PLOS GENETICS

RESEARCH ARTICLE

Identification of a non-canonical ciliate nuclear genetic code where UAA and UAG code for different amino acids

Jamie McGowan[1], Estelle S. Kilias[2], Elisabet Alacid[2], James Lipscombe[1], Benjamin H. Jenkins[2], Karim Gharbi[1], Gemy G. Kaithakottil[1], Iain C. Macaulay[1], Seanna McTaggart[1], Sally D. Warring[1], Thomas A. Richards[2]*, Neil Hall[1,3]*, David Swarbreck[1]*

1 Earlham Institute, Norwich Research Park, Norwich, United Kingdom, **2** Department of Biology, University of Oxford, Oxford, United Kingdom, **3** School of Biological Sciences, University of East Anglia, Norwich, United Kingdom

* thomas.richards@biology.ox.ac.uk (TAR); neil.hall@earlham.ac.uk (NH); david.swarbreck@earlham.ac.uk (DS)

Citation: McGowan J, Kilias ES, Alacid E, Lipscombe J, Jenkins BH, Gharbi K, et al. (2023) Identification of a non-canonical ciliate nuclear genetic code where UAA and UAG code for different amino acids. PLoS Genet 19(10): e1010913. https://doi.org/10.1371/journal.pgen.1010913

Editor: Kenneth H. Wolfe, University College Dublin, IRELAND

Received: July 4, 2023

Accepted: August 10, 2023

Published: October 5, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1010913

Data Availability Statement: All sequencing data and the genome assembly of Oligohymenophorea sp. PL0344 have been deposited to the European

Abstract

The genetic code is one of the most highly conserved features across life. Only a few lineages have deviated from the "universal" genetic code. Amongst the few variants of the genetic code reported to date, the codons UAA and UAG virtually always have the same translation, suggesting that their evolution is coupled. Here, we report the genome and transcriptome sequencing of a novel uncultured ciliate, belonging to the Oligohymenophorea class, where the translation of the UAA and UAG stop codons have changed to specify different amino acids. Genomic and transcriptomic analyses revealed that UAA has been reassigned to encode lysine, while UAG has been reassigned to encode glutamic acid. We identified multiple suppressor tRNA genes with anticodons complementary to the reassigned codons. We show that the retained UGA stop codon is enriched in the 3'UTR immediately downstream of the coding region of genes, suggesting that there is functional drive to maintain tandem stop codons. Using a phylogenomics approach, we reconstructed the ciliate phylogeny and mapped genetic code changes, highlighting the remarkable number of independent genetic code changes within the Ciliophora group of protists. According to our knowledge, this is the first report of a genetic code variant where UAA and UAG encode different amino acids.

Author summary

The genetic code is almost universal across life. The vast majority of organisms use the canonical genetic code, which has three stop codons (UAA, UAG, and UGA) and 61 sense codons that code for amino acids. Here, we report the discovery of an unexpected genetic code variant in an uncultured ciliate species from the Oligohymenophorea class, where the canonical stop codons UAA and UAG have been reassigned to code for lysine and glutamic acid, respectively. This is a particularly unusual genetic code reassignment

Nucleotide Archive under the study accession PRJEB58266. Additional supporting data have been deposited on Zenodo (10.5281/zenodo.7944379).

Funding: This work was funded by Wellcome though the Darwin Tree of Life Discretionary Award (218328 to NH and TAR) and supported by the Biotechnology and Biological Sciences Research Council (BBSRC), part of UK Research and Innovation, through the Core Capability Grant (BB/CCG1720/1 to NH), the National Capability in Genomics and Single Cell Analysis (BBS/E/T/000PR9816 to DS) and the National Capability in e-Infrastructure (BBS/E/T/000PR9814 to NH) at the Earlham Institute. TAR is supported by a Royal Society University Research Fellowship (URF/R/191005 to TAR). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

as UAA and UAG differ at the wobble position and their evolution is thought to be coupled. We also report that the remaining stop codon, UGA, is enriched immediately downstream of genes in the same reading frame, suggesting a possible role in minimising deleterious consequences in the event of translational readthrough. Our work documents, for the first time, a genetic code variant where the codons UAA and UAG specify two different amino acids and shows that there are still unexplored genetic code reassignments awaiting discovery.

Introduction

The genetic code is one of the most conserved features across life, emerging before the last universal common ancestor [1]. Virtually all organisms use the canonical genetic code which has three stop codons (UAA, UAG, and UGA) and 61 sense codons that specify one of 20 amino acids, including a translation start codon (AUG). Variants of the genetic code, while rare, have been reported in several lineages of bacteria, viruses, and eukaryotic organellar and nuclear genomes [2,3]. Ciliate nuclear genomes are a particular hotspot for genetic code variation. The phylum Ciliophora is a large group of single-celled eukaryotes (protists) that diverged from other Alveolates more than one billion years ago [4]. Ciliates are highly unusual in that they exhibit nuclear dimorphism whereby each cell has two types of nuclei, the germline micronucleus (MIC) and the somatic macronucleus (MAC), each of which contains its own distinct genome structure and function [5]. The MIC genome functions as the germline genome and is exchanged during sexual reproduction. MIC genomes are typically diploid and are transcriptionally inactive during vegetative growth. The MIC genome undergoes rearrangement and excision of micronucleus-limited sequences to serve as a template to generate the transcriptionally active MAC genome [6]. MAC genomes typically contain short, fragmented, gene-dense chromosomes that are present at high ploidy levels (up to tens of thousands of copies) [7].

Known genetic code changes in ciliates involve reassignment of one or more stop codons to specify for amino acids. Most reported ciliate genetic code changes involve reassignment of both the UAA and UAG codons to specify glutamine as in *Tetrahymena*, *Paramecium*, and *Oxytricha* [8], or glutamic acid in *Campanella umbellaria* and *Carchesium polypinum* or tyrosine in *Mesodinium* species [9]. Other known modifications include reassignment of the UGA stop codon to specify tryptophan in *Blepharisma* [8], or cysteine in *Euplotes* [10]. The most extreme example of genetic code remodelling is found in *Condylostoma magnum* where all three UAA, UAG, and UGA stop codons have been reassigned and can specify either an amino acid (glutamine for UAA and UAG, and tryptophan for UGA) or signal translation termination depending on their proximity to the mRNA 3' end [9,11]. Not all ciliates use non-canonical genetic codes. For example, *Fabrea salina*, *Litonotus pictus*, and *Stentor coeruleus* use the canonical genetic code [9,12].

Changing the meaning of codons from stop to sense requires modifications to the translational apparatus. In eukaryotes, the eukaryotic release factor 1 (eRF1) protein recognises the three standard stop codons in mRNA and triggers translation termination. Studies have shown that mutations in the N-terminus of eRF1 can alter stop codon specificity [8,13–15]. eRF1 specificity to recognise only the UGA codon has evolved independently via different molecular mechanisms at least twice in ciliates with reassigned UAA and UAG codons [14]. Acquisition of tRNA genes with anticodons that recognise canonical stop codons (suppressor tRNAs), via mutations, base modifications or RNA editing enables translation of canonical stop codons into amino acids [16,17].

Tandem stop codons are additional stop codons located in the 3'-UTR within a few positions downstream of a gene in the same reading frame [18]. They are thought to act as "back-up" stop codons in the event of readthrough, minimising the extent of erroneous protein elongation. For example, in yeast there is a statistical excess of stop codons in the third in-frame codon position downstream of genes with a UAA stop codon [18]. Tandem stop codons have been shown to be overrepresented in ciliates that only use UGA as a stop codon, compared to eukaryotes that use the canonical genetic code [19]. The level of overrepresentation is greater in highly expressed genes [20]. Tandem stop codons are thought to be particularly important in ciliates where, following stop codon reassignment, readthrough events might occur at a higher frequency due to mutations in eRF1 [20].

Several models have been proposed to describe genetic code changes. Under the "codon capture" model, a codon that is rarely used (e.g., due to GC content) is gradually eliminated from the genome followed by loss of the corresponding unused tRNA [21]. Due to random genetic drift the codon could reappear and be captured by a noncognate tRNA charged with a different amino acid, thus changing the genetic code. Such a process would be essentially neutral, not resulting in mistranslated protein products as the codon is eliminated from genes before the change in meaning occurs [17]. Alternatively, under the "ambiguous intermediate" model [22], reassignment of a codon takes place via an intermediate stage, where a codon is ambiguously translated via competing tRNAs charged with different amino acids, or in the context of stop codon reassignment, a suppressor tRNA competing with a release factor. This process would be driven by selection and result in the elimination of the cognate tRNA if the new meaning is advantageous. The "genome streamlining" model is more relevant to small genomes (e.g., organellar genomes or parasites) where there is pressure to minimise translational machinery [23]. More recently the "tRNA loss driven codon reassignment" mechanism was proposed to describe codon reassignments whereby tRNA loss, or alteration of release factor specificity, results in an unassigned codon that can be captured by another tRNA gene [24,25].

In virtually all genetic code changes reported to date, the codons UAA and UAG have the same meaning, i.e., they are either both used as canonical stop codons or are both reassigned to the same amino acid [25]. This suggests that evolutionary or mechanistic constraints couple the meaning of these two codons [26]. One such constraint is wobble binding of a suppressor tRNA gene with a UUA anticodon, where uracil in the first anticodon position can bind to either adenine or guanine in the third codon position of mRNA [27]. Thus, acquiring a suppressor tRNA gene with a UUA anticodon could potentially change the meaning of both the UAA and UAG codons. Wobble binding has been experimentally demonstrated in *Tetrahymena thermophila*, where tRNA-Sup(UUA) was shown to suppress both the UAA and UAG codons, whereas tRNA-Sup(CUA) suppressed only the UAG codon [16]. The first report of nuclear genetic code variants where UAA and UAG have different meanings were reported in transcriptomics analyses where a Rhizarian species (Rhizaria sp. exLh) was shown to use UAG to encode leucine and in a Fornicate (*Iotanema spirale*) where UAG has been reassigned to glutamine [26]. However, in both cases, the UAA codon was retained as a stop codon, thus avoiding the problem of genetic code ambiguity due to wobble binding.

Here, we report the discovery of a novel variant of the genetic code in a ciliate belonging to the Oligohymenophorea class, where the meaning of the UAA and UAG codons have changed to specify different amino acids. Using G&T-Seq [28], we performed parallel genome and transcriptome sequencing of small pools of ciliate cells. Combining genome and transcriptome sequencing data from multiple independently amplified samples enabled co-assembly of a highly complete macronuclear genome assembly and annotation. Genomic and transcriptomic analysis revealed that the UAA codon has been reassigned to specify lysine, while the

meaning of the UAG codon has changed to specify glutamic acid. We identified multiple suppressor tRNA genes of both types in the genome, supporting the genetic code changes. We show that UGA codons are significantly enriched in the 3'-UTR of genes suggesting that there is selective pressure to maintain tandem stop codons, which may play a role in minimising erroneous protein elongation in the event of translational readthrough. To our knowledge, this is the first report of a genetic code variant where UAA and UAG specify different amino acids.

Results & discussion

Genome assembly of an oligohymenophorean ciliate

We isolated a novel ciliate species Oligohymenophorea sp. PL0344 from a freshwater pond at Oxford University Parks, Oxford, UK. Attempts to establish a stable long-term culture were unsuccessful so we applied low input single-cell based approaches to generate genomic and transcriptomic data. Small pools of cells (5–50 cells) were sorted into a microplate using fluorescence-activated cell sorting (FACS). Parallel genome and transcriptome sequencing was performed using G&T-Seq, which relies on whole genome amplification using multiple displacement amplification (MDA) and transcriptome analysis using a modified Smart-seq2 protocol [28].

A *de novo* genome assembly was generated by co-assembling reads from 10 samples (totalling approximately 6 Gb). Following manual curation and removal of contaminant sequences, the resulting macronuclear genome assembly was 69.7 Mb in length, contained in 3671 scaffolds with an N50 of 59.6 Kb (**Table 1**). Approximately 89% of the corresponding RNA-Seq reads mapped to the genome assembly, indicating high completeness. GC content of the genome is low at 30.6% (**Table 1**), which is similar to previously sequenced ciliate genomes [12]. The mitochondrial genome was also recovered which is a linear molecule 35,635 bp in length with GC content of 25.33% and capped with repeats. The mitochondrial genome contains the small subunit (SSU) and large subunit (LSU) ribosomal RNA (rRNA) genes, 5 tRNA genes, 19 known protein-coding genes, and 13 open reading frames.

The nuclear encoded SSU rRNA gene sequence is 99.81% identical to an environmental sequence (AY821923) of an unnamed ciliate in the GenBank database, isolated from Orsay, France [29]. Maximum-likelihood phylogenetic analysis of the SSU rRNA gene placed it within a clade containing four unnamed ciliate species (AY821923, HQ219368, LR025746, HQ219418) and *Cinetochilum margaritaceum* (MW405094) with 100% bootstrap support (**S1 Fig**). Thus, based on the SSU rRNA gene, *C. margaritaceum* is the closest related named species. The SSU rRNA gene of *C. margaritaceum* is 96.03% identical to that of Oligohymenophorea sp. PL0344. *C. margaritaceum* belongs to the Loxocephalida order (Class Oligohymenophorea; Subclass Scuticociliatia), which is considered a controversial order due to its non-monophyly [30,31]. Our phylogenetic analysis places *C. margaritaceum* as a divergent branch relative to other members of Loxocephalida (**S1 Fig**), which is congruent with previous analyses [30,31], suggesting taxonomic revision is required.

Oligohymenophorea sp. PL0344 uses a novel genetic code

Preliminary analysis of the genome sequence revealed that many coding regions contained in-frame UAA and UAG codons. Consistent with codon reassignments in other ciliate species, this suggested that the UAA and UAG stop codons have been reassigned to code for amino acids. Surprisingly however, the meanings of these codons do not match any known genetic code. An example gene (tubulin gamma chain protein), showing six in-frame UAA codons and six in-frame UAG codons, translated and aligned to orthologous protein sequences with representatives from across Eukaryota is displayed in **Fig 1**. Five in-frame UAA codons

Table 1. Genome Assembly and Annotation Statistics.

Genome Assembly		
Total length	69,715,444 bp	
Contigs	3671	
N50	59,570 bp	
GC content	30.6%	
RNA-Seq mapping rate	89%	
BUSCO completeness[1]	Complete: 84.8% Complete and single copy: 80.1% Complete and duplicated: 4.7% Fragmented: 3.5% Missing: 11.7%	
Genome Annotation		
Number of genes	20,141	
Number of transcripts	22,084	
Number of monoexonic genes	7,080	
Exons per transcript	3.76	
GC content (CDS)	34.12%	
% of genome covered by CDS	63.8%	
BUSCO completeness[1]	Complete: 94.74% Complete and single copy: 87.72% Complete and duplicated: 7.02% Fragmented: 1.75% Missing: 3.51%	
	Median	**Mean**
CDS size (bp)	1,506	2,014.12
Intron size (bp)	57	80.39
5'UTR size (bp)	65	96.74
3'UTR size (bp)	95	137.61
Intergenic distances	156	574.43

[1]BUSCO completeness assessed using V4 with the Alveolata_obd10 dataset in genome mode for the genome assembly and in protein mode for the genome annotation.

https://doi.org/10.1371/journal.pgen.1010913.t001

correspond to highly conserved columns in the alignment where lysine is the consensus amino acid (Fig 1). Four in-frame UAG codons correspond to highly conserved columns in the alignment where glutamic acid is the consensus amino acid, and another corresponds to a column with a mix of glutamic acid and aspartic acid (Fig 1).

We used two complementary tools to analyse the genetic code further. First, we used the "genetic_code_examiner" utility from the PhyloFisher package [32], which predicts the genetic code by comparing codon positions in query sequences to highly conserved (> 70% conservation) positions in amino acid alignments from a database of 240 orthologous protein sequences. PhyloFisher identified 58 genes with 87 in-frame UAA codons that correspond to highly conserved amino acid sites. Of these, 74 UAA codons (85%) correspond to highly conserved lysine residues (Fig 2A). The second most numerous match was to arginine, another positively charged amino acid, with 9 (10%) hits. PhyloFisher identified 46 genes with 63 in-frame UAG codons that correspond to highly conserved amino acid sites. Of these, 56 UAG codons (89%) correspond to highly conserved glutamic acid residues (Fig 2B). The second most numerous match was to aspartic acid, another negatively charged amino acid, with 4 (6%) of hits. Amongst the genes identified by PhyloFisher, 27 contained both in-frame UAA codons and an in-frame UAG codons.

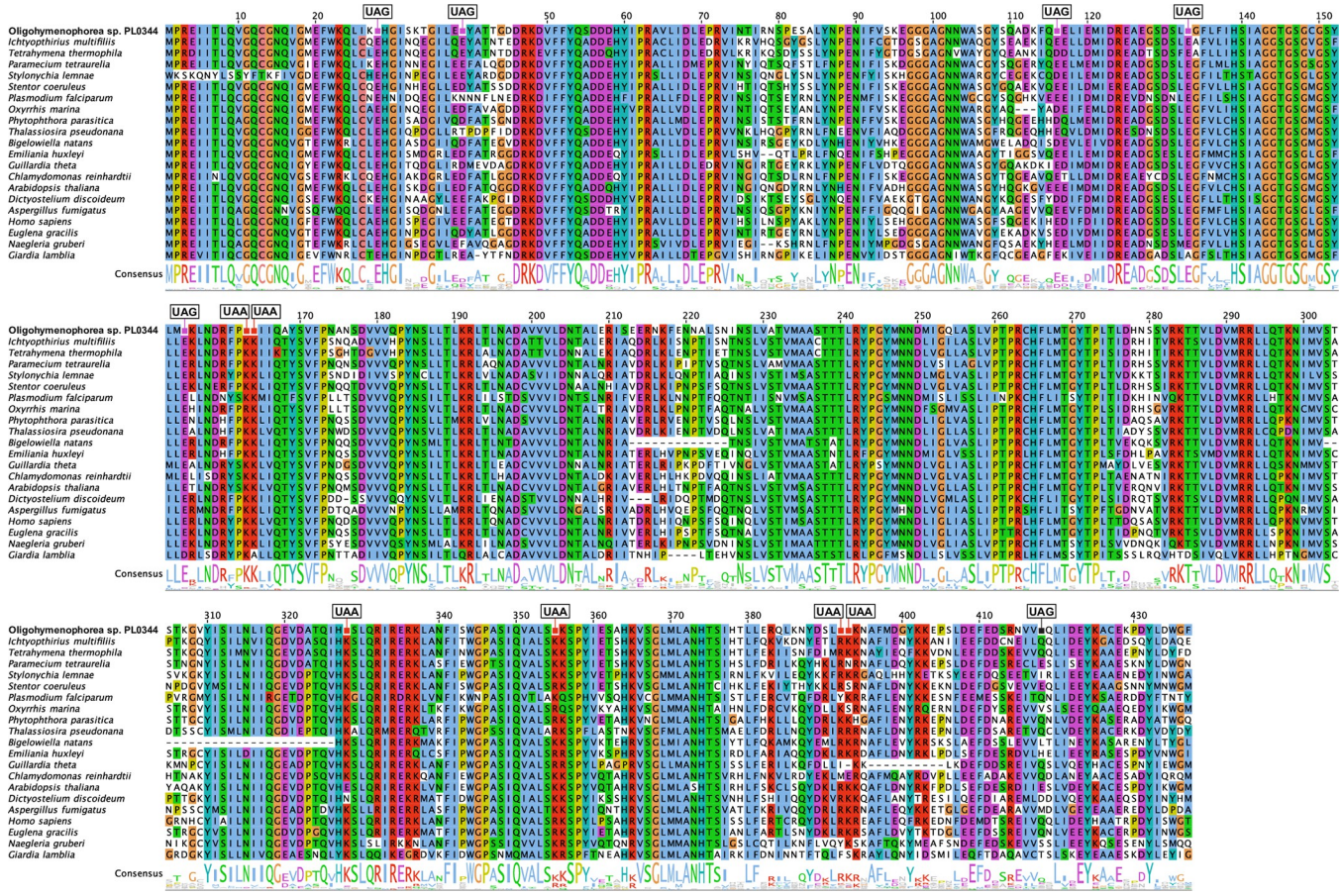
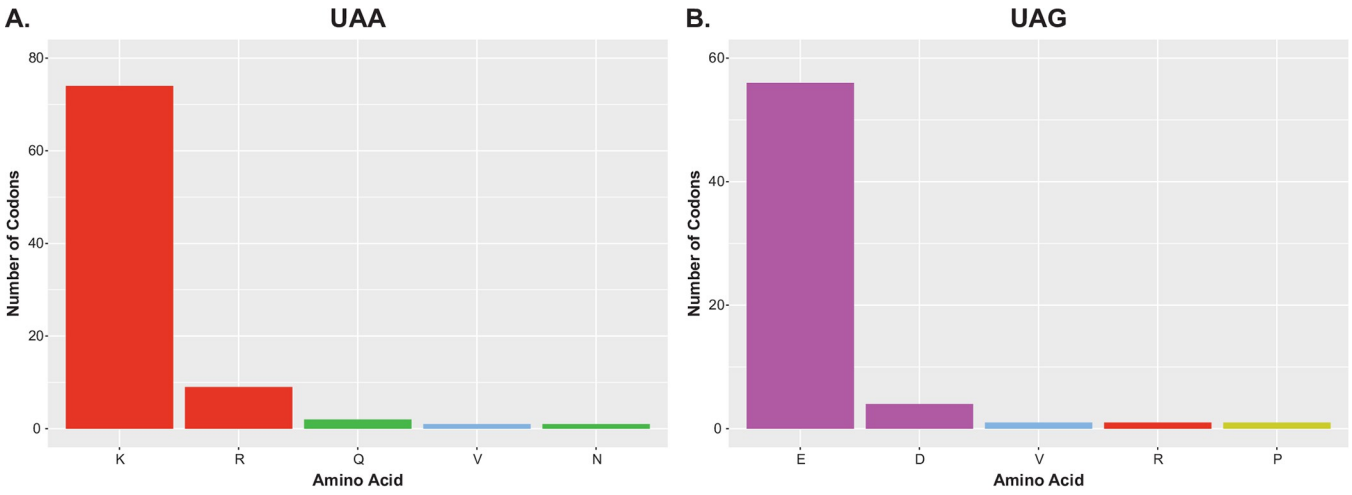


Fig 1. Genetic code change in Oligohymenophorea sp. PL0344. Example multiple sequence alignment of a tubulin gamma chain protein and orthologous sequences spanning Eukaryota. The alignment has been trimmed for visualisation purposes to remove poorly conserved regions and highlight internal UAA and UAG codons.

https://doi.org/10.1371/journal.pgen.1010913.g001

We also analysed the genetic code using Codetta [33,34]. Codetta predicts the genetic code by aligning profile hidden Markov models (HMMs) from the Pfam database against a six-frame translation of the query genome assembly. The meaning of each codon is inferred based on emission probabilities of the aligned HMM columns. From the whole genome sequence, 14,633 UAA codons and 10,160 UAG codons had a Pfam position aligned. Based on these alignments, Codetta also predicted that the UAA codon is translated as lysine and UAG translated as glutamic acid, each with a log decoding probability of zero (Fig 2C).

Thus, these results indicate that Oligohymenophorea sp. PL0344 uses a novel genetic code where UAA is translated as lysine and UAG is translated as glutamic acid. This is the first time this genetic code variant has been reported. Furthermore, according to our knowledge, this is the first report of a genetic code variant where UAA and UAG have been reassigned to specify different amino acids. Genetic code variants were previously reported where UAG was reassigned to specify an amino acid (either leucine or glutamine) but UAA was retained as a stop codon in both cases [32]. This is significant as it suggests that the genetic code variant reported herein has overcome mechanistic constraints linking the translation of these two codons.



C. Codetta Log Decoding Probabilities

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	?
UAA	-7247.7	-12835.5	-8178.5	-5787.8	-14432.1	-10288.2	-6864.4	-12385.8	0.0	-10974.3	-10337.2	-6106.2	-11220.1	-4178.8	-3215.4	-6325.0	-6863.7	-10629.3	-15433.8	-11716.0	-3255.1
UAG	-5129.3	-9789.2	-3214.5	0.0	-10447.8	-7064.8	-5266.2	-9348.6	-4089.2	-8450.4	-8002.1	-4528.5	-7525.4	-3136.4	-4933.1	-4482.9	-5141.5	-8054.1	-11143.0	-8505.1	-2540.6

Fig 2. Genetic code prediction for Oligohymenophorea sp. PL0344. PhyloFisher genetic code prediction for the **(A)** UAA and **(B)** UAG codons using the PhyloFisher database of 240 orthologs. Only well conserved (>70%) amino acids are considered. Colours correspond to amino acid properties and match the multiple sequence alignment in Fig 1. **(C)** Codetta genetic code prediction. Log decoding probabilities for the UAA and UAG codons are shown for each of the 20 standard amino acids.

https://doi.org/10.1371/journal.pgen.1010913.g002

Suppressor tRNA genes

tRNA genes were annotated using tRNAscan [35], resulting in the annotation of 320 tRNA genes, including 15 that are predicted to be pseudogenes. Amongst the annotated tRNA genes are 23 putative suppressor tRNA genes. These are tRNA genes with anticodons complementary to canonical stop codons (UAA, UAG, or UGA). The annotated suppressor tRNA genes include 12 tRNA-Sup(UUA) genes and 10 tRNA-Sup(CUA) genes. tRNAscan also predicted a single tRNA-Sup(UCA) gene, however this was low scoring and was not predicted by ARAGORN [36], an alternative tool to identify tRNA genes. tRNAscan also predicts the function of tRNAs. Many of the tRNAscan isotype predictions were consistent with the predicted genetic code (i.e., UAA = lysine and UAG = glutamic acid), however several putative tRNA genes had low-scoring or inconsistent isotype predictions. To better characterise the suppressor tRNA genes, we compared their sequences to the non-suppressor tRNA genes. Eight of the twelve predicted tRNA-Sup (UUA) genes were most similar to tRNA-Lys genes with UUU or CUU anticodons (68.49% to 80.95% identical) (**S1 Table**), consistent with the genetic code prediction that UAA has been reassigned to specify lysine. An example tRNA-Sup(UUA) predicted to function as a lysine tRNA is shown in **Fig 3A**. All ten tRNA-Sup(CUA) genes were most similar to tRNA-Glu genes with CUC or UUC anticodons (69.44% to 93.06% identical) (**S1 Table**), consistent with the genetic code prediction that UAG has been reassigned to specify glutamic acid. An example tRNA-Sup(CUA) predicted to function as a glutamic acid tRNA is shown in **Fig 3B**. Similarly, analysis using phylogenetic networks clusters most of the suppressor tRNA genes with tRNA genes of their predicted function (**S2 Fig**). We also identified a tRNA gene for selenocysteine, tRNA-SeC(UCA) (**Fig 3C**), suggesting that the UGA codon is used both as a stop codon and to encode selenocysteine. Thus, all 64 codons can specify amino acids as has been reported in other ciliate genomes [37].

UAA and UAG codons differ only in the wobble position. According to wobble-binding rules, uracil in the first tRNA anticodon position ("wobble position") (**Fig 3A**) can bind to

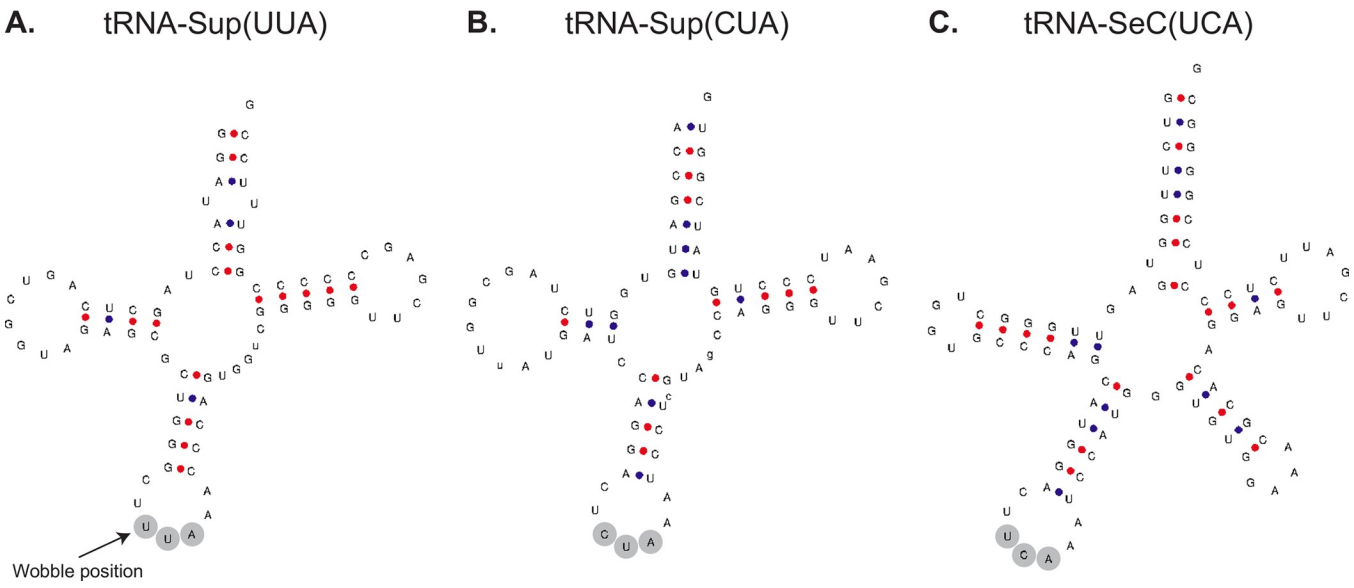


Fig 3. Example tRNA Genes. (A) Predicted secondary structure of an example tRNA-Sup(UUA) predicted to function as a lysine tRNA. The wobble position is highlighted. According to wobble-binding rules, uracil at this position can bind to either adenine or guanine in the third codon position of mRNA, allowing the suppressor tRNA to recognise both UAA and UAG stop codons. **(B)** Predicted secondary structure of an example tRNA-Sup(CUA) predicted to function as a glutamic acid tRNA. **(C)** Predicted secondary structure of the tRNA-SeC(UCA) for selenocysteine.

https://doi.org/10.1371/journal.pgen.1010913.g003

either adenine or guanine in the third codon position of mRNA [27], allowing tRNA with a UUA anticodon to recognise both UAA and UAG codons. It has been experimentally demonstrated that *T. thermophila* tRNA-Sup(UUA) can recognise both UAA and UAG codons [16]. It has been suggested that wobble binding is a possible explanation as to why UAA and UAG virtually always have the same meaning [9]. Considering that Oligohymenophorea sp. PL0344 has tRNA-Sup(UUA) genes for lysine and tRNA-Sup(CUA) genes for glutamic acid, this raises the question: are UAG codons ambiguously translated as both glutamic acid and lysine? If not, how has it overcome the mechanistic and evolutionary constraints that appear to couple the translation of these two codons? Presumably, if wobble binding allows tRNA-Sup(UUA) to recognise the UAG codon, it would be less efficient than tRNA-Sup(CUA) and outcompeted, possibly resulting in some degree of stochastically translated protein products with glutamic acid residues substituted by lysine at UAG codon positions. Attempts to establish a stable culture were unsuccessful, and while we can overcome this problem to generate a genome assembly using low-input sequencing methods designed for single-cell analysis, such low-input approaches are not available for proteomics. Without proteomics data, it is not possible to determine if UAG is ambiguously translated. Additionally, while the genomic and transcriptomic data provide strong evidence that lysine and glutamic acid are the major translation products of UAA and UAG codons, respectively, we cannot rule out the possibility that other amino acids are (mis)incorporated at these sites which could be detected using mass-spectrometry [38,39]. Furthermore, from suppressor tRNA gene sequences alone, it is not possible to determine if they incorporate modified nucleotides which could alter codon-anticodon binding specificity.

Genome annotation and codon usage analysis

Genome annotation incorporating RNA-Seq data and protein alignments from other ciliates resulted in the annotation of 22,048 transcripts from 20,141 gene models (**Table 1**). BUSCO

analysis estimates that the genome annotation is highly complete with 94.7% of BUSCO genes recovered as complete, which compares favourably to other high quality ciliate genomes (**S2 Table**). The median intron size of 57 bp (**Table 1**) is similar to previously sequenced ciliate genomes, such as *Tetrahymena thermophila* and *Oxytricha trifallax* [7,37] but not as short as the extremely short introns (15–25 bp) found in *Stentor coeruleus* or *Paramecium tetraurelia* [12,40]. We defined a subset of genes as "highly expressed" based on the 10% of genes with the highest transcripts per million (TPM) values for comparison below. Codon usage is biased towards using codons with lower GC content. This bias is reduced in highly expressed genes which have higher GC content compared to all genes (38.51% versus 34.12%), similar to previous reports in *Paramecium* and *Tetrahymena* [37,41]. The reassigned codons are widely used across genes with 95.9% of genes containing both a UAA codon and a UAG codon. However, their usage is reduced in highly expressed genes (**S3 Table**). Reduced codon usage in highly expressed genes could indicate translational inefficiency, or that selective pressure to retain canonical lysine and glutamic acid codons is higher in highly expressed genes.

Very little is known about translation termination efficiency in ciliates. This is particularly interesting for ciliates that use only UGA as a stop codon, as UGA is known to be the least robust stop codon and the most prone to translational readthrough [42]. The sequence composition surrounding a stop codon influences the rate of stop codon readthrough. The nucleotide immediately downstream of a stop codon (+4 position) is particularly important, with several studies demonstrating that presence of a cytosine following UGA substantially increases the rate of readthrough [43–45]. Interestingly, examining the sequence composition surrounding stop codons in Oligohymenophorea sp. PL0344, cytosine appears to be avoided following the stop codon (S**3 Fig**). This is particularly noticeable in highly expressed genes (S**3 Fig**) where the proportion of genes with a cytosine following UGA is significantly reduced (chi-squared test, p-value = 7.3e-10). This trend has also been observed in *P. tetraurelia* and *T. thermophila* [41]. Tandem stop codons potentially play an important role as "back-up" stop codons, minimising the extent of protein elongation in the event of readthrough [18]. Here, we analysed tandem stop codons by counting UGA codons in the first 20 in-frame codon positions downstream of genes. Our results show that UGA codons are significantly overrepresented (chi-squared test, p-value < 0.05) in the first four in-frame codons downstream of genes (**Fig 4**). 12.3% of genes have at least one UGA codon within the first six in-frame codon positions downstream of genes, similar to the proportion reported for *T. thermophila* (11.5%) where UAA and UAG have also been reassigned to encode amino acids [19]. For comparison, the reassigned UAA and UAG codons are not overrepresented in this region. The frequency of UGA codons at these positions is greater for highly expressed genes whereby 13.6% of highly expressed genes have at least one UGA codon within the first six in-frame codon positions downstream of genes (**Fig 4**). These data add support that there is selective pressure for ciliates with reassigned UAA and UAG codons to maintain tandem UGA stop codons at the beginning of the 3'-UTR. It is tempting to speculate that these additional UGA stop codons play a role in minimising deleterious consequences of readthrough events.

Phylogenomics analysis of genetic code changes in the Ciliophora

We carried out phylogenomics analyses to map genetic code changes onto the ciliate phylogeny. A phylogenomic dataset consisting of genomic and transcriptomic data from 46 ciliate species and 9 outgroup species was constructed (**S2 Table**). Phylogenomic reconstruction was performed on a concatenated alignment of 89 single-copy BUSCO proteins (40,289 amino acid sites) using maximum-likelihood (IQ-TREE under LG+F+I+R7 model) and Bayesian (PhyloBayes-MPI under CAT-GTR model) approaches. We also conducted a partitioned

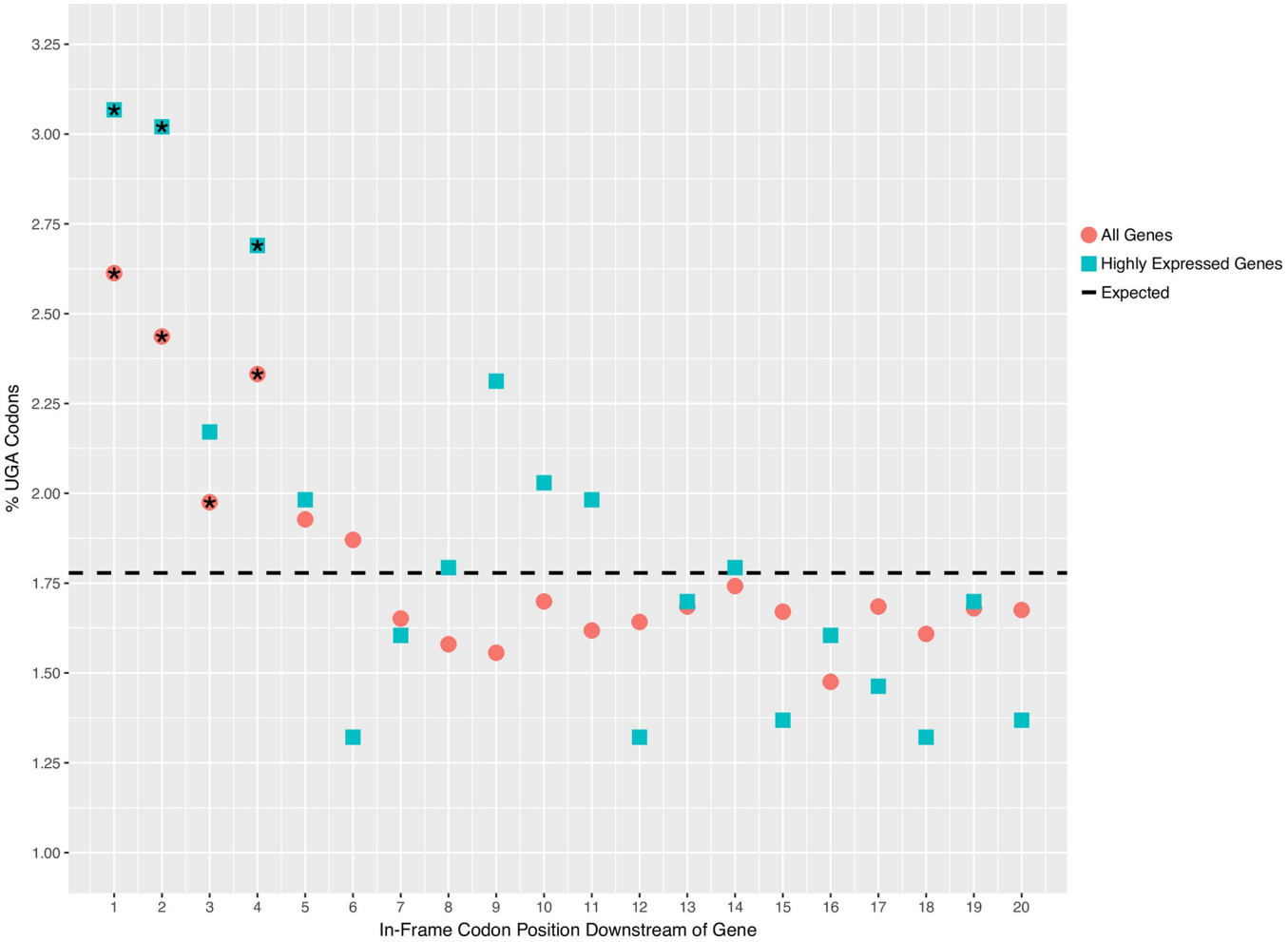


Fig 4. Enrichment of tandem stop codons. The proportion of codon positions occupied by UGA in the 20 in-frame codon positions immediately downstream of all genes and highly expressed genes. Positions where UGA is significantly overrepresented (chi-squared test, p-value < 0.05) are indicated with an asterisk.

https://doi.org/10.1371/journal.pgen.1010913.g004

analysis on the same dataset using IQ-TREE, with a partitioning scheme which merged the 89 proteins into 14 partitions. The three resulting phylogenies were largely in agreement with each other and with previously published analyses, with full or high support from all three methods at most branches (Fig 5). Oligohymenophorea sp. PL0344 was robustly placed within the Oligohymenophorea class in a clade containing Hymenostomatida and *Pseudocohnilembus persalinus* with full support from all three methods (Fig 5).

The position of *Paramecium* (order Peniculida) is unstable in our phylogenomic analyses. Both the LG+F+I+R7 and partitioned phylogeny group *Paramecium* as sister to the Peritrichia clade with 89% and 82% bootstrap support respectively (Fig 5). This is congruent with some previous phylogenomic analyses which recover Peniculida as sister to Peritrichia species [46–48]. However, the Bayesian phylogeny places *Paramecium* as sister to Hymenostomatida and Philasterida (S4 Fig). This grouping has been recovered in some previous phylogenomic analyses [49,50]. The correct placement of Peniculida is unclear based on the current datasets available. The *Paramecium* species included in our analysis have a high proportion of missing data (Fig 5). We anticipate that differences in topology may be influenced by varying levels of sensitivity to missing data in the models used. *Mesodinium rubrum* is another problematic taxon

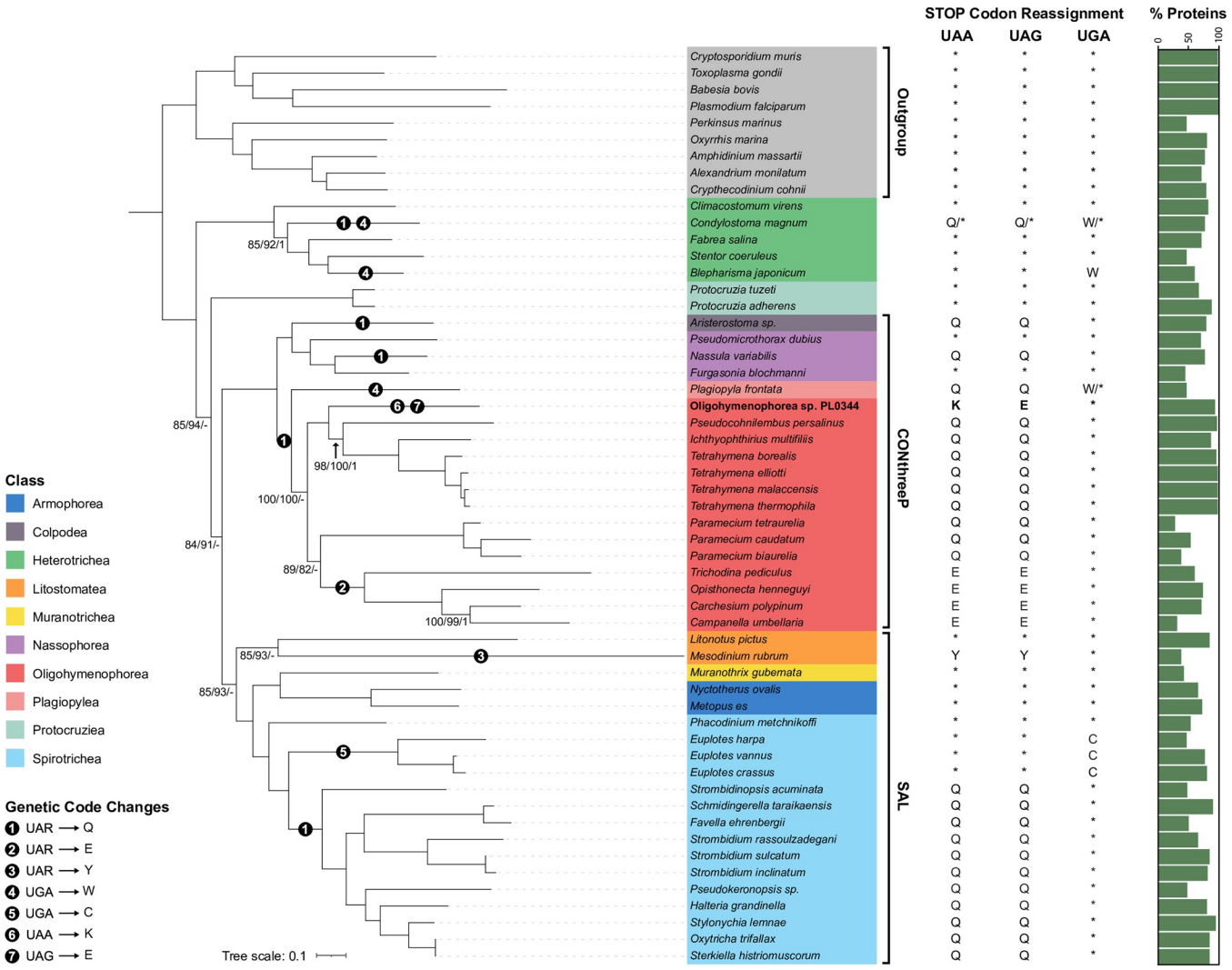


Fig 5. Phylogenomic analysis of genetic code changes in the Ciliophora. Maximum-likelihood phylogeny of 46 ciliate species and 9 outgroup species from the Alveolata, based on a concatenated alignment of 89 BUSCO proteins (40,289 amino acid sites) under the LG+F+I+R7 model using IQ-TREE. The values at branches represent statistical support from 100 non-parametric bootstraps with the LG+F+I+R7 model, 100 non-parametric bootstraps from the IQ-TREE partitioned analysis, and Bayesian posterior probabilities determined under the CAT-GTR model in PhyloBayes-MPI. Branches have full support from all three approaches (i.e., 100/100/1) except where indicated. Hyphens indicate branches that weren't recovered under a particular analysis. Stop codon reassignments are shown (*, STOP; Q, glutamine; W, tryptophan; K, lysine; E, glutamic acid; Y, tyrosine; C, cysteine). Numbers inside solid black circles along branches indicate when a genetic code change event was inferred to have occurred (UAR = UAA and UAG). The percentage of proteins included in the concatenated alignment is shown in the bar plot, highlighting the amount of missing data per species.

https://doi.org/10.1371/journal.pgen.1010913.g005

which is thought to be prone to long branch attraction (LBA) artefacts. Furthermore, existing *Mesodinium* transcriptomes are contaminated with sequences from their prey [51]. Some previous phylogenomic and phylogenetic analyses place it as an early branching ciliate [52,53], however these may have been influenced by contamination [51]. Here, we account for contamination by removing any sequences from the *M. rubrum* transcriptome with best BLAST hits outside of the Ciliophora (n = 3,574). Both the LG+F+I+R7 and partitioned phylogeny group *M. rubrum* with *Litonotus pictus*, another member of the Litostomatea class, with 85% and 93% bootstrap support respectively (**Fig 5**), while our Bayesian analysis places it as a deep branching ciliate branching before *Protocruzia* (S4 Fig). The grouping of *M. rubrum* with *L.*

pictus agrees with a recent phylogenomics analysis of *Mesodinium* species that accounts for LBA and contamination [51].

Where genome or transcriptome assemblies were available, or raw sequencing reads deposited in public databases, we validated the known genetic codes using Codetta and PhyloFisher. All species had the expected genetic code except for *Plagiopyla frontata*. Codetta and PhyloFisher both predicted that UAA and UAG are translated as glutamine in *P. frontata*, which is not surprising given how many ciliate species use this genetic code (**Fig 5**). Interestingly however, both methods predict that UGA has also been reassigned to specify tryptophan in *P. frontata*. From the PhyloFisher dataset of 240 query proteins, 3 (1.25%) contain internal UGA codons that correspond to highly conserved tryptophan residues in other species **(S5 Fig).** This suggests that *P. frontata* may use UGA both as a stop codon and also rarely as a sense codon to specify tryptophan, similar to the *Condylostoma* genetic code (**Fig 5**) [9,11].

We mapped genetic code reassignments onto the ciliate phylogeny, highlighting the remarkable number of independent genetic code changes within the ciliates (**Fig 5**). Based on our phylogeny, and assuming a non-canonical genetic code doesn't revert to the canonical genetic code, the translation of UAR (UAA and UAG) codons to glutamine is the most common genetic code variant and has independently evolved at least five times. From our analysis, translation of UGA to tryptophan has independently evolved at least three times in ciliate nuclear genomes. However, it has recently been reported that Karyorelict ciliates (not included in this analysis) use a context-dependent genetic code similar to *Condylostoma*, where UAR has been reassigned to glutamine and UGA specifies either tryptophan or stop depending on context, indicating a fourth independent origin of UGA being translated as tryptophan and a sixth independent origin of UAR being translated to glutamine in ciliates [54]. The translation of UGA to cysteine in *Euplotes*, UAR to tyrosine in *Mesodinium* and UAR to glutamic acid in Peritrichia have all evolved once. The Oligohymenophorea sp. PL0344 genetic code appears to be a relatively recent phenomenon and is unique in that the two codons have different meanings. The Oligohymenophorea class contains at least three different genetic code variants, and no sampled species which have retained usage of UAA or UAG as a stop codon. Our phylogeny suggests that the stop codons UAA and UAG were reassigned to glutamine in the ancestor of Oligohymenophorea (**Fig 5**). These codons were then reassigned to glutamic acid in the Peritrichs, or to lysine (UAA) and glutamic acid (UAG) in Oligohymenophorea sp. PL0344.

It remains unclear why Ciliate genomes are such a hotspot for stop codon reassignments. Our study shows that even within the Oligohymenophorea class, which is relatively well sampled compared to other ciliate clades, there remain novel genetic code reassignments to be discovered. Further sequencing of under-sampled ciliate lineages and other microbial eukaryotes may reveal more variant genetic code changes and help to better understand the evolution and mechanisms of genetic code changes.

Materials and methods

Sampling, Ciliate isolation, culturing, and cell-sorting

Surface water was collected from a margin of an artificial freshwater pond at Oxford University Parks (51˚45'51.0"N 1˚15'04.5"W), Oxford (UK) in April 2021 by directly submerging a 1 litre autoclaved glass collection bottle. 200 mL of the water sample were concentrated using a 5 μm filter into a final volume of 20 mL. Oligohymenophorea sp. PL0344 was identified using an inverted microscope (Olympus CKX41) and single cells were isolated manually using a glass micropipette by transferring them into successive drops of 0.2 μm pre-filtered and autoclaved environmental source water. When cells were free of any other eukaryote, they were transferred into a 96 well-plate containing filtered and autoclaved environmental source water. In

order to obtain a clonal culture, isolated cells were incubated during a week at 20˚C with a 12h:12h light:dark photo-cycle with a photon flux of 32 μmoles·m^{-2}·s^{-1}. When ciliate cells divided and a dense culture was observed in the well, the mini-culture was scaled-up during a month by successively transferring the cells into larger volumes until a non-axenic but mono-eukaryotic ciliate culture of 20 mL of volume was established. We attempted to establish a long-term culture of Oligohymenophorea sp. PL0344 but were unsuccessful. The culture grew but only for a relatively short period. We also attempted to re-isolate the same ciliate species from the same and surrounding water bodies but failed. We relied on a single-cell/low-input based approach to generate genomic and transcriptomic data. Pools of ciliate cells (5–50 cells) were sorted into a 384-well plate containing 5 μL of autoclaved source water using FACS (BD FACSMelody Cell Sorter, BD Biosciences). 10 μL of RLT+ lysis buffer (Qiagen) was then added to each well and the plate was sealed and centrifuged (2000 x g, 4˚C, 1 min) to remove bubbles and to ensure that the lysis buffer was at the bottom of each well. The sorted plate was stored at -80˚C until processed.

G&T-Seq, library preparation, and sequencing

Using a magnetic separator, Dynabeads MyOne Streptavidin C1 (Invitrogen) beads were washed according to the manufacturer's guidance and then incubated with 2 × Binding &Wash buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA, 2 M NaCl) and Biotinylated Oligo-dT primer (IDT, 5'-/BiotinTEG/AAG CAG TGG TAT CAA CGC AGA GTA CTT TTT TTT TTT TTT TTT TTT TTT TTT TVN-3') at 100 μM for 30 minutes at room temperature on a rotator. The oligo-treated beads were washed four times in 1 × Binding &Wash buffer (5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl) and then suspended in 1 × SuperScript II First Strand Buffer (Invitrogen) supplemented with SUPERase•In RNase Inhibitor (Invitrogen) to a final concentration of 1 U/μl. The lysate was thawed on ice. 10 μl of prepared oligo-dT beads was added to each well containing 12 μl cell lysate. The lysate plate was sealed and incubated on a ThermoMixer C (Eppendorf) at 21˚C for 20 minutes shaking at 1000 rpm. Using a Fluent 480 liquid handling robot (Tecan) and a Magnum FLX magnetic separator (Alpaqua), the lysate supernatant was transferred to a new plate and the beads were washed twice in a custom wash buffer (50 mM Tris-HCl pH 8.3, 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT, 0.5% Tween-20). The supernatant from the washes was added to the left-over cell lysate —containing the genomic DNA—which was stored at -20˚C overnight. The washed beads were suspended in a reverse transcription mastermix of 1 mM dNTPs, 0.01 M $MgCl_2$,1 × SuperScript II First Strand Buffer, 1 M Betaine, 5.4 M DTT, 1μM Template-Switching Oligo (5′-AAGCAGTGGTATCAACGCAGAGTACrGrG+G-3′, where "r" prefixes a ribonucleic acid base and "+" prefixes a locked nucleic acid base, Qiagen) then incubated using a ThermoMixer C with the following conditions: 42˚C for 2 minutes at 200 rpm, 42˚C for 60 minutes at 1500 rpm, 50˚C for 30 minutes at 1500 rpm, 60˚C for 10 minutes at 1500 rpm. The cDNA was amplified using HiFi Hotstart Ready Mix (KAPA) and IS Primers to a final concentration of 0.1 μM (IDT, 5'-AAG CAG TGG TAT CAA AGA GT-3') with the following thermocycling conditions: 98˚C for 3 minutes, then 21 cycles of 98˚C for 15 seconds, 67˚C for 20 seconds, 72˚C for 6 minutes and finally 72˚C for 5 minutes. The cDNA was then purified using 0.8 × vols Ampure XP (Beckman Coulter) and 80% ethanol on the Fluent 480 liquid handling robot and eluted in 10 mM Tris-HCl. The remaining cell lysate was thawed and subjected to a 0.6 × vols Ampure XP clean-up with 80% ethanol. The bead-bound gDNA was isothermally amplified for 3 hours at 30˚C then 10 minutes at 65˚C using a miniaturised (1/5 vols) Repli-g Single-Cell assay (Qiagen). The amplified gDNA was cleaned up with 0.8 × vols Ampure XP and 80% ethanol, then eluted in 10 mM Tris-HCl. The cDNA and gDNA were

quantified by fluorescence (Quant iT HS-DNA, Invitrogen) on an Infinite Pro 200 plate reader (Tecan) then normalised to a final concentration of 0.2 ng/μl in 10 mM Tris-HCl. Dual-indexed sequencing libraries (Nextera XT, Illumina) were prepared using Mosquito and Dragonfly liquid handling instruments (SPT Labtech). The libraries were pooled and cleaned up using 0.8 × vols Ampure XP and 80% ethanol. The libraries were eluted in 10mM Tris-HCl and assessed using a Bioanalyzer HS DNA assay (Agilent), HS DNA Qubit assay (Invitrogen) and finally an Illumina Library Quantification Kit assay (KAPA). Sequencing was conducted on a NovaSeq 6000 with a 300 cycle Reagent kit v1.5 (Illumina) to produce 150 bp paired-end, dual-indexed reads.

Genome assembly

Adapter and quality trimming were carried out using BBDuk (https://jgi.doe.gov/data-and-tools/bbtools). Reads which mapped to a database of common lab contaminants (human and mouse) were removed using BBMap. A co-assembly of genomic DNA reads from 10 samples was generated using SPAdes (v3.15.3) [55] with default settings except -k 21, 33, 55, 77 and single-cell mode (—sc) was enabled. The assembly was manually curated and contaminant contigs were removed using a combination of metagenomic binning with MetaBAT2 [56] based on tetra-nucleotide frequencies and taxonomic classification with CAT (v5.2) [57] and Tiara (v1.0.1) [58]. Independent genome assemblies were also generated for all 10 individual samples and average nucleotide identities were compared as a quality control step to confirm that they corresponded to the same species. Assembly statistics were calculated using Quast [59]. Genome completeness was assessed using BUSCO (v4.1.2) [60] with the Alveolata_obd10 dataset run in genome mode.

Genome annotation

The genetic code was predicted using Codetta (v2.0) [33,34] and also using the "Genetic Code Examiner" utility from the PhyloFisher package, with the included database of 240 orthologs [32].

Gene models were annotated via the Robust and Extendable eukaryotic Annotation Toolkit (REAT, https://github.com/EI-CoreBioinformatics/reat) and Minos (https://github.com/EI-CoreBioinformatics/minos) using a workflow incorporating repeat identification, RNA-Seq mapping / assembly, alignment of protein sequences from related species and evidence guided gene prediction with AUGUSTUS [61].

A de novo repeat annotation was created using the RepeatModeller [62] v1.0.11 -RepeatMasker v4.07 [63] pipeline with defaults settings and the—gff output option enabled. To ensure high copy number 'bonafide' genes were excluded from repeat masking, the RepeatModeler library was hard masked using protein coding genes from 11 ciliate species (detailed below). The protein coding genes were first filtered to remove any genes with descriptions indicating "transposon" or "helicase". TransposonPSI (r08222010) http://transposonpsi.sourceforge.net was then run to remove any transposon hits by hard-masking them and using the filtered gene set to mask the RepeatModeler library. RepeatMasker v4.0.7 was run with the Repbase Alveolata library (RepBaseRepeatMaskerEdition-20170127.tar.gz) and additionally with the filtered RepeatModeler library. The interspersed repeats were combined and used as evidence in the gene build.

The REAT transcriptome workflow was run with RNA-Seq (total 77 million read pairs) from 28 samples. As transcriptome assembly is sensitive to depth of RNA-Seq coverage samples were combined into sets of 28, 10, 10 and 8 samples to ensure reasonable coverage but also allow alternative assemblies to be created. Illumina RNA-seq reads were mapped to the

genome with HISAT2 v2.2.1 [64] and high-confidence splice junctions identified by Portcullis [65]. The aligned reads were assembled for each set of samples with StringTie2 v2.1.5 [66] and Scallop v0.10.5 [67]. From the combined set of RNA-Seq assemblies a filtered set of non-redundant gene-models were derived using Mikado [68]. The REAT homology workflow was used to generate gene models based on alignment of proteins from 11 ciliate species (**S2 Table**). These together with the transcriptome derived models were used to train the AUGUSTUS v3.4.0 gene predictor, with transcript and protein alignments plus repeat annotation provided as hints in evidence guided gene prediction using the REAT prediction workflow. Six alternative AUGUSTUS gene builds were generated using different evidence inputs or weightings for the protein, transcriptome and repeat annotation. The Minos pipeline was run to generate a consolidated set of gene models from the transcriptome, homology, and AUGUSTUS predictions. The pipeline utilises external metrics to assess how well supported each gene model is by available evidence, based on these and intrinsic characteristics of the gene models a final set of models is selected. For each gene model a confidence and biotype classification were determined based on the type and extent of supporting data.

Annotation completeness was assessed using BUSCO (v4.1.2) [60] with the Alveolata_obd10 dataset run in protein mode. tRNA genes were annotated using tRNAscan (v2.0.7) [35]. rRNA genes were annotated using barrnap (v0.9) (https://github.com/tseemann/barrnap).

Tandem stop codon analysis

To investigate if UGA stop codons are enriched in the 3'-UTR of genes, codon usage of the first 20 in-frame codons downstream of each gene's stop codon was calculated. Expected frequencies were determined by counting codons in all six reading frames in the 60 bp region downstream of each gene's stop codon. We also carried out this analysis for highly expressed genes which we defined as the 10% of genes with the highest transcripts per million (TPM) values, calculated using Kallisto [69]. Statistical significance was assessed using the chi-squared test.

Phylogenetic analysis of SSU rRNA genes

Small subunit ribosomal RNA sequences from related species were retrieved from GenBank (**S1 Fig**). Sequences were aligned using MAFFT (v7.490) with the G-INS-I algorithm [70]. Maximum-likelihood phylogenetic analysis was performed using IQ-TREE (v2.2.0) [71] under the GTR+F+R5 model, which was the best fit model according to ModelFinder [72], with 100 non-parametric bootstrap replicates.

Phylogenetic network analysis of tRNA genes

A multiple sequence alignment of tRNA genes was generated using MAFFT (v7.490) with the G-INS-I algorithm [70]. tRNA genes predicted to be pseudogenes, containing introns, or that were excessively truncated were excluded. A Neighbour-Net phylogenetic network was constructed using SplitsTree4 [73].

Phylogenomic analyses

A phylogenomic dataset of 55 species was assembled including previously published ciliate genomes and transcriptomes with outgroup species from the Alveolata, retrieved from databases and published phylogenomics analyses [74,75] (**S2 Table**). *De novo* transcriptome assemblies were generated for two species–*Campanella umbellaria* and *Carchesium polypinum*.

RNA-Seq reads were retrieved from the sequence read archive (SRR1768423 and SRR1768437) [46]. Transcriptome assemblies were generated using Trinity [76], redundancy was reduced using CD-HIT-EST [77] with an identity cut-off of 98% and protein coding transcripts were predicted using Transdecoder [78]. Coding sequences were translated into amino acids using the correct genetic code (UAR = E). The transcriptome assembly of *Mesodinium rubrum* is contaminated with sequences from its prey. We excluded any *M*. *rubrum* proteins with a best BLAST hit outside of the Ciliophora to account for this contamination which resulted in the removal of 3,574 (22%) proteins.

BUSCO analysis using the Alveolata_obd10 dataset identified 89 proteins that are present and single copy in at least 65% of species, i.e., at least 36 out of 55 species. Each BUSCO family was individually aligned using MAFFT (v7.490) [70] and then trimmed using trimAl (v1.4) with the "gappyout" parameter [79]. The trimmed alignments were concatenated together resulting in a supermatrix alignment of 40,289 amino acid sites. Maximum-likelihood phylogenetic reconstruction was performed using IQ-TREE (v2.2.0) [71] under the LG+F+I+R7 model, which was the best fitting model according to ModelFinder [72], and 100 non-parametric bootstrap replicates were used to assess branch support. We also conducted a partitioned analysis using IQ-TREE [80] with a partitioning scheme that merged the 89 proteins into 14 partitions with model selection performed by ModelFinder, with 100 non-parametric bootstrap replicates. Bayesian analyses were also performed on the supermatrix alignment using PhyloBayes-MPI (v1.8c) [81] under the CAT-GTR model. Constant sites (n = 3,299) were removed. Two independent Markov chain Monte Carlo (MCMC) chains were run for approximately 12,000 generations. Convergence was assessed using bpcomp and tracecomp with a burn-in of 20%.

Supporting information

S1 Fig. Maximum-likelihood phylogeny of small subunit ribosomal RNA genes under the GTR+F+R5 model using IQ-TREE with 100 non-parametric bootstraps.
(PDF)

S2 Fig. Neighbour-Net phylogenetic network analysis of tRNA genes. tRNA genes predicted to be pseudogenes, contain introns, or that were excessively truncated were excluded. Lysine, glutamic acid, and suppressor tRNA genes are highlighted.
(PDF)

S3 Fig. Sequence frequency logo showing the nucleotide composition surrounding stop codons in all genes and in the subset of highly expressed genes.
(PDF)

S4 Fig. Bayesian phylogenomic analysis of 46 ciliate species and 9 outgroup Alveolata species, based on a concatenated alignment of 89 BUSCO proteins under the CAT-GTR model using PhyloBayes-MPI.
(PDF)

S5 Fig. Example multiple sequence alignments of *Plagiopyla frontata* genes with internal UGA codons identified by PhyloFisher with orthologous sequences spanning Eukaryota. **A)** TM9SF1. **B)** PIK3C3. **C)** CRNL1.
(PDF)

S1 Table. tRNA genes pairwise identities.
(XLSX)

S2 Table. Datasets used for phylogenomics and genome annotation.
(XLSX)

S3 Table. Amino acid and codon usage.
(XLSX)

Acknowledgments

We would like to acknowledge members of the Genomics Pipelines, Single-Cell, Core Bioinformatics, and e-Infrastructure groups at the Earlham Institute, and note the specific contributions of Tom Barker, Vanda Knitlhoffer and Chris Watkins. We also acknowledge the Scientific Computing group, as well as support for the physical HPC infrastructure and data centre delivered via the NBI Research Computing group.

Author Contributions

Conceptualization: Thomas A. Richards, Neil Hall.

Formal analysis: Jamie McGowan, Gemy G. Kaithakottil, David Swarbreck.

Funding acquisition: Karim Gharbi, Iain C. Macaulay, Thomas A. Richards, Neil Hall, David Swarbreck.

Investigation: Jamie McGowan, Estelle S. Kilias, Elisabet Alacid, James Lipscombe, Benjamin H. Jenkins.

Project administration: Seanna McTaggart.

Supervision: Karim Gharbi, Iain C. Macaulay, Thomas A. Richards, Neil Hall, David Swarbreck.

Writing – original draft: Jamie McGowan, Estelle S. Kilias, James Lipscombe, Thomas A. Richards, Neil Hall, David Swarbreck.

Writing – review & editing: Jamie McGowan, Estelle S. Kilias, Elisabet Alacid, James Lipscombe, Benjamin H. Jenkins, Karim Gharbi, Gemy G. Kaithakottil, Iain C. Macaulay, Seanna McTaggart, Sally D. Warring, Thomas A. Richards, Neil Hall, David Swarbreck.

References

1. Knight RD, Freeland SJ, Landweber LF. Rewiring the keyboard: evolvability of the genetic code. Nat Rev Genet. 2001; 2: 49–58. https://doi.org/10.1038/35047500 PMID: 11253070
2. Ivanova NN, Schwientek P, Tripp HJ, Rinke C, Pati A, Huntemann M, et al. Stop codon reassignments in the wild. Science. 2014; 344: 909–913. https://doi.org/10.1126/science.1250691 PMID: 24855270
3. Keeling PJ. Genomics: Evolution of the Genetic Code. Current Biology. 2016; 26: R851–R853. https://doi.org/10.1016/j.cub.2016.08.005 PMID: 27676305
4. Parfrey LW, Lahr DJG, Knoll AH, Katz LA. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proceedings of the National Academy of Sciences. 2011; 108: 13624–13629. https://doi.org/10.1073/pnas.1110633108 PMID: 21810989
5. Prescott DM. The DNA of ciliated protozoa. 1994; 58.
6. Arnaiz O, Mathy N, Baudry C, Malinsky S, Aury J-M, Denby Wilkes C, et al. The Paramecium Germline Genome Provides a Niche for Intragenic Parasitic DNA: Evolutionary Dynamics of Internal Eliminated Sequences. Malik HS, editor. PLoS Genet. 2012; 8: e1002984. https://doi.org/10.1371/journal.pgen.1002984 PMID: 23071448
7. Swart EC, Bracht JR, Magrini V, Minx P, Chen X, Zhou Y, et al. The Oxytricha trifallax Macronuclear Genome: A Complex Eukaryotic Genome with 16,000 Tiny Chromosomes. Eisen JA, editor. PLoS Biol. 2013; 11: e1001473. https://doi.org/10.1371/journal.pbio.1001473 PMID: 23382650

8. Lozupone CA, Knight RD, Landweber LF. The molecular basis of nuclear genetic code change in ciliates. Current Biology. 2001; 11: 65–74. https://doi.org/10.1016/s0960-9822(01)00028-8 PMID: 11231122

9. Heaphy SM, Mariotti M, Gladyshev VN, Atkins JF, Baranov PV. Novel Ciliate Genetic Code Variants Including the Reassignment of All Three Stop Codons to Sense Codons in *Condylostoma magnum*. Mol Biol Evol. 2016; 33: 2885–2889. https://doi.org/10.1093/molbev/msw166 PMID: 27501944

10. Meyer F, Schmidt HJ, Plümper E, Hasilik A, Mersmann G, Meyer HE, et al. UGA is translated as cysteine in pheromone 3 of Euplotes octocarinatus. Proc Natl Acad Sci USA. 1991; 88: 3758–3761. https://doi.org/10.1073/pnas.88.9.3758 PMID: 1902568

11. Swart EC, Serra V, Petroni G, Nowacki M. Genetic Codes with No Dedicated Stop Codon: Context-Dependent Translation Termination. Cell. 2016; 166: 691–702. https://doi.org/10.1016/j.cell.2016.06.020 PMID: 27426948

12. 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. Current Biology. 2017; 27: 569–575. https://doi.org/10.1016/j.cub.2016.12.057 PMID: 28190732

13. Kryuchkova P, Grishin A, Eliseev B, Karyagina A, Frolova L, Alkalaeva E. Two-step model of stop codon recognition by eukaryotic release factor eRF1. Nucleic Acids Research. 2013; 41: 4573–4586. https://doi.org/10.1093/nar/gkt113 PMID: 23435318

14. Lekomtsev S, Kolosov P, Bidou L, Frolova L, Rousset J-P, Kisselev L. Different modes of stop codon restriction by the *Stylonychia* and *Paramecium* eRF1 translation termination factors. Proc Natl Acad Sci USA. 2007; 104: 10824–10829. https://doi.org/10.1073/pnas.0703887104 PMID: 17573528

15. Eliseev B, Kryuchkova P, Alkalaeva E, Frolova L. A single amino acid change of translation termination factor eRF1 switches between bipotent and omnipotent stop-codon specificity †. Nucleic Acids Research. 2011; 39: 599–608. https://doi.org/10.1093/nar/gkq759 PMID: 20860996

16. Hanyu N, Kuchino Y, Nishimura S, Beier H. Dramatic events in ciliate evolution: alteration of UAA and UAG termination codons to glutamine codons due to anticodon mutations in two *Tetrahymena* tRNAs^{Gln}. The EMBO Journal. 1986; 5: 1307–1311. https://doi.org/10.1002/j.1460-2075.1986.tb04360.x PMID: 16453685

17. Koonin EV, Novozhilov AS. Origin and evolution of the genetic code: The universal enigma. IUBMB Life. 2009; 61: 99–111. https://doi.org/10.1002/iub.146 PMID: 19117371

18. Liang H, Cavalcanti AR, Landweber LF. Conservation of tandem stop codons in yeasts. Genome Biol. 2005; 6: R31. https://doi.org/10.1186/gb-2005-6-4-r31 PMID: 15833118

19. Fleming I, Cavalcanti ARO. Selection for tandem stop codons in ciliate species with reassigned stop codons. Kapler GM, editor. PLoS ONE. 2019; 14: e0225804. https://doi.org/10.1371/journal.pone.0225804 PMID: 31770405

20. Adachi M, Cavalcanti ARO. Tandem Stop Codons in Ciliates That Reassign Stop Codons. J Mol Evol. 2009; 68: 424–431. https://doi.org/10.1007/s00239-009-9220-y PMID: 19294453

21. Osawa S, Jukes TH, Watanabe K. Recent Evidence for Evolution of the Genetic Code. MICROBIOL REV. 1992; 56. https://doi.org/10.1128/mr.56.1.229-264.1992 PMID: 1579111

22. Schultz DW, Yarus M. Transfer RNA Mutation and the Malleability of the Genetic Code. Journal of Molecular Biology. 1994; 235: 1377–1380. https://doi.org/10.1006/jmbi.1994.1094 PMID: 8107079

23. Andersson S, Kurland C. Genomic evolution drives the evolution of the translation system. Biochem Cell Biol. 1995; 73: 775–787. https://doi.org/10.1139/o95-086 PMID: 8721994

24. Mühlhausen S, Findeisen P, Plessmann U, Urlaub H, Kollmar M. A novel nuclear genetic code alteration in yeasts and the evolution of codon reassignment in eukaryotes. Genome Res. 2016; 26: 945–955. https://doi.org/10.1101/gr.200931.115 PMID: 27197221

25. Kollmar M, Mühlhausen S. Nuclear codon reassignments in the genomics era and mechanisms behind their evolution. BioEssays. 2017; 39: 1600221. https://doi.org/10.1002/bies.201600221 PMID: 28318058

26. Pánek T, Žihala D, Sokol M, Derelle R, Klimeš V, Hradilová M, et al. Nuclear genetic codes with a different meaning of the UAG and the UAA codon. BMC Biol. 2017; 15: 8. https://doi.org/10.1186/s12915-017-0353-y PMID: 28193262

27. Crick FHC. Codon—anticodon pairing: The wobble hypothesis. Journal of Molecular Biology. 1996.

28. Macaulay IC, Haerty W, Kumar P, Li YI, Hu TX, Teng MJ, et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. Nat Methods. 2015; 12: 519–522. https://doi.org/10.1038/nmeth.3370 PMID: 25915121

29. Šlapeta J, Moreira D, López-García P. The extent of protist diversity: insights from molecular ecology of freshwater eukaryotes. Proc R Soc B. 2005; 272: 2073–2081. https://doi.org/10.1098/rspb.2005.3195 PMID: 16191619

30. Poláková K, Čepička I, Bourland WA. Phylogenetic Position of Three Well-known Ciliates from the Controversial Order Loxocephalida Jankowski, 1980 (Scuticociliatia, Oligohymenophorea) and Urozona buetschlii (Schewiakoff, 1889) with Improved Morphological Descriptions. Protist. 2021; 172: 125833. https://doi.org/10.1016/j.protis.2021.125833 PMID: 34562740

31. Gao F, Katz LA, Song W. Multigene-based analyses on evolutionary phylogeny of two controversial ciliate orders: Pleuronematida and Loxocephalida (Protista, Ciliophora, Oligohymenophorea). Molecular Phylogenetics and Evolution. 2013; 68: 55–63. https://doi.org/10.1016/j.ympev.2013.03.018 PMID: 23541839

32. Tice AK, Žihala D, Pánek T, Jones RE, Salomaki ED, Nenarokov S, et al. PhyloFisher: A phylogenomic package for resolving eukaryotic relationships. Hejnol A, editor. PLoS Biol. 2021; 19: e3001365. https://doi.org/10.1371/journal.pbio.3001365 PMID: 34358228

33. Shulgina Y, Eddy SR. A computational screen for alternative genetic codes in over 250,000 genomes. eLife. 2021; 10: e71402. https://doi.org/10.7554/eLife.71402 PMID: 34751130

34. Shulgina Y, Eddy SR. Codetta: predicting the genetic code from nucleotide sequence. Bioinformatics. 2023; 39: btac802. https://doi.org/10.1093/bioinformatics/btac802 PMID: 36511586

35. Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Research. 2021; 49: 9077–9096. https://doi.org/10.1093/nar/gkab688 PMID: 34417604

36. Laslett D. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Research. 2004; 32: 11–16. https://doi.org/10.1093/nar/gkh152 PMID: 14704338

37. Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, et al. Macronuclear Genome Sequence of the Ciliate Tetrahymena thermophila, a Model Eukaryote. Gelfand M, editor. PLoS Biol. 2006; 4: e286. https://doi.org/10.1371/journal.pbio.0040286 PMID: 16933976

38. Krassowski T, Coughlan AY, Shen X-X, Zhou X, Kominek J, Opulente DA, et al. Evolutionary instability of CUG-Leu in the genetic code of budding yeasts. Nat Commun. 2018; 9: 1887. https://doi.org/10.1038/s41467-018-04374-7 PMID: 29760453

39. Mordret E, Dahan O, Asraf O, Rak R, Yehonadav A, Barnabas GD, et al. Systematic Detection of Amino Acid Substitutions in Proteomes Reveals Mechanistic Basis of Ribosome Errors and Selection for Translation Fidelity. Molecular Cell. 2019; 75: 427–441.e5. https://doi.org/10.1016/j.molcel.2019.06.041 PMID: 31353208

40. Aury J-M, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, et al. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature. 2006; 444: 171–178. https://doi.org/10.1038/nature05230 PMID: 17086204

41. Salim HMW, Ring KL, Cavalcanti ARO. Patterns of Codon Usage in two Ciliates that Reassign the Genetic Code: Tetrahymena thermophila and Paramecium tetraurelia. Protist. 2008; 159: 283–298. https://doi.org/10.1016/j.protis.2007.11.003 PMID: 18207458

42. Dabrowski M, Bukowy-Bieryllo Z, Zietkiewicz E. Translational readthrough potential of natural termination codons in eucaryotes–The impact of RNA sequence. RNA Biology. 2015; 12: 950–958. https://doi.org/10.1080/15476286.2015.1068497 PMID: 26176195

43. Bonetti B, Fu L, Moon J, Bedwell DM. The Efficiency of Translation Termination is Determined by a Synergistic Interplay Between Upstream and Downstream Sequences inSaccharomyces cerevisiae. Journal of Molecular Biology. 1995; 251: 334–345. https://doi.org/10.1006/jmbi.1995.0438 PMID: 7650736

44. Beznosková P, Wagner S, Jansen ME, von der Haar T, Valášek LS. Translation initiation factor eIF3 promotes programmed stop codon readthrough. Nucleic Acids Research. 2015; 43: 5099–5111. https://doi.org/10.1093/nar/gkv421 PMID: 25925566

45. Poole ES, Brown CM, Tate WP. The identity of the base following the stop codon determines the efficiency of in vivo translational termination in Escherichia coli. EMBO J. 1995; 14: 151–158. https://doi.org/10.1002/j.1460-2075.1995.tb06985.x PMID: 7828587

46. Feng J-M, Jiang C-Q, Warren A, Tian M, Cheng J, Liu G-L, et al. Phylogenomic analyses reveal subclass Scuticociliatia as the sister group of subclass Hymenostomatia within class Oligohymenophorea. Molecular Phylogenetics and Evolution. 2015; 90: 104–111. https://doi.org/10.1016/j.ympev.2015.05.007 PMID: 25999054

47. Jiang C-Q, Wang G-Y, Xiong J, Yang W-T, Sun Z-Y, Feng J-M, et al. Insights into the origin and evolution of Peritrichia (Oligohymenophorea, Ciliophora) based on analyses of morphology and phylogenomics. Molecular Phylogenetics and Evolution. 2019; 132: 25–35. https://doi.org/10.1016/j.ympev.2018.11.018 PMID: 30496843

48. Gentekaki E, Kolisko M, Gong Y, Lynn D. Phylogenomics solves a long-standing evolutionary puzzle in the ciliate world: The subclass Peritrichia is monophyletic. Molecular Phylogenetics and Evolution. 2017; 106: 1–5. https://doi.org/10.1016/j.ympev.2016.09.016 PMID: 27659723

49. Wang C, Gao Y, Lu B, Chi Y, Zhang T, El-Serehy HA, et al. Large-scale phylogenomic analysis provides new insights into the phylogeny of the class Oligohymenophorea (Protista, Ciliophora) with establishment of a new subclass Urocentria nov. subcl. Molecular Phylogenetics and Evolution. 2021; 159: 107112. https://doi.org/10.1016/j.ympev.2021.107112 PMID: 33609708

50. Rotterová J, Salomaki E, Pánek T, Bourland W, Žihala D, Táborský P, et al. Genomics of New Ciliate Lineages Provides Insight into the Evolution of Obligate Anaerobiosis. Current Biology. 2020; 30: 2037–2050.e6. https://doi.org/10.1016/j.cub.2020.03.064 PMID: 32330419

51. Lasek-Nesselquist E, Johnson MD. A Phylogenomic Approach to Clarifying the Relationship of Mesodinium within the Ciliophora: A Case Study in the Complexity of Mixed-Species Transcriptome Analyses. Zufall R, editor. Genome Biology and Evolution. 2019; 11: 3218–3232. https://doi.org/10.1093/gbe/evz233 PMID: 31665294

52. Lynn DH, Kolisko M. Molecules illuminate morphology: phylogenomics confirms convergent evolution among 'oligotrichous' ciliates. International Journal of Systematic and Evolutionary Microbiology. 2017; 67: 3676–3682. https://doi.org/10.1099/ijsem.0.002060 PMID: 28829032

53. Chen X, Zhao X, Liu X, Warren A, Zhao F, Miao M. Phylogenomics of non-model ciliates based on transcriptomic analyses. Protein Cell. 2015; 6: 373–385. https://doi.org/10.1007/s13238-015-0147-3 PMID: 25833385

54. Seah BKB, Singh A, Swart EC. Karyorelict ciliates use an ambiguous genetic code with context-dependent stop/sense codons. Peer Community Journal. 2022;2. https://doi.org/10.24072/pcjournal.141

55. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology. 2012; 19: 455–477. https://doi.org/10.1089/cmb.2012.0021 PMID: 22506599

56. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019; 7: e7359. https://doi.org/10.7717/peerj.7359 PMID: 31388474

57. von Meijenfeldt FAB, Arkhipova K, Cambuy DD, Coutinho FH, Dutilh BE. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. Genome Biol. 2019; 20: 217. https://doi.org/10.1186/s13059-019-1817-x PMID: 31640809

58. Karlicki M, Antonowicz S, Karnkowska A. Tiara: deep learning-based classification system for eukaryotic sequences. Birol I, editor. Bioinformatics. 2022; 38: 344–350. https://doi.org/10.1093/bioinformatics/btab672 PMID: 34570171

59. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013; 29: 1072–1075. https://doi.org/10.1093/bioinformatics/btt086 PMID: 23422339

60. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Kelley J, editor. Molecular Biology and Evolution. 2021; 38: 4647–4654. https://doi.org/10.1093/molbev/msab199 PMID: 34320186

61. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Research. 2005; 33: W465–W467. https://doi.org/10.1093/nar/gki458 PMID: 15980513

62. Hubley R, Smit A. RepeatModeler. [cited 23 May 2022]. Available: https://www.repeatmasker.org/RepeatModeler/

63. Smit AF, Hubley R, Green P. RepeatMasker. [cited 23 May 2022]. Available: https://www.repeatmasker.org/

64. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019; 37: 907–915. https://doi.org/10.1038/s41587-019-0201-4 PMID: 31375807

65. Mapleson D, Venturini L, Kaithakottil G, Swarbreck D. Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. GigaScience. 2018; 7: giy131. https://doi.org/10.1093/gigascience/giy131 PMID: 30418570

66. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biology. 2019; 20: 278. https://doi.org/10.1186/s13059-019-1910-1 PMID: 31842956

67. Shao M, Kingsford C. Accurate assembly of transcripts through phase-preserving graph decomposition. Nat Biotechnol. 2017; 35: 1167–1169. https://doi.org/10.1038/nbt.4020 PMID: 29131147

68. Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. GigaScience. 2018; 7: giy093. https://doi.org/10.1093/gigascience/giy093 PMID: 30052957

69. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016; 34: 525–527. https://doi.org/10.1038/nbt.3519 PMID: 27043002

70. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution. 2013; 30: 772–780. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

71. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Teeling E, editor. Molecular Biology and Evolution. 2020; 37: 1530–1534. https://doi.org/10.1093/molbev/msaa015 PMID: 32011700

72. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017; 14: 587–589. https://doi.org/10.1038/nmeth.4285 PMID: 28481363

73. Huson DH, Bryant D. Application of Phylogenetic Networks in Evolutionary Studies. Molecular Biology and Evolution. 2006; 23: 254–267. https://doi.org/10.1093/molbev/msj030 PMID: 16221896

74. Irwin NAT, Pittis AA, Mathur V, Howe LJ, Keeling PJ, Lynn DH, et al. The Function and Evolution of Motile DNA Replication Systems in Ciliates. Current Biology. 2021; 31: 66–76.e6. https://doi.org/10.1016/j.cub.2020.09.077 PMID: 33125869

75. Richter DJ, Berney C, Strassert JFH, Poh Y-P, Herman EK, Muñoz-Gómez SA, et al. EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. Peer Community Journal. 2022; 2. https://doi.org/10.24072/pcjournal.173

76. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011; 29: 644–652. https://doi.org/10.1038/nbt.1883 PMID: 21572440

77. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012; 28: 3150–3152. https://doi.org/10.1093/bioinformatics/bts565 PMID: 23060610

78. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013; 8: 1494–1512. https://doi.org/10.1038/nprot.2013.084 PMID: 23845962

79. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009; 25: 1972–1973. https://doi.org/10.1093/bioinformatics/btp348 PMID: 19505945

80. Chernomor O, von Haeseler A, Minh BQ. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Syst Biol. 2016; 65: 997–1008. https://doi.org/10.1093/sysbio/syw037 PMID: 27121966

81. Lartillot N, Rodrigue N, Stubbs D, Richer J. PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. Systematic Biology. 2013; 62: 611–615. https://doi.org/10.1093/sysbio/syt022 PMID: 23564032