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

Short title: Single-cell mito-genomics of heterotrophic flagellates

Jeremy G. Wideman[a,b,c,d,1,*], Adam Monier[a,1], Raquel Rodríguez-Martínez[a,e,1], Guy Leonard[a], Emily Cook[a], Camille Poirier[f,g], Finlay Maguire[a,h], David Milner[a], Nicholas A. T. Irwin[i], Karen Moore[a], Alyson E. Santoro[j], Patrick J. Keeling[i], Alexandra Z. Worden[f,g], and Thomas A. Richards[a,*]

[a]Living Systems Institute, University of Exeter, Stocker Road, Exeter EX4 4QD, United Kingdom.

[b]Wissenschaftskolleg zu Berlin, Wallotstrasse 19, 14193, Berlin, Germany.

[c]Department of Biochemistry & Molecular Biology, Dalhousie University, Halifax, Nova Scotia, B3H 4R2 Canada.

[d]Center for Mechanisms of Evolution, Biodesign Institute, School of Life Sciences, Arizona State University, Tempe, Arizona, 85287 USA.

[e]Laboratorio de Complejidad Microbiana y Ecología Funcional, Instituto Antofagasta, Universidad de Antofagasta, Antofagasta, Chile.

[f]Monterey Bay Aquarium Research Institute, 7700 Sandholdt Road, Moss Landing, CA 95039, USA.

[g]Ocean EcoSystems Biology Unit, Division of Marine Ecology, GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany.

[h]Faculty of Computer Science, Dalhousie University, 1459 Lemarchant Street, Halifax, NS B3H 3P8, Canada.

[i]Department of Botany, University of British Columbia, 3529-6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada.

[j]Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, CA 93106, USA.

[1]These authors contributed equally to this work.

*Corresponding authors: Jeremy G. Wideman, Center for Mechanisms of Evolution, Biodesign Institute, School of Life Sciences, Arizona State University, Tempe, Arizona, 85287 USA. E-mail: Jeremy.Wideman@asu.edu and Thomas A. Richards, Living Systems Institute, University of Exeter, Stocker Road, Exeter EX4 4QD, United Kingdom. E-mail:T.A.Richards@exeter.ac.uk

Abstract:

Most eukaryotic microbial diversity is uncultivated, under-studied, and lacks nuclear genome data. Mitochondrial genome sampling is more comprehensive, yet many phylogenetically important groups remain unsampled. 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 (SAGs) predominantly from under-represented branches on the tree of life. Seventy SAGs 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 (MASTs), more than doubling the available sampling of heterotrophic flagellate mitochondrial genomes. Collectively, these data identify a dynamic history of mitochondrial genome evolution including intron gain/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 only observed in plants.

Mitochondria originate from an alphaproteobacteria-like endosymbiont[1], often contain their own genomes, and make ATP via oxidative phosphorylation. Most of the 900-1100 different mitochondrial proteins are nuclear encoded[2]. The progenitor endosymbiont encoded many more genes than extant mitochondrial genomes (mtDNA), with numerous genes lost or transferred to the nucleus[3]. Mitochondria-encoded genes vary, but include those essential for mitochondrial transcription, 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 analyzed 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 latter, a few have attempted genome sequencing and assembly[11–15] while others have analyzed SSU-rRNA 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 SYBRGreen in preserved cells[15,17] combined with chlorophyll exclusion to enrich for putatively phagotrophic cells. Where genome sequencing has been attempted, insight into the genome sequences of a few eukaryotes has been provided; however, the highly fragmented incomplete nature of single amplified genomes (SAGs) has restricted their use for comparative genomics.

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

Results

Single-cell sampling of marine flagellates

Heterotrophic protists employ diverse lifestyles important for ecosystem function. Since heterotrophs are poorly sampled it is important that diverse methods are developed to recover diverse forms. Many feed by phagotrophy, employing 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 (e.g., obligate osmotrophs[19]), do not phagocytose; and furthermore, acidic vacuoles can be deployed for a diversity of alternative 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 (e.g., 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 sub-surface chlorophyll maximum (SCM, 30 m), 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 sub-cloning of the SSU-rDNA amplified-template. The Sanger chromatographs did not show evidence of mixed amplicons suggesting that the V9 sequences recovered 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 (e.g.[24]) demonstrating the efficacy of our approach (Figs. 1 and 2 and Supplementary Table 1). Six grouped with taxa containing photosynthetic/heterotrophic forms (e.g., haptophytes and ochrophytes), and eleven were derived from non-flagellated fungi previously sampled from marine environments[25] (Fig. 2, panel 1).

A recent TARA Oceans-related project presented a broad diversity of heterotrophic nSSUs from sorted cryopreserved SAGs using SYBR green and chlorophyll exclusion and enriched for different taxa compared to our analysis[17]. The majority of TARA heterotrophic flagellates recovered were MASTs (362, 71%); whereas, while our protocol recovered some MASTs (12, 6%), we predominantly recovered cercozoans (53, 26%), Marine ALVeolates (MALVs) and dinoflagellates (51, 25%), choanoflagellates (22, 11%), telonemids (13, 6%) and euglenozoans (20, 10%). Although from different geographic sites, the differences in taxa recovered highlights the importance of developing approaches that target specific 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 SAGs, were among the rarer taxa identified (Extended Data 1 and Supplementary Table 2). This was expected as the vast majority of eukaryotes at the SCM are photosynthetic. These data also show 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. We conclude that our sorting method was effective in targeting heterotrophic flagellates while excluding phototrophs and non-target cells/particles and can be applied to various environments (e.g. freshwater and potentially, with modification, in soils).

Genome sequencing of single-cell samples

Based on the phylogenetic affiliation of the 206 SAGs, we chose 99 cells from under-sampled lineages for DNA sequencing (Fig. 2). We generated 204 Gbps with a mean (median) sequencing depth of ~1.61 (1.35) Gbp/SAG. The resulting reads were assembled generating a mean assembly size of 14.5 Mbp (SD = 13.8) and N50 of 3.4 kbp (SD = 2.4)/SAG. Full length nSSU rRNAs recovered from these assemblies were used to confirm the V9 phylogenetic position of the SAGS sampled by BLAST[26] discussed above. In all cases only a single full-length eukaryotic SSU-type was recovered from each SAG, suggesting that co-sampling of multiple eukaryotic cells was minimal. After database curation, three of the nSSU-V9-types previously mapped to a tree were determined to be artefactual: As1 and As2, which mapped as ascomycete fungi (likely 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/median 11.4%/6.5%) (Tables S3-4), comparable to 2-45% recovery in other studies[11,13,14]. However, this approach to genome completion estimation is subject to a range of artefacts stemming from: i) sampling wells occupied by more than one cell, and ii) 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 (Supplemental Data S1 DOI: 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 > 50% predicted mitochondrial completion, the relative coverage of mtDNAs was higher and more variable (M = 17.0x SD = 17.2) compared to the SAG assemblies (M = 4.9x SD = 2.5) (Supplementary Table 5) consistent with their derivation from organellar genomes that are often present in higher copy numbers than nuclear genomes. Interestingly, 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/intermediate nuclear completion and mitochondrial coverage (Hotelling's T2-test[27] p = 9.07e-13), but no SAGs with both high nuclear and mtDNA recovery (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 in mitochondrial contigs, all suggest that we have sequenced bona fide mitochondrial genomes and not mitochondrial insertions into nuclear genomes.

A total of 10 unique, complete circular-mapping mtDNAs were assembled from individual SAGs. These include: two telonemids (T1 NCBI Accession: MK188946, T12 MN082145), a katablepharid (K4 MK188945), an unknown alveolate (see below As1 MK188935), two MAST3s (S11 MK188941, S18 MK188943), a MAST1 (S17 MK188942), a haptophyte (H2 MK188944), and two choanoflagellates (C14 MK188937, C15 MK188938) (Fig. 4- bold). Two cercozoan mtDNAs were assembled, judged linear, and likely to be complete based on protein repertoires (R17 MK188936, R32 MN082144, bold in 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[14,28], we assembled three additional complete mtDNAs: *Incisomonas marina*, a MAST4a, and a MAST4e (Fig. 4 asterisks). Additionally, we identified a likely complete MAST4a mtDNA (EU795181.1) misannotated 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 branches.

To confirm that the mtDNAs belong to the expected taxa, we used our complete and near-complete mitochondrial assemblies as BLAST queries into the NCBI non-redundant database (Supplementary Table 6). The choanoflagellate (C14, C15), katablepharid (K4), and haptophyte (H2) mtDNAs hit 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, 4, 6, 11, 14, 16, 18), 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 retrieving 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 multi-gene phylogeny using stramenopile and cercozoan mtDNAs (Fig. 6). Our cercozoans bifurcated with *Bigelowiella* and *Paracercomonas* and not stramenopiles with full support, confirming their likely-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% identital to the nSSU from picozoan MS5584-11[11] and a single circular mtDNA. The mtDNA encodes no tRNAs and only five putative genes including barely identifiable, fragmented, mitochondrial small and large ribosomal RNA genes, *cob, cox1*, and an unidentified open reading frame, but based on the predicted transmembrane architecture of the protein, is likely a divergent *cox3*[29] (Fig. 4). This repertoire is the same as myzozoan alveolates differing considerably from the picozoan MS5584-11 mtDNA[11,30]. Consistent with the BLAST results reported above, phylogenetic reconstruction using Cox1 demonstrated that the As1-derived protein branches within myzozoans (Extended Data 2). In contrast, the MS5584-11 Cox1 protein did not branch strongly with 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 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 allowed us to sample a wide diversity of eukaryotic lineages and compare repertoires of mitochondrial genes (Figs. 4-5). 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 3), *rpl31*, *rpl32*, and *tatC* thought to be rare. Whereas the katablepharids contain a single discontinuously distributed gene (*nad*8). 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, but some 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, *Incisomonas marina*, MAST4, and MAST8 mtDNAs but are absent in closely related lineages (MAST3i and MAST3e). Prior studies have noted *tatC* in labyrinthulomycete mtDNAs (KU183024.1 and AF288091.2[31,32]), which is absent in our thraustochytrid-related cells (S2 MK188939, S4 MK188940). We also identified the RNA component (*rnpB*) of RNase P encoded by MAST3e and MAST4e, *rps1* by MAST1c, and *rpl31* 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[33]. 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 choanoflagellate mtDNAs recovered have 97% identity but contain a different number of introns in the *cox1* gene (C14 = 4, C15 = 2, *M. brevicollis* = 3) (Fig. 4). The two encoded homing endonucleases in C15 are similar to two in C14 (89% and 98% amino acid identity), but none are similar to those in *M. brevicollis cox1* (AF538053.1), suggesting a complex pattern of replacement or rapid intron diversification[34].

Similarly, in the cercozoan mtDNAs, while 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 (e.g., R17 and R32 with 97% nSSU rRNA nucleotide identity, Fig. 4), the differences between the number of introns and the different positions of the introns (e.g., R32 has

4 large introns in *cox1* whereas R17 has no introns in *cox1*) suggests that most of the introns have been acquired recently or the genomes sampled have undergone repeated invasion by related introns coupled with differential loss of intron variants (e.g., [34]).

While the *P. bilix* and some cryptophyte mtDNAs contain no, or very few introns[35–37], the katablepharid K4 mtDNA contains seven introns (dark grey in Fig. 4). The published *L. marina* partial mtDNA sequence contains homing endonuclease-encoding group I introns in the *cob* and *cox1* genes at identical locations as introns identified in the katablepharid mtDNAs sampled here (49% and 73% amino acid identity, respectively). Our data confirm that multiple mitochondrial evolutionary lineages undergo a high turnover of self-splicing introns, whereas other lineages appear 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 4442-site concatenated protein phylogeny using members of cercozoans as an outgroup (Fig. 6). The phylogeny recovered previously established phylogenetic groups including Ochrophyta, Labyrinthulomycota, and Pseudofungi[28,38]. Similar to other mitochondrial phylogenies[39], and in contradiction to phylogenies based on nuclear proteins, we could not recover Ochrophyta-Pseudofungi sisterhood[28,40] suggesting there is either conflicting phylogenetic signal in mtDNA compared to nuclear markers, or some systematic phylogenetic artefact is present, discussed previously[39]. Our phylogeny recovered some support for the placement of MAST clades previously proposed from nSSU rRNA phylogenies[41] and partially corroborated in a recent multi-gene phylogeny of nuclear encoded genes[28]. 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, unexpectedly MAST1c, and labyrinthulomycetes (Fig. 6). Given previous evidence of contradictory relationships identified in stramenopile mitochondrial and nuclear gene phylogenies[39], the branching order presented here should be treated with caution. As such, additional sampling of stramenopile lineages is required to understand the conflict observed between mitochondrial and nuclear phylogenies.

Using the mitochondrial phylogeny, we sought to polarise mitochondrial traits onto the stramenopile tree. So far, recent and frequent functional mitochondria to nuclear gene transfers have been reported only in Archaeplastida[42] (i.e., green plants). Identification of closely related lineages

containing different mitochondrial genes (i.e. MAST4s, MAST1, and MAST8) suggests that genes have relatively recently been transferred 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[40,43]). 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: 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 like animals and fungi and resemble more closely the dynamic repertoires in the plant lineage[42]. 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 (e.g., Complex I components *nad7-11*) are more readily transferred to the nucleus than core (defined as energetically central) components (e.g., Complex I components *nad1-6*)[44]. 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[30].

Lastly, our stramenopile and cercozoan mtDNAs allowed us to trace the evolutionary history of three genetic code changes. Several mitochondrial code changes have been documented[45], the most common being TGA recoded 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 (e.g., see[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 strikingly, we identified a genetic code present in our thraustochytrid mtDNAs (two near-complete and three fragmented, S2, 4 and S1, 3, 15, respectively). In these mtDNAs, TGA and TTA (normally encodes

leucine) serve as the only termination codons, and TAG and TAA (normally termination codons) have been recoded to tyrosine (Extended Data 4). This finding is supported by the identification of a UUA anticodon tRNA encoded in the SAG mtDNAs (Extended Data 5). It is known that TTA was recoded as a stop codon in *Thraustochytrium aureum* (AF288091.2)[31] thus 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

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 allow for higher resolution studies of protist population structures and for effective sampling of multiple genes with different rates of sequence variation useful for phylogenetic analyses. The data reported here have substantially increased publicly available heterotrophic flagellate mtDNAs. NCBI reports 9520 complete mtDNAs, 8685 from animals, 406 from photosynthetic algae and plants, 334 from fungi, and 50 from animal/plant parasites (apicomplexans and oomycetes). Of the remaining 44 genomes of heterotrophic protists, only 17 are heterotrophic flagellates spread across the eukaryote tree of life. Our data more than doubles this representation, adding complete or near-complete genomes from 5 un- or under-represented groups (Telonemida (0 + 2), Katablepharida (~30% of 1 genome + 1), heterotrophic flagellated stramenopiles (2 + 11), Rhizaria (6 + 5) and Choanozoa (1 + 2)). 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 timeseries station M2, 56 km from shore) on 7 October 2014 using a Niskin rosette. Water was collected at 20 m and 30 m (sub-surface chlorophyll maximum as determined by *in vivo* chlorophyll fluorescence). For general community diversity analyses 500 mL of water was filtered on to a 0.2 µm pore size Supor filter (Pall cat# 60301) and extracted using a

modification of the DNeasy kit (Qiagen) including the addition of a mechanical lysis by bead-beating[46]. 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 ug/mL stock made in DMSO) at 10 µM (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 µg/ml. Stained samples were diluted into sterile artificial sea water in preparation for flow cytometry.

Cell sorting of marine heterotrophic flagellates

Cells were analyzed and sorted on a BD Influx flow cytometer equipped with a 488 nm and a 355 nm laser and using sterile nuclease-free PBS pH 7.4 as sheath fluid (ThermoFisher cat# AM9625). A combination of sort windows was applied to select the 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 [Blue-DNA], respectively) as compared to unstained control samples, and baseline red fluorescence (692/40 nm bandpass filter) indicating the absence of chlorophyll, allowing us to exclude the majority of photosynthetic cells (See Extended Data 6). Eighteen SAGs with recovered mitochondrial genomes were obtained following this strategy and originated from sort 34 and sort 36 (Supplementary Table 2). 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).

Targeted cells were sorted into 96-well plates so that all wells received one individual cell (Single-Cell sorting mode implemented in the BD FACS 'Sortware' sorter software v1.0.0.650), except for the outer column of wells which were left empty for negative controls. Duplicate plates were obtained for sort 34 and 36 and triplicate plates for sort 35. The plates were illuminated by UV radiation inside the sort chamber for 2 min prior to the sort, covered with foil and placed at -80°C immediately after the sort. The sort quality and correct drop delay was regularly checked

by sorting a known number of polystyrene beads (Polysciences, cat# 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 manufacturer's instructions for amplification of genomic DNA from single cells. After neutralization, samples were amplified using the Repli-g reagents for 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[47]. PCR products were run on 1% agarose gel 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 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 µL and a concentration of 7.7 ng/µL. 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 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 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). Nine repeated samples which were sequenced more deeply on an additional HiSeq 2500 lane in order 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 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 qPCR before 125 bp PE Illumina sequencing.

Single-cell genomic assembly

All SAG sample libraries were assembled using the automatic workflow available at https://zenodo.org/record/192677 (DOI: 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[48] in order 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 3.7.1[49] using the single-cell mode, the careful option and with a combination of k-mers (21, 33, 55). Quality assessment of the resulting scaffolds was computed with the analysis software QUAST[50] and completeness profiles were made using CEGMA[51]. A set of "blobtools" charts are also made with a combination of scaffolds, read mapping and megaBLAST hits to the NCBI 'nt' database[52]. Additional analyses including KRONA taxonomy charts and QUALIMAP reports of mapping were computed from these data[53].

Universal nSSU tree, V9 mapping, taxon identification from rDNA assemblies

For the SAG taxonomic classification, nSSU V9 sequences (primers 1389F-1510R[54], a single sequence from each sample) corresponding to each of our 206 SAG 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 v4.4[55]; 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/GenBank taxonomic data) were also discarded. To remove sequence redundancy, the PR2 database was then clustered using CD-HIT v4.6[56] 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 v1.2[57] using the nSSU seed alignment from Silva release v123[58] as a template alignment. The resulting alignment was then edited and trimmed using Trimal v1.4[59] to remove sites with gaps in more than 25% of the sequences, but conserving at least half of the original alignment (i.e., -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 (ML) tree was reconstructed using RAxML v8.2 (multithreaded version; PTHREADS-SSE3)[60] under the GTR model with CAT approximation. SAG V9 sequences were mapped onto the PR2 reference ML tree using the RAxML Evolutionary Placement Algorithm (EPA[61]) under GTR-CAT. To evaluate local node supports, a Shimodaira-Hasegawa (SH)-like test[62] was run using FastTree v2.1 (double precision build[63] 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 (i.e., 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[64]; 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[65].

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, 86 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 over-representation of these sequences in our SAG samples is a plausible source of the apparent technical contamination (i.e., 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 v1.14[66]. 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[67]; reads with no significant hit (E-value < 1e-5) against the reference database were

discarded. Reads were then processed using DADA2 v1.4[68]. Based on quality profiles, forward reads were truncated at 150 bp, reverse reads 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; i.e., 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 RDP naïve Bayesian classifier[69], as implemented in DADA2, and using PR2 as a reference database. ASVs classified as Bacteria, Archaea, Organelle, Metazoa or with no eukaryotic supergroup classification (i.e., classified only as "Eukaryota") were discarded. The final Monterey Bay V9 census dataset was comprised of 1,073 ASVs representing a total of 89,376 quality controlled, merged sequences (Supplementary Table 6). Comparison 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[70]. Subsequent V9 analyses were conducted using the R package Phyloseq v1.20[71].

Mitochondrial genome contig identification, re-assembly, 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 >90% identical to known bacterial, chloroplast, or contaminating (e.g. 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[72], SAMTOOLS[73], and BAMTOOLS[74]) and reassembled with SPAdes 3.7.1[49] 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[75] searches with HMMER v3.1 using Hidden Markov Models generated from alignments of known fragments (evalue < $1e^{-5}$)[76]. Complete or near-complete contigs (see below) were used as queries to identify shorter (i.e. 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 or nearly no coding sequence is missing based on comparisons with closely related taxa).

Samples T1, T12, K4, As1 (mapped near ascomycetes but top BLAST is picozoan), H2, C14, C15, S11, S17, S18, as well as *Incisomonas marina* and two MAST4 mitochondrial genomes from previous studies 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 likely-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 near 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, based on synteny, the missing stretches of DNA could be inferred since the sequences lacking 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 2 and 3 contigs respectively. S14 appears complete with respect to coding content, although the contigs could not be joined. S6 was incomplete, but when compared to the coding content of its closest sequenced relative *Cafeteria 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[77] 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 to 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). 51x coverage was the maximum coverage in the output of this tool. The relationship between relative mitochondrial genome coverage was compared with that of the nuclear coverage (as estimated by CEGMA%) using the 'ggplot2' (v2.2.1)[78] and 'DescTools' (v0.99.23)[79] packages in the R (v3.4.3) programming language[80]. A two-sided Hotelling's T2 test[27] (df1=2, df2=30, T.2=44.942, p=9.07e-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 assumptions 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 (S1-4 and 15) use TTA as a stop codon and contain in-frame TAG and TAA codons that align with conserved tyrosine residues when compared to homologues in other thraustochytrids (Extended Data 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[81] with publicly available *cob* genes from thraustochytrid mitochondrial genomes (KU183024.1 and AF288091.2) (Extended Data 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 5). Since *Thraustochytrium 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, 2, 3, 4, 4L, 5, 6, 7, 9, Cob, Cox1-3, and Atp6, 8, 9. After alignment and manual trimming using Mesquite v2.75, this resulted in a concatenated alignment with 4442 sites. IQ-Tree[82] was used for model testing resulting in LG as the highest scoring model by BIC. Phylogenetic tree reconstructions were performed using MrBayes v3.2.6 for Bayesian analysis[83]. 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 v8.2.10[84] under the LG model[85].

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[26]. Resulting sequences were aligned using MUSCLE[81], and manually trimmed to a resulting 402 sites. A phylogenetic reconstruction was conducted using RAxML v8.2.10[84] (100 bootstrap pseudoreplicates) under the LG model[85].

Data Availability:

Complete mtDNAs assembled from this study are found at NCBI under the accessions MK188935-47, MN082144-5. Sequencing data can be found under NCBI BioProject: PRJNA379597. Reads are deposited at NCBI SRA: SRP102236. Partial mtDNA contigs and other important contigs mentioned in text are available at DOI: 10.6084/m9.figshare.7314728. Nuclear SAG assemblies are available at DOI: 10.6084/m9.figshare.7352966. A public method can be accessed at: dx.doi.org/10.17504/protocols.io.ywpfxdn

Code availability:

Bioinformatic workflow is published at DOI: 10.5281/zenodo.192677. List of programs used in this study: 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.

Author contributions

JGW performed bioinformatic and phylogenetic analyses, and wrote the manuscript. RR-M performed molecular biological analyses. AM performed bioinformatic and phylogenetic analyses and GL performed bioinformatic analyses. EC and CP collected the samples and performed flow cytometry. FM performed statistical and bioinformatic analyses. DM performed molecular biological experiments and generated biochemical reagents. KM performed the genome sequencing. NATI analysed genomic data. TAR devised the project. JGW, AES, PJK, AZW, and TAR 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.

Corresponding authors

Jeremy G. Wideman: Jeremy.Wideman@asu.edu and Thomas A. Richards: T.A.Richards@exeter.ac.uk.

Acknowledgements

We would like to thank Franz Lang and Natacha Beck for annotation assistance and access to an unreleased version of mfannot, Dana Price for assistance with picozoan SAG data, and Cory Dunn for fruitful discussions and encouragement.

This project was supported by a Gordon and Betty Moore foundation grant (GBMF3307) to TAR, AES, AZW and PJK, and a Philip Leverhulme Award (PLP-2014-147) to TAR. Field sampling was supported by the David and Lucile Packard Foundation and GBMF3788 to AZW. TAR and AM are supported by Royal Society University Research Fellowships. JGW was supported by the European Molecular Biology

690 Organization Long-term Fellowship (ALTF 761-2014) co-funded by European Commission
691 (EMBOCOFUND2012, GA-2012-600394) support from Marie Curie Actions and a College for Life Sciences
692 Fellowship at the Wissenschaftskolleg zu Berlin. RRM is supported by CONICYT FONDECYT 11170748. FM
693 is supported by Genome Canada.

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 & Herrmann. 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**, 865–78 (1998).

5. Worden, A. Z. *et al.* Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science.* **347**, (2015).

6. Cuvelier, M. L. *et al.* Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14679–84 (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. Trans. B* (2019). doi:10.1098/rstb.2019.0090

23. Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci.* **99**, 5261–5266 (2002).

24. Richards, T. A. & Bass, D. Molecular screening of free-living microbial eukaryotes: diversity and distribution using a meta-analysis. *Curr. Opin. Microbiol.* **8**, 240–252 (2005).

25. Richards, T. A., Jones, M. D. M., Leonard, G. & Bass, D. Marine Fungi: Their Ecology and Molecular Diversity. (2011). doi:10.1146/annurev-marine-120710-100802

26. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–402 (1997).

27. Hotelling, H. The Generalization of Student's Ratio. *Ann. Math. Stat.* **2**, 360–378 (1931).

28. 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).

29. Flegontov, P. *et al.* Divergent mitochondrial respiratory chains in phototrophic relatives of apicomplexan parasites. *Mol. Biol. Evol.* **32**, 1115–1131 (2015).

30. Janouškovec, J. *et al.* A New Lineage of Eukaryotes Illuminates Early Mitochondrial Genome Reduction. *Curr. Biol.* **27**, 3717-3724.e5. (2017). doi:10.1016/j.cub.2017.10.051

31. Gray, M. W., Lang, B. F. & Burger, G. Mitochondria of Protists. *Annu. Rev. Genet.* **38**, 477–524 (2004).

32. Wang, Z. *et al.* Complete mitochondrial genome of a DHA-rich protist *Schizochytrium* sp. TIO1101. *Mitochondrial DNA Part B* **1**, 126–127 (2016).

33. Saldanha, R., Mohr, G., Belfort, M. & Lambowitz, A. M. Group I and group II introns. *FASEB J.* **7**, 15–24 (1993).

34. Goddard, M. R. & Burt, A. Recurrent invasion and extinction of a selfish gene. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 13880–5 (1999).

35. 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).

36. Kim, E. *et al.* Complete Sequence and Analysis of the Mitochondrial Genome of *Hemiselmis andersenii* CCMP644 (Cryptophyceae). *BMC Genomics* **9**, 215 (2008).

37. 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).

38. Riisberg, I. *et al.* Seven Gene Phylogeny of Heterokonts. *Protist* **160**, 191–204 (2009).

39. 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).

40. Leonard, G. *et al.* Comparative genomic analysis of the 'pseudofungus' *Hyphochytrium catenoides*. *Open Biol.* **8**, 170184 (2018).

41. 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).

42. 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).

43. Ševčíková, T. *et al.* A Comparative Analysis of Mitochondrial Genomes in Eustigmatophyte Algae. *Genome Biol. Evol.* **8**, 705–722 (2016).

44. Johnston, I. G. & Williams, B. P. Evolutionary Inference across Eukaryotes Identifies Specific Pressures Favoring Mitochondrial Gene Retention. *Cell Syst.* **2**, 101–111 (2016).

45. Keeling, P. J. Genomics: Evolution of the Genetic Code. *Curr. Biol.* **26**, R851–R853 (2016).

46. Demir-Hilton, E. *et al.* Global distribution patterns of distinct clades of the photosynthetic picoeukaryote Ostreococcus. *ISME J.* **5**, 1095–1107 (2011).

47. Logares, R. *et al.* Patterns of rare and abundant marine microbial eukaryotes. *Curr. Biol.* **24**, 813–21 (2014).

48. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).

49. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

50. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

51. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).

52. 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).

53. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, btv566 (2015).

54. 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).

55. 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**, D597-604 (2013).

56. 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).

57. Caporaso, J. G. *et al.* PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**, 266–267 (2010).

58. 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).

59. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–3 (2009).

60. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

61. 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).

62. Shimodaira, H. & Hasegawa, M. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol. Biol. Evol.* **16**, 1114–1116 (1999).

63. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* **5**, e9490 (2010).

64. Junier, T. & Zdobnov, E. M. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* **26**, 1669–1670 (2010).

65. 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).

66. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).

67. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

68. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).

69. 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. Microbiol.* **73**, 5261–7 (2007).

70. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–7 (2000).

71. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* **8**, e61217 (2013).

72. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

73. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

74. 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).

75. Wheeler, T. J. & Eddy, S. R. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**, 2487–2489 (2013).

76. Jackson, C. J. *et al.* Broad genomic and transcriptional analysis reveals a highly derived genome in dinoflagellate mitochondria. *BMC Biol.* **5**, 41 (2007).

77. Grant, J. R. & Stothard, P. The CGView Server: a comparative genomics tool for circular genomes. *Nucleic Acids Res.* **36**, W181–W184 (2008).

78. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2009).

79. Signorell, A. DescTools: Tools for descriptive statistics. R package version 0.99.23. (2017).

80. R Core Team. R: A language and environment for statistical computing. (2013). Available at: http://www.r-project.org/. (Accessed: 19th July 2018)

81. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

82. 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).

83. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–4 (2003).

84. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–90 (2006).

85. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–20 (2008).

Figure Legends

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

Fig. 2. Clade-specific ML subtrees. Six distinct eukaryotic clades from which numerous SAGs were recovered. Shown here are SAG nSSU V9 sequences mapped to a full-length reference tree that incorporated PR2 reference sequences. For each subtree, specific lineages that attracted SAG V9 sequences are highlighted in pink frames with the lineage name provided and in parentheses the number of SAG V9 that mapped onto the lineage. SAGs with mitochondrial contigs present are labelled: complete mtDNAs, white font on black circle; near complete, bold; partial genome, italics. SH-like local node supports are shown for > 0.9 (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.

Fig. 3. Distribution and groupings of mitochondrial sequence coverage relative to estimated nuclear genome completeness. Sequenced genomes showed either high nuclear completion (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

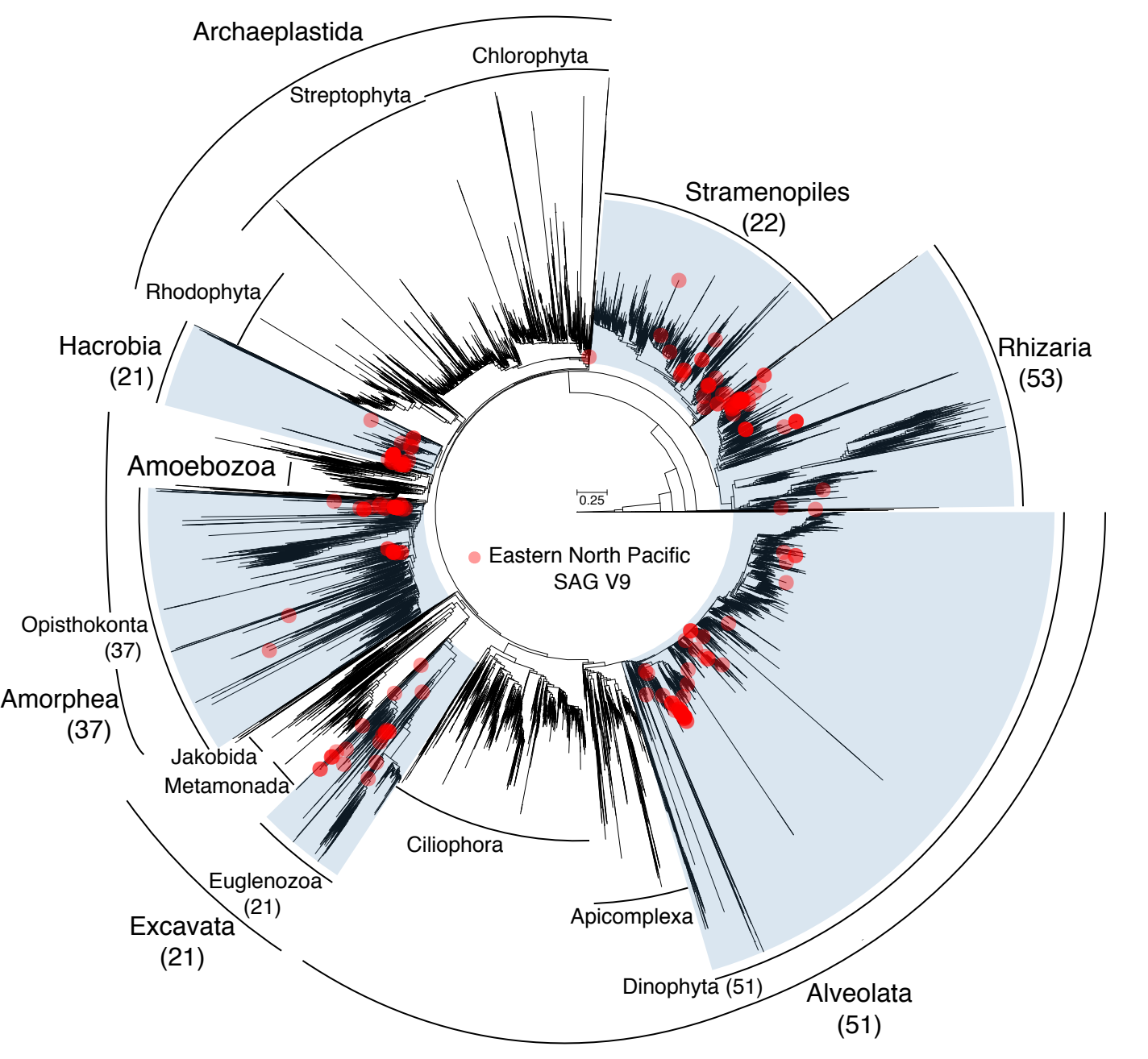
nuclear and mitochondrial coverage (blue). The blue density contours were plotted using 'ggplot2'. X-fold coverage of >80% of each sequenced mtDNA was calculated using BamQC in BAMTOOLS for each SAG with >50% estimated completion of the mtDNA. This score was plotted against the estimated CEGMA completion scores (%). The result of a Hotelling's T2-test to assess whether the SAGs with high mitochondrial coverage and those with high nuclear coverage supports rejection of the null hypothesis that these are sampled from the same population. The rejection of the null hypothesis suggests that the there is a fundamentaldifference between these two SAG sub populations.

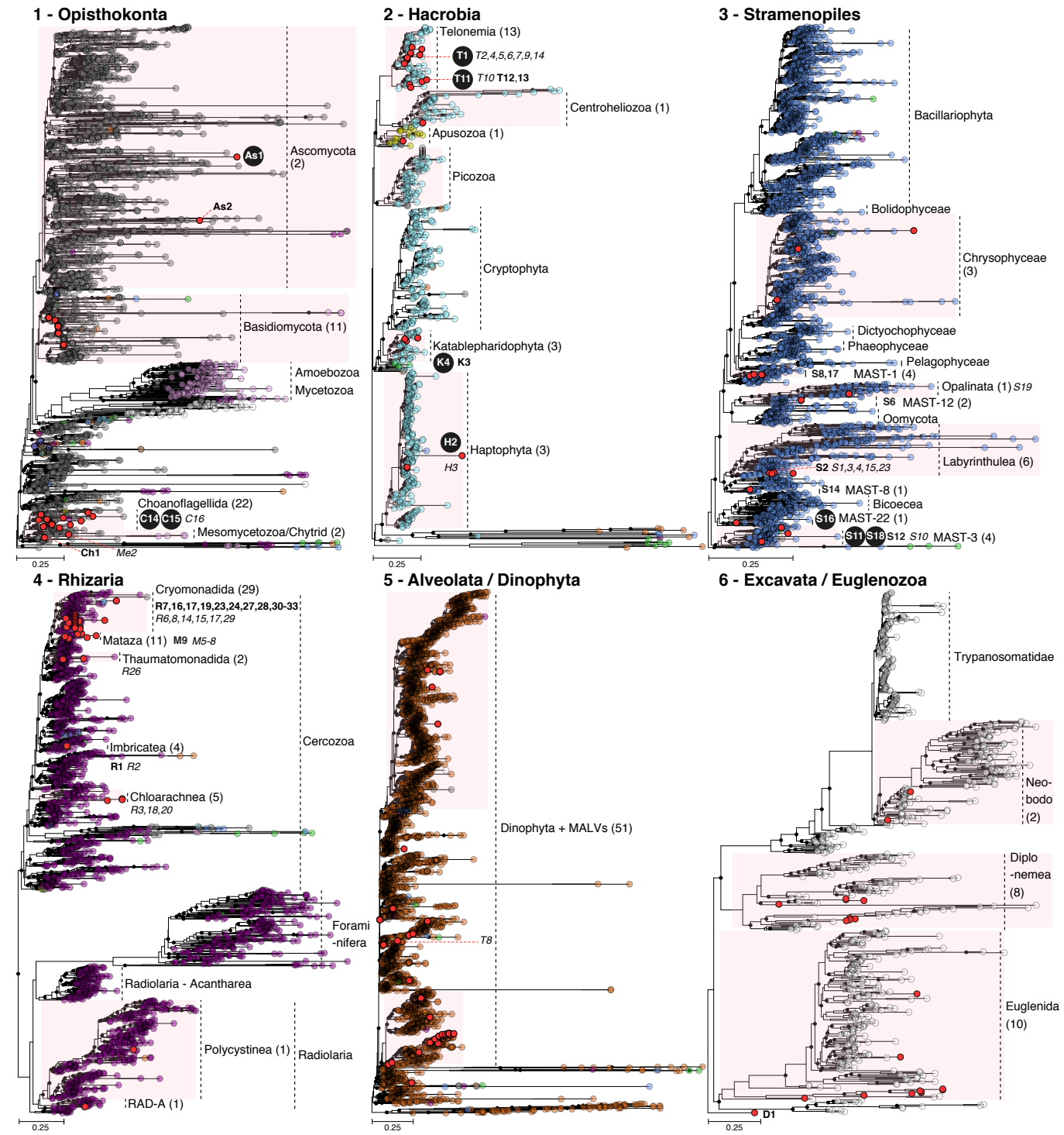
Fig. 4. Uncharacterized mtDNAs from underrepresented eukaryotic groups. Complete and near-complete mtDNAs assembled from heterotrophic marine flagellate SAGs. Mitochondrial contigs were annotated using mfannot with manual corrections as needed (http://megasun.bch.umontreal.ca/RNAweasel/). MtDNAs are represented as circular diagrams (bold central font) or broken circles if contigs could not be joined (regular central font). 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.

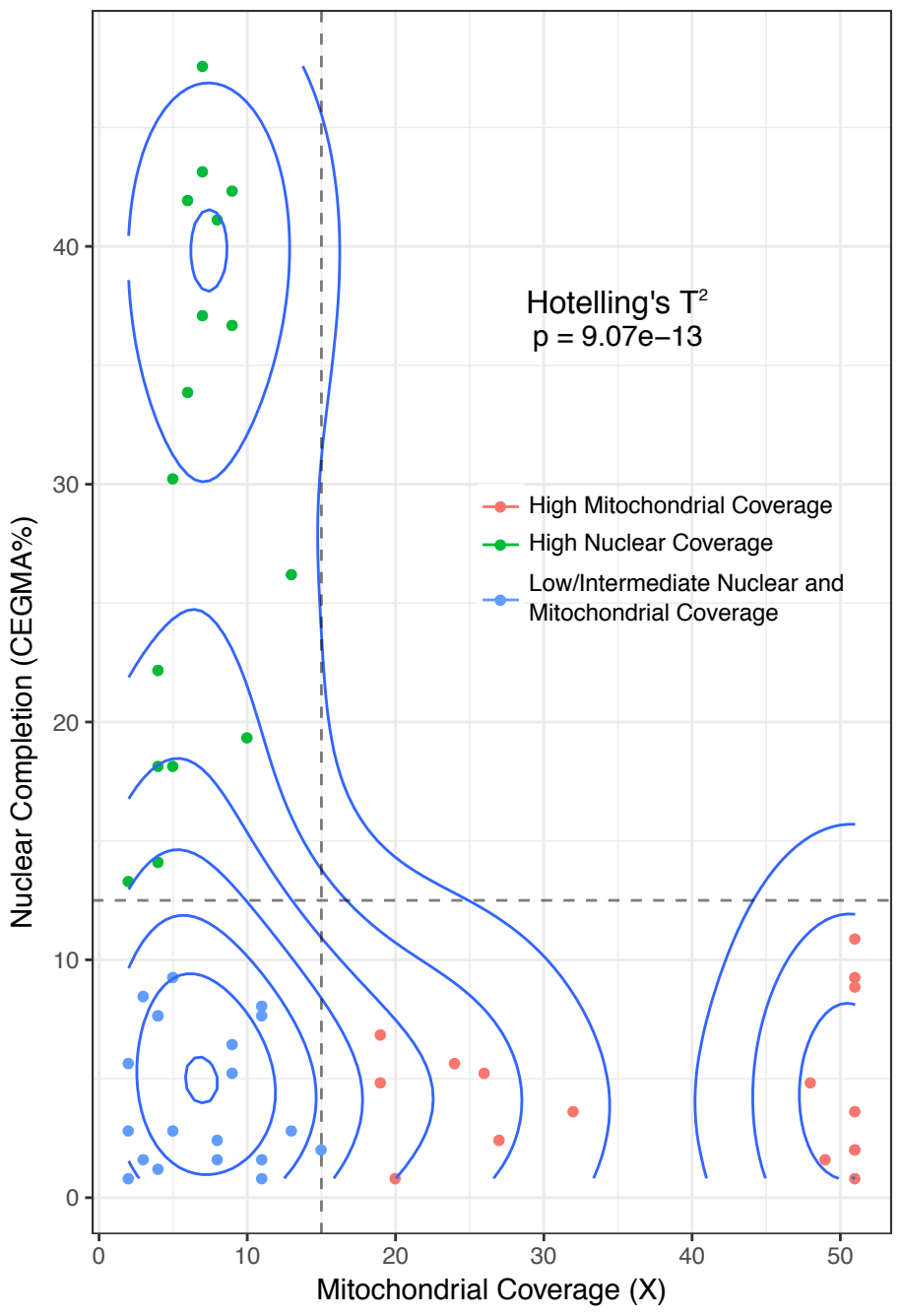
Fig. 5. Comparison between mtDNA gene repertoires. Mitochondrial genomes newly assembled in this study (bold font), previously sequenced mtDNAs (regular font), and ancestral reconstructions (L-Dia-CA, Last Diaphoretickes Common Ancestor; L-Amo-CA, Last Amorphean Common Ancestor - including malawimonads and collodictyonids); L-Jak-CA, Last Jakobid Common Ancestor; LECA, Last Eukaryote Common Ancestor) are shown. Black square, present; empty square, absent; red square, rare protein. # symbols indicate incomplete mtDNA. Asterisks indicate genomes assembled from publicly available datasets.

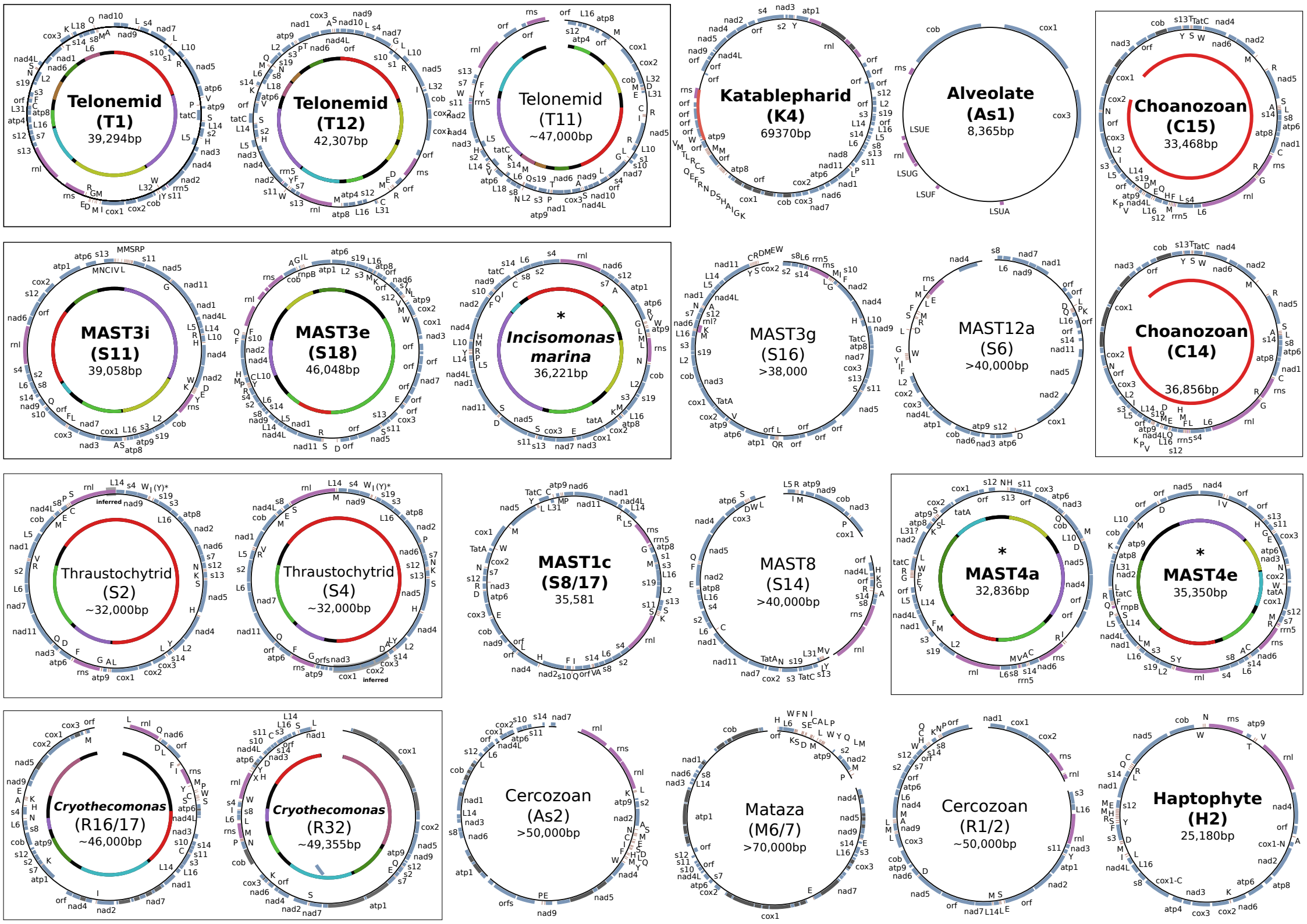
Fig. 6. Phylogenetic reconstruction of representative stramenopiles using concatenated conserved mitochondria-encoded electron transport chain proteins. Electron transport chain proteins encoded in publicly available mtDNAs and our newly sequenced mtDNAs of stramenopiles and rhizarians were collected, aligned, masked, and concatenated, resulting in a 16-protein 4442-site alignment. We excluded

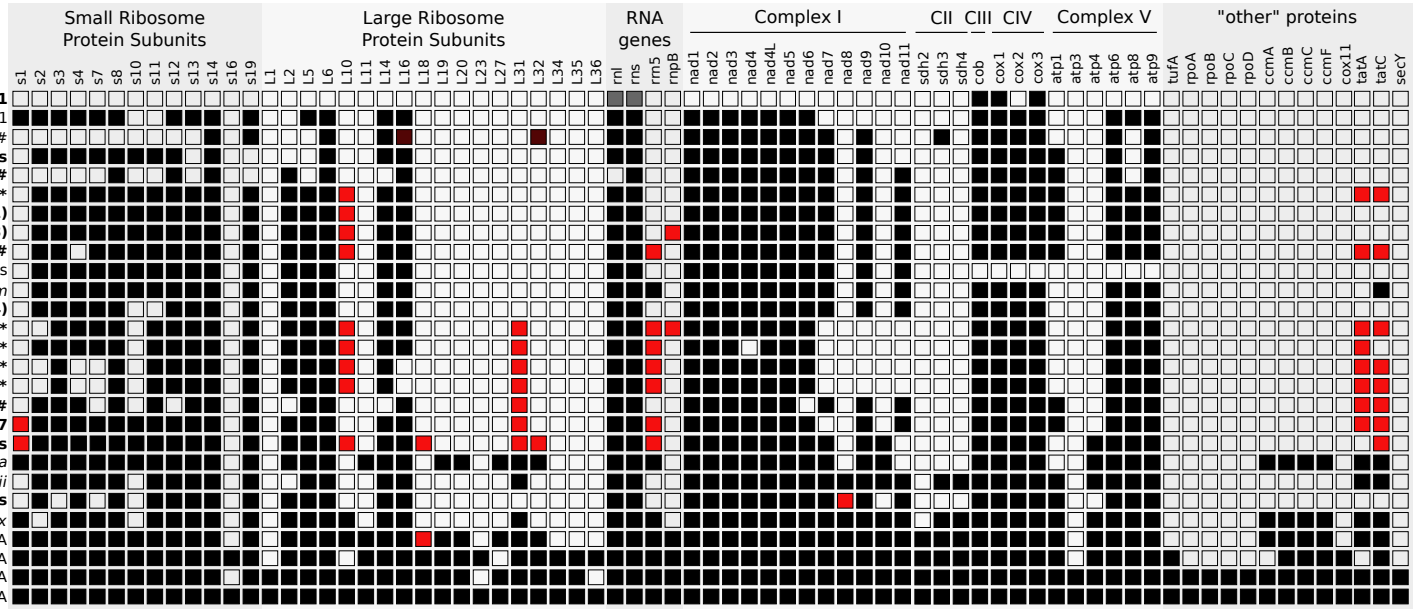
alveolate mtDNAs from this analysis because most of these datasets encode very few (e.g. dinoflagellates and apicomplexans) and/or highly divergent proteins (e.g. ciliates). Phylogenies were reconstructed and node support values were calculated using MrBayes v3.2.6 for posterior probability[83] and RAxML v8.2.10 for maximum likelihood[84] and presented as inset (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*, *nad11*) that have putatively moved to the nucleus are bolded and 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. MtDNAs presented in this study are indicated in bold. Percentages indicate the estimated completeness of each mtDNA presented in this study.

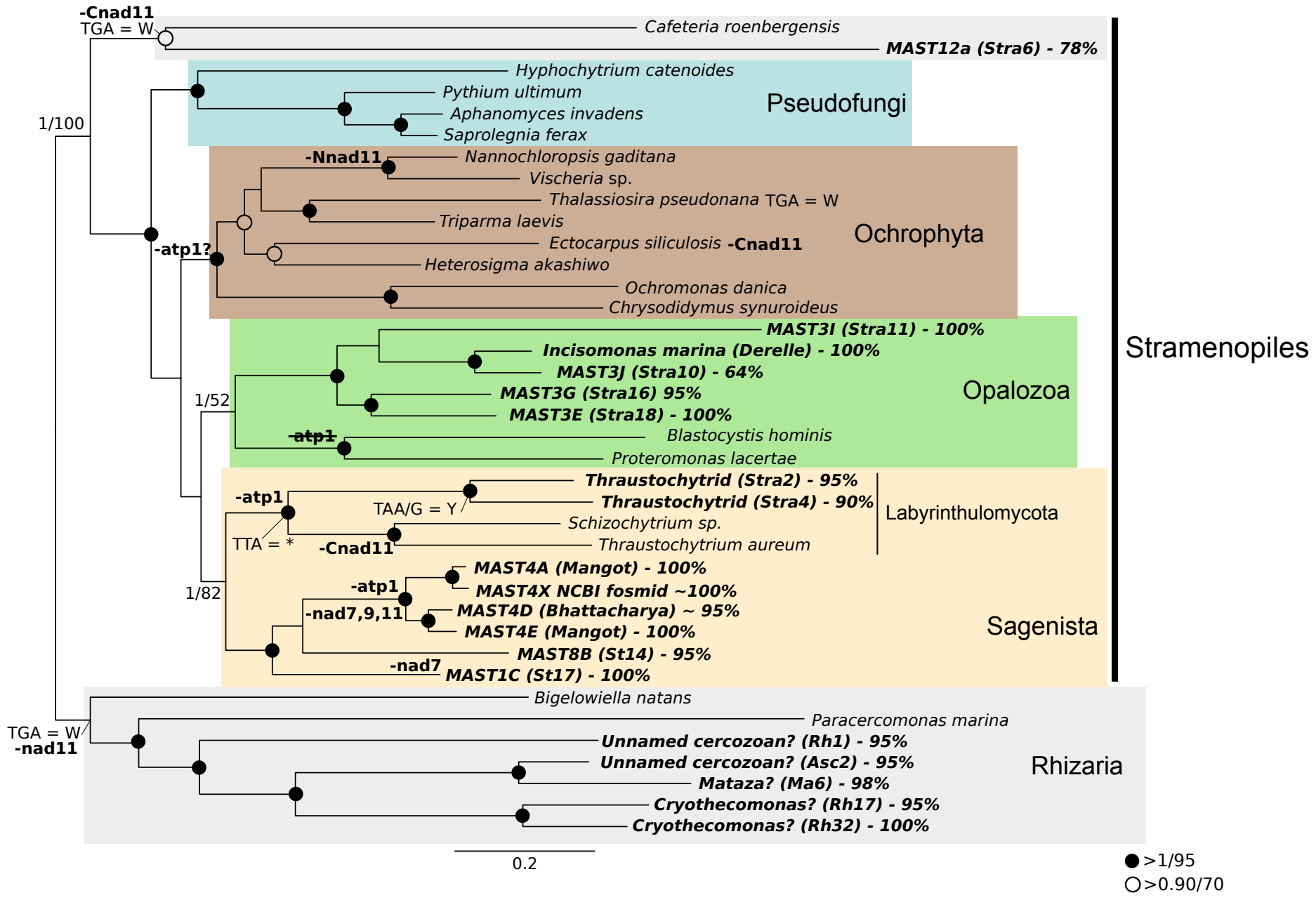

Archaeplastida
Chlorophyta
Streptophyta
Rhodophyta
Stramenopiles
(22)
Rhizaria
(53)
Hacrobia
(21)
Amoebozoa
0.25
Eastern North Pacific
SAG V9
Opisthokonta
(37)
Amorphea
(37)
Jakobida
Metamonada
Ciliophora
Euglenozoa
(21)
Excavata
(21)
Apicomplexa
Dinophyta (51)
Alveolata
(51)

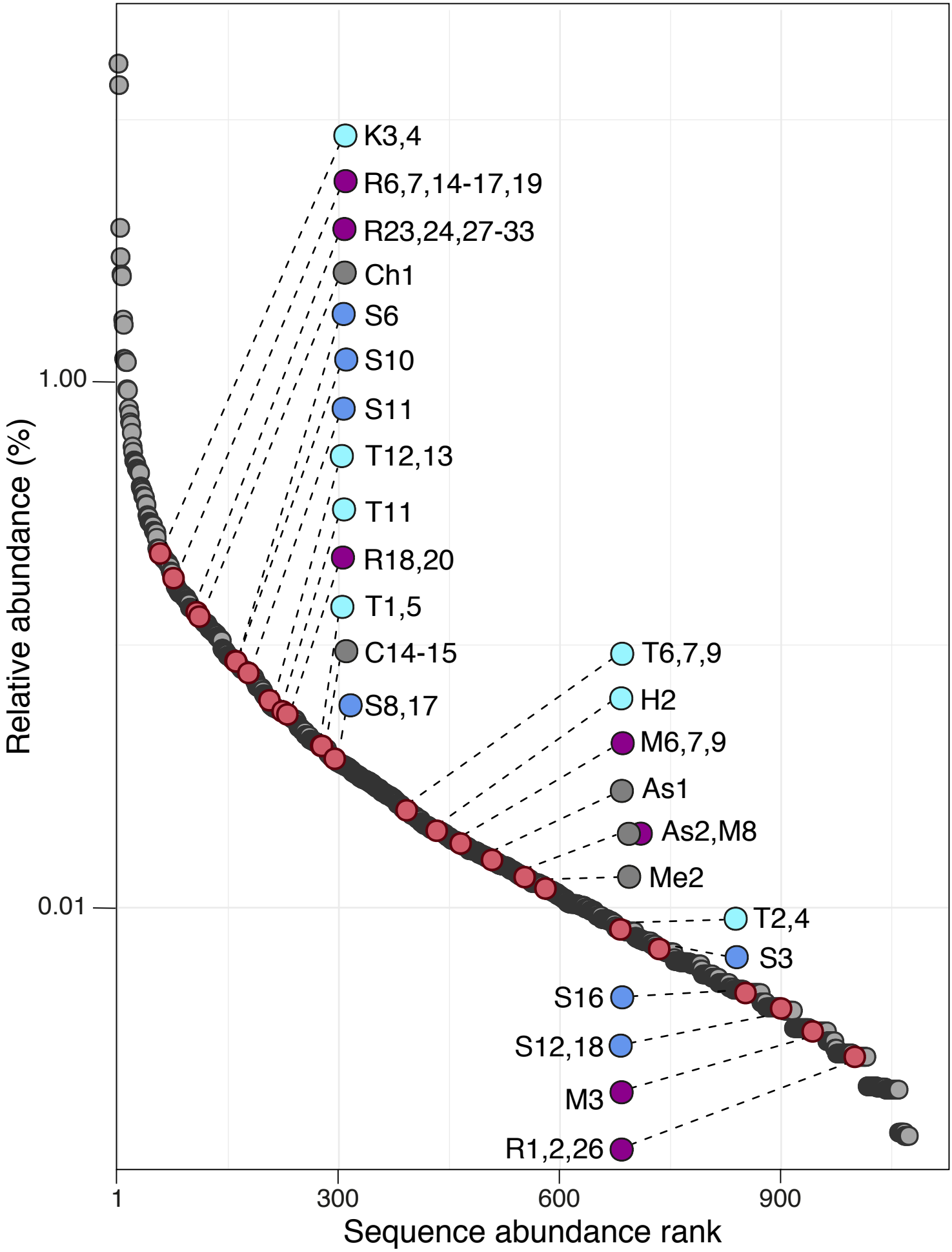

1 - Opisthokonta
As1
Ascomycota (2)
As2
Basidiomycota (11)
Amoebozoa
Mycetozoa
Choanoflagellida (22)
C14 C15 C16
Mesomycetozoa/Chytrid (2)
Ch1 Me2
0.25
2 - Hacrobia
Telonemia (13)
T1 T2,4,5,6,7,9,14
T11 T10 T12,13
Centroheliozoa (1)
Apusozoa (1)
Picozoa
Cryptophyta
Katablepharidophyta (3)
K4 K3
H2
Haptophyta (3)
H3
0.25
3 - Stramenopiles
Bacillariophyta
Bolidophyceae
Chrysophyceae (3)
Dictyochophyceae
Phaeophyceae
Pelagophyceae
S8,17 MAST-1 (4)
Opalinata (1) S19
S6 MAST-12 (2)
Oomycota
Labyrinthulea (6)
S2 S1,3,4,15,23
S14 MAST-8 (1)
Bicoecea
S16 MAST-22 (1)
S11 S18 S12 S10 MAST-3 (4)
0.25
4 - Rhizaria
Cryomonadida (29)
R7,16,17,19,23,24,27,28,30-33
R6,8,14,15,17,29
Mataza (11) M9 M5-8
Thaumatomonadida (2)
R26
Cercozoa
Imbricatea (4)
R1 R2
Chloarachnea (5)
R3,18,20
Foraminifera
Radiolaria - Acantharea
Polycystinea (1)
Radiolaria
RAD-A (1)
0.25
5 - Alveolata / Dinophyta
Dinophyta + MALVs (51)
T8
0.25
6 - Excavata / Euglenozoa
Trypanosomatidae
Neobodo (2)
Diplonemea (8)
Euglenida (10)
D1
0.25

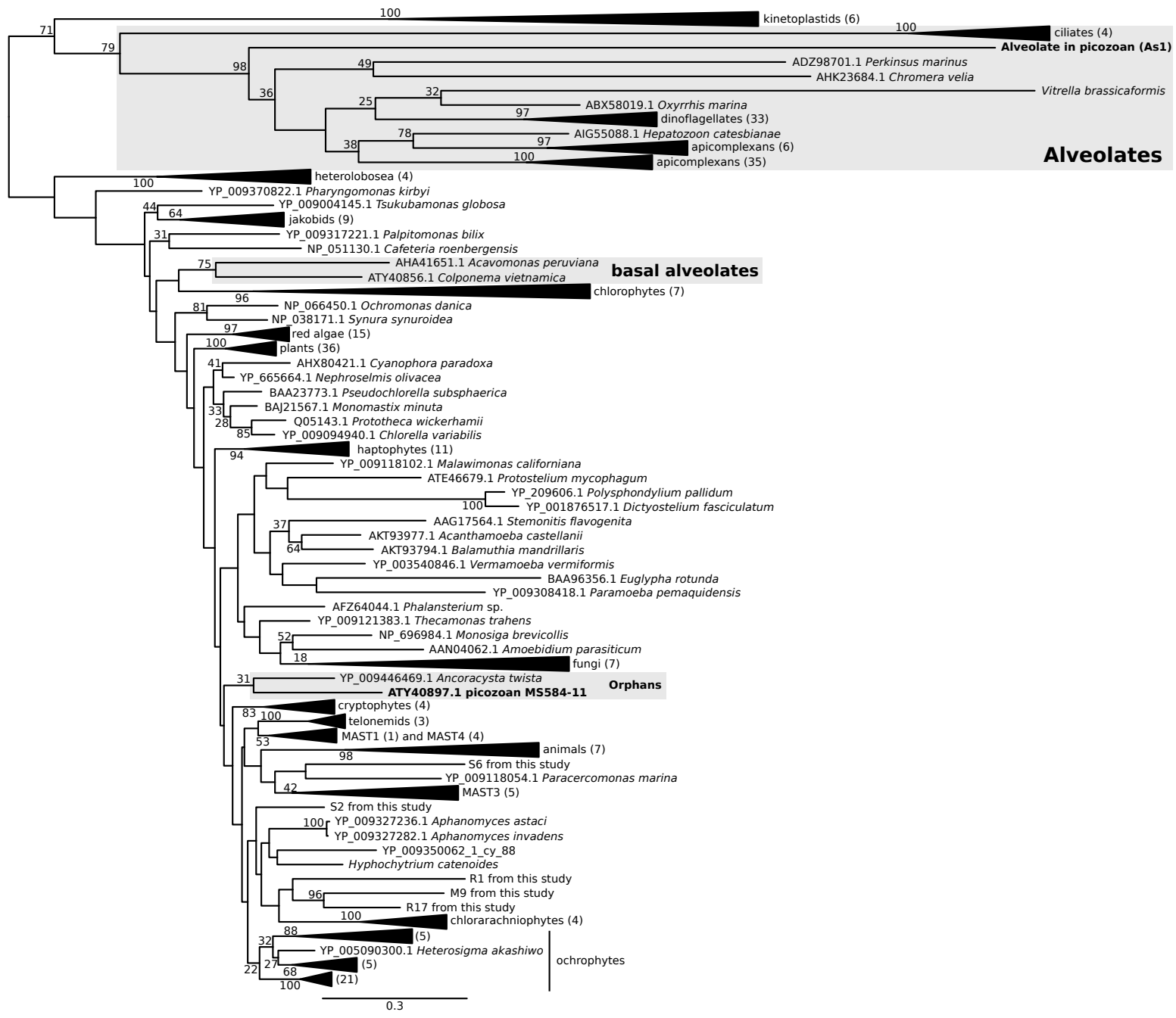

Hotelling's T²
p = 9.07e−13
High Mitochondrial Coverage
High Nuclear Coverage
Low/Intermediate Nuclear and Mitochondrial Coverage
Nuclear Completion (CEGMA%)
Mitochondrial Coverage (X)
0
10
20
30
40
50

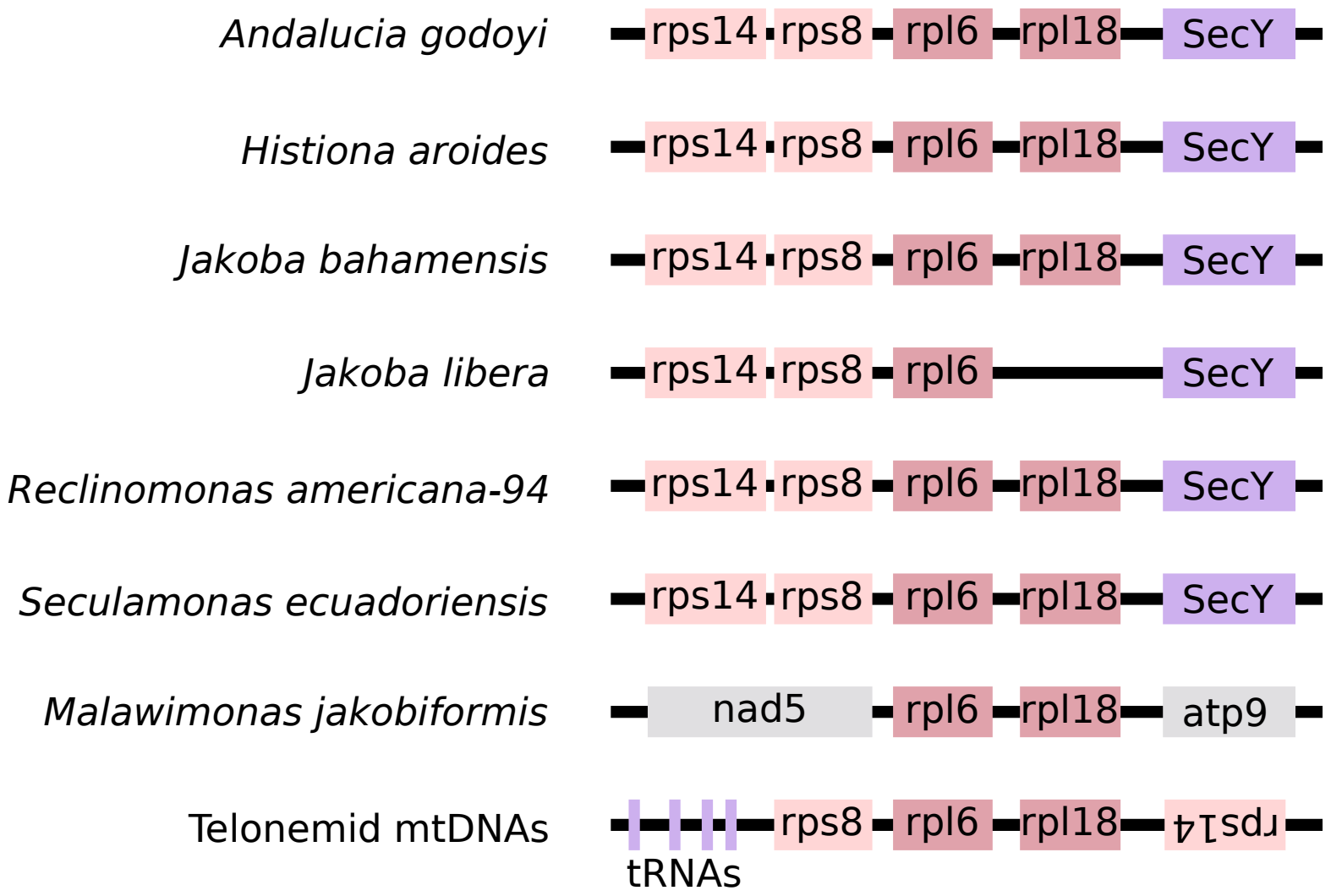

Telonemid
(T1)
39,294bp
Telonemid
(T12)
42,307bp
Telonemid
(T11)
~47,000bp
Katablepharid
(K4)
69370bp
Alveolate
(As1)
8,365bp
LSUE
LSUF
LSUG
LSUA
Choanozoan
(C15)
33,468bp
MAST3i
(S11)
39,058bp
MAST3e
(S18)
46,048bp
Incisomonas marina
36,221bp
MAST3g
(S16)
>38,000
MAST12a
(S6)
>40,000bp
Choanozoan
(C14)
36,856bp
Thraustochytrid
(S2)
~32,000bp
Thraustochytrid
(S4)
~32,000bp
MAST1c
(S8/17)
35,581
MAST8
(S14)
>40,000bp
*MAST4a
32,836bp
*MAST4e
35,350bp
Cryothecomonas
(R16/17)
~46,000bp
Cryothecomonas
(R32)
~49,355bp
Cercozoan
(As2)
>50,000bp
Mataza
(M6/7)
>70,000bp
Cercozoan
(R1/2)
~50,000bp
Haptophyte
(H2)
25,180bp

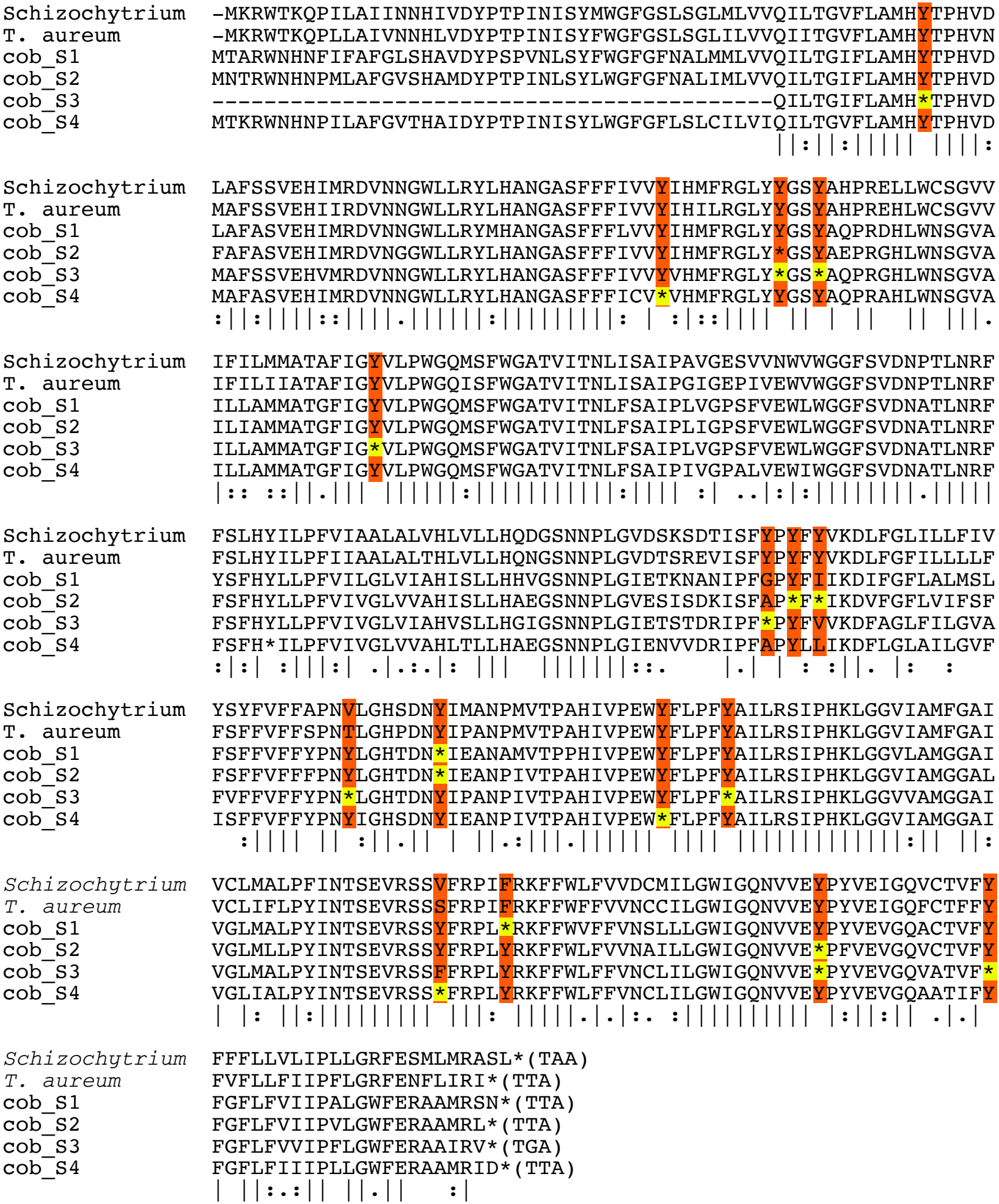

Small Ribosome Protein Subunits
Large Ribosome Protein Subunits
RNA genes
Complex I
CII
CIII
CIV
Complex V
"other" proteins
s1 s2 s3 s4 s7 s8 s10 s11 s12 s13 s14 s16 s19
L1 L2 L5 L6 L10 L11 L14 L16 L18 L19 L20 L23 L27 L31 L32 L34 L35 L36
rnl rns rnpB
nad1 nad2 nad3 nad4 nad4L nad5 nad6 nad7 nad8 nad9 nad10 nad11
sdh2 sdh3 sdh4
cob
cox1 cox2 cox3
atp1 atp3 atp4 atp6 atp8 atp9
tufA rpoA rpoB rpoC rpoD ccmA ccmB ccmC ccmF cox11 tatA tatC secY
Alveolate in As1
Picozoa MS5584-11
C. vietnamica#
Cercozoans
Mast12 (S6)#
I. marina*
Mast3i (S11)
Mast3e (S18)
Mast3g (S16)#
Blastocystis
T. aureum
Thrausto. (S2/S4)
Mast4e*
Mast4d#*
Mast4a*
Mast4a NCBI*
Mast8b S14#
Mast1c S17
Telonemids
Ancoracysta twista
H. andersenii
Katablepharids
P. bilix
L-Dia-CA
L-Amo-CA
L-Jak-CA
LECA

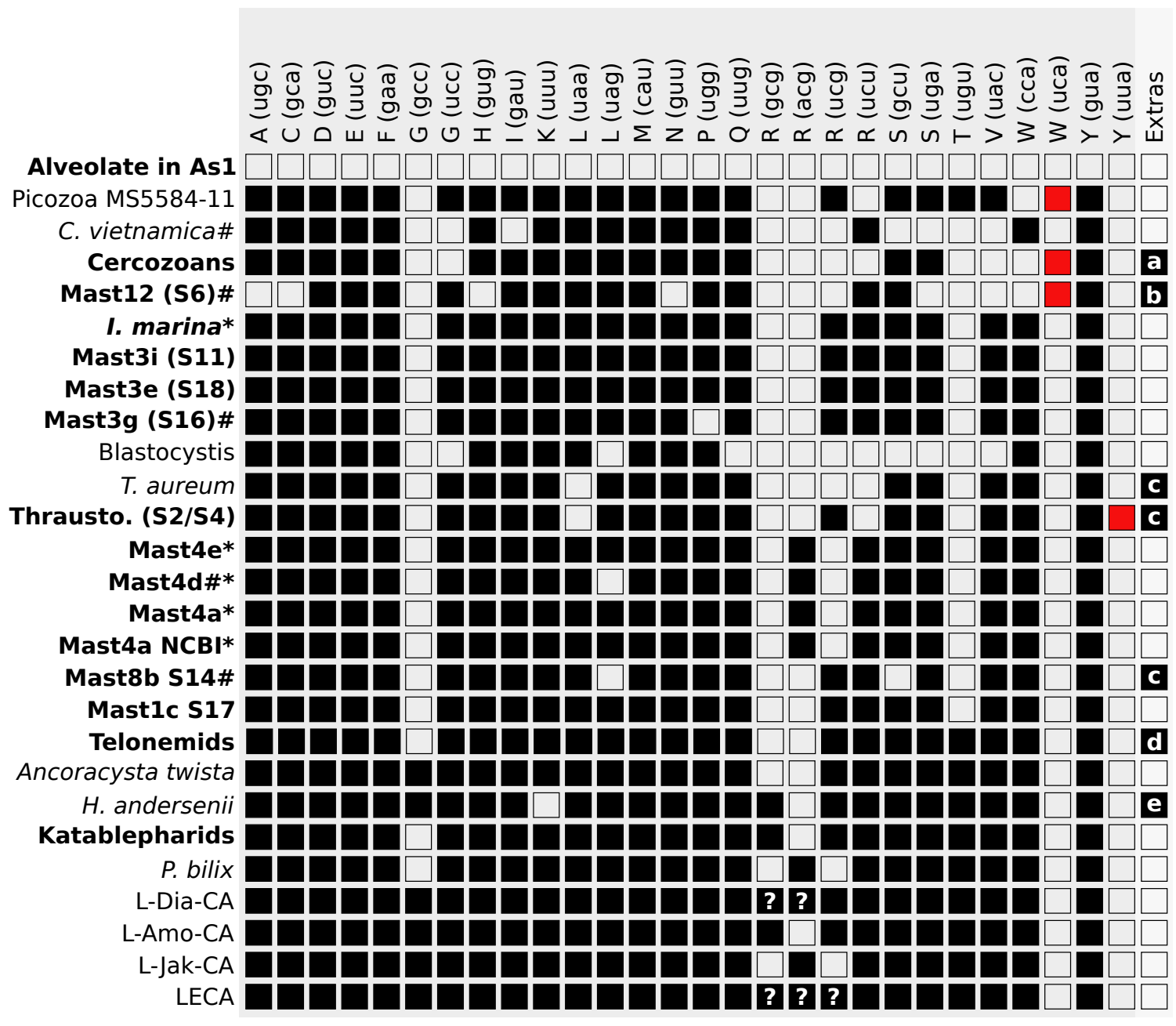

-Cnad11
TGA = W
Cafeteria roenbergensis
MAST12a (Stra6) - 78%
Hyphochytrium catenoides
Pythium ultimum
Aphanomyces invadens
Saprolegnia ferax
Pseudofungi
1/100
-Nnad11
Nannochloropsis gaditana
Vischeria sp.
Thalassiosira pseudonana TGA = W
Triparma laevis
Ectocarpus siliculosis -Cnad11
Heterosigma akashiwo
Ochromonas danica
Chrysodidymus synuroideus
Ochrophyta
-atp1?
MAST3I (Stra11) - 100%
Incisomonas marina (Derelle) - 100%
MAST3J (Stra10) - 64%
MAST3G (Stra16) 95%
MAST3E (Stra18) - 100%
1/52
-atp1
Blastocystis hominis
Proteromonas lacertae
Opalozoa
Stramenopiles
Thraustochytrid (Stra2) - 95%
-atp1
TAA/G = Y
Thraustochytrid (Stra4) - 90%
Schizochytrium sp.
TTA = *
-Cnad11
Thraustochytrium aureum
Labyrinthulomycota
MAST4A (Mangot) - 100%
1/82
-atp1
MAST4X NCBI fosmid ~100%
-nad7,9,11
MAST4D (Bhattacharya) ~ 95%
MAST4E (Mangot) - 100%
MAST8B (St14) - 95%
-nad7
MAST1C (St17) - 100%
Sagenista
Bigelowiella natans
TGA = W
-nad11
Paracercomonas marina
Unnamed cercozoan? (Rh1) - 95%
Unnamed cercozoan? (Asc2) - 95%
Mataza? (Ma6) - 98%
Cryothecomonas? (Rh17) - 95%
Cryothecomonas? (Rh32) - 100%
Rhizaria
0.2
>1/95
>0.90/70

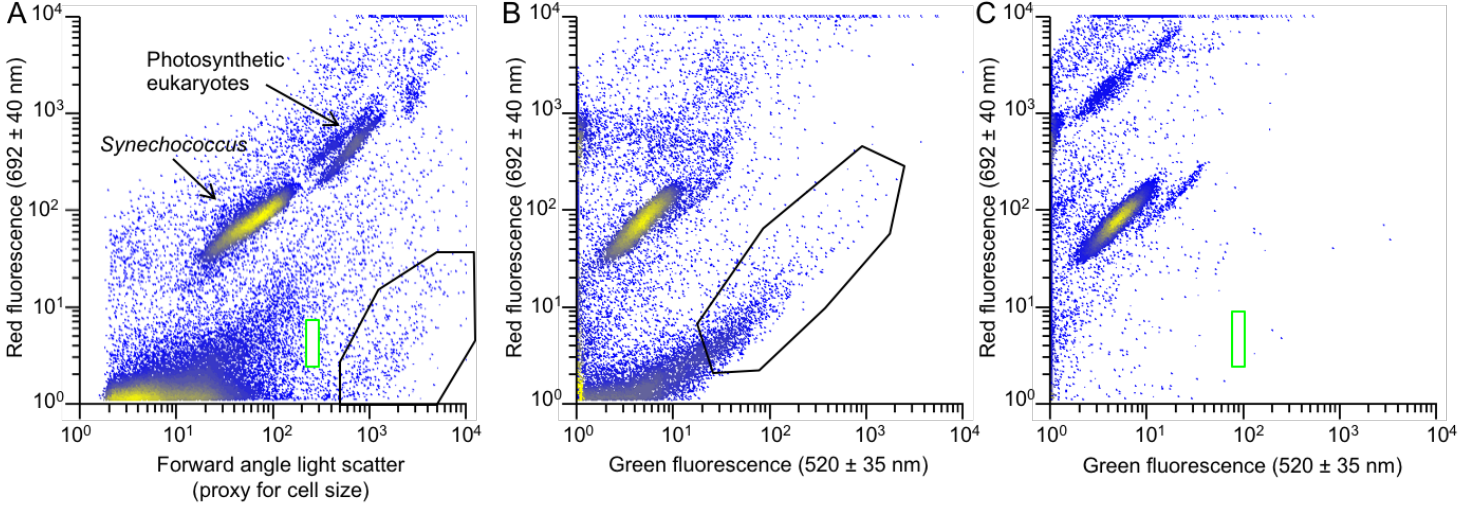

K3,4
R6,7,14-17,19
R23,24,27-33
Ch1
S6
S10
S11
T12,13
T11
R18,20
T1,5
C14-15
S8,17
T6,7,9
H2
M6,7,9
As1
As2,M8
Me2
T2,4
S3
S16
S12,18
M3
R1,2,26
1.00
0.01
Relative abundance (%)
1
300
600
900
Sequence abundance rank


kinetoplastids (6)
ciliates (4)
Alveolate in picozoan (As1)
ADZ98701.1 Perkinsus marinus
AHK23684.1 Chromera velia
Vitrella brassicaformis
ABX58019.1 Oxyrrhis marina
dinoflagellates (33)
AIG55088.1 Hepatozoon catesbianae
apicomplexans (6)
apicomplexans (35)
Alveolates
heterolobosea (4)
YP_009370822.1 Pharyngomonas kirbyi
YP_009004145.1 Tsukubamonas globosa
jakobids (9)
YP_009317221.1 Palpitomonas bilix
NP_051130.1 Cafeteria roenbergensis
AHA41651.1 Acavomonas peruviana
ATY40856.1 Colponema vietnamica
basal alveolates
chlorophytes (7)
NP_066450.1 Ochromonas danica
NP_038171.1 Synura synuroidea
red algae (15)
plants (36)
AHX80421.1 Cyanophora paradoxa
YP_665664.1 Nephroselmis olivacea
BAA23773.1 Pseudochlorella subsphaerica
BAJ21567.1 Monomastix minuta
Q05143.1 Prototheca wickerhamii
YP_009094940.1 Chlorella variabilis
haptophytes (11)
YP_009118102.1 Malawimonas californiana
ATE46679.1 Protostelium mycophagum
YP_209606.1 Polysphondylium pallidum
YP_001876517.1 Dictyostelium fasciculatum
AAG17564.1 Stemonitis flavogenita
AKT93977.1 Acanthamoeba castellanii
AKT93794.1 Balamuthia mandrillaris
YP_003540846.1 Vermamoeba vermiformis
BAA96356.1 Euglypha rotunda
YP_009308418.1 Paramoeba pemaquidensis
AFZ64044.1 Phalansterium sp.
YP_009121383.1 Thecamonas trahens
NP_696984.1 Monosiga brevicollis
AAN04062.1 Amoebidium parasiticum
fungi (7)
YP_009446469.1 Ancoracysta twista
ATY40897.1 picozoan MS584-11
Orphans
cryptophytes (4)
telonemids (3)
MAST1 (1) and MAST4 (4)
animals (7)
S6 from this study
YP_009118054.1 Paracercomonas marina
MAST3 (5)
S2 from this study
YP_009327236.1 Aphanomyces astaci
YP_009327282.1 Aphanomyces invadens
YP_009350062_1_cy_88
Hyphochytrium catenoides
R1 from this study
M9 from this study
R17 from this study
chlorarachniophytes (4)
(5)
YP_005090300.1 Heterosigma akashiwo
(5)
(21)
ochrophytes
0.3


Andalucia godoyi
rps14
rps8
rpl6
rpl18
SecY
Histiona aroides
rps14
rps8
rpl6
rpl18
SecY
Jakoba bahamensis
rps14
rps8
rpl6
rpl18
SecY
Jakoba libera
rps14
rps8
rpl6
SecY
Reclinomonas americana-94
rps14
rps8
rpl6
rpl18
SecY
Seculamonas ecuadoriensis
rps14
rps8
rpl6
rpl18
SecY
Malawimonas jakobiformis
nad5
rpl6
rpl18
atp9
Telonemid mtDNAs
rps8
rpl6
rpl18
rps14
tRNAs


Schizochytrium -MKRWTKQPILAIINNHIVDYPTPINISYMWGFGSLSGLMLVVQILTGVFLAMHYTPHVD
T. aureum -MKRWTKQPLLAIVNNHLVDYPTPINISYFWGFGSLSGLILVVQIITGVFLAMHYTPHVN
cob_S1 MTARWNHNFIFAFGLSHAVDYPSPVNLSYFWGFGFNALMMLVVQILTGIFLAMHYTPHVD
cob_S2 MNTRWNHNPMLAFGVSHAMDYPTPINLSYLWGFGFNALIMLVVQILTGIFLAMHYTPHVD
cob_S3 ----------QILTGIFLAMH*TPHVD
cob_S4 MTKRWNHNPILAFGVTHAIDYPTPINISYLWGFGFLSLCILVIQILTGVFLAMHYTPHVD
Schizochytrium LAFSSVEHIMRDVNNGWLLRYLHANGASFFFIVVYIHMFRGLYYGSYAHPRELLWCSGVV
T. aureum MAFSSVEHIIRDVNNGWLLRYLHANGASFFFIVVYIHILRGLYYGSYAHPREHLWCSGVV
cob_S1 LAFASVEHIMRDVNNGWLLRYMHANGASFFFLVVYIHMFRGLYYGSYAQPRDHLWNSGVA
cob_S2 FAFASVEHIMRDVNGGWLLRYLHANGASFFFIVVYIHMFRGLY*GSYAEPRGHLWNSGVA
cob_S3 MAFSSVEHVMRDVNNGWLLRYLHANGASFFFIVVYVHMFRGLY*GS*AQPRGHLWNSGVA
cob_S4 MAFASVEHIMRDVNNGWLLRYLHANGASFFFICV*VHMFRGLYYGSYAQPRAHLWNSGVA
Schizochytrium IFILMMATAFIGYVLPWGQMSFWGATVITNLISAIPAVGESVVNWVWGGFSVDNPTLNRF
T. aureum IFILIIATAFIGYVLPWGQISFWGATVITNLISAIPGIGEPIVEWVWGGFSVDNPTLNRF
cob_S1 ILLAMMATGFIGYVLPWGQMSFWGATVITNLFSAIPLVGPSFVEWLWGGFSVDNATLNRF
cob_S2 ILIAMMATGFIGYVLPWGQMSFWGATVITNLFSAIPLIGPSFVEWLWGGFSVDNATLNRF
cob_S3 ILLAMMATGFIG*VLPWGQMSFWGATVITNLFSAIPLVGPSFVEWLWGGFSVDNATLNRF
cob_S4 ILLAMMATGFIGYVLPWGQMSFWGATVITNLFSAIPIVGPALVEWIWGGFSVDNATLNRF
Schizochytrium FSLHYILPFVIAALALVHLVLLHQDGSNNPLGVDSKSDTISFYPYFYVKDLFGLILLFIV
T. aureum FSLHYILPFIIAALALTHLVLLHQNGSNNPLGVDTSREVISFYPYFYVKDLFGFILLLLF
cob_S1 YSFHYLLPFVILGLVIAHISLLHHVGSNNPLGIETKNANIPFGPYFIIKDIFGFLALMSL
cob_S2 FSFHYLLPFVIVGLVVAHISLLHAEGSNNPLGVESISDKISFAP*F*IKDVFGFLVIFSF
cob_S3 FSFHYLLPFVIVGLVIAHVSLLHGIGSNNPLGIETSTDRIPF*PYFVVKDFAGLFILGVA
cob_S4 FSFH*ILPFVIVGLVVAHLTLLHAEGSNNPLGIENVVDRIPFAPYLLIKDFLGLAILGVF
Schizochytrium YSYFVFFAPNVLGHSDNYIMANPMVTPAHIVPEWYFLPFYAILRSIPHKLGGVIAMFGAI
T. aureum FSFFVFFSPNTLGHPDNYIPANPMVTPAHIVPEWYFLPFYAILRSIPHKLGGVIAMFGAI
cob_S1 FSFFVFFYPNYLGHTDN*IEANAMVTPPHIVPEWYFLPFYAILRSIPHKLGGVLAMGGAI
cob_S2 FSFFVFFFPNYLGHTDN*IEANPIVTPAHIVPEWYFLPFYAILRSIPHKLGGVIAMGGAL
cob_S3 FVFFVFFYPN*LGHTDNYIPANPIVTPAHIVPEWYFLPF*AILRSIPHKLGGVVAMGGAI
cob_S4 ISFFVFFYPNYIGHSDNYIEANPIVTPAHIVPEW*FLPFYAILRSIPHKLGGVIAMGGAI
Schizochytrium VCLMALPFINTSEVRSSVFRPIFRKFFWLFVVDCMILGWIGQNVVEYPYVEIGQVCTVFY
T. aureum VCLIFLPYINTSEVRSSSFRPIFRKFFWFFVVNCCILGWIGQNVVEYPYVEIGQFCTFFY
cob_S1 VGLMALPYINTSEVRSSYFRPL*RKFFWVFFVNSLLLGWIGQNVVEYPYVEVGQACTVFY
cob_S2 VGLMLLPYINTSEVRSSYFRPLYRKFFWLFVVNAILLGWIGQNVVE*PFVEVGQVCTVFY
cob_S3 VGLMALPYINTSEVRSSFFRPLYRKFFWLFFVNCLILGWIGQNVVE*PYVEVGQVATVF*
cob_S4 VGLIALPYINTSEVRSS*FRPLYRKFFWLFFVNCLILGWIGQNVVEYPYVEVGQAATIFY
Schizochytrium FFFLLVLIPLLGRFESMLMRASL*(TAA)
T. aureum FVFLLFIIPFLGRFENFLIRI*(TTA)
cob_S1 FGFLFVIIPALGWFERAAMRSN*(TTA)
cob_S2 FGFLFVIIPVLGWFERAAMRL*(TTA)
cob_S3 FGFLFVVIPFLGWFERAAIRV*(TGA)
cob_S4 FGFLFIIIPLLGWFERAAMRID*(TTA)


A (ugc)
C (gca)
D (guc)
E (uuc)
F (gaa)
G (gcc)
G (ucc)
H (gug)
I (gau)
K (uuu)
L (uaa)
L (uag)
M (cau)
N (guu)
P (ugg)
Q (uug)
R (gcg)
R (acg)
R (ucg)
R (ucu)
S (gcu)
S (uga)
T (ugu)
V (uac)
W (cca)
W (uca)
Y (gua)
Y (uua)
Extras
Alveolate in As1
Picozoa MS5584-11
C. vietnamica#
Cercozoans
Mast12 (S6)#
I. marina*
Mast3i (S11)
Mast3e (S18)
Mast3g (S16)#
Blastocystis
T. aureum
Thrausto. (S2/S4)
Mast4e*
Mast4d#*
Mast4a*
Mast4a NCBI*
Mast8b S14#
Mast1c S17
Telonemids
Ancoracysta twista
H. andersenii
Katablepharids
P. bilix
L-Dia-CA
L-Amo-CA
L-Jak-CA
LECA
a
b
c
c
c
d
e
?
?
?
?
?


A
B
C
Photosynthetic eukaryotes
Synechococcus
Red fluorescence (692 ± 40 nm)
Forward angle light scatter (proxy for cell size)
Green fluorescence (520 ± 35 nm)
Green fluorescence (520 ± 35 nm)
10^0
10^1
10^2
10^3
10^4