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 diverse 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 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 (doi:10.17617/3.XWMBKT, Table S1). 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) (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 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.

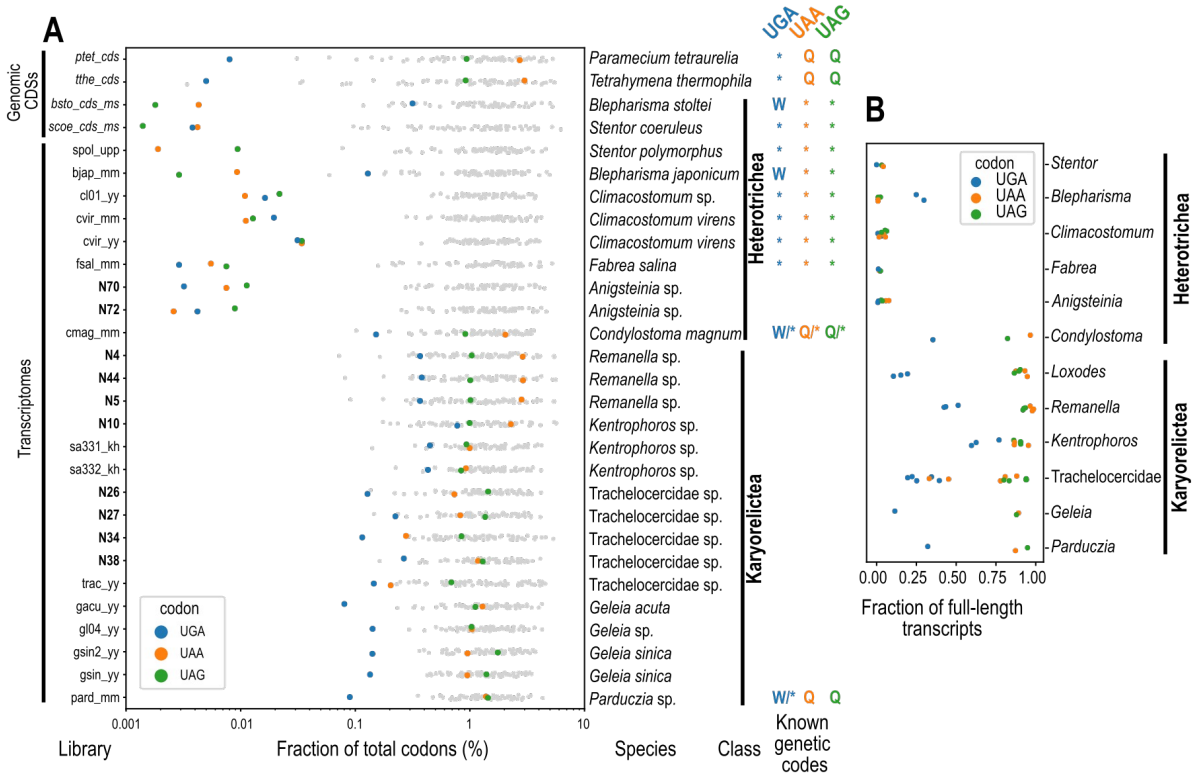


Figure 1. (A) Codon frequencies of canonical stop codons (UGA: blue, UAA: orange, UAG: green) and other codons (gray) in conserved protein domains found by hmmscan search in six-frame translations of transcriptome assemblies (doi:10.17617/3.XWMBKT, Table S1) or genomic CDSs (doi:10.17617/3.XWMBKT, Table S2) vs. Pfam. Names of libraries from this study are highlighted in bold. Assignments of canonical stops for organisms with known genetic codes follow Swart et al., 2016. (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).

105 In-frame UGAs were found in 10.5 to 76.9% of transcripts with putative coding regions
106 predicted by full-length Blastx hits per karyorelict library (Figure 1B). This frequency verified
107 that in-frame UGAs were not concentrated in a small fraction of potentially spurious
108 sequences but in fact found in many genes. Conserved “marker” genes that were generally
109 expected to be present in ciliate genomes (BUSCO orthologs, Alveolata marker set, [22])
110 also contained in-frame UGAs. The karyorelict transcriptome assemblies were relatively
111 incomplete, with 1.8% to 20.5% (median 12.0%) estimated completeness based on the
112 BUSCO markers, and a total of 91 of 171 BUSCO orthologs were found in these assemblies
113 (Figure 2A). Nonetheless, 46 BUSCO orthologs from 14 karyorelict assemblies were found
114 with in-frame UGAs in conserved alignment positions (e.g. Figure 2B, 2C), verifying that they
115 are not limited to poorly characterized or hypothetical proteins.

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

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

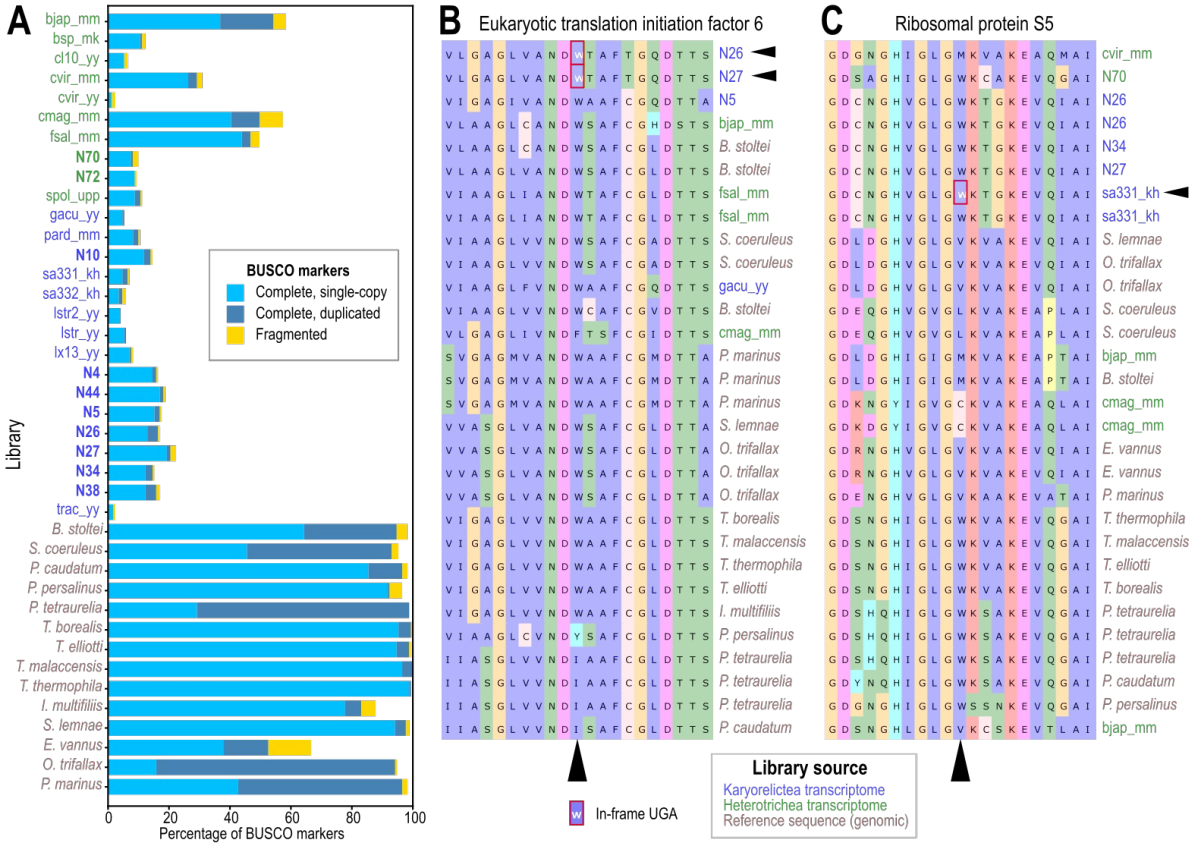


Figure 2. In-frame coding UGAs in conserved marker genes. (**A**) Completeness estimates of heterotrich and karyorelict transcriptomes (library names in green and blue respectively), compared with genomic reference sequences from other ciliates (doi:10.17617/3.XWMBKT, Table S3); BUSCO Alveolata marker 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.

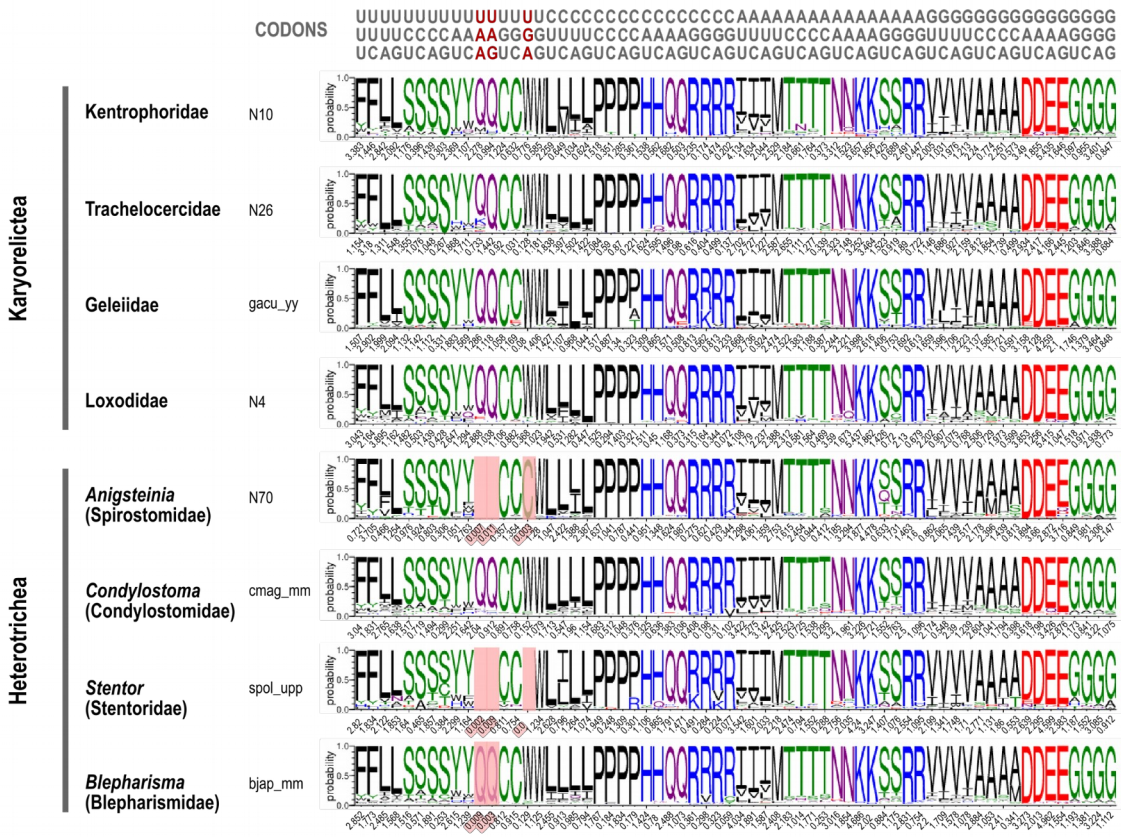
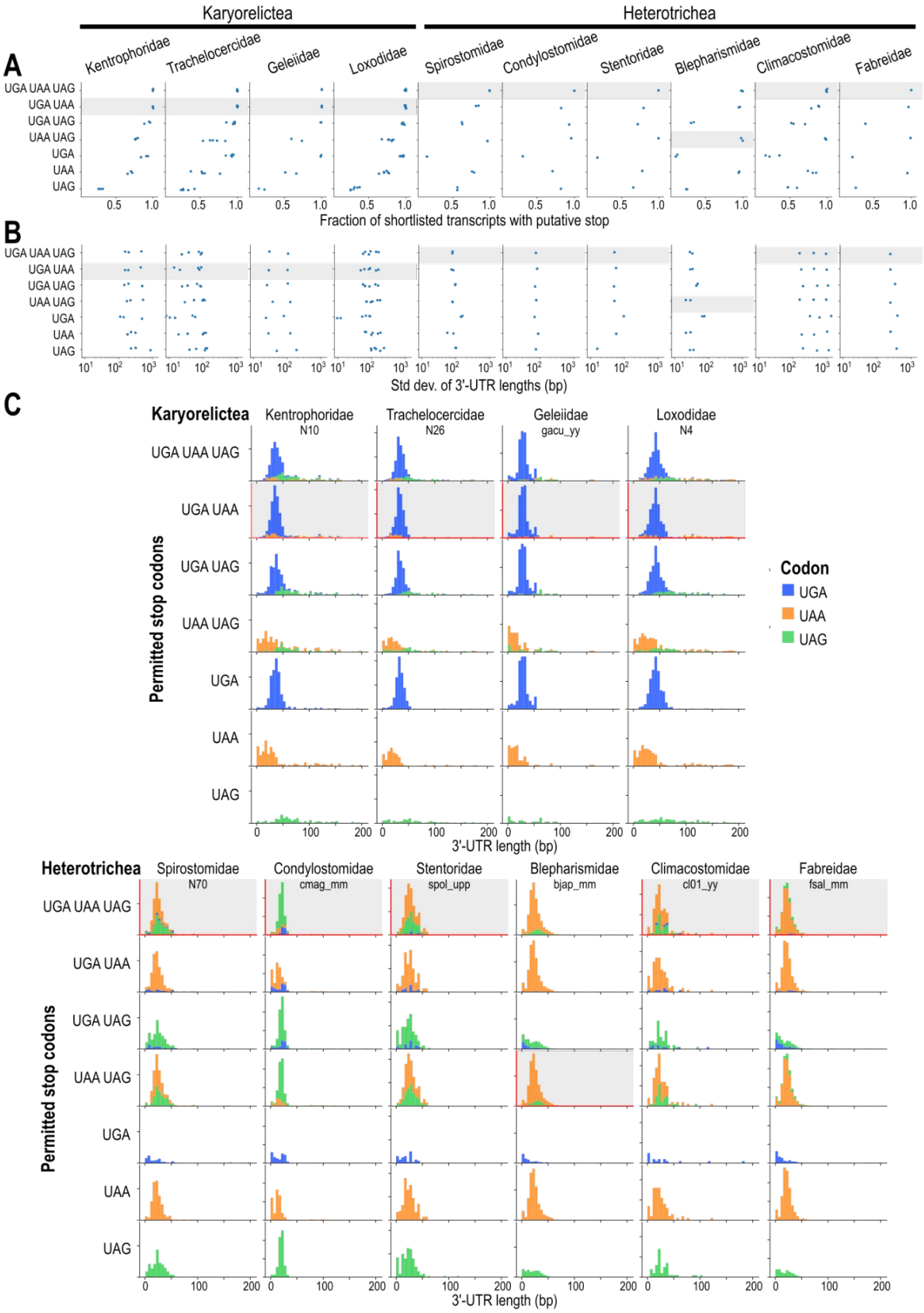


Figure 3. Weblogos representing the likely amino acid assignment of each codon in selected libraries (library with most coverage per taxon of interest). Heights of each letter represent the relative frequencies (all scaled to 100%) of each amino acid in conserved residues aligning to that codon. The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

141 *Stop codons in karyorelicts and heterotrichs*

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

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



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

170 codon usage for each taxon from this study or previous publications highlighted in gray. (**A**)
171 Strip plots for the fraction of full length contigs per transcriptome that have a putative stop
172 codon from that specific combination (rows), i.e. in-frame, downstream of full-length Blastx
173 hit vs. reference, and upstream of poly-A tail. Each point corresponds to one transcriptome
174 assembly, grouped by taxonomic family (columns). (**B**) Scatterplots for standard deviation of
175 3'-UTR lengths. (**C**) Histograms for 3'-UTR lengths, colored by putative stop codon (UGA:
176 blue, UAA: orange, UAG: green), one representative library per family.

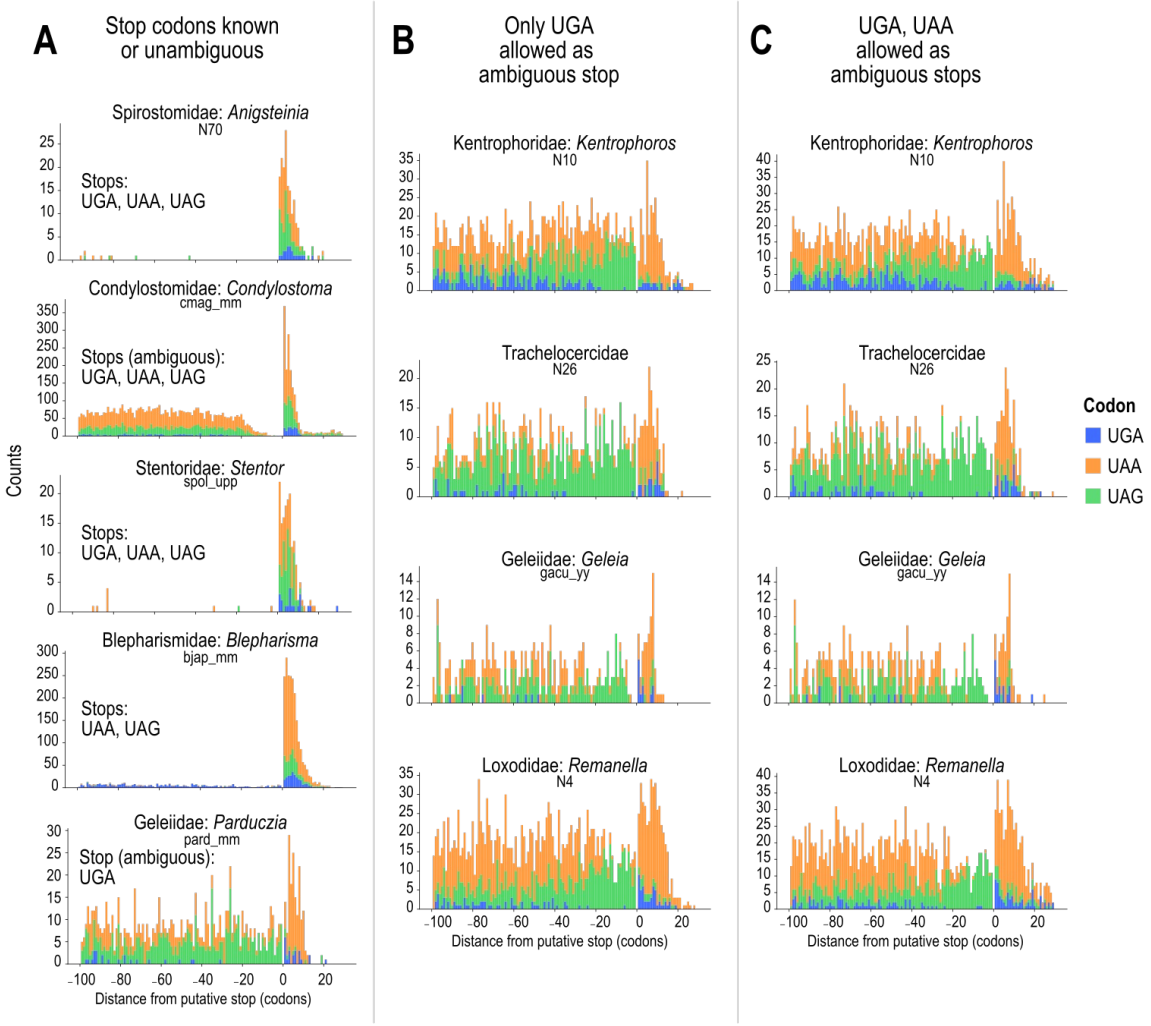


Figure 5. Depletion of in-frame coding "stop" codons in the coding sequence (negative coordinates) immediately before the putative true stop codon (position 0) and their enrichment in the 3'-UTR (positive coordinates). Representative library with highest number of assembled full length contigs chosen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

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

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

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).

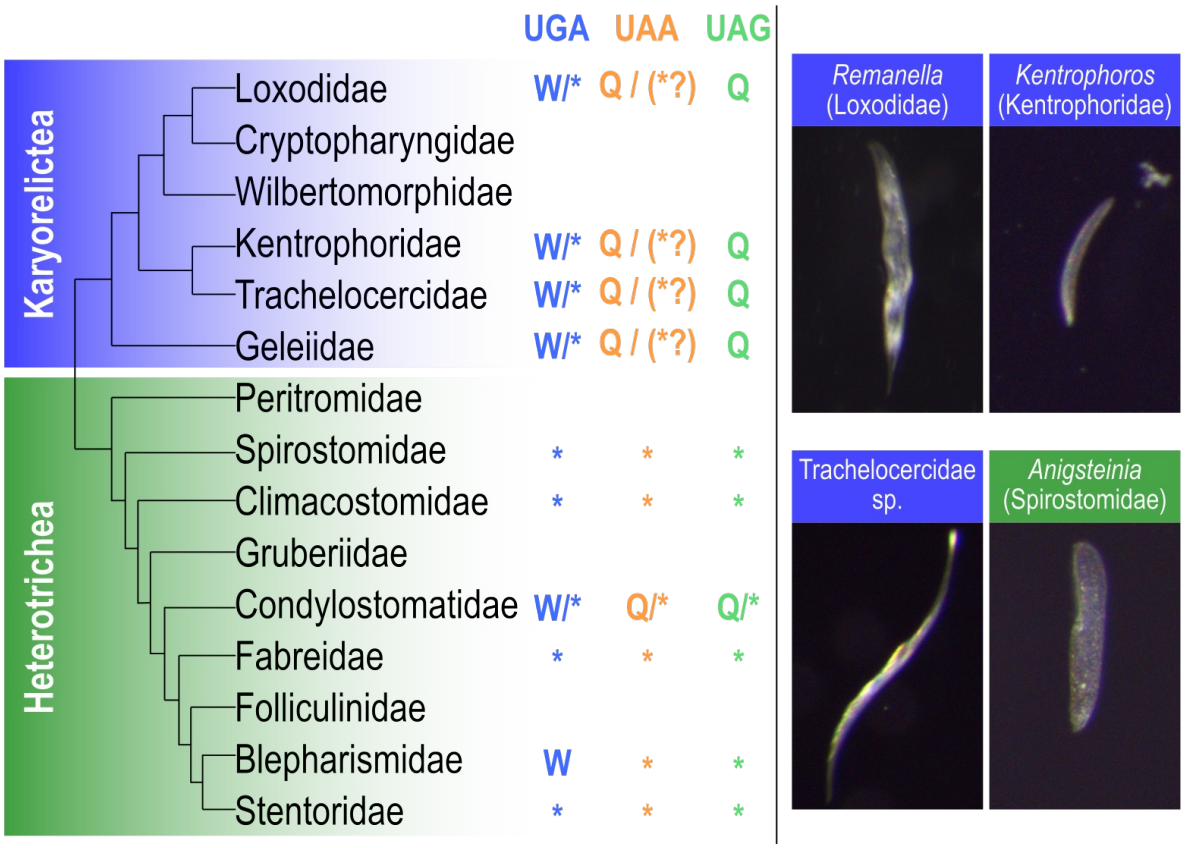


Figure 6. Genetic code diversity among karyorelict and heterotrich ciliates. (**Left**) Diagrammatic karyorelict + heterotrich tree with predicted stop codon reassignments mapped to each family. Subtree topologies are from Ma et al. (2022) and Fernandes et al. (2016) respectively. Branch lengths are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

218 Available data for *Cryptopharynx* (Karyorelictea: Cryptopharyngidae) were not conclusive.
219 The canonical stop codons had frequencies between 0.02 and 0.07%, lower than for other
220 karyorelicts, but higher than true stop codons, but Cryptopharyngidae was represented by a
221 single library that had high contamination from other eukaryotes (Supplementary Text) and
222 there were too few high-confidence, full length transcripts for a reliable conclusion on its
223 genetic code. No sequence data beyond rRNA genes were publicly available for the
224 remaining family, the monotypic Wilbertomorphidae, whose phylogenetic position in relation
225 to the other karyorelicts is unclear because of long branch lengths, and which has to our
226 knowledge only been reported once [23].

227 Ambiguous stop/sense codes are hence not just isolated phenomena, but are used by a
228 major taxon that is diverse, globally distributed, and common in its respective habitats. In
229 contrast, the heterotrichs, which constitute the sister group to Karyorelictea and are hence of
230 the same evolutionary age, use at least three different genetic codes, including one with
231 ambiguous stops (Figure 6). If organisms with ambiguous codes were isolated single
232 species whose nearest relatives have conventional stops, as appears to be the case for
233 *Blastocrithidia* spp. and *Amoebophrya* sp., we might conclude that these are uncommon
234 occurrences that do not persist over longer evolutionary time scales. However, the
235 karyorelict crown group diversified during the Proterozoic (posterior mean 455 Mya) and the
236 stem split from the Heterotrichea even earlier, in the Neo-Proterozoic [24].

237 This study has benefited from several technical improvements. A highly complete,
238 contiguous genome assembly with gene predictions is now available for the heterotrich
239 *Blepharisma stoltei* [25]. Because *Blepharisma* is more closely related to the karyorelicts
240 than other ciliate model species, which are mostly oligohymenophorans and spirotrichs, it
241 improved the reference-based annotation of the assembled transcriptomes. Single-cell RNA-
242 seq libraries in this study were also sequenced to a greater depth, with a lower fraction of
243 contamination from rRNA, and hence yielded more full length mRNA transcripts for analysis.

244 One proposed mechanism for how the cell correctly recognizes whether an ambiguous
245 codon is coding or terminal is based on the proximity of translation stops to the poly-A tail of
246 transcripts. In this model, tRNAs typically bind more efficiently to in-frame coding “stops”
247 than eukaryotic translation release factor 1 (eRF1), hence allowing these codons to be
248 translated. At the true termination stop codon, however, the binding of eRF1 can be
249 stabilized by interactions with poly-A interacting proteins like PABP bound to the nearby
250 poly-A tail, allowing it to outcompete tRNAs and hydrolyze the peptidyl-tRNA bond [10,26].
251 Consistent with this model, we found that karyorelict 3’-UTRs are also relatively short, and

252 that in-frame UGAs are depleted immediately before the putative true stop codon.
253 Nonetheless, karyorelict 3'-UTRs are actually about 10 nt longer on average than those of
254 heterotrichs.

255 Our results raised the possibility that UAA is also used as an ambiguous stop codon for ~1-
256 10% of karyorelict transcripts, in addition to the main stop codon UGA. eRF1 may retain a
257 weak affinity for UAA, and recognize UAA for terminating translation albeit with lower
258 efficiency. In *Blepharisma japonicum*, where UAA and UAG are non-ambiguous stops and
259 UGA encodes tryptophan (albeit at low frequency, 0.13%), heterologously expressed eRF1
260 could still recognize all three codons in an *in vitro* assay, although efficiency of peptidyl-tRNA
261 hydrolysis was lower with UGA than for UAA and UAG [27]. In species with non-ambiguous
262 stop codon reassignment, the effect of such "weak" ambiguity on the total pool of translated
263 protein may be negligible, but it shows that there is a latent potential that could account for
264 the repeated evolution of stop codon reassignments in ciliates. Furthermore, UAAs were
265 even more abundant than UGAs in ciliate 3'-UTRs, which can be attributed to the low GC%
266 of 3'-UTRs compared to coding sequences; other A/U-only codons were also enriched in 3'-
267 UTRs. Therefore, UAAs in the 3'-UTRs of karyorelicts may be a "backstop" mechanism that
268 prevents occasional stop-codon readthrough, as proposed for tandem stop codons (TSCs) in
269 other species with reassigned stop codons [28]. In the minority of transcripts where in-frame
270 stop-UGA is absent, the backstop may be adequate to terminate translation before the poly-
271 A tail and produce a functional protein most of the time. To verify our predictions that UGA is
272 the main stop codon and UAA a lower-frequency alternative stop, ribosome profiling and
273 mass spectrometry detection of peptide fragments corresponding to the expected 3'-ends of
274 coding sequences, e.g. as performed on Condylostoma [10], are the most applicable
275 experimental methods. If a karyorelict species can be developed into a laboratory model
276 amenable to genetic transformation, manipulation of the 3'-UTR length and sequence would
277 allow us to test the "backstop" hypothesis directly and tease apart the factors contributing to
278 translation termination in these organisms.

279 What selective pressures might favor the evolution and maintenance of an ambiguous
280 genetic code? One possibility is that context-dependent sense/stop codons confer
281 mutational robustness by eliminating substitutions that cause premature stop codons.
282 Ambiguous codes do not appear to be linked to a specific habitat: *Blastocrithidia* spp. and
283 *Amoebophrya* sp. are both parasites of eukaryotic hosts, but of insects and free-living
284 dinoflagellates respectively; whereas the karyorelict ciliates and *Condylostoma* are both
285 found in marine interstitial environments, but live alongside other ciliates that have

286 conventional codes, such as *Anigsteinia*. Having short 3'-UTRs may predispose ciliates to
287 adopt ambiguous codes by facilitating interactions between eRF1 and PABPs that could
288 enable stop recognition, but other factors, including simply contingent evolution, appear to
289 have led to their evolution because the 3'-UTRs of ciliates with conventional stop codons
290 are also comparably short, particularly among the heterotrichs.

291 Any adaptationist hypothesis for alternative and ambiguous codes will have to contend with
292 the existence of related organisms with conventional codes that have similar lifestyles.
293 Furthermore, once a stop codon has been reassigned to sense, it becomes increasingly
294 difficult to undo without the deleterious effects of premature translation termination, and may
295 function like a ratchet. Like the origins of the genetic code itself [29], we may have to be
296 content with the null hypothesis that they are "frozen accidents" that reached fixation
297 stochastically, and which are maintained because they do not pose a significant selective
298 disadvantage.

299 ## Materials and Methods

300 *Sample collection*

301 Surface sediment was sampled in September 2021 from two sites in the bay at Roscoff,
302 France when exposed at low tide. Site A: shallow swimming enclosure, 48.72451 N,
303 3.992294 W; Site B: adjacent to green algae tufts near freshwater outflow, 48.716169 N,
304 3.995626 W. Upper 1-2 cm of sediment was skimmed into glass beakers, and stored under
305 local seawater until use. Interstitial ciliates were collected by decantation: a spoonful of
306 sediment was stirred in seawater in a beaker. Sediment particles were briefly allowed to
307 settle out, and the overlying suspended organic material was decanted into Petri dishes.
308 Ciliate cells were preliminarily identified by morphology under a dissection microscope and
309 picked by pipetting with sterile, filtered pipette tips. Selected cells were imaged with incident
310 light under a stereo microscope (Olympus SZX10, Lumenera Infinity 3 camera).

311 NEBNext cell lysis buffer (NEB, E5530S) was premixed and filled into PCR tubes; per tube:
312 0.8 µL 10x cell lysis buffer, 0.4 µL murine RNAse inhibitor, 5.3 µL nuclease-free water.
313 Picked ciliate cells were transferred twice through filtered local seawater (0.22 µm, Millipore
314 SLGP033RS) to wash, then transferred with 1.5 µL carryover volume to 6.5 µL of cell lysis
315 buffer (final volume 8 µL), and snap frozen in liquid nitrogen. Samples were stored at -80 °C
316 before use.

317 *Single-cell RNAseq sequencing*

318 Samples collected in cell lysis buffer (doi:10.17617/3.XWMBKT, Table S1) were used for
319 RNAseq library preparation with the NEBNext Single Cell / Low Input RNA Library Prep Kit
320 for Illumina (NEB, E6420S), following the manufacturer's protocol for single cells, with the
321 following parameters adjusted: 17 cycles for cDNA amplification PCR, cDNA input for library
322 enrichment normalized to 3 ng (or all available cDNA used for libraries where total cDNA
323 was <3 ng), 8 cycles for library enrichment PCR. Libraries were dual-indexed (NEBNext
324 Dual Index Primers Set 1, NEB E7600S), and sequenced on an Illumina NextSeq 2000
325 instrument with P3 300 cycle reagents, with target yield of 10 Gbp per library.

326 *RNA-seq library quality control and transcriptome assembly*

327 Previously published karyorelict transcriptome data [17,30–32] were downloaded from the
328 European Nucleotide Archive (ENA) (doi:10.17617/3.XWMBKT, Table S1). Contamination
329 from non-target organisms was evaluated by mapping reads to an rRNA reference database

330 and summarizing the hits by taxonomy. Although RNAseq library construction enriches
331 mRNAs using poly-A tail selection, there is typically still sufficient rRNA present in the final
332 library to evaluate the taxonomic composition of the sample. All RNAseq read libraries
333 (newly sequenced and previously published) were processed with the same pipeline: The
334 taxonomic composition of each library was evaluated by mapping 1 M read pairs per library
335 against the SILVA SSU Ref NR 132 database [33], using phyloFlash v3.3b1 [34]. Newly
336 sequenced libraries were assigned to a genus or family using the mapping-based taxonomic
337 summary, or full-length 18S rRNA gene if it was successfully assembled.

338 Reads were trimmed with the program bbduk.sh from BBmap v38.22 (http://sourceforge.net/
339 projects/bbmap/) to remove known adapters (right end) and low-quality bases (both ends),
340 with minimum Phred quality 24 and minimum read length 25 bp. Trimmed reads were then
341 assembled with Trinity v2.12.0 [35] using default parameters. Assembled contigs were
342 aligned against the *Blepharisma stoltei* ATCC 30299 proteome [25] with NCBI Blastx v2.12.0
343 [36] using the standard genetic code and E-value cutoff 10^{-20}, parallelized with GNU Parallel
344 [37].

345 Morphological identifications of the newly collected samples were verified with 18S rRNA
346 sequences from the Trinity transcriptome assemblies. rRNA sequences were annotated with
347 barrnap v0.9. 18S rRNA sequences ≥80% of full length were extracted, except for two
348 libraries (N4, N26) where the longest sequences were <80% and for which the two longest
349 18S rRNA sequences were extracted instead. For comparison, reference sequences for
350 Karyorelictea and Heterotrichea above 1400 bp from the PR2 database v4.14.0 [38] were
351 used. Representative reference sequences were chosen by clustering at 99% identity with
352 the cluster_fast method using Vsearch v2.13.6 [39]. Extracted and reference sequences
353 were aligned with MAFFT v7.505 [40]. A phylogeny (Figure S3) was inferred from the
354 alignment with IQ-TREE v2.0.3 [41], using the TIM2+F+I+G4 model found as the best-fitting
355 model by ModelFinder [42]. Alignment and tree files are available from
356 doi:10.17617/3.QLWR38. 18S rRNA sequences were deposited in the European Nucleotide
357 Archive under accessions OX095806-OX095846.

358 Read pre-processing, quality control, and assembly were managed with a Snakemake
359 v6.8.1 [43] workflow (https://github.com/Swart-lab/karyocode-workflow, archived at
360 doi:10.5281/zenodo.6647650). Scripts for data processing described below were written in
361 Python v3.7.3 using Biopython v1.74 [44], pandas v0.25.0 [45], seaborn v0.11.0 [46] and
362 Matplotlib v3.1.1 [47] libraries unless otherwise stated.

363 *Prediction of stop codon reassignment to sense*

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

371 Codon frequencies and their corresponding amino acids were predicted with an updated
372 version of PORC (v2.1, https://github.com/Swart-lab/PORC, archived at
373 doi:10.5281/zenodo.6784075; managed with a Snakemake workflow,
374 https://github.com/Swart-lab/karyocode-analysis-porc, archived at
375 doi:10.5281/zenodo.6647652); the method has been previously described [10,48]. Briefly: a
376 six-frame translation was produced for each contig in the transcriptome assembly, and
377 searched against conserved domains in the Pfam-A database v32 [49] with hmmscan from
378 HMMer v3.3.2 (http://hmmer.org/). Overall codon frequencies were counted from alignments
379 with E-value ≤ 10^{-20}. To ensure that there was sufficient data underlying the codon and
380 amino acid frequencies, only those libraries with at least 100 observations for each of the
381 coding codons in the standard genetic code were used for comparison of codon frequencies
382 and for prediction of amino acid assignments.

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

387 In addition to the transcriptomes, genomic CDSs of selected model species with different
388 genetic codes [25,51–55] were also analyzed with PORC to obtain a reference baseline of
389 coding-codon frequencies (doi:10.17617/3.XWMBKT, Table S2). These model species have
390 non-ambiguous codes so they were not expected to have stop codons in the CDSs, except
391 for the terminal stop.

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

393 "Full-length" transcripts (with poly-A tail, intact 3'-UTR, and complete coding sequence) were
394 desirable to predict the stop codon, characterize 3'-UTR metrics, and verify genetic code
395 predictions. Contigs were therefore filtered with the following criteria: (i) poly-A tail ≥7 bp,

396 criterion following [10], (ii) contig contains a Blastx hit vs. *B. stoltei* protein sequence with E-
397 value ≤10^{-20} and where the alignment covers ≥80% of the reference *B. stoltei* sequence, (iii)
398 both poly-A tail and Blastx hit agree on the contig orientation. For contigs with multiple
399 isoforms assembled by Trinity, the isoform with the longest Blastx hit was chosen; in case of
400 a Blastx hit length tie, then the longer isoform was chosen. Only libraries with >100
401 assembled "full-length" transcripts were used for downstream analyses (Supplementary
402 Text).

403 *Metrics for evaluating potential stop codon combinations*

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

411 *Delimitation of putative coding sequences using Blastx hits*

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

420 *Verification of in-frame UGAs in conserved marker genes*

421 Full-length CDSs (see above) were translated with the karyorelict code (NCBI table 27).
422 Conserved marker genes were identified with BUSCO v5.2.2 (protein mode,
423 alveolata_odb10 marker set) [22], managed with a Snakemake workflow (https://github.com/
424 Swart-lab/karyocode-analysis-busco, archived at doi:10.5281/zenodo.6647679). Markers for
425 additional ciliate species where relatively complete genome assemblies and gene
426 predictions were available were also identified (doi:10.17617/3.XWMBKT, Table S3)
427 [52,54,56–63]. For each BUSCO marker, the ciliate homologs were aligned with Muscle

428 v3.8.1551 [64]. Alignment columns corresponding to in-frame putatively coding UGAs of
429 karyorelict sequences were identified. These positions were considered to be conserved if
430 ≥50% of residues were W or another aromatic amino acid (Y, F, or H).

431 Data availability

432 RNA-seq libraries sequenced for this study have been deposited at the European Nucleotide
433 Archive (https://www.ebi.ac.uk/ena/) under accession PRJEB50648. Lists of dataset
434 accessions for each analysis (doi:10.17617/3.XWMBKT) and the 18S rRNA phylogeny
435 (doi:10.17617/3.QLWR38) have been deposited at Edmond.

436 Supplementary Information

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

438 Funding

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

442 Acknowledgements

443 We thank S. Booker, C. Cabresin, E. Bourrigaud, and R. Garnier for facilitating our research
444 visit to the Station Biologique Roscoff; H. Budde for sequencing; H. Gruber-Vodicka for loan
445 of equipment; A. Peters for helpful suggestions and cake; A. Lemper for administrative
446 assistance; Y. Shulgina for thoughtful feedback on the manuscript, and members of the
447 Swart Lab for helpful feedback and support.

448 Conflict of interest disclosure

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

450 References

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

454 2. Shulgina Y, Eddy SR. A computational screen for alternative genetic codes in over
455 250,000 genomes. eLife. 2021;10. doi:10.7554/eLife.71402

456 3. Kollmar M, Mühlhausen S. Nuclear codon reassignments in the genomics era and

457 mechanisms behind their evolution. Bioessays. 2017;39. doi:10.1002/bies.201600221

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

460 5. Suzuki T, Ueda T, Watanabe K. The “polysemous” codon--a codon with multiple amino
461 acid assignment caused by dual specificity of tRNA identity. EMBO J. 1997;16: 1122–
462 1134. doi:10.1093/emboj/16.5.1122

463 6. Hatfield DL, Gladyshev VN. How selenium has altered our understanding of the
464 genetic code. Mol Cell Biol. 2002;22: 3565–3576. doi:10.1128/MCB.22.11.3565-
465 3576.2002

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

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

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

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

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

481 12. Preer JR, Preer LB, Rudman BM, Barnett AJ. Deviation from the universal code shown
482 by the gene for surface protein 51A in *Paramecium*. Nature. 1985;314: 188–190.
483 doi:10.1038/314188a0

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

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

489 15. Tourancheau AB, Tsao N, Klobutcher LA, Pearlman RE, Adoutte A. Genetic code
490 deviations in the ciliates: evidence for multiple and independent events. EMBO J.
491 1995;14: 3262–3267. doi:10.1002/j.1460-2075.1995.tb07329.x

492 16. Helftenbein E. Nucleotide sequence of a macronuclear DNA molecule coding for
493 alpha-tubulin from the ciliate *Stylonychia lemnae*. Special codon usage: TAA is not a
494 translation termination codon. Nucleic Acids Res. 1985;13: 415–433. doi:10.1093/nar/
495 13.2.415

496 17. Yan Y, Maurer-Alcalá XX, Knight R, Kosakovsky Pond SL, Katz LA. Single-Cell
497 Transcriptomics Reveal a Correlation between Genome Architecture and Gene Family

498 Evolution in Ciliates. MBio. 2019;10. doi:10.1128/mBio.02524-19

499 18. Raikov IB. Primitive never-dividing macronuclei of some lower ciliates. Int Rev Cytol.
500 1985;95: 267–325. doi:10.1016/s0074-7696(08)60584-7

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

505 20. Ma M, Li Y, Maurer-Alcalá XX, Wang Y, Yan Y. Deciphering phylogenetic relationships
506 in class Karyorelictea (Protista, Ciliophora) based on updated multi-gene information
507 with establishment of a new order Wilbertomorphida n. ord. Mol Phylogenet Evol.
508 2022; 107406. doi:10.1016/j.ympev.2022.107406

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

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

515 23. Xu Y, Li J, Song W, Warren A. Phylogeny and establishment of a new ciliate family,
516 Wilbertomorphidae fam. nov. (Ciliophora, Karyorelictea), a highly specialized taxon
517 represented by *Wilbertomorpha colpoda* gen. nov., spec. nov. J Eukaryot Microbiol.
518 2013;60: 480–489. doi:10.1111/jeu.12055

519 24. Fernandes NM, Schrago CG. A multigene timescale and diversification dynamics of
520 Ciliophora evolution. Mol Phylogenet Evol. 2019;139: 106521.
521 doi:10.1016/j.ympev.2019.106521

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

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

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

531 28. Fleming I, Cavalcanti ARO. Selection for tandem stop codons in ciliate species with
532 reassigned stop codons. PLoS ONE. 2019;14: e0225804.
533 doi:10.1371/journal.pone.0225804

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

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

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

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

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

549 34. Gruber-Vodicka HR, Seah BKB, Pruesse E. phyloFlash: Rapid Small-Subunit rRNA
550 Profiling and Targeted Assembly from Metagenomes. mSystems. 2020;5. doi:10.1128/
551 mSystems.00920-20

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

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

558 37. Tange O. Gnu Parallel 2018. Zenodo. 2018. doi:10.5281/zenodo.1146014

559 38. Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, et al. The Protist Ribosomal
560 Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA
561 sequences with curated taxonomy. Nucleic Acids Res. 2013;41: D597-604.
562 doi:10.1093/nar/gks1160

563 39. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source
564 tool for metagenomics. PeerJ. 2016;4: e2584. doi:10.7717/peerj.2584

565 40. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7:
566 improvements in performance and usability. Mol Biol Evol. 2013;30: 772–780.
567 doi:10.1093/molbev/mst010

568 41. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A,
569 et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the
570 genomic era. Mol Biol Evol. 2020;37: 1530–1534. doi:10.1093/molbev/msaa015

571 42. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder:
572 fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14: 587–
573 589. doi:10.1038/nmeth.4285

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

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

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

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

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

48. 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. doi:10.1093/bioinformatics/btr316

49. 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. doi:10.1093/nar/gkaa913

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

51. 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. doi:10.1038/nature05230

52. 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. doi:10.1016/j.cub.2016.12.057

53. 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. doi:10.1093/nar/gkz948

54. 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. doi:10.1007/s11427-020-1689-4

55. 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. doi:10.1093/nar/gkj054

56. 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. doi:10.1111/1755-0998.13023

57. 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. doi:10.1186/gb-2011-12-10-r100

58. Swart EC, Bracht JR, Magrini V, Minx P, Chen X, Zhou Y, et al. The *Oxytricha trifallax*

621 macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes.
622 PLoS Biol. 2013;11: e1001473. doi:10.1371/journal.pbio.1001473

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

627 60. Aeschlimann SH, Jönsson F, Postberg J, Stover NA, Petera RL, Lipps H-J, et al. The
628 draft assembly of the radically organized *Stylonychia lemnae* macronuclear genome.
629 Genome Biol Evol. 2014;6: 1707–1723. doi:10.1093/gbe/evu139

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

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

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

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