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

Alexei V Lobanov¹,⁷, Stephen M Heaphy²,⁷, Anton A Turanov¹, Maxim V Gerashchenko¹, Sandra Pucciarelli³, Raghu R Devaraj³, Fang Xie⁴, Vladislav A Petyuk⁴, Richard D Smith⁴, Lawrence A Klobutcher⁵, John F Atkins², Cristina Miceli³, Dolph L Hatfield⁶, Pavel V Baranov² & Vadim N Gladyshev¹

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, and +1 frameshifting at AAA was the most frequent. The frameshifts showed unusual plasticity and rapid evolution, and had little influence on translation rates. The 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 probably requires additional factors.

There are several known mRNAs for which translating ribosomes shift the reading frame at specific locations with a high efficiency; in very rare cases, the frameshifting frequency may even exceed the frequency of concurrent standard translation. This phenomenon is known as programmed ribosomal frameshifting and is observed primarily in virus decoding¹. Whereas programmed ribosomal frameshifting is an omnipresent translation process, it is usually considered to be a recoding mechanism. Recoding describes alterations in genetic decoding that take place at specific locations within particular mRNAs and are distinguished from codon reassignment². Except in the case of 40%-efficient programmed ribosomal frameshifting at a heptanucleotide site used during expression of the Ty1 transposon in *Saccharomyces cerevisiae³*, complex stimulatory signals, such as RNA pseudoknots, are required for a high efficiency of programmed ribosomal frameshifting⁴.

However, previous analyses of several sequenced genes of *Euplotes* ciliates have suggested that +1 ribosomal frameshifting may be more common in these organisms (reviewed in ref. 5). 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 has evolved as a consequence of TGA-codon reassignment from stop to cysteine, thus weakening release-factor recognition of the remaining stop codons, TAA and TAG⁵,⁶. Furthermore, it has been shown experimentally in a hybrid system that *Euplotes* release factors indeed recognize these stop codons inefficiently⁶.

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. focardii⁷,⁸. 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. The results revealed that ribosomes of the *Euplotes* ciliates are characterized by an inability to terminate at stop codons in internal positions of coding sequences, thus leading to frameshifting at these signals, whereas termination probably requires additional components in these organisms and occurs only at specific mRNA positions.

**RESULTS**

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

*Euplotes* DNA, similarly to that of other ciliates, 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 *Euplotes⁹,¹⁰*. These chromosomes are generated from the micronuclear DNA during the sexual cycle¹¹. 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.

¹Division of Genetics, Department of Medicine, Brigham and Women’s Hospital and Harvard Medical School, Boston, Massachusetts, USA. ²School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland. ³School of Biosciences and Veterinary Medicine, University of Camerino, Camerino, Italy. ⁴Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington, USA. ⁵Department of Molecular Biology and Biophysics, University of Connecticut Health Center, Farmington, Connecticut, USA. ⁶Molecular Biology of Selenium Section, Mouse Cancer Genetics Program, Center for Cancer Research, National Institutes of Health, Bethesda, Maryland, USA. ⁷These authors contributed equally to this work. Correspondence should be addressed to V.N.G. (vgladyshev@rics.bwh.harvard.edu) or P.V.B. (p.baranov@ucc.ie).

Received 11 March; accepted 31 October; published online 21 November 2016; doi:10.1038/nsmb.3330
Ribosomal frameshifting at AAATAA. (b) Beneath each plot. Stop codons (and adjacent upstream codons) where termination or frameshifting occur are also indicated. (a) Start (ATG, green vertical lines) and stop codons (TAA and TAG, red lines) are shown in each of the three reading frames for chromosomes. Splicing and noncanonical splice junctions. Some short introns, if accurate prediction of introns is complicated by instances of alternative splicing and noncanonical splice junctions. The general properties of the genomes are described in Supplementary Figure 1.

A large number of very short (20–30 nt) introns is a characteristic feature of macronuclear protein-coding genes in some ciliates, but accurate prediction of introns is complicated by instances of alternative splicing and noncanonical splice junctions. Some short introns, if not detected by annotation pipelines, may result in open reading frame disruption and thus may be misinterpreted as frameshift sites. To account for this possibility, we used experimentally confirmed, rather than predicted, mRNA transcripts (Supplementary Fig. 2).

Identification of ribosomal frameshifting through phylogenetic, ribosome-profiling and proteomic analyses

To identify sites of ribosomal frameshifting and to estimate frameshifting 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. It provides information on ribosome locations and their densities at the whole-transcriptome level. 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 the −1 or +1 direction, we compared the span of Ribo-seq coverage with ORF organization (Fig. 1). In certain cases in which unambiguous discrimination between potential events was difficult, we sought additional information. Using the basic local alignment search tool (BLAST), we explored which of the potential products were 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. We compared the sequences of the transcripts with 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).

To verify putative sites of frameshifting and to determine the associated mechanisms (i.e., the direction and identity of amino acids incorporated...
frameshifting sites in nature structural & molecular biology

frameshift sites (ribosome footprint density and BLAST hit alignments) stop. Altogether, we observed 47 out of 62 possible sense codons at the approximately three-quarters (1,368) consisted of an AAA codon followed by a

Among 1,765 putative frameshift sites detected with Ribo-seq, approximately three-quarters (1,368) consisted of an AAA codon followed by a

Figure 2 Identification of amino acids inserted at frameshift sites. (a) Lysine (K) and asparagine (N) are inserted at the AAATAA heptamer. Nucleotide sequence surrounding the AAATAA +1 frameshift site is shown in the middle. Amino acid sequence is shown above the schematic of reading frames. Peptide and frameshift site:

comp3835_c0_seq1 +1/6
comp3741_c0_seq1 +1/1
comp5116_c0_seq1 +1/2
comp7073_c0_seq1 +1/1
comp8412_c0_seq1 +2/4
comp3853_c0_seq1 +2/1
comp6651_c0_seq1 +2/1
comp973_c0_seq1 +2/0

Peptide and frameshift site:

comp3835_c0_seq1 +1/6
comp3741_c0_seq1 +1/1
comp5116_c0_seq1 +1/2
comp7073_c0_seq1 +1/1
comp8412_c0_seq1 +2/4
comp3853_c0_seq1 +2/1
comp6651_c0_seq1 +2/1
comp973_c0_seq1 +2/0

Supplementary Note 3 Earlier observations of frequent use of AAA TAA and AAA TAG as frameshifting, a result consistent with our observation of ribosomal frameshifting. XCorr and ∆Cn are the SEQUEST algorithm metrics reflecting the goodness of match between the predicted fragmentation pattern based on peptide sequence and the observed one. (∆Cn) Peptides detected by MS/MS analysis that were derived from the translation of 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 approximately eight times more frequently than expected (Fig. 3a). Moreover, six out of seven 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 peptidyl (P)-site tRNA and its codon in the initial frame increases the possibility of frameshifting. We also found that all XXX codons (i.e., codons with identical nucleotides) were also enriched (relative to non-AAA codons) at the frameshift sites (Fig. 3a), even though CCC and GGG were not the most frequent codons, owing to a relatively low GC content of Euplotes genomes. This finding suggests that the ability of P-site tRNAs to base-pair with a codon in +1 frameshifting also increases the chances of frameshifting because XXX codons would re-pair with XXT, thereby forming perfect Watson–Crick interactions with the first two subcodon positions.

Interestingly, XYX codons (with the same nucleotides at the first and third subcodon positions, but a different nucleotide in the second subcodon position) supported +2 ribosome frameshifting (ribosome density profile for an mRNA containing an ATA TAA frameshift site in Fig. 1c). It appears that the ribosomes shifted into the −1 frame. However, we found the mechanism to be +2 frameshifting on the basis of the MS/MS analysis (Supplementary Note 1). Additionally, +2 frameshifting appeared 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 nine XYX codons (out of 16 possible) in the +2 frameshift sites (Fig. 3a), and ATA was the most frequent. The other codons that appeared to support +2 frameshifting were XTA, which have T and A in the +2 and +3 positions.

Unexpectedly, 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; example of termination at AAA TAA in Supplementary Fig. 3a). Therefore, it is clear that whether the ribosome terminates at a particular stop codon does not depend solely on the identity of the 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 poly(A). This finding is consistent with Euplotes having very short 3′ untranslated regions (UTRs). Some mRNAs require longer 3′ UTRs; for example, selenoprotein mRNAs need to accommodate selenocysteine insertion sequence (SECIS) elements (Supplementary Fig. 3b). However, the ‘distance’ between the poly(A) tail and the genuine site of termination may be structural rather than sequence based, such that the SECIS structure could bring the poly(A) 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 frameshifting is not as detrimental in *Euplotes* as it is in other organisms. Metagene analysis (Fig. 4a and corresponding RNA-seq density in Supplementary Fig. 4) revealed similar ribosome density upstream and downstream of frameshift sites. Therefore, the efficiency of frameshifting was comparable to that of standard decoding. However, there was a substantial drop in density relative to stop codons identified as termination sites (Fig. 4b). At the same time, a peak of ribosome density was also present ~30 nt upstream of frameshift sites (Fig. 4a), the distance roughly corresponding to the distance between the aminoacyl (A) sites of the two stacked ribosomes. Such stacking would be expected if ribosomal frameshifting were slower than the standard decoding of sense codons. A slight depletion of ribosomes has also been observed immediately downstream of the frameshift sites (Fig. 4a). Therefore, it is plausible that, whereas ribosomal frameshifting does not impose considerable costs on the accuracy of synthesized proteins (for example, AAA TAA A would be decoded in the same way as AAA AAA), there is a cost to the speed of the ribosome and a subsequent increase in the number of ribosomes per mRNA. In this case, frameshifting would be expected to be harmful in genes expressed at high levels.

To test this hypothesis, we explored how frameshifting relates to gene expression levels, on the basis of RNA-seq and Ribo-seq signals (Fig. 4c,d). Indeed, we found that frameshifting was less frequent in highly expressed genes, a result 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) values, 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 to anticorrelate with elongation rates at that location. Therefore, although 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.

Because we found that particular codons are the most frequent at the frameshifting sites (mononucleotide and AT rich, with AAA being most overrepresented), we hypothesized that the frameshifting efficiency might vary depending on the identity of the codon upstream of a stop. To verify this hypothesis, we divided frameshifting sites into AAA and non-AAA groups and analyzed the distribution of footprint densities (Fig. 5a,b). The ribosome density did not change significantly downstream of frameshifting sites for either AAA or non-AAA frameshifting sites (Fig. 5c), although the pause at non-AAA-containing sites was 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. Although we have not observed a specific nucleotide context associated with non-AAA codons at the frameshifting sites, we noticed that TAG...
occurred almost three times more frequently (~29%) at non-AAA frameshifting sites than at AAA frameshifting sites (~12%) (Fig. 5a,b).

To analyze how TAA and TAG stop codons affect frameshifting, we compared footprint densities at the frameshifting sites according to which stop codon is used (Fig. 5d,e). Although we did not find significant differences in the changes in 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).

Frameshift patterns do not evolve under strong purifying selection

In most well-studied cases of programmed ribosomal frameshifting (for example, eukaryotic antizymes and bacterial release factor 2), the frameshifting sequence and its occurrence are remarkably conserved. In fact, evolutionary conservation of frameshift patterns is frequently used for the detection of recoded genes. In all these cases, the efficiency of frameshifting is below 100%, and two protein products are usually synthesized from the same mRNA: one is decoded according to the rules of standard genetic decoding, and the other is a product of frameshifting. The ratio between these two products is functionally important and is often tightly regulated. Therefore, there is a strong evolutionary pressure to preserve the frameshifting site and its regulatory capacity, thus leading to strong stabilizing selection acting on the sequences of frameshift sites and stimulatory signals. In contrast, frameshifting in two *Euplotes* species has often been characterized by cases in which only one of the two orthologous sites uses frameshifting (typical example in Fig. 6a). Whereas 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 in protein-coding regions, owing to their dramatic effects on the sequence of synthesized protein. In *Euplotes*, however, certain indels that could create an efficient site of ribosomal frameshifting irrespective of nucleotide context (for example, AAA AAA to AAA TAA A mutation) may be expected to 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 the evolution of indels, we analyzed the frequencies of sequences surrounding single-nucleotide indels. We generated pairwise alignments of orthologous sequences from the transcriptomes of both species by using Fasta and counted occurrences of each hexamer in

---

**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. The first nucleotide of a stop codon is shown as the zero coordinate. Whereas ribosome density upstream and downstream of frameshift sites is similar, there is a peak of density at the frameshift sites, which is accompanied by another peak 30 nt upstream. A sequence logo below represents the information content of sequences used for metagene alignment. The sequence AAATAA is predominant, and there are no other position-specific signals associated with frameshifting. “K” denotes thousand. (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 nt (poly(A) is not included) were used. (c) Frequencies of transcripts with the sites of ribosomal frameshifting (x axis) versus the transcripts ranked on the basis of the levels of protein synthesis (Ribo-seq density, y axis). (d) Similar to c, but with ranking based on RNA levels (RNA-seq density). (e) Distribution of transcripts with different Ribo-seq to RNA-seq ratios containing frameshift (FS) sites (red) or not containing frameshift sites (black).
which 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). The abundance of patterns matching AAATAA was striking (Fig. 6b). Indels in the center of the AAATAA pattern were strongly overrepresented in comparison with other patterns in both species, thus suggesting that frameshifting in Euplotes evolves essentially neutrally, thereby producing AAA-stop frameshifting sites, though this is unlikely to be the case for non-AAA frameshifting sites.

**Figure 5** Comparison of ribosomal frameshifting at AAA versus non-AAA frameshifting sites and TAA versus TAG frameshifting sites. (a–d) Aggregated densities of ribosome footprints around frameshift sites containing an AAA codon preceding stop (a), non-AAA codons (b), TAA stop codons (c) and TAG stop codons (d). (e,f) Comparison of footprint density changes observed at frameshift sites at each mRNA (D3 region) and downstream of frameshift sites (D2) relative to the footprint density upstream of frameshift sites (D1). D1 and D3 regions were chosen 60 nt upstream and downstream of frameshift sites to avoid aberrant densities inflicted by ribosome pauses at frameshifting sites. Box plots represent ratio distributions: horizontal line, median; box limits, 25th and 75th percentiles; whiskers, 5th and 95th percentiles. The comparison was carried out for AAA-containing (n = 1,368) versus non-AAA-containing (n = 397) frameshift sites (e) and TAA-containing (n = 1,488) versus TAG-containing (n = 277) frameshift sites (f). P values were calculated with unpaired two-tailed Wilcoxon rank-sum tests 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.
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. focardi*. (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. focardi* (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 ending with either TAA or TAG are shown in red. Two most frequent hexamers, AAATAA and AAATAG, are indicated.

**DISCUSSION**

In this work, we provide multiple lines of evidence for the frequent occurrence of ribosomal frameshifting during translation in *Euplotes* ciliates. Ribosomal frameshifting occurs at the stop codons at which tRNAs in the P site slip forward, predominantly by 1 or 2 nt. 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 in which frameshifting sites were predicted on the basis of genomic and transcriptomic sequences22, thus 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 little effect on the accuracy of protein synthesis, though it probably 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 poly(A) tails and that additional mechanisms are required for efficient termination. Thus, the presence of a stop codon is not a feature sufficient for translation termination in *Euplotes*. Instead, the default function of stop codons is ribosomal frameshifting. This finding is consistent with recent findings of reassignment of all stop codons in *Condylostoma magnum*, in which stop codons function as terminators only in proximity to mRNA 3' ends23,24. The substantial evolutionary distance between *Euplotes* and *Condylostoma* suggests the intriguing possibility that position-dependent termination may be a general property of ciliate decoding, thus potentially explaining the high frequency of changes in the genetic code in these species. A degree of positional preference in translation termination in other eukaryotes requires further exploration.

**METHODS**

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

**ACKNOWLEDGMENTS**

Supported by NIH GM061603 and GM065402 to V.N.G. S.M.H. and P.V.B. are supported by the grants from Wellcome Trust (094423) and Science Foundation Ireland (12/IA/1335). Portions of this research were also supported by NIH GM103493 and the WR. Wiley Environmental Molecular Science Laboratory (sponsored by DOE and located at Pacific Northwest National Laboratory). Pacific Northwest National Laboratory is operated by the Battelle Memorial Institute under the DOE contract DE-AC05-76RL01830. C.M. acknowledges the Italian PNRA and the COST action BM1102 for supporting a part of this work.

**AUTHOR CONTRIBUTIONS**

A.V.L., S.M.H., P.V.B. and V.N.G. analyzed the data and wrote the paper with advice from D.L.H. and J.F.A.; A.A.T. and M.V.G. prepared samples for sequencing; S.P., R.B.D., C.M. and L.A.K. performed cell culture maintenance and growth, E.X., V.A.P. and R.D.S. conducted MS analysis. All authors discussed the results and implications and commented on the manuscript at all stages.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.
19. Ivanov, I.P. & Atkins, J.F. 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 (2007).

20. Sharma, V. *et al.* 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 (2011).

21. Pearson, W. Finding protein and nucleotide similarities with FASTA. *Curr Protoc.* *Bioinformatics* **4**, 3.9 (2004).

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

23. Swart, E.C., Serra, V., Petroni, G. & Nowacki, M. Genetic codes with no dedicated stop codon: context-dependent translation termination. *Cell* **166**, 691–702 (2016).

24. Heaphy, S.M., Mariotti, M., Gladyshev, V.N., Atkins, J.F. & Baranov, P.V. Novel ciliate genetic code variants including the reassignment of all three stop codons to sense codons in *C. magnum*. *Mol. Biol. Evol.* **33**, 2885–2889 (2016).

ARTICLES
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 2,866,128,506 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 data set, we used DeconSeq. The following data sets were used: bacterial genomes (2,206 unique genomes, accessed on 02/12/11), archaean genomes (155 unique genomes, accessed on 02/12/11), *Salmomella enterica* genomes (52 strains, accessed on 12/16/10), bacterial genomes HMP (76,337 WGS sequences, accessed on 02/12/11), and viral genomes in RefSeq 45 (3,761 unique sequences, accessed on 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*, threshold values were changed to 50% coverage and 80% identity. Bacterial sequences in the genome data were not unexpected, given that both endosymbionts and endosymbionts have been reported in ciliates.

Several assembly programs were used to generate independent whole-genome assemblies, including ABYSS27, SOAP28, SSAKE29, Velvet30, Celera31, and ABySS27, SOAP28, SSAKE29, Velvet30, Celera31, producing 9,620,943 reads. The reads were aligned to the transcriptome with SAMtools package48, custom scripts and IGV40. Sequence pattern analysis.

Ribo-seq 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 X-100, 100 mg/l cycloheximide and 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, 100 mg/l cycloheximide and 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, 100 mg/l cycloheximide and protease inhibitors from Roche.

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 X-100, 100 mg/l cycloheximide and 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, 100 mg/l cycloheximide and 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, 100 mg/l cycloheximide and 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, 100 mg/l cycloheximide and protease inhibitors from Roche.

Proteomics and Ribo-seq analyses. Proteomics analysis used a conventional shotgun bottom-up approach described elsewhere41–43. Briefly, cells were resuspended in the lysis buffer (50 mM Tris-HCl, pH 8.0, 8 M urea, 10 mM DTT and 1 mM EDTA), pulverized in liquid nitrogen and then subjected to melting and sonication in a water bath for 1 min. The proteins were then digested with trypsin (samples 1 and 2) or Glu-C (sample 3, pH 7.5), then subjected to fractionation by SCX (trypsin sample, 25 fractions collected) and high-pH RP (trypsin and Glu-C samples; 24 concatenated fractions collected44). Analysis by liquid chromatographycoupled with LTQ Orbitrap (Thermo Fisher) mass spectrometry (LC–MS/MS) was performed on a 100-min LC gradient. The details of the gradient and mass spectrometer settings can be found elsewhere45. The data were preprocessed with DeconMSn46 and DtaRefinery46 tools and analyzed with MS-GF+47. The raw peak lists and MS/MS identification files have been deposited in the PRIDE database (PXD004333). Among all the 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 a maximum number of identifications without exceeding a false discovery rate of 5%. Spectra for peptides spanning the frameshift locations were manually verified. The details of MS/MS data analysis along with parameter files and an executable document reproducing all the post-search analysis steps have been deposited as an R package at https://github.com/vladpetyuk/EuclotesCrassus.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 lyase (measured at A260). After RNA digestion, sequencing libraries were prepared as previously described47, starting with gradient ultracentrifugation. Sequencing was performed on an Illumina HiSeq2000 platform.

**Sequence pattern analysis.** To analyze the frequency of indels that occurred since *E. crassus* and *E. focardii* diverged from their common ancestor, we generated a set of pairwise alignments by using FASTA21. The alignments were generated using Bowtie software v0.12.7 (ref. 39); up to two mismatches per read were allowed. We estimated positions of the ribosome A sites with an offset of 15 nt downstream of 5′ end of Ribo-seq data. Visualization and further manual analysis were conducted with the SAMtools package48, custom scripts and IGV48.

**Sequence pattern analysis**. To analyze the frequency of indels that occurred since *E. crassus* and *E. focardii* diverged from their common ancestor, we generated a set of pairwise alignments by using FASTA21. The alignments were generated using Bowtie software v0.12.7 (ref. 39); up to two mismatches per read were allowed. We estimated positions of the ribosome A sites with an offset of 15 nt downstream of 5′ end of Ribo-seq data. Visualization and further manual analysis were conducted with the SAMtools package48, custom scripts and IGV48.
rank-sum tests. The exact P values and degrees of freedom are provided in the figure legend.

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 and SRA SRP078897; for E. focardii, BioProject PRJNA329414, BioSample SAMN05412809 and SRA SRP078891. Proteomics data have been deposited in the PRIDE database (PXD004333). The whole-genome shotgun projects have been deposited at DDBJ/ENA/GenBank under accession codes MECR00000000 (Euplotes crassus, version MECR01000000) and MJUV00000000 (Euplotes focardii, version MJUV01000000). The interpretations of sequence data, such as coordinates of frameshifting sites are available upon request.

25. Schmieder, R. & Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One 6, e17288 (2011).
26. Dziallas, C., Allgaier, M., Monaghan, M.T. & Grossart, H.P. Act together-implications of symbioses in aquatic ciliates. Front. Microbiol. 3, 288 (2012).
27. Simpson, J.T. et al. ABySS: a parallel assembler for short read sequence data. Genome Res. 19, 1117–1123 (2009).
28. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1, 18 (2012).
29. Warren, R.L., Sutton, G.G., Jones, S.J. & Holt, R.A. Assembling millions of short reads—a practical guide. Curr. Protoc. Mol. Biol. 11, 3624–3639 (2013).
30. Gerashchenko, M.V., Lobanov, A.V. & Gladyshev, V.N. Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. Proc. Natl. Acad. Sci. USA 109, 17394–17399 (2012).
31. Myers, E.W. SOAPdenovo: A new parallel assembler for next-generation short-read sequencing. Nat. Biotechnol. 29, 890–894 (2011).
32. Gerashchenko, M.V., Lobanov, A.V. & Gladyshev, V.N. Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. Proc. Natl. Acad. Sci. USA 109, 17394–17399 (2012).
33. Huang, X. & Yang, S.P. Generating a genome assembly with PCAP. Gigascience 1, 18 (2012).
34. Huang, X., Wang, J., Aluru, S., Yang, S.P. & Hillier, L. PCAP: a whole-genome assembly program. Genome Res. 13, 2164–2170 (2003).
35. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955–964 (1997).
36. Laslett, D. & Canback, B. ARAGORN, a program to detect RNA genes and tRNA genes in nucleotide sequences. Nucleic Acids Res. 32, 1–16 (2004).
37. Simpson, J.T. ABySS: a parallel assembler for short read sequence data. Nat. Protoc. 8, 1494–1512 (2013).
38. Haas, B.J. et al. De novo transcript sequence reconstruction from RNA-Seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8, 1494–1512 (2013).
39. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25 (2009).
40. Thorvaldottir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief. Bioinform. 14, 178–192 (2013).
41. Depuydt, G. et al. Reduced insulin/insulin-like growth factor-1 signaling and dietary restriction inhibit translation but preserve muscle mass in Caenorhabditis elegans. Mol Cell Prot 12, 3624–3639 (2013).
42. Depuydt, G. et al. 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 (2014).
43. Petyuk, V.A. et al. Characterization of the mouse pancreatic islet proteome and comparative analysis with other mouse tissues. J. Proteome Res. 7, 3114–3126 (2008).
44. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25 (2009).
45. Mayampurath, A.M. et al. DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. Bioinformatics 24, 1021–1023 (2008).
46. Petyuk, V.A. et al. DtaRefinery, a software tool for elimination of systematic errors from parent ion mass measurements in tandem mass spectra data sets. Mol Cell Prot 9, 486–496 (2010).
47. Kim, S. & Pevzner, P.A. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun. 5, 5277 (2014).
48. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).