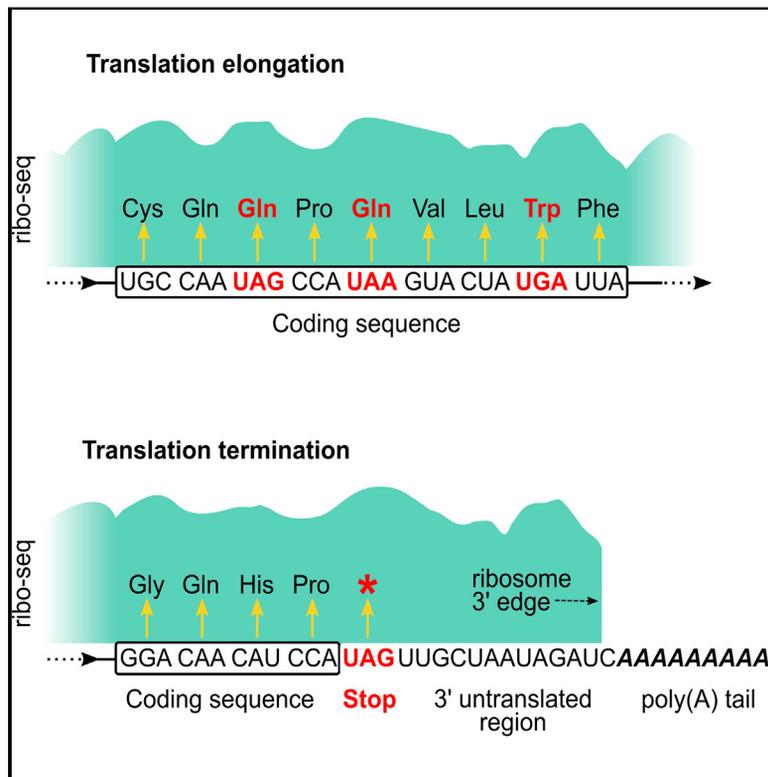


Genetic Codes with No Dedicated Stop Codon: Context-Dependent Translation Termination

Graphical Abstract



Authors

Estienne Carl Swart, Valentina Serra, Giulio Petroni, Mariusz Nowacki

Correspondence

mariusz.nowacki@izb.unibe.ch

In Brief

In some ciliates, all three “stop codons” can either terminate translation or code for an amino acid. Ribosomes may interpret this ambiguity using downstream features in the transcript, indicating that translational termination can be context-dependent.

Highlights

- Alternative nuclear genetic codes continue to be discovered in ciliates
- Genetic codes with stops and all their codons encoding standard amino acids exist
- Transcript ends may distinguish stop codons as such in ambiguous genetic codes
- The ability to resolve genetic code ambiguity may enable genetic code evolution

Genetic Codes with No Dedicated Stop Codon: Context-Dependent Translation Termination

Estienne Carl Swart,¹ Valentina Serra,² Giulio Petroni,² and Mariusz Nowacki^{1,*}

¹Institute of Cell Biology, University of Bern, 3012 Bern, Switzerland

²Department of Biology, University of Pisa, Pisa 56126, Italy

*Correspondence: mariusz.nowacki@izb.unibe.ch

<http://dx.doi.org/10.1016/j.cell.2016.06.020>

SUMMARY

The prevailing view of the nuclear genetic code is that it is largely frozen and unambiguous. Flexibility in the nuclear genetic code has been demonstrated in ciliates that reassign standard stop codons to amino acids, resulting in seven variant genetic codes, including three previously undescribed ones reported here. Surprisingly, in two of these species, we find efficient translation of all 64 codons as standard amino acids and recognition of either one or all three stop codons. How, therefore, does the translation machinery interpret a “stop” codon? We provide evidence, based on ribosomal profiling and “stop” codon depletion shortly before coding sequence ends, that mRNA 3' ends may contribute to distinguishing stop from sense in a context-dependent manner. We further propose that such context-dependent termination/readthrough suppression near transcript ends enables genetic code evolution.

INTRODUCTION

The first exceptions to the supposed universality of eukaryotic nuclear genetic codes were reported in ciliates (Caron and Meyer, 1985; Helftenbein, 1985; Horowitz and Gorovskiy, 1985; Preer et al., 1985). Subsequently, additional genetic codes were discovered in other ciliates, all due to stop codon reassignments, and appear to recur independently in different ciliate lineages (Lozupone et al., 2001; Sánchez-Silva et al., 2003; Tourancheau et al., 1995). Genetic code evolution is considered to have both an ancient phase, which gave rise to the standard genetic code before the radiation of bacteria, archaea, and eukaryotes, and a modern phase, which led to diversification from the standard code (Sengupta and Higgs, 2015). Thus far, alternative nuclear genetic codes have only been found in three major eukaryotic lineages other than ciliates. The first alternative nuclear genetic code, discovered in ciliates, with the UAA and UAG stop codons reassigned to glutamine, is also present in green algae (*Acetabularia* and *Batophora*) (Schneider and de Groot, 1991; Schneider et al., 1989) and diplomonads (Keeling and Doolittle, 1996). Alternative nuclear genetic codes, with CUG reassigned from leucine, also occur in the yeasts *Candida albicans* (predominantly to serine) and *Pachysolen tannophilus* (to alanine)

(Gomes et al., 2007; Mühlhausen et al., 2016; Santos and Tuite, 1995).

Other than the diversity of genetic codes in ciliates, the greatest number of variant genetic codes are found in mitochondria (Knight et al., 2001), whose diversification may have been facilitated by their small genomes and strong mutational biases, which increase the likelihood of loss and reassignment of rare codons (Osawa and Jukes, 1989). Expressed ciliate genomes (macronuclear genomes) are not especially small (typically 50–100 Mb) (Swart et al., 2013), and the manner in which changes in their genetic codes arose may not be as straightforward as that in smaller mitochondrial genomes. Alternative explanations for the evolution of ciliate genetic codes, such as the abolishment of recognition of certain stop codons by mutations in the stop-recognizing translation termination factor eukaryotic release factor 1 (eRF1) allowing codon reassignment have therefore been proposed (Lozupone et al., 2001).

While the genetic code is classically taught as being unambiguous, and indeed may largely be so, we now know this is an oversimplification. Since the original discovery of the standard genetic code, alternative translational interpretations of codons have been found, most notably in the use of the UGA codon for selenocysteine incorporation, in the context of special mRNA stem-loops in the UTRs of a small number of protein-coding genes (Nasim et al., 2000). An additional form of codon ambiguity, translational readthrough of stop codons, is now also recognized as pervasive, but usually weak, in eukaryotes, occurring at a few percent or less compared to the non-readthrough form (e.g., Dunn et al., 2013; Harrell et al., 2002; Roy et al., 2015). Translational readthrough usually gives rise to short protein extensions, e.g., a median length of 35 amino acids in *Drosophila* (Jungreis et al., 2011). Readthrough is enabled by near-cognate pairing of tRNAs to codons, with either the first or third anticodon base non-canonically paired (Blanchet et al., 2014). Thus, there is competition for the same codons between eRF1 and tRNAs.

Although the options for engineering of new genetic codes with artificial amino acids have been proliferating (Lemke, 2014), many important questions about natural genetic codes remain unresolved. Among these questions, are basic ones of how codons are recognized in variant genetic codes with stop codon reassignments and whether there is competition between eRF1 and stop-cognate tRNAs for the same codons. Experimental evidence attempting to address the former problem has been conflicting, supporting either loss or ongoing recognition of reassigned stop codons by eRF1 (Eliseev et al., 2011; Lekomtsev et al., 2007; Salas-Marco et al., 2006; Vallabhaneni et al., 2009).

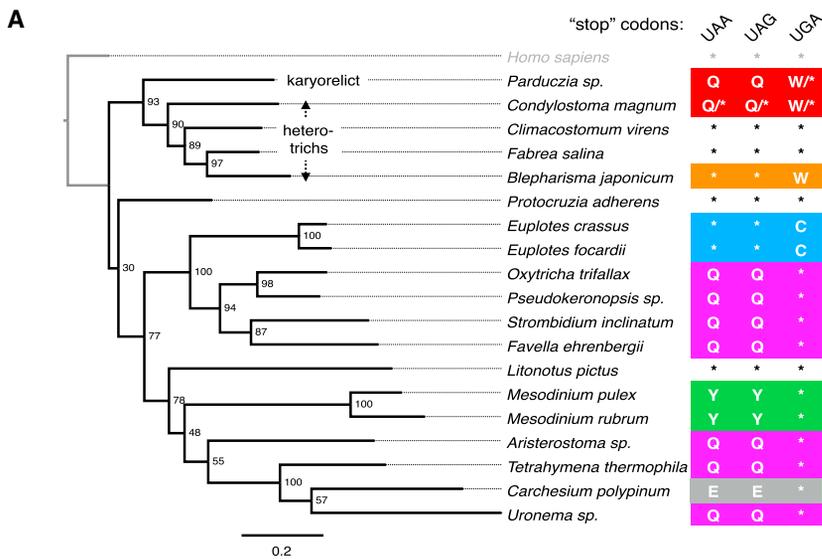
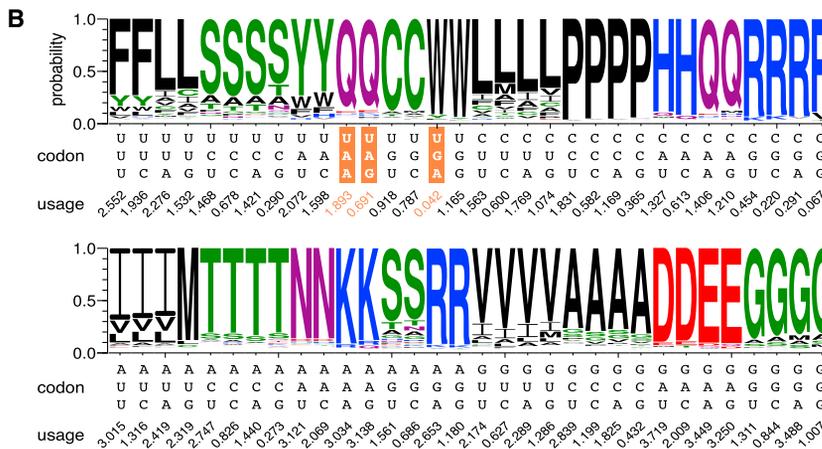


Figure 1. New Genetic Codes

(A) Stop codon reassignments (Q, glutamine; W, tryptophan; C, cysteine; Y, tyrosine; *, stop) are mapped onto an eRF1 maximum likelihood phylogeny. *Homo sapiens* (standard genetic code) is an outgroup. Bootstrap support for every node is shown. Scale bar indicates amino acid substitutions per site. UGA codons were previously found in the coding sequences of *Blepharisma americanum* and were predicted to encode tryptophan (Eliseev et al., 2011; Lozupone et al., 2001). Experimental assays in *Blepharisma japonicum* suggest its eRF1 recognizes all three standard stop codons (Eliseev et al., 2011). It should be noted that ciliates from the family Mesodiniidae have both a unique genetic code (UAG/UAA = UAR = tyrosine; UGA = stop) and extremely divergent rRNAs (Johnson et al., 2004). (B) Predicted *C. magnum* genetic code. Stop codons are highlighted in orange. Predicted amino acids are those with maximal heights. Codon usage inferred from translated BLAST matches is shown below the codons. UAA and UAG codons were previously predicted to encode glutamine (Lozupone et al., 2001; Tourancheau et al., 1995). See also Figure S1 and Table S1.



phan codon, although it does so at low levels (0.059%) and hence this reassignment may easily go undetected in small sequence samples (Figures 1B, S1A, and S1B). Thus, given this reassignment and previous experimental results (Eliseev et al., 2011), we deduce that *B. japonicum*'s eRF1 and at least one of its tryptophan tRNAs may be in competition for the same codon.

Because MMETSP represents the current broadest eukaryotic molecular diversity survey (Keeling et al., 2014) we

With extensive sequence data spanning a wide range of eukaryotes, including ciliates, now available, uncertain genetic codes may be properly determined, and consequently, the proposed basis for nuclear genetic code diversification is also ripe for reinvestigation. We present the new genetic codes we discovered in the course of screening a large collection of eukaryotic transcriptomes, how codons may have multiple meanings in two of these codes, and the consequences of tolerance of genetic code ambiguity for genetic code evolution.

RESULTS

Genetic Codes in which All 64 Codons Encode Standard Amino Acids

To identify and classify reassigned codons, we used a computational screening approach to search the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) transcriptomes (Keeling et al., 2014). We found that like *Bemidion americanum*, *Bradyrhizobium japonicum* uses UGA as a trypto-

screened all its transcriptomes to search for new genetic codes. In our screen, we discovered three new genetic codes among 24 ciliate species (Figures 1A, 1B and S1; Data S1A), but no new codes in the remaining 265 eukaryotes (Data S1B). Unexpectedly, in two of these genetic codes, belonging to the heterotrichous ciliate *Condylostoma magnum* and an unclassified karyorelict (18S rRNA 95% identical to that of *Parduzcia orbis* [Edgcomb et al., 2011]; *Parduzcia sp.* hereafter) all three “stop” codons are predicted to be reassigned to amino acids: UAA = Q, UAG = Q, UGA = W. As the remaining *C. magnum* and *Parduzcia sp.* codons encode standard amino acids (Figures 1A and S1A), all 64 of their codons are translated. Hence, the question is if and how translation termination occurs given these codes.

Because the UGA codon usage in *C. magnum*, *Parduzcia sp.*, and *B. japonicum* is relatively low (0.042%, 0.120%, and 0.059%, respectively), to computationally assess the hypothesis that the *C. magnum* and *Parduzcia sp.* genes with in-frame UGA codons are functional, and not simply pseudogenes with in frame

stops, we sought essential single copy genes with in-frame UGAs and examined their substitution rates. In-frame UGA codons are present in critical genes, such as *C. magnum* tryptophan-tRNA ligase (Figure 2B; MMETSP0210: CAMNT_0008287141) and eRF1 of *Parduzia sp.* (MMETSP1317: CAMNT_0047593165). Substitution rates of genes such as these support the hypothesis of functionality since they indicate strong purifying selection, e.g., for *C. magnum* tryptophan-tRNA ligase aligned to *Oxytricha trifallax* tryptophan-tRNA ligase, d_N/d_S is 0.013 (d_N/d_S = nonsynonymous substitutions per nonsynonymous site over synonymous substitutions per synonymous site; $d_N/d_S < 1$ indicates purifying selection) (Yang, 2007). The hypothesis that UGA codons are translated was assessed experimentally in two ways: we determined that UGA codons are translated as tryptophan by protein mass spectrometry (Data S1D and S1E); using ribosome profiling we observe that ribosomes efficiently translate through UGA codons, as they also do through UAG and UAA codons (Figures 2B and S3E).

The Genetic Codes of *C. magnum* and *Parduzia sp.* Are Ambiguous

Given evidence that all three “stop” codons in the *C. magnum* and *Parduzia sp.* genetic codes can be translated, we wished to assess how translation termination occurs. To investigate the nature of translation termination in *C. magnum* and *Parduzia sp.* we began by examining histone H4 coding sequence ends, since the proteins encoded by these sequences are among the most highly conserved proteins and typically have the same C-terminal residues (e.g., 95% of 105 reviewed UniProt histone H4 proteins end with two glycines; Feb 9, 2015). With respect to the conserved C-terminal amino acid of histone H4 homologs in other eukaryotes, each of the *C. magnum* histone H4 paralog coding sequences is expected to end with a C-terminal glycine codon (Figure 2C). The codon immediately following this, either UAG or UGA, is therefore a candidate stop. The coding sequence of the single histone H4 in the *Parduzia sp.* transcriptome is followed by a UGA codon at the expected stop position (Figure 2C). With respect to aligned homologs from other organisms, all the *Parduzia sp.* transcripts we inspected have a UGA where a stop codon would normally be expected. *C. magnum* also has transcripts that have only the possibility of UAA stops in proximity to where stops are expected (Figures S2B–S2D). From the sequence alignments, we therefore infer that *C. magnum*’s eRF1 recognizes all three standard stop codons and hence needs to outcompete stop cognate tRNAs to terminate translation.

To test whether translation termination occurs at the putative histone H4 stop codons, we used ribosome profiling (ribo-seq). For *C. magnum*’s histone H4.1b and H4.1c forms, it can be seen that translation terminates precisely at the predicted stop codons (Figure 2D), whereas it does so with a small amount of imprecision for H4.1d (Figure 3A; H4.1a was insufficiently covered by ribo-seq reads to assess termination). In general, translation terminating *C. magnum* translation terminating ribosome-protected fragments (RPFs) end 11/12 nucleotides (nt) after stop codon 3’ nt (Figure 3D—compare to sense codons in Figure 3C; Figure 2D is a typical example). Consequently, both the primary and secondary H4.1d stop codons, UAG and UAA, trigger trans-

lation termination, and the typical histone H4 C-terminus may occasionally be extended by one or more amino acids.

While readthrough is conventionally classified as translation of stop codons by near-cognate tRNAs, in *C. magnum*, which has stop cognate tRNAs (see next section), translation through stop codons by near-cognate tRNAs is effectively indistinguishable from translation by cognate tRNAs in ribo-seq data. Therefore, for the sake of simplicity, in *C. magnum*, we classify readthrough as translation through codons that typically trigger translation termination (as for H4.1d). It should be noted that in *C. magnum*, multiple translation termination opportunities often exist before the ribosome translates into poly(A) tails (on average approximately five codons intervene between the primary and additional downstream non-primary stops). As a consequence, if extensions result from readthrough they are typically expected to be very short. Even though multiple possible stop codons exist, examples of imprecise termination as in H4.1d are in the minority: ~90% of transcripts examined with >20 RPFs situated at their stops show no readthrough. Thus, overall readthrough is quite low, e.g., a mean of <1.8% and median of 0% (Figure S3K). The small amount of readthrough that does occur is most readily detected when the ribosome occupies downstream stops (Figure 3E).

Multiple lines of evidence therefore demonstrate that “stop” codons as a class in the *C. magnum* and *Parduzia sp.* genetic codes are ambiguous, whereas their individual codons are typically recognized unambiguously as either sense or stops, solving the translation termination paradox.

In Search of tRNAs that Enable “Stop” Codon Translation

All model ciliates have “suppressor” tRNAs that are complementary to and permit translation of reassigned stop codons (Eisen et al., 2006; Hanyu et al., 1986; Kuchino et al., 1985). Although we found a comprehensive set of tRNAs in our *C. magnum* genome assemblies, including glutamine tRNAs capable of recognizing UAA and UAG codons (Figures 4A and 4B; Data S1G), we were unable to detect tRNA^{Trp}s with UCA anticodons. Given the high sequence coverage of the *C. magnum* macronuclear genome, it is unlikely that we missed tRNA^{Trp}(UCA)s. Ciliates possess both a micronuclear and a macronuclear genome, with the former predominantly unsequenced in our *C. magnum* assembly due to its comparatively low ploidy. It is also unlikely that tRNA^{Trp}(UCA)s have gone undetected because they are micronuclear genome-encoded: although these genomes are transcriptionally active during ciliate sexual development they are generally inactive during vegetative growth (Chen et al., 2014; Nowacki et al., 2009) when many transcripts with UGA tryptophan codons are expressed. To test if CCA → UCA anticodon editing produces a UGA-cognate tRNA^{Trp}, we sequenced RT-PCR products targeting nuclear genome-encoded tRNA^{Trp}s and examined tRNA reads from small RNA sequencing data, but found no signs of significant anticodon editing (see Supplemental Experimental Procedures).

All sequenced ciliate mitochondrial genomes encode a UGA-cognate tRNA^{Trp}(UCA) (Swart et al., 2013) and so does that of *C. magnum* (Figure S4A). Experiments in cell-free lysates show cytoplasmic ribosomes can use yeast mitochondrial

tRNA^{Trp}(UCA) to translate UGA codons (Tuite and McLaughlin, 1982). Thus, to determine whether *C. magnum*'s mitochondrial tRNA^{Trp}(UCA)s are used to translate its mRNA UGA codons, it will be necessary to show these tRNAs are accessible to cytoplasmic ribosomes in quantities adequate for translation.

In standard genetic code organisms, readthrough UGA stop codons are preferentially translated as tryptophan (e.g., for *Saccharomyces cerevisiae*: UGA: 86% W, 7% C, 7% R) (Roy et al., 2015) by near-cognate tRNA^{Trp}(CCA)s. Near-cognate pairing of tRNA^{Trp}(CCA) to UGA may also be substantially enhanced through particular mutations, e.g., in *Escherichia coli* a tRNA^{Trp}(CCA) D-stem point mutation leads to 30× more tryptophan translation at UGA stop codons than the wild-type tRNA (Hirsh, 1971; Hirsh and Gold, 1971). *C. magnum* has three types of tRNA^{Trp}(CCA) (Figures S4B and S4C), and it will be necessary to experimentally assess if any of these tRNAs permits efficient translation of its mRNA UGA codons.

“Stop” Codon Recognition Switches from Sense in Coding Sequences to Stop Near Transcript Ends

We assessed two hypotheses for how sense codons are distinguished from stop codons in ambiguous codes: (1) that there are sequence-specific features (motifs) allowing discriminating protein factors to bind nearby sense and stop codons, and (2) that proximity to transcript ends results in recognition of stops. We reject the hypothesis that specific sequences are necessary for stop/sense discrimination for the following reasons: (1) the base composition around sense “stop” codons is not constrained (Figure S5A), and (2) although the bases flanking *C. magnum* stop codons are weakly biased (Figure S5B), and such biases exist in other eukaryotes, where they are associated with enhanced termination efficiency (McCaughan et al., 1995), it is trivial to find sense “stop” codons with the preferred stop codon flanking Us, thus flanking bases cannot be sufficient to distinguish stop codons.

We next assessed if the proximity of the “stop” codon to transcript ends might determine sense/stop state. While analyzing ciliate 3' UTRs we were struck by how short they are, with those of heterotrichs the shortest of all (median lengths, excluding the poly(A) tail and stop codon: 21–23 nt; Figure 5A). In the literature, we could find no eukaryotes with shorter 3' UTRs. In comparison, yeast, metazoan, and plant 3' UTRs typically have a >100 nt length mode and may be considerably longer (Aoki et al., 2010; Jan et al., 2011). Because poly(A) tails of certain *C. magnum* transcripts, especially those with UAA stop codons, start immediately after their stop codon (Figures 5B–5D) stops can be situ-

ated adjacent to poly(A)-binding proteins (PABPs) in vivo, and hence translation may be terminated with no additional information encoded by 3' UTRs. Because the ribosome occupies 11 or 12 nucleotides downstream of *C. magnum* stop codons, even for those transcripts with 3' UTRs, there may be little room for ribosomes to maneuver passed stop codons without displacing PABPs. Given such short 3' UTRs in ciliates, we therefore propose that nearby protein-bound poly(A) tails may contribute to discriminating stop from sense.

The very low readthrough levels detected in *C. magnum* by ribosome profiling imply that when “stop” codons are positioned close to transcript ends the probable outcome is termination. The few “stop” codons existing in the vicinity before stop codons (24–66 nt upstream; mean 50 nt upstream; 16 out of 1,672 transcripts) are efficiently translated and show no signs of appreciable premature translation termination (Figure S3I). Given the low tolerance of either readthrough or premature translation termination, the prediction is that when codons recognized inefficiently as either stop or sense arise in coding sequences, they are deleterious. Thus, in the hypothesis of discrimination of codons as stops close to transcript ends, if “stop” codons arise just upstream of the proper stops, where they might either be translated or result in premature termination, they will be counterselected and hence decrease in frequency. Consistent with this hypothesis, such a decrease in “stop” codon frequency exists in the upstream coding sequence vicinity of the stops in *C. magnum* (UAA, UAG, UGA) and *Parduczia* sp. (UGA) (Figures 6 and S6). Conversely, no codons other than “stop” codons become rare in coding sequences just before the actual stops (e.g., *C. magnum*; Figure S6). Furthermore, following cognate tRNA acquisition CAA and CAG frequencies are expected to remain higher near stops than distal coding sequence regions, since these codons may not freely mutate to UAA and UAG without causing premature translation termination (Figure 6D; unlike any other codons [Figure S6]; given the low UGA sense codon usage, only a small fraction of UGG codons has mutated to UGA, and UGG codon frequencies are not expected to be higher near stops).

DISCUSSION

Based on the observations of ribosome positioning and distribution of “stop” codons in transcripts, for translation in *C. magnum* and *Parduczia* sp. we propose a model where translation, rather than termination, is the default recognition mode for “stop” codons and where termination is due to the context-specific

Figure 2. “Stop” Codons in *C. magnum* and *Parduczia* sp.: Either Sense or Stop Codons

(A) *C. magnum* protein kinase alignment region highlighting putative sense “stop” codons. Standard genetic code stop codons are shown with stars, with larger stars for UGA. MMETSP0210 IDs: CAMNT_0008311047, CAMNT_0008316317, CAMNT_0008295895, CAMNT_0008281491, CAMNT_0008274923, CAMNT_0008274561, CAMNT_0008271577, CAMNT_0008291651, CAMNT_0008280967, CAMNT_0008289329.

(B) Ribosome-protected fragments (RPFs) mapped to a *C. magnum* tryptophan-tRNA ligase transcript (Data S1AC and S1AD). “RPF coverage” is calculated from all the bases of 25–32 nt RPFs.

(C) Histone H4 C-termini and stop codons (gray arrow, coding sequence) from *C. magnum*, *Parduczia* sp., and *Homo sapiens*. Poly(A) tails are visible at *C. magnum* and *Parduczia* sp. mRNA 3' termini. Histone H4.1a–H4.1d: MMETSP0210 IDs: CAMNT_0008274265, CAMNT_0008297091, CAMNT_0008284521, and CAMNT_0008296393; *Parduczia* sp. histone H4 is MMETSP137 CAMNT_0047598059. *H. sapiens* histone H4 is GenBank: M16707.1. Judging from paired-end read mapping, the 3' UTR of H4.1a is incorrectly fused to a downstream transcript.

(D) RPFs mapped to histone H4.1c (Data S1AE and S1AF).

See also Figure S2.

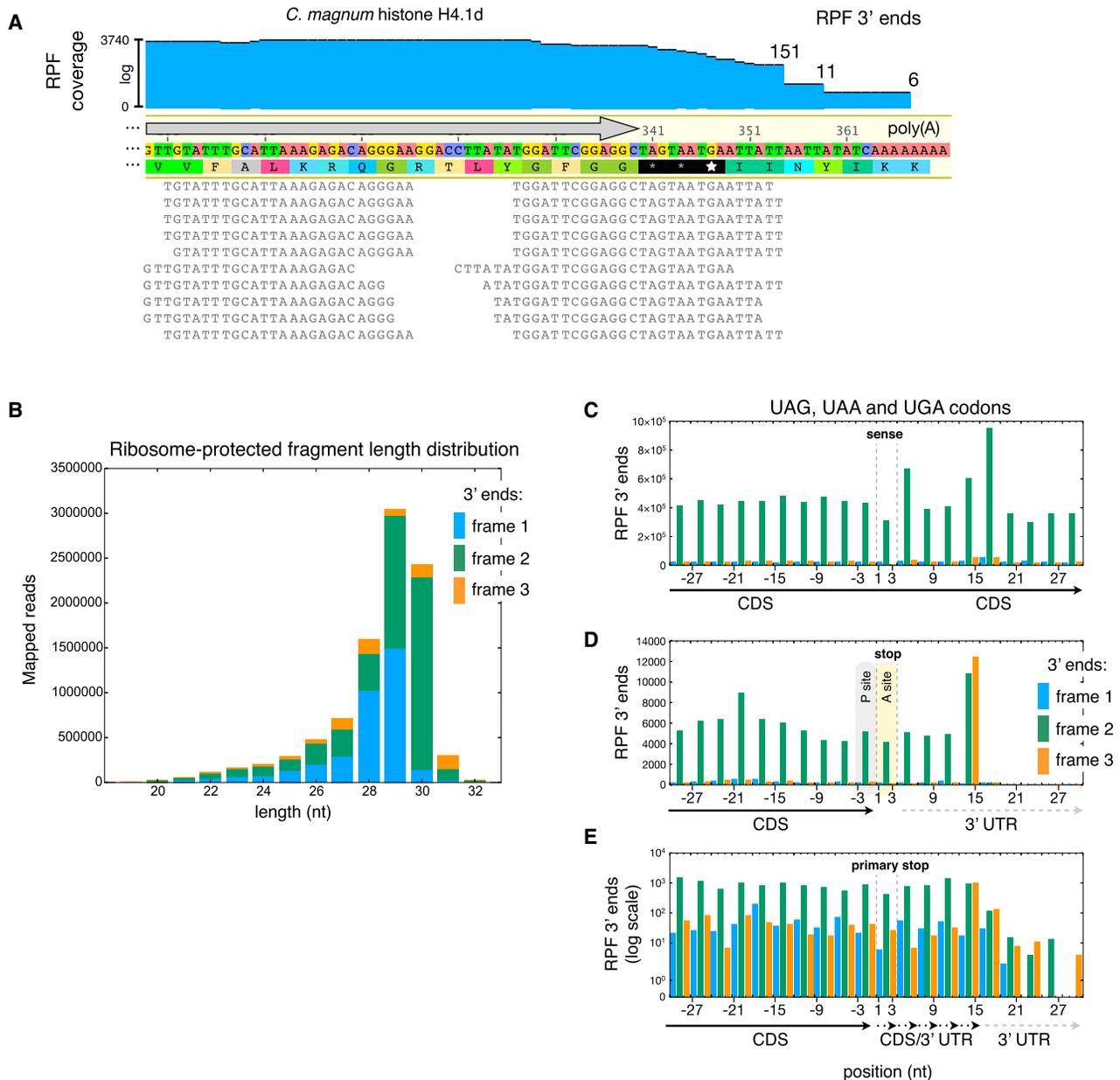


Figure 3. Ribosome Profiling Reveals Different Ribosome States at “Stop” Codons

(A) RPFs (25–32 nt) mapped to histone H4.1d (Data S1AG and S1AH). RPF 3' termini counts are given at the sequence coverage steps: the first and second steps correspond to ribosomes whose P-sites are the first and second stop codons, respectively.

(B) RPF read length distribution and frame distribution. For the 3U TruSeq ribo profile nuclease digestion more mRNA reads were present due to lower rRNA degradation, and most 30-nt RPFs have their 3' ends in frame 3 (compare to Figures S3A and S3B).

(C and D) Distribution of 30 nt RPF 3' ends around sense (C) and stop (D) UAG, UGA, and UAA codons (positions 1–3, indicated by dashed vertical lines) in Trinity assembled transcripts. CDS, coding sequence; UTR, untranslated region. Putative ribosomal P- and A-site locations of translation terminating RPFs situated at stop codons, based on that predicted for other eukaryotic ribosomes (Chung et al., 2015). Figures S3C–S3H show the distribution of RPF 3' ends around individual “stop” codons. Though the termination signal is most pronounced for 30-nt RPFs, it is also exhibited by other RPFs (Figure S3J).

(E) Distribution of 30-nt RPFs for transcripts with detected readthrough (≥ 13 nt downstream of the primary stop codon); additional stop codons are located downstream of the primary one, hence the region downstream of the primary stop may be either coding or untranslated.

See also Figure S3.

override provided by transcript ends (Figure 7). Thus, at sense “stop” codons, tRNAs outcompete eRF1, and at proper stop codons, eRF1 outcompetes tRNAs. The converse model (default

termination; context-specific translation), is not consistent with our results, and given preexisting surrounding coding sequence constraints, widespread context-specific translation signals

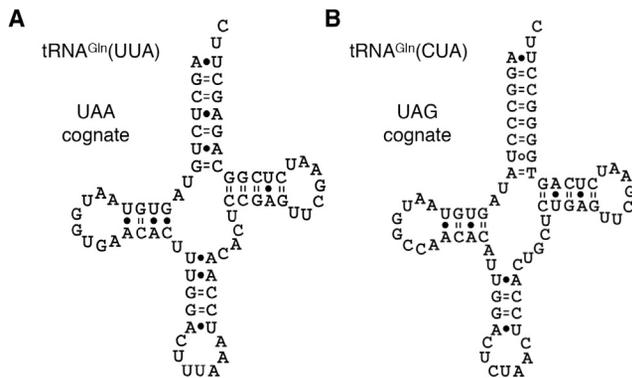


Figure 4. Predicted UAA- and UAG-cognate *C. magnum* tRNAs
 (A and B) UAA- and UAG-cognate glutamine tRNA secondary structures. Bonds shown are predicted by the RNAfold web server (Lorenz et al., 2011) (default parameters). See also Figure S4.

necessary to translate all the “stop” codons are exceedingly unlikely to arise.

Given the existence of transcripts without 3' UTRs, we deduce these regions are not essential for translation termination, and we propose that the close proximity of a poly(A) tail and poly(A)-interacting proteins, in particular PABPs, alone may be necessary to trigger termination. Three prior observations favor this hypothesis: (1) PABP overexpression enhances translation termination when it is weak, implying that PABPs may be involved in translation termination (Cosson et al., 2002), (2) tethering of a PABP 37–73 nt downstream of a premature stop codon substantially decreases NMD and results in recruitment of the translation termination factor eRF3, suggesting that PABP is involved in discriminating stops from premature stops (Amrani et al., 2004); and (3) PABPs bind to AU-rich RNA including 3' UTRs (Baejen et al., 2014; Kini et al., 2016; Sladic et al., 2004).

Reassigned “stop” codons in *C. magnum* and *Parduzia sp.* differ from conventional readthrough stops in standard genetic code organisms because they are efficiently translated and distributed throughout coding sequences, whereas conventional readthrough stops are the major termination signals whose disregard gives rise to modest levels of short protein extensions (Dunn et al., 2013; Jungreis et al., 2011). From their distribution throughout coding sequences, it is evident that most reassigned codons in ciliates arose from substitutions of codons that were already normally translated, rather than from readthrough stop codons. Upon acquisition of a stop cognate tRNA, a shift in balance from translation termination to readthrough at stop codons is expected. Normally this acquisition would immediately be deleterious, due to the creation of aberrant C-terminal peptide signals or the triggering of non-stop mRNA decay (Frischmeyer et al., 2002) upon translation into mRNA poly(A) tails. By enforcing proper translation termination close to transcript ends, ciliates with ambiguous genetic codes provide a way of getting around these problems.

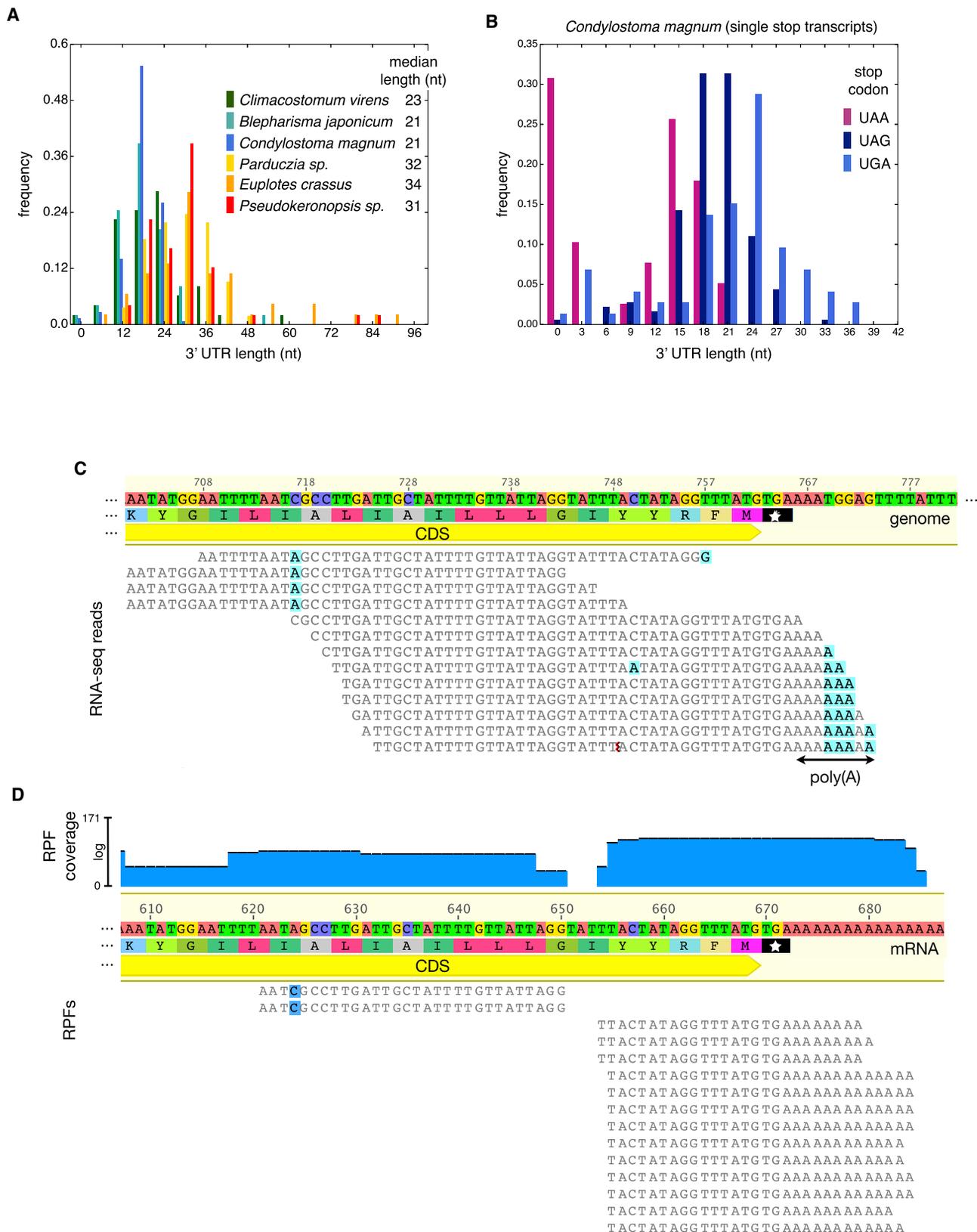
Given that we detected no new genetic codes in 265 diverse non-ciliate eukaryotic species from MMETSP, the abundance of alternative genetic codes within ciliates is all the more striking.

Two hypotheses for the origin of genetic codes in ciliates are that they were enabled by codon capture or eRF1 mutations. Under the “codon capture” hypothesis (Osawa and Jukes, 1989) when a codon disappears in a genome due to strong mutational biases it may then be reassigned when a suitable cognate tRNA arises (via tRNA duplication and anticodon mutation) and the codon subsequently reappears. To date, all sequenced ciliate genomes are AT rich (Aeschlimann et al., 2014; Aury et al., 2006; Coyne et al., 2011; Eisen et al., 2006; Swart et al., 2013; Wang et al., 2016). Reflecting their A/T mutational biases, among eukaryotes with the highest UAA stop codon usage are standard genetic code ciliates (Figures S7B–S7D; Data S1V). This suggests that the diversification of genetic codes from the standard one could have followed UAG and UGA stop codon depletion in ancestral ciliates with AT rich genomes. While codon capture is a reasonable explanation for the evolution of the *Blepharisma* genetic code (UAA stop codon usage 91%), it does not readily explain the origin of other ciliate genetic codes. For example, in *Euplotes sp.*, according to tRNA anticodon-codon wobble rules, UGG codons are expected to be misread as cysteine following the origin of a tRNA^{Cys}(UCA).

Even when relaxing the stop codon disappearance criterion (via genetic code ambiguity tolerance), codon capture cannot easily explain the general UAG and UAA reassignment trends seen in Figure 1A. In all ciliates with reassigned UAG and UAA codons and complete macronuclear genomes, both tRNAs with anticodon complements of these codons are present (Aeschlimann et al., 2014; Aury et al., 2006; Coyne et al., 2011; Eisen et al., 2006; Swart et al., 2013). In the event that the first acquisition during codon reassignment was a tRNA(UUA), by the codon-anticodon wobble rules UAA and UAG would both be translated; however, as this requires prior UAA stop codon disappearance, it is contrary to the ciliate mutational tendencies. If codon reassignment were to occur after a tRNA(CUA) acquisition, only UAG codons would be translated, and under the codon capture hypothesis, genetic codes with UAG reassignment alone should be common; however, this is not observed. Therefore, codon capture alone cannot explain the diversity of genetic codes in ciliates.

As eRF1 recognizes stop codons, this protein could be a determinant of genetic code reassignments in ciliates. Previously it was hypothesized that particular eRF1 amino acid substitutions are associated with each variant genetic code (Lozupone et al., 2001). The additional ciliate genetic codes and eRF1 diversity present in ciliates and other eukaryotes present multiple contradictions to the reported concordances between eRF1 amino acid substitutions and variant genetic codes (Lozupone et al., 2001) (Figure S7A). Because no obvious associations between single eRF1 substitutions and variant genetic codes are evident, any possible associations between genetic codes and eRF1 changes must be more complex than individual amino acid changes. The existence of the ambiguous ciliate genetic codes is also a challenge to explain by this hypothesis.

Because ciliate genetic code diversity does not seem to be adequately explained by codon capture or eRF1 changes, we instead propose that it is due to past genetic code ambiguity tolerance and resolution, as exemplified by *C. magnum* and *Parduzia sp.* Conversely, the inability to resolve ambiguity favors the “frozen” state of the genetic code in other eukaryotes.



(legend on next page)

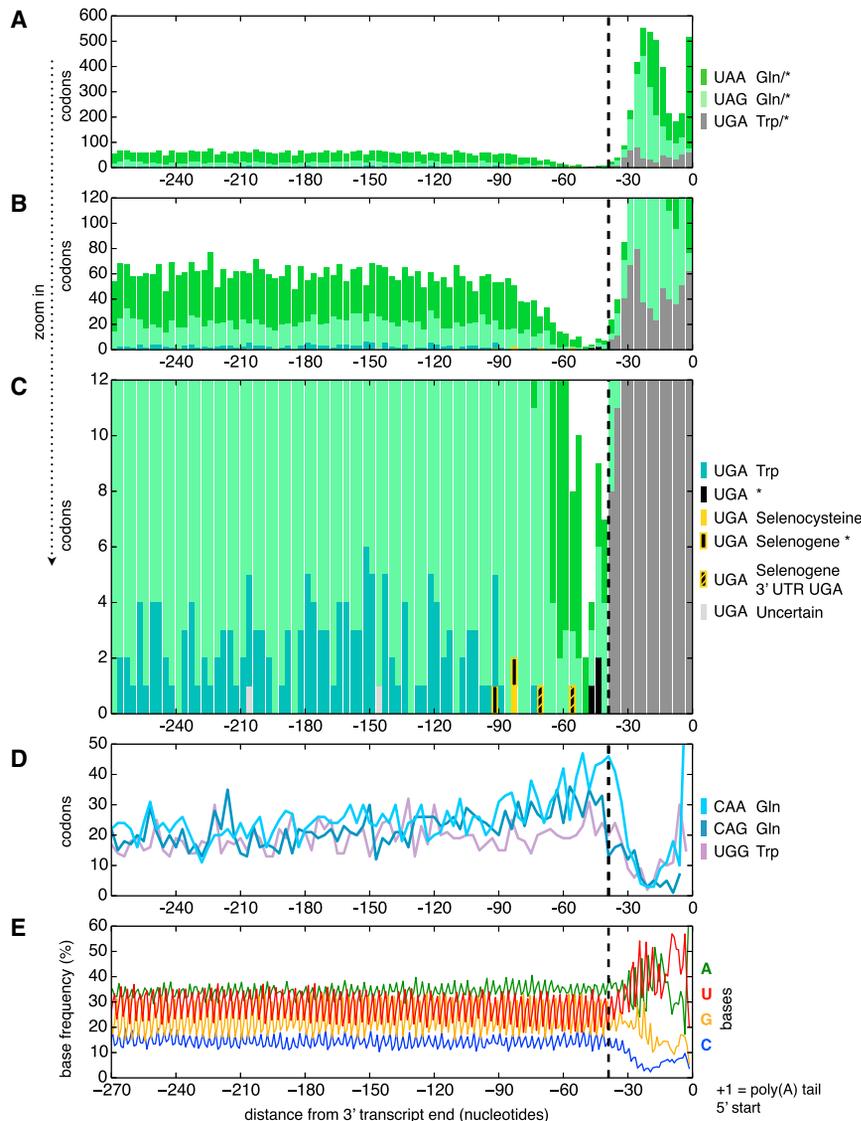


Figure 6. Terminal “Stop” Codon Decline Close to *C. magnum* Stops

Stacked bar graphs of “stop” codon counts are for the transcript regions upstream of poly(A) tails (position 0). Transcript ends include 0, 1, or 2 nucleotides of the poly(A) tail to complete the final “codon.” 3’ UTRs occur in the region to the right of the right-most dashed vertical line. Codons counted are those in the 1672 poly(A)-tailed single gene, single isoform Trinity assembled transcripts.

(A–C) The top three subgraphs are drawn in decreasing order of ordinate limits. Vertical line at –39 nt indicates approximately where most downstream “stops” are either stop codons or “codons” in 3’ UTRs. Codons whose sense/stop states have not been determined are indicated by “amino acid/*.” Transcripts with UGA codons upstream of –39 nt were visually classified based on BLASTX searches. Upstream of –39 nt, UGA codons predominantly code for tryptophan; downstream of –39 nt, UGA codons are predominantly stops or codons in 3’ UTRs downstream of primary stops (both indicated by gray bars). In the genetic codes of *C. magnum* and *Parduczia sp.* UGA is a codon triality (codon duality is reviewed in Atkins and Baranov, 2007), because in addition to being interpreted as a tryptophan codon and a stop codon, it also serves as a selenocysteine codon in the context of SECIS elements. Pale gray bars correspond to a transcript with an uncertain C-terminal, as judged by BLAST.

(D) Standard glutamine and tryptophan sense codon counts.

(E) Base frequencies are stable in the region of “stop” codon decline (~–90 to –42 bases upstream of poly-As).

See also Figures S5 and S6.

The codons in *C. magnum* and *Parduczia sp.* that are recognized either by tRNAs or eRF1 represent precisely the type of intermediate states with multiple meanings originally proposed to occur in the hypothesis of genetic code evolution through ambiguous translational intermediates (Schultz and Yarus, 1994). We furthermore propose that the evolution of very short, AU-rich 3’ UTRs and termination facilitated by poly(A) proximity have enabled codon reassignment, as translational ambiguity due to

the acquisition of stop cognate tRNAs could be suppressed at stops. In light of the ambiguous genetic codes presented here, it is worth reconsidering the idea that the standard genetic code is “one in a million” and is optimized to minimize the effects of errors arising from mutations (Freeland and Hurst, 1998) (although contested [Koonin and Novozhilov, 2009]). Naturally, organisms with only one or two stop codons due to reassignments are more robust to sense premature stop codon mutations than those with the standard genetic code. Given that, other than in the vicinity of transcript ends, “stop” codons are translated by default, the genetic codes of *C. magnum* and *Parduczia sp.* may confer very high resistance

Figure 5. Extremely Short and Nonexistent 3’ UTRs in Heterotrichs

(A) Ciliate 3’ UTR length distributions (lengths exclude the stop codon and poly(A) tail) for representatives of the ciliate genetic codes in Figure 1. (B) Length distribution of *C. magnum* 3’ UTRs. Lengths are from the putative primary stop in the 60 nt window upstream of poly(A) sites and exclude the stop and poly(A) tail lengths. (C) A 3’ UTR-less gene (synaptobrevin homolog). Poly(A) tail-ending reads mapped to the genomic region encoding this gene are shown, and no other reads extend beyond the poly(A) addition site. CDS, coding sequence (Data S1AI and S1AJ). (D) RPFs mapped to a transcript of the gene in (C) (Data S1AK and S1AL). See also Figure S6.

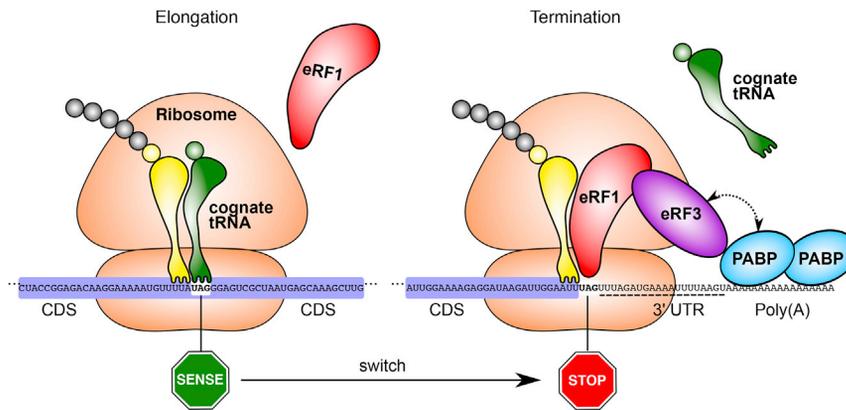


Figure 7. Model for Distinguishing Stops from Sense “Stops”

Representative regions from the same transcript (MMETSP0210: CAMNT_0008285195), with translation through a UAG sense codon and termination at a UAG stop codon (codon state verified by ribo-seq). CDS, coding sequence; 3' UTR, 3' UTR; eRF1, eukaryotic release factor 1; eRF3, eukaryotic release factor 3; PABP, poly(A)-binding protein; standard amino acids are indicated by circles. Putative interaction between eRF3 and PABPs, as inferred from experimental evidence in yeast (Cosson et al., 2002), is indicated by a dotted bidirectional arrow. Ribosome position and the protected mRNA span are illustrated as inferred from *C. magnum* RPFs and from estimates of other eukaryotic ribosomes (Chung et al., 2015).

to substitutions that would cause premature translation termination in the standard genetic code. A potential drawback of such robustness is that large insertions at 3' transcript ends may expose stops that were previously translated. However, large insertions likely occur much less often than substitutions, and the strong purifying selection governing non-protein-coding regions in the heterotrich and karyorelict genomes will inhibit progressive transcript end lengthening.

In summary, we propose that ambiguous ciliate genetic codes are resolved by context-dependent translation termination, and the reason why ciliates possess such diverse genetic codes is that their ancestors had the ability to thrive for extended periods with ambiguous genetic codes, as epitomized by *C. magnum*. Together with the other variant genetic codes, these codes show that the standard nuclear genetic code is not necessarily an evolutionary dead end and that genetic codes can occasionally be observed in a state of flux. As highlighted here, the ambiguous genetic codes of *C. magnum* and *Parduczia sp.* also have ramifications for our understanding of the suppression of translational readthrough, as well as how nonsense-mediated decay (NMD) and selenocysteine translation operate (conserved proteins from both of these pathways are present in ciliates with ambiguous genetic codes; see e.g., Figure S2E). To facilitate future investigations concerning how sense is distinguished from stop and related questions about codon disambiguation, we have made a draft *C. magnum* macronuclear genome available under the accession number European Nucleotide Archive: GCA_001499635.1.

EXPERIMENTAL PROCEDURES

See the [Supplemental Experimental Procedures](#) for additional detailed protocols.

Transcriptomes Analyzed

Transcriptomes for *C. magnum* (MMETSP0210), *Parduczia sp.* (MMETSP1317), and other eukaryotes assembled as part of MMETSP (Gentekaki et al., 2014; Keeling et al., 2014) were used to identify genetic codes and analyze stop codon usage. We also predicted genetic codes after de novo assembling the transcriptomes of two peritrichous ciliates: *Campanella umbellaria* and *Carchesium polypinum* (NCBI short read archive: SRR1768423 and SRR1768437, respectively; data from a recent phylogenomic study) (Feng et al., 2015) with Trinity (Grabher et al., 2011) (default parameters, version: trinitymaseq_r20140717).

Prediction of Alternative Stop Codon Reassignments

To predict codon reassignments, we simplified and refined the key steps of a method developed for such prediction (Dutilh et al., 2011), which identifies codons aligned to conserved amino acids in hidden Markov models inferred from multiple sequence alignments. Dutilh et al. (2011) may be consulted for a graphical outline and more details of the method. This method builds upon and advances the classical method of inspecting conserved positions in multiple sequence alignments of homologous protein sequences to infer codon reassignments. First, we generated a database of peptide sequences by translating nucleotide sequences in all six frames with the standard genetic code, recording standard stop codons as “X” (any amino acid). Next, we used HMMER 3.1b (<http://hmmmer.org>) to search and align the hidden Markov models from the Pfam-A protein domain database (release 27) (Finn et al., 2014) against the translated sequences. Using a custom Python script, the alignment outputs were filtered at a conditional e-value threshold $< 1e-10$. We then simultaneously scanned through the Pfam consensus, aligned database match and its underlying coding sequence, recording the codon and consensus amino acid for well-conserved amino acids at $\geq 50\%$ frequency in columns of the multiple sequence alignment used to build the Pfam model. From the resultant counts of aligned amino acid/codon pairs (m_{ij} ; $i = 1..64$ codons, $j = 1..20$ amino acids) a 20 amino acid by 64 codon matrix, M , was created, with each entry scaled by the sum of the counts for each amino acid (i.e., $M = m_{ij} / \sum_i m_{ij}$). This matrix was used to generate a sequence logo with WebLogo 3.3 (Crooks et al., 2004) (command line switches: “--scale-width no -c chemistry -U probability -A protein”). Note that the lower frequency amino acids shown in the genetic code logos generated by this procedure typically reflect the underlying codon mutational space, but may also be subject to noise, and the focus for codon reassignment prediction should be on the highest frequency amino acid. Genetic code sequence logos for all MMETSP transcriptomes are provided as [Data S1A](#) (ciliates) and [Data S1B](#) (nonciliates). See [Table S1](#) for a summary of the ciliate genetic code predictions. An explanation of stop codon identification is provided in the [Supplemental Experimental Procedures](#).

Ribosome Profiling

Illumina’s TruSeq Ribo Profile (Mammalian) kit was used for ribosome profiling. A total of 32,000 *C. magnum* cells (strain COL2) were isolated, gently pelleted at $280 \times g$ for 2 min in 100 ml pear-shaped centrifuge tubes, then washed in clean saline solution and centrifuged again at $280 \times g$ for 2 min to remove excess algae. The cleaned *C. magnum* cell pellet was incubated in saline solution with 0.1 mg/ml cycloheximide for 1 min. Cells were rinsed with 10 ml PBS, 0.1 mg/ml cycloheximide, pelleted at $280 \times g$, and excess liquid was removed with a micropipette. Pelleted cells were lysed in TruSeq Ribo Profile lysis buffer using a syringe with a 21G needle. The TruSeq Ribo Profile protocol was followed for the remaining ribosome profiling steps. Three concentrations of TruSeq Ribo Profile Nuclease (3 U, 10 U, and 30 U) were used to generate ribosome-protected fragments (RPFs), which were purified with MicroSpin S-400 columns. Ribo-Zero Gold Yeast rRNA depletion was performed on purified RPFs. DNA libraries isolated from 15 (10 U) or 17 (3 U, 10 U) cycle PCRs

were multiplexed and sequenced on one lane of a HiSeq 2500 sequencer by FASTER SA (Switzerland). Ribosome profiling data are available from the European Nucleotide Archive: ERS1066482–ERS1066484. After adaptor trimming, reads were mapped to 1,672 poly(A)-tailed, translation frame inferred Trinity assembled transcripts (see the [Supplemental Experimental Procedures](#)) with STAR (parameters: “–alignIntronMin 12 –alignIntronMax 25”). Reads with 0 or 1 mismatches to the transcripts were used in ribo-seq analyses.

ACCESSION NUMBERS

The accession number for the draft of the *C. magnum* macronuclear genome reported in this paper is European Nucleotide Archive: GCA_001499635.1.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, one table, and supplemental data and can be found with this article online at <http://dx.doi.org/10.1016/j.cell.2016.06.020>.

AUTHOR CONTRIBUTIONS

E.C.S. performed the computational analyses and assisted in laboratory experiments. V.S. cultured *C. magnum*, isolated nucleic acids and proteins, and performed laboratory experiments searching for tRNAs. E.C.S. and V.S. performed ribosome profiling. M.N. supervised the project. E.C.S. drafted the manuscript with input from V.S., G.P., and M.N.

ACKNOWLEDGMENTS

We thank Letizia Modeo for collecting *C. magnum*, Vittorio Boscaro for the original *C. magnum* RNA isolation, Sophie Braga-Lagache and Manfred Heller from the Mass Spectrometry and Proteomics Laboratory at the Children’s University Hospital in Bern for mass spectrometry support, Deis Haxholli for initial genetic code inspections and the M.N. lab members for support and discussion. This research was supported by grants from the European Research Council (ERC) (EPIGENOME) and National Center of Competence in Research (NCCR) RNA and Disease to M.N., and the European COST Action BM1102. Cluster computing was performed at the Vital-IT Center for High-Performance Computing (<http://www.vital-it.ch>) of the Swiss Institute of Bioinformatics.

Received: August 11, 2015

Revised: April 19, 2016

Accepted: June 6, 2016

Published: July 14, 2016

REFERENCES

Aeschlimann, S.H., Jönsson, F., Postberg, J., Stover, N.A., Petera, R.L., Lipps, H.J., Nowacki, M., and Swart, E.C. (2014). The draft assembly of the radically organized *Stylonychia lemnae* macronuclear genome. *Genome Biol. Evol.* **6**, 1707–1723.

Amrani, N., Ganesan, R., Kervestin, S., Mangus, D.A., Ghosh, S., and Jacobson, A. (2004). A faux 3′-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay. *Nature* **432**, 112–118.

Aoki, K., Yano, K., Suzuki, A., Kawamura, S., Sakurai, N., Suda, K., Kurabayashi, A., Suzuki, T., Tsugane, T., Watanabe, M., et al. (2010). Large-scale analysis of full-length cDNAs from the tomato (*Solanum lycopersicum*) cultivar Micro-Tom, a reference system for the Solanaceae genomics. *BMC Genomics* **11**, 210.

Atkins, J.F., and Baranov, P.V. (2007). Translation: duality in the genetic code. *Nature* **448**, 1004–1005.

Aury, J.M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Ségurens, B., Daubin, V., Anthonard, V., Aiach, N., et al. (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444**, 171–178.

Baejen, C., Torkler, P., Gressel, S., Essig, K., Söding, J., and Cramer, P. (2014). Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition. *Mol. Cell* **55**, 745–757.

Blanchet, S., Cornu, D., Argentini, M., and Namy, O. (2014). New insights into the incorporation of natural suppressor tRNAs at stop codons in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **42**, 10061–10072.

Caron, F., and Meyer, E. (1985). Does *Paramecium primaurelia* use a different genetic code in its macronucleus? *Nature* **314**, 185–188.

Chen, X., Bracht, J.R., Goldman, A.D., Dolzhenko, E., Clay, D.M., Swart, E.C., Perlman, D.H., Doak, T.G., Stuart, A., Amemiya, C.T., et al. (2014). The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. *Cell* **158**, 1187–1198.

Chung, B.Y., Hardcastle, T.J., Jones, J.D., Irigoyen, N., Firth, A.E., Baulcombe, D.C., and Brierley, I. (2015). The use of duplex-specific nuclease in ribosome profiling and a user-friendly software package for Ribo-seq data analysis. *RNA* **21**, 1731–1745.

Cosson, B., Couturier, A., Chabelskaya, S., Kiktev, D., Inge-Vechtomov, S., Philippe, M., and Zhouravleva, G. (2002). Poly(A)-binding protein acts in translation termination via eukaryotic release factor 3 interaction and does not influence [PSI(+)] propagation. *Mol. Cell. Biol.* **22**, 3301–3315.

Coyne, R.S., Hannick, L., Shanmugam, D., Hostetler, J.B., Bami, D., Joardar, V.S., Johnson, J., Radune, D., Singh, I., Badger, J.H., et al. (2011). 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.* **12**, R100.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190.

Dunn, J.G., Foo, C.K., Belletier, N.G., Gavis, E.R., and Weissman, J.S. (2013). Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila melanogaster*. *eLife* **2**, e01179.

Dutilh, B.E., Jurgelenaite, R., Szklarczyk, R., van Hijum, S.A., Harhangi, H.R., Schmid, M., de Wild, B., François, K.J., Stunnenberg, H.G., Strous, M., et al. (2011). FACIL: Fast and Accurate Genetic Code Inference and Logo. *Bioinformatics* **27**, 1929–1933.

Edgcomb, V.P., Leadbetter, E.R., Bourland, W., Beaudoin, D., and Bernhard, J.M. (2011). Structured multiple endosymbiosis of bacteria and archaea in a ciliate from marine sulfidic sediments: a survival mechanism in low oxygen, sulfidic sediments? *Front Microbiol* **2**, 55.

Eisen, J.A., Coyne, R.S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J.R., Badger, J.H., Ren, Q., Amedeo, P., Jones, K.M., et al. (2006). Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol.* **4**, e286.

Eliseev, B., Kryuchkova, P., Alkalaeva, E., and Frolova, L. (2011). A single amino acid change of translation termination factor eRF1 switches between bipotent and omnipotent stop-codon specificity. *Nucleic Acids Res.* **39**, 599–608.

Feng, J.M., Jiang, C.Q., Warren, A., Tian, M., Cheng, J., Liu, G.L., Xiong, J., and Miao, W. (2015). Phylogenomic analyses reveal subclass Scuticociliatia as the sister group of subclass Hymenostomatia within class Oligohymenophorea. *Mol. Phylogenet. Evol.* **90**, 104–111.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230.

Freeland, S.J., and Hurst, L.D. (1998). The genetic code is one in a million. *J. Mol. Evol.* **47**, 238–248.

Frischmeyer, P.A., van Hoof, A., O’Donnell, K., Guerrero, A.L., Parker, R., and Dietz, H.C. (2002). An mRNA surveillance mechanism that eliminates transcripts lacking termination codons. *Science* **295**, 2258–2261.

Gentekaki, E., Kolisko, M., Boscaro, V., Bright, K.J., Dini, F., Di Giuseppe, G., Gong, Y., Miceli, C., Modeo, L., Molestina, R.E., et al. (2014). Large-scale phylogenomic analysis reveals the phylogenetic position of the problematic taxon *Protocruzia* and unravels the deep phylogenetic affinities of the ciliate lineages. *Mol. Phylogenet. Evol.* **78**, 36–42.

- Gomes, A.C., Miranda, I., Silva, R.M., Moura, G.R., Thomas, B., Akoulitchev, A., and Santos, M.A. (2007). A genetic code alteration generates a proteome of high diversity in the human pathogen *Candida albicans*. *Genome Biol.* **8**, R206.
- Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652.
- Hanyu, N., Kuchino, Y., Nishimura, S., and Beier, H. (1986). Dramatic events in ciliate evolution: alteration of UAA and UAG termination codons to glutamine codons due to anticodon mutations in two *Tetrahymena* tRNAs. *EMBO J.* **5**, 1307–1311.
- Harrell, L., Melcher, U., and Atkins, J.F. (2002). Predominance of six different hexanucleotide recoding signals 3' of read-through stop codons. *Nucleic Acids Res.* **30**, 2011–2017.
- Helftenbein, E. (1985). Nucleotide sequence of a macronuclear DNA molecule coding for alpha-tubulin from the ciliate *Stylonychia lemnae*. Special codon usage: TAA is not a translation termination codon. *Nucleic Acids Res.* **13**, 415–433.
- Hirsh, D. (1971). Tryptophan transfer RNA as the UGA suppressor. *J. Mol. Biol.* **58**, 439–458.
- Hirsh, D., and Gold, L. (1971). Translation of the UGA triplet in vitro by tryptophan transfer RNA's. *J. Mol. Biol.* **58**, 459–468.
- Horowitz, S., and Gorovsky, M.A. (1985). An unusual genetic code in nuclear genes of *Tetrahymena*. *Proc. Natl. Acad. Sci. USA* **82**, 2452–2455.
- Jan, C.H., Friedman, R.C., Ruby, J.G., and Bartel, D.P. (2011). Formation, regulation and evolution of *Caenorhabditis elegans* 3'UTRs. *Nature* **469**, 97–101.
- Johnson, M.D., Tengs, T., Oldach, D.W., Delwiche, C.F., and Stoecker, D.K. (2004). Highly divergent SSU rRNA genes found in the marine ciliates *Myrionecta rubra* and *Mesodinium pulex*. *Protist* **155**, 347–359.
- Jungreis, I., Lin, M.F., Spokony, R., Chan, C.S., Negre, N., Victorsen, A., White, K.P., and Kellis, M. (2011). Evidence of abundant stop codon readthrough in *Drosophila* and other metazoa. *Genome Res.* **21**, 2096–2113.
- Keeling, P.J., and Doolittle, W.F. (1996). A non-canonical genetic code in an early diverging eukaryotic lineage. *EMBO J.* **15**, 2285–2290.
- Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., et al. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.* **12**, e1001889.
- Kini, H.K., Silverman, I.M., Ji, X., Gregory, B.D., and Liebhaber, S.A. (2016). Cytoplasmic poly(A) binding protein-1 binds to genomically encoded sequences within mammalian mRNAs. *RNA* **22**, 61–74.
- Knight, R.D., Freeland, S.J., and Landweber, L.F. (2001). Rewiring the keyboard: evolvability of the genetic code. *Nat. Rev. Genet.* **2**, 49–58.
- Koonin, E.V., and Novozhilov, A.S. (2009). Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* **61**, 99–111.
- Kuchino, Y., Hanyu, N., Tashiro, F., and Nishimura, S. (1985). *Tetrahymena thermophila* glutamine tRNA and its gene that corresponds to UAA termination codon. *Proc. Natl. Acad. Sci. USA* **82**, 4758–4762.
- Lekomtsev, S., Kolosov, P., Bidou, L., Frolova, L., Rousset, J.P., and Kisselev, L. (2007). Different modes of stop codon restriction by the *Stylonychia* and *Paramecium* eRF1 translation termination factors. *Proc. Natl. Acad. Sci. USA* **104**, 10824–10829.
- Lemke, E.A. (2014). The exploding genetic code. *ChemBioChem* **15**, 1691–1694.
- Lorenz, R., Bernhart, S.H., Höner Zu Siederdisen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26.
- Lozupone, C.A., Knight, R.D., and Landweber, L.F. (2001). The molecular basis of nuclear genetic code change in ciliates. *Curr. Biol.* **11**, 65–74.
- McCaughan, K.K., Brown, C.M., Dalphin, M.E., Berry, M.J., and Tate, W.P. (1995). Translational termination efficiency in mammals is influenced by the base following the stop codon. *Proc. Natl. Acad. Sci. USA* **92**, 5431–5435.
- Mühlhausen, S., Findeisen, P., Plessmann, U., Urlaub, H., and Kollmar, M. (2016). A novel nuclear genetic code alteration in yeasts and the evolution of codon reassignment in eukaryotes. *Genome Res.* Published online May 6, 2016. <http://dx.doi.org/10.1101/gr.200931.115>.
- Nasim, M.T., Jaenecke, S., Belduz, A., Kollmus, H., Flohé, L., and McCarthy, J.E. (2000). Eukaryotic selenocysteine incorporation follows a nonprocessive mechanism that competes with translational termination. *J. Biol. Chem.* **275**, 14846–14852.
- Nowacki, M., Higgins, B.P., Maquilan, G.M., Swart, E.C., Doak, T.G., and Landweber, L.F. (2009). A functional role for transposases in a large eukaryotic genome. *Science* **324**, 935–938.
- Osawa, S., and Jukes, T.H. (1989). Codon reassignment (codon capture) in evolution. *J. Mol. Evol.* **28**, 271–278.
- Preer, J.R., Jr., Preer, L.B., Rudman, B.M., and Barnett, A.J. (1985). Deviation from the universal code shown by the gene for surface protein 51A in *Paramecium*. *Nature* **314**, 188–190.
- Roy, B., Leszyk, J.D., Mangus, D.A., and Jacobson, A. (2015). Nonsense suppression by near-cognate tRNAs employs alternative base pairing at codon positions 1 and 3. *Proc. Natl. Acad. Sci. USA* **112**, 3038–3043.
- Salas-Marco, J., Fan-Minogue, H., Kallmeyer, A.K., Klobutcher, L.A., Farabaugh, P.J., and Bedwell, D.M. (2006). Distinct paths to stop codon reassignment by the variant-code organisms *Tetrahymena* and *Euplotes*. *Mol. Cell Biol.* **26**, 438–447.
- Sánchez-Silva, R., Villalobo, E., Morin, L., and Torres, A. (2003). A new non-canonical nuclear genetic code: translation of UAA into glutamate. *Curr. Biol.* **13**, 442–447.
- Santos, M.A., and Tuite, M.F. (1995). The CUG codon is decoded in vivo as serine and not leucine in *Candida albicans*. *Nucleic Acids Res.* **23**, 1481–1486.
- Schneider, S.U., and de Groot, E.J. (1991). Sequences of two *rbcS* cDNA clones of *Batophora oerstedii*: structural and evolutionary considerations. *Curr. Genet.* **20**, 173–175.
- Schneider, S.U., Leible, M.B., and Yang, X.P. (1989). Strong homology between the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase of two species of *Acetabularia* and the occurrence of unusual codon usage. *Mol. Gen. Genet.* **218**, 445–452.
- Schultz, D.W., and Yarus, M. (1994). Transfer RNA mutation and the malleability of the genetic code. *J. Mol. Biol.* **235**, 1377–1380.
- Sengupta, S., and Higgs, P.G. (2015). Pathways of Genetic Code Evolution in Ancient and Modern Organisms. *J. Mol. Evol.* **80**, 229–243.
- Sladic, R.T., Lagnado, C.A., Bagley, C.J., and Goodall, G.J. (2004). Human PABP binds AU-rich RNA via RNA-binding domains 3 and 4. *Eur. J. Biochem.* **271**, 450–457.
- Swart, E.C., Bracht, J.R., Magrini, V., Minx, P., Chen, X., Zhou, Y., Khurana, J.S., Goldman, A.D., Nowacki, M., Schotanus, K., et al. (2013). The *Oxytricha trifallax* macronuclear genome: a complex eukaryotic genome with 16,000 tiny chromosomes. *PLoS Biol.* **11**, e1001473.
- Tourancheau, A.B., Tsao, N., Klobutcher, L.A., Pearlman, R.E., and Adoutte, A. (1995). Genetic code deviations in the ciliates: evidence for multiple and independent events. *EMBO J.* **14**, 3262–3267.
- Tuite, M.F., and McLaughlin, C.S. (1982). Endogenous read-through of a UGA termination codon in a *Saccharomyces cerevisiae* cell-free system: evidence for involvement of both a mitochondrial and a nuclear tRNA. *Mol. Cell Biol.* **2**, 490–497.
- Vallabhaneni, H., Fan-Minogue, H., Bedwell, D.M., and Farabaugh, P.J. (2009). Connection between stop codon reassignment and frequent use of shifty stop frameshifting. *RNA* **15**, 889–897.
- Wang, R., Xiong, J., Wang, W., Miao, W., and Liang, A. (2016). High frequency of +1 programmed ribosomal frameshifting in *Euplotes octocarinatus*. *Sci. Rep.* **6**, 21139.
- Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.