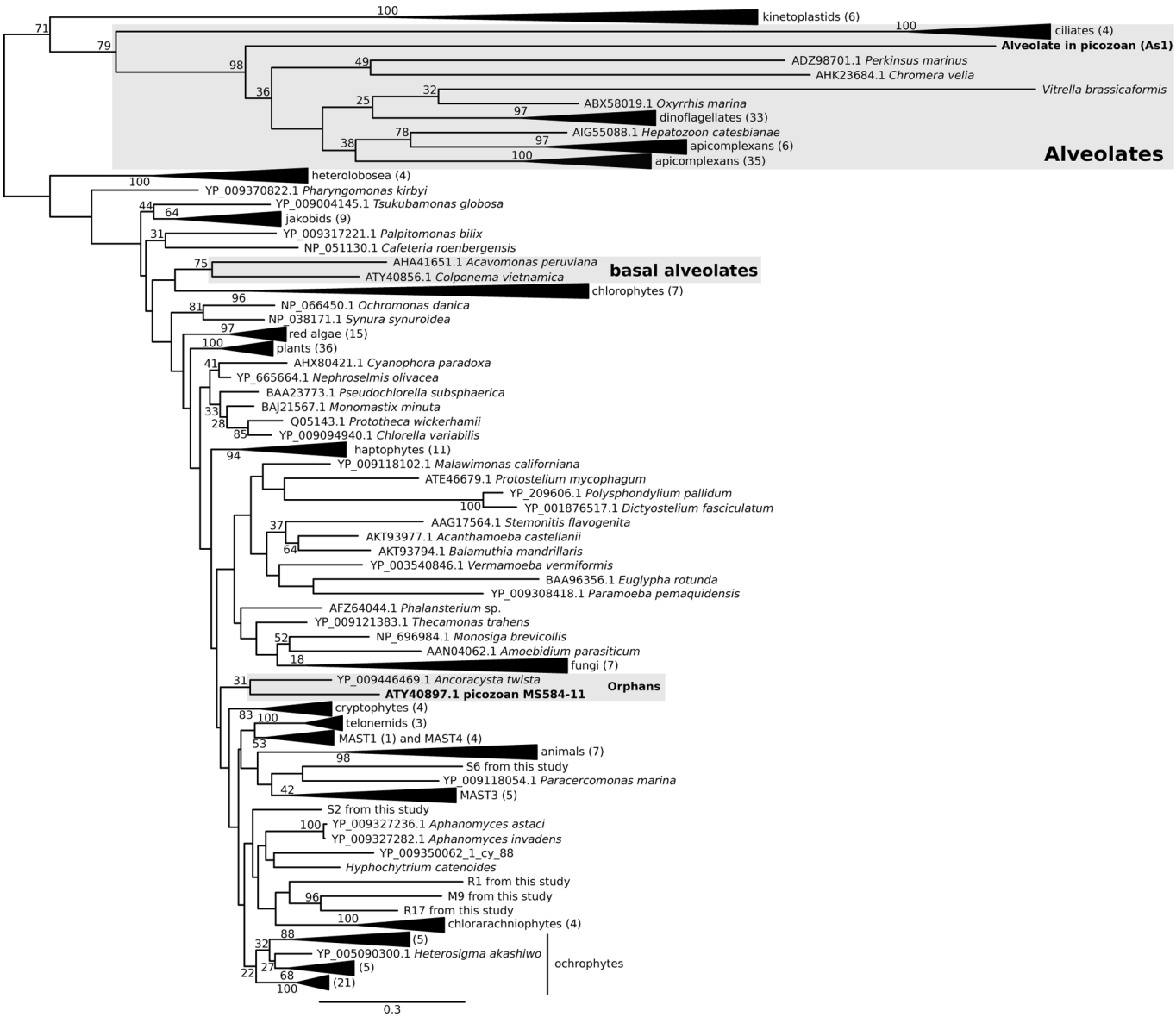
Correspondence and requests for materials should be addressed to J.G.W. or T.A.R.

Reprints and permissions information is available at www.nature.com/reprints.

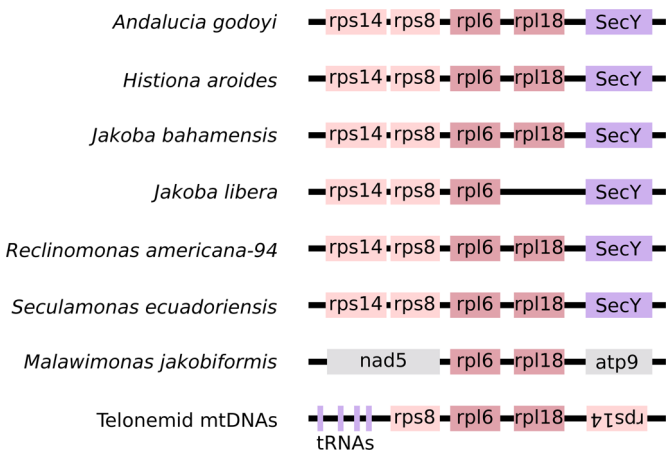
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



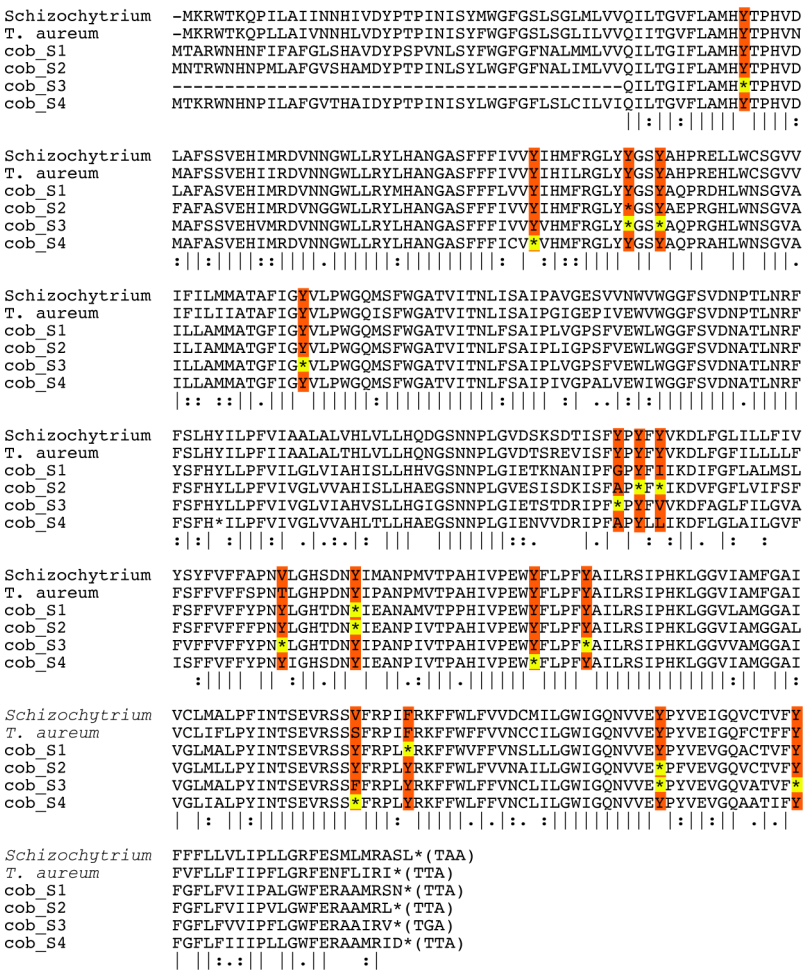
Extended Data Fig. 1 | Rank abundance curve of amplicon sequence variants (ASVs) from the Monterey Bay nSSU-V9 environmental census. Relative abundances correspond to the mean relative abundance of each ASV in samples from two depths (20 m and 30 m) of eastern North Pacific station M2 (SAGs were recovered from 30 m depth). ASV sequences identical to V9 sequences from SAGs with recovered mitochondrial genomic information are represented by red circles; ASVs with no identical sequence match to V9 SAGs with mitochondrial data are represented by grey circles. For each ASV identical to a SAG V9, the corresponding SAG codenames are provided (in some cases there are multiple of each type). Samples are coloured according to taxonomic affiliation in V9 sorting. Blue, stramenopile; teal, hacrobian; purple, rhizarian; brown, opisthokont. See Supplemenatry Table 7 for details on ASV relative abundance.



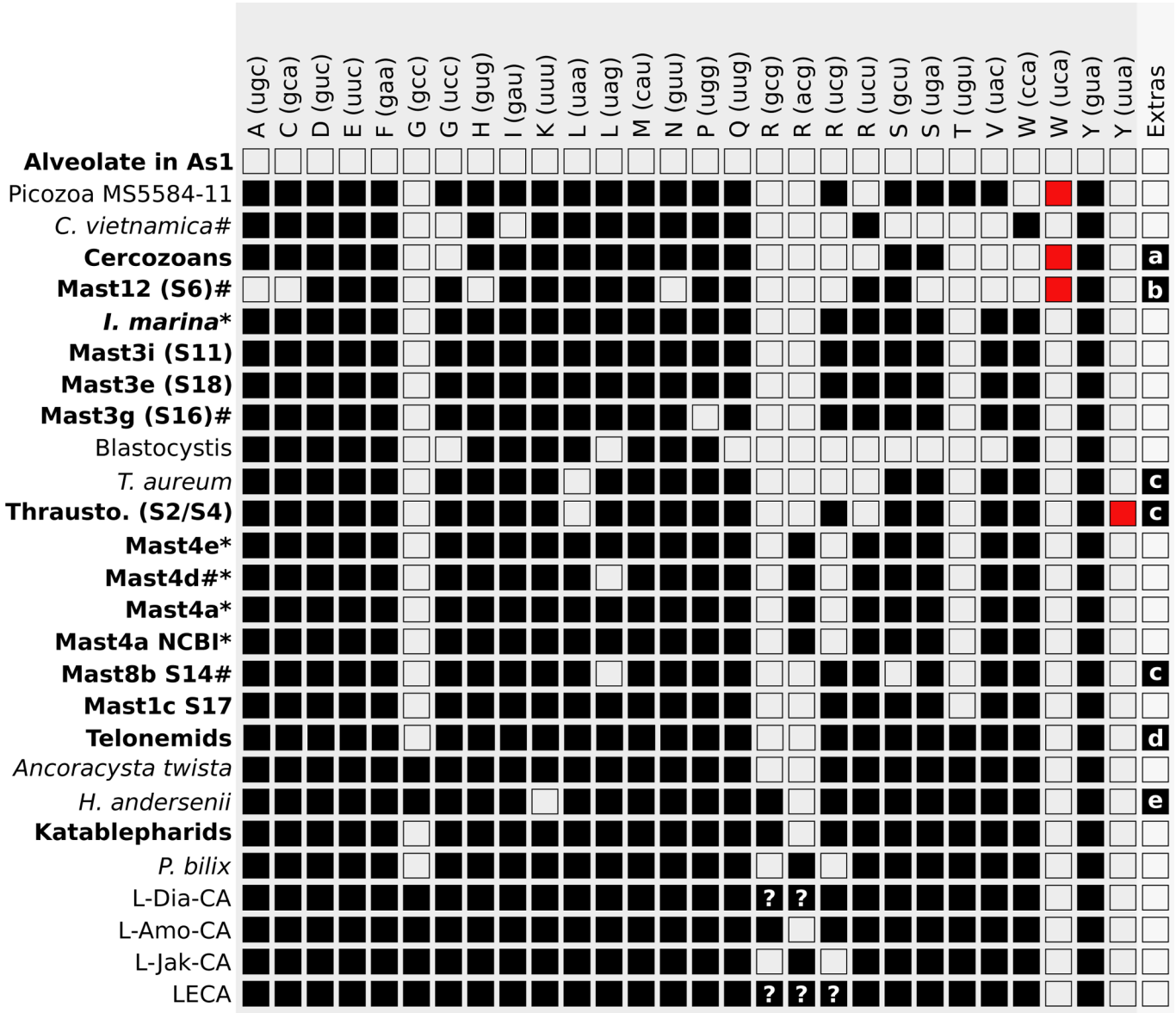
Extended Data Fig. 2 | Cox1 protein phylogeny. Cox1 proteins were collected from representative eukaryote groups using BLAST[26], aligned using MUSCLE[81], and manually trimmed to a resulting 402 sites. We reconstructed the phylogeny of Cox1 using RAxML v8.2.10[84] (100 bootstrap pseudoreplicates) under the LG model[85]. Maximum likelihood support values are indicated above each branch. The Cox1 from As1 grouped within the myzozoan alveolates within a fully supported clade comprising dinoflagellates, apicomplexans, and 'chromerid' algae. Picozoan M5584–11 Cox1 does not branch strongly with any eukaryotic group. Numbers in brackets indicate number of sequences collapsed.



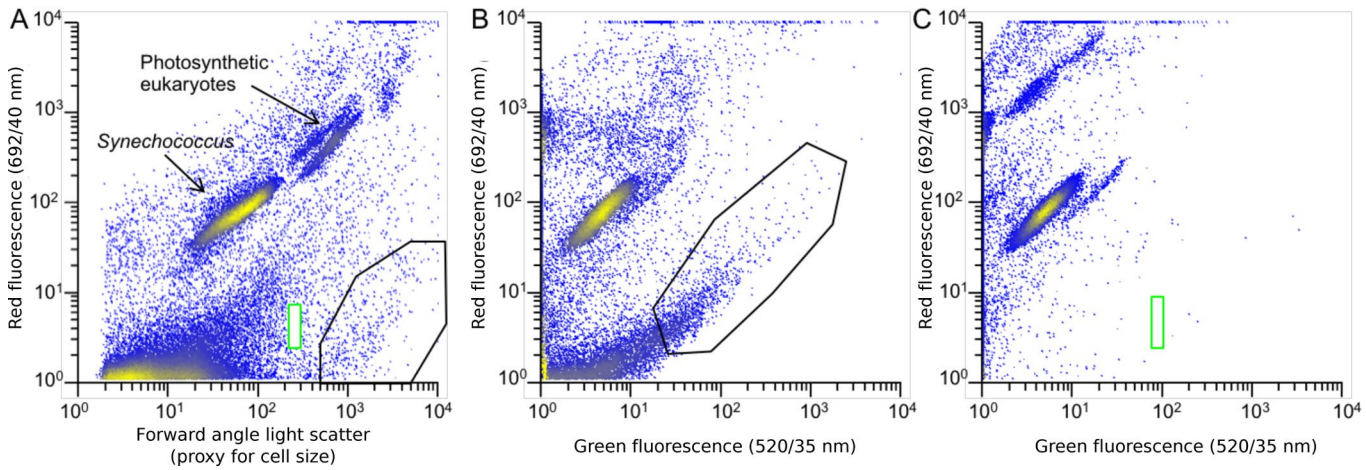
Extended Data Fig. 3 | Telonemid mtDNAs encode a putative *rpl18* and retain partial synteny with the bacterial-like genomes of jakobids. In all telonemid mitochondrial DNAs examined *rps8*, *rpl6*, and *rpl18* were found in synteny as in mtDNAs of jakobids. *Malawimonas jakobiformis* is somewhat similar as *rpl6* and *rpl18* are found adjacent to one another. Genbank: *Andalucia godoyi* NC_021124.1, *Histiona aroides* NC_021125.1, *Jakoba bahamiensis* NC_021126.1, *Jakoba libera* NC_021127.1, *Reclinomonas americana* NC_001823.1, *Seculamonas ecuadoriensis* NC_021128.1, *Malawimonas jakobiformis* NC_002553.1. Small subunit ribosomal genes are coloured in pink, large subunit ribosomal genes in red, SecY in purple, and electron transport chain components in grey.



Extended Data Fig. 4 | Thraustochytrid mtDNAs harbour a unique genetic code. Alignment of mitochondria-encoded Cob proteins from *Thraustochytrium aureum*, *Schizochytrium* sp., and four putative thraustochytrid SAGs. *Cob* genes with internal stop codons were identified in mitochondrial contigs from each SAG and translated using the standard genetic code. These proteins were aligned using MUSCLE[81] with proteins from publicly available thraustochytrid mtDNAs (KU183024.1 and AF288091.2). Positions occupied by TAG or TAA codons are marked with yellow asterisks and aligned most often with tyrosine or other hydrophobic residues (marked in orange). Relatively few TAA and TAG codons were conserved between genome sequences suggesting that these changes occurred during the recent radiation of this lineage.



Extended Data Fig. 5 | Distribution of mitochondria-encoded tRNAs. Comparison of mtDNA tRNA coding capacities from: new assemblies from this study (bold font), previously sequenced mtDNAs (regular font), and ancestral reconstructions (L-Dia- CA, Last Diaphoretickes Common Ancestor; L-Amo-CA, Last Amorphean Common Ancestor - including malawimonads and collodictyonids)); L-Jak-CA, Last Jakobid Common Ancestor; LECA, Last Eukaryote Common Ancestor. # symbols indicate incomplete mtDNA. Asterisks indicate genomes assembled from publicly available datasets. Black filled square, present; empty square, absent. Red filled squares indicate an independent codon reassignment. In some lineages extra tRNAs are also present other than the common tRNAs presented: a, I (uau), one cercozoan lineage (R32) contained a possible suppressor tRNA (gcaa); b, I (uau); c, L (caa); d, I (aau); e, L (gag), N (auu).



Extended Data Fig. 6 | Gating strategy for cell sort 35 from which most SAGs originated. A combination of gates (black polygons) was applied to select. **a**. cells larger than Synechococcus displaying low red fluorescence to exclude photosynthetic eukaryotes and **b**. cells stained with Oregon Green as compared to **c**. an unstained sample. The green rectangles show the position of 0.75 μm yellow-green beads.

nature research

Corresponding author(s): Jeremy Wideman
Thomas Richards

Last updated by author(s): Sep 30, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The bioinformatic workflow is available here: DOI: 10.5281/zenodo.192677
Data analysis	The bioinformatic workflow is available here: DOI: 10.5281/zenodo.192677. Programs used included: BD FACS 'Sortware' sorter software v1.0.0.650, PEAR 0.9.8, Trim Galore! www.bioinformatics.babraham.ac.uk/projects/trim_galore/, SPAdes 3.7.1, QUAST 5.0.2, CEGMA v2, Blobtools v1.0, Qualimap v2.2.1, BWA 0.7.17, SAMTOOLS 1.9, BAMTOOLS 2.4.0, PyNAST v1.2, Trimal v1.4, RAxML v8.2 , FastTree v2.1 , DADA2 v1.4, Phyloseq v1.20, mfannot http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl, HMMER v3.1 , ggplot2 v2.2.1, DescTools v0.99.23, MUSCLE https://www.ebi.ac.uk/Tools/msa/muscle/, MrBayes v3.2.6

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequencing data can be found under NCBI BioProject: PRJNA379597. Reads are deposited at NCBI SRA: SRP102236 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA379597). Complete mitochondrial genomes have been deposited in NCBI: Accessions will be available upon publication. For review follow the private link: https://figshare.com/s/e80871f5eb07bbd984de DOI: 10.6084/m9.figshare.7314734. Partial mitochondrial genome contigs and other important contigs mentioned in text are available at DOI: 10.6084/m9.figshare.7314728 private for review: https://figshare.com/s/78f0c96767b5c554a163. Nuclear SAG assemblies are available at DOI: 10.6084/m9.figshare.7352966 private link for review: https://figshare.com/s/e37b32acbaae40502459. Bioinformatic workflow is published at DOI: 10.5281/zenodo.192677.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Single cell heterotrophic flagellates were sorted by flow cytometry. 206 sorted single cells were subjected to whole genome amplification. After screening 99 of these were chosen for genome sequencing and analysis.
Research sample	206 single cells sorted from seawater by flow cytometry.
Sampling strategy	Seawater was collected in Monterey Bay at 36.6893°N; 122.384°W (Monterey Bay Aquarium Research Institute timeseries station M2, 56 km from shore) on 7 October 2014 using a Niskin rosette. Water was collected at 30 m (sub-surface chlorophyll maximum as determined by in vivo chlorophyll fluorescence).
Data collection	For Illumina library preparation an aliquot of each chosen MDA sample (including 6 negative controls) was purified with AMPureXP magnetic beads (Beckmann) following the manufacturer's instructions, quantified with a Qubit and diluted in 10mM TrisCl (pH 8.0) to a final volume of 130 uL and a concentration of 7.7 ng/uL. DNA was fragmented using focused acoustic waves (Covaris E220), concentrated, and libraries made with Nextflex Rapid DNA library preparation kit and indexes (BIOO Scientific) without PCR amplification. For a subset of samples, 3 uL of each was pooled and concentrated for 450-650 bp size selection using a Blue Pippin 1.5% agarose cassette with R2 marker. The average size of the recovered libraries was 420 bp (with 295 bp inserts). For a second subset, libraries were prepared similarly but used bead-based size selection (420-620 bp), rather than Blue Pippin, quantified by qPCR and equimolar pooled at 2 nM. Library pools were denatured, diluted and 250 paired-end sequenced across two lanes on a HiSeq 2500 using Rapid Run SBS v2 reagents (Illumina).
Timing and spatial scale	7 October 2014 samples were collected. This was an exploratory study.
Data exclusions	No data were excluded from the analysis.
Reproducibility	The major hypothesis tested in this project is that mitochondrial genomes are reproducibly recovered from single cell genomic data. That we recovered mitochondrial contigs from 70% of our samples suggests that this study is reproducible. Furthermore, we were able to recover mitochondrial genomes from previously published single-cell studies confirming that our results are reproducible.
Randomization	This wasn't relevant to our study as it is a single-cell genomics exploratory investigation. No samples to randomize.
Blinding	Blinding was not necessary for this study as it is a genomic investigation into unknown diversity.

Did the study involve field work? ☒ Yes ☐ No

Field work, collection and transport

Field conditions	The field conditions on the day are not relevant to our analysis.
Location	Seawater was collected in Monterey Bay at 36.6893°N; 122.384°W (Monterey Bay Aquarium Research Institute timeseries station M2, 56 km from shore) on 7 October 2014 using a Niskin rosette. Water was collected at 30 m (sub-surface chlorophyll maximum as determined by in vivo chlorophyll fluorescence), pre-filtered through a 30 um mesh, then concentrated by gravity ~70-100 times onto a 0.8 um filter
Access and import/export	No sampling permits were required.
Disturbance	No disturbance occurred.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Seawater was collected in Monterey Bay at 36.6893°N; 122.384°W (Monterey Bay Aquarium Research Institute timeseries station M2, 56 km from shore) on 7 October 2014 using a Niskin rosette. Water was collected at 30 m (sub-surface chlorophyll maximum as determined by in vivo chlorophyll fluorescence), pre-filtered through a 30 um mesh, then concentrated by gravity ~70-100 times onto a 0.8 um filter and stained with Paclitaxel, Oregon Green® 488 Conjugate (ThermoFisher, 100 ug/mL stock made in DMSO) at 10 uM (targeting tubulin from cytoskeleton). Cells were washed twice with sterile artificial sea water to remove unbound dye, then stained with Hoechst 33342 (targeting DNA) at 2 ug/ml. Stained samples were diluted into sterile artificial sea water in preparation for flow cytometry.
Instrument	BD InFlux mounted with 488 and 355 nm lasers. Sheath Fluid: sterile nuclease-free PBS pH 7.4 as sheath fluid (ThermoFisher cat# AM9625).
Software	BD FACS(TM) Software v 1.2.0.142 (run software); Verity Software House WinList 9.0 (figure display software)
Cell population abundance	There is no population abundance analysis in this manuscript. The sorting was used to separate cells into individual wells that were then sequenced (as described in methods) and no quantitative information is discussed or implied. The pre-concentration methods used preclude derivation of numerical information.
Gating strategy	A combination of sort windows was applied to select the cells that showed green and blue fluorescence (captured by a 520/35nm and a 460/50nm bandpass filter for Oregon Green [tubulin] and Hoechst 33342 [Blue-DNA], respectively) as compared to unstained control samples, and baseline red fluorescence (692/40nm bandpass filter) indicating the absence of chlorophyll, allowing exclusion of photosynthetic cells. Eighteen SAGs with recovered mitochondrial genomes were obtained following this strategy and originated from sort 34 and sort 36 (Table S2). A majority of SAGs (52) were recovered from sort 35 where cells were targeted based on Oregon Green fluorescence only and regardless of Hoechst fluorescence, however sort windows were refined using the forward angle light scatter (used as a proxy for cell size) to select cells larger than cyanobacterial cells present in the sample (i.e., Synechococcus, recognizable by the orange fluorescence of the phycoerythrin present in the cells detected in a 572/27 nm bandpass filter).

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.