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

Brandon Kwee Boon Seah, Aditi Singh, Estienne Carl Swart

Max Planck Institute for Biology, 72076 Tübingen, Germany

Correspondence:

Brandon Kwee Boon Seah, kb.seah@tuebingen.mpg.de

Estienne Carl Swart, estienne.swart@tuebingen.mpg.de

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'-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 abundant group of eukaryotes.

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 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. This 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 alveolate *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*.

58 The karyorelicts are a class-level taxon within the ciliates, and sister group to the
59 heterotrichs. They are globally distributed and abundant in the sediment interstitial habitat of
60 marine coastal environments [18]. At least ~150 species have been formally described but
61 this is believed to be a severe underestimate of the true diversity [19,20], and they are also
62 poorly represented in sequence databases.

63 We therefore sequenced additional karyorelict transcriptomes and reanalyzed published
64 data to assess whether karyorelicts other than *Parduczia* could be using ambiguous genetic
65 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). After filtering for quality and sufficient data, 25 transcriptome assemblies were used to evaluate stop codon reassignment (15 previously published), vs. 26 assemblies (16 previously published) for inferring the actual stop codon(s) (Supplementary Information).

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 observed for coding codons in organisms with known genetic codes (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 1). 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 1). Nonetheless, these 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.

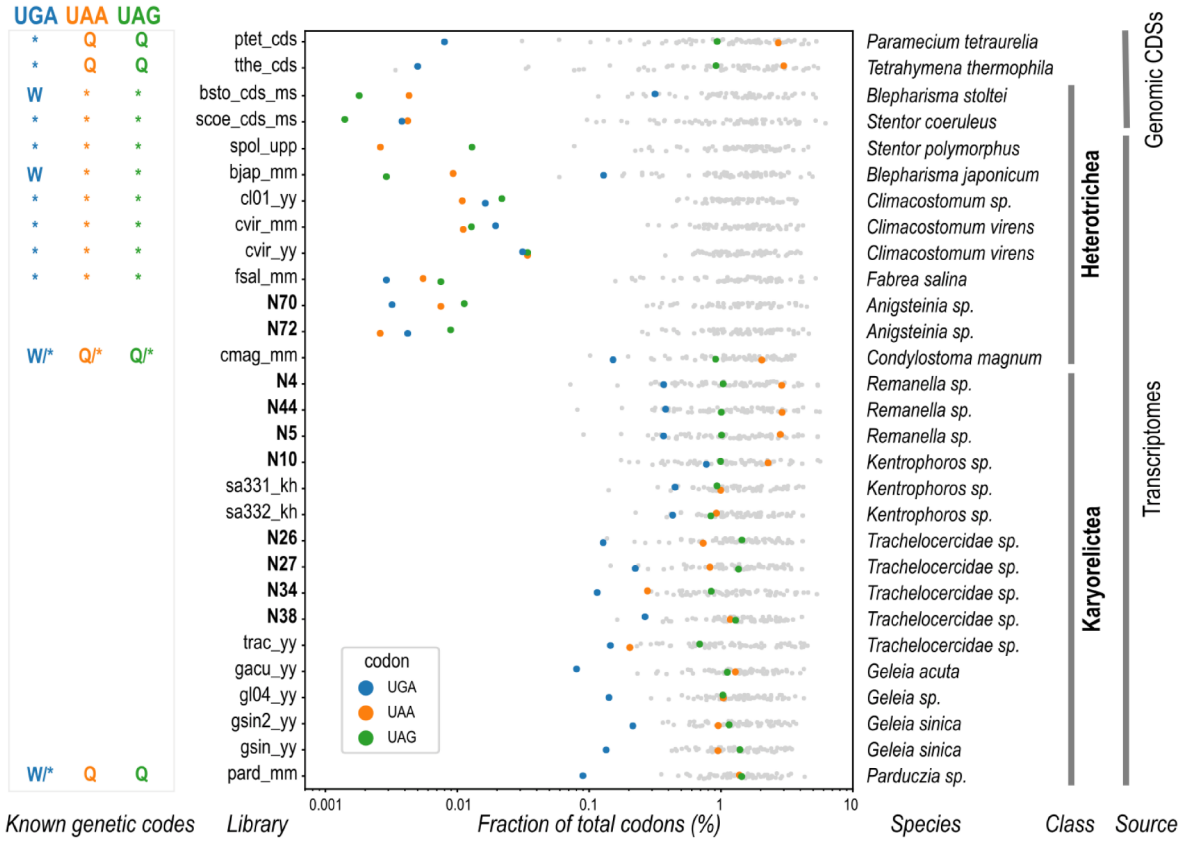
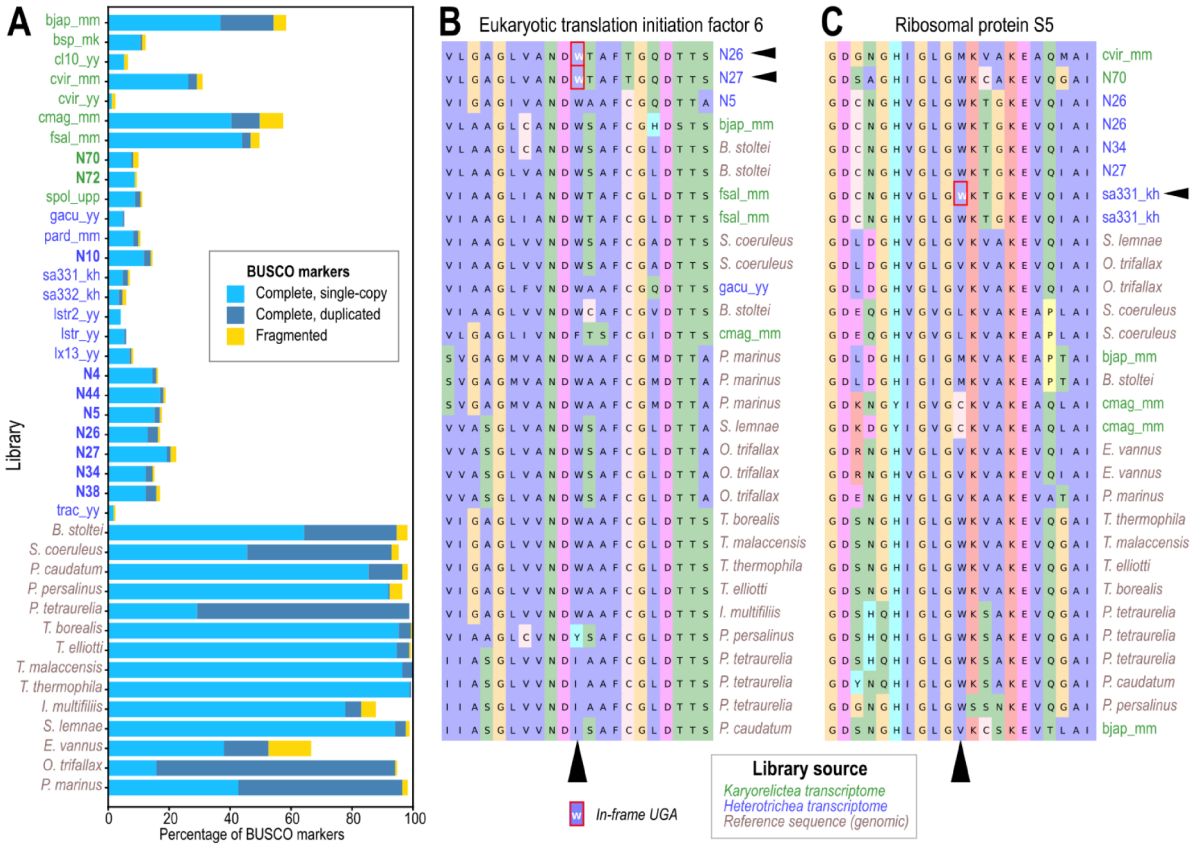


Figure 1. Codon frequencies of canonical stop codons (UGA: blue, UAA: orange, UAG: green) and other codons (gray) in conserved protein domains found by hmmscan search in six-frame translations of transcriptome assemblies (Table S1) or genomic CDSs (Table S2) vs. Pfam. (**left**) Assignments of canonical stops for organisms with known genetic codes, following Swart et al., 2016. Names of libraries from this study are highlighted in bold.

96 In-frame UGAs were found in 10.5 to 76.9% of transcripts with putative coding regions
97 predicted by full-length Blastx hits per karyorelict library (Figure 4D). This verified that in-
98 frame UGAs were not concentrated in a small fraction of potentially spurious sequences but
99 in fact found in many genes. Conserved 'marker' genes that were generally expected to be
100 present in ciliate genomes (BUSCO orthologs, Alveolata marker set) also contained in-frame
101 UGAs. The karyorelict transcriptome assemblies were relatively incomplete, with 1.8% to
102 20.5% (median 12.0%) estimated completeness based on the BUSCO markers, and a total
103 of 91 of 171 BUSCO orthologs were found in these assemblies (Figure 2A). Nonetheless, 46
104 BUSCO orthologs from 14 karyorelict assemblies were found with in-frame UGAs in
105 conserved alignment positions (e.g. Figure 2B, 2C), verifying that they are not limited to
106 poorly characterized or hypothetical proteins.

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

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



115 **Figure 2.** In-frame coding UGAs in conserved marker genes. (**A**) Completeness estimates of
116 heterotrich and karyorelict transcriptomes (library names in green and blue respectively),
117 compared with genomic reference sequences from other ciliates (Table S3); BUSCO
118 Alveolata marker set. (**B, C**) Two examples of alignments (excerpts) for conserved
119 orthologous protein-coding genes (orthologs 20320at33630 and 23778at33630), which
120 contain in-frame UGAs translated as W in karyorelict sequences, flanked by conserved
121 alignment blocks.

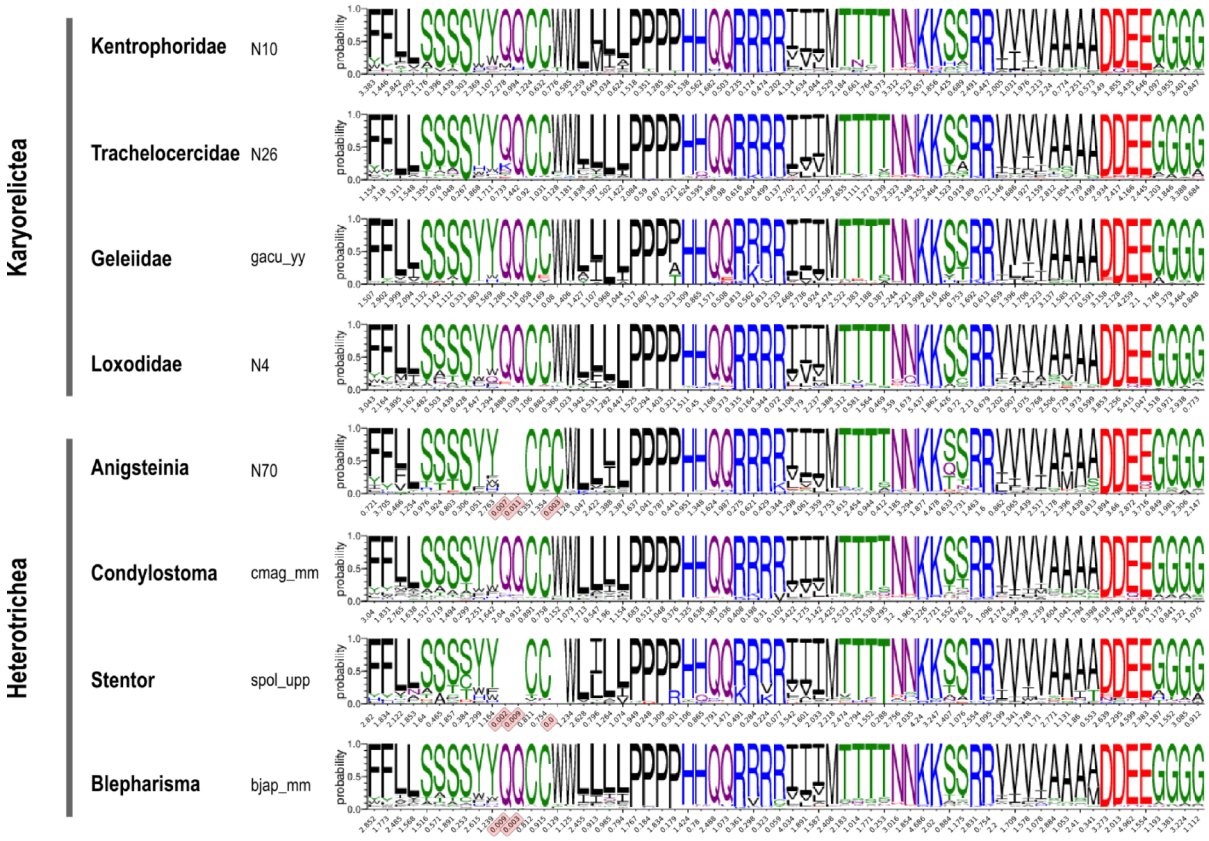
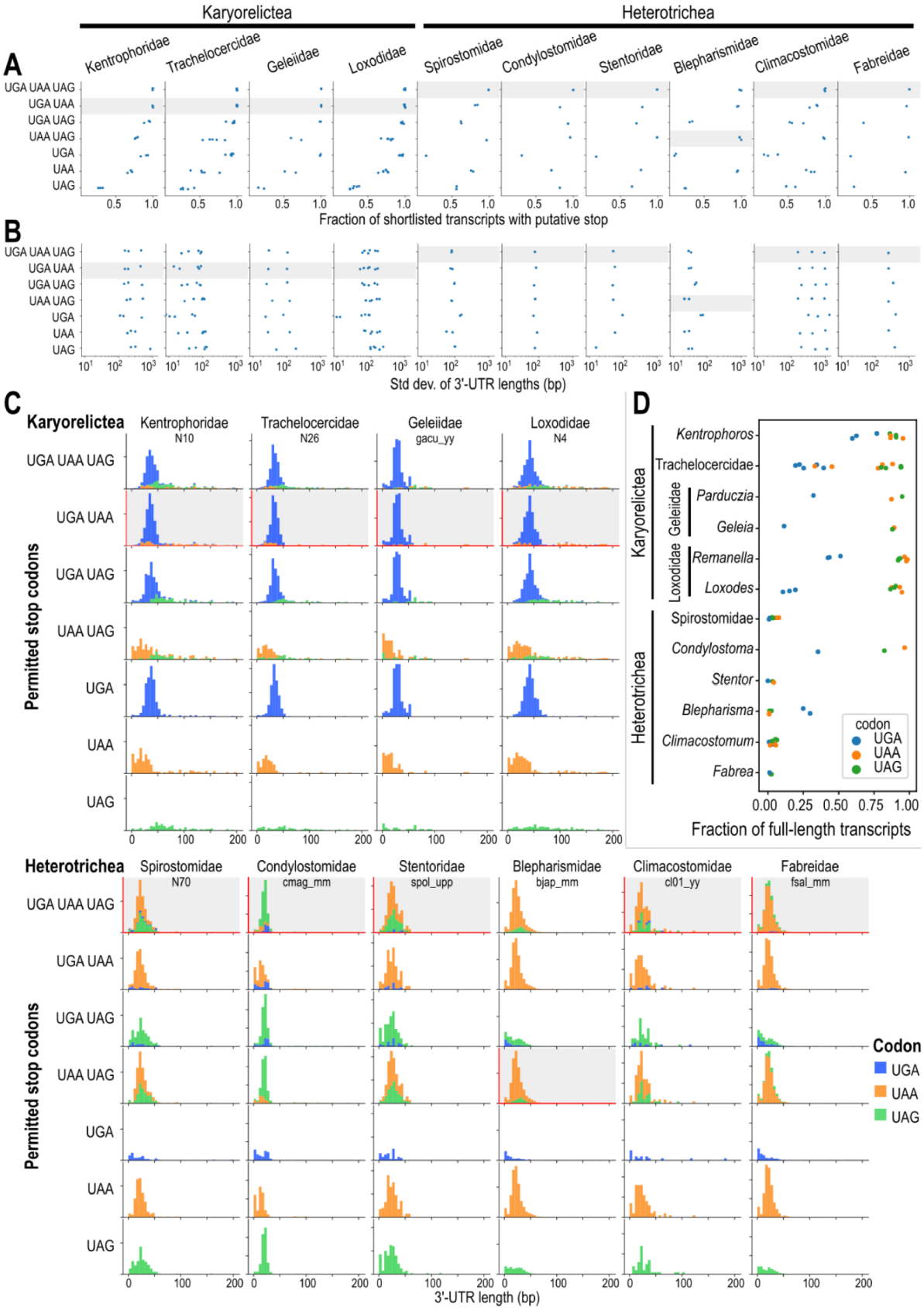


Figure 3. Weblogos representing the likely amino acid assignment of each codon in selected libraries (library with most coverage per taxon of interest). Heights of each letter represent the relative frequencies (all scaled to 100%) of each amino acid in conserved residues aligning to that codon. The 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).

129 *Stop codons in karyorelicts and heterotrichs*

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

144 With these criteria, the candidate stop codons for karyorelicts could be narrowed to two
145 possibilities: UGA alone or UGA + UAA. If only UGA was permitted as a stop codon, 84-98%
146 of transcripts per library had a putative true stop, but if both UGA and UAA were permitted
147 as stop codons, the proportion was over 98% (Figure 4A). This was at the expense of
148 somewhat more variance in the 3'-UTR length distribution, although we found that this metric
149 was of limited usefulness because UGA was always the majority in all stop codon
150 combinations where it was present (Figures 4B, 4C). Both karyorelicts and heterotrichs in
151 this study had short and narrowly distributed 3'-UTR lengths (median 28 nt, interquartile
152 range 18 nt) (Figure 4C). The heterotrichs were shortest overall, with median lengths per
153 taxon between 21 nt (*Condylostoma*) and 26 nt (*Stentor*), followed by the karyorelict families
154 Trachelocercidae (33 nt), Geleiidae (31 nt), Kentrophoridae (37 nt), and Loxodidae (43 nt).



155 **Figure 4.** Effect of different stop codon combinations on assembly metrics. Predicted stop

157 codon usage for each taxon from this study or previous publications highlighted in gray. (**A**)
158 Strip plots for the fraction of full length contigs per transcriptome that have a putative stop
159 codon from that specific combination (rows), i.e. in-frame, downstream of full-length Blastx
160 hit vs. reference, and upstream of poly-A tail. Each point corresponds to one transcriptome
161 assembly, grouped by taxonomic family (columns). (**B**) Scatterplots for standard deviation of
162 3'-UTR lengths. (**C**) Histograms for 3'-UTR lengths, colored by putative stop codon (UGA:
163 blue, UAA: orange, UAG: green), one representative library per family. (**D**) Fraction of full-
164 length transcripts that have at least one canonical stop codon in the putative coding region,
165 grouped by family (further split to genus for Loxodidae and Geleiidae).

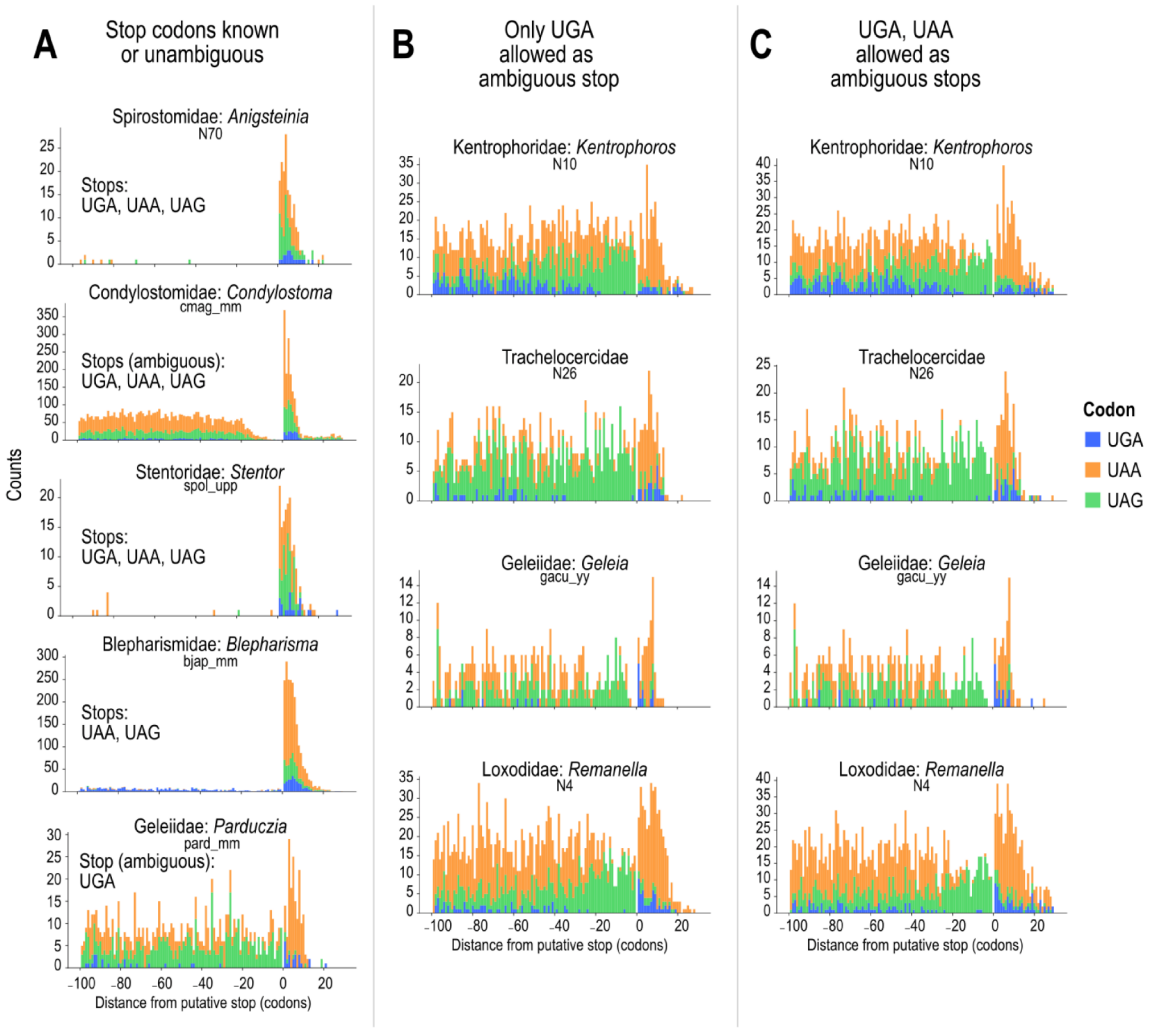


Figure 5. Depletion of in-frame coding "stop" codons in the coding sequence (negative coordinates) immediately before the putative true stop codon (position 0) and their enrichment in the 3'-UTR (positive coordinates). Representative library with highest number of assembled full length contigs chosen 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.

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

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

Discussion

We have found 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 is mapped to an up-to-date phylogeny [19], we can infer that this ambiguous code, formerly reported only for *Parduczia* sp. (Geleiidae) among ciliates, was actually acquired at the root of the karyorelict clade (Figure 6).

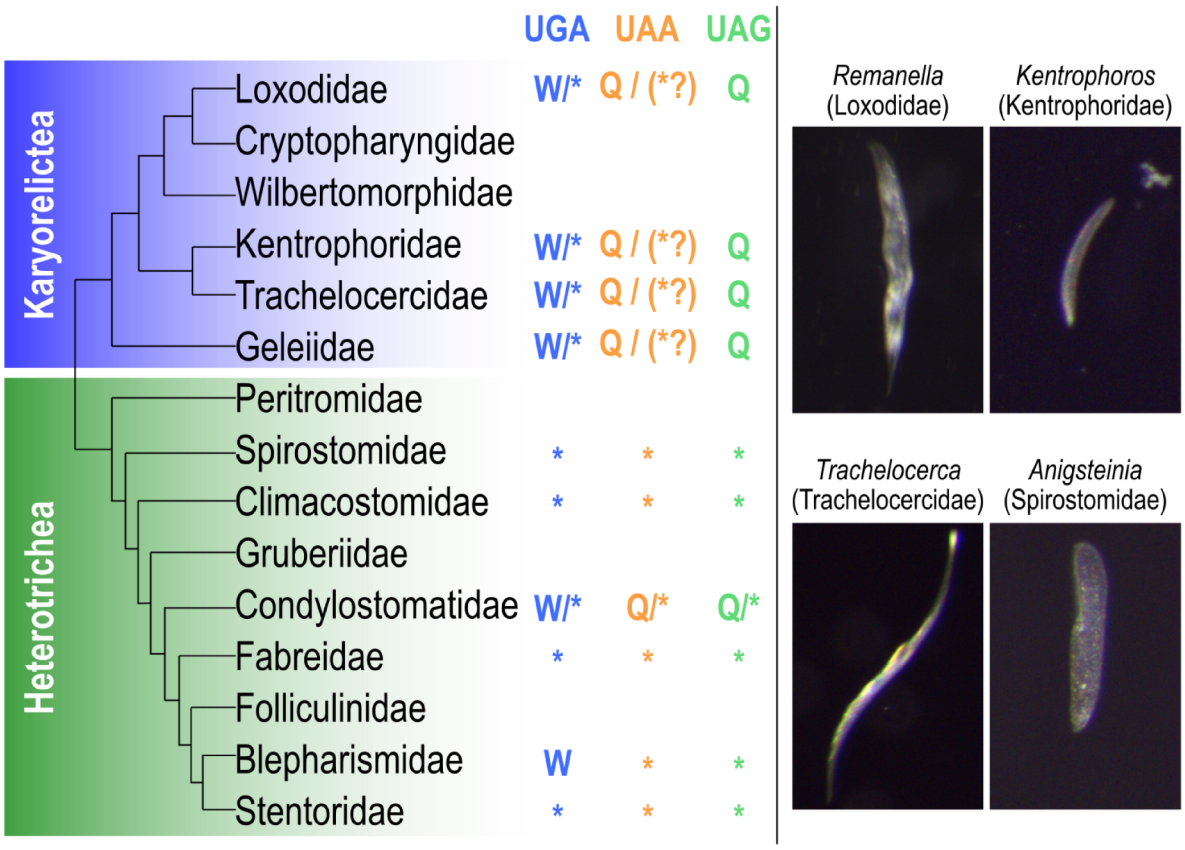


Figure 6. Genetic code diversity among karyorelict and heterotrich ciliates. (**Left**) Diagrammatic karyorelict + heterotrich tree with predicted stop codon reassignments mapped to each family. Subtree topologies are from Ma et al. (2022) and Fernandes et al. (2016) respectively. Branch lengths are not representative of evolutionary distances. (**Right**) Photomicrographs of ciliates (incident light) collected in this study from Roscoff, France; height of each panel 50 µm.

207 Available data for *Cryptopharynx* (Karyorelictea: Cryptopharyngidae) were not conclusive.
208 The canonical stop codons had frequencies between 0.02 and 0.07%, lower than for other
209 karyorelicts, but higher than true stop codons, but this family was represented by a single
210 library that had high contamination from other eukaryotes (Supplementary Text) and there
211 were too few high-confidence, full length transcripts for a reliable conclusion on its genetic
212 code. No sequence data beyond rRNA genes were publicly available for the remaining
213 family, Wilbertomorphidae, whose phylogenetic position in relation to the other karyorelicts is
214 unclear because of long branch lengths [21].

215 Ambiguous stop/sense codes are hence not just isolated phenomena, but are used by a
216 major taxon that is diverse, globally distributed, and abundant in its respective habitats. In
217 contrast, the heterotrichs, which constitute the sister group to Karyorelictea and are hence of
218 the same evolutionary age, use at least three different genetic codes, including one with
219 ambiguous stops (Figure 6). If organisms with ambiguous codes were isolated single
220 species whose nearest relatives have conventional stops, as appears to be the case for
221 *Blastocrithidia* spp. and *Amoebophrya* sp., we might conclude that these are uncommon
222 occurrences that do not persist over longer evolutionary time scales. However, the
223 karyorelict crown group diversified during the Proterozoic (posterior mean 455 Mya) and the
224 stem split from the Heterotrichea even earlier, in the Neo-Proterozoic [22].

225 This study has benefited from several technical improvements. A highly complete,
226 contiguous genome assembly with gene predictions is now available for the heterotrich
227 *Blepharisma stoltei* [23]. Because *Blepharisma* is more closely related to the karyorelicts
228 than other ciliate model species, which are mostly oligohymenophorans and spirotrichs, it
229 improved the reference-based annotation of the assembled transcriptomes. Single-cell RNA-
230 seq libraries in this study were also sequenced to a greater depth, with a lower fraction of
231 contamination from rRNA, and hence yielded more full length mRNA transcripts for analysis.

232 One proposed mechanism for how the cell correctly recognizes whether an ambiguous
233 codon is coding or terminal is based on the proximity of translation stops to the poly-A tail of
234 transcripts. In this model, tRNAs typically bind more efficiently to in-frame coding “stops”
235 than eukaryotic translation release factor 1 (eRF1), hence allowing these codons to be
236 translated. At the true termination stop codon, however, the binding of eRF1 can be
237 stabilized by interactions with poly-A interacting proteins like PABP bound to the nearby
238 poly-A tail, allowing it to outcompete tRNAs and hydrolyze the peptidyl-tRNA bond [10,24].
239 Consistent with this model, we found that karyorelict 3’-UTRs are also relatively short, and
240 that in-frame UGAs are depleted immediately before the putative true stop codon.

241 Nonetheless, karyorelict 3'-UTRs are actually about 10 nt longer on average than those of
242 heterotrichs.

243 Our results also raised the possibility that UAA is also used as an ambiguous stop codon for
244 ~1-10% of karyorelict transcripts, in addition to the main stop codon UGA. eRF1 may retain
245 a weak affinity for UAA, and recognize UAA for terminating translation albeit with lower
246 efficiency. In *Blepharisma japonicum*, where UAA and UAG are non-ambiguous stops and
247 UGA encodes tryptophan (albeit at low frequency, 0.13%), heterologously expressed eRF1
248 could still recognize all three codons in an *in vitro* assay, although efficiency of peptidyl-tRNA
249 hydrolysis was lower with UGA than for UAA and UAG [25]. In species with non-ambiguous
250 stop codon reassignment, the effect of such "weak" ambiguity on the total pool of translated
251 protein may be negligible, but it shows that there is a latent potential that could account for
252 the repeated evolution of stop codon reassignments in ciliates. Furthermore, UAAs were
253 even more abundant than UGAs in ciliate 3'-UTRs, which can be attributed to the low GC%
254 of 3'-UTRs compared to coding sequences; other A/U-only codons were also enriched in 3'-
255 UTRs. Therefore, UAAs in the 3'-UTRs of karyorelicts may be a 'backstop' mechanism that
256 prevents occasional stop-codon readthrough, as proposed for tandem stop codons (TSCs) in
257 other species with reassigned stop codons [26]. In the minority of transcripts where in-frame
258 stop-UGA is absent, the backstop may be adequate to terminate translation before the poly-
259 A tail and produce a functional protein most of the time.

260 What selective pressures might favor the evolution and maintenance of an ambiguous
261 genetic code? One possibility is that they confer mutational robustness by eliminating
262 substitutions that cause premature stop codons. They do not appear to be linked to a
263 specific habitat: *Blastocrithidia* spp. and *Amoebophrya* sp. are both parasites of eukaryotic
264 hosts, but of insects and free-living dinoflagellates respectively; whereas the karyorelict
265 ciliates and *Condylostoma* are both found in marine interstitial environments, but live
266 alongside other ciliates that have conventional codes, such as *Anigsteinia*. Having short 3'-
267 UTRs may predispose ciliates to adopt ambiguous codes by facilitating interactions between
268 eRF1 and PABPs that could enable stop recognition, but it is not the only deciding factor
269 because the 3'-UTRs of ciliates with conventional stop codons are also comparably short,
270 particularly among the heterotrichs.

271 Any adaptationist hypothesis for alternative and ambiguous codes will have to contend with
272 the existence of related organisms with conventional codes that have similar lifestyles.
273 Furthermore, once a stop codon has been reassigned to sense, it becomes increasingly
274 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 [27], 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 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) 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 [17,28–30] were downloaded from the European Nucleotide Archive (ENA) (Table S1). Contamination from non-target organisms was evaluated by mapping reads to an rRNA reference database and summarizing the hits

310 by taxonomy. Although RNAseq library construction enriches mRNAs using poly-A tail
311 selection, there is typically still sufficient rRNA present in the final library to evaluate the
312 taxonomic composition of the sample. All RNAseq read libraries (newly sequenced and
313 previously published) were processed with the same pipeline: The taxonomic composition of
314 each library was evaluated by mapping 1 M read pairs per library against the SILVA SSU
315 Ref NR 132 database [31], using phyloFlash v3.3b1 [32]. Newly sequenced libraries were
316 assigned to a genus or family using the mapping-based taxonomic summary, or full-length
317 18S rRNA gene if it was successfully assembled.

318 Reads were trimmed with the program bbduk.sh from BBmap v38.22 (http://sourceforge.net/
319 projects/bbmap/) to remove known adapters (right end) and low-quality bases (both ends),
320 with minimum Phred quality 24 and minimum read length 25 bp. Trimmed reads were then
321 assembled with Trinity v2.12.0 [33] using default parameters. Assembled contigs were
322 aligned against the *Blepharisma stoltei* ATCC 30299 proteome [23] with NCBI Blastx v2.12.0
323 [34] using the standard genetic code and E-value cutoff 10^{-20}, parallelized with GNU Parallel
324 [35].

325 Read pre-processing, quality control, and assembly were managed with a Snakemake
326 v6.8.1 [36] workflow (https://github.com/Swart-lab/karyocode-workflow). Scripts for data
327 processing described below were written in Python v3.7.3 using Biopython v1.74 [37],
328 pandas v0.25.0 [38], seaborn v0.11.0 [39] and Matplotlib v3.1.1 [40] libraries unless
329 otherwise stated.

330 *Prediction of stop codon reassignment to sense*

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

338 Codon frequencies and their corresponding amino acids were predicted with an updated
339 version of PORC (v2.1, https://github.com/Swart-lab/PORC, managed with a Snakemake
340 workflow, https://github.com/Swart-lab/karyocode-analysis-porc); the method has been
341 previously described [10,41]. Briefly: a six-frame translation was produced for each contig in
342 the transcriptome assembly, and searched against conserved domains in the Pfam-A

343 database v32 [42] with hmmscan from HMMer v3.3.2 (http://hmmer.org/). Overall codon
344 frequencies were counted from alignments with E-value ≤ 10^{-20}. To ensure that there was
345 sufficient data underlying the codon and amino acid frequencies, only those libraries with at
346 least 100 observations for each of the coding codons in the standard genetic code were
347 used for comparison of codon frequencies and for prediction of amino acid assignments.

348 Frequencies of amino acids aligning to a given codon were counted from columns where the
349 HMM model consensus was ≥50% identity in the alignment used to build the model (upper-
350 case positions in the HMM consensus). Sequence logos of amino acid frequencies per
351 codon for each library were drawn with Weblogo v3.7.5 [43].

352 In addition to the transcriptomes, genomic CDSs of selected model species with different
353 genetic codes [23,44–48] were also analyzed with PORC to obtain a reference baseline of
354 coding-codon frequencies (Table S2). These model species have non-ambiguous codes so
355 they were not expected to have stop codons in the CDSs, except for the terminal stop.

356 *Prediction of coding frame in full-length transcripts*

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

367 *Metrics for evaluating potential stop codon combinations*

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

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

384 *Verification of in-frame UGAs in conserved marker genes*
385 Full-length CDSs (see above) were translated with the karyorelict code (NCBI table 27).
386 Conserved marker genes were identified with BUSCO v5.2.2 (protein mode,
387 alveolata_odb10 marker set) [49], managed with a Snakemake workflow (https://github.com/
388 Swart-lab/karyocode-analysis-busco). Markers for additional ciliate species where relatively
389 complete genome assemblies and gene predictions were available were also identified
390 (Table S3) [45,47,50–57]. For each BUSCO marker, the ciliate homologs were aligned with
391 Muscle v3.8.1551 [58]. Alignment columns corresponding to in-frame putatively coding
392 UGAs of karyorelict sequences were identified. These positions were considered to be
393 conserved if ≥50% of residues were W or another aromatic amino acid (Y, F, or H).

394 Data availability

395 RNA-seq libraries sequenced for this study have been deposited at the European Nucleotide
396 Archive (https://www.ebi.ac.uk/ena/) under accession PRJEB50648.

397 Supplementary Information

398 **Supplementary Text**. Quality metrics of single-cell transcriptome assemblies.

399 **Table S1**. Transcriptomic RNAseq libraries from karyorelict and heterotrich ciliates analyzed
400 in this project.

401 **Table S2**. Genomic CDS sequences of cultivated model ciliates with unambiguous stop
402 codons, used for baseline comparison of coding vs. stop codon frequencies in HMMer
403 searches of six-frame translations.

404 **Table S3**. High completeness proteomes of ciliate model organisms used for BUSCO

marker comparison and alignment.

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.

Acknowledgements

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, and members of the Swart Lab for helpful feedback and support.

416 ## References

417 1. Ivanova NN, Schwientek P, Tripp HJ, Rinke C, Pati A, Huntemann M, et al. Stop
418 codon reassignments in the wild. Science. 2014;344: 909–913.
419 doi:10.1126/science.1250691

420 2. Shulgina Y, Eddy SR. A computational screen for alternative genetic codes in over
421 250,000 genomes. eLife. 2021;10. doi:10.7554/eLife.71402

422 3. Kollmar M, Mühlhausen S. Nuclear codon reassignments in the genomics era and
423 mechanisms behind their evolution. Bioessays. 2017;39. doi:10.1002/bies.201600221

424 4. Schueren F, Thoms S. Functional translational readthrough: A systems biology
425 perspective. PLoS Genet. 2016;12: e1006196. doi:10.1371/journal.pgen.1006196

426 5. Suzuki T, Ueda T, Watanabe K. The “polysemous” codon--a codon with multiple amino
427 acid assignment caused by dual specificity of tRNA identity. EMBO J. 1997;16: 1122–
428 1134. doi:10.1093/emboj/16.5.1122

429 6. Hatfield DL, Gladyshev VN. How selenium has altered our understanding of the
430 genetic code. Mol Cell Biol. 2002;22: 3565–3576. doi:10.1128/MCB.22.11.3565-
431 3576.2002

432 7. Záhonová K, Kostygov AY, Ševčíková T, Yurchenko V, Eliáš M. An Unprecedented
433 Non-canonical Nuclear Genetic Code with All Three Termination Codons Reassigned
434 as Sense Codons. Curr Biol. 2016;26: 2364–2369. doi:10.1016/j.cub.2016.06.064

435 8. Bachvaroff TR. A precedented nuclear genetic code with all three termination codons
436 reassigned as sense codons in the syndinean *Amoebophrya* sp. ex *Karlodinium*
437 *veneficum*. PLoS ONE. 2019;14: e0212912. doi:10.1371/journal.pone.0212912

438 9. Heaphy SM, Mariotti M, Gladyshev VN, Atkins JF, Baranov PV. Novel Ciliate Genetic
439 Code Variants Including the Reassignment of All Three Stop Codons to Sense Codons
440 in *Condylostoma magnum*. Mol Biol Evol. 2016;33: 2885–2889.
441 doi:10.1093/molbev/msw166

442 10. Swart EC, Serra V, Petroni G, Nowacki M. Genetic Codes with No Dedicated Stop
443 Codon: Context-Dependent Translation Termination. Cell. 2016;166: 691–702.
444 doi:10.1016/j.cell.2016.06.020

445 11. Lozupone CA, Knight RD, Landweber LF. The molecular basis of nuclear genetic code
446 change in ciliates. Curr Biol. 2001;11: 65–74. doi:10.1016/s0960-9822(01)00028-8

447 12. Preer JR, Preer LB, Rudman BM, Barnett AJ. Deviation from the universal code shown

448 by the gene for surface protein 51A in *Paramecium*. Nature. 1985;314: 188–190.
449 doi:10.1038/314188a0

450 13. Horowitz S, Gorovsky MA. An unusual genetic code in nuclear genes of *Tetrahymena*.
451 Proc Natl Acad Sci USA. 1985;82: 2452–2455. doi:10.1073/pnas.82.8.2452

452 14. Meyer F, Schmidt HJ, Plümper E, Hasilik A, Mersmann G, Meyer HE, et al. UGA is
453 translated as cysteine in pheromone 3 of *Euplotes octocarinatus*. Proc Natl Acad Sci
454 USA. 1991;88: 3758–3761. doi:10.1073/pnas.88.9.3758

455 15. Tourancheau AB, Tsao N, Klobutcher LA, Pearlman RE, Adoutte A. Genetic code
456 deviations in the ciliates: evidence for multiple and independent events. EMBO J.
457 1995;14: 3262–3267. doi:10.1002/j.1460-2075.1995.tb07329.x

458 16. Helftenbein E. Nucleotide sequence of a macronuclear DNA molecule coding for
459 alpha-tubulin from the ciliate *Stylonychia lemnae*. Special codon usage: TAA is not a
460 translation termination codon. Nucleic Acids Res. 1985;13: 415–433. doi:10.1093/nar/
461 13.2.415

462 17. Yan Y, Maurer-Alcalá XX, Knight R, Kosakovsky Pond SL, Katz LA. Single-Cell
463 Transcriptomics Reveal a Correlation between Genome Architecture and Gene Family
464 Evolution in Ciliates. MBio. 2019;10. doi:10.1128/mBio.02524-19

465 18. Fenchel T. The ecology of marine microbenthos IV. Structure and function of the
466 benthic ecosystem, its chemical and physical factors and the microfauna commuities
467 with special reference to the ciliated protozoa. Ophelia. 1969;6: 1–182.
468 doi:10.1080/00785326.1969.10409647

469 19. Ma M, Li Y, Maurer-Alcalá XX, Wang Y, Yan Y. Deciphering phylogenetic relationships
470 in class Karyorelictea (Protista, Ciliophora) based on updated multi-gene information
471 with establishment of a new order Wilbertomorphida n. ord. Mol Phylogenet Evol.
472 2022; 107406. doi:10.1016/j.ympev.2022.107406

473 20. Foissner W. The karyorelictids (Protozoa: Ciliophora), a unique and enigmatic
474 assemblage of marine, interstitial ciliates: a review emphasizing ciliary patterns and
475 evolution. Evolutionary relationships among Protozoa. 1998; 305–325.

476 21. Xu Y, Li J, Song W, Warren A. Phylogeny and establishment of a new ciliate family,
477 Wilbertomorphidae fam. nov. (Ciliophora, Karyorelictea), a highly specialized taxon
478 represented by *Wilbertomorpha colpoda* gen. nov., spec. nov. J Eukaryot Microbiol.
479 2013;60: 480–489. doi:10.1111/jeu.12055

480 22. Fernandes NM, Schrago CG. A multigene timescale and diversification dynamics of
481 Ciliophora evolution. Mol Phylogenet Evol. 2019;139: 106521.
482 doi:10.1016/j.ympev.2019.106521

483 23. Singh M, Seah BKB, Emmerich C, Singh A, Woehle C, Huettel B, et al. The
484 *Blepharisma stoltei* macronuclear genome: towards the origins of whole genome
485 reorganization. BioRxiv. 2021. doi:10.1101/2021.12.14.471607

486 24. Alkalaeva E, Mikhailova T. Reassigning stop codons via translation termination: How a
487 few eukaryotes broke the dogma. Bioessays. 2017;39. doi:10.1002/bies.201600213

488 25. Eliseev BD, Alkalaeva EZ, Kryuchkova PN, Lekomtsev SA, Wang W, Liang A-H, et al.
489 Translation termination factor eRF1 of the ciliate *Blepharisma japonicum* recognizes all
490 three stop codons. Mol Biol (NY). 2011;45: 614–618.
491 doi:10.1134/S0026893311040030

492 26. Fleming I, Cavalcanti ARO. Selection for tandem stop codons in ciliate species with
493 reassigned stop codons. PLoS ONE. 2019;14: e0225804.
494 doi:10.1371/journal.pone.0225804

495 27. Crick FH. The origin of the genetic code. J Mol Biol. 1968;38: 367–379.
496 doi:10.1016/0022-2836(68)90392-6

497 28. Seah BKB, Antony CP, Huettel B, Zarzycki J, Schada von Borzyskowski L, Erb TJ, et
498 al. Sulfur-Oxidizing Symbionts without Canonical Genes for Autotrophic CO2 Fixation.
499 MBio. 2019;10: e01112-19. doi:10.1128/mBio.01112-19

500 29. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, et al. The
501 Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP):
502 illuminating the functional diversity of eukaryotic life in the oceans through
503 transcriptome sequencing. PLoS Biol. 2014;12: e1001889.
504 doi:10.1371/journal.pbio.1001889

505 30. Kolisko M, Boscaro V, Burki F, Lynn DH, Keeling PJ. Single-cell transcriptomics for
506 microbial eukaryotes. Curr Biol. 2014;24: R1081-2. doi:10.1016/j.cub.2014.10.026

507 31. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA
508 ribosomal RNA gene database project: improved data processing and web-based
509 tools. Nucleic Acids Res. 2013;41: D590-6. doi:10.1093/nar/gks1219

510 32. Gruber-Vodicka HR, Seah BKB, Pruesse E. phyloFlash: Rapid Small-Subunit rRNA
511 Profiling and Targeted Assembly from Metagenomes. mSystems. 2020;5. doi:10.1128/

512 mSystems.00920-20

513 33. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length
514 transcriptome assembly from RNA-Seq data without a reference genome. Nat
515 Biotechnol. 2011;29: 644–652. doi:10.1038/nbt.1883

516 34. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al.
517 BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421.
518 doi:10.1186/1471-2105-10-421

519 35. Tange O. Gnu Parallel 2018. Zenodo. 2018. doi:10.5281/zenodo.1146014

520 36. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al.
521 Sustainable data analysis with Snakemake. F1000Res. 2021;10: 33.
522 doi:10.12688/f1000research.29032.2

523 37. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely
524 available Python tools for computational molecular biology and bioinformatics.
525 Bioinformatics. 2009;25: 1422–1423. doi:10.1093/bioinformatics/btp163

526 38. McKinney W. Data structures for statistical computing in python. Proceedings of the
527 9th Python in Science Conference. SciPy; 2010. pp. 56–61. doi:10.25080/Majora-
528 92bf1922-00a

529 39. Waskom M. seaborn: statistical data visualization. JOSS. 2021;6: 3021.
530 doi:10.21105/joss.03021

531 40. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007;9: 90–95.
532 doi:10.1109/MCSE.2007.55

533 41. Dutilh BE, Jurgelenaite R, Szklarczyk R, van Hijum SAFT, Harhangi HR, Schmid M, et
534 al. FACIL: Fast and Accurate Genetic Code Inference and Logo. Bioinformatics.
535 2011;27: 1929–1933. doi:10.1093/bioinformatics/btr316

536 42. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al.
537 Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49: D412–
538 D419. doi:10.1093/nar/gkaa913

539 43. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo
540 generator. Genome Res. 2004;14: 1188–1190. doi:10.1101/gr.849004

541 44. Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, et al. Global trends of whole-
542 genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature. 2006;444:
543 171–178. doi:10.1038/nature05230

544 45. Slabodnick MM, Ruby JG, Reiff SB, Swart EC, Gosai S, Prabakaran S, et al. The
545 Macronuclear Genome of *Stentor coeruleus* Reveals Tiny Introns in a Giant Cell. Curr
546 Biol. 2017;27: 569–575. doi:10.1016/j.cub.2016.12.057

547 46. Arnaiz O, Meyer E, Sperling L. ParameciumDB 2019: integrating genomic data across
548 the genus for functional and evolutionary biology. Nucleic Acids Res. 2020;48: D599–
549 D605. doi:10.1093/nar/gkz948

550 47. Sheng Y, Duan L, Cheng T, Qiao Y, Stover NA, Gao S. The completed macronuclear
551 genome of a model ciliate *Tetrahymena thermophila* and its application in genome
552 scrambling and copy number analyses. Sci China Life Sci. 2020;63: 1534–1542.
553 doi:10.1007/s11427-020-1689-4

554 48. Stover NA, Krieger CJ, Binkley G, Dong Q, Fisk DG, Nash R, et al. Tetrahymena
555 Genome Database (TGD): a new genomic resource for Tetrahymena thermophila
556 research. Nucleic Acids Res. 2006;34: D500-3. doi:10.1093/nar/gkj054

557 49. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al.
558 BUSCO applications from quality assessments to gene prediction and phylogenomics.
559 Mol Biol Evol. 2018;35: 543–548. doi:10.1093/molbev/msx319

560 50. Chen X, Jiang Y, Gao F, Zheng W, Krock TJ, Stover NA, et al. Genome analyses of
561 the new model protist *Euplotes vannus* focusing on genome rearrangement and
562 resistance to environmental stressors. Mol Ecol Resour. 2019;19: 1292–1308.
563 doi:10.1111/1755-0998.13023

564 51. Coyne RS, Hannick L, Shanmugam D, Hostetler JB, Brami D, Joardar VS, et al.
565 Comparative genomics of the pathogenic ciliate *Ichthyophthirius multifiliis*, its free-
566 living relatives and a host species provide insights into adoption of a parasitic lifestyle
567 and prospects for disease control. Genome Biol. 2011;12: R100. doi:10.1186/gb-2011-
568 12-10-r100

569 52. Swart EC, Bracht JR, Magrini V, Minx P, Chen X, Zhou Y, et al. The *Oxytricha trifallax*
570 macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes.
571 PLoS Biol. 2013;11: e1001473. doi:10.1371/journal.pbio.1001473

572 53. Xiong J, Wang G, Cheng J, Tian M, Pan X, Warren A, et al. Genome of the facultative
573 scuticociliatosis pathogen *Pseudocohnilembus persalinus* provides insight into its
574 virulence through horizontal gene transfer. Sci Rep. 2015;5: 15470.
575 doi:10.1038/srep15470

576 54. Aeschlimann SH, Jönsson F, Postberg J, Stover NA, Petera RL, Lipps H-J, et al. The

577 draft assembly of the radically organized *Stylonychia lemnae* macronuclear genome.
578 Genome Biol Evol. 2014;6: 1707–1723. doi:10.1093/gbe/evu139

579 55. Hamilton EP, Kapusta A, Huvos PE, Bidwell SL, Zafar N, Tang H, et al. Structure of
580 the germline genome of *Tetrahymena thermophila* and relationship to the massively
581 rearranged somatic genome. eLife. 2016;5. doi:10.7554/eLife.19090

582 56. McGrath CL, Gout J-F, Doak TG, Yanagi A, Lynch M. Insights into three whole-
583 genome duplications gleaned from the *Paramecium caudatum* genome sequence.
584 Genetics. 2014;197: 1417–1428. doi:10.1534/genetics.114.163287

585 57. Arnaiz O, Van Dijk E, Bétermier M, Lhuillier-Akakpo M, de Vanssay A, Duharcourt S,
586 et al. Improved methods and resources for *Paramecium* genomics: transcription units,
587 gene annotation and gene expression. BMC Genomics. 2017;18: 483.
588 doi:10.1186/s12864-017-3887-z

589 58. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high
590 throughput. Nucleic Acids Res. 2004;32: 1792–1797. doi:10.1093/nar/gkh340