Recoding and reassignment in protists

By

Stephen Heaphy

Thesis in fulfilment for the degree of

PhD (Science)

National University of Ireland, Cork

January 2018

Head of School: Rosemary O'Connor

Supervisor: Pavel Baranov

Declaration

This thesis is my own work and has not been submitted for another degree, either at University College Cork or elsewhere.

____

Stephen Heaphy

Acknowledgements

The work that went into this thesis and others like it cannot be done by one individual alone. A debt of gratitude is owed to my colleagues, collaborators, friends and family. I would like to thank Pasha Baranov for giving me the opportunity to pursue this peculiar endeavour through his Science Foundation Ireland awarded grant. The atmosphere he created in the lab supported creativity and discovery that I'm sure would it be hard to find elsewhere. His door was always open for light chat and serious discussion.

I would like to thank all the members of the lapti lab, in particular Patrick O'Connor and Audrey Michel who were immense help throughout.

In addition I would like to thank our colleagues in the recode lab in particular John Atkins who provided much encouragement and advice, along with Gary Loughran and Sinead O'Loughlin.

Table of Contents

Preface

Proteins are synthesised in the cell by a process known as translation, which is comprised of three different stages; initiation, elongation and termination. Translation initiation in eukaryotes involves the recruitment of the ribosome and initiation complex to an mRNA molecule which is facilitated by the 5' cap and poly-A tail, ensuring an initiating methionine tRNA is brought to the P-site of the ribosome corresponding to an initiating ATG codon on the mRNA. The ribosome moves by translocating from codon to codon on the mRNA, an aminoacyl- tRNA synthetase charges each tRNA with one of 20 standard amino acid bringing it to the A-site of the ribosome forming a chain. Finally the process of termination is brought about when a class I release factor recognises one of three signals for termination (i.e. stop codons), triggering hydrolysis and releasing the chain of amino acids, i.e. the polypeptide. The polypeptide then folds into a functional protein. These biomolecules are at the core of life as we know it. Their existence in cellular biology is paramount for many purposes including; catalysing metabolic reactions, acting as transport molecules, structural purposes and for DNA replication and repair to name but a few. The synthesis of proteins is an enormously energy expensive process. Due to their extreme importance and high production costs they are also highly regulated.

The standard genetic code which comprises 61 amino acid specifying sense codons and three stop codons, was long considered to be unchangeable. Since there are only 20 standard amino acids to choose from, the genetic code expresses a level of redundancy, for example both CAA and CAG codons specify glutamine. A change to the 3rd codon position, i.e. the 'wobble position' from A to G or vice versa, in the case of glutamine codons will not change the amino acid being incorporated. The earliest variations to standard decoding were discovered in the 1960's and 1970's in bacteria and viruses, where ribosomal frameshifting and stop codon readthrough were identified. It was hypothesised that organisms with small genomes could utilize such 'recoding' events to maximise their coding potential. In more recent decades various other decoding anomalies were uncovered such as two additional non-standard amino acids, (selenocysteine and pyrrolysine), translational bypassing and stop codon reassignment. These events are rare in occurrence and are often regulatory in function. For example, the gene orinithine decarboxylase antizyme (*OAZ*) in eukarotes and release factor 2 in bacteria, both require a frameshift during mRNA translation to synthesise a full length protein. Recent large scale sequencing projects have provided a wealth of new examples of frameshifting in bacteria and other organisms and provide the basis for new repositories such as the Recode Database (Bekaert et al. 2010). Additionally, new techniques make identifying recode events easier.

One such recently developed technique is ribosome profiling (Ribo-seq), which captures the ~30nt protected mRNA 'footprint' of translating ribosomes, where it is then isolated and sequenced. This concept was hypothesized decades before its subsequent development in the lab of J. Weissman at the University of California, Santa Cruz (UCSF). The first publically available Ribo-seq dataset was published in April 2009, "Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling" (Ingolia et al. 2009). By applying a specific 15nt offset to the 30nt footprint it is possible to identify which codon is in the ribosome A-site of elongating ribosomes. In combination with transcript data (RNA-seq), the level of translation efficiency could be determined (Ribo-seq/RNA-seq). Ribosome profiling provided a novel way of quantifying cellular translation.

The practical uses and applications of this new technique were not lost on the scientific community and our lab was one of the first to realise its potential. I commenced my PhD in the summer of 2013 and contributed to the development of a genome browser, Genome Wide Information on Protein Synthesis (GWIPS-viz), which provides analysis and visualization of Ribo-seq data (https://gwips.ucc.ie/). I was involved in developing a pipeline for processing Ribo-seq and RNA-seq data and generating the relevant tracks for each genome. Initially we made Ribo-seq data available from published studies for ten different genomes plus the corresponding RNA-seq. It became apparent that this tool was very useful for identifying recoding events, as large peaks in the profiling data corresponded to pausing ribosomes were observed at frameshift sites of; *dnaX* from *E.coli*, the *gp* gene in *Bacteriophage lambda* and antizyme 1 (*OAZ1*) in human and mouse genomes. It also provided evidence for stop codon readthrough in different organisms and contributed to identifying a novel regulatory mechanism in the *AMD1* gene in humans. Currently there are 24 genomes including the protists *Plasmodium falciparum* and *Trypanosoma brucei* available on GWIPS-viz.

Protists, which form an informal grouping of eukaryotes are neither animals, plants nor fungi. They include certain microscopic algae and red algae, which are often included in the Archaeplastida group of eukaryotes. Protists include other eukaryotic supergroups including; Excavata, Amoebozoa, Hacrobia, Apusozoa and certain Opisthokonta. The term protist also includes the very well-studied SAR supergroup, which consists of the superphyla: Stramenopiles, Alveolata and Rhizaria. Stramenopiles are also known as heterokonts, they include a major group of algae commonly known as diatoms, which are routinely used in environmental monitoring. The alveolates include numerous phyla such as: ciliates, dinoflagellates and apicomplexa. The latter phylum includes *Plasmodium,* the causative agent of malaria.

The ciliates are of particular interest. Named for their cilia, they have provided the scientific community with an array of exceptional insights into cell biology. Ciliates contain two nuclei; a micronucleus active during reproduction and a macronucleus which is active during cell growth. Previous analyses of the species *Tetrahymena thermophila* which led to the discovery of telomeres and ribozymes, resulted in the awarding of two Nobel prizes. Other discoveries from this species include; the identification of self-splicing RNA and the reassignment of stop codons TAA and TAG to incorporate the amino acid glutamine while TGA functions as the only stop codon.

Samples of transcripts taken from the ciliate of genus *Euplotes*, demonstrated an unusually high level of ribosomal frameshifting to produce full length protein products (Aigner et al. 2000; Tan et al. 2001; Wang et al. 2002). The frameshift site identified as a conserved lysine (AAA) codon directly 5' of a TAA or TAG stop codon (5'-AAA_TAR-3'), was discovered to induce a +1 shift in reading frame, while it was believed the efficiency of frameshifting might be close to 100% (Klobutcher and Farabaugh 2002). The TGA stop codon is reassigned to cysteine in this species. The original estimate was put at ~10% of genes requiring a +1 frameshift. My supervisor, Pasha Baranov and I began a collaboration with the lab of Vadim Gladyshev to quantify the level of frameshifting in *Euplotes*. When I began working on this project the genomes of two closely related species *E. crassus* and *E. focardii* had already been sequenced and assembled. I was responsible for quantifying the number of chromosomes present in each species, based on telomere repeat sequences and carried out intron analysis. I obtained RNA-seq and Ribo-seq data for *E. crassus* from our collaborator. I assembled the *E. crassus* transcriptome and mapped the corresponding Ribo-seq data to the new assembly. I performed all corresponding Ribo-seq analysis and all pairwise transcriptome analysis on both species. I identified all frameshift sites and provided candidates for mass spectrometry investigation.

While analysing the *Euplotes* data I came across The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (Keeling et al. 2014), which would dramatically influence the direction of my research over the next two years. The MMETSP provides over 650 assembled transcriptomes from the major protist lineages and are publically available. I employed a wide range of different bioinformatics tools to mine the MMETSP datasets. Initially, I began looking for similar cases of *Euplotes* like frameshifting in the ciliate *Blepharisma* where TGA is reassigned to tryptophan, the results of which proved inconclusive. I then proceeded to look for variant genetic codes as many ciliates are known to reassign stop codons to sense codons. I assembled ciliate transcriptomes from the MMETSP and from other studies, then carried out large scale analyses on the assemblies. I was looking for reassigned stop codons and the amino acids they incorporate. From here I identified two new genetic codes, in *Condylostoma* and *Mesodinium/Myrionecta* and redefined a number of others. I was very fortunate to recruit the skills of Marco Mariotti who analysed the termination sites of *Condylostoma* genes. The true value of the MMETSP quickly became apparent.

In light of discoveries I made in *Euplotes* and other ciliates I began looking for orthologs of the protein ornithine decarboxylase antizyme (OAZ) in other protists. This gene requires a ribosomal frameshift during mRNA translation, which has regulatory functions. Limited information on this highly important and conserved protein was known about its existence in protists. It was previously identified in the ciliate *Tetrahymena thermophile,* but as a single open reading frame protein without the required frameshift. I developed a bioinformatics pipeline to mine MMETSP transcriptomes for antizyme signals and identified the protein in new protist phyla, including a unique mechanism of translation in dinoflagellates, experimentally validated by Gary Loughran.

When the MMETSP was published it provided a wealth of new data however, such large datasets bring new challenges, especially around the annotation of new organisms with high quantities of unknown sequences. The importance of correctly annotating new data sets will become a challenge to database curators, as more large scale sequencing projects become available.

In this thesis I provide examples of discoveries I made during my time as a PhD student. It provides an insight into the translation machinery of protists, the novel mechanisms by which certain groups have evolved and to the plasticity of genetic decoding. Two of the chapters contained in this thesis are published while another is currently being prepared for journal submission. To show my contribution I highlight in red each figure I generated or a panel of a figure i.e. **Figure 1** or Figure 3 **(c)**. Here I list the publications with a short description of my findings:

I identified frameshifting at an unprecedented rate of 22% of genes in *E. crassus*, where the codon upstream of a stop directs the ribosome to either the +1 frame or novel +2 frame. I also observed the requirement of multiple frameshifts per transcript (>3). Frameshifting is the standard function of stop codons when identified in the coding sequence, whereas termination only occurs in the context of the poly-A tail.

Lobanov AV*, **Heaphy SM***, Turanov AA, Gerashchenko MV, Pucciarelli S, Devaraj RR, Xie F, Petyuk VA, Smith RD, Klobutcher LA, Atkins JF, Miceli C, Hatfield DL, Baranov PV, Gladyshev VN. (2017) Position dependent termination and widespread obligatory frameshifting in Euplotes translation. *Nature Structural & Molecular Biology* 24:61-68

* Equal contribution

Here I discovered a novel mechanism of stop codon readthrough to regulate antizyme production in dinoflagellates and also additional examples of +1 frameshifting and single ORF sequences from other protist phyla.

Heaphy SM, Loughran G, Atkins JF, Baranov PV. Diversity of antizyme decoding mechanisms among protists: classical +1 frameshifting, stop codon readthrough and single ORFs. *Manuscript in preparation for journal submission*

I discovered two novel genetic codes in ciliates, one where all three stop codons are reassigned to sense codons in the *Condylostoma magnum* i.e. 64 sense codon genetic code and where TAA and TAG are reassigned to tyrosine in *Mesodinium/Myrionecta*. Both new codes were recognised by NCBI and assigned genetic code numbers 28 and 29 respectively. Additionally I redefined the genetic codes of 9 other ciliates. Like *Euplotes,* termination for *C. magnum* also only occurs in the context of the poly-A tail.

Heaphy SM, Mariotti M, Gladyshev VN, Atkins JF, Baranov PV. (2016) Novel Ciliate Genetic Code Variants Including the Reassignment of All Three Stop Codons to Sense Codons in *Condylostoma magnum*. *Molecular Biology & Evolution.* 33:2885-2889

The publications on frameshifting in *Euplotes* and reassignment of all three stop codons in *Condylostoma* were the subject of a perspective published in the journal Science on 2nd of December 2016 entitled '*When stop makes sense'* written by Boris Zinshteyn and Rachel Green.

I have also contributed to the following publication however, I have not included it in the main body of the thesis but as an appendix.

Michel AM, Fox G, Kiran AM, De Bo C, O'Connor PBF, **Heaphy SM**, Mullan JPA, Donohue CA, Higgins DG and Baranov PV. (2014) GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Research*. 42(Database issue): D859–D864.

Abstract

During mRNA translation the ribosome reads each codon (nucleotide triplet) with a specific meaning. The standard genetic code comprises 61 sense-codons for specifying the 20 standard amino acids during elongation and three anti-sense codons which signal termination. While variations to the standard rules of genetic decoding are widely acknowledged, recent advances in next generation sequencing techniques have provided a wealth of new examples across many species. In this thesis, I provide evidence of novel decoding mechanisms in protists, as identified through bioinformatics analysis. To begin with I analysed the genomes of two ciliate species, *Euplotes* crassus and *E. focardii*. In combination with the analysis of *E. crassus* transcriptome using ribosome profiling, I determined over 1,700 cases of ribosomal frameshifting (22% of genes analysed) in *E. crassus*. I identified 47 codons upstream of a stop signal which directs the ribosome to either the +1 or +2 reading frames. Termination only occurs in the context of the poly-A tail. In addition I analysed the transcriptomes of over 200 diverse protist species for the protein ornithine decarboxylase antizyme, a key negative regulator of cellular polyamine synthesis. The synthesis of this protein usually requires a +1 ribosomal frameshift at the end of the first open reading frame. In this study I identified a novel mechanism of stop codon readthrough to regulate antizyme production in dinoflagellates and single ORF sequences from other protist phyla. Further I analysed transcriptomes of diverse ciliate organisms to characterize stop codon reassignments in their genetic codes. In addition to finding novel stop codon reassignments, I identified an organism, *Condylostoma magnum* where all three stop codons TAA, TAG & TGA have been reassigned to sense codons. All three stop codons are enriched at the expected positions of translation termination sites which occur at a short distance from the 3’ poly-A tail.

Introduction

The standard genetic code, which consists of 61 sense codons for incorporating the 20 standard amino acids during translation elongation and three anti-sense codons (TAA, TAG & TGA), which signal translation termination, was long considered to be immutable while its ancestry was believed to be a ‘frozen accident’ (Crick 1968). Variations to the standard rules of genetic decoding were soon identified, including; frameshifting (Riyasaty and Atkins 1968), stop codon readthrough (Weiner and Weber 1971) and translational bypassing (Huang et al. 1988). Two additional non-standard proteinogenic amino acids were also identified; selenocysteine (Sec) (Chambers et al. 1986) and pyrrolysine (Pyl) (Srinivasan et al. 2002). It is possible that additional amino acids are still waiting to be discovered (Ambrogelly et al. 2007). These recoding events are always in direct competition with standard decoding and as a result the efficiency of their translation varies. These events usually employ signals in the mRNA or nascent peptide which act to stimulate or supress the recoding process and can result in more than one synthesised protein product from a single mRNA transcript (reviewed by Atkins and Gesteland 2010). Variant recoding events are expected to occur in most organisms. Recent bioinformatics analyses of bacterial genomes identified many examples of ribosomal frameshifting (Sharma et al. 2011; Antonov et al. 2013).

In addition to the standard genetic code a number of variant codes exist. According to the National Center for Biotechnology Information (NCBI), there are 23 alternative codes (as of December 2017), the majority of which are mitochondrial in origin. Such variant genetic codes evolve in response to an altered meaning of a codon and these anomalies are known as codon reassignments. The majority of these are stop codon reassignments and are facilitated by changes to the translation machinery, namely; tRNAs, aminoacyl-tRNA synthetase or release factors (reviewed by Baranov et al. 2015). Codon reassignments are pervasive in the genomes where they are found, no additional elements are involved and they are perfectly efficient. A recent study highlighted the immense level of stop codon reassignments in prokaryotes and phages from environmental samples (Ivanova et al. 2014). Stop codon reassignments may have evolved as a defensive mechanism to protect the host against invading pathogens (Li et al. 2013). The differences between recoding and reassignment are highlighted by (Atkins and Baranov 2010).

Stop codons are the most versatile of codons and they are found in the majority of recoding and reassignment cases. Stop codons are extremely rare in the protein coding sequences of transcripts compared to sense codons. The usage of individual stop codons varies considerably also (Korkmaz et al. 2014) and as a result, changing the meaning of a less frequent stop codon may not alter the function of a protein. Stop codon triplets are found in comparatively equal frequencies in the non-coding regions of transcripts to other nucleotide triplets, such that if a readthrough event occurs, the ribosome will find another stop codon shortly downstream resulting in a short C-terminal extension to the protein.

Translation termination relies upon the class I release factor (eRF1 in eukaryotes and aRF1 in archaea; RF1 & RF2 in bacteria) to recognise one of three stop codons (TAA, TAG & TGA) on the mRNA and facilitate release of the polypeptide. During termination the nucleotide sequence downstream of a stop codon can affect the efficiency of termination (Namy et al. 2001) and weak termination at stop codons can facilitate readthrough and frameshifting. The 21st amino acid identified, the non-standard selenocysteine, which is found in all three domains of life, is incorporated into the polypeptide at a TGA codon (Chambers et al. 1986), and facilitated by a downstream secondary structure known as a SECIS element (Berry et al. 1993). The ciliate *Euplotes crassus* can support the incorporation of both cysteine and selenocysteine at a TGA codon (Turanov et al. 2009). It was recently observed that selenocysteine may also be incorporated into TAA and TAG in bacterial cells (Mukai et al. 2016). Pyrrolysine, the 22nd amino acid identified, is incorporated into a TAG stop codon in archaea and methanogenic bacteria (Srinivasan et al. 2002).

Reassignment

During mRNA translation sense codons and stop codons have different functions and machinery to decipher their meaning. Elongating ribosomes add amino acids to a growing chain upon identifying one of 61 sense codons on the mRNA, and this process is facilitated by a tRNA that recognises the codon along with aminoacyl- tRNA synthetase that charges the tRNA with one of 20 standard amino acids. A change to the meaning of a sense codon here would involve altering either one or both of these two proteins, or subsequently if they were lost by the organism. Translation termination is facilitated by the class I release factor which triggers hydrolysis resulting in the release of the nascent peptide. In eukaryotes and archaea RF1 (eRF1 & aRF1) will recognise all three stop codons; TAA, TAG, and TGA. The majority of bacteria and other organelles have two release factors. RF1 recognises stop codons TAA and TAG, while RF2 recognises TAA and TGA (Duarte et al. 2012). For eRF1 the recognition patterns involved with identifying stop codons were found on the N-terminal domain of the protein (Nakamura and Ito 1998; Bertram et al. 2000). Subsequently a variety of cross-linking and mutagenesis studies have identified highly conserved motifs within the N-terminal domain such as TASNIKS, for stop codon recognition, while YxCxxxF and GTS motifs are implicated in purine recognition in the second and third sub-codon positions (Chavatte et al. 2002; Frolova et al. 2002; Ito et al. 2002; Kolosov et al. 2005; Bulygin et al. 2010; Conard et al. 2012). The structural basis for these motifs were determined by cryo-EM (Brown et al. 2015). An additional two residues critical for eRF1 recognition of all three stops has been reported (Blanchet et al. 2015).

Of the variant genetic codes that are currently available on NCBI, nine are eukaryotic nuclear in origin, while seven of these are stop codon reassignments. The two exceptions are changes to the CTG codon; from leucine to serine in the *Candida* genus of yeast (Ohama et al. 1993) and more recently leucine to alanine in another yeast species *Pachysolen tannophilus* (Mühlhausen et al. 2016). The earliest stop codon reassignments were reported in ciliates of genera; *Paramecium* (Caron and Meyer 1985), *Stylonychia* (Helftenbein 1985) and *Tetrahymena* (Horowitz and Gorovsky 1985), where glutamine was found inserted at stop codons TAA and TAG. Subsequently TGA was also reported to be reassigned to cysteine in *Euplotes* (Meyer et al. 1991) and to tryptophan in *Blepharisma* (Liang and Heckmann 1993). Additional ciliate genetic codes in other lineages were also identified (Lozupone et al. 2001; Sánchez-Silva et al. 2003). Apart from ciliates, similar reassignments of TAA & TAG to glutamine were found in two green algae species *Acetabularia* (Schneider et al. 1989) and *Batophora* (Schneider and de Groot 1991) and also in diplomonads (Keeling and Doolittle 1996).

The majority of stop codon reassignments occur in mitochondrial genomes where TGA is reassigned to tryptophan, summarized by (Knight et al. 2001). A variant mitochondrial genetic code also reassigns TAG to leucine (Laforest et al. 1997; Fučíková et al. 2014). Such reassignments may be influenced by their small genome sizes; the mitochondrial genome of *Arabidopsis thaliana* contains a mere 57 genes and is 0.37 megabases (Mb) in size (Unseld et al. 1997). Stop codon reassignments in bacterial genomes may be explained through a mechanism known as codon capture (Osawa and Jukes 1989). Here, the loss of a codon along with its corresponding machinery, such as the stop codon TGA and its release factor (RF2), could support the re-emergence of TGA as a near cognate sense codon i.e. tryptophan. However, this is less likely to explain stop codon reassignment in ciliates, where genomes are much larger. The macronucleus of *Tetrahymena* is 104Mb in size spanning 225 chromosomes (Eisen et al. 2006). It was previously hypothesised that mutated eRF1 sequences which are common in ciliates, do not recognise stop codons as signals for termination which lead to their reassignment (Lozupone et al. 2001). It is not clear how stop codons are recognised by eRF1 in variant genetic codes, and a number of studies attempting to identify the residues involved provide alternative mechanisms (Salas-Marco et al. 2006; Lekomtsev et al. 2007; Vallabhaneni et al. 2009).

In 2016, five additional genetic codes were added to the NCBI list. Three studies reported on the reassignment of all three stop codons, i.e. genetic codes with 64 sense codons. They include, two ciliate species *Parduczia sp.* and *Condylostoma magnum* (Heaphy et al. 2016; Swart et al. 2016) and the trypanosomatid *Blastocrithidia* (Záhonová et al. 2016). These three studies redefine our understanding of protein synthesis; in particular termination, the applications of which may have far reaching implications in the field of synthetic biology (Bezerra et al. 2015).

In eukaryotes where stop codons TAA and TAG were previously reported as reassigned, they both acquired the same new meaning. A recent study, however, reports for the first time that these stop codons can have different functions independent of each other: TAG is reassigned to leucine in a rhizarian species, while TAG to glutamine in the fornicate *Iotanema spirale.* In both cases the corresponding TAA codon functions as a signal for termination, with variant eRF1 residues implicated in the reassignment (Pánek et al. 2017).

Recoding

Frameshifting

One of the earliest conflicts to the 'frozen accident' theory of standard genetic decoding was that of frameshifting. Here ribosomes break from triplet decoding and shift forwards (+) or backwards (-) into a different reading frame at specific sites. These rare events can be exploited for regulatory purposes or for the synthesis of additional proteins and are often programmed to occur in response to signals located within the mRNA, generally referred to as programmed ribosomal frameshifting (PRF). The synthesis of additional proteins from a single mRNA is a common feature of translation in mobile genetic elements and viruses. They utilize frameshifting in order to expand the coding potential of their small genomes. Regulatory frameshifting is found in cellular genes such as ornithine decarboxylase antizyme in eukaryotes and release factor 2 in bacteria (Gesteland and Atkins 1996).

Most recoding events are stimulated by signals in the RNA or the nascent peptide. RNA secondary structures such as stem loops are found both 5' and 3' of the recode sites, however, the vast majority are found downstream (Kim et al. 2014). Other structures include pseudoknots, which generally contain multiple stem loops, and G-quadruplexes, both act as 3' stimulators (Endoh and Sugimoto 2013). Long range stimulators have been reported to act up to 4Kb downstream of the recode site (Wang and Miller 1995). The nascent peptide was shown to act as a stimulator when interacting with the ribosome exit tunnel in fungal antizyme (Yordanova et al. 2015). Secondary structures that stimulate frameshifting must not completely block the ribosome (Tholstrup et al. 2012). Translating ribosomes will unwind pseudoknots and stemloops slowing the rate of translation, thus increasing the time it spends on a slippery sequence allowing for greater opportunity to facilitate a shift in frame (Farabaugh 2000).

The first examples of frameshifting were reported in viruses and were of -1 in mechanism. They included phages; MS2 (Atkins et al. 1979) and T7 (Dunn and Studier 1983). The slippery sequence is usually seven nucleotides taking the form of X.XXY.YYZ (where X, Y & Z can be either nucleotide and '.' defines the codon boundary), tRNAs in the A and P sites shift -1 to a new reading frame XXX.YYY in a process known as tandem slippage (Jacks et al.

1988a). In *E.coli* the gene *dnaX* which encodes the gamma and tau subunits of DNA polymerase III requires a -1 frameshift to produce the gamma subunit (Blinkowa and Walker 1990). The slippery sequence A.AAA.AAG directs the ribosome with an efficiency of 50% into the -1 frame, terminating at a stop codon shortly downstream and producing the gamma subunit (Flower and McHenry 1990). The efficiency of frameshifting is supported by a Shine-Dalgarno sequence upstream and an RNA stemloop structure downstream of the slippery site (Tsuchihashi 1991; Larsen et al. 1994). The Human immunodeficiency virus 1 (HIV-1) also requires a -1 frameshift to produce the gag-pol fusion protein (Jacks et al. 1988b; Parkin et al. 1992). The slippery sequence T.TTT.TTA is crucial to the efficiency of frameshifting making it a target for anti-viral therapy (Biswas et al. 2004).

In contrast to -1 frameshifting, examples of known +1 frameshifting are far less abundant. The mechanisms involved are more diverse and far more difficult to identify. In bacteria, stop codons TAA and TGA are recognised by RF2 and for many species it is regulated by a +1 frameshift during protein synthesis (Craigen and Caskey 1986; Bekaert et al. 2006). In *E. coli* the +1 frameshift occurs at the sequence CTT.TGA. When RF2 levels are high, termination occurs at the TGA stop codon, however, when RF2 levels are low efficient termination decreases inducing a +1 frameshift resulting in more RF2 synthesis and therefore autoregulating its own production (Adamski et al. 1993). In the yeast *Saccharomyces cerevisiae,* +1 ribosomal frameshifting is required for expression of the Ty1 transposon at the frameshift site GCG.AGT.T in response to a rare arginine codon (AGT), with an efficiency of 40% (Belcourt and Farabaugh 1990).

Comparable to RF2 in bacteria, ornithine decarboxylase antizyme (OAZ), which is a key regulator of cellular polyamine levels, requires an unusual and rare +1 frameshift during mRNA translation. Cellular polyamine levels are tightly regulated, as expected from their multiple important roles in cell functioning. The first identified *OAZ* gene, encoding rat antizyme 1 (Matsufuji et al. 1990), has its coding sequence in two different and partially overlapping ORFs with synthesis of functional antizyme involving a programmed ribosomal frameshift event at the end of ORF1 (Miyazaki et al. 1992). The frameshift sequence of mammalian *OAZ1* is a conserved TCC.TGA. There are five known cases of frameshifting genes in humans, three of these are antizymes *OAZ1*, *OAZ2* & *OAZ3* (Ivanov and Atkins 2007).

The efficiency of ribosomes shifting to the +1 reading frame to enter ORF2 and so of the amount of antizyme synthesized, is dependent on cellular polyamine levels (Matsufuji et al. 1995). Elevated polyamine levels enhance frameshifting efficiency, closing an autoregulatory negative feedback loop. Antizyme translation is also dependent on a diverse range of cis-acting stimulatory signals which facilitate frameshifting (Ivanov and Atkins 2007). For mammalian antizymes 1 and 2, a sequence 5’ of the frameshift site has been shown to enhance the efficiency rate (Matsufuji et al. 1996;Ivanov et al. 2000), while the most common 3’ stimulator is an RNA pseudoknot located directly downstream of the frameshift site (Matsufuji et al. 1995). Some eukaryotes most likely have lost antizyme, e.g. plants. However, where antizyme is found, the requirement for a +1 frameshift during mRNA translation is nearly universal, a single known exception is in the ciliate *Tethramymena thermophile* where it is encoded in a single ORF (Ivanov and Atkins 2007).

A number of studies which sequenced genes from the ciliate of genus *Euplotes* reported the on the unusually high level of frameshifting observed. It was proposed the that frameshifting may be more frequent in this organism than any other, with estimates of ~10% of genes and near 100% efficient (Aigner et al. 2000; Klobutcher 2005). The frameshift motif consists of an AAA (lysine) codon directly 5’ of a stop codon, either TAA or TAG (5’-AAA.TAR-3’), while TGA is reassigned to cysteine in *Euplotes*. This motif is known to induce a +1 frameshift during mRNA translation and more than one frameshift per gene has been reported (Karamysheva et al. 2003). It has been proposed that frameshifting in *Euplotes* evolved as a result of TGA reassignment from a termination signal to cysteine, reducing eRF1 recognition of the remaining stop codons (Klobutcher and Farabaugh 2002; Giedroc and Cornish 2009). It has been demonstrated experimentally in a hybrid system that *Euplotes* eRF1 does not efficiently recognise the remaining stop codons (Vallabhaneni et al. 2009). A more recent study searched the entire transcriptome and genome of *E. octocarinatus* and identified a frameshift frequency of >11%, containing alternative shift sites (Wang et al. 2016). This type of high frequency frameshifting has not been reported in any other genus of ciliate.

Readthrough

The three stop codons TAA, TAG & TGA, generally function as signals for translation termination. However, an elongating ribosome can continue through a stop codon in a process known as stop codon readthrough. It has been demonstrated that translation termination is slower and less accurate than translation elongation and as a result, low efficiency stop codon read through may occur in the absence of any stimulatory signals such as RNA secondary structures (Freistroffer et al. 2000; Bertram et al. 2001). When readthrough occurs the ribosome inserts an amino acid at the stop codon, and translation proceeds in the same frame until it encounters another stop codon, resulting in a proportion of proteins with C-terminal extensions (Namy and Rousset 2010). Observations of readthrough in chromosomal genes in *Drosophila* estimated an average C-terminus extension of 35 amino acids (Jungreis et al. 2011). In cases where a near cognate tRNA is inserted at a stop codon, the readthrough efficiency can be greater than 5% of transcripts (Namy et al. 2001). Readthrough of TGA codons, has shown that near cognate tRNAs are in competition with the eukaryotic release factor 1 (eRF1) for pairing with codons and cysteine, tryptophan and arginine have been observed here (Blanchet et al. 2014). The readthrough of stop codons, which extends the length of the protein in the C-terminal, is akin to 'leaky' ATG initiation which affects the length of the protein at the N-terminal, and both types of 'leakiness' contribute to the variety of protein isoforms synthesised in the cell (Touriol et al. 2003).

Like frameshifting, stop codon readthrough can be utilised for regulatory purposes particularly by viruses in order to expand their coding potential by producing functional C-terminal extensions or additional proteins from single mRNAs. One of the earliest examples was the readthrough of the coat protein (CP) stop codon of phage Qβ, where the extended protein is used for viral propagation (Weiner and Weber 1971; Hofstetter et al. 1974). Similar readthrough in tobacco mosaic tobamovirus (TMV) utilizes readthrough to produce the replicase protein (Gesteland et al. 1976; Pelham 1978). Many studies have highlighted the importance of stimulatory factors promoting readthrough, such as translation of the *gag-pol* gene of the murine leukemia virus (MuLV) (Feng et al. 1992; Firth et al. 2011).

Until recently only a few examples of readthrough in eukaryotic organisms were known outside of selenocysteine insertion. In *Drosophila* the kelch *kel* gene, which is involved in viable egg production in females, requires a TGA readthrough to produce a full-length functional protein (Robinson and Cooley 1997) and also the headcase, *hdc* gene, where a hairpin loop was proposed to stimulate readthrough of a TAA codon (Steneberg and Samakovlis 2001). Two cases were reported in the yeast *Saccharomyces cerevisiae* (Namy et al. 2003). Recently the number of readthrough cases in *Drosophila* has increased dramatically based on analyses of 12 genomes (Jungreis et al. 2011) and additional cases confirmed by ribosome profiling (Dunn et al. 2013). The number of cases in humans is also increasing (Eswarappa et al. 2014; Loughran et al. 2014; Schueren et al. 2014) Recent observations have identified readthrough in four mammalian genes with the highly conserved tetranucleotide motif CTAG downstream from a TGA stop codon as an essential component to facilitate readthrough (Loughran et al. 2014).

Selenocysteine

Similar to readthrough, selenocysteine, which is a non-standard amino acid is inserted at a TGA codon facilitated by a 3’ secondary structure referred to as the selenocysteine insertion sequence (SECIS). Selenocysteine, which is an analogue of cysteine contains the element selenium which plays a great number of roles related to human health and development (Rayman 2000). The biochemical properties are also enhanced in the proteins where selenocysteine is inserted (Lee et al. 2000; Zhong et al. 2000). Proteins containing selenocysteine are referred to as selenoproteins and are found in the three domains of life; however, they are not present in fungi or land plants (Lobanov et al. 2009). Protists display a scattered distribution of selenoproteins (Mariotti et al. 2015), while selenoproteins specific to certian protists were identified (Cassago et al. 2006; Lobanov et al. 2006; Novoselov et al. 2007). Some protists have reported high quantities of selenoproteins in their genomes; as many as 60 in the algae *Aureococcus anophagefferens* (Gobler et al. 2011).

For eukaryotes the SECIS element is located in the 3’ untranslated region of the mRNA transcript (Berry et al. 1993). A specific Sec tRNA and a complex of proteins, specifically SBP2 are involved to facilitate incorporation of selenocysteine (Copeland et al. 2000; Fletcher et al. 2001). The human selenoprotein SelP can facilitate incorporating up to ten selenocysteine amino acids due to the same number of TGA codons in the mRNA (Hill et al. 1993). In bacteria the SECIS element is located within the coding sequence, a stem loop 3' of the TGA codon (Heider et al. 1992). As previously noted, selenocysteine was shown to be incorporated at the other stop codons (TAA and TAG), in bacterial cells (Mukai et al. 2016).

Recoding

Position dependent termination and widespread obligatory frameshifting in *Euplotes* translation

This chapter has been published as a research article in Nature Structural and Molecular Biology, 2017 Jan;24(1):61-68. Journal Article.

https://doi.org/10.1038/nsmb.3330

Abstract

The ribosome can change its reading frame during translation in a process known as programmed ribosomal frameshifting. These rare events are supported by complex mRNA signals. However, we found that the ciliates *Euplotes crassus* and *Euplotes focardii* exhibit widespread frameshifting at stop codons. 47 different codons preceding stop signals resulted in either +1 or +2 frameshifts, with the +1 frameshifting at AAA being the most frequent. The frameshifts show unusual plasticity and rapid evolution, and have little influence on translation rates. Proximity of a stop codon to the 3'-mRNA end rather than its occurrence or sequence context appeared to designate termination. Thus, a stop codon is not a sufficient signal for translation termination, and the default function of stop codons in *Euplotes* is frameshifting, whereas termination is specific to certain mRNA positions and likely requires additional factors.

Introduction

There are several known mRNAs where translating ribosomes shift reading frame at specific locations with high efficiency that in very rare cases may even exceed the rate of concurrent standard translation. This phenomenon is known as programmed ribosomal frameshifting and is mostly observed in viruses (Atkins et al. 2016). While programmed ribosomal frameshifting is an omnipresent translation process, it is usually considered as a recoding mechanism. Recoding describes alterations in genetic decoding that take place at specific locations within particular mRNAs and is distinguished from codon reassignment

(Baranov et al. 2015). With an exception of 40% efficient programmed ribosomal frameshifting at a heptanucleotide site in *Saccharomyces cerevisiae* that is used during expression of the Ty1 transposon (Belcourt and Farabaugh 1990), complex stimulatory signals, such as RNA pseudoknots, are required for a high efficiency of programmed ribosomal frameshifting (Giedroc and Cornish 2009).

However, previous analyses of several sequenced genes of the ciliates *Euplotes*, suggested that +1 ribosomal frameshifting may be more common in these organisms, reviewed by (Klobutcher and Farabaugh 2002). All frameshift motifs in *Euplotes* identified until recently consist of an AAA codon followed by a stop codon, either TAA or TAG. It has been hypothesized that frameshifting evolved as a consequence of TGA codon reassignment from stop to cysteine, which weakened release factor recognition of the remaining stop codons, TAA and TAG (Klobutcher and Farabaugh 2002; Vallabhaneni et al. 2009). Furthermore, it has been shown experimentally in a hybrid system that *Euplotes* release factors indeed recognize these stop codons inefficiently (Vallabhaneni et al. 2009).

To understand this unusual case of frameshifting and the molecular mechanisms involved, we sequenced and analyzed the macronuclear genomes of two *Euplotes* species: *E. crassus* and *E. focardii* (Valbonesi and Luporini 1993; Pucciarelli et al. 2009). We also sequenced the transcriptome of *E. crassus* and carried out ribosome profiling and proteomic analyses. The genomic and high-throughput biochemical analyses allowed us to identify and characterize over a thousand frameshift sites. This revealed that ribosomes of the *Euplotes* ciliates are characterized by inability to terminate at stop codons in internal positions of coding sequences and instead frameshift at these signals, whereas termination likely requires additional components in these organisms and occurs only at specific mRNA positions.

Macronuclear genomes of *E. crassus* and *E. focardii* and their transcriptomes.

Similar to other ciliates, *Euplotes* DNA is distributed among its two compartments: the macronucleus, which controls all cell functions during vegetative growth, and the micronucleus, which is needed for reproduction. The macronuclear genome consists of many small chromosomes. The copy number of individual chromosomes in ciliates may range from 100 to 10,000, with an average of 2,000 per macronucleus in *Euplotes* (Baird and Klobutcher 1991; Prescott 1994). These chromosomes are generated from the micronuclei DNA following sexual reproduction (Wong and Landweber 2006). It is the macronuclear DNA that is actively transcribed and is used as a template for mRNA synthesis, and therefore we were interested primarily in the macronuclear genomes.

To understand how *Euplotes* genes are translated, it was beneficial to examine at least two genomes, thereby allowing comparative sequence analysis. Thus, we sequenced macronuclear genomes of two related *Euplotes*. One is *E. crassus*, a sand-dwelling hypotrichous ciliate of the marine intertidal zone. The other is a recently isolated *E. focardii*, which is endemic to the Antarctic (Valbonesi and Luporini 1993). The strain TN1 was obtained from the samples collected in Terra Nova Bay, and its psychrophilic phenotypes (optimal survival and multiplication rates at 4–5 °C) suggest adaptation to the stably cold Antarctic waters (Valbonesi and Luporini 1993). The general properties of their genomes are described in Supplementary Figure 1.

A large number of very short (20-30 nts) introns is a characteristic feature of macronuclear protein coding genes in some ciliates (Ricard et al. 2008; Swart et al. 2013), but accurate prediction of introns is complicated by instances of alternative splicing and non-canonical splice junctions (Vinogradov et al. 2012). Some short introns, if not detected by annotation pipelines, may result in ORF disruption and thus be misinterpreted as frameshift sites. To account for this possibility, we utilized experimentally confirmed rather than predicted mRNA transcripts (Supplementary Fig. 2).

Identification of ribosomal frameshifting using phylogenetics, ribosome profiling and proteomic analyses.

To identify sites of ribosomal frameshifting and estimate its efficiency, we first carried out ribosome profiling (Ribo-seq) in *E. crassus*. Ribosome profiling is based on sequencing of mRNA fragments protected by the translating ribosomes from nuclease digestion (Ingolia et al. 2009). It provides information on ribosome locations and their densities at the whole transcriptome level (Michel and Baranov 2013; Ingolia 2014). Ribosome-protected fragments are expected to occur immediately downstream of stop codons only in cases of efficient stop codon readthrough or ribosomal frameshifting. To discriminate between readthrough and ribosomal frameshifting in -1 or +1 direction we compared the span of Ribo-seq coverage with ORF organization (Fig. 1). In certain cases, where unambiguous discrimination between potential events was difficult, we sought additional information. Using BLAST, we explored which of the potential products is more likely to have closely related homologs. Overall, we identified 1,765 putative frameshift sites spanning 1,326 transcripts from a total of 6,087, with at least 100 Ribo-seq reads per transcript. In a number of transcripts we found more than one site of ribosomal frameshifting (Fig. 1b). In addition to +1 frameshifting, we detected frameshifting into the -1/+2 frame (Fig. 1c). However, we did not find a single example of stop codon readthrough. The sequences of the transcripts were compared to the sequences of genomic contigs to exclude the possibility of identifying frameshifting as a result of misidentification of sequencing errors during RNA-seq analysis (Fig. 1a,d).

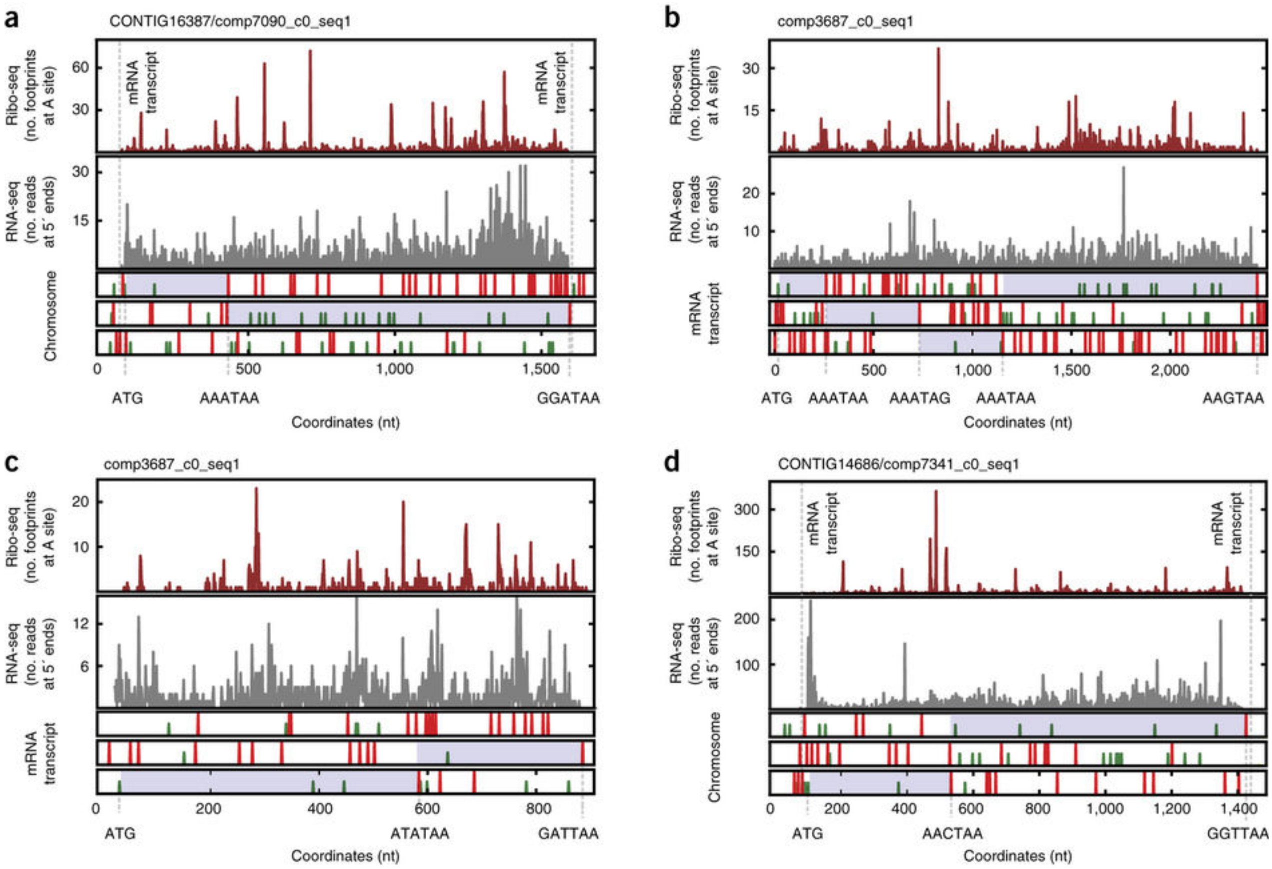


Figure 1. Frequent frameshifting in *Euplotes*. Ribo-seq profiles of individual mRNAs are shown in the upper panels, RNA-seq in the middle panels, and features of reading frames in the lower panels. Start (ATG, green vertical lines) and stop codons (TAA, TAG, red lines) are shown in each of the three reading frames for chromosomes (a, d) and transcripts (b, c). Inferred translated regions are highlighted in blue. ATG codons corresponding to translation initiation sites are indicated beneath each plot. Stop codons (and adjacent upstream codons) where termination or frameshifting occur are also indicated. (a) Example of +1 ribosomal frameshifting at AAA_TAA. (b) Example of mRNA with several ribosomal frameshifting sites. (c) Example of +2 frameshifting at the ATA_TAA. (d) Example of +1 frameshifting at AAC_TAA.

To verify putative sites of frameshifting and determine the associated mechanisms (i.e. direction and identity of amino acids incorporated at frameshift sites), we carried out LC-MS/MS proteomics analyses of soluble *E. crassus* fractions, following trypsin and Glu-C digestions (the latter was used to preserve peptides with internal Lys). We examined if any of these peptides covered two different frames within the same gene and detected 13 such peptides with validated MS/MS spectra (Fig. 2, Supplementary Note 1, Supplementary Note 2). In addition to +1 frameshifting, some peptides were the products of +2 ribosomal frameshifting, consistent with our observation of ribosomal frameshifting into the -1/+2 frame based on Ribo-seq data.

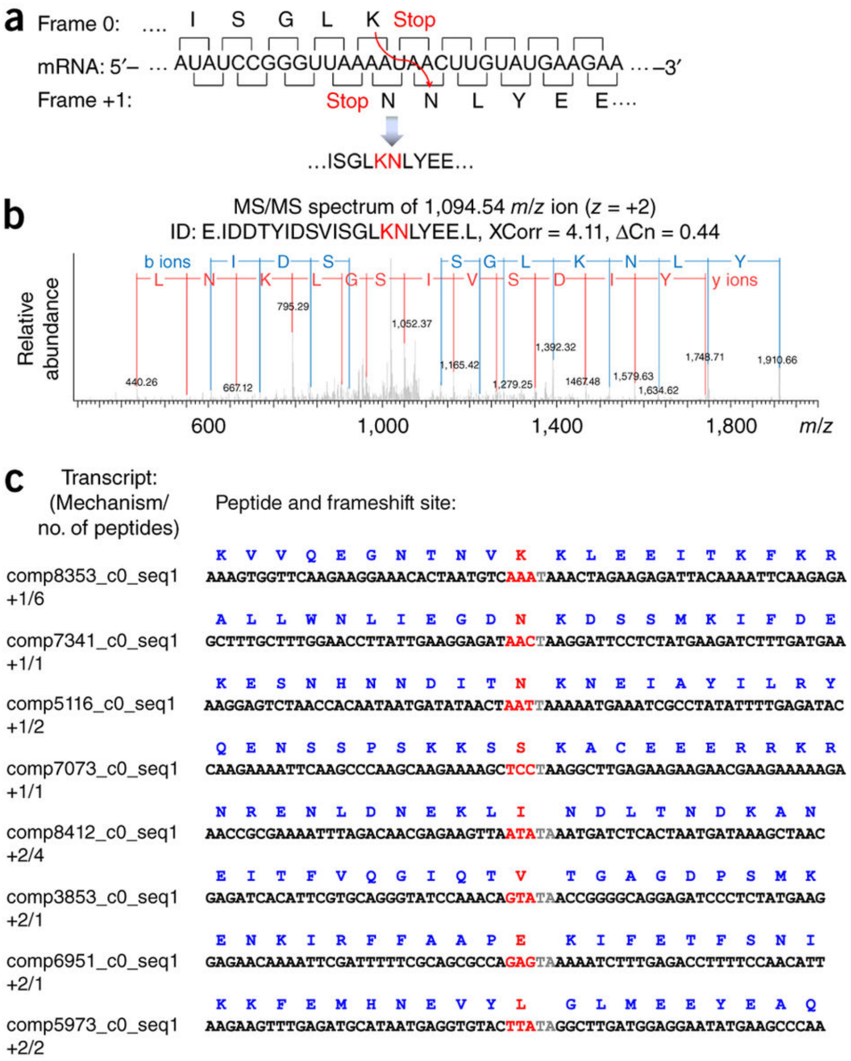

a
Frame 0: .... I S G L K Stop
mRNA: 5′– ... AUAUCCGGGUUAAAAUAACUUGUAUGAAGAA ... –3′
Frame +1: Stop N N L Y E E....
...ISGLKNLYEE...
b
MS/MS spectrum of 1,094.54 m/z ion (z = +2)
ID: E.IDDTYIDSVISGLKNLYEE.L, XCorr = 4.11, ΔCn = 0.44
b ions
y ions
Relative abundance
440.26
667.12
795.29
1,052.37
1,165.42
1,279.25
1,392.32
1467.48
1,579.63
1,634.62
1,748.71
1,910.66
600
1,000
1,400
1,800
m/z
c
Transcript: (Mechanism/ no. of peptides)
Peptide and frameshift site:
K V V Q E G N T N V K K L E E I T K F K R
comp8353_c0_seq1 +1/6
AAAGTGGTTCAAGAAGGAAACACTAATGTCAAATAAACTAGAAGAGATTACAAAATTCAAGAGA
A L L W N L I E G D N K D S S M K I F D E
comp7341_c0_seq1 +1/1
GCTTTGCTTTGGAACCTTATTGAAGGAGATAACTAAGGATTCCTCTATGAAGATCTTTGATGAA
K E S N H N N D I T N K N E I A Y I L R Y
comp5116_c0_seq1 +1/2
AAGGAGTCTAACCACAATAATGATATAACTAATTAAAAATGAAATCGCCTATATTTTGAGATAC
Q E N S S P S K K S S K A C E E E R R K R
comp7073_c0_seq1 +1/1
CAAGAAAATTCAAGCCCAAGCAAGAAAAGCTCCTAAGGCTTGAGAAGAAGAACGAAGAAAAAGA
N R E N L D N E K L I N D L T N D K A N
comp8412_c0_seq1 +2/4
AACCGCGAAAATTTAGACAACGAGAAGTTAATATAAATGATCTCACTAATGATAAAGCTAAC
E I T F V Q G I Q T V T G A G D P S M K
comp3853_c0_seq1 +2/1
GAGATCACATTCGTGCAGGGTATCCAAACAGTATAACCGGGGCAGGAGATCCCTCTATGAAG
E N K I R F F A A P E K I F E T F S N I
comp6951_c0_seq1 +2/1
GAGAACAAAATTCGATTTTTCGCAGCGCCAGAGTAAAAATCTTTGAGACCTTTTCCAACATT
K K F E M H N E V Y L G L M E E Y E A Q
comp5973_c0_seq1 +2/2
AAGAAGTTTGAGATGCATAATGAGGTGTACTTATAGGCTTGATGGAGGAATATGAAGCCCAA

Figure 2. Identification of amino acids inserted at frameshift sites. (a) Lysine (K) and asparagine (N) are inserted at the AAA_TAA_C heptamer. Nucleotide sequence surrounding the AAA_TAA +1 frameshift site is shown in the middle. Amino acid sequence is shown above for the zero frame and below for the +1 frame. (b) Recorded MS/MS spectrum confirming the presence of a peptide derived from predicted frameshifting. **(c)** Peptides detected by MS/MS analysis that were derived from the translation of frameshift sites are shown along with the corresponding nucleotide templates. Nucleotides "skipped" as a result of frameshifting are highlighted in gray. Codons preceding stop codons are shown in red, and the amino acids inserted at frameshifting sites are indicated.

Sequence properties of +1 and +2 frameshifting sites.

Among 1,765 putative frameshift sites detected with Ribo-seq, about three quarters (1,368) consisted of an AAA codon followed by a stop codon, and a quarter (397) contained other codons preceding stop. Altogether, we observed 47 out of 62 possible sense codons at the frameshift sites. The supporting information (ribosome footprint density and BLAST hit alignments) for various types of frameshifting sites is shown in Supplementary Note 3.

Earlier observations of frequent use of AAA_TAA and AAA_TAG as frameshifting sites in *Euplotes* prompted researchers to speculate that there is something special about AAA that allows frameshifting to take place at this codon (Klobutcher and Farabaugh 2002). Our comparison of codon frequencies upstream of stop codons in the frameshift sites and in the sites of termination revealed that AAA was not only the most frequent codon at the frameshift sites (Fig. 3a), but also was the second most frequent codon at the termination sites (Fig. 3b). However, high frequency of AAA codons at frameshift sites cannot be explained simply by their high frequency upstream of stop codons. The AAA codon was overrepresented at the frameshift sites in comparison with its usage in internal positions of coding frames, occurring ~8 times more frequently than expected (Fig. 3a). Moreover, 6 out of 7 AT-only codons were the most frequent codons at the frameshift sites, and they were also overrepresented at the frameshift sites in comparison with internal positions (Fig. 3a). A higher frequency of AT-rich codons among frameshift sites suggests that weak interactions between P-site tRNA and its codon in the initial frame increases possibility of frameshifting. We also found that all XXX codons (i.e. codons with identical nucleotides) were also enriched (relative to most non-AAA codons) at the frameshift sites (Fig. 3a, right), even though CCC and GGG were not the most frequent ones, owing to a relatively low GC content of *Euplotes* genomes. This suggests that the ability of P-site tRNAs to form base pairing with a codon in +1 frameshifting also increases chances of frameshifting because XXX codons would re-pair with XXT forming perfect Watson-Crick interactions with the first two subcodon positions.

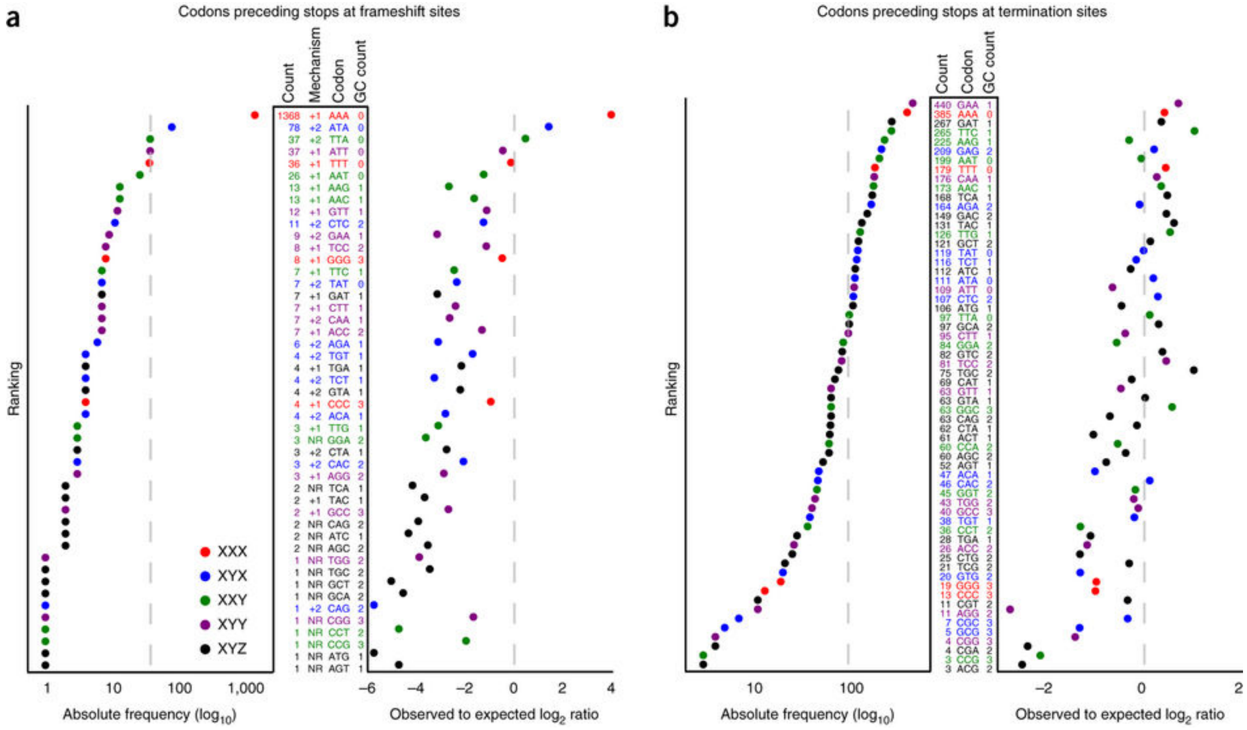


Figure 3. Distribution of codons upstream of stop codons at the frameshift sites and at the sites of translation termination. (a) Frameshift sites. The plot on the left shows absolute frequency of each sense codon ranked based on its frequency. Identity of codons is given by Codon in the middle table. GC content and the inferred mechanism of frameshifting (+1 or +2) are also indicated (nr indicates that the mechanism was not resolved). The absolute number of frameshift sites is listed in Count. Plot on the right shows frequency of codons relative to their expected occurrence based on their usage in internal positions of coding regions. Rows are colored according to codon type. (b) Sites of translation termination. See panel (a) for details. Broken lines indicate average values for absolute frequencies and expected values for normalized frequencies.

Interestingly, XYX codons (same nucleotides at the 1st and 3rd subcodon position, but a different nucleotide in the 2nd subcodon position) supported +2 ribosome frameshifting. Figure 1c shows a ribosome density profile for an mRNA containing an ATA_TAA frameshift site. It appears that the ribosomes shifted into the -1 frame. However, the mechanism was found to be +2 frameshifting based on the MS/MS analysis (Supplementary Note 1). Also, +2 frameshifting seemed to be more likely because in this case the isoleucine tRNA decoding the ATA codon would re-pair with the same ATA codon. We found 9 XYX codons (out of 16 possible) in the +2 frameshift sites (Fig. 3a) with ATA being the most frequent. The other codons that seemed to support +2 frameshifting were XTA that have T and A in the +2 and +3 positions.

Surprisingly, we did not observe noticeable underrepresentation of “shifty” codons upstream of stop codons that are recognized as terminators. The AAA codon was the second most frequent codon preceding terminator stop codons (Fig. 3b). An example of termination at AAA_TAA is shown in Supplementary Fig. 3a. Therefore, it is clear that whether the ribosome terminates or not at a particular stop codon does not depend solely on the identity of a codon preceding it, and that additional signals should be in place. Examination of information content surrounding frameshift sites and termination sites did not reveal position-specific sequence signals (Fig. 4a). Instead, it appears that the translation machinery senses the end of the mRNA and terminates only at the stop codons close to polyA. This is consistent with *Euplotes* having very short 3’ UTRs. Some mRNAs require longer 3’UTRs, e.g. selenoprotein mRNAs need to accommodate SECIS elements (Supplementary Fig. 3b). However, the “distance” between the polyA tail and the genuine site of termination could be structural rather than sequence-based such that the SECIS structure could bring the polyA tail close to the position of the termination site. Indeed, we observed highly structured 3’UTRs in all selenoprotein genes and found only a single example of a long 3’UTR other than that coding for selenoproteins (Supplementary Fig. 3c), but even in this case there is a possibility of a functional RNA secondary structure in its 3’UTR.

The effect of frameshifting on gene expression.

The high frequency of ribosomal frameshifting in *Euplotes* suggested that it was not as detrimental as in other organisms. Metagene analysis (Fig. 4a, see Supplementary Fig. 4 for corresponding RNA-seq density) revealed similar ribosome density upstream and downstream of frameshift sites. Therefore, the efficiency of frameshifting was comparable to that of standard decoding. On the other hand, there was a substantial drop of density relative to stop codons identified as termination sites (Fig. 4b). At the same time, a peak of ribosome density was also present about 30 nts upstream of frameshift sites (Fig. 4a), the distance roughly corresponding to the distance between A-sites of the two stacked ribosomes. Such stacking would be expected if ribosomal frameshifting is slower than standard decoding of sense codons. A slight depletion of ribosomes was also observed immediately downstream of the frameshift sites (Fig. 4a). Therefore, it is plausible that while ribosomal frameshifting does not impose considerable costs on the accuracy of synthesized proteins (e.g. AAA_TAA_A would be decoded in the same way as AAA_AAA), there is a cost to the speed of the ribosome and subsequently increase the number of ribosomes per mRNA. In this case frameshifting would be expected to be harmful in genes expressed at high levels.

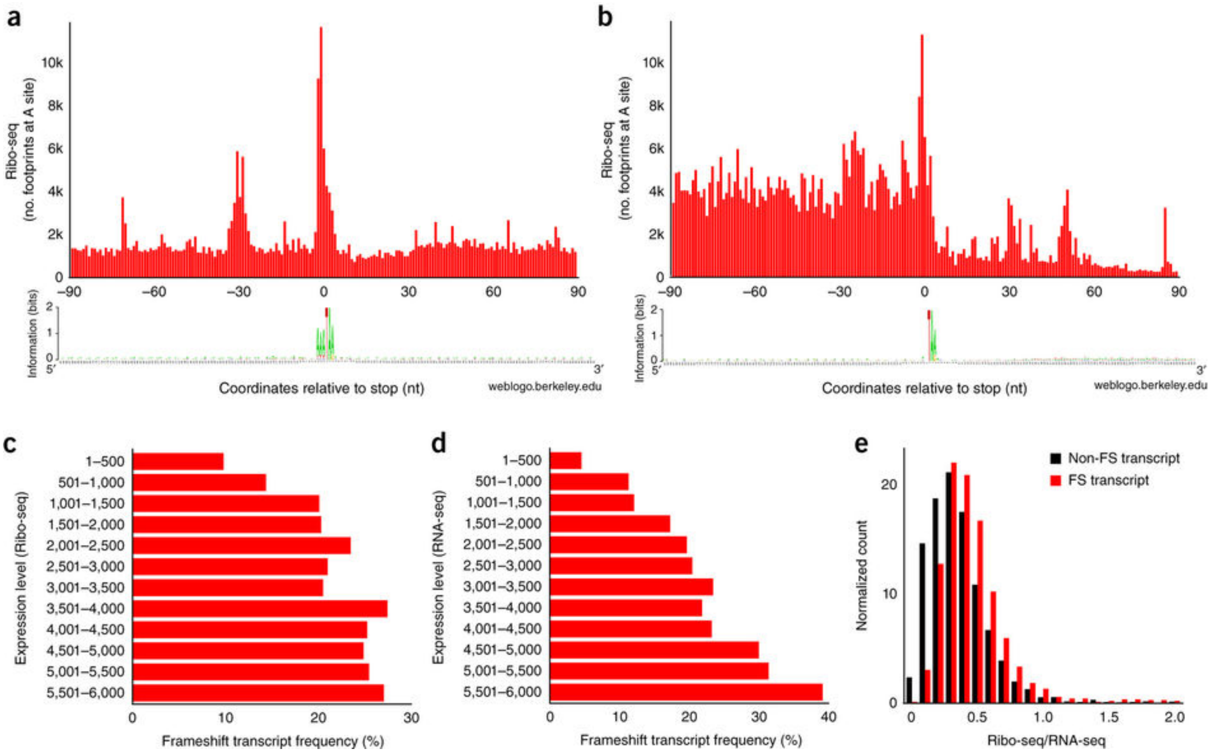


Figure 4. Metagene analysis of ribosome profiling and distribution of frameshifting according to transcript levels. (a) Metagene analysis of ribosome density in the vicinity of frameshift sites. First nucleotide of a stop codon is shown as zero coordinate. Note that while ribosome density upstream and downstream of frameshift sites is similar, there is a peak of density at the frameshift sites and this is accompanied by another peak 30 nucleotides upstream. A sequence logo below represents the information content of sequences used for metagene alignment. The sequence AAA_TAA is predominant, and there are no other position-specific signals associated with frameshifting. (b) Metagene analysis of ribosome density in the vicinity of translation termination sites. A drop in ribosome density is evident downstream of stop codons. A sequence logo representing information content in the sequences used for metagene analysis is given below. Only mRNAs with 3'UTRs longer than 90 nts (polyA is not included) were used. (c) Frequency of transcripts with the sites of ribosomal frameshifting (axis X) versus the transcripts ranked based on the levels of protein synthesis (Ribo-seq density), axis Y. (d) Similar to (c), but ranking is based on RNA levels (RNA-seq density). (e) Distribution of transcripts with different Ribo-seq to RNA-seq ratios containing frameshift sites (red) and not containing frameshift sites (black).

To test this hypothesis, we explored how frameshifting relates to gene expression levels based on RNA-seq and Ribo-seq signals (Fig. 4c,d). Indeed, we found that frameshifting was less frequent in highly expressed genes, supporting the idea that frameshifting is somewhat harmful in highly expressed genes. However, when we measured frequency of frameshifting in genes with different translation efficiency (TE) measured as the ratio of Ribo-seq signal to RNA-seq signal, we found that frameshifting was more frequent in genes with high TE (Fig. 4e). The ribosome density at any given location is expected to positively correlate with translation initiation rates and anticorrelate with elongation rates at that location. Therefore, while we cannot exclude the possibility that frameshifting is more frequent in genes with high initiation rates, a much more likely explanation is that the high Ribo-seq to RNA-seq ratio in mRNAs expressed with ribosome frameshifting was due to increased ribosome density caused by ribosome pauses and queuing induced by ribosomal frameshifting.

Since we found that particular codons are the most frequent at the frameshifting sites (mononucleotide and AT-rich with AAA being overrepresented the most), we hypothesized that frameshifting efficiency may vary depending on the identity of a codon upstream of a stop. To verify the hypothesis, we split frameshifting sites on AAA and non-AAA and analyzed the distribution of footprint densities (Fig. 5a,b). It appeared that the ribosome density does not change significantly downstream of frameshifting sites neither for AAA nor for non-AAA frameshifting sites (Fig. 5c), although the pause at non-AAA containing sites is less frequent (Fig. 5e). Why then are AAA codons preferred at frameshifting sites? A possible explanation is that the efficiency of frameshifting at non-AAA codons is context dependent and only efficient frameshifting sites are selected during evolution. While we have not observed a specific nucleotide context associated with non-AAA codons at the frameshifting sites, we noticed that TAG occurs almost three times more frequently (~29%) at non-AAA frameshifting sites than at AAA frameshifting sites (~12%) (Fig. 5a,b). To analyse how TAA and TAG stop codons affect frameshifting we compared footprint densities at the frameshifting sites depending on which stop codon is used (Fig. 5d,e). While we did not find a significant difference in a change of density downstream of frameshifting sites, it appeared that the peak of density associated with presumed ribosome pausing at the frameshifting sites was significantly greater for TAA codons than for TAG codons (Fig. 5f).

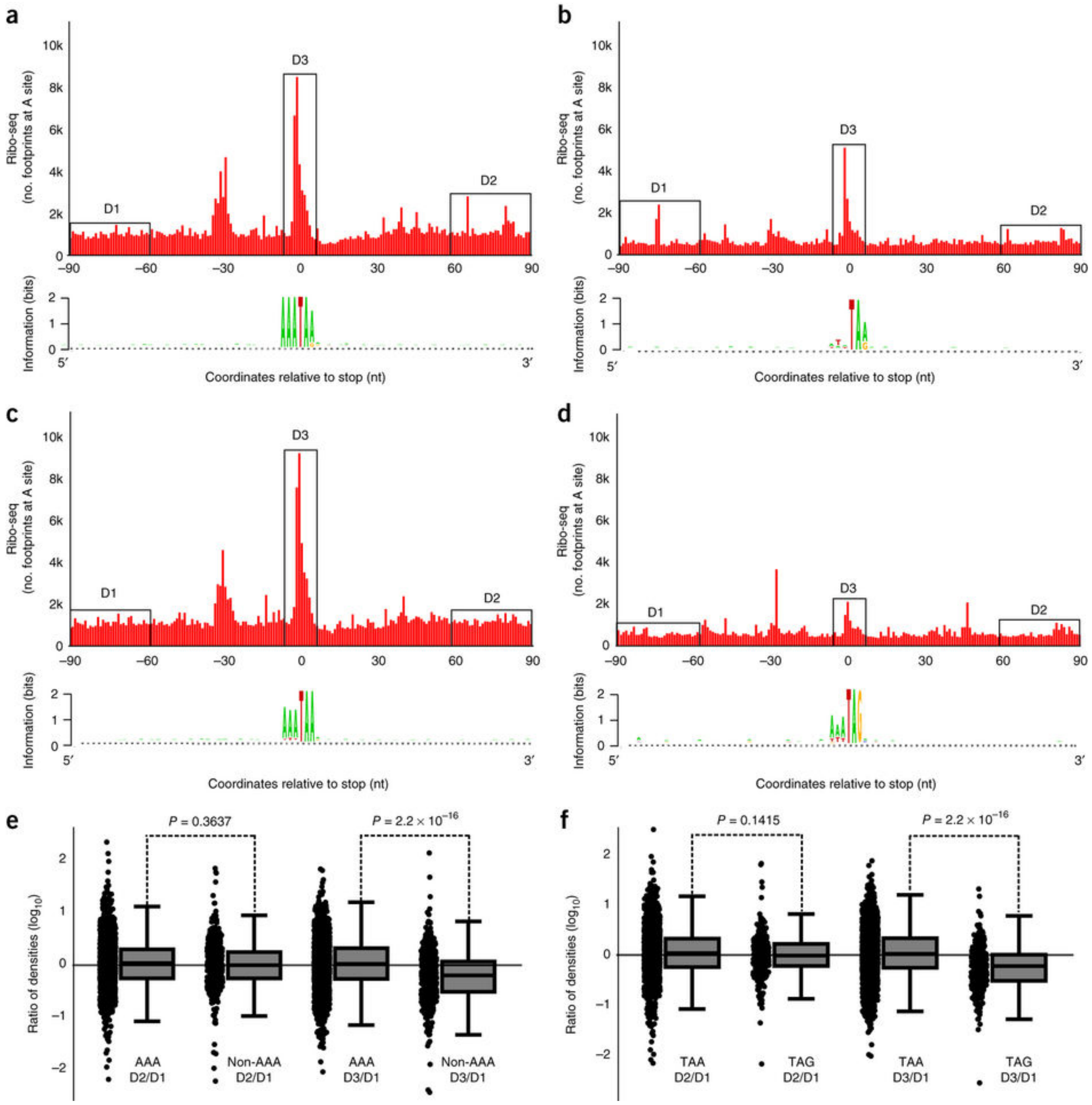


Figure 5. Comparison of ribosomal frameshifting at AAA vs non-AAA frameshifting sites and TAA vs TAG frameshifting sites. Aggregated densities of ribosome footprints around frameshift sites containing AAA codon preceding stop (a), non-AAA codons (b), TAA stop codons (d) and TAG stop codons (e). Comparison of footprint density changes observed at frameshift sites at each mRNA (D3 region) and downstream of frameshift sites (D2) relative to footprint density upstream of frameshift sites (D1). D1 and D3 regions were chosen 60 nts upstream and downstream of frameshift sites in order to avoid aberrant densities inflicted by ribosome pauses at frameshifting sites. Box plots represent ratio distributions with horizontal line corresponding to the median, box representing 25th and 75th percentiles and whiskers 5th and 95th percentiles. The comparison was carried out for AAA (n=1368) vs non-AAA (n=397) containing frameshift sites (e) and TAA (n=1488) vs TAG (n=277) containing frameshifting sites (f). P-values were calculated using unpaired Wicloxon rank-sum test on log ratios. The data suggest that the frameshifting efficiencies are similar at all frameshift sites, but strong pauses (D3/D1) are less frequent in non-AAA and TAG containing sites.

Frameshift patterns do not evolve under strong purifying selection.

In most well-studied cases of programmed ribosomal frameshifting (e.g. eukaryotic antizymes and bacterial release factor 2), the frameshift sequence and its occurrence are remarkably conserved (Baranov et al. 2002; Ivanov and Atkins 2007). In fact, evolutionary conservation of frameshift patterns is frequently used for the detection of recoded genes (Sharma et al. 2011). In all these cases, the efficiency of frameshifting is below 100%, and two protein products are usually synthesized from the same mRNA, one being decoded according to the rules of standard genetic decoding and another being a product of frameshifting. The ratio between these two products is functionally important and is often tightly regulated (Atkins et al. 2016). Therefore, there is a strong evolutionary pressure to preserve the frameshift site and its regulatory capacity, leading to strong stabilizing selection acting on the sequences of frameshift sites and stimulatory signals. In contrast, frameshifting in two *Euplotes* species was often characterized by cases where only one of the two orthologous sites used frameshifting (a typical example is shown in Fig. 6a). While the amino acid sequences of two orthologous genes were conserved, the corresponding nucleotide sequences differed by a single indel. Thus, frameshifting in *Euplotes* is not regulatory and the phenotypic difference between gene variants with and without frameshift sites is unlikely to be high.

Normally, there is a strong negative selection acting on single nucleotide indels inside protein coding regions due to their dramatic effects on the sequence of synthesized protein. In *Euplotes*, however, it could be expected that certain indels that likely create an efficient site of ribosomal frameshifting irrespective of nucleotide context (e.g. AAA_AAA to AAA_TAA_A mutation) would have no effect on the sequence of the synthesized protein. Therefore, indels would be expected to evolve under different evolutionary selection depending on where they occur. To explore evolution of indels, we analyzed the frequency of sequences surrounding single nucleotide indels. We generated pairwise alignments of orthologous sequences from the transcriptomes of both species using FASTA (Pearson 2004) and counted occurrences of each hexamer where a gap in the alignment corresponded to the fourth position (from the 5’ end) of the hexamer (highlighted sequence in Fig. 6a). Then, we normalized the frequency of such patterns in gapped alignments to the total number of their occurrence in the two transcriptomes (Fig. 6b,c). The abundance of patterns matching

AAATAA was striking (Fig. 6b,c). Indels in the center of the AAATAA pattern were strongly overrepresented in comparison with other patterns in both species, suggesting that frameshifting in *Euplotes* evolves essentially neutrally to produce AAA-stop frameshifting sites, though this is unlikely to be the case for non-AAA frameshifting sites.

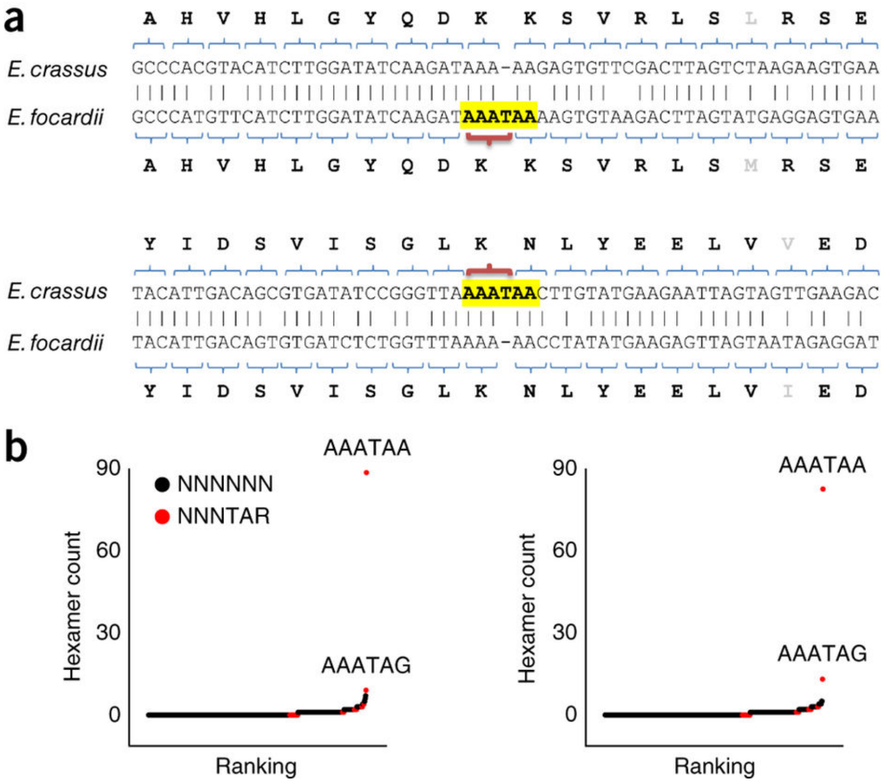


Figure 6. Cross-species comparison and frequency of nucleotide deletions in different hexamers. (a) Two typical pairwise alignments containing single nucleotide gaps in one of two orthologous sequences in *E. crassus* and *E. focardii*. **(b)** Frequency analysis of all possible hexamer patterns corresponding to deletions (as highlighted in yellow in a) in pairwise alignments for *E. crassus* (left) and *E. focardii* (right). The Y axis shows the frequency of each hexamer found in the pairwise alignments with a gap corresponding to the fourth position of the hexamer. Hexamers that end with either TAA or TAG are shown in red. Two most frequent hexamers, AAATAA and AAATAG, are indicated.

Conclusions.

In this work, we provide manifold evidence for the frequent occurrence of ribosomal frameshifting during translation in *Euplotes* ciliates. Ribosomal frameshifting occurs at the stop codons where tRNAs in the P-site slip forward predominantly either by 1 or 2 nucleotides. The most frequent type of frameshifting is +1 at AAA codons preceding stop; however, frameshifting also occurs at many other sense codons. While this work was under review, a study of two other *Euplotes* was published where frameshifting sites were predicted based on genomic and transcriptomic sequences (Wang et al. 2016), supporting our findings. Our analyses further show that ribosomal frameshifting in *Euplotes* is plastic and rapidly evolves, that it is the predominant process at stop codons and that it has no or low impact on the accuracy of protein synthesis, though it likely affects ribosome speed. Interestingly, sequences that trigger ribosomal frameshifting are also found as genuine termination sites. The data suggest that the function of stop codons as frameshifters or terminators is determined by their proximity to polyA tails and that additional mechanisms are required for efficient termination. Thus, the presence of a stop codon is not a sufficient feature for translation termination in *Euplotes*. Instead, the default function of stop codons is ribosomal frameshifting. This is consistent with recent findings of reassignment of all stop codons in *Condylostoma magnum* where stop codons function as terminators only in close proximity to mRNA 3' ends (Heaphy et al. 2016; Swart et al. 2016). A significant evolutionary distance between *Euplotes* and *Condylostoma* suggests an intriguing possibility that it may be a general property of ciliate decoding. If so, it may explain high frequency of changes in the genetic code in these species. A degree of positional preference of translation termination in other eukaryotes requires further exploration.

Accession Codes.

PRJNA329413; SAMN05412464; SRP078897; PRJNA329414; SAMN05412809; SRP078901; MJUV00000000; MECR00000000; PXD004333; .

Data availability.

Sequence data that support the findings of this study have been deposited in the following repositories: for *E. crassus* (BioProject: PRJNA329413; BioSample: SAMN05412464; SRA: SRP078897) and for *E. focardii* (BioProject: PRJNA329414; BioSample: SAMN05412809; SRA: SRP078901). Proteomics data were deposited to PRIDE (PXD004333) the interpretations of sequence data, such as coordinates of frameshifting sites are available upon request.

Online Methods

Genome sequencing and assembly. The nucleotide sequence of the *E. crassus* strain CT5 macronuclear genome was obtained by using a combination of Roche 454 (a total of 2,550,648 reads covering 577,513,019 bp, with an average read length of 236 bp) and Illumina (27,092,578 reads with an average read length of 77 bp, totaling 2,086,128,506 bp) sequencing. The macronuclear genome of *E. focardii* was generated through Illumina paired-end sequencing (a total of 43,588,788 reads covering 4,402,467,588 bp, with an average read length of 100 bp).

To identify sequences of other organisms within the dataset, we utilized DeconSeq (Schmieder and Edwards 2011). The following datasets were used: bacterial genomes (2,206 unique genomes, 02/12/11), archaeal genomes (155 unique genomes, 02/12/11), *Salmonella enterica* genomes (52 strains, 12/16/10), bacterial genomes HMP (76,337 WGS sequences, 02/12/11), and viral genomes in RefSeq 45 (3,761 unique sequences, 02/12/11). Whereas very little contamination was observed in *E. crassus* samples, bacterial sequences were found in *E. focardii* samples. To filter them out, we applied the following procedure: for *E. crassus* threshold values were left at default values (80% coverage and 95% identity), whereas for *E. focardii* they were changed to 50% coverage and 80% identity. Bacterial sequences in the genome data are not unexpected, considering that both ectosymbionts and endosymbionts have been reported in ciliates (Dziallas et al. 2012).

Several assembly programs were used to generate independent whole-genome assemblies, including ABYSS (Simpson et al. 2009), SOAP (Luo et al. 2012), SSAKE (Warren et al. 2007), Velvet (Zerbino and Birney 2008), Celera (Myers et al. 2000), 454 Newbler v.2.7, and PCAP (Huang et al. 2003; Huang and Yang 2005). To perform the assembly, we followed the instruction manuals for Newbler and Celera and the published protocols for other programs. A hybrid assembly (short reads pre-assembled using Velvet, with the final assembly done using Newbler) was chosen for further analyses (designated as "Newbler" in Supporting data Table 1). The *E. crassus* genome assembly consisted of 56,588 contigs, with N50 of 1.6 kb. The *E. focardii* genome assembly consisted of 109,492 contigs, of which 36,663 contigs (59M) were larger than 500 bp with the N50 of 2.1 kb.

Separately, selenoprotein genes were analyzed as described (Turanov et al. 2009). tRNA prediction was carried out using tRNAscan-SE (Lowe and Eddy 1996)and ARAGORN (Laslett and Canback 2004).

Transcriptome analysis. Frozen *E. crassus* pellets were cryogenically ground in a Biospec bead homogenizer. Cell powder was lysed in 1 ml of lysis buffer (20 mM Tris-HCl, pH 7.5, 140 mM KCl, 10 mM MgCl2, 0.25% Triton, 100 mg/l cycloheximide, protease inhibitors from Roche). Lysate was loaded on a 2 ml cushion of 1 M sucrose in 20 mM Tris-HCl, pH 7.5, 140 mM KCl, 5 mM MgCl2, 100 mg/l cycloheximide). Samples were centrifuged for 2 h at 45,000 rpm in a SW55 rotor. Pellets were recovered and resuspended in lysis buffer, and then incubated for 1 h with 750 U of RNAse I (Ambion) per 30 U of lysate (measured at A260). Following RNA digestion, sequencing libraries were prepared as described (Gerashchenko et al. 2012), starting with gradient ultracentrifugation. There were several additional changes to the procedure. Instead of polyadenylation, we attached a 3’ adapter (IDT, miRNA linker #1) as a handle for subsequent reverse transcription step using T4 RNA ligase 2 (NEB). The reverse transcription primer was changed accordingly: (5’-GATCGTCGGACTGTAGAACTCTGAACCTGTCGGTGGTCGCCGTATCATT/iSp18/CAAGCAGAAGACGGCATACGAATTGATGGTGCCTACAG-3’), which allowed us to keep the 3’ ends of footprints unperturbed. The following are the sequences of forward and reverse primers for the final PCR: CAAGCAGAAGACGGCATACGA and AATGATACGGCGACCACCGA. Sequencing was performed on an Illumina HiSeq2000 platform. The transcriptome assembly was carried out using de novo assembler Trinity (Haas et al. 2013), producing 33,701 unique transcripts.

Identification of frameshift sites. Sequences of ribosome footprint cDNAs (Ribo-seq) from *E. crassus* obtained in three replicates were aggregated producing 9,620,943 reads.

They were aligned to the transcriptome using Bowtie software v.0.12.839 allowing ambiguous mapping and up to 3 mismatches per read (-v 3). 8,353,221of reads (86.2%) were aligned to the transcriptome. The Integrative Genomics Viewer (IGV) 40 was used to visualize reads aligned to each transcript. Using IGV we visually analyzed all transcripts where the number of mapped footprints was ≥ 100 reads. Supplementary Note 4 shows examples of IGV screenshots in the vicinity of frameshifting sites whose productive translation was directly supported by peptides matching mass spectra (shown in Supplementary Note 1b). The obtained alignments were used to determine the boundaries of translated segment within a transcript. Frameshift sites were identified by analyzing ORF organization within the translated region at internal stop codons using maximum parsimony as a guiding principle in determining the direction of frameshifting to yield the minimal number of frameshift sites per transcript in most cases. Transcripts with frameshift sites were aligned to corresponding genomic contigs to verify sequence identity and avoid misinterpretation of indel sequencing errors as ribosomal frameshifting sites.

Proteomic and Ribo-Seq analyses. Proteomics analysis employed conventional shotgun bottom-up approach described elsewhere (Petyuk et al. 2008; Depuydt et al. 2013; Depuydt et al. 2014). Briefly, cells were resuspended in the lysis buffer (50 mM Tris-HCl pH 8.0, 8 M urea, 10 mM DTT, 1 mM EDTA), pulverized in liquid nitrogen followed by melting and sonication in a water bath for 1 min. The proteins were then digested using trypsin (samples 1 and 2) and Glu-C (sample 3, pH 7.5), followed by fractionation by SCX (trypsin sample, 25 fractions collected) and High-pH RP (trypsin and Glu-C samples, 24 concatenated fractions collected (Yang et al. 2012)). Analysis by liquid chromatography coupled with LTQ Orbitrap (Thermo Fisher, San Jose, CA) mass spectrometry (LC-MS/MS) was performed using a 100 min LC gradient. The details on the gradient and mass spectrometer settings can be found elsewhere41. The data were pre-processed with DeconMSn (Mayampurath et al. 2008) and DtaRefinery (Petyuk et al. 2010) tools, and analyzed using MS-GF+ (Kim and Pevzner 2014). The raw, peak lists and MS/MS identification files were deposited at PRIDE (dx.doi.org/10.6019/PXD004333).   Amongst the all peptide identifications, we retained only those that uniquely matched protein sequences originating from the frameshift events. The tolerances on parent ion mass measurement and MS/MS spectrum matching scores were optimized to achieve maximum number of identifications while not exceeding false discovery rate of 5%. Spectra for peptides spanning the frameshift locations were manually verified. The details on MS/MS data analysis along with parameter files and executable document reproducing all the post-search analysis steps were deposited as an R package at GitHub https://github.com/vladpetyuk/EuplotesCrassus.proteome.

For Ribo-Seq analysis, frozen *E. crassus* pellets were cryogenically ground in a Biospec bead homogenizer. Pellets were recovered and resuspended in lysis buffer, and then incubated for 1 h with 750 U of RNAse I (Ambion) per 30 U of lysate (measured at A260). Following RNA digestion, sequencing libraries were prepared as described 37, starting with gradient ultracentrifugation. Sequencing was performed on an Illumina HiSeq2000 platform.

E. crassus genome and transcriptome sequences were used as references for read alignments. The alignments were generated using Bowtie software v.0.12.7 (Langmead et al. 2009); up to two mismatches per read were allowed. We estimated positions of the ribosome A-sites with an offset of 15 nucleotides downstream of 5' ends of Ribo-seq data. Visualization and further manual analysis were conducted by using SAMtools package (Li et al. 2009), custom scripts and IGV (Thorvaldsdóttir et al. 2013).

Sequence patterns analysis. To analyze for frequency of indels that occurred since *E. crassus* and *E. focardii* split from their common ancestor we generated a set of pairwise alignments using FASTA (Pearson 2004). The alignments were generated by searching *E. crassus* sequences as query against *E. focardii* and also in a reverse order. The sequence pairs with the best scores were considered as true orthologous sequences and were used in further analysis. To minimize the potential effect from misalignments, or highly diverged sequence pairs, only those indels were analyzed that occurred exactly in the center of a 41-nucleotide stretch of the alignment containing no other indels. For each gap a hexamer pattern was registered whose fourth position (counting from the 5' end) corresponds to a gap in the alignment, e.g. PPPPPP pattern in the schematic alignment below

NNPPPPPPNN

NNNNN-NNNN

The observed-to-expected ratio of deletions in hexamers was calculated as the following

$$(gi\sum f)/(fi\sum g)$$

where gi is the number of gaps corresponding to pattern i and fi is the number of patterns i in the fraction of the genome predicted as coding.

Statistics. For the data shown in Figure 5 to estimate statistical significance between distributions of changes in footprint densities downstream of, upstream of and at the frameshifting sites. log(D2/D1) and log(D3/D1) we used Wilcoxon rank test. The exact p-values and degrees of freedom are provided in figure legend.

Diversity of antizyme decoding mechanisms among protists: classical +1 frameshifting, stop codon readthrough and single ORFs

Manuscript in preparation for journal submission

Abstract

Ornithine decarboxylase antizyme (OAZ) is a key negative regulator of cellular polyamine synthesis. From yeast to human antizyme it is encoded in two overlapping open reading frames (ORFs) and its synthesis requires a very rare and unusual decoding mechanism, +1 ribosomal frameshifting at the end of its first ORF. Elevated polyamine levels enhance frameshifting efficiency, closing an autoregulatory negative feedback loop.

We took advantage of recently sequenced transcriptomes and explored the occurrence of antizymes in over 200 diverse genera from the major protist supergroups. Most antizymes were found in the Alveolata superphylum and certain algae. Interestingly we did not find any antizymes in the other major protist groups; Excavata, Stramenopiles and Rhizaria. Surprisingly, while many antizyme genes were found to utilize +1 frameshifting, we also found antizymes encoded in a single ORF. Most strikingly we found a highly conserved in-frame stop codon in dinoflagellate species suggesting the use of stop codon readthrough as an alternative decoding mechanism. Comparative sequence analysis of flanking regions indicates the existence of conserved RNA structures that probably function as stimulators of stop codon readthrough. We tested stop codon readthrough by introducing a sequence containing the stop codon into a dual luciferase vector and expressing it in mammalian cells. Despite being expressed in a heterologous system, the insert supported stop codon readthrough which appeared to be sensitive to polyamine levels. In addition we observed the highly conserved and essential ‘NIKS’ motif of eukaryotic release factor 1 (eRF1), involved in stop codon recognition to be mutated at the ‘N’ residue to either ‘S’ or ‘C’ for each dinoflagellate species.

Introduction

Cellular polyamine levels are tightly regulated, as expected from their multiple important roles in cell functioning. The protein ornithine decarboxylase antizyme (OAZ), referred to as antizyme, is a key regulator of cellular polyamine levels, and also has cross pathway effects via its interaction with ATP citrate lyase (Tajima et al. 2016). The first identified antizyme gene, encoding rat antizyme 1 (Matsufuji et al. 1990), has its coding sequence in two different and partially overlapping open reading frames (ORFs) with synthesis of functional antizyme involving a programmed ribosomal frameshift event at the end of ORF1 (Miyazaki et al. 1992). The efficiency of ribosomes shifting to the +1 reading frame to enter ORF2 and so of the amount of antizyme synthesized, is dependent on cellular polyamine levels (Matsufuji et al. 1995). Elevated polyamine levels result in enhanced frameshifting and as antizyme is a negative regulator of polyamine levels, a homeostatic reduction of cellular polyamine levels. Antizyme translation is also dependent on a diverse range of cis-acting stimulatory signals which facilitate frameshifting (Ivanov and Atkins 2007). For mammalian antizymes 1 and 2 a sequence 5' of the frameshift site has been shown to enhance the efficiency rate (Matsufuji et al. 1996; Ivanov et al. 2000a), while the most common 3' stimulator is an RNA pseudoknot located directly downstream of the frameshift site (Matsufuji et al. 1995). Additional examples of regulatory frameshifting and their stimulators have recently been reviewed (Atkins et al. 2016).

While most vertebrates possess two antizyme paralogs which are widely expressed (Ivanov et al. 1998), a third mammalian antizyme paralog was found with specific male germ cell functions (Ivanov et al. 2000b) and additional paralogs expressed in specific tissues have been found in fish (Ivanov et al. 2007). Some eukaryotes most likely have lost antizyme, e.g. plants. However, where antizyme is found, the requirement for a +1 frameshift during mRNA translation is nearly universal; a single known exception is in the ciliate *Tethramymena thermophile* where it is encoded in a single ORF (Ivanov and Atkins 2007).

This single exception from an outgroup to fungi and protozoa is an important exception that suggests that the decoding strategies of antizymes may deviate from classical polyamine dependent +1 ribosomal frameshifting. This anticipation of antizyme decoding diversity is further supported by the breadth of genetic decoding in protsists in general that is exemplified by the recent discoveries of organisms with genetic codes supporting incorporation of amino acids at all 64 codons (Heaphy et al. 2016; Swart et al. 2016; Záhonová et al. 2016) and of the genetic code with ribosomal frameshifting as a standard feature (Lobanov et al. 2017). To explore the genetic diversity of antizymes in protists we took advantage of large scale sequencing projects, specifically The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (Keeling et al. 2014) which generated publicly available transcriptome data for hundreds of marine microbes.

The majority of antizyme coding transcripts were found in the Alveolata superphylum and certain algae. We did not find antizyme transcripts in either of the other major protist groups; Excavata, Stramenopiles and Rhizaria. Classical +1 frameshifting dependent antizymes were found in ciliates, chromerid, algae and one amoeba species. Single ORF antizymes were found in ciliates and dinoflagellates. Strikingly most antizyme coding sequences from dinoflagellates contain a conserved in-frame stop codon. In this manuscript we provide evidence for its functional utilization for regulated stop codon readthrough. These findings contribute to the steady growth of alternative genetic decoding examples (Baranov et al. 2015).

Results

Using a HMMER based approach (Finn et al. 2011), see Materials and Methods, we explored the transcriptomes of over 200 representative species from diverse genera across the major protist supergroups including: Archaeplastida (e.g. algae), Stramenopiles (e.g. diatoms), Alveolate (e.g. ciliates), Rhizaria (e.g. foraminifera), Excavata (e.g. euglenozoa), Amoebozoa and Opisthokonta (e.g. choanoflagellates). In total we identified 95 sequences matching the OAZ profile HMM, including four ciliate species matching more than one, see Figure S1 for the distribution of HMMER scores. To confirm the presence of the OAZ domain we searched databases; phmmer using the protein sequences and blastx of each transcript to find the optimal coding sequence, see Materials and Methods. Open reading frames containing the OAZ profile HMM are defined as stop codon to stop codon ORFs on the transcript. Phmmer and blastx results identified low HMMER scoring sequences (<20), do not contain OAZ protein domains or are uncharacterised, we identified 26 such sequences, summarized in Table S1. A total of 69 sequences contain an OAZ domain or closely related to antizymes, summarised in Table 1. For antizyme, the conserved residues are located in the C-terminal of the protein. To investigate the presence of this domain, we generated a multiple sequence alignment of the 69 candidates and built a sequence logo using WebLogo (Crooks et al. 2004). We compared the alignments to the OAZ domain of the ciliate *Tetrahymena thermophilla*, according to Pfam (Finn et al. 2016), Figure S2. We observe highly conserved residues in the C-terminal that are essential for antizyme interaction with ornithine decarboxylase (ODC) (Hoffman et al. 2005). To infer relatedness of the sequences we carried out maximum likelihood phylogenetic analysis using MEGA6 (Tamura et al. 2013), on the protein coding ORFs and included reference antizyme sequences from *Tetrahymena thermophilla, Schizosaccharomyces pombe* and *Homo sapiens OAZ1*. The resulting tree shows groups of orthologs branching together by phylum, with accompanying bootstrap values, Figure S3.

Contaminant sequences in the past have contributed to speculation surrounding a possible antizyme 4 and to the elusive plant antizyme (Ivanov and Atkins 2007). Figure S3 alludes to potential contaminant sequences, marked with an '*'. We indicate nine sequences that do not branch with other species of their phyla. We performed PSI-BLAST searches on sequences from; *Campanella umbellaria*, *Tiarina fusa1* and *Goniomonas pacifica*, the results indicate they are of animal origin; *Octopus bimaculoides* E-value 4e^{-53}, *Anopheles darling* E-value 2e^{-4} and *Papilio polytes* E-value 3e^{-6} respectively. Pairwise analysis of ciliate sequences *Tiarina fusa2, Myrionecta rubra* and *Strombidinopsis acuminata1* with their respective branch neighbours showed that they are 99% identical to each. Furthermore, these ciliate species utilize variant genetic codes, i.e. stop codon reassignments; however, the sequences here do not contain in-frame stop codons suggesting they are not of ciliate origin. We removed these six sequences as contaminants.

For the single represented amoeba sequence, *Filamoeba nolandi*, which shows low scoring database results (PSI-BLAST top result; *Neocallimastix californiae,* E-value 0.005), we cannot determine if it is a contaminant. The sequence from the ciliate *Protocruzia adherens*, does not branch with other ciliates. However, a recent study reported the ambiguous classification of this genus as it does not belong to any known ciliate class (Jiang et al. 2016). We believe the sequence from *Mesodinium pulex* is from an unidentified dinoflagellate, it is 69% identical to its closest branch neighbour, *Heterocapsa triquestra*, and refer to this as *ex.Mesodinium.sp*.

The functional domains of antizymes are encoded in the second ORF downstream of ribosomal frameshifting site, making its protein sequence more evolutionarily conserved (Ivanov and Atkins 2007). Therefore, the profile HMM can be used only for the identification of the part of antizyme that is encoded by the downstream part of the ORF sequence. To identify ORF architecture of sequences coding for antizyme and translation mechanisms involved, we developed an approach that is based on the analysis of the first ORF extensions, see Materials and Methods. Essentially we measured the distance between the stop codon of the first ORF (defined by the most 5' ATG codon as a start) and the next stop codon in all three frames; the frame that is the longest most likely encodes the antizyme. We further verified that with profile HMMs. The results are shown in Figure. 1

In all sequences of cryptophyta, glaucocystophytes, rhodophyta, amoebozoa and chromerida and five ciliate sequences, the OAZ profile HMM signal is found in a second downstream ORF, which is in the +1 frame relative to ORF1. The presence of the OAZ signal is indicated by a red circle, Figure 1. These findings are indicative of the requirement for a +1 ribosomal frameshift in these transcripts. The majority of ciliate sequences, 13, have the OAZ signal in the first ORF, for example *Tetrahymena elliotti* a sister species of *T. thermophilla*. Ciliate species that synthesise single ORF antizyme proteins utilize variant genetic codes, i.e. stop codon reassignments, while four out of the six species displaying a +1 frameshift sequence utilize the standard genetic code.

Interestingly, we observe 22 dinoflagellate sequences contain the OAZ signal in a downstream ORF which is in the same reading frame as ORF1, indicated by red crosses in Figure 1. Such observations are suggestive of stop codon readthrough of the first ORF. It also appears that three dinoflagellate sequences contain the OAZ signal in the first ORF, similar to the majority of ciliate species, while *Dinophysis acuminata* contains the OAZ signal in the +2 frame. Analysis of this transcript reveals it is most likely truncated at the 5' end, resulting in an unlikely ATG initiating codon, closely followed by a +2 shift. A summary of ORF lengths are provided in Table 1.

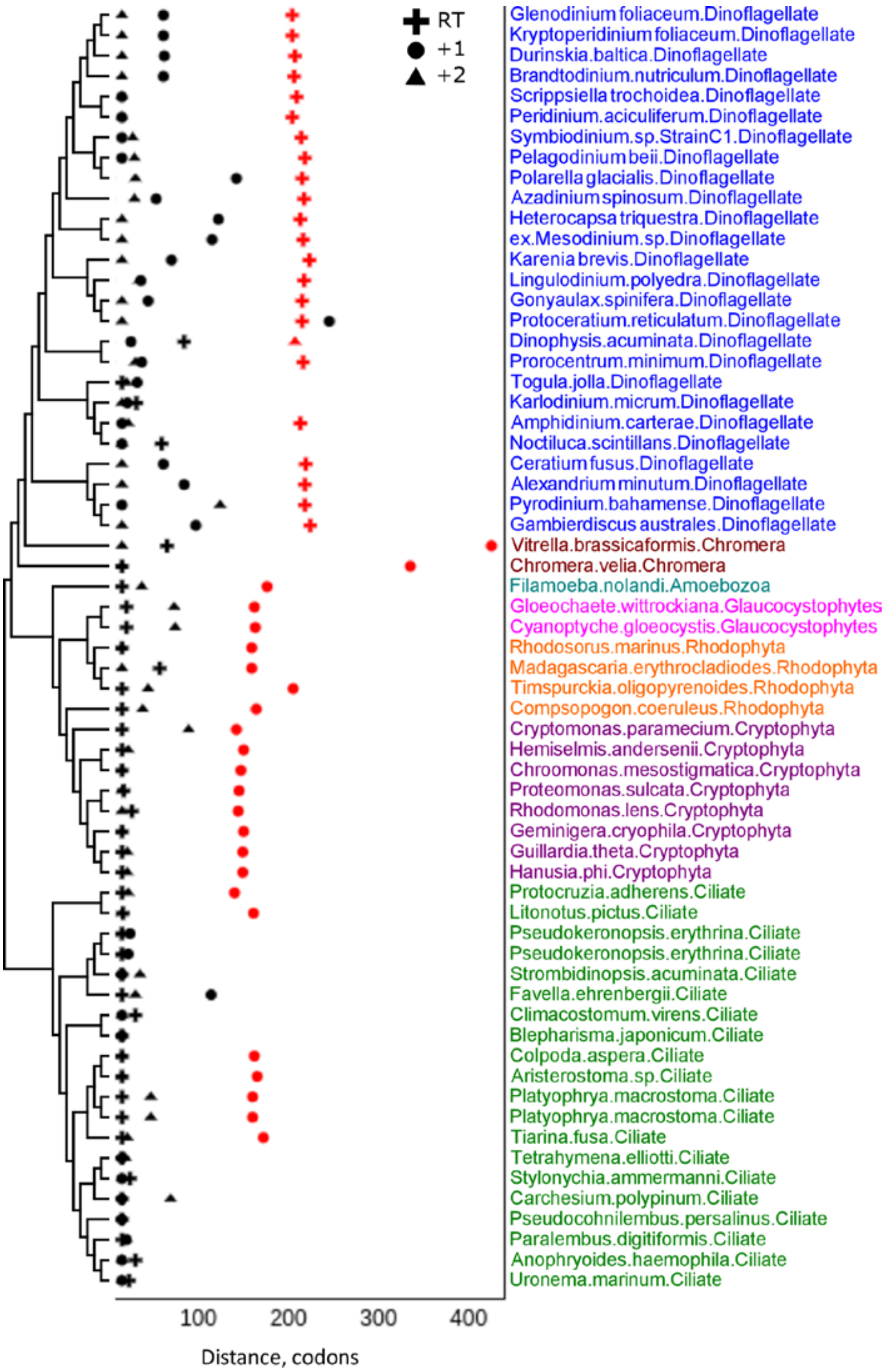


Figure 1. ORF extensions in all three frames. Extensions measured in codons is the distance from ORF1, X-axis. Sequences with the profile HMM OAZ in an extended frame (ORF2) are indicated with a red shape. Species without a red shape have their OAZ signal in the first ORF. The cross represents a readthrough (RT), circle is a +1 extended ORF2, and triangle is a +2 extended ORF2. The phyla are colour coded; blue = Dinoflagellate, maroon = Chromera, green = Ciliate, purple = Cryptophyta, orange = Rhodophyta, fuchsia = Glaucocystophytes and teal = Amoebozoa. The phylogenetic tree is constructed from HMMER derived OAZ protein sequences. The maximum likelihood (ML) tree was inferred with MEGA6, using the Jones-Taylor-Thornton (JTT) substitution model.

The nucleotide sequence downstream of a stop codon can affect the efficiency of termination and as a result weak termination at stop codons can facilitate readthrough and frameshifting (Namy et al. 2001). Using MACSE (Ranwez et al. 2011) we built multiple sequence alignments and investigated a region 13 codons upstream and downstream from the ORF1 and ORF2 boundaries. The alignments highlight the codons surrounding the +1 frameshift candidates Figure 2(a) and readthrough candidates Figure 2(b). For the +1 frameshift candidates we removed the 'T' from the stop codon at the end of ORF1 to maintain the reading frame during alignment. The +1 frameshift candidates display a high level of conservation in the region upstream of the ORF1 stop codon. In particular a highly conserved tryptophan at position -2/-3, glycine at -8 and cysteine at -13, positions are measured in codons relative to the ORF1 stop codon. These residues are found in other antizymes and believed to facilitate frameshifting, however speculation exists as to the influence of the nascent peptide on recoding (Atkins et al. 2016). For the vast majority of antizymes TGA is the stop codon at the ORF1 frameshift site, due to their ambiguous genetic codes ciliates employ either TGA or TAG at this position. Three known antizyme frameshift sites were identified in the position directly 5' of the ORF1 stop; GTT, TCC and TTT. The sequence for *Filamoeba nolandi* GAT.TAA, is not a known antizyme frameshift site (Ivanov and Atkins 2007).

For the dinoflagellate readthrough candidates, Figure 2 (b) we see highly conserved aspartic acid at position -3, glycine at -7 and glutamic acid at -13. We do not see any tryptophan residues in this region. The presence of the tryptophan residue directly upstream from the +1 frameshift site may slow the ribosome at a particular point on the mRNA facilitating the shift of reading frame, while not required for effective stop codon readthrough. For three species the ORF1 stop codon TGA, is substituted by sense codons; TCA, AGC and TGC. The suspected readthrough site is a highly conserved TGA followed by a leucine (CTC/T). This context is highly similar to known readthrough examples from viruses to mammals (Firth et al. 2011; Jungreis et al. 2011; Dunn et al. 2013; Loughran et al. 2014; Hofhuis et al. 2016). Studies have shown that translation termination is slower and less accurate than translation elongation (Freistroffer et al. 2000; Bertram et al. 2001) and as a result, low efficiency stop codon read through may occur in the absence of any stimulatory signals such as RNA secondary structures. However, in the region +6 (codons) downstream of the readthrough site, there are highly conserved GC nucleotides that could have the potential to base-pair forming a secondary structure Figure 2 (b).

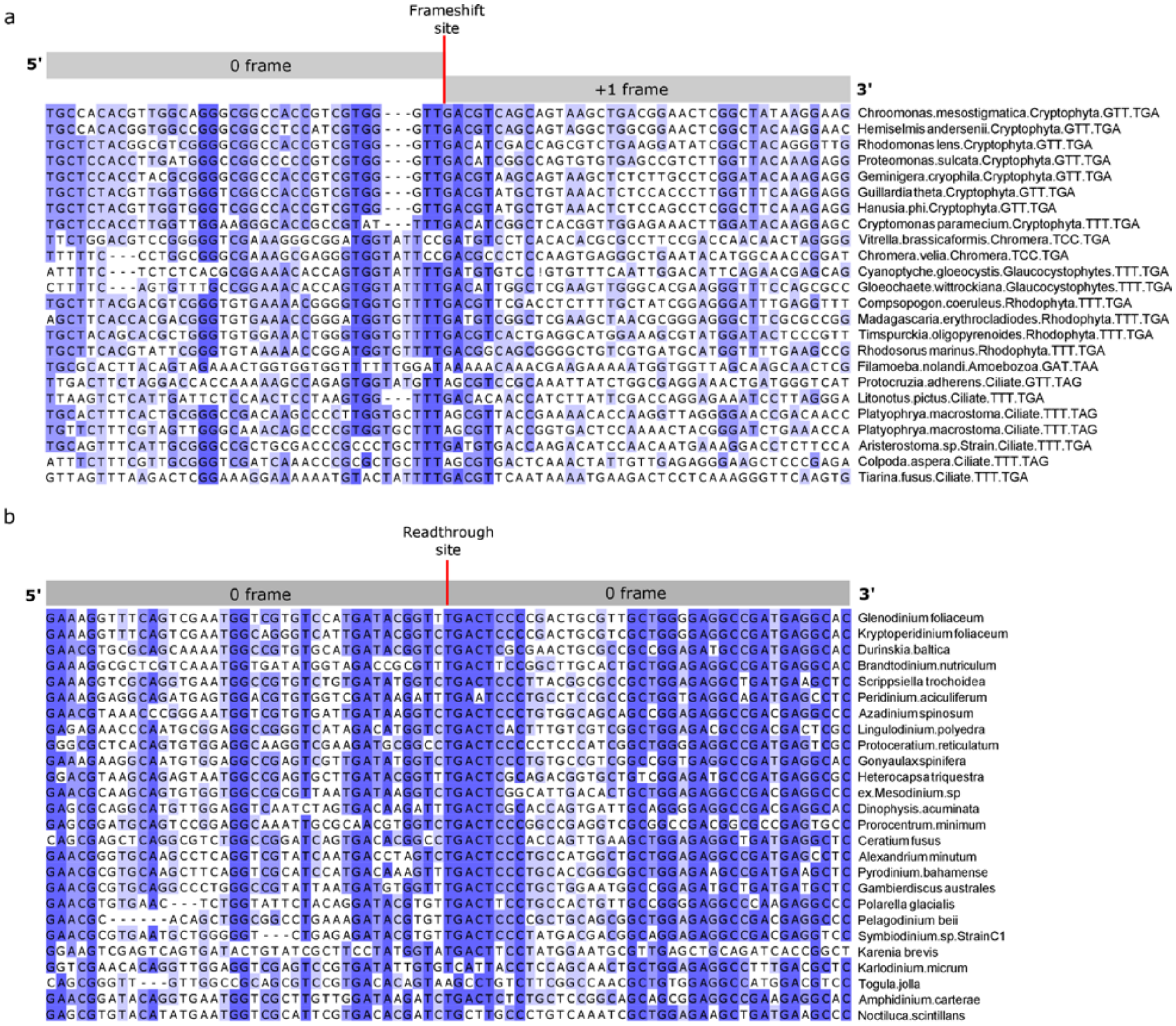

a
Frameshift site
5'
0 frame
+1 frame
3'
Chroomonas.mesostigmatica.Cryptophyta.GTT.TGA
Hemiselmis andersenii.Cryptophyta.GTT.TGA
Rhodomonas lens.Cryptophyta.GTT.TGA
Proteomonas.sulcata.Cryptophyta.GTT.TGA
Geminigera.cryophila.Cryptophyta.GTT.TGA
Guillardia theta.Cryptophyta.GTT.TGA
Hanusia.phi.Cryptophyta.GTT.TGA
Cryptomonas paramecium.Cryptophyta.TTT.TGA
Vitrella.brassicaformis.Chromera.TCC.TGA
Chromera.velia.Chromera.TCC.TGA
Cyanoptyche.gloeocystis.Glaucocystophytes.TTT.TGA
Gloeochaete.wittrockiana.Glaucocystophytes.TTT.TGA
Compsopogon.coeruleus.Rhodophyta.TTT.TGA
Madagascaria.erythrocladiodes.Rhodophyta.TTT.TGA
Timspurckia.oligopyrenoides.Rhodophyta.TTT.TGA
Rhodosorus marinus.Rhodophyta.TTT.TGA
Filamoeba.nolandi.Amoebozoa.GAT.TAA
Protocruzia.adherens.Ciliate.GTT.TAG
Litonotus.pictus.Ciliate.TTT.TGA
Platyophrya.macrostoma.Ciliate.TTT.TAG
Platyophrya.macrostoma.Ciliate.TTT.TAG
Aristerostoma.sp.Strain.Ciliate.TTT.TGA
Colpoda.aspera.Ciliate.TTT.TAG
Tiarina.fusus.Ciliate.TTT.TGA
b
Readthrough site
5'
0 frame
0 frame
3'
Glenodinium foliaceum
Kryptoperidinium foliaceum
Durinskia.baltica
Brandtodinium.nutriculum
Scrippsiella trochoidea
Peridinium.aciculiferum
Azadinium spinosum
Lingulodinium.polyedra
Protoceratium.reticulatum
Gonyaulax spinifera
Heterocapsa triquestra
ex.Mesodinium.sp
Dinophysis.acuminata
Prorocentrum.minimum
Ceratium fusus
Alexandrium minutum
Pyrodinium.bahamense
Gambierdiscus australes
Polarella glacialis
Pelagodinium beii
Symbiodinium.sp.StrainC1
Karenia brevis
Karlodinium.micrum
Togula.jolla
Amphidinium.carterae
Noctiluca.scintillans

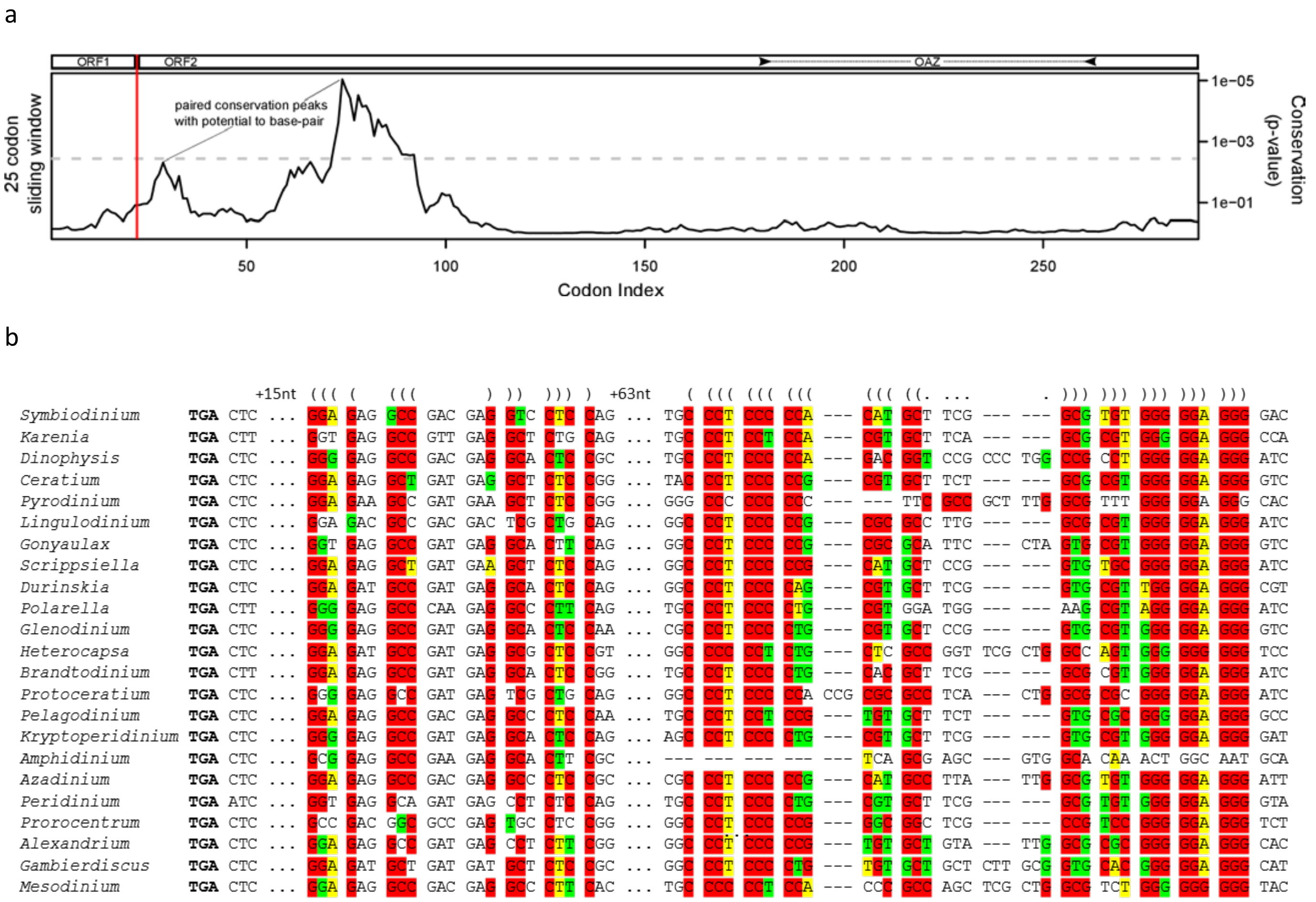
Figure 2. Sequence alignments surrounding ORF1 and ORF2 boundaries. 13 codons upstream and downstream of a) +1 frameshift sites, showing each species and phylum along with the last codon and stop codon of ORF1 and b) dinoflagellate readthrough sites plus three non-readthrough candidates. The location of the frameshift and readthrough sites are indicated with a red line. Alignments were generated using MACSE and presented using Jalview. Sequences are coloured by % identidy.

Secondary structure prediction

Previous studies have highlighted the importance of stimulatory factors promoting readthrough, such as translation of the *gag-pol* gene of the murine leukemia virus (MuLV) (Feng et al. 1992; Firth et al. 2011). To investigate the presence of conserved secondary structures we first looked for conservation at synonymous sites across the 23 dinoflagellate readthrough sequences, we used to perform the analysis SynPlot2 (Firth 2014), see Materials and Methods. Interestingly, we identified two peaks in the transcripts, both in contexts with potential to base-pair, Figure 3(a). From a strong sequence consensus we predicted two RNA stem-loops, at codon positions +6 and +21, Figure 3(b). We used IPknot (Sato et al. 2011), to predict the location and structure type. Both structures correspond to the locations of the peaks identified in Figure 3(a). We highlighted the predicted consensus structure, highlighting Watson-Crick base-pairing, G:C pairings are highlighted red, A:T are yellow and the wobble base-pair G:T are green. Remarkably both structures show a high level of similarity at the sequence level considering the diversity of dinoflagellate species involved. The observed location of the 'pair of peaks' and possible RNA stem-loop structures is more consistent with viral readthrough mechanisms rather than with antizyme frameshifting (Firth et al. 2011). While the RNA structures enhance readthrough in viruses the precise mechanisms by which these structures function is still to be resolved. We do not report any additional secondary structures from the other phyla.

Experimental validation of polyamine stimulated recoding of dinoflagelate antizymes

To determine whether predicted single and overlapping ORF dinoflagelate antizymes permit polyamine induced translation of their second ORFs we tested representative antizyme cassettes flanked by dual luciferase reporters in a heterlogous system. Recently, we described a modification to the classical dual luciferase reporter system that avoids potential distortions, sometimes observed using fused dual reporters, by incorporating 'StopGo' sequences on either side of the polylinker (Loughran et al. 2017). The advantage is that reporter activities and/or stabilities are not influenced by the product/s of the test sequences. At least 140 3' nucleotides were included for each dinoflagelate construct and 5' boundaries. HEK-293T cells were depleted of polyamines by treatment with the ornithine decarboxylase inhibitor (DFMO: D,L-alpha-difluoromethylornithine) before transfection. Transfectants were then either left untreated or else supplemented with spermidine prior to lysis and dual luciferase assay. Recoding efficiencies were determined by comparing relative luciferase activities (firefly/Renilla) of test constructs against in-frame controls for each construct. Although spermidine did not stimulate stop codon readthrough for *H. triquestra* and *K. brevis*, we did observe 6% of readthrough for *C. fusus* Figure 4(a), which corresponded to a threefold increase in stimulation. In contrast, we observed robust polyamine-induced +1 frameshifting for both *C. gloeocystis* and *C. coeruleus* Figure 4(a).



```
                       +15nt ((( (   (((       ) ))  ))) ) +63nt   ( ((( ((( (((     ((( ((. ...       . ))) ))) ))) ))) )))
Symbiodinium     TGA CTC ... GGA GAG GCC GAC GAG GTC CTC CAG ... TGC CCT CCC CCA --- CAT GCT TCG --- --- GCG TGT GGG GGA GGG GAC
Karenia          TGA CTT ... GGT GAG GCC GTT GAG GCT CTG CAG ... TGC CCT CCT CCA --- CGT GCT TCA --- --- GCG CGT GGG GGA GGG CCA
Dinophysis       TGA CTC ... GGG GAG GCC GAC GAG GCA CTC CGC ... TGC CCT CCC CCA --- GAC GGT CCG CCC TGG CCG CCT GGG GGA GGG ATC
Ceratium         TGA CTC ... GGA GAG GCT GAT GAG GCT CTC CGG ... TAC CCT CCC CCG --- CGT GCT TCT --- --- GCG CGT GGG GGA GGG GTC
Pyrodinium       TGA CTC ... GGA GAA GCC GAT GAA GCT CTC CGG ... GGG CCC CCC CCC --- --- TTC GCC GCT TTG GCG TTT GGG GGA GGG CAC
Lingulodinium    TGA CTC ... GGA GAC GCC GAC GAC TCG CTG CAG ... GGC CCT CCC CCG --- CGC GCC TTG --- --- GCG CGT GGG GGA GGG ATC
Gonyaulax        TGA CTC ... GGT GAG GCC GAT GAG GCA CTT CAG ... GGC CCT CCC CCG --- CGC GCA TTC --- CTA GTG CGT GGG GGA GGG GTC
Scrippsiella     TGA CTC ... GGA GAG GCT GAT GAA GCT CTC CAG ... GGC CCT CCC CCG --- CAT GCT CCG --- --- GTG TGC GGG GGA GGG ATC
Durinskia        TGA CTC ... GGA GAT GCC GAT GAG GCA CTC CAG ... GGC CCT CCC CAG --- CGT GCT TCG --- --- GTG CGT TGG GGA GGG CGT
Polarella        TGA CTT ... GGG GAG GCC CAA GAG GCC CTT CAG ... TGC CCT CCC CTG --- CGT GGA TGG --- --- AAG CGT AGG GGA GGG ATC
Glenodinium      TGA CTC ... GGG GAG GCC GAT GAG GCA CTC CAA ... CGC CCT CCC CTG --- CGT GCT CCG --- --- GTG CGT GGG GGA GGG GTC
Heterocapsa      TGA CTC ... GGA GAT GCC GAT GAG GCG CTC CGT ... GGC CCC CCT CTG --- CTC GCC GGT TCG CTG GCC AGT GGG GGG GGG TCC
Brandtodinium    TGA CTT ... GGA GAG GCC GAT GAG GCA CTC CGG ... TGC CCT CCC CTG --- CAC GCT TCG --- --- GCG CGT GGG GGA GGG ATC
Protoceratium    TGA CTC ... GGG GAG GCC GAT GAG TCG CTG CAG ... GGC CCT CCC CCA CCG CGC GCC TCA --- CTG GCG CGC GGG GGA GGG ATC
Pelagodinium     TGA CTC ... GGA GAG GCC GAC GAG GCC CTC CAA ... TGC CCT CCT CCG --- TGT GCT TCT --- --- GTG CGC GGG GGA GGG GCC
Kryptoperidinium TGA CTC ... GGG GAG GCC GAT GAG GCA CTC CAG ... AGC CCT CCC CTG --- CGT GCT TCG --- --- GTG CGT GGG GGA GGG GAT
Amphidinium      TGA CTC ... GCG GAG GCC GAA GAG GCA CTT CGC ... --- --- --- --- --- TCA GCG AGC --- GTG GCA CAA ACT GGC AAT GCA
Azadinium        TGA CTC ... GGA GAG GCC GAC GAG GCC CTC CGC ... CGC CCT CCC CCG --- CAT GCC TTA --- TTG GCG TGT GGG GGA GGG ATT
Peridinium       TGA ATC ... GGT GAG GCA GAT GAG CCT CTC CAG ... TGC CCT CCC CTG --- CGT GCT TCG --- --- GCG TGT GGG GGA GGG GTA
Prorocentrum     TGA CTC ... GCC GAC GGC GCC GAG TGC CTC CGG ... GGC CCT CCC CCG --- GGC GGC TCG --- --- CCG TCC GGG GGA GGG TCT
Alexandrium      TGA CTC ... GGA GAG GCC GAT GAG CCT CTT CGG ... GGC CCT CCC CCG --- TGT GCT GTA --- TTG GCG CGC GGG GGA GGG CAC
Gambierdiscus    TGA CTC ... GGA GAT GCT GAT GAT GCT CTC CGC ... GGC CCT CCC CTG --- TGT GCT GCT CTT GCG GTG CAC GGG GGA GGG CAT
Mesodinium       TGA CTC ... GGA GAG GCC GAC GAG GCC CTT CAC ... TGC CCC CCT CCA --- CCC GCC AGC TCG CTG GCG TCT GGG GGG GGG TAC
```

Figure 3. RNA secondary structures. (a) Synonymous site conservation for readthrough coding sequences of dinoflagellate OAZ. Codon index within the RT coding sequence is plotted on the X axis (the codon index starts from ORF1) and the conservation P-value is plotted on the Y axis. The RT site is indicated by a vertical red line. (b) Predicted RNA structures of RT cassettes. Genera are listed on the Y axis. The RT stop codon is indicated in bold, predicted base pairings are indicated with ‘()’s. G:C pairings in the step loop structure are highlighted in red, while A:T are highlighted in yellow and G:T in green.

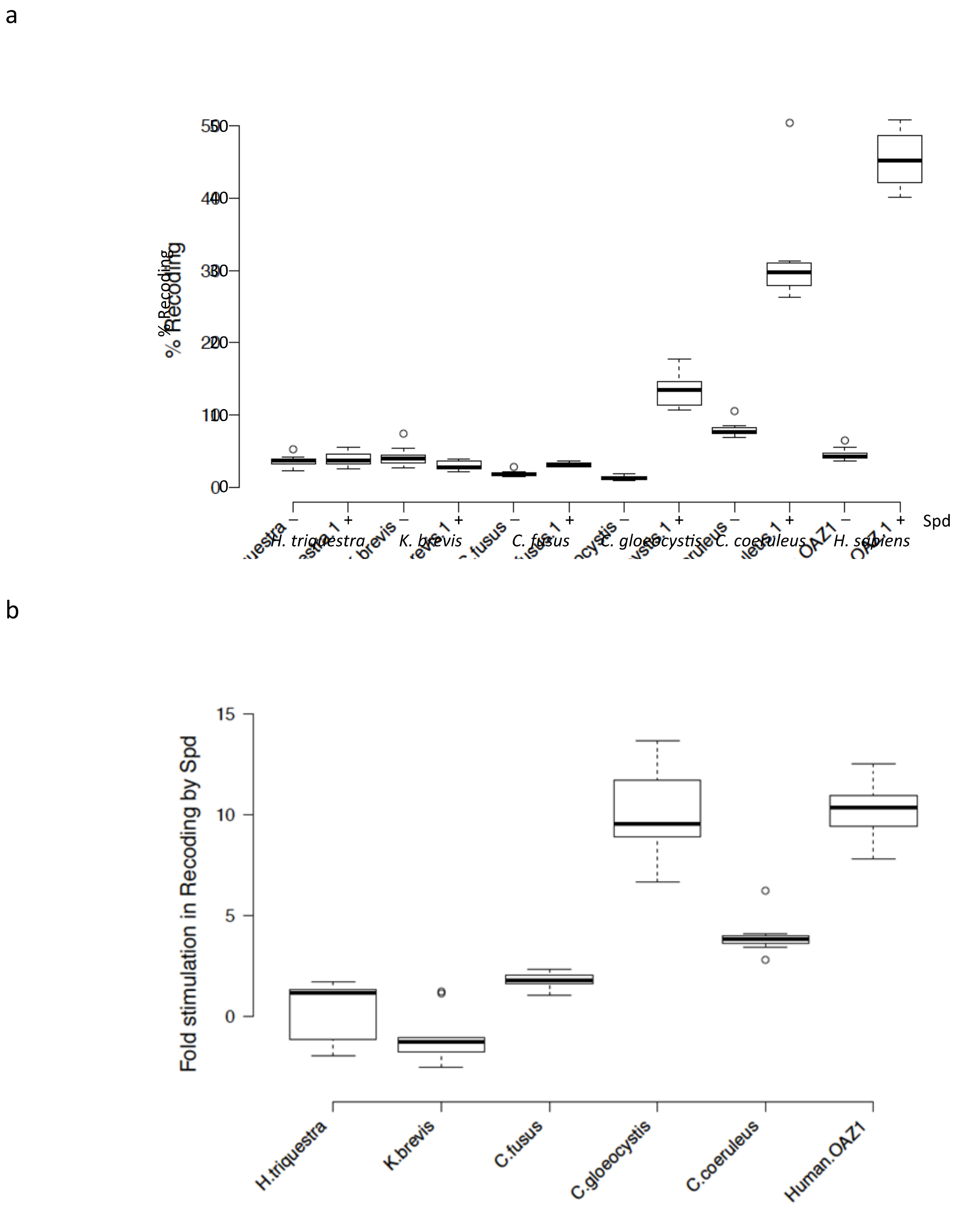


Figure 4. Figure legend – Analysis of dinoflagelate oaz recoding by dual luciferase assay. HEK293T cells were depleted of polyamines by treating with DFMO prior to transfection with plasmids encoding Renilla and firefly luciferases flanking oaz casettes from the indicated dinoflagelates. After 24 h, transfectants were either maintained in DFMO treated media (- spd) or media supplemented with 2 mM spermidine (+ spd). For H. triquestra, K. brevis and C. fusus, ribosomes can only access the second ORF by stop codon readthrough whereas for C. gloeocystis, C. coeruleus and the H. sapiens oaz1 positive control, ribosomes can only access the second ORF by +1 frameshifting. Seperate in-frame control constructs were generated and tested for each test construct and percent recoding determined as

descibed in the materials and methods. Centre lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by dots. n = 12.

Variant eukaryotic release factor 1

In eukaryotes translation termination is facilitated by eukaryotic release factor 1 (eRF1) which recognises all three stop codons (TAA, TAG, and TGA). A variety of cross-linking and mutagenesis studies have identified highly conserved motifs within the N-terminal domain such as TASNIKS, for stop codon recognition, while YxCxxxF and GTS motifs are implicated in purine recognition in the second and third sub-codon positions (Chavatte et al. 2002; Ito et al. 2002; Seit-Nebi et al. 2002; Kolosov et al. 2005; Bulygin et al. 2010; Conard et al. 2012). The structural basis for these motifs were determined using cryo-EM (Brown et al. 2015). A recent study has indicated new individual residues important for TGA recognition (Blanchet et al. 2015).

Employing a similar profile HMM method we identified eRF1 protein sequences for each of the dinoflagellate species. Interestingly upon analysis of the ‘NIKS’ motif we observed in all cases, the ‘N’ was replaced by either a serine (S) or cysteine (C), Figure S5. Where variant genetic codes have been identified, mostly in ciliates, the eRF1 sequences show mutated residues in the conserved motifs (Pánek et al. 2017). Previous studies aiming at identifying how stop codons are recognised by eRF1 in variant genetic codes provide differing observations (Salas-Marco et al. 2006; Lekomtsev et al. 2007; Vallabhaneni et al. 2009). However, no stop codon reassignments were reported in dinoflagellates (Swart et al. 2016), while the TGA stop codon is observed at a greater frequency (66%) than TAA (14%) and TAG (20%) at ORF stop sites, Figure S5.

Discussion

In this study we provide strong bioinformatics and experimental evidence in support of newly identified antizymes. Recoding of antizyme transcripts during translation occurs in response to elevated levels of cellular polyamines, and the classical +1 frameshifting mechanism was identified in six different phyla. Ciliates display the greatest diversity in antizyme decoding, where we identified single ORF and +1 frameshifting. We propose that the diversity of observed antizyme decoding is related to variant genetic code usage by ciliate species. For antizyme synthesis, a regulatory short ORF, produced when polyamine levels are low may not be possible for ciliates that utilize context specific termination. The majority of ciliates that utilize the standard genetic code appear to recode antizyme transcripts in the classical +1 frameshift manner. The idea of context dependent termination was already proposed for ciliates *Euplotes* and *Condylostoma* (Heaphy et al. 2016; Swart et al. 2016; Lobanov et al. 2017) and for the trypanosome *Blastocrithidia* (Záhonová et al. 2016). It is not clear how single ORF antizymes function in the absence of regulatory frameshifting, but it is possibly regulated at the transcript level.

We observe stop codon readthrough bioinformatically in dinoflagellate species and confirmed experimentally for *C. fusus*. Although the experimental validation showed much higher levels of frameshifting for algae *C. gloeocystis* and *C. coeruleus* than for readthrough of dinoflagellate sequences, we suspect that release factor machinery in HEK-293T cells will out-compete readthrough of ORF1, resulting in termination and synthesis of short ORF1 proteins. The influence of viral-like RNA secondary structures is still to be alluded to. We provide evidence of a highly conserved mutation in the 'NIKS' motif in dinoflagellate eRF1 proteins, a possible contributing factor to the highly conserved TGA readthrough site.

In addition to finding alternative mechanisms of antizyme translation, we identified the protein in a number of algae phyla such as; red algae rhodophyta and microscopic algae glaucocystophyte which belong to the Archaeplastida/plant kingdom. However, we did not identify antizyme in the chlorophyte phylum, a green species of algae which shares a common ancestor with land plants. It is likely that antizyme was lost at this this very early stage of plant development and plants have evolved a different mechanism of polyamine regulation which is still to be identified.

In total we identified antizyme in ~30% of protist transcriptomes, while over 140 species from the Excavata, Stramenopiles and Rhizaria do not contain the protein. These organisms possibly regulate polyamine synthesis through alternative methods or the antizyme sequence in these species is so dramatically different that our methods would not pick up the signal.

Table 1. Summary of 69 identified antizymes including; organism details, HMMER score, length of each ORF, Translation mechanism (+1FS: +1 Frameshift, 1ORF: single open reading frame, RT: readthrough,) Transcript ID and Data source. Species with an '*' we considered as contaminants identified by phylogenetic analysis.

Phylum	Genus & Species	HMMER Score	ORF1 Length	ORF2 Length	Mechanism	Transcript ID	Source
Amoebozoa	Filamoeba nolandi	55.6	33	176	+1 FS	CAMNT_0008574259	Keeling et al. 2014
Chromerida	Chromera velia	36.4	30	335	+1 FS	CAMNT_0027034471	Keeling et al. 2014
Chromerida	Vitrella brassicaformis	41.1	33	425	+1 FS	CAMNT_0044104631	Keeling et al. 2014
Ciliophora	Carchesium polypinum	28.2	87	-	1 ORF	c35565_g1_i1	Feng et al. 2015
Ciliophora	Paralembus digitiformis	31	190	-	1 ORF	c16049_g1_i1	Feng et al. 2015
Ciliophora	Colpoda aspera	38.9	28	162	+1 FS	c7701_g1_i1	Feng et al. 2015
Ciliophora	Campanella umbellaria*	111.5	27	201	+1 FS	c74226_g2_i3	Feng et al. 2015
Ciliophora	Blepharisma japonicum	24	166	-	1 ORF	CAMNT_0049652285	Keeling et al. 2014
Ciliophora	Protocruzia adherens	27.5	35	140	+1 FS	CAMNT_0002344635	Keeling et al. 2014
Ciliophora	Favella ehrenbergii	27.6	-	243	1 ORF	CAMNT_0028240967	Keeling et al. 2014
Ciliophora	Anophryoides haemophila	32.3	181	-	1 ORF	CAMNT_0051934985	Keeling et al. 2014
Ciliophora	Strombidinopsis acuminata	33.1	238	-	1 ORF	CAMNT_0017742655	Keeling et al. 2014
Ciliophora	Pseudokeronopsis sp.OXSA	33.4	206	-	1 ORF	CAMNT_0010847671	Keeling et al. 2014
Ciliophora	Mesodinium pulex*	33.5	17	216	RT	CAMNT_0004588833	Keeling et al. 2014
Ciliophora	Uronema sp.Bbcil	37.3	181	-	1 ORF	CAMNT_0038733395	Keeling et al. 2014
Ciliophora	Tiarina fusus	38.1	41	172	+1 FS	CAMNT_0004720167	Keeling et al. 2014
Ciliophora	Aristerostoma sp. ATCC	38.9	27	165	+1 FS	CAMNT_0001770011	Keeling et al. 2014
Ciliophora	Platyophrya macrostoma	40.6	33	160	+1 FS	CAMNT_0017817839	Keeling et al. 2014
Ciliophora	Pseudocohnilembus persalinus	41.2	183	-	1 ORF	c5048_g1_i1	Xiong et al. 2015
Ciliophora	Pseudokeronopsis sp.OXSA	43.7	210	-	1 ORF	CAMNT_0010846041	Keeling et al. 2014
Ciliophora	Platyophrya macrostoma	46.4	34	160	+1 FS	CAMNT_0017806185	Keeling et al. 2014
Ciliophora	Tiarina fusus*	53.7	28	129	+1 FS	CAMNT_0004750939	Keeling et al. 2014
Ciliophora	Tiarina fusus*	60	41	241	+1 FS	CAMNT_0004757239	Keeling et al. 2014
Ciliophora	Myrionecta rubra*	62	47	150	+1 FS	CAMNT_0021103209	Keeling et al. 2014
Ciliophora	Strombidinopsis acuminate*	79.2	143	-	1 ORF	CAMNT_0017760475	Keeling et al. 2014
Ciliophora	Stylonychia lemnae	46		-	1 ORF	c42594_g1_i1	Mu et al. 2016
Ciliophora	Tetrahymena elliotti	49.5	194	-	1 ORF	c34953_g1_i1	PRJNA167917
Ciliophora	Litonotus pictus	20.6	26	162	+1 FS	CAMNT_0010803269	Keeling et al. 2014
Ciliophora	Climacostomum virens	27.7	134	-	1 ORF	CAMNT_0052081759	Keeling et al. 2014
Cryptophyta	Cryptomonas paramecium	55.2	40	142	+1 FS	CAMNT_0000629511	Keeling et al. 2014
Cryptophyta	Goniomonas pacifica	57.2	55	129	+1 FS	CAMNT_0017255013	Keeling et al. 2014
Cryptophyta	Hemiselmis andersenii	72.3	40	150	+1 FS	CAMNT_0001218315	Keeling et al. 2014
Cryptophyta	Geminigera cryophila	78	47	150	+1 FS	CAMNT_0021191549	Keeling et al. 2014
Cryptophyta	Rhodomonas lens	81.6	43	145	+1 FS	CAMNT_0026626701	Keeling et al. 2014
Cryptophyta	Chroomonas mesostigmatica	81.8	39	147	+1 FS	CAMNT_0053642491	Keeling et al. 2014
Cryptophyta	Guillardia theta	82.6	48	149	+1 FS	CAMNT_0000746277	Keeling et al. 2014
Cryptophyta	Hanusia phi	86.3	48	149	+1 FS	CAMNT_0009611707	Keeling et al. 2014
Cryptophyta	Proteomonas sulcata	86.7	43	145	+1 FS	CAMNT_0026626701	Keeling et al. 2014
Dinoflagellata	Noctiluca scintillans	42.7	229	-	1 ORF	CAMNT_0039346585	Keeling et al. 2014
Dinoflagellata	Amphidinium carterae	47.3	28	213	RT	CAMNT_0017964279	Keeling et al. 2014
Dinoflagellata	Symbiodinium StrainC1	55.8	16	214	RT	CAMNT_0049187311	Keeling et al. 2014
Dinoflagellata	Pyrodinium bahamense	57	17	219	RT	CAMNT_0020738001	Keeling et al. 2014
Dinoflagellata	Gambierdiscus australes	64.4	17	224	RT	CAMNT_0011552387	Keeling et al. 2014
Dinoflagellata	Karlodinium micrum	65	221	-	1 ORF	CAMNT_0009192537	Keeling et al. 2014
Dinoflagellata	Dinophysis acuminata	67	13	211	RT	CAMNT_0021007677	Keeling et al. 2014
Dinoflagellata	Ceratium fusus	67.1	17	219	RT	CAMNT_0013653805	Keeling et al. 2014
Dinoflagellata	Togula jolla	67.8	220	-	1 ORF	CAMNT_0010981165	Keeling et al. 2014
Dinoflagellata	Durinskia baltica	68	18	207	RT	CAMNT_0010483421	Keeling et al. 2014
Dinoflagellata	Protoceratium reticulatum	68.3	17	215	RT	CAMNT_0008499395	Keeling et al. 2014
Dinoflagellata	Alexandrium minutum	69.9	20	218	RT	CAMNT_0000786083	Keeling et al. 2014
Dinoflagellata	Prorocentrum minimum	70	17	216	RT	CAMNT_0052904047	Keeling et al. 2014
Dinoflagellata	Brandtodinium nutriculum	71.6	19	206	RT	CAMNT_0044221217	Keeling et al. 2014
Dinoflagellata	Lingulodinium polyedra	76.4	17	217	RT	CAMNT_0006338445	Keeling et al. 2014
Dinoflagellata	Polarella glacialis	77.2	16	215	RT	CAMNT_0002482521	Keeling et al. 2014
Dinoflagellata	Azadinium spinosum	77.8	19	217	RT	CAMNT_0022492459	Keeling et al. 2014
Dinoflagellata	Karenia brevis	78.1	22	223	RT	CAMNT_0014679917	Keeling et al. 2014
Dinoflagellata	Peridinium aciculiferum	78.3	17	204	RT	CAMNT_0025617643	Keeling et al. 2014
Dinoflagellata	Heterocapsa triquestra	78.8	17	213	RT	CAMNT_0040099907	Keeling et al. 2014
Dinoflagellata	Scrippsiella trochoidea	82.9	17	209	RT	CAMNT_0002639107	Keeling et al. 2014
Dinoflagellata	Pelagodinium beii	83.8	15	218	RT	CAMNT_0043231573	Keeling et al. 2014
Dinoflagellata	Glenodinium foliaceum	84.8	17	204	RT	CAMNT_0007781229	Keeling et al. 2014
Dinoflagellata	Kryptoperidinium foliaceum	89.7	17	204	RT	CAMNT_0017406019	Keeling et al. 2014
Dinoflagellata	Gonyaulax spinifera	93.3	17	215	RT	CAMNT_0043521955	Keeling et al. 2014
Glaucocystophyte	Gloeochaete wittrockiana	76.6	34	162	+1 FS	CAMNT_0026656637	Keeling et al. 2014
Glaucocystophyte	Cyanoptyche gloeocystis	78.1	35	112	+1 FS	CAMNT_0041987049	Keeling et al. 2014
Rhodophyta	Timspurckia oligopyrenoides	51	109	205	+1 FS	CAMNT_0024647303	Keeling et al. 2014
Rhodophyta	Madagascaria erythrocladiodes	55.3	86	159	+1 FS	CAMNT_0044076049	Keeling et al. 2014
Rhodophyta	Rhodosorus marinus	59	75	159	+1 FS	CAMNT_0000965445	Keeling et al. 2014
Rhodophyta	Compsopogon caeruleus	62.7	106	164	+1 FS	CAMNT_0027127547	Keeling et al. 2014

Materials and Methods

Data Sources and Assembly

We obtained 196 assembled transcriptomes of protist genera from The Marine Microbial Eukaryote Transcriptome Sequencing Project (Keeling et al. 2014). Additionally we downloaded transcriptomics data for species representing four additional genera from (Feng et al. 2015) and one each from (Kodama et al. 2014; Xiong et al. 2015; Mu et al. 2016; Lobanov et al. 2017; Roy et al. 2017), summarized in Table 1 and assembled each transcriptome de novo using Trinity version r20140413p1 (Haas et al. 2014).

HMM profile of OAZ and identification of its homologs

Using a collection of 218 OAZ amino acid sequences assembled by (Bekaert et al. 2008), we generated a multiple sequence alignment (MSA) with Clustal Omega (Sievers et al. 2014). This alignment was then used to build a profile HMM with HMMER 3.1b2 (Finn et al. 2011). Using default settings (E-value = 10) of hmmsearch we searched each transcriptome (conceptually translated in all 6 frames) for sequence motifs matching OAZ profile HMM. We performed a total of five iterative searches of each transcriptome generating a new profile HMM from new ORFs containing the OAZ signal with a HMMER score of ≥ 20. The scores obtained with the last OAZ HMM profile are shown for each phylum in Figure S1.

In order to determine if sequences were likely to be antizymes, especially low scoring candidates we searched each protein sequence using phmmer (hmmer.org), for the presence of an OAZ pfam domain or where the query sequence matched a target sequence described as an '*Ornithine decarboxylase antizyme protein*'. We also performed blastx searches of the candidate nucleotide sequences to identify the optimal coding sequence and functional domains. We provide a summary of 69 sequences with evidence of an OAZ domain in Table 1 and 26 sequences which we identified containing non-OAZ domain proteins or uncharacterised proteins are tabulated in Table S1.

We defined the protein sequence at this stage as an ORF containing the OAZ signal from a stop codon to stop codon on the mRNA transcript. Taking the OAZ containing ORFs we built a MSA using Clustal Omega (Sievers et al. 2014) of the 69 newly identified antizymes and built a sequence logo using WebLogo (Crooks et al. 2004) and compared the C-terminal regions to the OAZ domain of the ciliate *Tetrahymena thermophilla* amino acid positions 95-192 as per Pfam (Finn et al. 2016), Figure S2.

Taking the 69 protein sequences we carried out maximum likelihood phylogenetic analysis using MEGA6 (Tamura et al. 2013) and built a tree of molecular phylogeny using Jones-Taylor-Thornton (JTT) model of amino acid substitution and bootstrap method for test of phylogeny with 100 bootstrap replications. We included the OAZ sequences from; *Tetrahymena thermophilla, Schizosaccharomyces pombe* and *Homo sapiens OAZ1*.

Contaminant sequence identification

We carried out additional searches of potential contaminant sequences, identified from Figure S3. We performed PSI-BLAST searches of the sequences *Campanella umbellaria* (Ciliate), *Tiarina fusa1* (Ciliate), *Goniomonas pacifica* (Cryptophyta), *Filamoeba nolandi* (Amoebozoa), *Protocruzia adherens* (Ciliate) as they do not branch with the other sequences from their phyla, we looked for the presence of *T.thermophilla* in the results list as a reference organism, we reported the highest scoring PSI-BLAST hit. We performed four pairwise alignments using EMBOSS Needle (Protein Alignment) (McWilliam et al. 2013) of the protein sequences between potential contaminants and their closest branch neighbour. 1. (*Tiarina fusa2* and *Rhodomonas lens)*, 2. (*Myrionecta rubra* and *Geminigera cryophila)*, 3. (*Strombidinopsis acuminata1* and *Heterocapsa triquestra*) and 4. (*Mesodinium pulex* and *Heterocapsa triquestra*)

Frame extensions

Using a custom Python script we simulated translation initiation from the most 5' ATG codon of each mRNA transcript to the next in-frame stop codon (TAA, TAG or TGA). From here we then calculated the distance measured in codons from the position of the 'T', at the ORF1 stop codon to the next in-frame stop. For ciliates with ambiguous genetic codes we altered the script accordingly to recognise the appropriate termination codons. From this output we plotted all extensions for 63 OAZ transcripts and built a tree of molecular phylogeny, same methods as above.

Multiple sequence alignments

We built multiple sequence alignments of the transcript coding regions, initiating at the most 5’ ATG on the mRNA, using MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons, (Ranwez et al. 2011). We lowered the parameters for the cost of a frameshift to -10 (from -30) and cost of a stop codon to -10 (from -100), with all other inputs remaining at the default settings. For the alignments of +1 frameshift candidates we removed the ‘T’ at the first sub-codon position of the ORF1 stop codon to maintain the reading frame for ease of alignment, the location is marked with a red line in Figure 2(a). The alignments are visualized using Jalview2 (Waterhouse et al. 2009).

Secondary structure identification

Using the multiple sequence alignment output from MACSE above we looked for synonymous site conservation in coding regions of all the frameshifting candidates and all individual phyla. We used SynPlot2 (Firth 2014) with default sliding window size of 25 codons. From this output we used IPknot with default settings, searching for nested pseudoknots, McCaskill scoring model (Sato et al. 2011), which looks for a consensus secondary structure when a multiple sequence alignment is provided.

Plasmids

Dinoflagelate oaz sequences were synthesized by IDT as G blocks (see Supplementary Data 1 for sequences), digested with incorporated 5’ XhoI and 3’ BglII restriction sites and cloned into PspXI / BglII digested pSGDluc (PMID:28442579). Human oaz1 was PCR amplified from human genomic DNA with primers incorporating flanking 5’ XhoI and 3’ BglII restriction sites then cloned into PspXI / BglII digested pSGDluc. All constructs were verified by sequencing.

Cell Culture and Transfections

HEK-293T cells (ATCC) were maintained in DMEM supplemented with 10% FBS, 1 mM L-glutamine and antibiotics. For polyamine depletion experiments, 4 x 106 cells were plated in a 10 cm petri dish in medium supplemented with 2.5 mM □-difluoromethylornithine (DFMO; a kind gift from P. Woster via Dr. Michael Howard, University of Utah). Cells were incubated for 5 days in DFMO-supplemented media at 37°C in 5% CO2 and then transfected using Lipofectamine 2000 reagent (Invitrogen) and the one-day protocol in which suspended cells are added directly to the DNA complexes in white half-area 96-well plates (Costar). For each transfection the following were added to each well: 25 ng of each plasmid plus 0.2 µl Lipofectamine 2000 in 25 µl Opti-Mem (Gibco). The transfecting DNA complexes in each well were incubated with 2.5 × 104 cells suspended in 25 µl DFMO-supplemented media. Transfected cells were incubated at 37°C in 5% CO2 for 18 h before the addition of 50 µl of media supplemented to final concentrations of 2.5 mM DFMO, 1 mM aminoguanidine hydrochloride (Sigma), or the same plus 2 mM spermidine (Sigma) for a further 24 h.

Dual Luciferase Assay

Firefly and Renilla luciferase activities were determined using the Dual Luciferase Stop & Glo® Reporter Assay System (Promega). Relative light units were measured on a Veritas Microplate Luminometer with two injectors (Turner Biosystems). Transfected cells were lysed in 12.6 µl of 1 × passive lysis buffer and light emission was measured following injection of 25 µl of either Renilla or firefly luciferase substrate. Recoding efficiency (% Recoding) was determined by comparing the relative luciferase activities (firefly/Renilla) of test constructs to their respective in-frame controls.

eRF1 identification and stop codon frequencies

Using a hmmer based method described above, we built a profile hmm of eRF1 sequences obtained from (Pánek et al. 2017) and searched each of the dinoflagellate transcriptomes. We generated a MSA using Clustal Omega (Sievers et al. 2014) and generated a sequence logo using WebLogo (Crooks et al. 2004) of the N-terminal domain, including the GTS, TASNIKS, YxCxxxF and GGQ motifs.

To investigate stop codon frequencies of unannotated transcriptomes, we searched for open reading frames ≥ 600 nts, that begin with an ATG and stop at either, TAA, TAG or TGA. We then calculated the frequency of each stop observed as a percentage (%), we performed this analysis for transcriptomes of; *C. fusus, K. brevis, T. elliotti, C. gloeocystis* and *C. coeruleus*, for T. elliotti we changed the settings for reassignments of stop codons TAA and TAG. We used the Galaxy Project platform to perform the analysis (Afgan et al. 2016).

Supplementary Figures

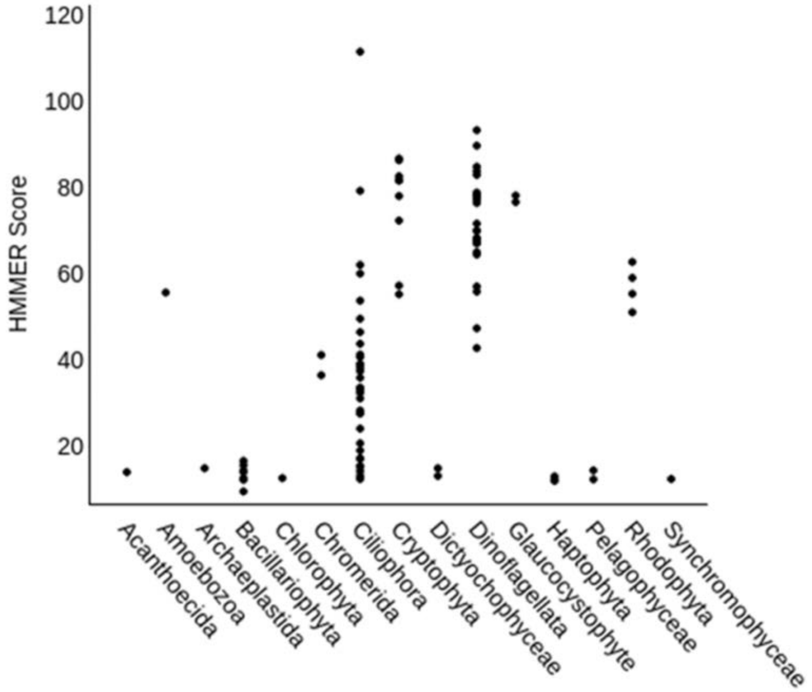


Figure S1. The distribution of OAZ profile HMM scores (Y axis) for protein sequences resulted from conceptual translation of 95 transcripts grouped by origin phylum (X-axis).

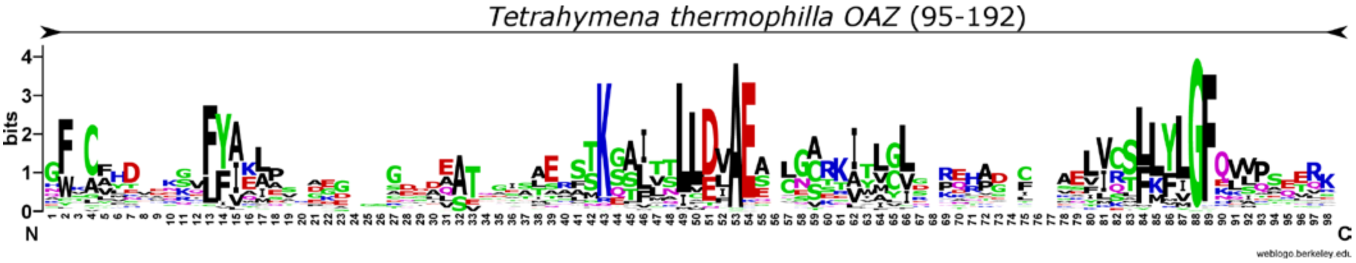


Figure S2. Sequence logo representation of the region corresponding to the OAZ signal for 69 sequences we identified as antizymes. Compared to *Tetrahymena thermophilla*, coordinates 95-192 refer to the location of the OAZ domain according to Pfam.

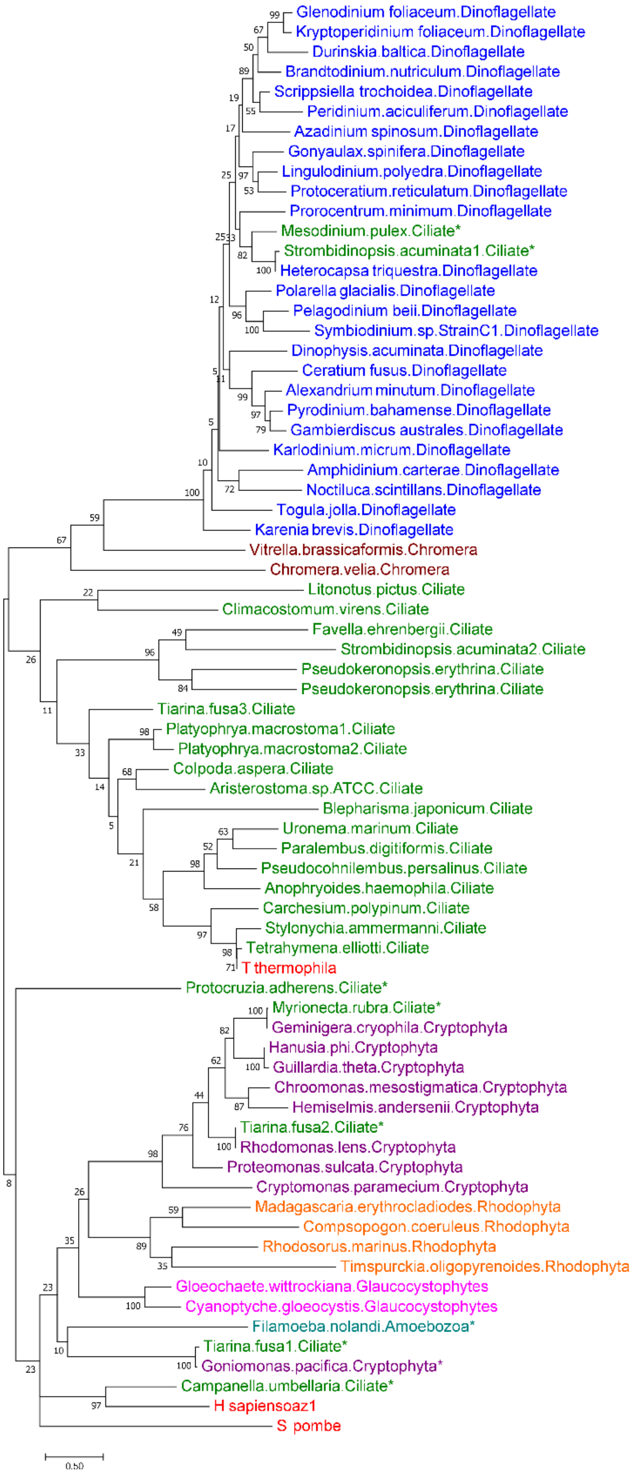

Glenodinium foliaceum.Dinoflagellate
Kryptoperidinium foliaceum.Dinoflagellate
Durinskia.baltica.Dinoflagellate
Brandtodinium.nutriculum.Dinoflagellate
Scrippsiella trochoidea.Dinoflagellate
Peridinium.aciculiferum.Dinoflagellate
Azadinium spinosum.Dinoflagellate
Gonyaulax.spinifera.Dinoflagellate
Lingulodinium.polyedra.Dinoflagellate
Protoceratium.reticulatum.Dinoflagellate
Prorocentrum.minimum.Dinoflagellate
Mesodinium.pulex.Ciliate*
Strombidinopsis.acuminata1.Ciliate*
Heterocapsa triquestra.Dinoflagellate
Polarella glacialis.Dinoflagellate
Pelagodinium beii.Dinoflagellate
Symbiodinium.sp.StrainC1.Dinoflagellate
Dinophysis.acuminata.Dinoflagellate
Ceratium fusus.Dinoflagellate
Alexandrium minutum.Dinoflagellate
Pyrodinium.bahamense.Dinoflagellate
Gambierdiscus australes.Dinoflagellate
Karlodinium.micrum.Dinoflagellate
Amphidinium.carterae.Dinoflagellate
Noctiluca.scintillans.Dinoflagellate
Togula.jolla.Dinoflagellate
Karenia brevis.Dinoflagellate
Vitrella.brassicaformis.Chromera
Chromera.velia.Chromera
Litonotus.pictus.Ciliate
Climacostomum.virens.Ciliate
Favella.ehrenbergii.Ciliate
Strombidinopsis.acuminata2.Ciliate
Pseudokeronopsis.erythrina.Ciliate
Pseudokeronopsis.erythrina.Ciliate
Tiarina.fusa3.Ciliate
Platyophrya.macrostoma1.Ciliate
Platyophrya.macrostoma2.Ciliate
Colpoda.aspera.Ciliate
Aristerostoma.sp.ATCC.Ciliate
Blepharisma.japonicum.Ciliate
Uronema.marinum.Ciliate
Paralembus.digitiformis.Ciliate
Pseudocohnilembus.persalinus.Ciliate
Anophryoides.haemophila.Ciliate
Carchesium.polypinum.Ciliate
Stylonychia.ammermanni.Ciliate
Tetrahymena.elliotti.Ciliate
T thermophila
Protocruzia.adherens.Ciliate*
Myrionecta.rubra.Ciliate*
Geminigera.cryophila.Cryptophyta
Hanusia.phi.Cryptophyta
Guillardia.theta.Cryptophyta
Chroomonas.mesostigmatica.Cryptophyta
Hemiselmis.andersenii.Cryptophyta
Tiarina.fusa2.Ciliate*
Rhodomonas.lens.Cryptophyta
Proteomonas.sulcata.Cryptophyta
Cryptomonas.paramecium.Cryptophyta
Madagascaria.erythrocladiodes.Rhodophyta
Compsopogon.coeruleus.Rhodophyta
Rhodosorus.marinus.Rhodophyta
Timspurckia.oligopyrenoides.Rhodophyta
Gloeochaete.wittrockiana.Glaucocystophytes
Cyanoptyche.gloeocystis.Glaucocystophytes
Filamoeba.nolandi.Amoebozoa*
Tiarina.fusa1.Ciliate*
Goniomonas.pacifica.Cryptophyta*
Campanella.umbellaria.Ciliate*
H sapiensoaz1
S pombe
0.50

Figure S3. Phylogenetic positions of 69 OAZ sequences identified. Constructed from HMMER derived OAZ protein sequences, for each species. The maximum likelihood (ML) tree was inferred with MEGA6, using Jones-Taylor-Thornton (JTT) substitution model, with bootstrap values at each branch. Included are three reference OAZ sequences highlighted in red; *Tetrahymena thermophilla*, *Schizosaccharomyces pombe*, and *Homo sapiens OAZ1*. Sequences derived from each phyla are colour coded; blue = Dinoflagellate, maroon = Chromera, green = Ciliate, purple = Cryptophyta, orange = Rhodophyta, fuchsia = Glaucocystophytes and teal = Amoebozoa. Possible contaminants are indicated with an '*'

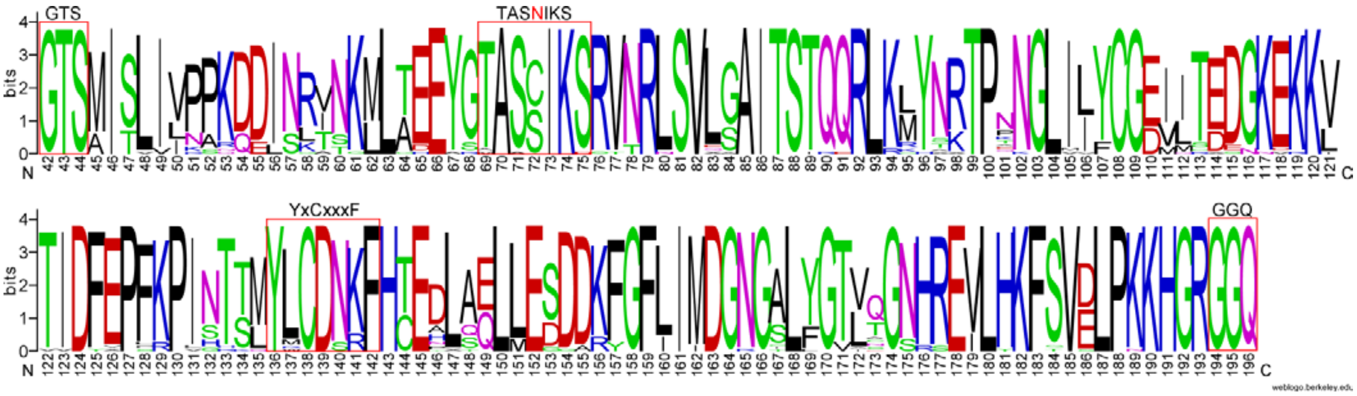


Figure S4. **Sequence logo representation of dinoflagellate eRF1 N-terminal domains.** Highly conserved motifs GTS, TASNIKS, YxCxxxF and GGQ are indicated in red boxes. The mutated 'N' residue of TASNIKS is highlighted in red indicating a substitution to either C or S in each species.

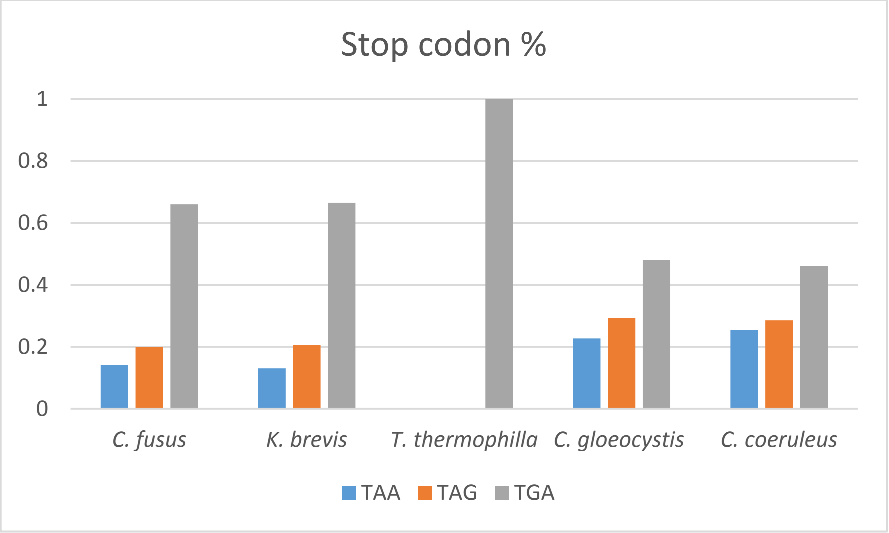


Figure S5. **Stop codon frequencies of selected organisms.** Dinoflagellates *C. fusus* & *K. brevis*, ciliate *T. thermophilla* and algae *C. gloeocystis* and *C. coeruleus*.

Table S1. **Summary non-antizyme sequences**. Blastx output indicates their likely protein product.

Phylum	Genus & Species	HMMER Score	blastx	ORF Length	Accession No.	Source
Acanthoecida	Acanthoeca-like.sp	13.9	-	339	CAMNT_0025061575	Keeling et al 2014
Archaeplastida	Palpitomonas bilix	14.8	Asparaginyl beta-hydroxylase	581	CAMNT_0000852767	Keeling et al 2014
Bacillariophyta	Minutocellus polymorphus	9.4	Ig-like_fold	52	CAMNT_0023117269	Keeling et al 2014
Bacillariophyta	Grammatophora oceanica	12.1	Dimer_Tnp_hAT	407	CAMNT_0038674929	Keeling et al 2014
Bacillariophyta	Skeletonema costatum	12.5	-	41	CAMNT_0000281321	Keeling et al 2014
Bacillariophyta	C. paradoxa	13.9	HATPase_c	293	CAMNT_0043789695	Keeling et al 2014
Bacillariophyta	Ditylum brightwellii	15.5	-	118	CAMNT_0051839761	Keeling et al 2014
Bacillariophyta	Rhizosolenia setigera	16.9	-	76	CAMNT_0020663699	Keeling et al 2014
Chlorophyta	Tetraselmis striata	12.5	-	41	CAMNT_0026121605	Keeling et al 2014
Ciliophora	Condylostoma magnum	12.3	-	115	CAMNT_0008316563	Keeling et al 2014
Ciliophora	Paramecium bursaria	13	Ubiquitin-like	173	c29182_g1_i1	Kodama et al. 2014
Ciliophora	Climacostomum virens	13.2	TPR_repeat	138	CAMNT_0052074341	Keeling et al 2014
Ciliophora	Euplotes crassus	14.1	Ubiquitin hydrolase	389	comp8579_c0_seq1	Lobonov et al 2017
Ciliophora	Fabrea salina	15.2	-	450	c3183_g1_i1	Keeling et al 2014
Ciliophora	Paramecium tetraurelia	15.3	-	216	c23414_g1_i1	SAMEA2053275
Ciliophora	Climacostomum virens	17	Putative ankyrin repeat	134	CAMNT_0052081759	Keeling et al 2014
Ciliophora	Stentor coeruleus	17	-	149	c47594_g1_i1	Slabodnick et al. 2017
Ciliophora	Blepharisma japonicum	18.9	ABC_permease	162	CAMNT_0049652285	Keeling et al 2014
Dictyochophyceae	Pseudopedinella elastica	13	-	79	CAMNT_0013398503	Keeling et al 2014
Dictyochophyceae	Florenciella parvula	14.8	elongation factor EF1B	132	CAMNT_0007549543	Keeling et al 2014
Haptophyta	Scyphosphaera apsteinii	11.9	FG-GAP_2	147	CAMNT_0007332251	Keeling et al 2014
Haptophyta	Phaeocystis.sp	12	-	51	CAMNT_0006772947	Keeling et al 2014
Haptophyta	Prymnesium parvum	12.9	He_PIG	246	CAMNT_0026721579	Keeling et al 2014
Pelagophyceae	Pelagococcus subviridis	12.2	IGPS	330	CAMNT_0034305685	Keeling et al 2014
Pelagophyceae	Chrysoreinhardia sp	14.3	-	108	CAMNT_0006835463	Keeling et al 2014
Synchromophyceae	Synchroma pusillum	12.3	HSF-type DNA-binding	213	CAMNT_0044162495	Keeling et al 2014

Supplementary Data 1 - Sequences of dinoflagelate and human *oaz* casettes cloned into pSGDLuc. *Xho*I sites are highlighted in light blue, *Bgl*II sites are highlighted in teal and stop codons highlighted in purple. In-frame controls were generated by standard two step PCR using primers that changed the TGA stop codon to either TGG for the readthrough candidates (*Heterocapsa. triquestra*, *Karenia. brevis* and *Ceratium. fusus*) or GA for the +1 frameshifting candidates (*Cyanoptyche. Gloeocystis*, *Compsopogon. Coeruleus* and *Homo. sapiens*).

>Heterocapsa. triquestra

```
CTCGAGACTACAGTACGGACGTAAGCAGAGTAATGGCCGAGTGCTTGATACGGTTTGACTCGCAGACGGTGCTGT
CGGAGATGCCGATGAGGCGCTCCGTGTTGCTGGCTTCGTTCTCACCCAGGAGGTGGACGAGGGTGCCGATGACGA
GGATGGAAGTCTCGGCCCCCCTCTGCTCGCCGGTTCGCTGGCCAGTGGGAGATCT
```

>Karenia. brevis

```
CTCGAGACAAAAATTTGAAAGGCGACAGGTGGGAAGTCGAGTCAGTGATACTGTATCGCTTCCTATGGTATGACT
TCCTATGGAATGCGCCGGTGAGGCCGTTGAGGCTCTGCAGATCACCGGCTTCGCCCCATCCGGTGAGGACATCGG
ATGTGCTGACGACGAGGATGGCAGTCGCTGCCCTCCTCCACGTGCTTCAGCGCGTGGGGGAGGGCCACCGAGTGT
ACAGGGCAGCGAAAGCTACCGAGCTGTGCCTGTTGAGCTATGGTCGAGATCT
```

>Ceratium. fusus

```
CTCGAGAGTGAAGTATCAGCGAGCTCAGGCGTCTGGCCGGATCAGTGACACGGCCTGACTCCCACCAGTTGAAGC
TGGAGAGGCTGATGAGGCTCTCCGGCGCGCTGGCTTCTGCGGGGATGCACTCGCGGAGGAGGCTGCCGATGATTC
GGATGGGTGTATCTACCCTCCCCCGCGTGCTTCTGCGCGTGGGGGAGGGGTCCAACTCCTTGCCCGAGTCGGTGA
GCGAGCTTCTCTCGTGAATTTGTCGACTGTCAGATCT
```

>Cyanoptyche. gloeocystis

```
CTCGAGAATCCAGAAACAGACTCAGAGATTCGTCAAGACAACGAAGGCGAACTTTGCCAGTGAGCTGAGCATTTT
CTCTCTCACGCGGAAACACCAGTGGTATTTTTGATGTTGGCAACAAGCTGGGGAAGAAAGGATTCCAGAAGCTTC
AGCCCAGCTTGGTCAGCCCTGAAACGGTCTCTGCTGCATCTGTCTTCTACAAGAGGATCTTTGGTCATCTTCTCG
CCGAAGGTGAAGCAGAAGATGCAAAGGTTACAGATGATGTGTCCGTGTTTCAATTGGACATTCAGAACGAGCAGC
AAGAAACCGAGATCT
```

>Compsopogon. coeruleus

```
CTCGAGAGAAGTCTGCGCGTTTGTAAAGTGCGTGAATAATCGAAGGAGGTACAGTGAAAACGTCACTTGCTTTAC
GACGTCGGGTGTGAAAACGGGGTGGTGTTTTTGACGTTCGACCTCTTTTGCTATCGGAGGGATTTGAGGTTTCCG
CTCCAGAGGCCATCTCGGAGAAATTCAGTCTCCTGTTCCGTCACATGGGGGACTCACAATCACATCAAACCCTGA
CCGATTGGAGGGTCGCCTTCCATGTTACGACGACCGATATTGAAAGCGGAGCAAGAGTCTCAGAGTGGACTCTCC
TCATTTCCGAGATCT
```

>Homo. sapiens oaz1

```
CTCGAGGGTCTCCCTCCACTGCTGTAGTAACCCGGGTCCGGGGCCTCGGTGGTGCTCCTGATGCCCCT
CACCCACCCCTGAAGATCCCAGGTGGGCGAGGGAATAGTCAGAGGGATCACAATCTTTCAAGATCT
```

Reassignment

Novel ciliate genetic code variants including the reassignment of all three stop codons to sense codons in *Condylostoma magnum*

This chapter has been published as a Letter Discovery in Molecular Biology and Evolution, Volume 33, Issue 11, 1 November 2016, Pages 2885–2889, https://doi.org/10.1093/molbev/msw166

Abstract

mRNA translation in many ciliates utilises variant genetic codes where stop codons are reassigned to specify amino acids. To characterise the repertoire of ciliate genetic codes we analysed ciliate transcriptomes from marine environments. Using codon substitution frequencies in ciliate protein-coding genes and their orthologs we inferred the genetic codes of 24 ciliate species. Nine did not match genetic code tables currently assigned by NCBI. Surprisingly, we identified a novel genetic code where all three standard stop codons (TAA, TAG, TGA) specify amino acids in *Condylostoma magnum*. We provide evidence suggesting that the functions of these codons in *C. magnum* depends on their location within mRNA. They are decoded as amino acids at internal positions, but specify translation termination when in close proximity to an mRNA 3' end. The frequency of stop codons in protein coding sequences of closely related *Climacostomum virens* suggest that it may represent a transitory state.

Key words: the genetic code, ciliates, translation termination, stop codon reassignment, alternative genetic decoding.

The standard genetic code contains 61 amino acid specifying codons and 3 codons that specify translation termination. It was long considered to be unchangeable and its origin was described as a ‘frozen accident’ (Crick 1968). Since then a number of variant genetic codes have been discovered, and the National Center for Biotechnology Information (NCBI) currently reports 18 additional genetic codes alongside the standard one (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi). The majority of them have been found in mitochondrial and bacterial genomes. The rise of variant genetic codes is due to a change in codon meaning which is referred to as codon reassignment. This phenomenon can occur due to alterations in the components of translation machinery (tRNAs, aminoacyl-tRNA synthetases or release factors), see (Baranov et al. 2015) for a review.

Stop codon reassignments are a particularly common feature of mRNA translation in ciliates (Knight et al. 2001; Lozupone et al. 2001). Species belonging to the genera *Paramecium*, *Tetrahymena* and *Oxytricha* are known to translate TAA and TAG as glutamine (Q) and only recognise TGA as a signal for termination (Horowitz and Gorovsky 1985), while *Blepharisma* translates TGA as tryptophan (W) and recognises TAA and TAG as signals for translation termination (Liang and Heckmann 1993). In *Euplotes*, TGA is reassigned to cysteine (C) (Meyer et al. 1991) and high frequency of +1 frameshifting during mRNA translation occurs at TAA and TAG codons (Klobutcher and Farabaugh 2002; Wang et al. 2016). It has been conjectured recently that *Euplotes* species use additional mechanisms to discriminate between TAA/TAG codons specifying ribosomal frameshifting and termination of translation (Lobanov et al. 2017).

To obtain a more detailed picture of stop codon reassignment events in ciliates, we took advantage of recent advances in large scale sequencing projects. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (Keeling et al. 2014) provides transcriptomic data for over 650 different marine microbes including ciliates. We obtained RNA-seq from 18 different ciliate genera from the MMETSP. In addition, transcriptomics for four additional genera were obtained from (Feng et al. 2015), one from (Kodama et al. 2014) and one from (Lobanov et al. 2017). We assembled each transcriptome de novo using Trinity (Grabherr et al. 2011), see Methods.

Using BLAST (Altschul et al. 1997), we searched each transcript against the NCBI Reference Sequence (RefSeq) protein database with an e-value of 10^{-30} as a threshold for significant sequence similarity for individual transcript hits. Table 1 summarises characteristics of each transcriptome composition and provides information on the number of transcripts with statistically significant similarity hits.

Table 1. Summary of 24 species analysed; including the assembled transcriptome size and the number of significant alignment hits. Comparison between NCBI genetic codes and the genetic codes inferred in this study (separated with /). - refers to no reassignment and '?' shows that the function of the codon cannot be classified based on threshold used in this study. (Q = Glutamine, E = Glutamic Acid, W = Tryptophan, C = Cysteine, Y = Tyrosine)

Genus & Species	Assembled Transcripts	Transcripts E= 10^{-30}	TAA NCBI/Here	TAG NCBI/Here	TGA NCBI/Here	Source
Anophryoides haemophila	14,853	2,189	Q/Q	Q/Q	-/-	Keeling et al. 2014
Aristerostoma sp. ATCC	30,326	3,950	Q/Q	Q/Q	-/-	Keeling et al. 2014
Blepharisma japonicum	32,295	6,392	-/-	-/-	W/W	Keeling et al. 2014
Campanella umbellaria	171,018	16,384	Q/E	Q/E	-/-	Feng et al. 2015
Carchesium polypinum	87,362	8,610	Q/E	Q/E	-/-	Feng et al. 2015
Climacostomum virens	23,177	5,718	-/?	-/?	C/?	Keeling et al. 2014
Colpoda aspera	87,297	9,079	-/-	-/-	C/-	Feng et al. 2015
Condylostoma magnum	29,437	4,510	Q/Q	Q/Q	-/W	Keeling et al. 2014
Euplotes focardii	34,984	3,939	-/?	-/?	C/C	Keeling et al. 2014
Euplotes crassus	33,701	3,619	-/-	-/-	C/C	Lobanov et al. 2017
Fabrea salina	15,706	4,340	-/-	-/-	C/-	Keeling et al. 2014
Favella ehrenbergii	31,448	3,387	Q/Q	Q/Q	-/-	Keeling et al. 2014
Litonotus pictus	30,341	2,692	-/-	-/-	-/-	Keeling et al. 2014
Mesodinium pulex	84,288	7,615	-/Y	-/Y	-/-	Keeling et al. 2014
Myrionecta rubra	40,881	3,579	-/Y	-/Y	-/-	Keeling et al. 2014
Paralembus digitiformis	108,308	5,579	Q/Q	Q/Q	-/-	Feng et al. 2015
Paramecium bursaria	128,693	13,341	Q/Q	Q/Q	-/-	Kodama et al. 2014
Platyophrya macrostoma	46,111	7,407	Q/-	Q/-	-/-	Keeling et al. 2014
Protocruzia adherens	42,999	4,835	Q/-	Q/-	-/-	Keeling et al. 2014
Pseudokeronopsis sp.OXSA	32,771	3,919	Q/Q	Q/Q	-/-	Keeling et al. 2014
Strombidinopsis acuminata	66,812	7,693	Q/Q	Q/Q	-/-	Keeling et al. 2014
Strombidium inclinatum	38,510	3,545	Q/Q	Q/Q	-/-	Keeling et al. 2014
Tiarina fusus	77,484	6,261	Q/Q	Q/?	-/-	Keeling et al. 2014
Uronema sp.Bbcil	14,501	2,843	Q/Q	Q/Q	-/-	Keeling et al. 2014

To infer stop codon reassignment events, we first calculated the density of stop codons in pairwise alignments of conceptually translated ciliate mRNAs (with stop codons translated as an unknown amino acid) for each dataset. Figure 1 shows the densities of each stop codon (see Methods for the description of the pipeline). *Blepharisma* and *Paramecium* were used as reference organisms for determining a threshold for discrimination between stop codons that were reassigned to code for amino acids and stop codons that function as signals for termination. The threshold is shown as a grey shaded area in Figure 1. It can be seen that the distribution of stop codon frequencies is bimodal with a clear separation between two classes. The few stop codons falling into the grey area may represent very recent stop codon reassignments, transitory states, or may correspond to organisms with a large number of pseudogenes in their genomes or frequent utilization of recoding mechanisms in translation of their transcriptomes. Most species have either 1 or 2 stop codons reassigned to amino acids. It is also clear that evolution of TAA and TAG codon meanings is coupled, i.e. if one of these codons is reassigned the other codon is also reassigned. This is most likely because these two codons differ at the wobble position and could be recognized by the same tRNA. A few exceptions where one of these two codons occur in the grey area could be due to inability of the threshold used to provide a clear discrimination (see discussion below). Most striking, however, is that all three stop codons in *Condylostoma magnum* show frequencies indicative of reassignment to sense codons.

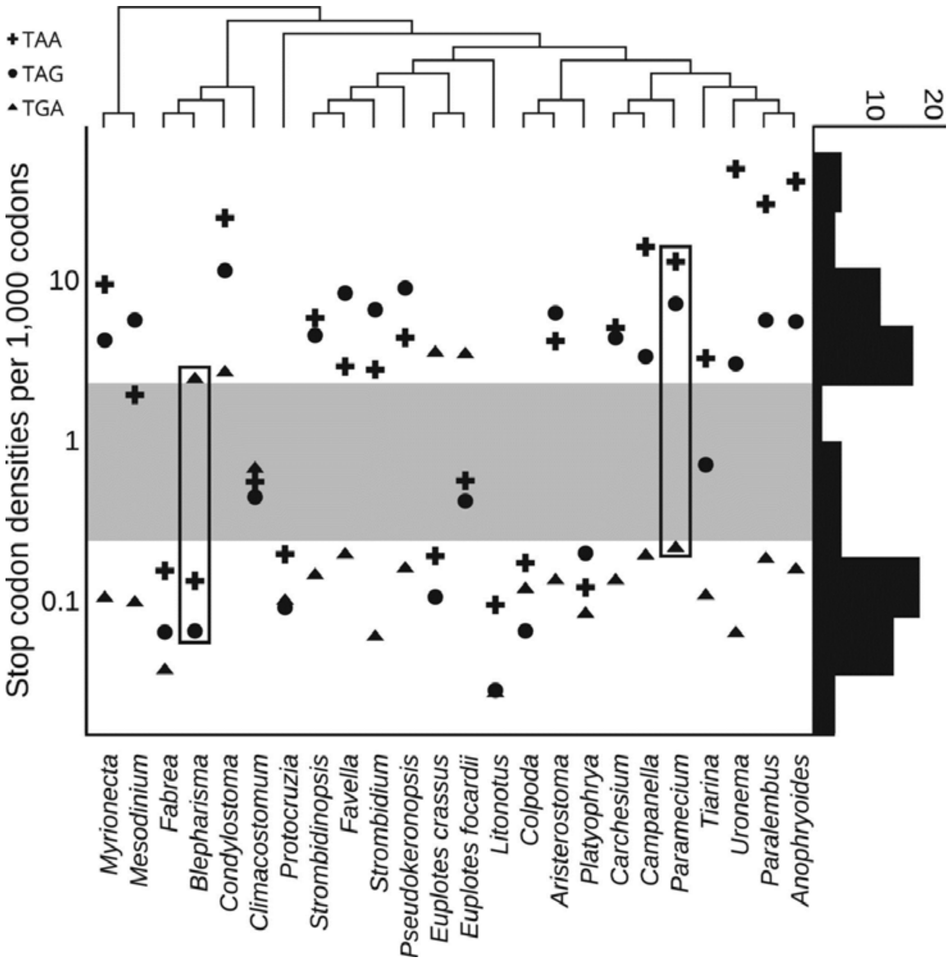


Figure 1. Classification of ciliate stop codons. Stop codon densities (axis y) in protein coding sequences are indicated for each species (bottom). Rectangles specify stop codons of the organisms used for defining a threshold (grey area) for discriminating reassigned codons (above grey area) from those that retained their function as signals for termination (below grey area). The phylogenetic tree constructed with 18S rRNA sequences above indicates the relatedness of each species. The histogram on the right shows distribution of stop codon densities.

To determine the meaning of reassigned stop codons, we evaluated the frequency of amino acid substitutions in pairwise alignments of translated mRNAs and their close homologs from other species. Occasional matching of a ciliate stop codon (functioning as a terminator) to a sense amino acid in a homolog may occur close to N- or C- termini if a ciliate homolog is shorter, in the case of transcribed pseudogenes containing nonsense mutations, when a ciliate transcript contains a sequencing error or when a specific stop codon is recoded to an amino acid in the context of a specific mRNA. However, if a stop codon reassignment took place, it is expected that the reassigned stop codon would frequently match the specific amino acid to which it was reassigned. We provide the total substitution values of all three stop codons for each ciliate in supplementary tables S1-S3. Supplementary Figure S1 shows z-scores of amino acid substitution frequencies for each likely reassigned stop codon. It can be seen that for each reassigned stop codon there is only a single amino acid with exceptionally high Z-score. An even clearer picture is obtained when substitution frequencies are calculated only for amino acid residues evolving under strong stabilizing selection (Figure 2 and Supplementary Table S4).

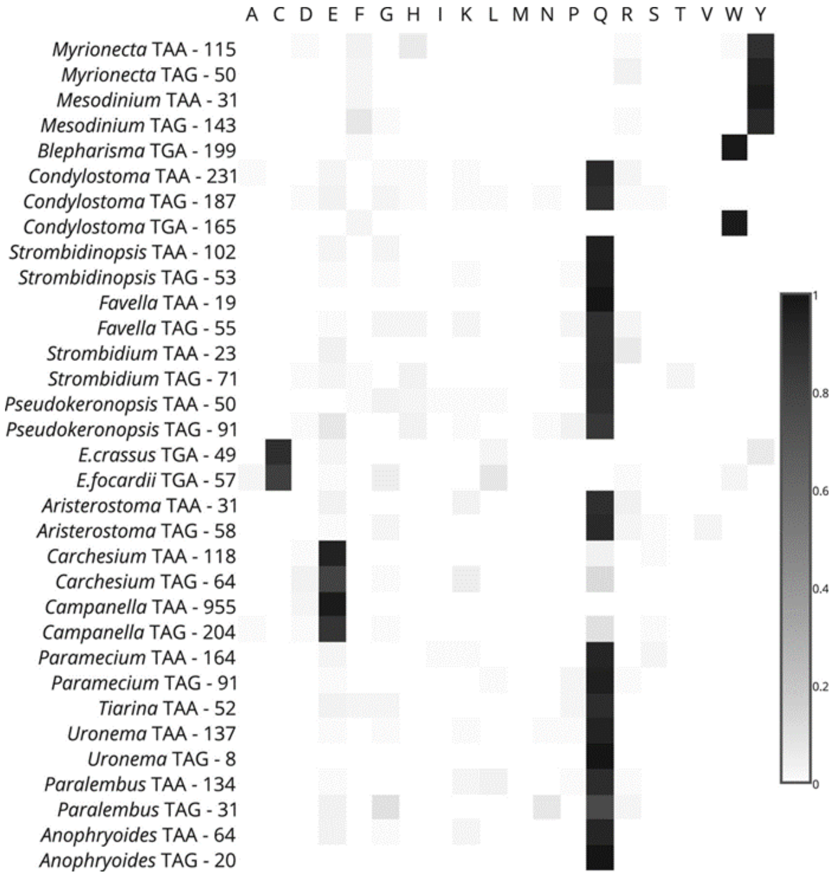


Figure 2. Identification of amino acid specifications of the reassigned codons. Each row corresponds to a single reassigned codon. The organism source of a codon, its identity and the total number of occurrences at highly conserved positions of aligned sequences are indicated on the left. The normalized frequencies of amino acid substitutions are shown as heatmaps.

For *Paramecium* we observe that Q is the most frequently substituted amino acid for both TAA and TAG, and for *Blepharisma* and both *Euplotes* species tryptophan (W) and cysteine (C) are the most frequently substituted amino acids for TGA, respectively. With the same frequency as *Blepharisma* we can clearly see that TGA in *Condylostoma* is likely reassigned to W along with TAA and TAG also reassigned to Q. The specificity of substitutions in *Condylostoma* further supports the notion that all three codons are reassigned in this organism. In addition, we report novel stop codon reassignments in *Mesodinium* and *Myrionecta* where TAA and TAG appear to code for tyrosine (Y). In total, we provide evidence in support of redefining the genetic codes of nine ciliates. Table 1 compares the genetic code of each ciliate species analysed with the NCBI assigned code.

The unclassified, grey shaded region of Figure 1 requires additional attention. It is likely that *Mesodinium* TAA is reassigned to Y. It is very close to the threshold and such reassignment would be consistent with the function of TAG in *Mesodinium*. *Climacostomum* is closely related to *Condylostoma* and may represent a transitory state that potentially could provide an answer to how *Condylostoma* emerged as an organism with the genetic code composed of 64 sense codons. Recently we carried out ribosome profiling analysis of *E. crassus* translatome and mass-spectrometry analysis of its proteome (Lobanov et al. 2017). While the analysis revealed thousands of ribosomal frameshifting occurrences at TAA/TAG codons, it revealed no cases of stop codon readthrough that preserved the frame. As can be seen in Figure 1, the density of TAA/TAG codons is much higher in *E. focardii* than in *E. crassus* and this could be due to potential utilization of stop codon readthrough in addition to ribosomal frameshifting.

Identification of an organism with all stop codons reassigned to sense codons poses a question of how translation termination is accomplished in *Condylostoma*. A theoretical possibility is a regulated termination where stop codon function would depend on specific ligands whose expression is regulated by a specific condition. Such a situation has been observed previously in *Acetohalobium arabaticum*, where the function of TAG codon as a signal for termination or as a codon for pyrrolysine depends on the energy source used by these bacteria (Prat et al. 2012). This, however, seems an unlikely possibility because of very high frequency of stop codons in protein coding genes and tremendous impact of such switches on the whole proteome. An alternative possibility is that the function of stop codon depends on its position within mRNA. Based on our recent characterization of *E. crassus* translatome and proteome (Lobanov et al. 2017) we proposed that the translational machinery of *E. crassus* is able to discriminate stop codons in internal positions of mRNAs from those at the ends and use only the latter for termination of translation. Such a mechanism could also explain the enigmatic reassignment of all three stop codons in *Condylostoma*. To address this possibility, we analysed codon frequencies relative to the expected ends of protein coding regions (CDS). For this purpose, the *Condylostoma* transcriptome was aligned to the most conserved eukaryotic proteins using eukaryotic orthologous groups KOGs (Tatusov et al. 2003). The positions in pairwise alignments matching the stop codons of homologous sequences were considered as expected locations of stop codons in the corresponding *Condylostoma* sequence. Figure 3 shows frequencies of all 64 codons relative to expected CDS ends. It can be seen that stop codons (TAG and TGA in particular) are overrepresented at the expected locations of CDS ends. Importantly, it also can be seen that ~15 nt downstream of expected termination locations there is overrepresentation of AAAs which probably reflects locations of mRNA polyA tails. This suggests that 3’UTRs in *Condylostoma* are very short and conserved and may be implicated in recognition of stop codons signalling for termination of translation. Consistent with this hypothesis, a depletion of stop codons is observed within the upstream ~30 nt of expected locations of termination sites, which probably is due to selection to avoid premature termination due to close location of stop codons to the poly-A tails.

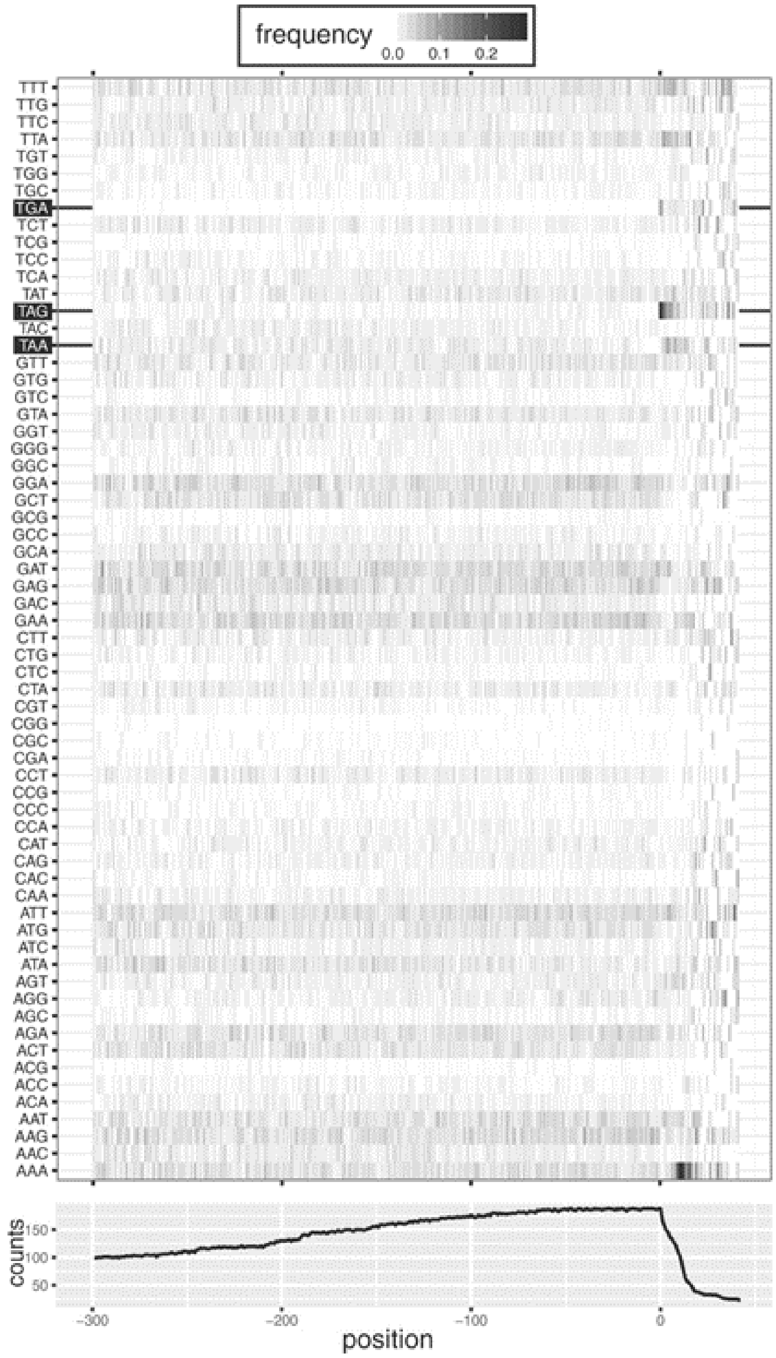


Figure. 3. Frequency of codons relative to expected positions (zero on axis x) of translation termination in *Condylostoma*. Top panel – frequency of each out of 64 codons (stop codons are highlighted). Bottom panel – total number of codons found at corresponding location. The total number differs due to variance in transcript and CDS lengths and also due to presence of ambiguous nucleotides (codons with ambiguous nucleotides were ignored).

Since the strength of stop codons as signals for termination is highly dependent on the identity of the nucleotide adjacent at the 3' end (McCaughan et al. 1995; Poole et al. 1995) we explored the possibility of a particular context preference at internal (reassigned) or terminal positions of coding regions. We observed that in both cases A and T occur more frequently than G and C consistent with AT richness of the *Condylostoma* genome (Supplementary Fig. S2). However, Ts downstream of TAGs and TGAs are more frequent than As at the sites of termination, but not at the internal positions.

Given that *Euplotes* and *Condylostoma* are distant relatives within the ciliophora phylum, it is possible that a polyA distance mechanism of translation termination has emerged in the course of convergent evolution; however, it is also conceivable that the mechanism evolved earlier in the evolution and is common to all ciliates. If the latter is true, it could explain the high propensity of ciliates for stop codon reassignment. The difference in genetic codes among ciliates would depend primarily on the availability of specific tRNAs for recognition of stop codons when those occur in internal positions. Emergence of such tRNAs is not an unlikely event in the light of a recent discovery of substantial variability in identity of codons recoded as selenocysteine in bacteria (Mukai et al. 2016). Sequence analysis of ciliate tRNAs and future experimental studies may shed a light on this intriguing possibility and disclose the molecular machinery used by ciliates to discriminate between stop codons at different positions. Possibilities include interactions between poly-A biding proteins (PABP) and ribosome complexes with release factors, as it has been shown recently that PABPs enhance termination in a mammalian system in vitro (Ivanov et al. 2016). It is also conceivable that the first ribosome reading through all stop codons could stall in the beginning of poly-A tails and serve as a barrier for trailing ribosomes favouring termination of translation when the trailing ribosomes are located at stop codons shortly upstream of ploy-A tails. Ribosome stalling at the beginnings of ploy-A tails have been observed in a yeast strain lacking ribosome rescue factor Dom34 (Guydosh and Green 2014).

METHODS

Data sources and assembly

We obtained RNA-seq data for 19 of the MMETSP ciliate species from iMicrobe (http://data.imicrobe.us/), along with four sequence read archive (SRA) files from (Feng et al. 2015) and one SRA file from (Kodama et al. 2014). SRA files were converted to fastq with FASTQ-DUMP. We used RNA-seq forward strand reads to assemble a transcriptome de novo using Trinity version r20140413p1 (Grabherr et al. 2011)for each species. A summary of the assemblies is tabulated in Table 1.

Stop codon densities and substitution frequencies

We performed pairwise alignments of conceptually translated ciliate mRNAs using standalone BLASTX 2.2.31 (Altschul et al. 1997) for each transcriptome against NCBI Reference Sequence (RefSeq) protein sequences database with an e-value of 10^{-30} as a threshold for significant sequence similarity for individual transcript hits. In order to indicate each stop codon individually we performed pseudo reassignments of two stop codons to amino acids with the one remaining stop codon translated as an unknown amino acid, denoted by ‘*’ . In total we carried out three alignments for each of the species analysed, one per stop codon. The alignments were output in format option 2 ‘query-anchored no identities’. We removed alignments where hits were originating from mitochondrial and bacterial species to reduce contamination from unintended assembled transcripts. From this output we were able to calculate the density of stop codons in each query sequence, based on the frequency in the pairwise alignments and the length of the alignment size, as illustrated in Figure 1.

Using a custom Python script, we calculated the frequency of amino acid substitutions (20 standard amino acids) in pairwise alignments for each stop codon classified as reassigned (Fig. 1). For each amino acid substitution, we calculated corresponding z-scores which are displayed as a heatmap in Supplementary Figure S1. For Figure 2, the analysis was limited only to substitutions at the positions evolving under strong stabilizing selection, i.e. the positions that are at least 95% identical across 100 closest homologs found in RefSeq database. The absolute substitution counts among conserved positions is summarised in Supplementary Table S4.

Position specific codon frequencies in *Condylostoma*.

Individual transcripts from MMETSP0210 *Condylostoma* magnum, strain COL2 from iMicrobe (http://data.imicrobe.us/) were searched using a collection of eukaryotic orthologous groups, KOGs (Tatusov et al. 2003). One “profile” alignment was built for each KOG and the pipeline (Mariotti and Guigó 2010) was used to perform protein-to-RNA alignments. The hits were filtered with a blast e-value threshold 10^{-10} and a minimum profile coverage of 90% (i.e. the predicted *Condylostoma* protein sequence was required to span at least 90% of the input KOG alignment). When multiple transcripts matched the same KOGs family, only the best scoring sequence was chosen for further analysis. The sequences at the 3’ of the homologous regions identified in this way in the *Condylostoma* transcriptome were treated as expected locations of translation termination. The frequency of each of the 64 codons was counted at each position relative to the expected location of termination (Fig. 3). The analysis of nucleotide context at the 3’ ends of stop codons was performed in the same way, except that quadruplets (stop codon and adjacent 3’ end nucleotide) were used instead of codons.

Conclusions

In this thesis I have directed my efforts to investigating alternative genetic decoding in protists. I provide additional examples of decoding plasticity including; novel mechanisms of frameshifting, stop codon readthrough and stop codon reassignment. I identified frameshifting in *Euplotes* at a frequency of over 20%, twice the previous known level. This type of frameshifting is independent of stimulatory signals and is perfectly efficient. I identified stop codon readthrough in the highly regulated gene *OAZ* in dinoflagellates. This type of recoding is rare for chromosomal genes, especially for regulatory purposes. The most significant discovery was that of a 64 sense codon genetic code in the ciliate *Condylostoma magnum*. The applications of alternative genetic decoding are becoming more relevant to the field of synthetic biology (Bezerra et al. 2015; Haimovich et al. 2015). Genetic codes are being designed to incorporate engineered amino acids into site specific locations (Chin 2014), while other engineered modifications include the development of a quadruplet-decoding ribosome (Neumann et al. 2010).

Much of the data for this thesis came from publically available datasets, in particular from The Marine Microbial Eukaryote Transcriptome Sequencing Project, which I have mined extensively. It provided an insight into a largely undiscovered area for scientists studying translation and recoding. With advances in sequencing technologies and techniques, more large scale sequencing projects are being made available. With new datasets providing opportunities for novel discoveries, the variety and frequency of recoding and reassignment examples will surely increase. However, mass sequencing of new transcriptomes and genomes from diverse organisms such as protists, provide databases with vast quantities of uncharacterised sequences. Mining these largely unknown sequences may become an issue for future comparative sequence analysis.

.

References

Adamski FM, Donly BC, Tate WP. 1993. Competition between frameshifting, termination and suppression at the frameshift site in the escherichia coli release factor-2 mRNA. Nucleic Acids Res 21:5074–5078.

Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Eberhard C, et al. 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res 44:W3–W10.

Aigner S, Lingner J, Goodrich KJ, Grosshans CA, Shevchenko A, Mann M, Cech TR. 2000. Euplotes telomerase contains an La motif protein produced by apparent translational frameshifting. EMBO J 19:6230–6239.

Altschul SF, Madden TL, Schäffer a a, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PS I-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Ambrogelly A, Palioura S, Söll D. 2007. Natural expansion of the genetic code. Nat Chem Biol 3:29–35.

Antonov I, Coakley A, Atkins JF, Baranov P V., Borodovsky M. 2013. Identification of the nature of reading frame transitions observed in prokaryotic genomes. Nucleic Acids Res 41:6514–6530.

Atkins J, Baranov V. 2010. The Distinction Between Recoding and Codon Reassignment. Genetics 185:1535–1536.

Atkins J, Gesteland RF. 2010. Recoding: Expansion of Decoding Rules Enriches Gene Expression. 1st ed. (Atkins JF, Gesteland RF, editors.). Springer-Verlag New York

Atkins JF, Loughran G, Bhatt PR, Firth AE, Baranov V. 2016. Ribosomal frameshifting and transcriptional slippage : From genetic steganography and cryptography to adventitious use. Nucleic Acids Res 44:7007–7078.

Atkins JF, Steitz J, Anderson C, Model P. 1979. Binding of mammalian ribosomes to ms2 phage rna reveals an overlapping gene encoding a lysis function. Cell 18:247–256.

Baird S, Klobutcher L. 1991. Differential DNA amplification and copy number control in the hypotrichous ciliate Euplotes crassus. J Protozool 38:136–140.

Baranov P V, Atkins JF, Yordanova MM. 2015. Augmented genetic decoding: global, local and temporal alterations of decoding processes and codon meaning. Nat Rev Genet 16:517–529.

Baranov P V., Gesteland RF, Atkins JF. 2002. Release factor 2 frameshifting sites in different bacteria. EMBO Rep 3:373–377.

Bekaert M, Atkins JF, Baranov PVP V. 2006. ARFA: A program for annotating bacterial release factor genes, including prediction of programmed ribosomal frameshifting. Bioinformatics 22:2463–2465.

Bekaert M, Ivanov IP, Atkins JF, Baranov P V. 2008. Ornithine decarboxylase antizyme finder (OAF): fast and reliable detection of antizymes with frameshifts in mRNAs. BMC Bioinformatics 9:178.

Belcourt MF, Farabaugh PJ. 1990. Ribosomal frameshifting in the yeast retrotransposon Ty: tRNAs induce slippage on a 7 nucleotide minimal site. Cell 62:339–352.

Berry MJ, Banu L, Harney JW, Larsen PR. 1993. Functional characterization of the eukaryotic SECIS elements which direct selenocysteine insertion at UGA codons. EMBO J 12:3315–3322.

Bertram G, Bell HA, Ritchie DW, Fullerton G, Stansfield I. 2000. Terminating eukaryote translation: Domain 1 of release factor eRF1 functions in stop codon recognition. Rna 6:1236–1247.

Bertram G, Innes S, Minella O, Richardson JP, Stansfield I. 2001. Endless possibilities: Translation termination and stop codon recognition. Microbiology 147:255–269.

Bezerra A, Guimarães A, Santos M. 2015. Non-Standard Genetic Codes Define New Concepts for Protein Engineering. Life 5:1610–1628.

Biswas P, Jiang X, Pacchia AL, Joseph P, Peltz SW, Dougherty JP. 2004. The Human Immunodeficiency Virus Type 1 Ribosomal Frameshifting Site Is an Invariant Sequence Determinant and an Important Target for Antiviral Therapy The Human Immunodeficiency Virus Type 1 Ribosomal Frameshifting Site Is an Invariant Sequence Determina. J Virol 78:2082–2087.

Blanchet S, Cornu D, Argentini M, Namy O. 2014. New insights into the incorporation of natural suppressor tRNAs at stop codons in Saccharomyces cerevisiae. Nucleic Acids Res 42:1–12.

Blanchet S, Rowe M, Haar T Von Der, Howard MJ, Namy O. 2015. New insights into stop codon recognition by eRF1. Nucleic Acids Res 43:3298–3308.

Blinkowa a L, Walker JR. 1990. Programmed ribosomal frameshifting generates the Escherichia coli DNA polymerase III gamma subunit from within the tau subunit reading frame. Nucleic Acids Res 18:1725–1729.

Brown A, Shao S, Murray J, Hegde RS, Ramakrishnan V. 2015. Structural basis for stop codon recognition in eukaryotes. Nature 524:493–496.

Bulygin KN, Khairulina YS, Kolosov PM, Ven’yaminova AG, Graifer DM, Vorobjev YN, Frolova LY, Kisselev LL, Karpova GG. 2010. Three distinct peptides from the N domain of translation termination factor eRF1 surround stop codon in the ribosome. RNA 16:1902–1914.

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

Cassago A, Rodrigues EM, Prieto EL, Gaston KW, Alfonzo JD, Iribar MP, Berry MJ, Cruz AK, Thiemann OH. 2006. Identification of Leishmania selenoproteins and SECIS element. Mol Biochem Parasitol 149:128–134.

Chambers I, Frampton J, Goldfarb P, Affara N, McBain W, Harrison PR. 1986. The structure of the mouse glutathione peroxidase gene: the selenocysteine in the active site is encoded by the “termination” codon, TGA. EMBO J 5:1221–1227.

Chavatte L, Seit-Nebi A, Dubovaya V, Favre A. 2002. The invariant uridine of stop codons contacts the conserved NIKSR loop of human eRF1 in the ribosome. EMBO J 21:5302–5311.

Chin J. 2014. Expanding and Reprogramming the Genetic Code of Cells and Animals. Annu Rev Biochem 83:379–408.

Conard SE, Buckley J, Dang M, Bedwell GJ, Carter RL, Khass M, Bedwell DM. 2012. Identification of eRF1 residues that play critical and complementary roles in stop codon recognition. RNA 18:1210–1221.

Copeland PR, Fletcher JE, Carlson BA, Hatfield DL, Driscoll DM. 2000. A novel RNA binding protein, SBP2, is required for the translation of mammalian selenoprotein mRNAs. EMBO J 19:306–314.

Craigen W, Caskey C. 1986. Expression of peptide chain release factor 2 requires high-efficiency frameshift. Nature 322:273–275.

Crick F. 1968. The origin of the genetic code. J Mol Biol 38:367–379.

Crooks G, Hon G, Chandonia J, Brenner S. 2004. WebLogo: a sequence logo generator. Genome Res 14:1188–1190.

Depuydt G, Xie F, Petyuk V a, Smolders A, Brewer M, Camp DG, Smith RD, Braeckman BP. 2014. LC-MS proteomics analysis of the insulin/IGF-1-deficient Caenorhabditis elegans daf-2(e1370) mutant reveals extensive restructuring of intermediary metabolism. J Proteome Res 13:1938–1956.

Depuydt G, Xie F, Petyuk VA, Shanmugam N, Smolders A, Dhondt I, Brewer HM, Camp DG, Smith RD, Braeckman BP. 2013. Reduced Insulin/Insulin-like Growth Factor-1 Signaling and Dietary Restriction Inhibit Translation but Preserve Muscle Mass in *Caenorhabditis elegans*. Mol Cell Proteomics 12:3624–3639.

Duarte I, Nabuurs SB, Magno R, Huynen M. 2012. Evolution and diversification of the organellar release factor family. Mol Biol Evol 29:3497–3512.

Dunn J, Studier F. 1983. Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. J Mol Biol 4:477–535.

Dunn JG, Foo CK, Belletier NG, Gavis ER, Weissman JS. 2013. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. Elife 2013:1–32.

Dziallas C, Allgaier M, Monaghan MT, Grossart HP. 2012. Act together-implications of symbioses in aquatic ciliates. Front Microbiol 3:1–17.

Eisen JA, Coyne RS, Wu M, Wu D, Thiagarajan M, Wortman JR, Badger JH, Ren Q, Amedeo P, Jones KM, et al. 2006. Macronuclear genome sequence of the ciliate Tetrahymena thermophila, a model eukaryote. PLoS Biol 4:1620–1642.

Endoh T, Sugimoto N. 2013. Unusual -1 ribosomal frameshift caused by stable RNA G-quadruplex in open reading frame. Anal Chem 85:11435–11439.

Eswarappa SM, Potdar AA, Koch WJ, Fan Y, Vasu K, Lindner D, Willard B, Graham LM, Dicorleto PE, Fox PL. 2014. Programmed translational readthrough generates antiangiogenic VEGF-Ax. Cell 157:1605–1618.

Farabaugh PJ. 2000. Translational frameshifting: implications for the mechanism of translational frame maintenance. Prog Nucleic Acid Res Mol Biol 64:131–170.

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

Feng YX, Yuan H, Rein A, Levin JG. 1992. Bipartite signal for read-through suppression in murine leukemia virus mRNA: an eight-nucleotide purine-rich sequence immediately downstream of the gag termination codon followed by an RNA pseudoknot. J Virol 66:5127–5132.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: Interactive sequence similarity searching. Nucleic Acids Res 39:29–37.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. 2016. The Pfam protein families database: Towards a more sustainable future. Nucleic Acids Res 44:D279–D285.

Firth AE, Wills NM, Gesteland RF, Atkins JF. 2011. Stimulation of stop codon readthrough: Frequent presence of an extended 3’ RNA structural element. Nucleic Acids Res 39:6679–6691.

Firth AE. 2014. Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. Nucleic Acids Res 42:12425–12439.

Fletcher JE, Copeland PR, Driscoll DM, Krol a. 2001. The selenocysteine incorporation machinery: interactions between the SECIS RNA and the SECIS-binding protein SBP2. RNA 7:1442–1453.

Flower AM, McHenry CS. 1990. The gamma subunit of DNA polymerase III holoenzyme of Escherichia coli is produced by ribosomal frameshifting. Proc Natl Acad Sci U S A 87:3713–3717.

Freistroffer D V., Kwiatkowski M, Buckingham RH, Ehrenberg M. 2000. The accuracy of codon recognition by polypeptide release factors. Proc Natl Acad Sci 97:2046–2051.

Frolova L, Seit-Nebi A, Kisselev L. 2002. Highly conserved NIKS tetrapeptide is functionally essential in eukaryotic translation termination factor eRF1. RNA 8:129–136.

Fučíková K, Leliaert F, Cooper ED, Å kaloud P, D’Hondt S, De Clerck O, Gurgel CFD, Lewis LA, Lewis PO, Lopez-Bautista JM, et al. 2014. New phylogenetic hypotheses for the core Chlorophyta based on chloroplast sequence data. Front Ecol Evol 2:1–12.

Gerashchenko M V., Lobanov A V., Gladyshev VN. 2012. Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. Proc Natl Acad Sci 109:17394–17399.

Gesteland R., Wolfner M, Grisafi P, Fink G, Botstein D, Roth J. 1976. Yeast suppressors of UAA and UAG nonsense codons work efficiently in vitro via tRNA. Cell 7:381–390.

Gesteland RF, Atkins JF. 1996. Recoding: Dynamic Reprogramming of Translation. Annu Rev Biochem 65:741–768.

Giedroc DP, Cornish P V. 2009. Frameshifting RNA pseudoknots: Structure and mechanism. Virus Res 139:193–208.

Gobler CJ, Berry DL, Dyhrman ST, Wilhelm SW, Salamov A, Lobanov A V., Zhang Y, Collier JL, Wurch LL, Kustka AB, et al. 2011. Niche of harmful alga Aureococcus anophagefferens revealed through ecogenomics. Proc Natl Acad Sci 108:4352–4357.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, 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.

Guydosh NR, Green R. 2014. Dom34 rescues ribosomes in 3' untranslated regions. Cell 156:950–962.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8:1494–1512.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Philip D, Bowden J, Couger MB, Eccles D, Li B, Macmanes MD, et al. 2014. reference generation and analysis with Trinity.

Haimovich AD, Muir P, Isaacs FJ. 2015. Genomes by design. Nat Rev Genet 16:501–516.

Heaphy SM, Mariotti M, Gladyshev VN, Atkins JF, Baranov P V. 2016. Novel Ciliate Genetic Code Variants Including the Reassignment of All Three Stop Codons to Sense Codons in Condylostoma magnum. Mol Biol Evol 33:2885–2889.

Heider J, Baron C, Böck a. 1992. Coding from a distance: dissection of the mRNA determinants required for the incorporation of selenocysteine into protein. EMBO J 11:3759–3766.

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 terminationcodon. Nucleic Acids Res 13:415–433.

Hill KE, Lloyd RS, Burk RF. 1993. Conserved nucleotide sequences in the open reading frame and 3' untranslated region of selenoprotein P mRNA. Proc Natl Acad Sci U S A 90:537–541.

Hoffman DW, Carroll D, Martinez N, Hackert ML, June R V, Re V, Recei M, July V. 2005. Solution Structure of a Conserved Domain of Antizyme : A Protein Regulator of Polyamines. Biochemistry 44:11777–11785.

Hofhuis J, Schueren F, No C, Jahn O, Thoms S, Thoms S. 2016. The functional readthrough extension of malate dehydrogenase reveals a modification of the genetic code. Open Biol 6:1–13.

Hofstetter H, Monstein J, Weissmann C. 1974. The readthrough protein A1 is essential for the formation of viable Qβ particles. Biochim Biophys Acta - Nucleic Acids Protein Synth 374:238–251.

Horowitz S, Gorovsky M a. 1985. An unusual genetic code in nuclear genes of Tetrahymena. Proc Natl Acad Sci U S A 82:2452–2455.

Huang WM, Ao SZ, Casjens S, Orlandi R, Zeikus R, Weiss R, Winge D, Fang M. 1988. A persistent untranslated sequence within bacteriophage T4 DNA topoisomerase gene 60. Science (80-) 239:1005–1012.

Huang X, Wang J, Aluru S, Yang S, Hillier L. 2003. PCAP : A Whole-Genome Assembly Program. :2164–2170.

Huang X, Yang S. 2005. Generating a Genome Assembly with PCAP. Curr Protoc Bioinforma Chapter 11.

Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324:218–223.

Ingolia NT. 2014. Ribosome profiling: New views of translation, from single codons to genome scale. Nat Rev Genet 15:205–213.

Ito K, Frolova L, Seit-Nebi A, Karamyshev A, Kisselev L, Nakamura Y. 2002. Omnipotent decoding potential resides in eukaryotic translation termination factor eRF1 of variant-code organisms and is modulated by the interactions of amino acid sequences within domain 1. Proc Natl Acad Sci U S A 99:8494–8499.

Ivanov A, Mikhailova T, Eliseev B, Yeramala L, Sokolova E, Susorov D, Shuvalov A, Schaffitzel C, Alkalaeva E. 2016. PABP enhances release factor recruitment and stop codon recognition during translation termination. Nucleic Acids Res 44:7766–7776.

Ivanov IP, Atkins JF. 2007. Ribosomal frameshifting in decoding antizyme mRNAs from yeast and protists to humans: Close to 300 cases reveal remarkable diversity despite underlying conservation. Nucleic Acids Res 35:1842–1858.

Ivanov IP, Gesteland RF, Atkins JF. 1998. A second mammalian antizyme: conservation of programmed ribosomal frameshifting. Genomics 52:119–129.

Ivanov IP, Gesteland RF, Atkins JF. 2000. Antizyme expression: a subversion of triplet decoding, which is remarkably conserved by evolution, is a sensor for an autoregulatory circuit. Nucleic Acids Res 28:3185–3196.

Ivanov IP, Pittman AJ, Chien C Bin, Gesteland RF, Atkins JF. 2007. Novel antizyme gene in Danio rerio expressed in brain and retina. Gene 387:87–92.

Ivanov IP, Rohrwasser A, Terreros DA, Gesteland RF, Atkins JF. 2000. Discovery of a spermatogenesis stage-specific ornithine decarboxylase antizyme: Antizyme 3. Proc Natl Acad Sci 97:4808–4813.

Ivanova NN, Schwientek P, Tripp H, Rinke C. 2014. Stop codon reassignments in the wild. 909:909–914.

Jacks T, Madhani H, Masiarz F, Varmus H. 1988. Signals for ribosomal frameshifting in the rous sarcoma virus gag-pol region. Cell 55:447–458.

Jacks T, Power M, Masiarz F, Luciw P, Barr P, Varmus H. 1988. Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. Nature 331:280–283.

Jiang J, Huang J, Al-Farraj SA, Lin X, Hu X. 2016. Morphology and Molecular Phylogeny of Two Poorly Known Species of *Protocruzia* (Ciliophora: Protocruziida). J Eukaryot Microbiol:1–9.

Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, White KP, Kellis M. 2011. Evidence of abundant stop codon readthrough in Drosophila and other metazoa. Genome Res 21:2096–2113.

Karamysheva Z, Wang L, Shrode T, Bednenko J, Hurley LA, Shippen DE. 2003. Developmentally Programmed Gene Elimination in Euplotes crassus Facilitates a Switch in the Telomerase Catalytic Subunit. 113:565–576.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler L a., Armbrust EV, Archibald JM, Bharti AK, Bell CJ, 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.

Keeling PJ, Doolittle WF. 1996. A non-canonical genetic code in an early diverging eukaryotic lineage. Embo J 15:2285–2290.

Kim H-K, Liu F, Fei J, Bustamante C, Gonzalez RL, Tinoco I. 2014. A frameshifting stimulatory stem loop destabilizes the hybrid state and impedes ribosomal translocation. Proc Natl Acad Sci 111:5538–5543.

Kim S, Pevzner PA. 2014. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 5:1–10.

Klobutcher LA, Farabaugh PJ. 2002. Shifty Ciliates : Frequent Programmed Translational Frameshifting in Euplotids Minireview. 111:763–766.

Klobutcher LA. 2005. Sequencing of random Euplotes crassus macronuclear genes supports a high frequency of +1 translational frameshifting. Eukaryot Cell 4:2098–2105.

Knight RD, Freeland SJ, Landweber LF. 2001. Rewiring the keyboard: evolvability of the genetic code. Nat Rev Genet 2:49–58.

Kodama Y, Suzuki H, Dohra H, Sugii M, Kitazume T, Yamaguchi K, Shigenobu S, Fujishima M. 2014. Comparison of gene expression of *Paramecium bursaria* with and without *Chlorella variabilis* symbionts. BMC Genomics 15:1–8.

Kolosov P, Frolova L, Seit-Nebi A, Dubovaya V, Kononenko A, Oparina N, Justesen J, Efimov A, Kisselev L. 2005. Invariant amino acids essential for decoding function of polypeptide release factor eRF1. Nucleic Acids Res 33:6418–6425.

Korkmaz G, Holm M, Wiens T, Sanyal S. 2014. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. J Biol Chem 289:30334–30342.

Laforest MJ, Roewer I, Franz Lang B. 1997. Mitochondrial tRNAs in the lower fungus Spizellomyces punctatus: TRNA editing and UAG “stop” codons recognized as leucine. Nucleic Acids Res 25:626–632.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:1–10.

Larsen B, Wills NM, Gesteland RF, Atkins JF. 1994. rRNA-mRNA base pairing stimulates a programmed -1 ribosomal frameshift. J Bacteriol 176:6842–6851.

Laslett D, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res 32:11–16.

Lee S, Bar-noy S, Kwon J, Levine RL, Stadtman TC, Rhee SG. 2000. Mammalian thioredoxin reductase: Oxidation of the C-terminal cysteine/selenocysteine active site forms a thioselenide, and replacement of selenium with sulfur markedly reduces catalytic activity. Proc Natl Acad Sci 97:2521–2526.

Lekomtsev S, Kolosov P, Bidou L, Frolova L, Rousset J-P, Kisselev L. 2007. Different modes of stop codon restriction by the Stylonychia and Paramecium eRF1 translation termination factors. Proc Natl Acad Sci U S A 104:10824–10829.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Li Y, Kim OTP, Ito K, Saito K, Suzaki T, Harumoto T. 2013. A Single Amino Acid Substitution Alters Omnipotent eRF1 of Dileptus to Euplotes-type Dualpotent eRF1: Standard Codon Usage May be Advantageous in Raptorial Ciliates. Protist 164:440–449.

Liang and Heckmann. 1993. Blepharisma uses UAA as a termination codon. Naturwissenschaften 80:225–226.

Lobanov A V, Heaphy SM, Turanov AA, Gerashchenko M V, Pucciarelli S, Devaraj RR, Xie F, Petyuk VA, Smith RD, Klobutcher LA, et al. 2017. Position-dependent termination and widespread obligatory frameshifting in Euplotes translation. Nat Struct Mol Biol 24:61–68.

Lobanov A V., Gromer S, Salinas G, Gladyshev VN. 2006. Selenium metabolism in Trypanosoma: Characterization of selenoproteomes and identification of a Kinetoplastida-specific selenoprotein. Nucleic Acids Res 34:4012–4024.

Lobanov A V., Hatfield DL, Gladyshev VN. 2009. Eukaryotic selenoproteins and selenoproteomes. Biochim Biophys Acta - Gen Subj 1790:1424–1428.

Loughran G, Chou M, Ivanov IP, Jungreis I, Kellis M, Kiran AM, Baranov P V, Atkins JF. 2014. Evidence of efficient stop codon readthrough in four mammalian genes. Nucleic Acids Res 42:8928–8938.

Loughran G, Howard MT, Firth AE, Atkins JF. 2017. Avoidance of reporter assay distortions from fused dual reporters. RNA 23:1285–1289.

Lowe TM, Eddy SR. 1996. TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964.

Lozupone CA, Knight RD, Landweber LF. 2001. The molecular basis of nuclear genetic code change in ciliates. Curr Biol 11:65–74.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Tang J, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient sort read de novo assembler. Gigascience 1:1–6.

Mariotti M, Guigo R. 2010. Selenoprofiles: Profile-based scanning of eukaryotic genome sequences for selenoprotein genes. Bioinformatics 26:2656–2663.

Mariotti M, Santesmasses D, Capella-Gutierrez S, Mateo A, Arnan C, Johnson R, D'Aniello S, Yim SH, Gladyshev VN, Serras F, et al. 2015. Evolution of selenophosphate synthetases: Emergence and relocation of function through independent duplications and recurrent subfunctionalization. Genome Res 25:1256–1267.

Matsufuji S, Matsufuji T, Miyazaki Y, Murakami Y, Atkins JF, Gesteland RF, Hayashi S ichi. 1995. Autoregulatory frameshifting in decoding mammalian ornithine decarboxylase antizyme. Cell 80:51–60.

Matsufuji S, Matsufuji T, Wills NM, Gesteland RF, Atkins JF. 1996. Reading two bases twice: mammalian antizyme frameshifting in yeast. EMBO J 15:1360–1370.

Matsufuji S, Miyazaki Y, Kanamoto R, Kameji T, Murakami Y, Baby G, Fujita K, Ohno T. 1990. Analyses of Ornithine Decarboxylase Antizyme mRNA with a cDNA Cloned from Rat Liver1. J Biochem 108:365–371.

Mayampurath AM, Jaitly N, Purvine SO, Monroe ME, Auberry KJ, Adkins JN, Smith RD. 2008. DeconMSn: A software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. Bioinformatics 24:1021–1023.

McCaughan KK, Brown CM, Dalphin ME, Berry MJ, Tate WP. 1995. Translational termination efficiency in mammals is influenced by the base following the stop codon. Proc Natl Acad Sci U S A 92:5431–5435.

McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, Cowley AP, Lopez R. 2013. Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res 41:597–600.

Meyer F, Schmidt HJ, Plümper E, Hasilik a, Mersmann G, Meyer HE, Engström a, Heckmann K. 1991. UGA is translated as cysteine in pheromone 3 of Euplotes octocarinatus. Proc Natl Acad Sci U S A 88:3758–3761.

Michel AM, Baranov P V. 2013. Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. Wiley Interdiscip Rev RNA 4:473–490.

Miyazaki Y, Matsufuji S, Hayashi S. 1992. Cloning and characterization of a rat gene encoding ornithine decarboxylase antizyme. Gene 113:191–197.

Mu W, Wang Q, Bourland WA, Jiang C, Yuan D, Pan X, Miao W, Chen Y, Xiong J. 2016. Epidermal growth factor-induced stimulation of proliferation and gene expression changes in the hypotrichous ciliate , Stylonychia lemnae. Gene 592:186–192.

Mühlhausen S, Findeisen P, Plessmann U, Urlaub H, Kollmar M. 2016. A novel nuclear genetic code alteration in yeasts and the evolution of codon reassignment in eukaryotes. Genome Res 26:945–955.

Mukai T, Englert M, Tripp HJ, Miller C, Ivanova NN, Rubin EM, Kyrpides NC, Soll D. 2016. Facile Recoding of Selenocysteine in Nature. Angew Chemie - Int Ed 55:5337–5341.

Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KHJ, Remington KA, et al. 2000. A Whole-Genome Assembly of Drosophila. 287:2196–2205.

Nakamura Y, Ito K. 1998. How protein reads the stop codon and terminates translation. Genes to Cells 3:265–278.

Namy O, Duchateau-Nguyen G, Hatin I, Hermann-Le Denmat S, Termier M, Rousset JP. 2003. Identification of stop codon readthrough genes in Saccharomyces cerevisiae. Nucleic Acids Res 31:2289–2296.

Namy O, Hatin I, Rousset JP. 2001. Impact of the six nucleotides downstream of the stop codon on translation termination. EMBO Rep 2:787–793.

Namy O, Rousset J. 2010. Specification of Standard Amino Acids by Stop Codons. In Recoding: Expansion of Decoding Rules Enriches Gene Expression. In: Recoding: Expansion of Decoding Rules Enriches Gene Expression. Springer US. p. 79–100.

Neumann H, Wang K, Davis L, Garcia-alai M, Chin JW. 2010. Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. Nature 464:441–444.

Novoselov S V, Lobanov A V, Hua D, Kasaikina M V, Hatfield DL, Gladyshev VN. 2007. A highly efficient form of the selenocysteine insertion sequence element in protozoan parasites and its use in mammalian cells. Proc Natl Acad Sci U S A 104:7857–7862.

Ohama T, Suzuki T, Mori M, Osawa S, Ueda T, Watanabe K, Nakase T. 1993. Non-universal decoding of the leucine codon CUG in several Candida species. Nucleic Acids Res 21:4039–4045.

Osawa S, Jukes T. 1989. Codon reassignment (codon capture) in evolution. J Mol Evol 28:271–278.

Pánek T, Žihala D, Sokol M, Derelle R, Klimeš V, Hradilová M, Zadrobílková E, Susko E, Roger AJ, Čepička I, et al. 2017. Nuclear genetic codes with a different meaning of the UAG and the UAA codon. BMC Biol 15:8.

Parkin NT, Chamorro M, Varmus HE. 1992. Human immunodeficiency virus type 1 gag-pol frameshifting is dependent on downstream mRNA secondary structure: demonstration by expression in vivo. J Virol 66:5147–5151.

Pearson W. 2004. Finding Protein and Nucleotide Similarities with FASTA. Curr Protoc Bioinforma Chapter 3.

Pelham HR. 1978. Leaky UAG termination codon in tobacco mosaic virus RNA. Nature 272:469–471.

Petyuk VA, Mayampurath AM, Monroe ME, Polpitiya AD, Purvine SO, Anderson GA, Camp DG, Smith RD. 2010. DtaRefinery, a Software Tool for Elimination of Systematic Errors from Parent Ion Mass Measurements in Tandem Mass Spectra Data Sets. Mol Cell Proteomics 9:486–496.

Petyuk VA, Qian WJ, Hinault C, Gritsenko MA, Singhal M, Monroe ME, Camp DG, Kulkarni RN, Smith RD. 2008. Characterization of the mouse pancreatic islet proteome and comparative analysis with other mouse tissues. J Proteome Res 7:3114–3126.

Poole ES, Brown CM, Tate WP. 1995. The identity of the base following the stop codon determines the efficiency of in vivo translational termination in Escherichia coli. EMBO J 14:151–158.

Prat L, Heinemann IU, Aerni HR, Rinehart J, O'Donoghue P, Söll D. 2012. Carbon source-dependent expansion of the genetic code in bacteria. Proc Natl Acad Sci U S A 109:21070–21075.

Prescott DM. 1994. The DNA of Ciliated Protozoa. Microbiol Rev 58:233–267.

Pucciarelli S, La Terza A, Ballarini P, Barchetta S, Yu T, Marziale F, Passini V, Methé B, Detrich HW, Miceli C. 2009. Molecular cold-adaptation of protein function and gene regulation: The case for comparative genomic analyses in marine ciliated protozoa. Mar Genomics 2:57–66.

Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: Multiple alignment of coding SEquences accounting for frameshifts and stop codons. PLoS One 6:e22594–e22594.

Rayman MP. 2000. The importance of selenium to human health. [Review] [84 refs]. Lancet 356:233–241.

Ricard G, de Graaf RM, Dutilh BE, Duarte I, van Alen TA, van Hoek AHAM, Boxma B, van der Staay GWM, Moon-van der Staay S, Chang WJ, et al. 2008. Macronuclear genome structure of the ciliate Nyctotherus ovalis: Single-gene chromosomes and tiny introns. BMC Genomics 9:1–15.

Riyasaty S, Atkins J. 1968. External suppression of a frameshift mutant in Salmonella. J Mol Biol 34:541–557.

Robinson DN, Cooley L. 1997. Examination of the function of two kelch proteins generated by stop codon suppression. Development 124:1405–1417.

Roy SW, Ruby JG, Reiff SB, Swart EC, Gosai S, Prabakaran S, Witkowska E, Larue GE, Fisher S, Freeman RM, et al. 2017. The Macronuclear Genome of Stentor coeruleus Reveals Tiny Introns in a Giant Cell. Curr Biol 27:569–575.

Salas-Marco J, Fan-Minogue H, Kallmeyer AK, Klobutcher LA, Farabaugh PJ, Bedwell DM. 2006. Distinct paths to stop codon reassignment by the variant-code organisms Tetrahymena and Euplotes. Mol Cell Biol 26:438.

Sánchez-Silva R, Villalobo E, Morin L, Torres A. 2003. A new noncanonical nuclear genetic code: Translation of UAA into glutamate. Curr Biol 13:442–447.

Sato K, Kato Y, Hamada M, Akutsu T, Asai K. 2011. IPknot: Fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. Bioinformatics 27:85–93.

Schmieder R, Edwards R. 2011. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One 6:e17288–e17288.

Schneider S, de Groot E. 1991. Sequences of two rbcS cDNA clones of Batophora oerstedii: structural and evolutionary considerations. Curr Genet 20:173–175.

Schneider S, Leible M, Yang X. 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 Genet Genomics 218:445–452.

Schueren F, Lingner T, George R, Hofhuis J, Dickel C, Gärtner J, Thoms S. 2014. Peroxisomal lactate dehydrogenase is generated by translational readthrough in mammals. Elife 3:e03640.

Seit-Nebi A, Frolova L, Kisselev L. 2002. Conversion of omnipotent translation termination factor eRF1 into ciliate-like UGA-only unipotent eRF1. EMBO Rep 3:881–886.

Sharma V, Firth AE, Antonov I, Fayet O, Atkins JF, Borodovsky M, Baranov P V. 2011. A pilot study of bacterial genes with disrupted ORFs reveals a surprising profusion of protein sequence recoding mediated by ribosomal frameshifting and transcriptional realignment. Mol Biol Evol 28:3195–3211.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, et al. 2014. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539–539.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM. 2009. ABySS : A parallel assembler for short read sequence data ABySS : A parallel assembler for short read sequence data. :1117–1123.

Srinivasan G, James CM, Krzycki JA. 2002. Pyrrolysine encoded by UAG in archaea: Charging of a UAG-decoding specialized tRNA. Science 296:1459–1462.

Steneberg P, Samakovlis C. 2001. A novel stop codon readthrough mechanism produces functional headcase protein in Drosophila trachea. EMBO Rep 2:593–597.

Swart EC, Bracht JR, Magrini V, Minx P, Chen X, Zhou Y, Khurana JS, Goldman AD, 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–e1001473.

Swart EC, Serra V, Petroni G, Nowacki M. 2016. Genetic Codes with No Dedicated Stop Codon: Context-Dependent Translation Termination. Cell 166:691–702.

Tajima A, Murai N, Murakami Y, Iwamoto T, Migita T, Matsufuji S. 2016. Polyamine regulating protein antizyme binds to ATP citrate lyase to accelerate acetyl-CoA production in cancer cells. Biochem Biophys Res Commun 471:646–651.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: Molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729.

Tan M, Heckmann K, Brunen-nieweler C. 2001. Analysis of Micronuclear, Macronuclear and cDNA Sequences Encoding the Regulatory Subunit of CAMP-Dependent Protein Kinase of. 69:80–87.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin E V, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.

Tholstrup J, Oddershede LB, Sørensen MA. 2012. MRNA pseudoknot structures can act as ribosomal roadblocks. Nucleic Acids Res 40:303–313.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief Bioinform 14:178–192.

Touriol C, Bornes S, Bonnal S, Audigier S, Prats H, Prats AC, Vagner S. 2003. Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. Biol Cell 95:169–178.

Tsuchihashi Z. 1991. Translational frame shifting in the Escherichia coli dnaX gene in vitro. Nucleic Acids Res 19:2457–2462.

Turanov A a, Lobanov A V, Fomenko DE, Morrison HG, Sogin ML, Klobutcher L a, Hatfield DL, Gladyshev VN. 2009. Genetic code supports targeted insertion of two amino acids by one codon. Science 323:259–261.

Unseld M, Marienfeld JR, Brandt P, Brennicke A. 1997. The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides. Nat Genet 15:57–61.

Valbonesi A, Luporini P. 1993. Biology of Euplotes focardii, an Antarctic ciliate. Polar Biol 13:489–493.

Vallabhaneni H, Fan-minogue H, Bedwell DM, Fan-minogue HUA, Farabaugh PJ. 2009. Connection between stop codon reassignment and frequent use of shifty stop frameshifting. RNA 15:889–897.

Vinogradov D V, Tsoi O V, Zaika A V, Lobanov A V, Turanov AA, Gladishev VN. 2012. Draft Macronucleus Genome of Euplotes crassus Ciliate. Mol Biol 46:328–333.

Wang L, Dean SR, Shippen DE. 2002. Oligomerization of the telomerase reverse transcriptase from Euplotes crassus. Nucleic Acids Res 30:4032–4039.

Wang R, Xiong J, Wang W, Miao W, Liang A. 2016. High frequency of +1 programmed ribosomal frameshifting in Euplotes octocarinatus. Sci Rep 6:21139.

Wang S, Miller WA. 1995. A sequence located 4.5 to 5 kilobases from the 5′ end of the barley yellow dwarf virus (PAV) genome strongly stimulates translation of uncapped mRNA. J Biol Chem 270:13446–13452.

Warren RL, Sutton GG, Jones SJM, Holt RA. 2007. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23:500–501.

Waterhouse AM, Procter JB, Martin DM a, Clamp M, Barton GJ. 2009. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191.

Weiner AM, Weber K. 1971. Natural read-through at the UGA termination signal of Q-β coat protein cistron. Nat New Biol 234:206–209.

Wong LC, Landweber LF. 2006. Evolution of programmed DNA rearrangements in a scrambled gene. Mol Biol Evol 23:756–763.

Xiong J, Wang G, Cheng J, Tian M, Pan X, Warren A, Jiang C, Yuan D, Miao W. 2015. Genome of the facultative scuticociliatosis pathogen Pseudocohnilembus persalinus provides insight into its virulence through horizontal gene transfer. Sci Rep 5:15470.

Yang F, Shen Y, Camp DG, Smith RD. 2012. High pH reversed-phase chromatography with fraction concatenation as an alternative to strong-cation exchange chromatography for two-dimensional proteomic analysis. Expert Rev Proteomics 9:129–134.

Yordanova MM, Wu C, Andreev DE, Sachs MS, Atkins JF. 2015. A nascent peptide signal responsive to endogenous levels of polyamines acts to stimulate regulatory frameshifting on antizyme mRNA. J Biol Chem 290:17863–17878.

Záhonová K, Kostygov AY, Ševčíková T, Yurchenko V, Eliáš M. 2016. An Unprecedented Non-canonical Nuclear Genetic Code with All Three Termination Codons Reassigned as Sense Codons. Curr Biol 26:2364–2369.

Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829.

Zhong L, Arner ESJ, Holmgren A. 2000. Structure and mechanism of mammalian thioredoxin reductase: The active site is a redox-active selenolthiol/selenenylsulfide formed from the conserved cysteine-selenocysteine sequence. Proc Natl Acad Sci 97:5854–5859.

Appendix

GWIPS-viz: Development of a ribo-seq genome browser

This appendix has been published as a research article Nucleic Acids Res. 2014 Jan 1; 42(Database issue): D859–D864.

https://doi.org/10.1093/nar/gkt1035

ABSTRACT

Ribosome profiling (ribo-seq) is a recently developed technique that provides Genome Wide Information on Protein Synthesis (GWIPS) in vivo. It is based on the deep sequencing of ribosome protected mRNA fragments which allows the ribosome density along all mRNA transcripts present in the cell to be quantified. Since its inception, ribo-seq has been carried out in a number of eukaryotic and prokaryotic organisms. Due to the increasing interest in ribo-seq, there is a pertinent demand for a dedicated ribo-seq genome browser. Therefore we have developed GWIPS-viz, an online genome browser for viewing ribosome profiling data. GWIPS-viz is based on the UCSC Genome Browser. Ribo-seq tracks coupled with mRNA-seq tracks are currently available for several genomes: Human, Mouse, Zebrafish, Nematode, Yeast, Bacteria (Escherichia coli K12, Bacillus subtilis), Human Cytomegalovirus and Bacteriophage lambda. Our objective is to continue incorporating published ribo-seq datasets so that the wider community can readily view ribosome profiling information without the need to carry out computational processing.

Database URL: http://gwips.ucc.ie

INTRODUCTION

Ribosome profiling is based on the isolation of mRNA fragments protected by ribosomes followed by massively parallel sequencing of the protected fragments or footprints. This allows the measurement of ribosome density along all mRNA transcripts present in the cell providing genome-wide information on protein synthesis (GWIPS) in vivo (1). The ribosome profiling technique, also known as ribo-seq, was first carried out in Saccharomyces cerevisiae (2). Since the original publication, the technique has been carried out in many organisms including Homo sapiens (3-10) Mus musculus (3,7,9,11,12) Danio rerio (13), Caenorhabditis elegans (4,14), Saccharomyces cerevisiae (15,16), Escherichia coli (17,18), Bacillus subtilis (18), human cytomegalovirus (19) and, Bacteriophage lambda (20). .

To date, there have been two main strategies of ribosome profiling: ribosome profiling of initiating ribosomes and ribosome profiling of elongating ribosomes. For a review on the usages and advantages of each approach, please see (21).

The majority of published studies using ribosome profiling provide the raw sequencing data in NCBI's Sequence Read Archive (SRA)(22). In addition, most published ribosome profiling experiments have corresponding naked mRNA control, where total mRNA is randomly degraded to yield fragments of a size similar to ribosome protected fragments. For simplicity here we refer to it as mRNA-seq. mRNA-seq is carried out under the same experimental conditions. It helps to take into account the differential abundance of mRNA between experimental conditions and to monitor technical biases associated with cDNA libraries generation and sequencing.

Due to the increasing popularity of the ribo-seq technique, the number of ribosome profiling experiments is expected to increase dramatically in the near future. However, the visualization of ribosome profiling data in a browser first requires pre-processing and aligning the raw sequencing reads. As with any type of next-generation sequencing data (NGS), demands are placed on biomedical researchers in terms of time, data storage, computational knowledge and prototyping of computational pipelines (23). Web-based integrative framework tools such as Galaxy (24) provide centralized platforms for researchers to carry out NGS alignment pipelines. However, due to decreasing costs, the coverage depth of ribo-seq and corresponding mRNA-seq data is continually increasing resulting in ever larger datasets. Consequently the computational resources required to process such data and the computer memory required to store such data may not be available to many biologists. Indeed, the time required to download, pre-process and align the raw data may be the most limiting factor of all for time-poor researchers.

To address these issues, we introduce GWIPS-viz (http://gwips.ucc.ie), a free online browser which is pre-populated with published ribo-seq data. The aim of GWIPS-viz is to provide an intuitive graphical interface of translation in the genomes for which ribo-seq data are available. Users can readily view alignments from many of the published ribo-seq studies without the need to carry out any computational processing. GWIPS-viz is based on a customized version of the UCSC Genome Browser (http://genome.ucsc.edu) (25). Riboseq tracks, coupled with mRNA-seq tracks, are currently available for Human, Mouse, Zebrafish, Nematode Yeast, two bacterial species (Escherichia coli K12 and Bacillus subtilis) and two viral genomes (Human Cytomegalovirus and Bacteriophage lambda).

USAGE

In GWIPS-viz, users can search for their gene(s) of interest in the genome(s) for which riboseq data is available and view a snapshot of the gene's translation under the conditions of the experiment. Ribosome coverage plots (red) and mRNA-seq coverage plots (green) display the number of reads that cover a given genomic coordinate. Figure 1 provides coverage plots for the S. cerevisiae genome locus containing ABP140, MET7, SSP2, and PUS7 (from Ingolia et al. PMID:19213877) and illustrates how differential translation can be viewed in GWIPS-viz.

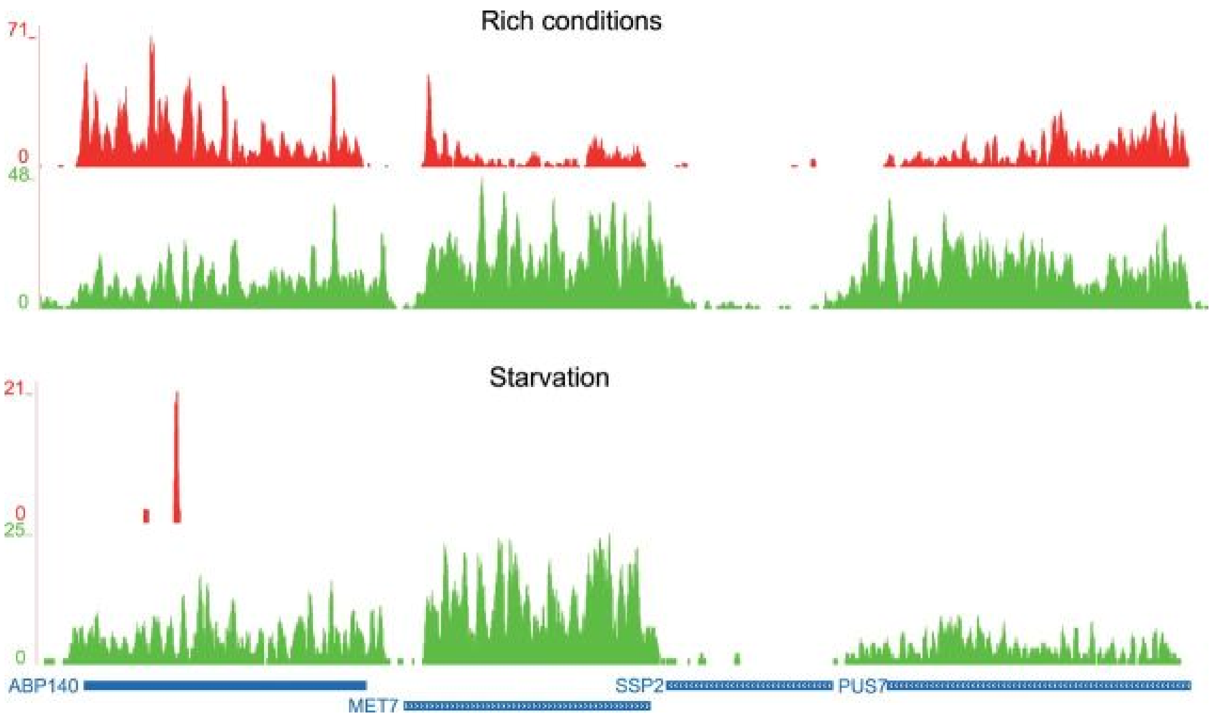


Figure 1. Observing differential translation in GWIPS-viz. Ribo-seq (red) and RNA-seq (green) coverage plots for the S. cerevisiae genome locus containing ABP140, MET7, SSP2

and PUS7 genes from (2). Under starvation conditions (right panel), ABP140, MET7 and PUS7 are transcribed, but not trnslated.

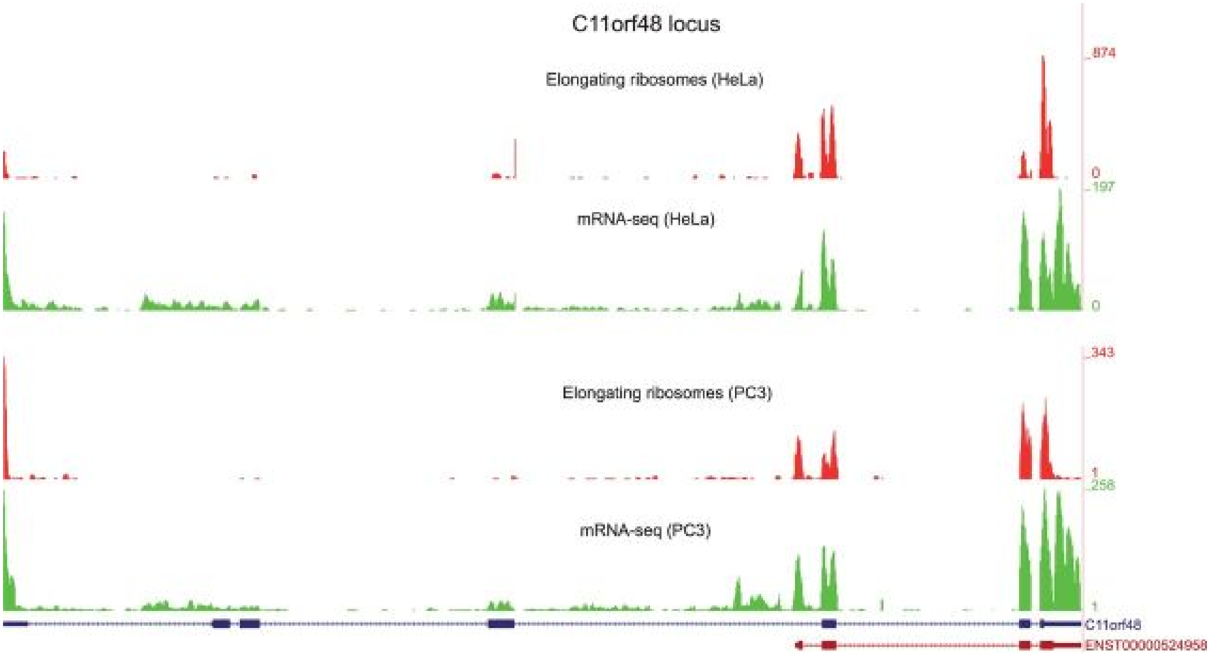


Figure 2. Comparing profiles from independent studies. Data from different studies and different organisms can be compared in GWIPS-viz. The C11orf48 locus in the human genome is shown where translation of an ENSEMBL transcript (brown bars) not annotated in RefSeq (blue bars) has been identified in HeLa cells (26). As can be seen, translation of the Ensembl transcript occurs in both, HeLa (3) and human PC3 cells (6).

Users can visually identify which isoform(s) of a gene is transcribed and translated and also compare translation of the gene between different ribo-seq studies. For example, Figure 2 provides a comparison of two ribo-seq datasets obtained in different tissuecultured human cells, HeLa (3) and PC3 cells (6). It can be seen that translation of a nonRefseq ENSEMBL transcript, reported based on the analysis of HeLa cell data (26), is observed in both datasets

For the eukaryotic datasets, ribosome profiles display the number of footprint reads at a particular genomic coordinate that align to the A-site (elongating ribosomes) or P-site (initiating ribosomes) of the ribosome, depending on the study. For the prokaryotic datasets, a weighted centred approach (17) is used to indicate the positions of ribosomes. Figure 3 shows ribosome profile densities in a region of the E. coli genome that includes the gene dnaX (b0470). The ribosome density is scaled relative to the maximum density present within the displayed genomic segment. As a result at the zoom allowing visualization of neighbour genes (top), dnaX appears as lowly expressed. However, at a range covering only the dnaX locus, it can be seen that nearly all codons in the dnaX mRNA are covered with footprints. Moreover the coverage is sufficient to allow visual detection of decreased ribosome density downstream of the site of programmed ribosomal frameshifting which is known to causes about 50% of translating ribosomes to terminate prematurely (27,28).

Figure 4 provides an example of how ribo-seq tracks for elongating and initiating ribosomes can be compared. The example illustrates the data obtained in Human HEK293 cells (7) mapped to TOMM6 and SFPQ genes, the latter gene apparently uses two sites of translation initiation for its expression.

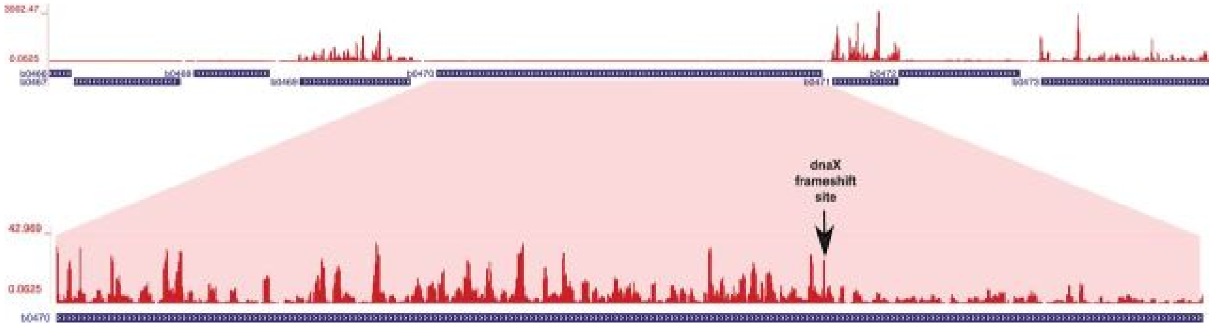


Figure 3. Ribo-seq data for the dnaX locus in the E.coli genome. The top panel corresponds to a segment containing neighbouring genes. The bottom panel contains the dnaX coordinates only. The displayed ribosome density is scaled relative to the maximum density within the selected region. The position of the programmed ribosomal frameshifting site in dnaX is indicated with an arrow.

DATABASE DESIGN AND IMPLEMENTATION

GWIPS-viz is a customized version of the UCSC Genome Browser (25) version 269, and runs on Ubuntu Linux version 12.04.1, with Apache version 2.2.22 and MySQL 5.2.24. Static HTML and CSS files of the UCSC Genome Browser were downloaded from

http://hgdownload.cse.ucsc.edu/ and rehosted on our local server, while C source code for the CGI executables was downloaded and compiled using gcc 4.6.3. Selected parts of the

MySQL databases were synced from the UCSC browser for the majority of organisms included in GWIPS-viz.

Our partial mirror of the UCSC Genome Browser hosted on our server displays tracks for human (hg19), mouse (mm10), S. cerevisiae (sacCer3), zebrafish (danRer7), C. elegans (ce10), Escherichia coli K12 (eschColi_K12), Bacillus subtilis (baciSubt2), human cytomegalovirus (Human herpesvirus 5 strain Merlin (HHV5)) and Bacteriophage lambda (NC_001416) assemblies. While several genome assemblies are available for many of the organisms, we chose to include only the most recent genome assembly for each organism.

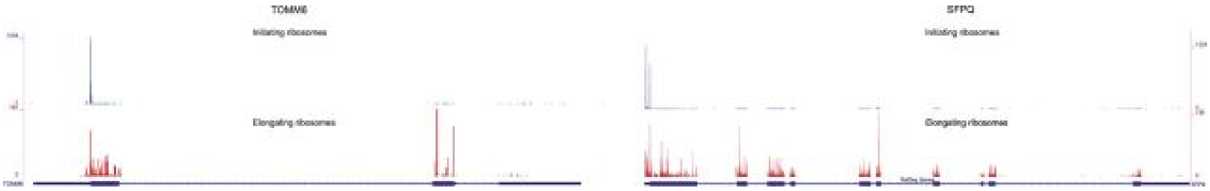


Figure 4. Combining profiles of initiating and elongating ribosomes. Profiles of initiating (blue) and elongating (red) ribosomes generated in Human HEK 293 cells (7). Locations of elongating and initiating ribosomes are consistent with the annotated coding region of the TOMM6 gene (left). However, ribosome profiles of the SFPQ gene points to the existence of an additional start codon (stronger peak) upstream of the annotated start codon (weaker peak).

Since the goal of GWIPS-viz is to be a browser for ribo-seq data, rather than a mirror of the UCSC browser, some of the functionality of the UCSC browser was removed in order to streamline the interface of GWIPS-viz. For example, the ‘clade’ menu in the genome selection menu was removed. In the browser window, a link was added in the top bar to allow the user to view the current genome position in the UCSC browser.

Depending on the organism, certain tracks were retained from the UCSC browser (25) and were consolidated into one group called ‘Annotation Tracks’. Examples include RefSeq (29), Ensembl (30), CCDS (31), Conservation (32), RepeatMasker (Smit et al., unpublished data, www.repeatmasker.org), Mouse ESTs (33), SGD genes (34), tRNA genes (35).

Ribo-seq and mRNA-seq tracks were added by incorporating the outputs of our RUM (36) alignment pipeline into the MySQL database. These tracks are divided into groups by publication and data type (ribo-seq and mRNA-seq). Tracks generated from uniquely mapping reads are colour coded according to their experiment type (elongating ribosome footprints are red, initiating ribosome footprints are blue, mRNA-seq reads are green).

Raw sequencing data retrieval

Published Ribo-seq and mRNA-seq datasets are downloaded from the NCBI Sequence Read Archive (SRA) (22) and converted to FASTQ format using the fastq-dump utility (SRA Handbook citation, not in PubMed). Data from replicate experiments are consolidated into one dataset so as to have one browser track for each experimental condition. An additional "All" track is generated for each study by aggregating the short reads from all available ribo-seq or mRNA-seq experiments for the given study.

Alignment pipeline

As there are no specific tools as yet for aligning ribo-seq data, RNA-seq tools are used in our pre-processing and alignment pipeline.

Depending on the study, adaptor linker sequence or poly-(A) tails are trimmed from the 3' ends of reads using Cutadapt version 1.1 (37). Trimmed reads shorter than 25 nucleotides are discarded.

Contamination from ribosomal RNA may account for a significant proportion of the raw reads even after depletion by subtractive hybridization during the experiment. Hence it is desirable to remove rRNA reads from the dataset before performing alignments in order to increase the proportion of informative sequences and improve alignment efficiency. To detect reads which are the result of ribosomal RNA contamination, trimmed reads are aligned to rRNA sequences using Bowtie (38). Bowtie version 0.12.8 is run using the -v option allowing three or fewer mismatches between the read sequence and the reference (rRNA) sequence. All reads that align to rRNA are discarded.

In most eukaryotes, a proportion of ribosome footprints will span splice junctions, i.e. the read will span the 3' end of one exon and the 5' end of another. There is the added complexity that ribo-seq reads are typically ~30 nucleotides in length. Hence the short-read alignment program needs to be capable of aligning reads of ~30nt across splice junctions.

We use the RNA-seq Unified Mapper (RUM), (current version 2.0.5_05) (36). RUM handles splice junctions by using the short read aligner Bowtie (38) to align sequence reads to both the genome and transcriptome and merging the results, before attempting to map remaining unaligned reads using another existing short-read aligner, BLAT (39).

Due to the relatively short lengths of ribosome footprint reads, a read may align to two or more distinct genomic locations due to sequence similarity. RUM outputs information separately for uniquely mapping reads and non-uniquely mapping reads (reads which align to several positions in the genome). Currently we provide tracks of uniquely mapping reads only in GWIPS-viz.

RUM's output files include a SAM alignment file showing the alignment(s) for each read, files giving the span of the alignment in genomic coordinates (RUM_Unique and RUM_NU) and coverage files (RUM.cov and RUM_NU.cov) listing the depth of coverage of reads across the genome.

The coverage files generated by the RUM alignment, RUM_Unique.cov and RUM_NU.cov, are in 4 column bedGraph format. The bedGraph data are converted into bigWig format, an indexed binary format that results in higher performance (40).

Ribosome profiles are generated from the RUM_Unique and RUM_NU files by obtaining the number of footprint reads whose 5' ends align at a given genomic coordinate (with an offset of 12nt designating the ribosome P-site for initiating ribosomes or 15nt for the ribosome A-site for elongating ribosomes).

FUTURE PERSPECTIVES

We plan to expand the existing repertoire of ribo-seq tracks by integrating publically available ribosome profiling experiments as they become available.

GWIPS-viz currently displays positions of the ribosomes mapped to the reference genomes. In case of eukaryotic organisms that extensively use RNA splicing, visualization of ribosome positions in GWIPS-viz could be problematic due to a large number of long exons. Therefore, visualization of ribosome positions mapped to individual RNA transcripts is among our top priorities.

We currently provide ribo-seq and mRNA-seq tracks of uniquely mapping reads only. In the future, we wish to provide a differential display that will incorporate non-unique mapping reads (mapping to two or more locations in the genome) with uniquely mapping reads.

We also aim to provide access to the Galaxy platform from within GWIPS-viz so that researchers who generate their own ribo-seq experimental data can pre-process and align their data with the tools provided within Galaxy and then view the alignments in GWIPS-viz.

Our overall objective is to continuously improve the service we provide in GWIPS-viz. As GWIPS-viz is under intensive development, some of the features described in this article could become outdated soon. Hence we encourage users to post their questions, comments and feedback on the GWIPS-viz forum.

REFERENCES

1. Weiss, R.B. and Atkins, J.F. (2011) Molecular biology. Translation goes global.

Science, 334, 1509-1510.

2. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. and Weissman, J.S. (2009) Genomewide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science, 324, 218-223.

3. Guo, H., Ingolia, N.T., Weissman, J.S. and Bartel, D.P. (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature, 466, 835-840.

4. Stadler, M. and Fire, A. (2011) Wobble base-pairing slows in vivo translation elongation in metazoans. RNA, 17, 2063-2073.

5. Reid, D.W. and Nicchitta, C.V. (2012) Primary role for endoplasmic reticulum-bound ribosomes in cellular translation identified by ribosome profiling. The Journal of biological chemistry, 287, 5518-5527.

6. Hsieh, A.C., Liu, Y., Edlind, M.P., Ingolia, N.T., Janes, M.R., Sher, A., Shi, E.Y., Stumpf, C.R., Christensen, C., Bonham, M.J. et al. (2012) The translational landscape of mTOR signalling steers cancer initiation and metastasis. Nature, 485, 55-61.

7. Lee, S., Liu, B., Huang, S.X., Shen, B. and Qian, S.B. (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proceedings of the National Academy of Sciences of the United States of America, 109, E2424-2432.

8. Fritsch, C., Herrmann, A., Nothnagel, M., Szafranski, K., Huse, K., Schumann, F., Schreiber, S., Platzer, M., Krawczak, M., Hampe, J. et al. (2012) Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. Genome research, 22, 2208-2218.

9. Shalgi, R., Hurt, J.A., Krykbaeva, I., Taipale, M., Lindquist, S. and Burge, C.B. (2013) Widespread regulation of translation by elongation pausing in heat shock. Molecular cell, 49, 439-452.

10. Liu, B., Han, Y. and Qian, S.B. (2013) Cotranslational response to proteotoxic stress by elongation pausing of ribosomes. Molecular cell, 49, 453-463.

11. Ingolia, N.T., Lareau, L.F. and Weissman, J.S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell, 147, 789-802.

12. Thoreen, C.C., Chantranupong, L., Keys, H.R., Wang, T., Gray, N.S. and Sabatini, D.M. (2012) A unifying model for mTORC1-mediated regulation of mRNA translation. Nature, 485, 109-113.

13. Bazzini, A.A., Lee, M.T. and Giraldez, A.J. (2012) Ribosome profiling shows that miR430 reduces translation before causing mRNA decay in zebrafish. Science, 336, 233237.

14. Stadler, M., Artiles, K., Pak, J. and Fire, A. (2012) Contributions of mRNA abundance, ribosome loading, and post- or peri-translational effects to temporal repression of C. elegans heterochronic miRNA targets. Genome research, 22, 2418-2426.

15. Brar, G.A., Yassour, M., Friedman, N., Regev, A., Ingolia, N.T. and Weissman, J.S. (2012) High-resolution view of the yeast meiotic program revealed by ribosome profiling. Science, 335, 552-557.

16. Gerashchenko, M.V., Lobanov, A.V. and Gladyshev, V.N. (2012) Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress.

Proceedings of the National Academy of Sciences of the United States of America, 109, 17394-17399.

17. Oh, E., Becker, A.H., Sandikci, A., Huber, D., Chaba, R., Gloge, F., Nichols, R.J., Typas, A., Gross, C.A., Kramer, G. et al. (2011) Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. Cell, 147, 1295-1308.

18. Li, G.W., Oh, E. and Weissman, J.S. (2012) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature, 484, 538-541.

19. Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V.T., Hein, M.Y., Huang, S.X., Ma, M., Shen, B., Qian, S.B., Hengel, H. et al. (2012) Decoding human cytomegalovirus.

Science, 338, 1088-1093.

20. Liu, X., Jiang, H., Gu, Z. and Roberts, J.W. (2013) High-resolution view of bacteriophage lambda gene expression by ribosome profiling. Proceedings of the

National Academy of Sciences of the United States of America, 110, 11928-11933.

21. Michel, A.M. and Baranov, P.V. (2013) Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. Wiley interdisciplinary reviews. RNA.

22. Shumway, M., Cochrane, G. and Sugawara, H. (2010) Archiving next generation sequencing data. Nucleic acids research, 38, D870-871.

23. Nekrutenko, A. and Taylor, J. (2012) Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nature reviews. Genetics, 13, 667-672.

24. Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J. et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome research, 15, 1451-1455.

25. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. Genome research, 12, 996-1006.

26. Michel, A.M., Choudhury, K.R., Firth, A.E., Ingolia, N.T., Atkins, J.F. and Baranov, P.V. (2012) Observation of dually decoded regions of the human genome using ribosome profiling data. Genome research, 22, 2219-2229.

27. Larsen, B., Gesteland, R.F. and Atkins, J.F. (1997) Structural probing and mutagenic analysis of the stem-loop required for Escherichia coli dnaX ribosomal frameshifting: programmed efficiency of 50%. Journal of molecular biology, 271, 47-60.

28. Tsuchihashi, Z. and Kornberg, A. (1990) Translational frameshifting generates the gamma subunit of DNA polymerase III holoenzyme. Proceedings of the National Academy of Sciences of the United States of America, 87, 2516-2520.

29. Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research, 33, D501-504.

30. Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T. et al. (2002) The Ensembl genome database project. Nucleic acids research, 30, 38-41.

31. Pruitt, K.D., Harrow, J., Harte, R.A., Wallin, C., Diekhans, M., Maglott, D.R., Searle, S., Farrell, C.M., Loveland, J.E., Ruef, B.J. et al. (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome research, 19, 1316-1323.

32. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. and Siepel, A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome research, 20, 110-121.

33. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2004) GenBank: update. Nucleic acids research, 32, D23-26.

34. Cherry, J.M., Ball, C., Weng, S., Juvik, G., Schmidt, R., Adler, C., Dunn, B., Dwight, S., Riles, L., Mortimer, R.K. et al. (1997) Genetic and physical maps of Saccharomyces cerevisiae. Nature, 387, 67-73.

35. Pavesi, A., Conterio, F., Bolchi, A., Dieci, G. and Ottonello, S. (1994) Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions. Nucleic acids research, 22, 1247-1256.

36. Grant, G.R., Farkas, M.H., Pizarro, A.D., Lahens, N.F., Schug, J., Brunk, B.P., Stoeckert, C.J., Hogenesch, J.B. and Pierce, E.A. (2011) Comparative analysis of RNA-Seq

alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics, 27, 2518-2528.

37. Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal, 17, 10-12.

38. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memoryefficient alignment of short DNA sequences to the human genome. Genome biology, 10, R25.

39. Kent, W.J. (2002) BLAT--the BLAST-like alignment tool. Genome research, 12, 656664.

40. Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S. and Karolchik, D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics, 26, 22042207.