Peer Community Journal

Section: Genomics

RESEARCH ARTICLE

Published
2022-07-19

Cite as
Brandon Kwee Boon Seah, Aditi Singh and Estienne Carl Swart (2022) *Karyorelict ciliates use an ambiguous genetic code with context-dependent stop/sense codons*, Peer Community Journal, 2: e42.

Correspondence
kb.seah@tuebingen.mpg.de
estienne.swart@tuebingen.mpg.de

Peer-review
Peer reviewed and recommended by PCI Genomics,
https://doi.org/10.24072/pci.genomics.100019

Karyorelict ciliates use an ambiguous genetic code with context-dependent stop/sense codons

Brandon Kwee Boon Seah[1], Aditi Singh[1], and Estienne Carl Swart[1]

Volume **2** (2022), article e42

https://doi.org/10.24072/pcjournal.141

Abstract

In ambiguous stop/sense genetic codes, the stop codon(s) not only terminate translation but can also encode amino acids. Such codes have evolved at least four times in eukaryotes, twice among ciliates (*Condylostoma magnum* and *Parduczia* sp.). These have appeared to be isolated cases whose next closest relatives use conventional stop codons. However, little genomic data have been published for the Karyorelictea, the ciliate class that contains *Parduczia* sp., and previous studies may have overlooked ambiguous codes because of their apparent rarity. We therefore analyzed single-cell transcriptomes from four of the six karyorelict families to determine their genetic codes. Reassignment of canonical stops to sense codons was inferred from codon frequencies in conserved protein domains, while the actual stop codon was predicted from full-length transcripts with intact 3'-untranslated regions (3'-UTRs). We found that all available karyorelicts use the *Parduczia* code, where canonical stops UAA and UAG are reassigned to glutamine, and UGA encodes either tryptophan or stop. Furthermore, a small minority of transcripts may use an ambiguous stop-UAA instead of stop-UGA. Given the ubiquity of karyorelicts in marine coastal sediments, ambiguous genetic codes are not mere marginal curiosities but a defining feature of a globally distributed and diverse group of eukaryotes.

[1] Max Planck Institute for Biology, 72076 Tübingen, Germany

e-ISSN 2804-3871

Introduction

In addition to the "standard" genetic code used by most organisms, there are numerous variant codes across the tree of life, and new ones continue to be discovered [1–3]. The differences between codes lie in which amino acids are coded by which codon, as well as which codons are used to start and terminate translation (stop codons). Much of the variation is concentrated in a small number of codons, particularly the canonical stop codons UAA, UAG, and UGA, which have repeatedly been reassigned to encode amino acids. The most striking variants are ambiguous codes where one codon can have multiple meanings. The outcome during translation can be stochastic, such as in stop codon readthrough [4], or translation of CUG as either leucine or serine by *Candida* spp. [5]. Alternatively, they can be context-dependent, such as UGA encoding selenocysteine only in selenoproteins [6], meaning that the translation system is able to interpret the codon correctly as either an amino acid or a stop.

Other context-dependent stop/sense codes have been discovered where all the stop codons used by the cell are potentially also sense codons. These have evolved independently several times among the eukaryotes [7–10]: parasitic trypanosomes of the genus *Blastocrithidia* (three different species) use UAA and UAG to encode stop/glutamate (NCBI Genetic Codes ftp.ncbi.nih.gov/entrez/misc/data/gc.prt, table 31); a strain of the marine parasitic dinoflagellate *Amoebophrya* and a marine karyorelict ciliate, *Parduczia* sp., have convergently evolved to use UGA for stop/tryptophan (table 27); and the marine heterotrich ciliate *Condylostoma magnum* uses UGA for stop/tryptophan and UAA/UAG for stop/glutamine (table 28).

The ciliates are a clade with an unusual propensity for variant genetic codes [11]. At least eight different nuclear genetic codes are used by ciliates [10], including some of the first examples of variant codes documented in nuclear genomes [12–16]. At first glance, organisms that use these ambiguous stop/sense codes appear to be isolated single species or strains embedded among relatives with conventional codes. For example, other heterotrichs related to *Condylostoma* use the standard code (e.g. *Stentor*) or the *Blepharisma* code. Additionally, a previous survey of genetic codes across the ciliate tree, including numerous uncultivated heterotrichs and karyorelicts, did not report any new examples of organisms that use ambiguous stop/sense codes, nor appeared to have accounted for such a possibility in their methods [17]. However our own preliminary studies appeared to contradict this, finding other karyorelicts that use the same genetic code as *Parduczia*.

The karyorelicts are a class-level taxon within the ciliates, and sister group to the heterotrichs. Unlike other ciliates, the somatic nuclei (macronuclei) of karyorelicts do not divide but must differentiate anew from germline nuclei (micronuclei) every time, even during vegetative division [18]. They are globally distributed and commonly encountered in the sediment interstitial habitat of marine coastal environments [19]. At least ~150 species have been formally described but this is believed to be a severe underestimate of the true diversity [20,21], and they are also poorly represented in sequence databases.

We therefore sequenced additional karyorelict transcriptomes and reanalyzed published data to assess whether karyorelicts other than *Parduczia* could be using ambiguous genetic codes.

Results

Ten new single-cell RNA-seq libraries from karyorelicts and heterotrichs were sequenced in this study, representing interstitial species from marine sediment at Roscoff, France. These were analyzed alongside 33 previously published RNA-seq libraries (Table_S1.xlsx in [22]). After filtering for quality and sufficient data, 25 transcriptome assemblies (of which 15 were previously published) were used to evaluate stop codon reassignment, vs. 26 assemblies (16 previously published) for inferring the actual stop codon(s) (Appendix).

Reassignment of all three canonical stop codons to sense codons in karyorelicts

Codon frequencies in protein-coding sequences were calculated from sequence regions that aligned to conserved Pfam domains, in transcripts with poly-A tails. Transcriptomes and genomic coding sequences (CDSs) from ciliates with known genetic codes were used as a comparison to estimate the false positive rate of stop codons being found in these alignments, e.g. because of misalignments, misassembly, or pseudogenes.

Among karyorelicts, all three canonical stop codons (UAA, UAG, UGA) were observed in conserved protein domains, with frequencies between 0.08-2.9%, which fell within the range of codon frequencies observed for unambiguous sense codons in other ciliates where the genetic code is known (0.03-6.8%, excluding the outlier CGG in *Tetrahymena thermophila* with only 0.003%). This range was also similar to frequencies of the ambiguous stops in *Parduczia* and the heterotrich *Condylostoma* (Figure 1A). UGA was generally less frequent than UAA/UAG in all karyorelicts, but the frequencies varied between taxa, reflecting their individual codon usage biases or which genes are assembled in the transcriptome because of sequencing depth. UGA was the least-frequent codon in most Trachelocercidae and Geleiidae, but was more frequent in Loxodidae and Kentrophoridae than some other codons, especially C/G-rich ones like CGG (Figure 1A). Nonetheless, frequencies of the UGA codon in karyorelicts were all still one to two orders of magnitude higher than the observed frequencies of in-frame actual stops from other ciliate species in the reference set.

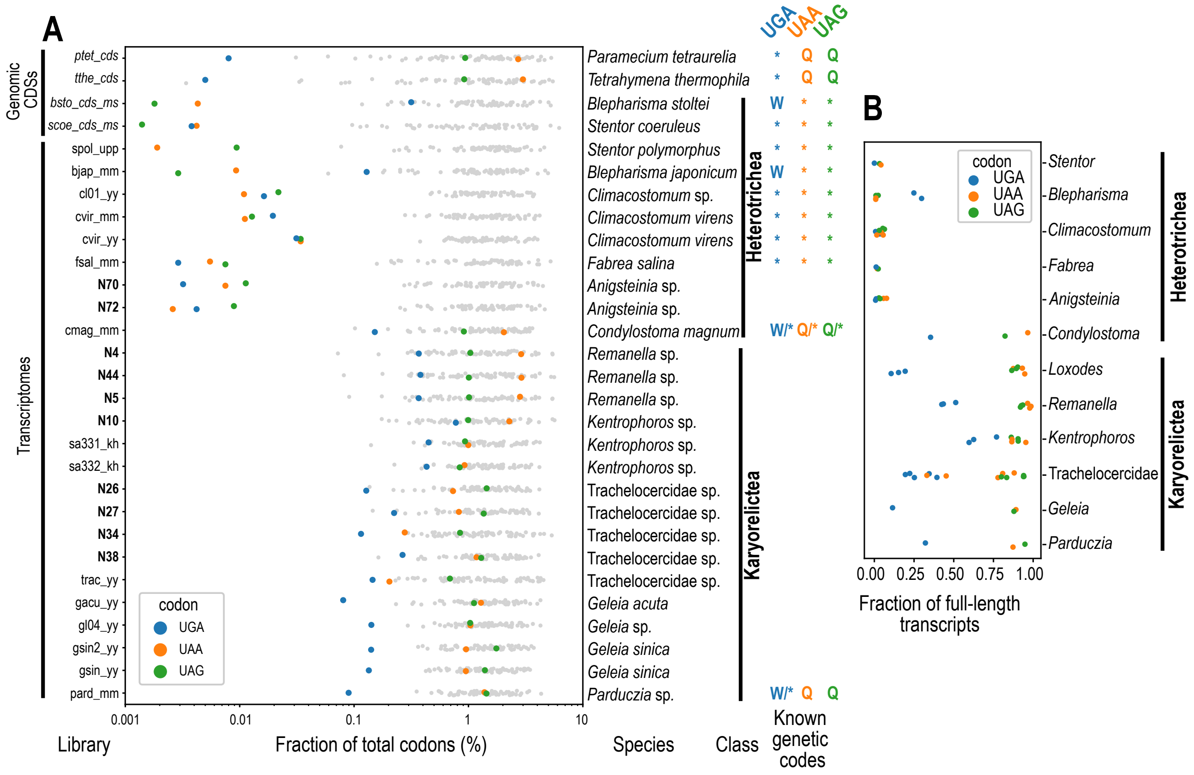


Figure 1. (**A**) Codon frequencies of canonical stop codons (UGA: blue, UAA: orange, UAG: green) and other codons (gray) in conserved protein domains found by hmmscan search in six-frame translations of transcriptome assemblies (Table_S1.xlsx in [22]) or genomic CDSs (Table_S2.xlsx in [22]) vs. Pfam. Names of libraries from this study are highlighted in bold. Assignments of canonical stops for organisms with known genetic codes follow ref. [10]. (**B**) Fraction of full-length transcripts that have at least one canonical stop codon in the putative coding region, grouped by genus (except Trachelocercidae, where classification was unclear).

In-frame UGAs were found in 10.5 to 76.9% of transcripts with putative coding regions predicted by full-length Blastx hits per karyorelict library (Figure 1B). This frequency verified that in-frame UGAs were not concentrated in a small fraction of potentially spurious sequences but in fact found in many genes. Conserved "marker" genes that were generally expected to be present in ciliate genomes (BUSCO orthologs, Alveolata marker set, [23]) also contained in-frame UGAs. The karyorelict transcriptome assemblies were relatively incomplete, with 1.8% to 20.5% (median 12.0%) estimated completeness based on the BUSCO markers, and a total of 91 of 171 BUSCO orthologs were found in these assemblies (Figure 2A). Nonetheless, 46 BUSCO orthologs from 14 karyorelict assemblies were found with in-frame UGAs in conserved alignment positions (e.g. Figure 2B, 2C), verifying that they are not limited to poorly characterized or hypothetical proteins.

In comparison, the heterotrich *Anigsteinia*, for which two new sequence libraries were also produced and which was found in the same habitats as karyorelicts, had in-frame frequencies of ≤0.011% for all three canonical stop codons, which were comparable to frequencies of the known stop codons in *Blepharisma* (UAA, UAG) and *Stentor* (UAA, UAG, UGA) (max. 0.09%). Hence *Anigsteinia* probably does not have ambiguous sense/stop codons.

All karyorelicts had the same inferred amino acid reassignments for the three canonical stops: glutamine (Q) for UAA and UAG, and tryptophan (W) for UGA (Figure 3), matching previous predictions for *Parduczia* sp. and *Condylostoma magnum* [9,10].

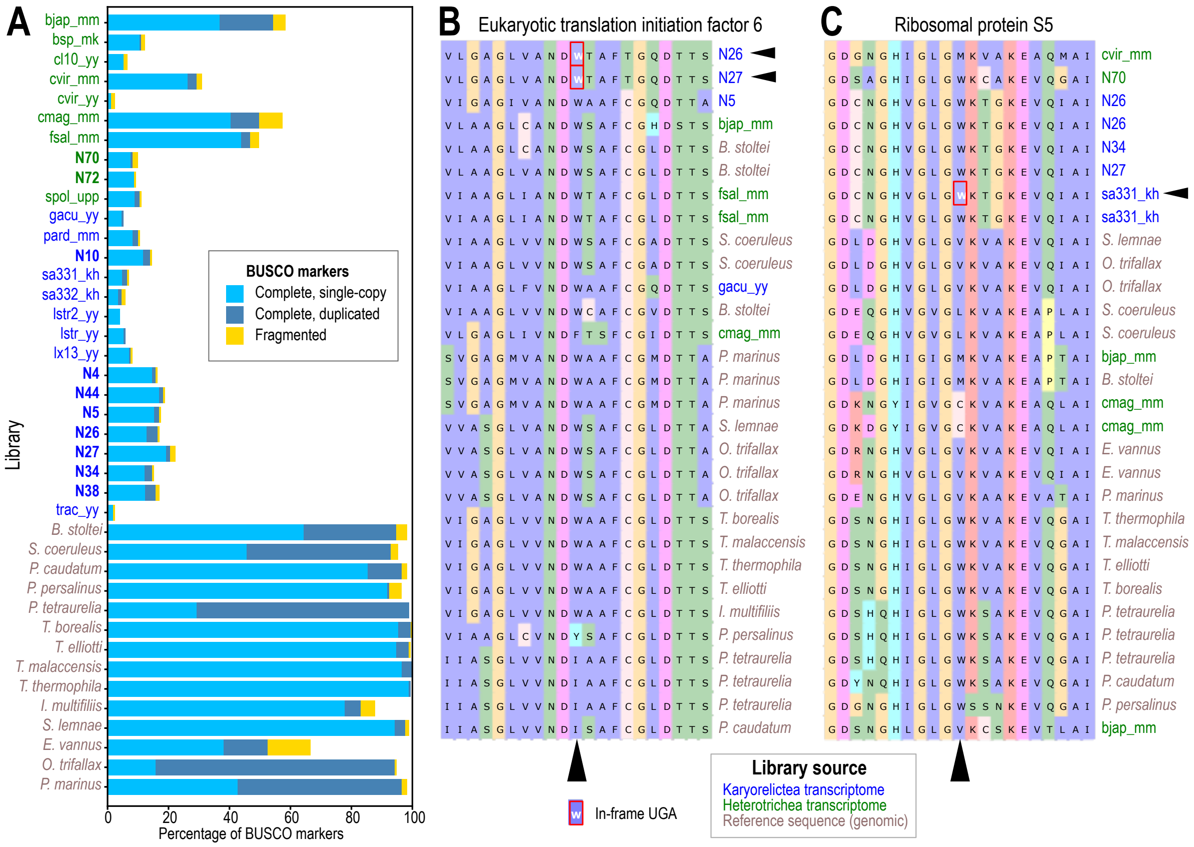


Figure 2. In-frame coding UGAs in conserved marker genes. (**A**) Completeness estimates of heterotrich and karyorelict transcriptomes (library names in green and blue respectively), compared with genomic reference sequences from other ciliates (Table_S3.xlsx in [22]); BUSCO Alveolata marker set. (**B**, **C**) Two examples of alignments (excerpts) for conserved orthologous protein-coding genes (orthologs 20320at33630 and 23778at33630), which contain in-frame UGAs translated as W in karyorelict sequences, flanked by conserved alignment blocks.

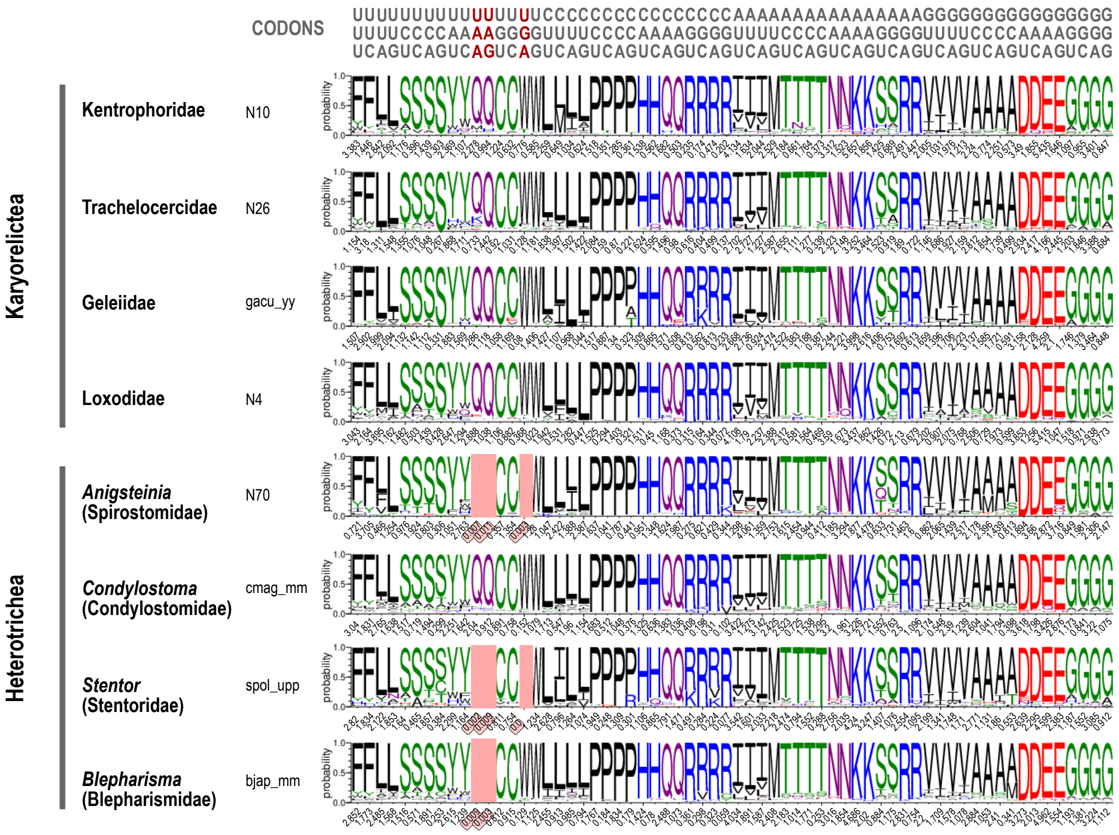


Figure 3. Weblogos representing the likely amino acid assignment of each codon in selected libraries (library with most coverage per taxon of interest). Heights of each letter represent the relative frequencies (all scaled to 100%) of each amino acid in conserved residues aligning to that codon. The observed codon frequency (in %) is indicated below. Codons with frequencies <0.02% are highlighted in red, representing either non-ambiguous stops or unassigned codons. Assignment of cysteine (C) for UGA in *Anigsteinia* is based on only 16 alignments, of which 14 are to a likely selenoprotein (Pfam domain GSHPx); assignment of glutamine (Q) for UAA and UAG in *Blepharisma* may represent recent paralogs or translational readthrough.

Stop codons in karyorelicts and heterotrichs

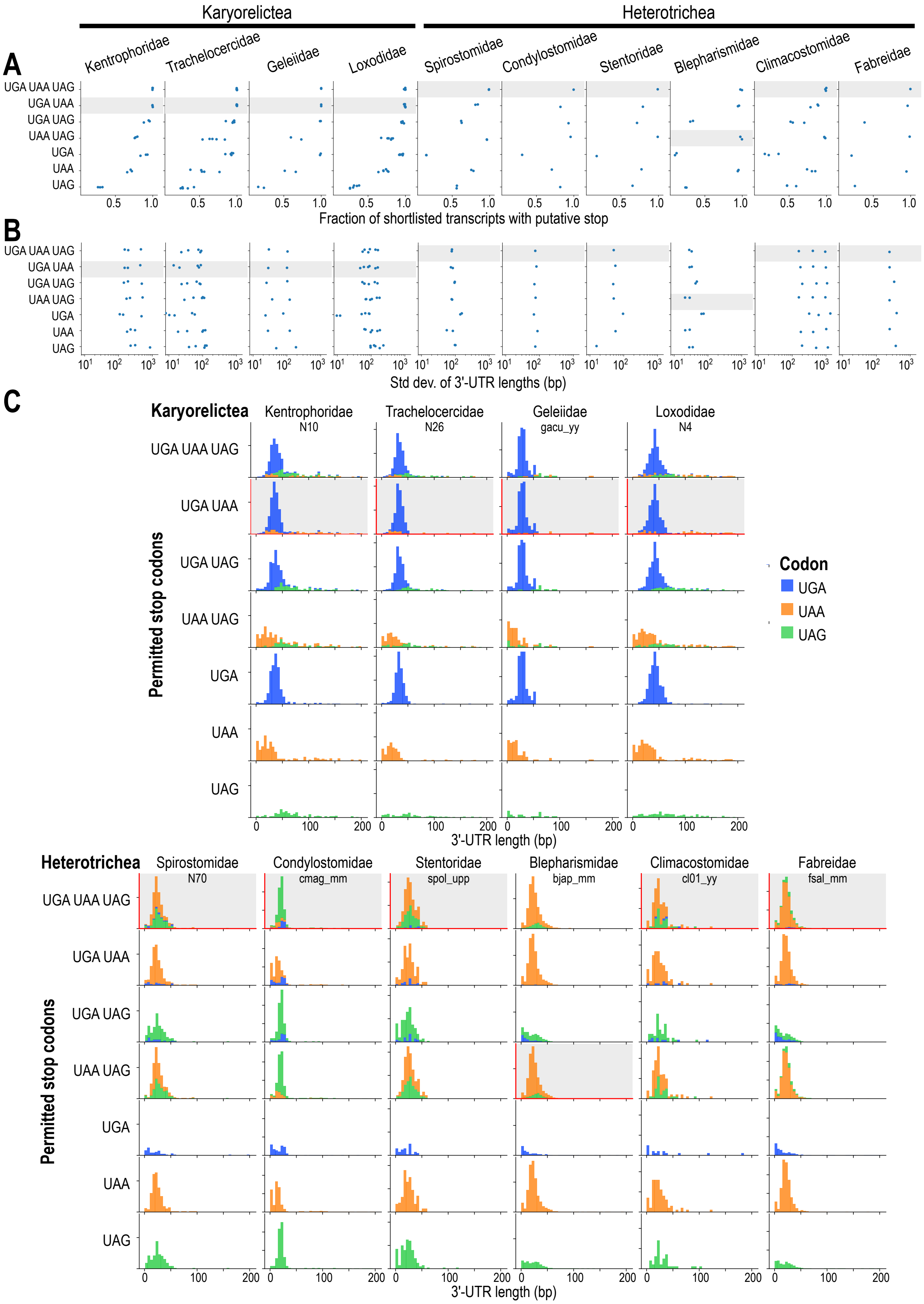
Frequency of a codon in coding regions can be used to infer if it is a sense codon but not whether it can terminate translation, especially for ambiguous codes where codons that can terminate translation also frequently appear in coding sequences. Therefore we used full length transcripts with both a high quality Blastx alignment to a reference protein and a poly-A tail to predict the likely stop codon(s) used in each sample. To avoid double counting, only one isoform was used per gene. We assumed that the true stop codon(s) were one or more of the three canonical stops UGA, UAA, UAG, and that if a contig has a high quality Blastx hit to a reference protein sequence, the true stop should lie somewhere between the last codon at the 3’ end of the hit region and the beginning of the poly-A. We reasoned that if the true stop codon set was used for annotation, (i) the number of transcripts without a putative true stop should be minimized; (ii) the variance of the 3’-untranslated region (3’-UTR) length should also be minimized because ciliate 3’-UTRs are known to be short (mostly <100 bp); and (iii) if there was more than one stop codon, the length distributions of the putative 3’-UTRs for each stop codon should be centered on the same value.

With these criteria, the candidate stop codons for karyorelicts could be narrowed to two possibilities: UGA alone or UGA + UAA. If only UGA was permitted as a stop codon, 84-98% of transcripts per library had a putative true stop, but if both UGA and UAA were permitted as stop codons, the proportion was over 98% (Figure 4A). Permitting both UGA+UAA as stops in karyorelicts resulted in a higher variance in 3'-UTR lengths compared to permitting only UGA. Although this was contrary to criterion (ii) above, we judged that this metric was not as useful in deciding whether UAA was also a stop codon, because the difference was small, and transcripts with putative UAA stops were relatively few (Figures 4B, 4C). Both karyorelicts and heterotrichs in this study had short and narrowly distributed 3'-UTR lengths (median 28 nt, interquartile range 18 nt) (Figure 4C). The heterotrichs were shortest overall, with median lengths per taxon between 21 nt (*Condylostoma*) and 26 nt (*Stentor*), followed by the karyorelict families Trachelocercidae (33 nt), Geleiidae (31 nt), Kentrophoridae (37 nt), and Loxodidae (43 nt).

In previous analyses of the ambiguous stop codons in *Condylostoma* and *Parduczia*, a distinct depletion of in-frame coding "stop" codons immediately upstream of the actual terminal stop was observed [10]. We could reproduce this depletion of all three canonical stops in *Condylostoma* and of UGA in *Parduczia*, about 10 to 20 codon positions before the putative terminal stop, in our reanalysis of the same data (Figure 5A). For the karyorelicts, if only UGA was permitted as a stop codon, we observed depletion of coding-UGA but also of coding-UAAs before the terminal stop-UGA (Figure 5B). If UGA + UAA were permitted as stops, the depletion of coding-UGA before terminal stops was still observed, and the depletion of coding-UAA was even more pronounced (Figure 5C). Unfortunately, there were only a limited number of full-length karyorelict transcripts with putative stop-UAAs (max. 47 contigs per library). We therefore concluded that UGA is the predominant stop codon in karyorelicts, but UAA may also function as a stop codon for about 1-10% of transcripts.

UAA and UAG were predicted as stop codons of *Anigsteinia* (Spirostomidae), consistent with their near-absence from coding regions in this genus (see above, Figure 1A). UGA was not only near-absent from coding regions, but also rarely encountered as a putative stop codon, although it was not uncommon in 3'-UTRs. Similar rarity of UGAs as putative stops was also observed in *Stentor* and other heterotrichs that are said to use the standard code. Either (i) these heterotrichs use the standard genetic code with all three canonical stop codons but a strong bias against using UGA for stop, or (ii) UGA is an unassigned codon in these organisms.

Figure 4 (next page). Effect of different stop codon combinations on assembly metrics. Predicted stop codon usage for each taxon from this study or previous publications highlighted in gray. (**A**) Strip plots for the fraction of full length contigs per transcriptome that have a putative stop codon from that specific combination (rows), i.e. in-frame, downstream of full-length Blastx hit vs. reference, and upstream of poly-A tail. Each point corresponds to one transcriptome assembly, grouped by taxonomic family (columns). (**B**) Scatterplots for standard deviation of 3'-UTR lengths. (**C)** Histograms for 3'-UTR lengths, colored by putative stop codon (UGA: blue, UAA: orange, UAG: green), one representative library per family.


Karyorelictea
Heterotrichea
A
Kentrophoridae
Trachelocercidae
Geleiidae
Loxodidae
Spirostomidae
Condylostomidae
Stentoridae
Blepharismidae
Climacostomidae
Fabreidae
UGA UAA UAG
UGA UAA
UGA UAG
UAA UAG
UGA
UAA
UAG
0.5
1.0
Fraction of shortlisted transcripts with putative stop
B
10^1
10^2
10^3
Std dev. of 3'-UTR lengths (bp)
C
Karyorelictea
N10
N26
gacu_yy
N4
Permitted stop codons
Codon
UGA
UAA
UAG
0
100
200
3'-UTR length (bp)
Heterotrichea
N70
cmag_mm
spol_upp
bjap_mm
cl01_yy
fsal_mm

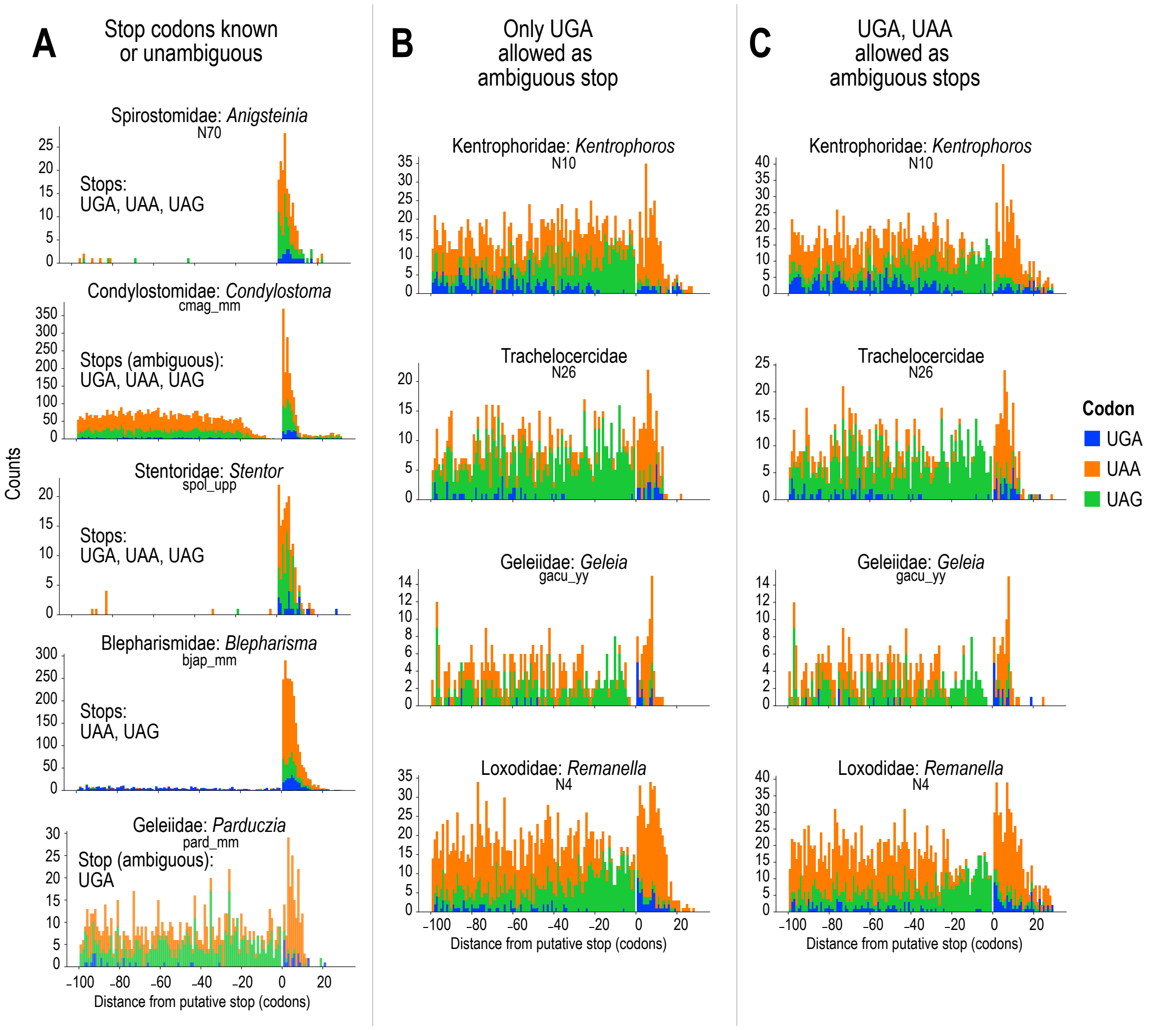


Figure 5. Depletion of in-frame coding "stop" codons in the coding sequence (negative coordinates) immediately before the putative true stop codon (position 0) and their enrichment in the 3'-UTR (positive coordinates). Representative library with highest number of assembled full length contigs chosen per taxon. (**A**) Codon counts for UGA (blue), UAA (orange), and UAG (green) before and after putative true stop in *Condylostoma magnum* (uses all three as ambiguous stops), and three heterotrichs with unambiguous stops. (**B**) Codon counts for karyorelicts if only UGA is permitted as a stop codon. (**C**) Codon counts for karyorelicts if both UGA and UAA are permitted as stop codons.

Discussion

We have found evidence that the codon UGA is used as both a stop codon and to code for tryptophan by karyorelictean ciliates. The taxa sampled represent four of the six families of karyorelicts: Loxodidae, Trachelocercidae, Geleiidae, and Kentrophoriidae. When this distribution of genetic codes is mapped to an up-to-date phylogeny [20], we can infer that the ambiguous code formerly reported only for *Parduczia* sp. (Geleiidae) among ciliates was actually acquired at the root of the karyorelict clade (Figure 6).

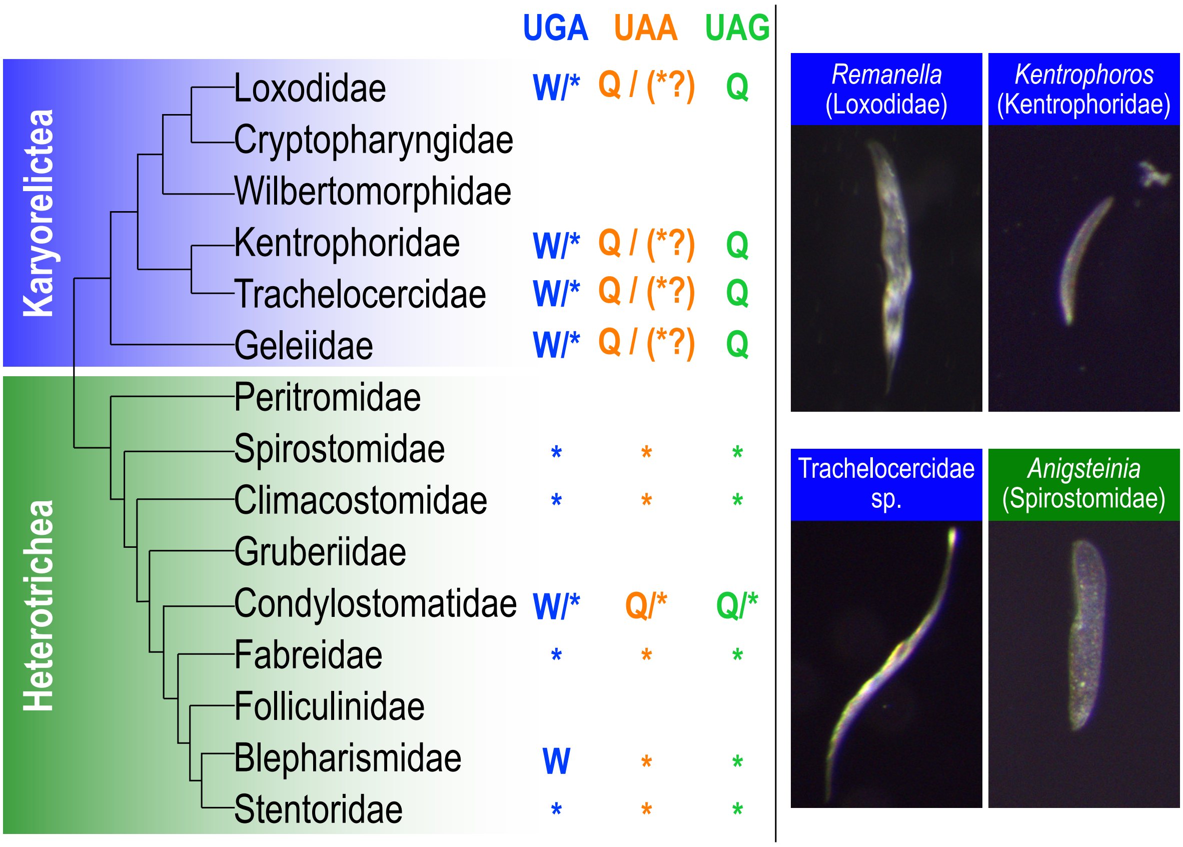


Figure 6. Genetic code diversity among karyorelict and heterotrich ciliates. (**Left**) Diagrammatic karyorelict + heterotrich tree with predicted stop codon reassignments mapped to each family. Subtree topologies are from refs. [20] and [25] respectively. Branch lengths are not representative of evolutionary distances. (**Right**) Photomicrographs of ciliates (incident light) collected in this study from Roscoff, France; height of each panel 50 µm.

Available data for *Cryptopharynx* (Karyorelictea: Cryptopharyngidae) were not conclusive. The canonical stop codons had frequencies between 0.02 and 0.07%, lower than for other karyorelicts, but higher than true stop codons, but Cryptopharyngidae was represented by a single library that had high contamination from other eukaryotes (Appendix) and there were too few high-confidence, full length transcripts for a reliable conclusion on its genetic code. No sequence data beyond rRNA genes were publicly available for the remaining family, the monotypic Wilbertomorphidae, whose phylogenetic position in relation to the other karyorelicts is unclear because of long branch lengths, and which has to our knowledge only been reported once [24].

Ambiguous stop/sense codes are hence not just isolated phenomena, but are used by a major taxon that is diverse, globally distributed, and common in its respective habitats. In contrast, the heterotrichs, which constitute the sister group to Karyorelictea and are hence of the same evolutionary age, use at least three different genetic codes, including one with ambiguous stops (Figure 6). If organisms with ambiguous codes were isolated single species whose nearest relatives have conventional stops, as appears to be the case for *Blastocrithidia* spp. and *Amoebophrya* sp., we might conclude that these are uncommon occurrences that do not persist over longer evolutionary time scales. However, the karyorelict crown group diversified during the Proterozoic (posterior mean 455 Mya) and the stem split from the Heterotrichea even earlier, in the Neo-Proterozoic [25].

This study has benefited from several technical improvements. A highly complete, contiguous genome assembly with gene predictions is now available for the heterotrich *Blepharisma stoltei* [26]. Because *Blepharisma* is more closely related to the karyorelicts than other ciliate model species, which are mostly oligohymenophorans and spirotrichs, it improved the reference-based annotation of the assembled transcriptomes. Single-cell RNA-seq libraries in this study were also sequenced to a greater depth, with a lower fraction of contamination from rRNA, and hence yielded more full length mRNA transcripts for analysis.

One proposed mechanism for how the cell correctly recognizes whether an ambiguous codon is coding or terminal is based on the proximity of translation stops to the poly-A tail of transcripts. In this model, tRNAs typically bind more efficiently to in-frame coding “stops” than eukaryotic translation release factor 1 (eRF1), hence allowing these codons to be translated. At the true termination stop codon, however, the binding of eRF1 can be stabilized by interactions with poly-A interacting proteins like PABP bound to the nearby poly-A tail, allowing it to outcompete tRNAs and hydrolyze the peptidyl-tRNA bond [10,27]. Consistent with this model, we found that karyorelict 3’-UTRs are also relatively short, and that in-frame UGAs are depleted immediately before the putative true stop codon. Nonetheless, karyorelict 3’-UTRs are actually about 10 nt longer on average than those of heterotrichs.

Our results raised the possibility that UAA is also used as an ambiguous stop codon for ~1-10% of karyorelict transcripts, in addition to the main stop codon UGA. eRF1 may retain a weak affinity for UAA, and recognize UAA for terminating translation albeit with lower efficiency. In *Blepharisma japonicum*, where UAA and UAG are non-ambiguous stops and UGA encodes tryptophan (albeit at low frequency, 0.13%), heterologously expressed eRF1 could still recognize all three codons in an in vitro assay, although efficiency of peptidyl-tRNA hydrolysis was lower with UGA than for UAA and UAG [28]. In species with non-ambiguous stop codon reassignment, the effect of such “weak” ambiguity on the total pool of translated protein may be negligible, but it shows that there is a latent potential that could account for the repeated evolution of stop codon reassignments in ciliates. Furthermore, UAAs were even more abundant than UGAs in ciliate 3’-UTRs, which can be attributed to the low GC% of 3’-UTRs compared to coding sequences; other A/U-only codons were also enriched in 3’-UTRs. Therefore, UAAs in the 3’-UTRs of karyorelicts may be a “backstop” mechanism that prevents occasional stop-codon readthrough, as proposed for tandem stop codons (TSCs) in other species with reassigned stop codons [29]. In the minority of transcripts where in-frame stop-UGA is absent, the backstop may be adequate to terminate translation before the poly-A tail and produce a functional protein most of the time. To verify our predictions that UGA is the main stop codon and UAA a lower-frequency alternative stop, ribosome profiling and mass spectrometry detection of peptide fragments corresponding to the expected 3’-ends of coding sequences, e.g. as performed on *Condylostoma* [10], are the most applicable experimental methods. If a karyorelict species can be developed into a laboratory model amenable to genetic transformation, manipulation of the 3’-UTR length and sequence would allow us to test the “backstop” hypothesis directly and tease apart the factors contributing to translation termination in these organisms.

What selective pressures might favor the evolution and maintenance of an ambiguous genetic code? One possibility is that context-dependent sense/stop codons confer mutational robustness by eliminating substitutions that cause premature stop codons. Ambiguous codes do not appear to be linked to a specific habitat: *Blastocrithidia* spp. and *Amoebophrya* sp. are both parasites of eukaryotic hosts, but of insects and free-living dinoflagellates respectively; whereas the karyorelict ciliates and *Condylostoma* are both found in marine interstitial environments, but live alongside other ciliates that have conventional codes, such as *Anigsteinia*. Having short 3'-UTRs may predispose ciliates to adopt ambiguous codes by facilitating interactions between eRF1 and PABPs that could enable stop recognition, but other factors, including simply contingent evolution, appear to have led to their evolution because the 3'-UTRs of ciliates with conventional stop codons are also comparably short, particularly among the heterotrichs.

Any adaptationist hypothesis for alternative and ambiguous codes will have to contend with the existence of related organisms with conventional codes that have similar lifestyles. Furthermore, once a stop codon has been reassigned to sense, it becomes increasingly difficult to undo without the deleterious effects of premature translation termination, and may function like a ratchet. Like the origins of the genetic code itself [30], we may have to be content with the null hypothesis that they are "frozen accidents" that reached fixation stochastically, and which are maintained because they do not pose a significant selective disadvantage.

Materials and Methods

Sample collection

Surface sediment was sampled in September 2021 from two sites in the bay at Roscoff, France when exposed at low tide. Site A: shallow swimming enclosure, 48.72451 N, 3.992294 W; Site B: adjacent to green algae tufts near freshwater outflow, 48.716169 N, 3.995626 W. Upper 1-2 cm of sediment was skimmed into glass beakers, and stored under local seawater until use. Interstitial ciliates were collected by decantation: a spoonful of sediment was stirred in seawater in a beaker. Sediment particles were briefly allowed to settle out, and the overlying suspended organic material was decanted into Petri dishes. Ciliate cells were preliminarily identified by morphology under a dissection microscope and picked by pipetting with sterile, filtered pipette tips. Selected cells were imaged with incident light under a stereo microscope (Olympus SZX10, Lumenera Infinity 3 camera).

NEBNext cell lysis buffer (NEB, E5530S) was premixed and filled into PCR tubes; per tube: 0.8 µL 10x cell lysis buffer, 0.4 µL murine RNAse inhibitor, 5.3 µL nuclease-free water. Picked ciliate cells were transferred twice through filtered local seawater (0.22 µm, Millipore SLGP033RS) to wash, then transferred with 1.5 µL carryover volume to 6.5 µL of cell lysis buffer (final volume 8 µL), and snap frozen in liquid nitrogen. Samples were stored at -80 °C before use.

Single-cell RNAseq sequencing

Samples collected in cell lysis buffer (Table_S1.xlsx in [22]) were used for RNAseq library preparation with the NEBNext Single Cell / Low Input RNA Library Prep Kit for Illumina (NEB, E6420S), following the manufacturer's protocol for single cells, with the following parameters adjusted: 17 cycles for cDNA amplification PCR, cDNA input for library enrichment normalized to 3 ng (or all available cDNA used for libraries where total cDNA was <3 ng), 8 cycles for library enrichment PCR. Libraries were dual-indexed (NEBNext Dual Index Primers Set 1, NEB E7600S), and sequenced on an Illumina NextSeq 2000 instrument with P3 300 cycle reagents, with target yield of 10 Gbp per library.

RNA-seq library quality control and transcriptome assembly

Previously published karyorelict transcriptome data [17,31–33] were downloaded from the European Nucleotide Archive (ENA) (Table_S1.xlsx in [22]). Contamination from non-target organisms was evaluated by mapping reads to an rRNA reference database and summarizing the hits by taxonomy. Although RNAseq library construction enriches mRNAs using poly-A tail selection, there is typically still sufficient rRNA present in the final library to evaluate the taxonomic composition of the sample. All RNAseq read libraries (newly sequenced and previously published) were processed with the same pipeline: The taxonomic composition of each library was evaluated by mapping 1 M read pairs per library against the SILVA SSU Ref NR 132 database [34], using phyloFlash v3.3b1 [35]. Newly sequenced libraries were assigned to a genus or family using the mapping-based taxonomic summary, or full-length 18S rRNA gene if it was successfully assembled.

Reads were trimmed with the program bbduk.sh (https://sourceforge.net/projects/bbmap/, BBmap v38.22) to remove known adapters (right end) and low-quality bases (both ends), with minimum Phred quality 24 and minimum read length 25 bp . Trimmed reads were then assembled with Trinity v2.12.0 [36] using default parameters. Assembled contigs were aligned against the *Blepharisma stoltei* ATCC 30299 proteome [26] with NCBI Blastx v2.12.0 [37] using the standard genetic code and E-value cutoff 10-20, parallelized with GNU Parallel [38].

Morphological identifications of the newly collected samples were verified with 18S rRNA sequences from the Trinity transcriptome assemblies (Appendix). rRNA sequences were annotated with barrnap v0.9. 18S rRNA sequences ≥80% of full length were extracted, except for two libraries (N4, N26) where the longest sequences were <80% and for which the two longest 18S rRNA sequences were extracted instead. For comparison, reference sequences for Karyorelictea and Heterotrichea above 1400 bp from the PR2 database v4.14.0 [39] were used. Representative reference sequences were chosen by clustering at 99% identity with the cluster_fast method using Vsearch v2.13.6 [40]. Extracted and reference sequences were aligned with MAFFT v7.505 [41]. A phylogeny (Figure S3) was inferred from the alignment with IQ-TREE v2.0.3 [42], using the TIM2+F+I+G4 model found as the best-fitting model by ModelFinder [43]. Alignment and tree files are available from [44]. 18S rRNA sequences were deposited in the European Nucleotide Archive under accessions OX095806-OX095846.

Read pre-processing, quality control, and assembly were managed with a Snakemake v6.8.1 [45] workflow https://github.com/Swart-lab/karyocode-workflow [46]. Scripts for data processing described below were written in Python v3.7.3 using Biopython v1.74 [47], pandas v0.25.0 [48], seaborn v0.11.0 [49] and Matplotlib v3.1.1 [50] libraries unless otherwise stated.

Prediction of stop codon reassignment to sense

Only contigs with poly-A tails ≥7 bp were used for genetic code prediction, to exclude potential bacterial contaminants, especially because several species (*Kentrophoros* spp., *Parduczia* sp., Appendix) are known to have abundant bacterial symbionts. Presence and lengths of poly-A tails in assembled transcripts were evaluated with a Python regular expression. Library preparation was not strand-specific, hence contigs starting with poly-T were reverse-complemented, and contigs with both a poly-A tail and a poly-T head (presumably fused contig) were excluded.

Codon frequencies and their corresponding amino acids were predicted with an updated version of PORC v2.1 https://github.com/Swart-lab/PORC [51]; managed with a Snakemake workflow https://github.com/Swart-lab/karyocode-analysis-porc [52]; the method has been previously described [10,53]. Briefly: a six-frame translation was produced for each contig in the transcriptome assembly, and searched against conserved domains in the Pfam-A database v32 [54] with hmmscan from HMMer v3.3.2 (http://hmmer.org/). Overall codon frequencies were counted from alignments with E-value ≤ 10-20. To ensure that there was sufficient data underlying the codon and amino acid frequencies, only those libraries with at least 100 observations for each of the coding codons in the standard genetic code were used for comparison of codon frequencies and for prediction of amino acid assignments.

Frequencies of amino acids aligning to a given codon were counted from columns where the HMM model consensus was ≥50% identity in the alignment used to build the model (upper-case positions in the HMM consensus). Sequence logos of amino acid frequencies per codon for each library were drawn with Weblogo v3.7.5 [55].

In addition to the transcriptomes, genomic CDSs of selected model species with different genetic codes [26,56–60] were also analyzed with PORC to obtain a reference baseline of coding-codon frequencies (Table_S2.xlsx in [22]). These model species have non-ambiguous codes so they were not expected to have stop codons in the CDSs, except for the terminal stop.

Prediction of coding frame in full-length transcripts

"Full-length" transcripts (with poly-A tail, intact 3'-UTR, and complete coding sequence) were desirable to predict the stop codon, characterize 3'-UTR metrics, and verify genetic code predictions. Contigs were therefore filtered with the following criteria: (i) poly-A tail ≥7 bp, criterion following [10], (ii) contig contains a Blastx hit vs. *Blepharisma stoltei* protein sequence with E-value ≤10^{-20} and where the alignment covers ≥80% of the reference *B. stoltei* sequence, (iii) both poly-A tail and Blastx hit agree on the contig orientation. For contigs with multiple isoforms assembled by Trinity, the isoform with the longest Blastx hit was chosen; in case of a Blastx hit length tie, then the longer isoform was chosen. Only libraries with >100 assembled "full-length" transcripts were used for downstream analyses (Appendix).

Metrics for evaluating potential stop codon combinations

For each of the 7 possible combinations of the 3 canonical stop codons (UGA, UAA, UAG), we treated the first in-frame stop downstream of the Blastx hit in each full-length transcript (including the last codon of the hit) as the putative stop codon, and recorded the number of full-length transcripts with a putative stop, the length of the 3'-UTR (distance from stop to beginning of the poly-A tail), as well as the codon frequencies for each position from 150 codons upstream of the putative stop to the last in-frame three-nucleotide triplet before the poly-A tail.

Delimitation of putative coding sequences using Blastx hits

The start codon was more difficult to evaluate because the 5' end of the transcript may not have been fully assembled, and there was no straightforward way to recognize its boundaries, unlike the 3'-poly-A tail. We used the following heuristic criteria to define the start of the CDS: first in-frame ATG upstream of the Blastx hit (including first codon of the hit), or first in-frame stop codon encountered upstream (to avoid potential problems with ORFs containing in-frame stops), whichever comes first. Otherwise, the transcript was assumed to be incomplete at the 5'-end and simply truncated with the required 1 or 2 bp offset to keep the CDS in frame.

Verification of in-frame UGAs in conserved marker genes

Full-length CDSs (see above) were translated with the karyorelict code (NCBI table 27). Conserved marker genes were identified with BUSCO v5.2.2 (protein mode, alveolata_odb10 marker set) [23], managed with a Snakemake workflow https://github.com/Swart-lab/karyocode-analysis-busco [61]. Markers for additional ciliate species where relatively complete genome assemblies and gene predictions were available were also identified (Table_S3.xlsx in [22]) [57,59,62–69]. For each BUSCO marker, the ciliate homologs were aligned with Muscle v3.8.1551 [70]. Alignment columns corresponding to in-frame putatively coding UGAs of karyorelict sequences were identified. These positions were considered to be conserved if ≥50% of residues were W or another aromatic amino acid (Y, F, or H).

Acknowledgements

A preprint version of this article has been peer-reviewed and recommended by Peer Community In Genomics (https://doi.org/10.24072/pci.genomics.100019). We thank S. Booker, C. Cabresin, E. Bourrigaud, and R. Garnier for facilitating our research visit to the Station Biologique Roscoff; H. Budde for sequencing; H. Gruber-Vodicka for loan of equipment; A. Peters for helpful suggestions and cake; A.

Lemper for administrative assistance; Y. Shulgina for thoughtful feedback on the manuscript; the recommender and peer reviewers for PCI Genomics for their reviews; and members of the Swart Lab for helpful feedback and support.

Data, scripts, and codes availability

RNA-seq libraries sequenced for this study have been deposited at the European Nucleotide Archive (https://www.ebi.ac.uk/ena/) under accession PRJEB50648. Lists of dataset accessions for each analysis [22] and the 18S rRNA phylogeny [44] have been deposited at Edmond. Software and computational workflows have been deposited at Zenodo [46, 51, 52, 61].

Conflict of interest disclosure

The authors declare that they have no conflict of interest relating to the content of this article.

Funding

This work was supported by the Max Planck Society. The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 730984, Assemble Plus project.

References

1. Ivanova NN, Schwientek P, Tripp HJ, Rinke C, Pati A, Huntemann M, et al. Stop codon reassignments in the wild. Science. 2014;344: 909–913. https://doi.org/10.1126/science.1250691
2. Shulgina Y, Eddy SR. A computational screen for alternative genetic codes in over 250,000 genomes. eLife. 2021;10. https://doi.org/10.7554/eLife.71402
3. Kollmar M, Mühlhausen S. Nuclear codon reassignments in the genomics era and mechanisms behind their evolution. Bioessays. 2017;39. https://doi.org/10.1002/bies.201600221
4. Schueren F, Thoms S. Functional translational readthrough: A systems biology perspective. PLoS Genet. 2016;12: e1006196. https://doi.org/10.1371/journal.pgen.1006196
5. Suzuki T, Ueda T, Watanabe K. The “polysemous” codon--a codon with multiple amino acid assignment caused by dual specificity of tRNA identity. EMBO J. 1997;16: 1122–1134. https://doi.org/10.1093/emboj/16.5.1122
6. Hatfield DL, Gladyshev VN. How selenium has altered our understanding of the genetic code. Mol Cell Biol. 2002;22: 3565–3576. https://doi.org/10.1128/MCB.22.11.3565-3576.2002
7. Záhonová K, Kostygov AY, Ševčíková T, Yurchenko V, Eliáš M. An Unprecedented Non-canonical Nuclear Genetic Code with All Three Termination Codons Reassigned as Sense Codons. Curr Biol. 2016;26: 2364–2369. https://doi.org/10.1016/j.cub.2016.06.064
8. Bachvaroff TR. A precedented nuclear genetic code with all three termination codons reassigned as sense codons in the syndinean *Amoebophrya* sp. ex *Karlodinium veneficum*. PLoS ONE. 2019;14: e0212912. https://doi.org/10.1371/journal.pone.0212912
9. Heaphy SM, Mariotti M, Gladyshev VN, Atkins JF, Baranov PV. Novel Ciliate Genetic Code Variants Including the Reassignment of All Three Stop Codons to Sense Codons in *Condylostoma magnum*. Mol Biol Evol. 2016;33: 2885–2889. https://doi.org/10.1093/molbev/msw166
10. Swart EC, Serra V, Petroni G, Nowacki M. Genetic Codes with No Dedicated Stop Codon: Context-Dependent Translation Termination. Cell. 2016;166: 691–702. https://doi.org/10.1016/j.cell.2016.06.020
11. Lozupone CA, Knight RD, Landweber LF. The molecular basis of nuclear genetic code change in ciliates. Curr Biol. 2001;11: 65–74. https://doi.org/10.1016/s0960-9822(01)00028-8
12. Preer JR, Preer LB, Rudman BM, Barnett AJ. Deviation from the universal code shown by the gene for surface protein 51A in *Paramecium*. Nature. 1985;314: 188–190. https://doi.org/10.1038/314188a0

13. Horowitz S, Gorovsky MA. An unusual genetic code in nuclear genes of *Tetrahymena*. Proc Natl Acad Sci USA. 1985;82: 2452–2455. https://doi.org/10.1073/pnas.82.8.2452
14. Meyer F, Schmidt HJ, Plümper E, Hasilik A, Mersmann G, Meyer HE, et al. UGA is translated as cysteine in pheromone 3 of *Euplotes octocarinatus*. Proc Natl Acad Sci USA. 1991;88: 3758–3761. https://doi.org/10.1073/pnas.88.9.3758
15. Tourancheau AB, Tsao N, Klobutcher LA, Pearlman RE, Adoutte A. Genetic code deviations in the ciliates: evidence for multiple and independent events. EMBO J. 1995;14: 3262–3267. https://doi.org/10.1002/j.1460-2075.1995.tb07329.x
16. Helftenbein E. Nucleotide sequence of a macronuclear DNA molecule coding for alpha-tubulin from the ciliate *Stylonychia lemnae*. Special codon usage: TAA is not a translation termination codon. Nucleic Acids Res. 1985;13: 415–433. https://doi.org/10.1093/nar/13.2.415
17. Yan Y, Maurer-Alcalá XX, Knight R, Kosakovsky Pond SL, Katz LA. Single-Cell Transcriptomics Reveal a Correlation between Genome Architecture and Gene Family Evolution in Ciliates. MBio. 2019;10. https://doi.org/10.1128/mBio.02524-19
18. Raikov IB. Primitive never-dividing macronuclei of some lower ciliates. Int Rev Cytol. 1985;95: 267–325. https://doi.org/10.1016/s0074-7696(08)60584-7
19. Fenchel T. The ecology of marine microbenthos IV. Structure and function of the benthic ecosystem, its chemical and physical factors and the microfauna commuities with special reference to the ciliated protozoa. Ophelia. 1969;6: 1–182. https://doi.org/10.1080/00785326.1969.10409647
20. Ma M, Li Y, Maurer-Alcalá XX, Wang Y, Yan Y. Deciphering phylogenetic relationships in class Karyorelictea (Protista, Ciliophora) based on updated multi-gene information with establishment of a new order Wilbertomorphida n. ord. Mol Phylogenet Evol. 2022; 107406. https://doi.org/10.1016/j.ympev.2022.107406
21. Foissner W. The karyorelictids (Protozoa: Ciliophora), a unique and enigmatic assemblage of marine, interstitial ciliates: a review emphasizing ciliary patterns and evolution. Evolutionary relationships among Protozoa. 1998; 305–325.
22. Seah BKB, Singh A, Swart EC. Dataset accessions for comparative analysis of ciliate genetic codes, V1. Edmond. 2022. https://doi.org/10.17617/3.XWMBKT
23. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2018;35: 543–548. https://doi.org/10.1093/molbev/msx319
24. Xu Y, Li J, Song W, Warren A. Phylogeny and establishment of a new ciliate family, Wilbertomorphidae fam. nov. (Ciliophora, Karyorelictea), a highly specialized taxon represented by *Wilbertomorpha colpoda* gen. nov., spec. nov. J Eukaryot Microbiol. 2013;60: 480–489. https://doi.org/10.1111/jeu.12055
25. Fernandes NM, Schrago CG. A multigene timescale and diversification dynamics of Ciliophora evolution. Mol Phylogenet Evol. 2019;139: 106521. https://doi.org/10.1016/j.ympev.2019.106521
26. Singh M, Seah BKB, Emmerich C, Singh A, Woehle C, Huettel B, et al. The *Blepharisma stoltei* macronuclear genome: towards the origins of whole genome reorganization. BioRxiv. 2021. https://doi.org/10.1101/2021.12.14.471607
27. Alkalaeva E, Mikhailova T. Reassigning stop codons via translation termination: How a few eukaryotes broke the dogma. Bioessays. 2017;39. https://doi.org/10.1002/bies.201600213
28. Eliseev BD, Alkalaeva EZ, Kryuchkova PN, Lekomtsev SA, Wang W, Liang A-H, et al. Translation termination factor eRF1 of the ciliate *Blepharisma japonicum* recognizes all three stop codons. Mol Biol (NY). 2011;45: 614–618. https://doi.org/10.1134/S0026893311040030
29. Fleming I, Cavalcanti ARO. Selection for tandem stop codons in ciliate species with reassigned stop codons. PLoS ONE. 2019;14: e0225804. https://doi.org/10.1371/journal.pone.0225804
30. Crick FH. The origin of the genetic code. J Mol Biol. 1968;38: 367–379. https://doi.org/10.1016/0022-2836(68)90392-6
31. Seah BKB, Antony CP, Huettel B, Zarzycki J, Schada von Borzyskowski L, Erb TJ, et al. Sulfur-Oxidizing Symbionts without Canonical Genes for Autotrophic CO2 Fixation. MBio. 2019;10: e01112-19. https://doi.org/10.1128/mBio.01112-19

32. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, 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. 2014;12: e1001889. https://doi.org/10.1371/journal.pbio.1001889
33. Kolisko M, Boscaro V, Burki F, Lynn DH, Keeling PJ. Single-cell transcriptomics for microbial eukaryotes. Curr Biol. 2014;24: R1081-2. https://doi.org/10.1016/j.cub.2014.10.026
34. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41: D590-6. https://doi.org/10.1093/nar/gks1219
35. Gruber-Vodicka HR, Seah BKB, Pruesse E. phyloFlash: Rapid Small-Subunit rRNA Profiling and Targeted Assembly from Metagenomes. mSystems. 2020;5. https://doi.org/10.1128/mSystems.00920-20
36. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29: 644–652. https://doi.org/10.1038/nbt.1883
37. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421. https://doi.org/10.1186/1471-2105-10-421
38. Tange O. Gnu Parallel 2018. Zenodo. 2018. https://doi.org/10.5281/zenodo.1146014
39. Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner 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. 2013;41: D597-604. https://doi.org/10.1093/nar/gks1160
40. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4: e2584. https://doi.org/10.7717/peerj.2584
41. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30: 772–780. https://doi.org/10.1093/molbev/mst010
42. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37: 1530–1534. https://doi.org/10.1093/molbev/msaa015
43. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14: 587–589. https://doi.org/10.1038/nmeth.4285
44. Seah BKB, Singh A, Swart EC. 18S rRNA phylogeny for ciliate single cell environmental samples, V1. Edmond. 2022. https://doi.org/10.17617/3.QLWR38
45. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Res. 2021;10: 33. https://doi.org/10.12688/f1000research.29032.2
46. Seah BKB. Swart-lab/karyocode-workflow. Zenodo. 2022. https://doi.org/10.5281/zenodo.6647650
47. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25: 1422–1423. https://doi.org/10.1093/bioinformatics/btp163
48. McKinney W. Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference. SciPy; 2010. pp. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a
49. Waskom M. seaborn: statistical data visualization. JOSS. 2021;6: 3021. https://doi.org/10.21105/joss.03021
50. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007;9: 90–95. https://doi.org/10.1109/MCSE.2007.55
51. Swart EC, Seah BKB. Swart-lab/PORC. Zenodo. 2022. https://doi.org/10.5281/zenodo.6784075
52. Seah BKB Swart-lab/karyocode-analysis-porc. Zenodo. 2022. https://doi.org/10.5281/zenodo.6647652
53. Dutilh BE, Jurgelenaite R, Szklarczyk R, van Hijum SAFT, Harhangi HR, Schmid M, et al. FACIL: Fast and Accurate Genetic Code Inference and Logo. Bioinformatics. 2011;27: 1929–1933. https://doi.org/10.1093/bioinformatics/btr316

54. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49: D412–D419. https://doi.org/10.1093/nar/gkaa913
55. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14: 1188–1190. https://doi.org/10.1101/gr.849004
56. Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, et al. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature. 2006;444: 171–178. https://doi.org/10.1038/nature05230
57. Slabodnick MM, Ruby JG, Reiff SB, Swart EC, Gosai S, Prabakaran S, et al. The Macronuclear Genome of *Stentor coeruleus* Reveals Tiny Introns in a Giant Cell. Curr Biol. 2017;27: 569–575. https://doi.org/10.1016/j.cub.2016.12.057
58. Arnaiz O, Meyer E, Sperling L. ParameciumDB 2019: integrating genomic data across the genus for functional and evolutionary biology. Nucleic Acids Res. 2020;48: D599–D605. https://doi.org/10.1093/nar/gkz948
59. Sheng Y, Duan L, Cheng T, Qiao Y, Stover NA, Gao S. The completed macronuclear genome of a model ciliate *Tetrahymena thermophila* and its application in genome scrambling and copy number analyses. Sci China Life Sci. 2020;63: 1534–1542. https://doi.org/10.1007/s11427-020-1689-4
60. Stover NA, Krieger CJ, Binkley G, Dong Q, Fisk DG, Nash R, et al. Tetrahymena Genome Database (TGD): a new genomic resource for *Tetrahymena thermophila* research. Nucleic Acids Res. 2006;34: D500-3.https://doi.org/10.1093/nar/gkj054
61. Seah BKB. Swart-lab/karyocode-analysis-busco. Zenodo. 2022. https://doi.org/10.5281/zenodo.6647679
62. Chen X, Jiang Y, Gao F, Zheng W, Krock TJ, Stover NA, et al. Genome analyses of the new model protist *Euplotes vannus* focusing on genome rearrangement and resistance to environmental stressors. Mol Ecol Resour. 2019;19: 1292–1308. https://doi.org/10.1111/1755-0998.13023
63. Coyne RS, Hannick L, Shanmugam D, Hostetler JB, Brami D, Joardar VS, et al. Comparative genomics of the pathogenic ciliate *Ichthyophthirius multifiliis*, its free-living relatives and a host species provide insights into adoption of a parasitic lifestyle and prospects for disease control. Genome Biol. 2011;12: R100. https://doi.org/10.1186/gb-2011-12-10-r100
64. Swart EC, Bracht JR, Magrini V, Minx P, Chen X, Zhou Y, et al. The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. PLoS Biol. 2013;11: e1001473. https://doi.org/10.1371/journal.pbio.1001473
65. Xiong J, Wang G, Cheng J, Tian M, Pan X, Warren A, et al. Genome of the facultative scuticociliatosis pathogen *Pseudocohnilembus persalinus* provides insight into its virulence through horizontal gene transfer. Sci Rep. 2015;5: 15470. https://doi.org/10.1038/srep15470
66. Aeschlimann SH, Jönsson F, Postberg J, Stover NA, Petera RL, Lipps H-J, et al. The draft assembly of the radically organized *Stylonychia lemnae* macronuclear genome. Genome Biol Evol. 2014;6: 1707–1723. https://doi.org/10.1093/gbe/evu139
67. Hamilton EP, Kapusta A, Huvos PE, Bidwell SL, Zafar N, Tang H, et al. Structure of the germline genome of *Tetrahymena thermophila* and relationship to the massively rearranged somatic genome. eLife. 2016;5. https://doi.org/10.7554/eLife.19090
68. McGrath CL, Gout J-F, Doak TG, Yanagi A, Lynch M. Insights into three whole-genome duplications gleaned from the *Paramecium caudatum* genome sequence. Genetics. 2014;197: 1417–1428. https://doi.org/10.1534/genetics.114.163287
69. Arnaiz O, Van Dijk E, Bétermier M, Lhuillier-Akakpo M, de Vanssay A, Duharcourt S, et al. Improved methods and resources for *Paramecium* genomics: transcription units, gene annotation and gene expression. BMC Genomics. 2017;18: 483. https://doi.org/10.1186/s12864-017-3887-z
70. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32: 1792–1797. https://doi.org/10.1093/nar/gkh340
71. Yan Y, Xu Y, Al-Farraj SA, Al-Rasheid KAS, Song W. Morphology and phylogeny of three trachelocercids (Protozoa, Ciliophora, Karyorelictea), with description of two new species and insight into the evolution of the family Trachelocercidae. Zool J Linn Soc. 2016;177: 306–319. https://doi.org/10.1111/zoj.12364

72. Xu Y, Gao S, Hu X, Al-Rasheid KAS, Song W. Phylogeny and systematic revision of the karyorelictid genus *Remanella* (Ciliophora, Karyorelictea) with descriptions of two new species. Eur J Protistol. 2013;49: 438–452. https://doi.org/10.1016/j.ejop.2012.12.001

Appendix

Quality metrics of single-cell transcriptome assemblies

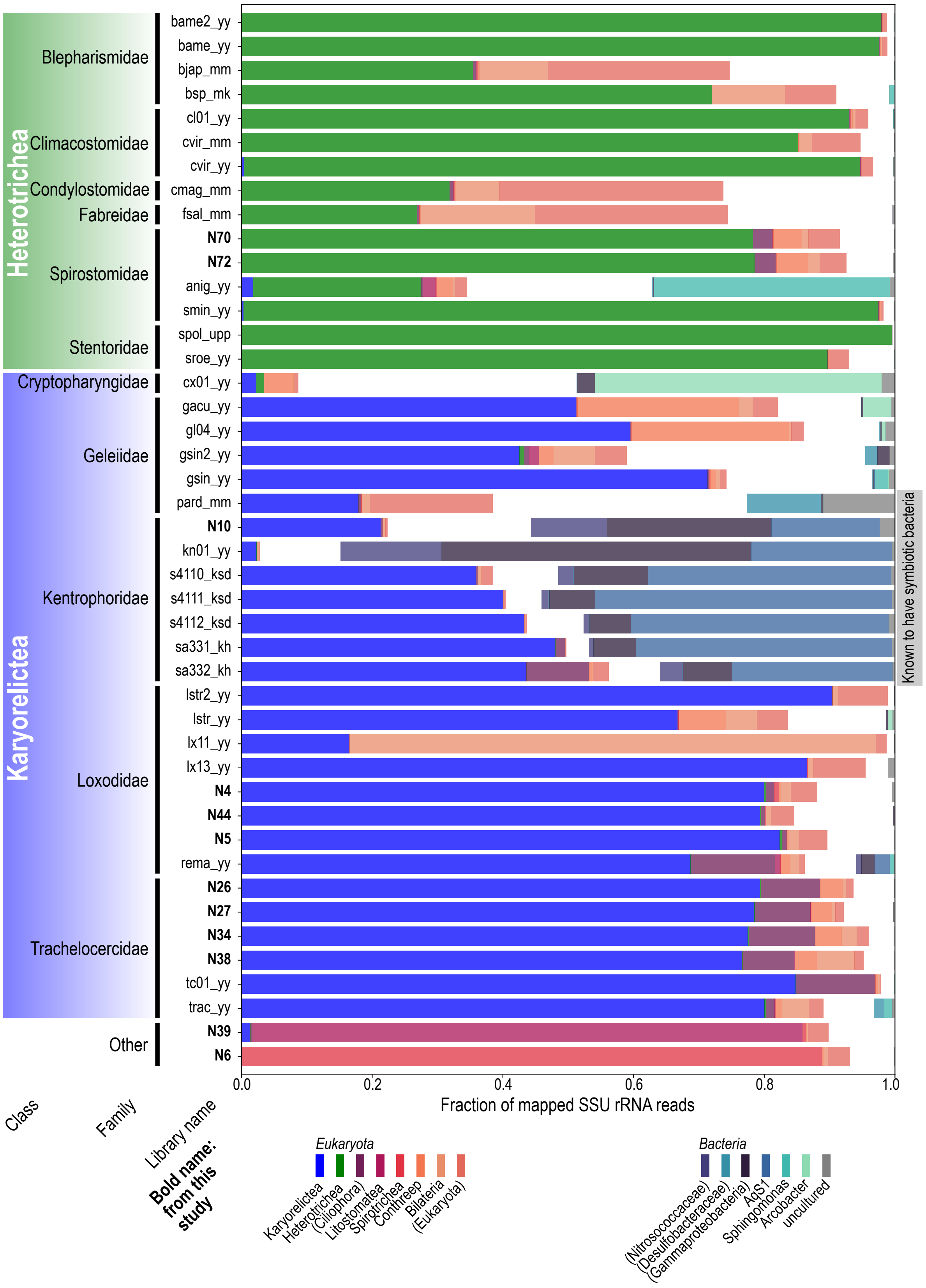
Taxonomic composition was variable between samples. Samples with the lowest contamination from non-target taxa were those from cultured isolates, while single-cell environmental samples newly sequenced for this study had consistently about 80% of rRNA reads assigned to the expected target taxon (Figure S1). Furthermore, the proportion of the library composed of SSU rRNA reads was also relatively low in the newly sequenced samples (0.03 to 1.7%). The number of contigs per assembly was highly variable (15389 to 162815), but when only contigs with poly-A tails ≥7 bp were counted, samples from this study had more poly-A tailed contigs (4135 to 13427), compared to assemblies from previously published libraries (48 to 4932) (Figure S2). Presence of poly-A tails can be used to exclude bacterial and rRNA sequences. Contigs with both a poly-A tail and a putatively full-length coding sequence were most abundant for four heterotrich libraries that were prepared from bulk cultured cells instead of single cells (Figure S2, Table_S1.xlsx in [22]). As heterotrichs they were also more closely related to the species used for the reference protein set (*Blepharisma stoltei*).

Different filtering criteria were used to shortlist transcriptome assemblies for prediction of stop codon reassignment to sense vs. prediction of the actual stop codon usage. All ten newly sequenced karyorelict and heterotrich libraries from this study were shortlisted for both analyses. Of the 33 previously published libraries, 15 were used for the former and 16 for the latter (Table_S1.xlsx in [22]).

Confirmation of phylogenetic identity with 18S rRNA sequences

During collection, each sample was preliminarily identified to a family or genus by morphology under the dissection microscope. Two libraries (N6, N39) were found to be neither karyorelicts nor heterotrichs during the initial screen with phyloFlash (Figure S1). The morphology-based identification of the remaining sequences was verified by a tree of 18S rRNA sequences from the transcriptome assembly (≥80% full length) alongside reference sequences (Figure S3, Table_S1.xlsx in [22]). Trachelocercidae spp. (samples N26, N27, N34, N38) could not be identified more specifically to genus, because the taxonomy of several reference sequences were only to family level, and some genera also do not appear monophyletic with 18S rRNA phylogeny [71]. *Remanella* may also be paraphyletic [72] but we chose to retain the name *Remanella* for our samples (N4, N5, N44) because there are only two genera in the family Loxodidae, and the marine species have conventionally been designated *Remanella*.

Figure S1 (next page). Taxonomic composition of RNAseq libraries, derived from mapping of reads to the SILVA SSU rRNA database, summarized at class level. Only taxa comprising ≥10% of the total in at least one library are shown. Bars representing eukaryotic taxa are aligned to the left, while prokaryotic taxa are aligned to the right.


Heterotrichea
Blepharismidae
Climacostomidae
Condylostomidae
Fabreidae
Spirostomidae
Stentoridae
Karyorelictea
Cryptopharyngidae
Geleiidae
Kentrophoridae
Loxodidae
Trachelocercidae
Other
bame2_yy
bame_yy
bjap_mm
bsp_mk
cl01_yy
cvir_mm
cvir_yy
cmag_mm
fsal_mm
N70
N72
anig_yy
smin_yy
spol_upp
sroe_yy
cx01_yy
gacu_yy
gl04_yy
gsin2_yy
gsin_yy
pard_mm
N10
kn01_yy
s4110_ksd
s4111_ksd
s4112_ksd
sa331_kh
sa332_kh
lstr2_yy
lstr_yy
lx11_yy
lx13_yy
N4
N44
N5
rema_yy
N26
N27
N34
N38
tc01_yy
trac_yy
N39
N6
Known to have symbiotic bacteria
0.0
0.2
0.4
0.6
0.8
1.0
Fraction of mapped SSU rRNA reads
Class
Family
Library name
Bold name: from this study
Eukaryota
Karyorelictea
Heterotrichea
(Ciliophora)
Litostomatea
Spirotrichea
Conthreep
Bilateria
(Eukaryota)
Bacteria
(Nitrosococcaceae)
(Desulfobacteraceae)
(Gammaproteobacteria)
AqS1
Sphingomonas
Arcobacter
uncultured

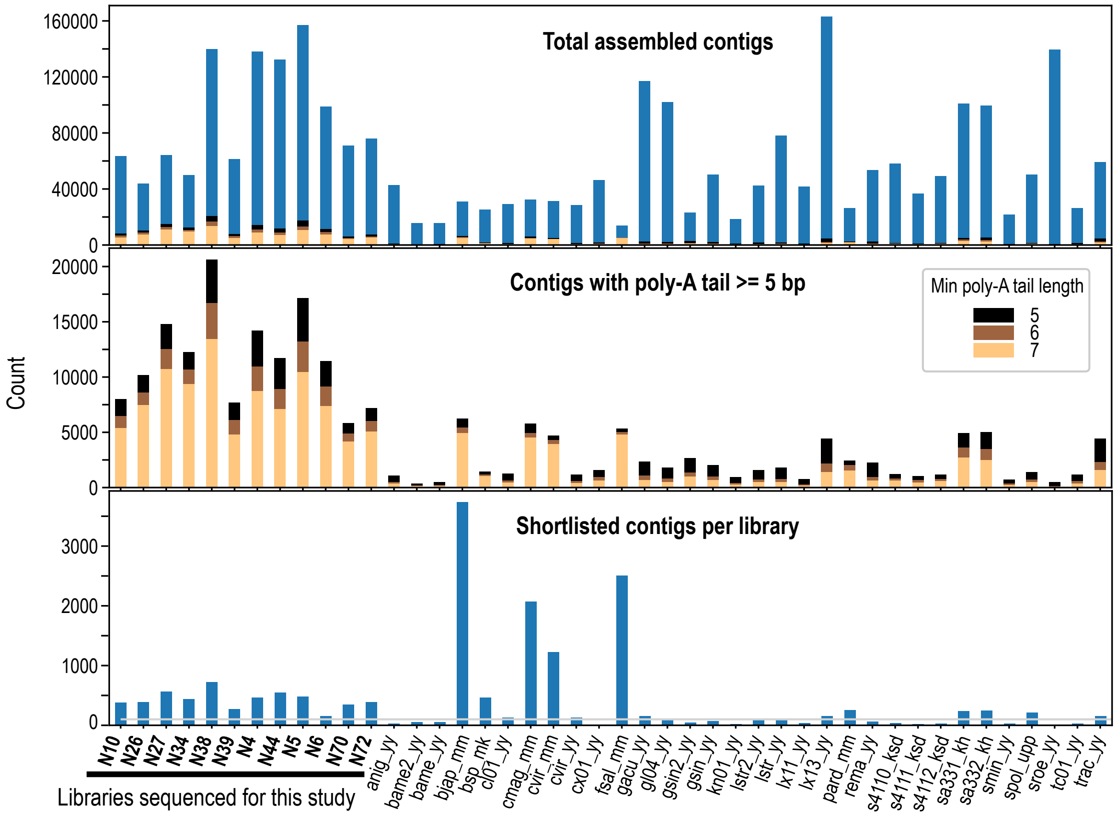
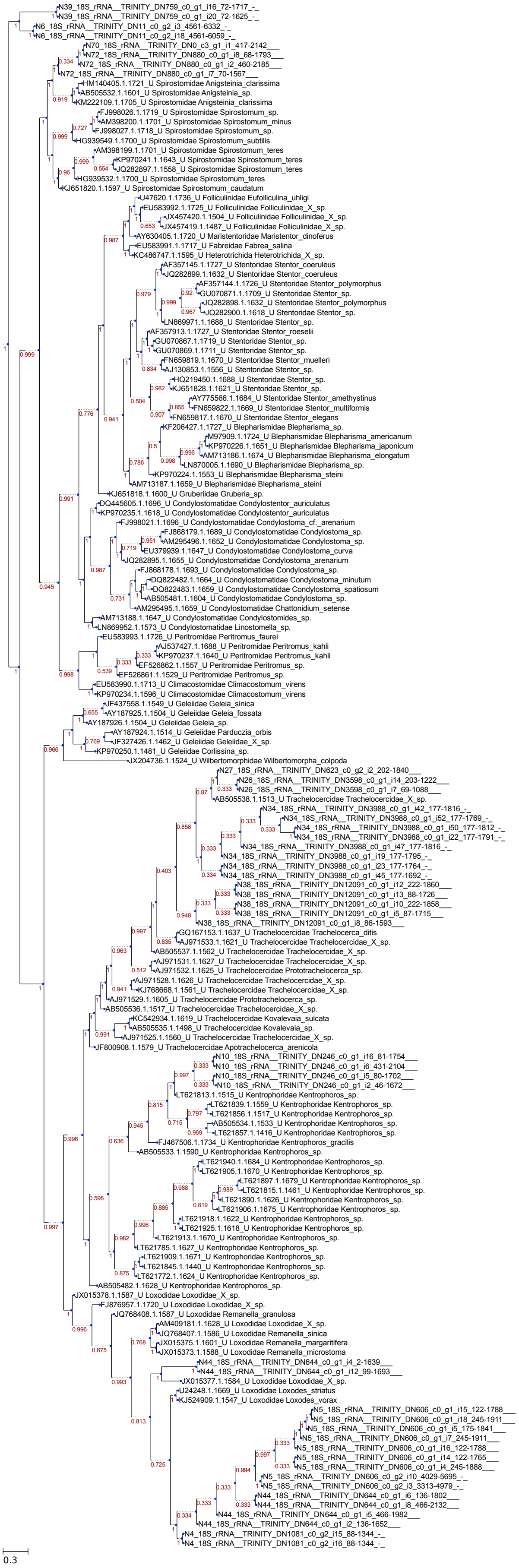


Figure S2. Number of contigs per transcriptome assembly:(top) total number of contigs, (middle) contigs with poly-A-tails, when different minimum lengths were applied, (bottom) putative full-length transcripts with both a poly-A tail (≥7 bp) and >80% Blastx hit vs. reference *Blepharisma stoltei* protein sequence (grey horizontal line: 100 sequences cutoff value).

Figure S3 (next page). Phylogenetic tree of 18s rRNA sequences from newly sequenced libraries (sequence names beginning with "N") vs. reference sequences of Karyorelictea and Heterotrichea from the PR2 database (identifier, family, species). Node labels: aBayes support values. Scale bar: Substitutions per site. Dotted lines: Branches spaced to accommodate node labels.



N39_18S_rRNA__TRINITY_DN759_c0_g1_i16_72-1717_-_
N39_18S_rRNA__TRINITY_DN759_c0_g1_i20_72-1625_-_
N6_18S_rRNA__TRINITY_DN11_c0_g2_i3_4561-6332_-_
N6_18S_rRNA__TRINITY_DN11_c0_g2_i18_4561-6059_-_
N70_18S_rRNA__TRINITY_DN0_c3_g1_i1_417-2142___
N72_18S_rRNA__TRINITY_DN880_c0_g1_i8_68-1793___
N72_18S_rRNA__TRINITY_DN880_c0_g1_i2_460-2185___
N72_18S_rRNA__TRINITY_DN880_c0_g1_i7_70-1567___
HM140405.1.1721_U Spirostomidae Anigsteinia_clarissima
AB505532.1.1601_U Spirostomidae Anigsteinia_sp.
KM222109.1.1705_U Spirostomidae Anigsteinia_clarissima
FJ998026.1.1719_U Spirostomidae Spirostomum_sp.
AM398200.1.1701_U Spirostomidae Spirostomum_minus
FJ998027.1.1718_U Spirostomidae Spirostomum_sp.
HG939549.1.1700_U Spirostomidae Spirostomum_subtilis
AM398199.1.1701_U Spirostomidae Spirostomum_teres
KP970241.1.1643_U Spirostomidae Spirostomum_teres
JQ282897.1.1558_U Spirostomidae Spirostomum_teres
HG939532.1.1700_U Spirostomidae Spirostomum_teres
KJ651820.1.1597_U Spirostomidae Spirostomum_caudatum
U47620.1.1736_U Folliculinidae Eufolliculina_uhligi
EU583992.1.1725_U Folliculinidae Folliculinidae_X_sp.
JX457420.1.1504_U Folliculinidae Folliculinidae_X_sp.
JX457419.1.1487_U Folliculinidae Folliculinidae_X_sp.
AY630405.1.1720_U Maristentoridae Maristentor_dinoferus
EU583991.1.1717_U Fabreidae Fabrea_salina
KC486747.1.1595_U Heterotrichida Heterotrichida_X_sp.
AF357145.1.1727_U Stentoridae Stentor_coeruleus
JQ282899.1.1632_U Stentoridae Stentor_coeruleus
AF357144.1.1726_U Stentoridae Stentor_polymorphus
GU070871.1.1709_U Stentoridae Stentor_sp.
JQ282898.1.1632_U Stentoridae Stentor_polymorphus
JQ282900.1.1618_U Stentoridae Stentor_sp.
LN869971.1.1688_U Stentoridae Stentor_sp.
AF357913.1.1727_U Stentoridae Stentor_roeselii
GU070867.1.1719_U Stentoridae Stentor_sp.
GU070869.1.1711_U Stentoridae Stentor_sp.
FN659819.1.1670_U Stentoridae Stentor_muelleri
AJ130853.1.1556_U Stentoridae Stentor_sp.
HQ219450.1.1688_U Stentoridae Stentor_sp.
KJ651828.1.1621_U Stentoridae Stentor_sp.
AY775566.1.1684_U Stentoridae Stentor_amethystinus
FN659822.1.1669_U Stentoridae Stentor_multiformis
FN659817.1.1670_U Stentoridae Stentor_elegans
KF206427.1.1727_U Blepharismidae Blepharisma_sp.
M97909.1.1724_U Blepharismidae Blepharisma_americanum
KP970226.1.1651_U Blepharismidae Blepharisma_japonicum
AM713186.1.1674_U Blepharismidae Blepharisma_elongatum
LN870005.1.1690_U Blepharismidae Blepharisma_sp.
KP970224.1.1553_U Blepharismidae Blepharisma_steini
AM713187.1.1659_U Blepharismidae Blepharisma_steini
KJ651818.1.1600_U Gruberiidae Gruberia_sp.
DQ445605.1.1696_U Condylostomatidae Condylostentor_auriculatus
KP970235.1.1618_U Condylostomatidae Condylostentor_auriculatus
FJ998021.1.1696_U Condylostomatidae Condylostoma_cf._arenarium
FJ868179.1.1689_U Condylostomatidae Condylostoma_sp.
AM295496.1.1652_U Condylostomatidae Condylostoma_sp.
EU379939.1.1647_U Condylostomatidae Condylostoma_curva
JQ282895.1.1655_U Condylostomatidae Condylostoma_arenarium
FJ868178.1.1693_U Condylostomatidae Condylostoma_sp.
DQ822482.1.1664_U Condylostomatidae Condylostoma_minutum
DQ822483.1.1659_U Condylostomatidae Condylostoma_spatiosum
AB505481.1.1604_U Condylostomatidae Condylostoma_sp.
AM295495.1.1659_U Condylostomatidae Chattonidium_setense
AM713188.1.1647_U Condylostomatidae Condylostomides_sp.
LN869952.1.1573_U Condylostomatidae Linostomella_sp.
EU583993.1.1726_U Peritromidae Peritromus_faurei
AJ537427.1.1688_U Peritromidae Peritromus_kahli
KP970237.1.1640_U Peritromidae Peritromus_kahli
EF526862.1.1557_U Peritromidae Peritromus_sp.
EF526861.1.1529_U Peritromidae Peritromus_sp.
EU583990.1.1713_U Climacostomidae Climacostomum_virens
KP970234.1.1596_U Climacostomidae Climacostomum_virens
JF437558.1.1549_U Geleiidae Geleia_sinica
AY187925.1.1504_U Geleiidae Geleia_fossata
AY187926.1.1504_U Geleiidae Geleia_sp.
AY187924.1.1514_U Geleiidae Parduczia_orbis
JF327426.1.1462_U Geleiidae Geleiidae_X_sp.
KP970250.1.1481_U Geleiidae Corlissina_sp.
JX204736.1.1524_U Wilbertomorphidae Wilbertomorpha_colpoda
N27_18S_rRNA__TRINITY_DN623_c0_g2_i2_202-1840___
N26_18S_rRNA__TRINITY_DN3598_c0_g1_i14_203-1222___
N26_18S_rRNA__TRINITY_DN3598_c0_g1_i7_69-1088___
AB505538.1.1513_U Trachelocercidae Trachelocercidae_X_sp.
N34_18S_rRNA__TRINITY_DN3988_c0_g1_i42_177-1816_-_
N34_18S_rRNA__TRINITY_DN3988_c0_g1_i52_177-1769_-_
N34_18S_rRNA__TRINITY_DN3988_c0_g1_i50_177-1812_-_
N34_18S_rRNA__TRINITY_DN3988_c0_g1_i22_177-1791_-_
N34_18S_rRNA__TRINITY_DN3988_c0_g1_i47_177-1816_-_
N34_18S_rRNA__TRINITY_DN3988_c0_g1_i19_177-1795_-_
N34_18S_rRNA__TRINITY_DN3988_c0_g1_i23_177-1764_-_
N34_18S_rRNA__TRINITY_DN3988_c0_g1_i45_177-1692_-_
N38_18S_rRNA__TRINITY_DN12091_c0_g1_i12_222-1860___
N38_18S_rRNA__TRINITY_DN12091_c0_g1_i13_88-1726___
N38_18S_rRNA__TRINITY_DN12091_c0_g1_i10_222-1858___
N38_18S_rRNA__TRINITY_DN12091_c0_g1_i5_87-1715___
N38_18S_rRNA__TRINITY_DN12091_c0_g1_i8_86-1593___
GQ167153.1.1637_U Trachelocercidae Trachelocerca_ditis
AJ971533.1.1621_U Trachelocercidae Trachelocercidae_X_sp.
AB505537.1.1562_U Trachelocercidae Trachelocercidae_X_sp.
AJ971531.1.1627_U Trachelocercidae Trachelocercidae_X_sp.
AJ971532.1.1625_U Trachelocercidae Prototrachelocerca_sp.
AJ971528.1.1626_U Trachelocercidae Trachelocercidae_X_sp.
KJ768668.1.1561_U Trachelocercidae Trachelocercidae_X_sp.
AJ971529.1.1605_U Trachelocercidae Prototrachelocerca_sp.
AB505536.1.1517_U Trachelocercidae Trachelocercidae_X_sp.
KC542934.1.1619_U Trachelocercidae Kovalevaia_sulcata
AB505535.1.1498_U Trachelocercidae Kovalevaia_sp.
AJ971525.1.1560_U Trachelocercidae Trachelocercidae_X_sp.
JF800908.1.1579_U Trachelocercidae Apotrachelocerca_arenicola
N10_18S_rRNA__TRINITY_DN246_c0_g1_i16_81-1754___
N10_18S_rRNA__TRINITY_DN246_c0_g1_i6_431-2104___
N10_18S_rRNA__TRINITY_DN246_c0_g1_i5_80-1702___
N10_18S_rRNA__TRINITY_DN246_c0_g1_i2_46-1672___
LT621813.1.1515_U Kentrophoridae Kentrophoros_sp.
LT621839.1.1559_U Kentrophoridae Kentrophoros_sp.
LT621856.1.1517_U Kentrophoridae Kentrophoros_sp.
AB505534.1.1533_U Kentrophoridae Kentrophoros_sp.
LT621857.1.1416_U Kentrophoridae Kentrophoros_sp.
FJ467506.1.1734_U Kentrophoridae Kentrophoros_gracilis
AB505533.1.1590_U Kentrophoridae Kentrophoros_sp.
LT621940.1.1684_U Kentrophoridae Kentrophoros_sp.
LT621905.1.1670_U Kentrophoridae Kentrophoros_sp.
LT621897.1.1679_U Kentrophoridae Kentrophoros_sp.
LT621815.1.1461_U Kentrophoridae Kentrophoros_sp.
LT621890.1.1626_U Kentrophoridae Kentrophoros_sp.
LT621906.1.1675_U Kentrophoridae Kentrophoros_sp.
LT621918.1.1622_U Kentrophoridae Kentrophoros_sp.
LT621925.1.1618_U Kentrophoridae Kentrophoros_sp.
LT621913.1.1670_U Kentrophoridae Kentrophoros_sp.
LT621785.1.1627_U Kentrophoridae Kentrophoros_sp.
LT621909.1.1671_U Kentrophoridae Kentrophoros_sp.
LT621845.1.1440_U Kentrophoridae Kentrophoros_sp.
LT621772.1.1624_U Kentrophoridae Kentrophoros_sp.
AB505482.1.1628_U Kentrophoridae Kentrophoros_sp.
JX015378.1.1587_U Loxodidae Loxodidae_X_sp.
FJ876957.1.1720_U Loxodidae Loxodidae_X_sp.
JQ768408.1.1587_U Loxodidae Remanella_granulosa
AM409181.1.1628_U Loxodidae Loxodidae_X_sp.
JQ768407.1.1586_U Loxodidae Remanella_sinica
JX015375.1.1601_U Loxodidae Remanella_margaritifera
JX015373.1.1588_U Loxodidae Remanella_microstoma
N44_18S_rRNA__TRINITY_DN644_c0_g1_i4_2-1639___
N44_18S_rRNA__TRINITY_DN644_c0_g1_i12_99-1693___
JX015377.1.1584_U Loxodidae Loxodidae_X_sp.
U24248.1.1669_U Loxodidae Loxodes_striatus
KJ524909.1.1547_U Loxodidae Loxodes_vorax
N5_18S_rRNA__TRINITY_DN606_c0_g1_i15_122-1788___
N5_18S_rRNA__TRINITY_DN606_c0_g1_i18_245-1911___
N5_18S_rRNA__TRINITY_DN606_c0_g1_i5_175-1841___
N5_18S_rRNA__TRINITY_DN606_c0_g1_i7_245-1911___
N5_18S_rRNA__TRINITY_DN606_c0_g1_i16_122-1788___
N5_18S_rRNA__TRINITY_DN606_c0_g1_i14_122-1765___
N5_18S_rRNA__TRINITY_DN606_c0_g1_i4_245-1888___
N5_18S_rRNA__TRINITY_DN606_c0_g2_i10_4029-5695_-_
N5_18S_rRNA__TRINITY_DN606_c0_g2_i3_3313-4979_-_
N44_18S_rRNA__TRINITY_DN644_c0_g1_i6_136-1802___
N44_18S_rRNA__TRINITY_DN644_c0_g1_i8_466-2132___
N44_18S_rRNA__TRINITY_DN644_c0_g1_i5_466-1982___
N44_18S_rRNA__TRINITY_DN644_c0_g1_i2_136-1652___
N4_18S_rRNA__TRINITY_DN1081_c0_g2_i15_88-1344_-_
N4_18S_rRNA__TRINITY_DN1081_c0_g2_i16_88-1344_-_
0.3