PNAS RESEARCH ARTICLE | EVOLUTION OPEN ACCESS

Nontriplet feature of genetic code in *Euplotes* ciliates is a result of neutral evolution

Sofya A. Gaydukova[a,1,2], Mikhail A. Moldovan[b,1,3], Adriana Vallesi[c], Stephen M. Heaphy[d,4], John F. Atkins[d,e], Mikhail S. Gelfand[b,5], and Pavel V. Baranov[d,5]

Edited by Eugene Koonin, NIH, Bethesda, MD; received December 21, 2022; accepted April 12, 2023

The triplet nature of the genetic code is considered a universal feature of known organisms. However, frequent stop codons at internal mRNA positions in *Euplotes* ciliates ultimately specify ribosomal frameshifting by one or two nucleotides depending on the context, thus posing a nontriplet feature of the genetic code of these organisms. Here, we sequenced transcriptomes of eight *Euplotes* species and assessed evolutionary patterns arising at frameshift sites. We show that frameshift sites are currently accumulating more rapidly by genetic drift than they are removed by weak selection. The time needed to reach the mutational equilibrium is several times longer than the age of *Euplotes* and is expected to occur after a several-fold increase in the frequency of frameshift sites. This suggests that *Euplotes* are at an early stage of the spread of frameshifting in expression of their genome. In addition, we find the net fitness burden of frameshift sites to be noncritical for the survival of *Euplotes*. Our results suggest that fundamental genome-wide changes such as a violation of the triplet character of genetic code can be introduced and maintained solely by neutral evolution.

ciliates | genetic code | ribosomal frameshifting

The sequential nonoverlapping triplet nature of genetic decoding was established by Crick, Brenner and their colleagues in early 60s (1). Almost all proteins are encoded by such sequential nucleotide triplets, codons. Thus, the decoding ribosome moves along mRNA in one of the three-periodic phases known as the reading frames. Errors in maintaining the reading frame are more detrimental than missense errors as they affect the entire downstream part of the protein (2). Consequently, spontaneous shifts between reading frames are highly infrequent (3). As the accuracy of triplet decoding is sequence-dependent, frameshifting-prone sequences are selected against in protein-coding genes (4, 5). However, the sequence dependence of frameshifting efficiency enabled evolution of genes that exploit this phenomenon to regulate their expression in the process known as programmed ribosomal frameshifting (6). Although genes requiring ribosomal frameshifting for their expression have been found in most organisms, such genes are generally extremely rare, though common in viruses (7) and transposable elements (8). To achieve higher efficiency, programmed ribosomal frameshifting often requires the presence of elaborate stimulatory signals such as RNA pseudoknots altering progression of the ribosome (9, 10) or nascent peptides interfering with the ribosome function from within (11, 12). Even with the assistance of such stimulators, the efficiency of ribosomal frameshifting is usually lower than that of the competing triplet decoding (6). Thus, the product of ribosomal frameshifting is synthesized in addition to the product of standard translation.

Ribosomal frameshifting observed during mRNA translation in ciliates of the genus *Euplotes* (13–17) is often described as programmed ribosomal frameshifting, but, as we will argue below, strikingly contrasts to the programmed ribosomal frameshifting in other organisms. It has been proposed to occur whenever a stop codon is encountered by the ribosome (Fig. 1*A*). Unlike programmed ribosomal frameshifting, it is highly efficient with virtually no products of termination at stop codons at internal positions being detected (13). Termination of translation occurs only near the 3′ ends of mRNAs in close proximity to the polyA tails (13) similarly to the situation in those species where all three stop codons have been reassigned to code for amino acids (18–23) as outlined in ref. 24. Therefore, ribosomal frameshifting has been suggested to be a part of the standard *Euplotes* genetic code (13) (Fig. 1*C*).

It is not clear, however, how such a nontriplet feature of the *Euplotes* genetic code has emerged and which processes enable its persistence. To address these intriguing questions, we explored the evolution of frameshift site (FS) occurrences across genomes of several *Euplotes* species.

Global and Local Alterations of the Genetic Code

The standard readout of the genetic information could be altered globally, affecting genetic code of an entire organism, or locally, affecting decoding of a specific mRNA (25). Global

Significance

In this work, we provide compelling evidence that *Euplotes* genetic code violates the triplet nature of the genetic decoding that was thought to be universal. Thus, *Euplotes* possess the most extreme example of genetic code variation described so far. The nontriplexy arises from abundant ribosomal frameshift sites with no regulatory function, where stop-codons distant from the 3′ transcript end specify +1 or +2 ribosomal frameshifting with high accuracy. We show that this violation of the triplet coding in *Euplotes* is brought about and further maintained by neutral evolution rather than selective processes but still is irreversible.

Competing interest statement: P.V.B. is a founder and a shareholder of EIRNA Bio.

This article is a PNAS Direct Submission.

[1] S.A.G. and M.A.M. contributed equally to this work.

[2] Present address: Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139.

[3] Present address: Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115.

[4] Present address: Life Sciences Deptartment, Enterprise Ireland, Bishopstown, Cork, Ireland, T12WCH2.

[5] To whom correspondence may be addressed. Email: mikhail.gelfand@gmail.com or p.baranov@ucc.ie.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2221683120/-/DCSupplemental.

Published May 22, 2023.

alterations involve a change of a molecular component that is required for decoding of most or all mRNAs in the cell or an organelle. For example, the loss of a gene encoding release factor 2 in vertebrate mitochondria resulted in the reassignment of UGA stop codon to tryptophan, so that all UGA codons in coding regions of mitochondrial mRNAs are decoded as tryptophan (26). While most codon reassignments involve stop codons (20, 27), sense codon reassignments have been also observed (28).

In addition, the meaning of a codon could also be altered locally in a specific mRNA. For example, the presence of a special RNA secondary structure, SECIS element, in the 3′ UTR of a eukaryotic mRNA can redefine a specific UGA codon in that mRNA to encode a selenocysteine (29). Selenocysteine is incorporated into the proteins synthesized from only 25 human genes whose mRNAs contain SECIS elements (30); mRNAs from other human genes lack SECIS and do not encode selenoproteins. An extreme example is the selenoprotein P mRNA in bivalve molluscs where 132 UGA codons are decoded as selenocysteines (31). To distinguish between global and local alterations of codon meanings, a change in codon meaning that affects the entire genetic code is termed *codon reassignment*, while a site- or mRNA- specific change of meaning is termed *codon redefinition* (25, 32). Dynamic codon redefinition is an instance of a more general phenomenon called *recoding* which encompasses numerous translational deviations from the standard genetic code occurring in the decoding of specific mRNAs with no overall effect on the genetic code (25, 33–35).

Programmed Ribosomal Frameshifting

One of the recoding mechanisms is programmed ribosomal frameshifting which is extremely rare in cellular genes (6). For example, in humans, translation of only a handful of genes of nonviral origin is known to utilize ribosomal frameshifting. These are three paralogous genes encoding ornithine decarboxylase antizyme

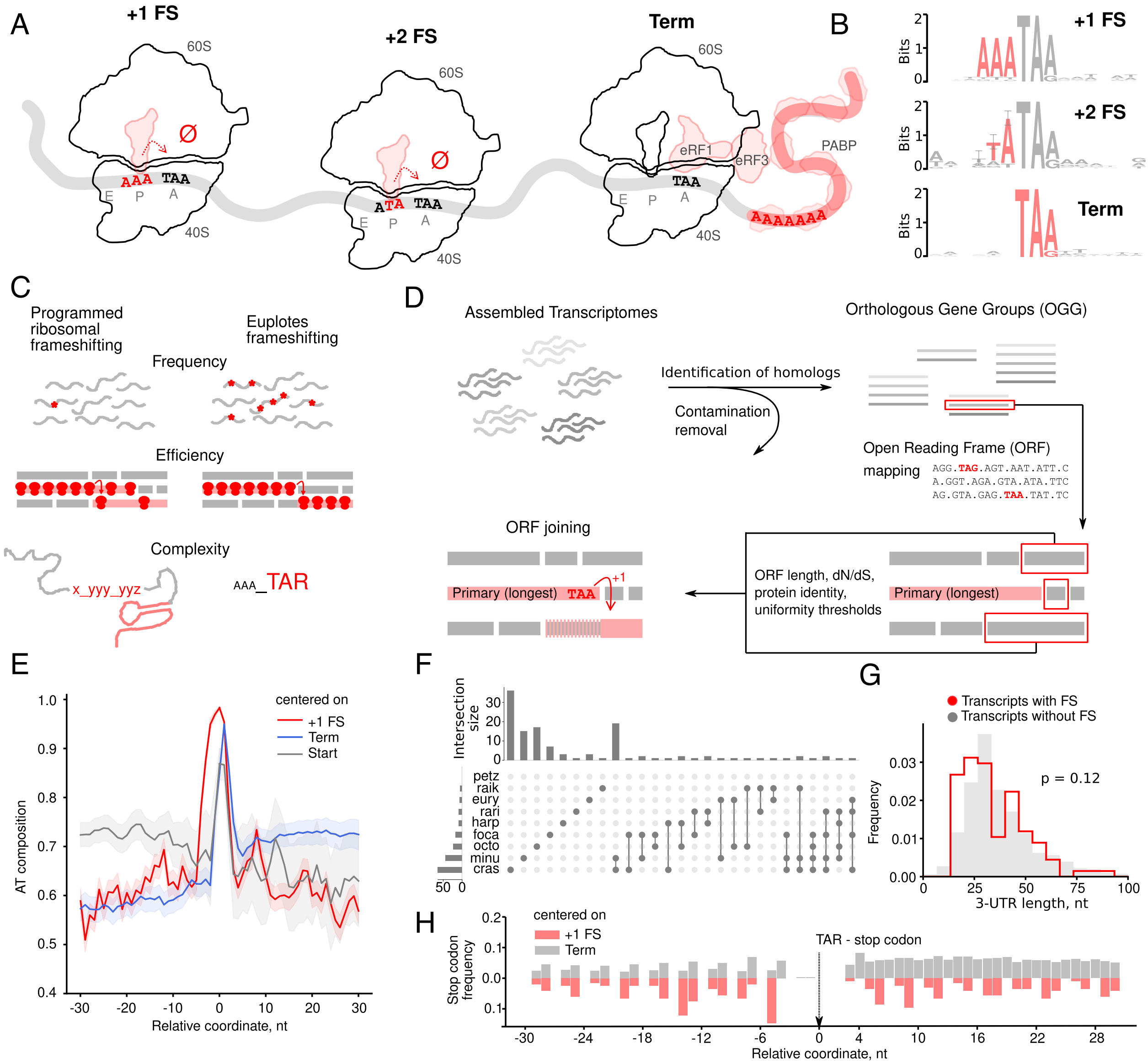

Fig. 1. *Euplotes* transcriptomes and identification of frameshift sites (T specifies uridine in RNA). (*A*) The proposed mechanism of ribosomal frameshifting for codons supporting tRNA repairing with either +1 (*AAA TAA*) or +2 (*ATA TAA*) codons (in italics) and translation termination in *Euplotes* spp. Conserved bases are in red. (*B*) Sequence LOGOs of frameshift sites and terminating stop codons contexts. (*C*) Properties of ribosomal frameshifting in *Euplotes* spp. in contrast to programmed ribosomal frameshifting. (*D*) Schematic representation of the algorithm for identification of frameshift sites. (*E*) Positional AT-content relative to start and stop codons (FS and terminators) centered at zero. (*F*) UpSetR representation of the distribution of identified frameshift sites across species. Rows correspond to the species and columns to the number of unique or shared frameshift sites. (*G*) Distributions of 3′ UTR lengths (nt) in transcripts with (red) and without (gray) frameshifts. (*H*) Stop-codon frequency around frameshifting (red) and terminating (gray) stop-codons.

whose regulation via polyamine-dependent +1 ribosomal frameshifting is nearly universal (36). +1 and –2 ribosomal frameshifting is also suspected to occur in the expression of two paralogous genes *ASXL1* and *ASXL2* (37). The three genes that use –1 frameshifting are of the viral ancestry, i.e. *PEG10*/*EDR* (38, 39), *PNMA3*, and *PNMA5* (40). While there were reports of –1 frameshifting in human genes of nonviral origin (*CCR5* (41) and *ATP7B* (42)), they appeared to be due to misinterpretations of artifacts of the reporters used (43–45). The best studied bacterium, *Escherichia coli*, has only three genes of nonviral origin whose expression is currently known to utilize frameshifting, *prfB* (46–48), *dnaX* (49–51), and a more recently discovered *copA* (42). Even if these genes are combined with sequences of prophages and Insertion Sequence elements, the programmed ribosomal frameshifting is estimated to occur on average in only four genes per bacterial genome (52). Ribosomal frameshifting is more common in viruses, for which it provides an attractive mechanism to produce more than one product from the same mRNA, to maintain a fixed stoichiometric ratio between its products or for regulatory purposes (53), and to yield a more compact organization of protein-coding information (7). However, even in viruses, frameshifting rarely occurs at more than a single location per genome.

The efficient frameshifting may also be observed in so-called pseudo-pseudogenes (54, 55), genes that are expressed despite having nonsense or frameshift mutations. The latter happens because the sequence downstream of a frameshift-causing indel mutation is translated in a new frame and has not been optimized by evolution for accurate triplet decoding, thus containing frameshifting-prone sequences (4, 5).

Abundant Ribosomal Frameshifting in *Euplotes* spp. Is Not Programmed Ribosomal Frameshifting

The frequency of ribosomal frameshifting in *Euplotes* spp. is in striking contrast with the frequency of programmed frameshifting. This was initially observed by Lawrence Klobutcher and Phil Farabaugh (15) who noted that, at the time when sequences of only 67 genes from *Euplotes* species were available, frameshifting was reported in four genes (56–59) suggesting that it occurred in more than 5% of *Euplotes* genes. Subsequent ribosome profiling and proteomics studies did confirm the high frequency of ribosomal frameshfting in *Euplotes* spp. by identifying thousands of instances of frameshifting and revealing that about one-fifth of all *Euplotes* genes use frameshifting in their expression, with some genes having up to eight frameshift sites (13, 16, 17).

In addition to its high frequency, frameshifting in *Euplotes* spp. is highly efficient and provides deterministic readout, again arguing that it is a feature of a *Euplotes* standard genetic code. Indeed, all known cases of programmed ribosomal frameshifting are not 100% efficient (6). Only a fraction of ribosomes shift the reading frame, while the remaining and usually a larger proportion, continue translation in the same reading frame or terminate when a stop codon is a part of the frameshift site. This optionality is at the core of the functional role of programmed frameshifting in gene expression, as it creates a bifurcation in the process of mRNA decoding yielding two different products of the same gene (34). Even when only one of these products is a functional product, the sensitivity of the frameshifting efficiency to the cellular environment provides an opportunity for regulation. The best-known eukaryotic example is the polyamine-sensitive +1 frameshifting required for synthesis of the protein antizyme. This is a key part of a negative feedback loop, since antizyme is a negative regulator of polyamine synthesis and transport (10). A similar negative control loop operates in the *prfB* gene encoding bacterial release factor 2 (RF2) (46). In decoding RF2 mRNA, frameshifting competes with termination at a UGA stop codon which is recognized exclusively by RF2. Thus, a drop in the RF2 levels leads to increased frameshifting efficiency which in turn is required for RF2 synthesis.

The situation with ribosomal frameshifting in *Euplotes* spp. is clearly different. A ribosome profiling study did not reveal any significant drop in ribosome densities downstream of frameshift sites (13), that would be expected if the frameshifting efficiency was substantially different from 100%. Thus, unlike programmed ribosomal frameshifting, the frameshifting in *Euplotes* spp. is deterministic, as it does not compete with triplet decoding. Hence, a failure of the ribosome to shift could be considered as an error in the translation process, exactly as spontaneous frameshifting during triplet mRNA decoding in other species.

Furthermore, programmed ribosomal frameshifting requires the presence of stimulatory signals increasing its efficiency, such as RNA secondary structures (9, 60, 61), complementary interactions between ribosomal RNA and mRNA (62, 63) or nascent peptides (12, 64), or protein factors interacting with mRNA (65–67). Sequences known to trigger highly efficient ribosomal frameshifting in the absence of additional stimulators are known only in yeast (68). This largely is enabled by a severe imbalance in the concentration of in-frame and out-of-frame tRNAs, as in the frameshift site in the TY1 element consisting of CUU_AGG_C. Only one gene copy of AGG-decoding tRNA exists in *Saccharomyces cerevisiae* while there are sixteen copies of genes for tRNA recognizing the +1 GGC codon (69). This imbalance favors decoding of the +1 frame codon (69). Sequence requirements for frameshifting in *Euplotes* spp. are even weaker. Seemingly, all that is needed for frameshifting is either of the two stop codons, UAA or UAG [with UGA reassigned to cysteine (70)]. In early reports, all identified *Euplotes* frameshifting sites had the AAA codon preceding the stop, which prompted Klobutcher and Farabaugh to suggest that Lys-tRNA decoding the AAA codon had some special properties enabling ribosomal frameshifting at this codon via repairing with overlapping +1 AAU codon (15). However, subsequent high-throughput studies have revealed that while the AAA codon is by far the most frequently used codon in the frameshifting sites, it is not absolutely required for frameshifting. However, the identity of the codon preceding stop codons determines the mechanism of frameshifting, with AUA_UAR resulting in a +2 frameshifting, presumably via repairing of tRNA decoding AUA with the identical overlapping codon in the +2 frame (13). No additional signals or stimulators were found to be required for frameshifting, again suggesting that frameshifting is a standard default meaning of stop codons in the *Euplotes* genetic code (13).

This raises an immediate question. If the standard meaning of stop codons in the *Euplotes* genetic code is frameshifting, how does it terminate protein synthesis? Here, the situation in *Euplotes* spp. seems to be similar to that occurring in species with recently discovered genetic codes where all stop codons are reassigned to code for amino acids (18–23). It has been shown that in ciliate *Condylostoma magnum* the same stop codons are used as terminators, but in a strictly position-specific manner (18, 20). Ribosomes terminate only in close proximity to polyA tails, with 3′ UTRs in *C. magnum* being very short. Changes in the sequence of a specific tRNA and release factor have been recently attributed to the reassigned stop codons in trypanosomatid *Blastocrithidia nonstop* which includes position-specific meaning of UAA codon (22). The situation is similar in *Euplotes* spp. with only a single exception found so far, where termination takes place far upstream of the polyA tail (13). This exception is a mRNA encoding a selenoprotein. As mentioned earlier, selenocysteine incorporation requires the presence of a SECIS structure and thus 3′ UTR is sufficiently long to accommodate this structure. Possibly, in this case the polyA proximity to the terminating stop is steric rather than purely sequence-length dependent and it is likely that the formation of SECIS structure brings the two into proximity. The requirement for polyA proximity is most likely due to involvement of polyA binding proteins (PABPs) in the process of termination. The stimulatory effect of PABPs on termination is well documented across a variety of biological systems (71–74), and ciliates likely represent an extreme case of such a requirement. Indeed the ~100% efficiency of frameshifting in *Euplotes* suggests that it does not compete with termination as happens in most other cases involving frameshifting at stop codons (75). This is also indirectly supported by experiments that tested the stop-codon specificity of *Euplotes* eRF1 in yeast (76). A hybrid release factor needed to be created where the *Euplotes* eRF3 recognition domain had been replaced with the yeast version. Presumably this alleviated the requirement for the polyA proximity.

More generally, ciliates are famous for extravagant ways to organize and express their genetic material, such as multiple nuclei and extensive structural genome rearrangements during transfer of genetic information between nuclei such as gene unscrambling (77–79), exceptionally extra short introns (80–82), as well as exceptionally frequent stop-codon reassignment (25, 27). Likely, the strict positional dependence of translation termination is one of the features enabling this reassignment, as stop codons in the middle of mRNAs would be expected to be highly inefficient and would need to be resolved either via codon reassignment or ribosomal frameshifting as in *Euplotes* spp. (83). While position-specific termination and frequent genetic code alterations in the forms of codon reassignment and standard frameshifting, seem to be connected, the direction of their causality remains to be elucidated.

Results

Detection of Frameshift Sites. We have sequenced transcriptomes of nine species from six major clades of the *Euplotes* phylogenetic tree (*SI Appendix*, Fig. S3): freshwater *Euplotes octocarinatus,* brackish waters *Euplotes harpa,* and marine *Euplotes focardii*, *Euplotes petzi*, *Euplotes euryhalinus*, *Euplotes rariseta*, *Euplotes minuta*, *Euplotes raikovi,* and *Euplotes crassus*. The transcriptomes were assembled and combined into 1,614 orthologous groups containing 4,903 sequences (*SI Appendix*, Table S1 and *Materials and Methods*). In all subsequent analyses, we considered only transcripts with identifiable orthologs, as we would not be able to assess the evolutionary trajectories of FSs from individual sequences. This also reduced contamination with sequences derived from other organisms found in the environment (*Materials and Methods*).

To detect FSs in these transcriptomes, we developed a systematic and unbiased procedure (*SI Appendix, Supplementary Methods*). We assumed absolute unambiguity of frameshifting in *Euplotes* spp., which means that ORFs following a frameshifting event were expected to be translated in a single reading frame. The procedure starts with the longest ORF that is then extended with adjacent or overlapping ORFs joined with either +1 or +2 frameshifting or stop codon readthrough (Fig. 1*D*). Candidate ORFs were detected using stringent criteria relying on minimal ORF length, high sequence similarity with their orthologs, signatures of purifying selection typical for the evolution of protein-coding sequences, and uniformity of the two latter characteristics along the candidate protein-coding sequence (*SI Appendix, Supplementary Methods*). To avoid arbitrariness in thresholds underlying these criteria, we considered these thresholds as parameters and fine-tuned them to obtain robust results. Subsequent validation using ribosome profiling data generated in *E. crassus* (13) demonstrated high specificity of the algorithm, with true- and false-positive discovery rates of 54% and 0%, respectively (*SI Appendix, Supplementary Methods*). Using this approach, we identified translated regions in the generated transcriptomes and found 3.9% of transcripts to contain FSs (8.3% orthogroups), i.e. 197 instances of +1 and 16 instances of +2 frameshift sites distributed across 192 sequences from the total 4,903 (Fig. 1 *B* and *F*). No stop codon readthrough events could be identified. No frameshifting events were identified in *E. petzi,* though manual analysis of alignments allows one to identify several frameshifts. This is likely due to stringent pipeline parameters combined with a small number of *E. petzi* transcripts with identifiable orthologs, as compared with other species. The set of predicted FSs and the predicted terminating stop codons (terminators) confirm previously reported features of *Euplotes* coding sequences, such as the prevalence of +1 shifts (predominantly at AAA preceding stop codons), short 3′ UTRs, enhanced AT-content in 3′ and 5′ UTRs and high frequency of stop codons in 3′ UTRs (Fig. 1 *E*, *G*, and *H*). The efficiency of termination is often influenced by the local context (84), however, we do not observe significant biases at FSs relative to terminating stops (Fig. 1*B*).

Evolution of Frameshifts in *Euplotes* spp. The unusually large number of tolerated, highly efficient FSs in *Euplotes* spp., a feature not observed in any other group of cellular organisms studied so far, calls for an evolutionary explanation. Although the current amount of data renders any population-based methods inefficient in this case, we can assess some basic evolutionary features of FSs, such as their general effects on fitness, from the FS gain and loss rates along the phylogeny.

We inferred frameshift site gain and loss events based on the reconstructed ancestry (Fig. 2*A*). We found that the frequency of gains exceeds that of losses about ten-fold (39 vs. 4). This sharp asymmetry may be explained by positive selection, however, to test for selection we have to consider probabilities of FS gains and losses rather than the numbers of respective events. These probabilities depend on the number of contexts suitable for FS gains and the number of existing FSs (that may be lost). The evolution of FSs is likely to be context-dependent since ~72% of all gain and loss events occurred due to the insertion or deletion of T at AAA_[T]AR (underscore separates codons, R = A or G). Thus, we initially focused on the analysis of evolution of FSs conforming to this specific pattern. We define the probabilities of frameshift gains and losses as $P_g = n_i/K$ and $P_l = n_d/F$, respectively, where n_i is the number of observed insertions of T yielding new FSs, K is the number of ancestral AAA_AR sequence motifs, n_d is the number of observed deletions of T leading to the loss of FSs, and F is the number of AAA_TAR FSs. As opposed to simple counts of FS gains and losses, FS loss probability exceeds FS gain probability by 3 to 30-fold (Fig. 2*D*, P = 0.028, permutation test). This discrepancy may be explained by selection against novel FSs combined with increased mutational pressure favoring FS gains. Concerning the latter, we observe the numbers of FS gain contexts to be on average 74-fold higher than the numbers of existing FSs. This inflates the probability of FS-gain mutations compared to FS losses (85). The selection against FSs may be estimated from differences in the observed probabilities of FS gain and loss (*Materials and Methods*), with the scaled selection coefficient calculated as the average $S = 4sN_e$ of FSs, where s is the selection coefficient and N_e, the effective population size (85).

The calculated S (Fig. 2*E*) differed for frameshifts in the AAA_[T]AR and non-AAA_[T]AR contexts, and while S for frameshifts in the latter context are in the highly deleterious range (upper bound S = −10±1, the CI derived from a permutation analysis), frameshifts in the AAA_[T]AR context seem to be only slightly deleterious (S = −2 ± 1), which is consistent with fitness-reducing

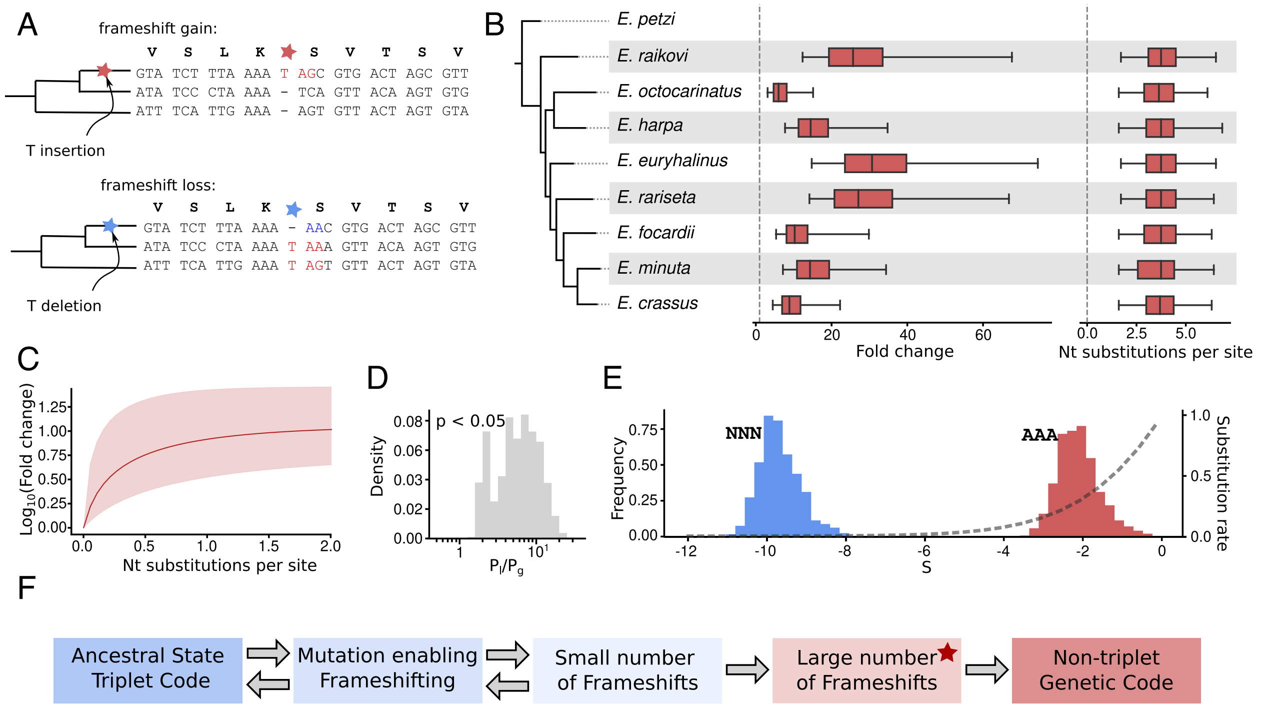

Fig. 2. Increasing numbers of frameshift sites during the evolution of *Euplotes* spp. (*A*) Inference of gains (*Top*) and losses (*Bottom*) of FSs based on the ancestral states. (*B*, *Left*) Phylogenetic tree of studied *Euplotes* species. (*Middle*) Distributions of the fold changes in FSs numbers upon reaching equilibrium. (*Right*) Distributions of time intervals (in nucleotide substitutions) required to reach 95% of the FSs number expected at equilibrium. (*C*) Projected fold change of the number of FSs over time. Red shading indicates the 95% CI obtained from permutations. (*D*) Ratio of FS gain and loss probabilities for AAA_[T]AR contexts. (*E*) Permutation distributions of *S* values for selection against FSs arising in AAA_[T]AR contexts (red) and of upper bounds for *S* in non-AAA_[T]AR contexts (blue). Dashed line is the theoretical dependence of the normalized substitution rate for variants affected by selection and genetic drift on the *S* value. (*F*) Proposed scheme of the emergence and subsequent entrenchment of the nontriplet genetic code. A star indicates the current stage of the genetic code of *Euplotes*.

FS effect in *Euplotes* spp. that may be due to ribosome pausing (13). Indeed, the observed $S = -2$ may result in both negative selection and drift influencing accumulation of mutations, whereas $S = -10$ effectively indicates only negative selection acting against FSs arising in the NNN_[T]AR contexts (Fig. 2*E*) (85).

Thus, the large number of contexts suitable for FS gains compared to the numbers of already existing FSs yields a relatively large mutational target size for FS accumulation, which counters selection against FSs in favorable contexts. But are these processes at equilibrium and, consequently, are the frequencies of euplotid FSs constant in time? And, if there is no equilibrium, how distant from it are *Euplotes* spp.? The following differential equation describes the change of the number of FSs over time where u_g and u_l are, respectively, the rates of gain and loss derived from their probabilities (*Materials and Methods*), and *K* and *F* are, as earlier, the (constant) number of ancestral AAA_AR sequence motifs and the (variable) number of current AAA_TAR FSs, respectively:

$$dF = (K - F)\, u_g dt - F u_l dt.$$

Solving this equation (*Materials and Methods*) we projected changes in the number of FSs over time. At infinite time, this function reaches the upper asymptote which corresponds to the number of FSs when gains and losses are at equilibrium. We find the asymptote to be (depending on species) 5 to 25-fold larger than the current numbers of FSs ($P < 0.0001$, permutation test, *Materials and Methods*) (Fig. 2 *B*, *Middle*).

Next, we estimated the time required to reach equilibrium. As formally this time is infinite, we consider the effective equilibrium as the time point when the number of FSs is 95% of its asymptotic value. We found these times to be consistent across species and constitute 1.7 to 6.3 nucleotide substitutions per site on average (Fig. 2*B*), which is about 0.68 to 2.52 of the estimated age of the considered group of *Euplotes* spp. (86). Thus, our results indicate that the euplotid FSs are at an early stage of the FS accumulation process.

The number of euplotid FSs per genome at the equilibrium is expected to be between approximately 17 and 71 thousand. Thus, if we consider independent effects of FSs on fitness, the net lag load (loss of relative fitness) *L* conveyed by the total body of FSs becomes $L = 1 - (1 - S/N_e)^F$, which, depending on the N_e value, may be either substantial and pose a potential hazard for the survival of *Euplotes*, or negligible. To assess this, we calculated the dependence between the net lag load of eventual frameshifts calculated from the obtained per-site $4sN_e$ values and N_e. We observe that the net lag load would drastically decrease fitness of *Euplotes* spp. on $N_e < 10^5$, whereas on $N_e > 10^6$ there is only a minor fitness decrease (*SI Appendix*, Fig. S6). Although the exact N_e values of *Euplotes* spp. cannot be calculated in the absence of micronuclei genomic data, we may assume the euplotid N_e values to be larger than 10^6, as smaller organisms typically have much larger N_e [for comparison, $N_e \sim 10^6$ for the current human population (87)]. Thus, although there will eventually be some significant load associated with frameshifts, it should not impact the survival of *Euplotes* spp.

Discussion

The findings presented here suggest that *Euplotes* spp. are only starting their evolution towards a balanced use of frameshifting as a part of their genetic code. Since the ability of efficient frameshifting at internal stop codons is shared among all *Euplotes* spp., it likely existed in their last common ancestor. However, as we show here, the process of FS accumulation is very slow, and the number of currently observed FSs constitutes only about 4 to 20% of its expected maximum at equilibrium.

Is frameshifting itself useful for *Euplotes* spp.? Does it increase their fitness and at least partially compensate for the detrimental effects of FSs? It has been suggested that alternative genetic codes are frequently used in ciliates to protect their macronuclear genomes from foreign genetic elements (88). In addition, efficient frameshifting at out-of-frame stops in *Euplotes* spp. makes their genes resistant to single nucleotide insertions in protein-coding regions. Indeed, we found several instances of insertions that disrupt the protein-coding reading frame, but then the reading frame is restored due to frameshifting at a premature stop codon downstream (*SI Appendix*, Fig. S7). Thus, single nucleotide insertions in *Euplotes* protein-coding sequences result in a change of only a short segment of the encoded protein (between insertion and newly formed premature stop codon) rather than the entire downstream part of the protein.

In addition to these and other possible benefits of frameshifting, our study suggests an alternative, more likely evolutionary explanation of the abundant frameshifting in *Euplotes*. Once a species develops an ability to frameshift ribosomes at in-frame stop codons with high efficiency, the number of in-frame stop codons would start growing even if frameshifting is mildly deleterious. Once a certain number of FSs is reached, the process is expected to become entrenched and hence irreversible, as a reversal of the change that has enabled frameshifting would result in mistranslation of many genes (Fig. 2*F*). Hence, the high number of FSs in *Euplotes* does not necessarily imply that they are beneficial, but simply that they are not a limiting factor in the evolution and are not, individually or collectively, under sufficiently strong enough selection to be eliminated. However, depending on the long-term dynamics of strength and efficiency of selection, the exact frequencies of FSs occurrence in protein-coding sequences may vary. Thus we conclude that changes in the genetic code, even as profound as the violation of its triplet character, may be the result of neutral evolution. To what extent the standard genetic code is a product of adaptation and to what extent it is a product of neutral evolution possibly remains a matter of debate (89–93).

Our finding has an unexpected implication to the evolvability of the genetic code. Francis Crick's Frozen Accident Hypothesis (94) was partly based on the necessity of the Discontinuity Principle. A change in the feature represented by a codon (amino acid or stop) would have sudden and severe consequences on the composition of the entire proteome to which a species would not be able to adapt in a short period of time (95). The later discoveries of many genetic code variants have revealed numerous possibilities for intermediate states such as ambiguous codons that prevent violations of the Discontinuity Principle (25, 27). Interestingly, the evolution towards frameshifting use as described here (Fig. 2*F*) does not violate the Discontinuity Principle since codons that specify (or alternatively "lead to") termination of protein synthesis codons do not occur in protein-coding regions and the change of the meaning of UAG, UAA, and UGA at internal positions should not alter the composition of the proteome. It is the reversal process that would violate the discontinuity principle and thus should not be possible. A counterintuitive corollary of this asymmetry is that the number of organisms with such nontriplet features might increase with time.

Materials and Methods

Species Selection and Cell Cultures. *Euplotes* strains used in this study were obtained from a large collection maintained in the laboratories of the Universities of Pisa and Camerino. Each species was selected based on the position it occupies within the *Euplotes* phylogenetic tree, which is commonly regarded as forming six major clades (numbered I to VI from the bottom) (*SI Appendix*, Fig. S3) (96, 97). *E. petzi* forms the most basal clade I. *E. raikovi* clusters with several other species into clade IV. *E. octocarinatus* and *E. harpa* lie in the well-supported clade V which includes most of the freshwater *Euplotes* species, but they belong to two different subclades. *E. euryhalinus*, *E. rariseta*, *E. minuta*, *E. crassus* and *E. focardii* cluster together into the poorly resolved and species-richest clade VI. However, these species branch into three different subclades, *E. euryhalinus* with *E. rariseta*; *E. focardii*; *E. minuta* with *E. crassus*. The latter two species are the most closely related among all the species analyzed here, as also demonstrated by previous breeding studies (98).

The selected *Euplotes* species have different ecologies. *E. octocarinatus* and *E. harpa* are temperate freshwater and brackish species, respectively (99, 100), while all the others are marine species. *E. focardii* is endemic to Antarctica (101). *E. petzi* and *E. euryhalinus* have a bipolar (Antarctic and Arctic) distribution (102, 103). *E. rariseta, E. crassus*, *E. minuta*, and *E. raikovi* are virtually ubiquitous in temperate coastal areas (104).

RNA Preparation and Sequencing. Cultures were grown under a daily cycle of 12 h of dark and 12 h of very weak light, at 4 to 6 °C (polar species) or 18 to 20 °C (nonpolar species) and fed on green algae *Dunaliella* (marine species) and *Chlorogonium* (freshwater species). They were expanded by daily food additions up to a cell density of about 10^4 cells/mL, then washed free of food and debris, and re-suspended for 3 (temperate species) or 6 (polar species) d in fresh marine or distilled water before being harvested. The TRIzol plus purification kit (Thermo Fisher Scientific) was used to purify total RNA, following the manufacturer's recommendations. Samples of about 10^7 cells were concentrated by mild centrifugation and lysed by rapid resuspension in 1 mL TRIzol reagent containing phenol and guanidine. After chloroform addition and centrifugation, an equal volume of 70% ethanol was added to the aqueous phase containing RNA, which was next purified using silica cartridges. As a rule, an on-column DNase treatment was carried out for 45 min at room temperature to obtain DNA-free RNA preparations. After washing, RNA was eluted with 30 μL RNase-free water and stored at −80 °C before use. RNA concentration and purity were estimated by NanoDrop One Spectrophotometer (Thermo Fisher Scientific), while RNA integrity was analyzed by agarose gel electrophoresis.

cDNA library preparation and sequencing were carried out by BGI using Illumina TruSeq library construction and sequencing at Illumina HiSeq 2000 using 101PE.

Assembly of Transcriptomes and Construction of Orthologous Gene Groups. The resulting read libraries were trimmed with Trimmomatic (105) with parameters: ILLUMINACLIP:TruSeq3-SE:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36. The transcriptomes were assembled with Trinity (106) using the default set of parameters. For the expression analyses, the expression rate of each transcript was calculated as Transcripts Per Million (TPM) using the Kallisto software with the default parameters (107).

Orthologous gene groups (OGGs) were constructed using ProteinOrtho v. 5.15 (108) with parameters: −p = blastn −e = 1e−25 −identity = 70 on obtained transcriptomes.

The constructed OGGs were aligned with MUSCLE (109) with the default parameters. The coding regions were aligned by TranslatorX (110) with parameters: −p M −t F −w 1 −c 10. The predicted proteins were aligned with the Smith–Waterman algorithm (111).

To determine whether a sequence of a transcript is in the sense or antisense orientation, we relied on the locations of polyA and polyT tails, polyA tails at the 3′ ends were used as an indicator of the sense strand, while transcripts with polyT were classified as antisense and reverse complements of these transcripts were used instead. In the absence of polyA/T tails (truncated transcripts), we selected the orientation yielding the lowest number of indels in the longest ORF aligned to one of its orthologs.

Calculation of Identity and d_N/d_S Values. To calculate pairwise protein identity, ORFs were translated using the euplotid genetic code (#10 from ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi) and aligned with the Smith-Waterman algorithm (111). All gap-containing positions were removed. The identity was calculated as the number of identical amino acids divided by the number of nongap positions in the alignment.

We employed two ways to calculate the d_N/d_S ratio. The Nei-Gojobori method (112) yields values which can be straightforwardly compared by simple statistics (*SI Appendix, Supplementary Methods*). For the correction and validation of the obtained d_S values (Fig. 2C), we additionally calculated this statistic with the Nielsen–Yang method (113) implemented in the PAML software (114).

For the analysis of d_N/d_S uniformity (*SI Appendix, Supplementary Methods*), random sampling was performed from compared ORFs (115). At each permutation round all four values needed to calculate the d_N/d_S ratio, that is *dN*, *N*, *dS*, *S* [in the Nei–Gojobori notation (112)], were sampled from the respective Poisson distributions with parameters derived from the data. The d_N/d_S ratios obtained for two ORFs were then compared. The percent of permutations which resulted in d_N/d_S of adjacent (shorter) ORF being higher than d_N/d_S of the primary (longer) ORF was used as the metric of uniformity.

Transcriptome Quality Assessment. Ciliates are obligate heterotrophs (116) mostly feeding on algae and other microorganisms, the remnants of which remain in their cytoplasm. They also may have intracellular symbionts (117). Hence the experimental separation of the ciliate DNA from the DNA of symbionts and prey is currently not feasible (118–120). This necessitated removal of contaminant sequences from assembled transcriptomes. We employed a four-step filtering procedure with subsequent controls to ensure that the transcripts considered in downstream analyses were indeed ciliate transcripts:

1. AT-content was calculated for each transcript and the resulting distribution was assessed (*SI Appendix*, Fig. S4). Since some distributions were bimodal, we filtered out all transcripts with the AT content less than the distribution mean minus 4 SD. This constraint follows observations that ciliate transcriptomes are AT-rich (121).
2. All transcripts having no homologs in other transcriptomes obtained in this study (singletons) were filtered out (see "*Assembly of transcriptomes and construction of orthologous gene groups*" above).
3. For each sequence, the nucleotide identity with its closest homolog from the sample was calculated. The resulting distribution appeared sharply trimodal (*SI Appendix*, Fig. S5) with ciliate sequences found only in the middle peak, as verified with BLASTn search (see below). From this distribution, we obtained the boundaries of the middle peak corresponding to the between-ciliate identity range. Then we discarded all transcripts with identity below the obtained lower bound of 0.65 and above the obtained upper bound of 0.95, which corresponded to the boundaries of the middle peak.
4. For the remaining sequences, the taxonomy of candidate contaminating species was identified. For this, we randomly selected 1,000 transcripts from each transcriptome and performed online BLASTn (122) search against the Genbank nonredundant database using the NCBI implementation at https://blast.ncbi.nlm.nih.gov/Blast.cgi. A nonciliate species was considered as being closely related to a contaminating species if it produced the best hit with at least 50% identity and alignment length of at least 50% relative to the query at E value below 10^{-5}. A total of 41 contaminating (or closely related) organisms were identified (Dataset S5). All transcripts in our transcriptome were tested for sequence similarity to the genomic and transcriptomic sequences of these species using local BLASTn. All transcripts with hits exceeding 70% identity at E-value below 10^{-25} were removed.

To assess the reliability of this procedure, we searched the discarded sequences against known ciliate genomes and did not obtain significant alignments for any sequence from 111 discarded OGGs. We further queried sequences from the remaining 1,614 OGGs against the nonredundant Genbank database and did not obtain statistically significant alignments.

Finally, we checked for the presence of chimeric transcripts, i.e., transcripts assembled from reads originating from different organisms. We considered a transcript to be chimeric, if it yielded at least two BLAST hits (E-value < 10^{-10}, Identity > 70%) that do not originate from the same organism with the overlap of at most 15 nt. No chimeric transcripts satisfying this criterion were identified.

Our contamination removal procedure and the restriction of the analyses to only genes with established orthologs produces a set of generally conserved genes with higher expression levels (*SI Appendix*, Fig. S9). Along with that, FSs tend to be more frequent in genes with lower expression levels (13), an effect that is also to some degree observed in our filtered transcriptomes (*SI Appendix*, Fig. S10). However, these effects should not result in substantial biases in our analyses, firstly, due to the weak FS tendency towards genes with low expression levels (*SI Appendix*, Fig. S10) and secondly due to FSs expected to have larger fitness effects in genes with higher levels of expression.

Phylogenetic Analysis. To validate the phylogenetic tree obtained using 18S rRNA (99) (*SI Appendix*, Fig. S3), we also constructed a tree using concatenated coding sequences obtained in this study. The tree was built using paralog-free OGGs containing genes from all nine studied euplotid species with the sequences from *Tetrahymena thermophila* as an outgroup. The coding regions were aligned with MUSCLE (109) and then the resulting alignments were concatenated. The tree was constructed with the Maximum Likelihood algorithm implemented in the PhyML package (123) using the automatic model selection feature (124) and 100 bootstrap replicates. To avoid biases in the estimation of neutral mutation rates arising from conserved regions in 18S rRNA, we estimated neutral divergence times from synonymous substitutions in protein sequences. Divergence times (per-branch *dS* values) were estimated with the m1 model of the PAML package. The ancestral sequences corresponding to frameshift sites were inferred using maximum parsimony (MP) (125, 126).

Gains and Losses of Frameshift Sites. To avoid misalignment errors we considered only frameshift sites within ±10 indel-free alignment blocks. The numbers of frameshift site gains and losses were calculated as the numbers of respective mutations, which all appeared as insertions and deletions of thymines in stop codons (Fig. 2A). Hence, the probability of frameshift site gain/loss at a specific context *i* was calculated as the number of gain/loss events *n* normalized over the number of the contexts *K*:

$$P_i = \frac{n_i}{K_i}$$

The contexts are defined by the codon preceding gained or lost stop codon, e.g., the context for frameshift sites is defined as NNN within NNN_TAR and in NNN_AR (potential gain upon insertion of T), R is purine (A or G), and the underscore separates codons in the coding reading phase.

Calculating the Inflation of Frameshift Sites. The temporal dynamics of the per-genome numbers of FSs is given by a logistic differential equation:

$$dF = (K - F)\ u_{\text{gain}}\ dt\ -\ Fu_{\text{loss}}\ dt\ =\ Ku_{\text{gain}}\ dt\ -\ F\left(u_{\text{gain}} +\ u_{\text{loss}}\right)\ dt$$

where $F = F(t)$ is the (variable) number of FSs, u_{gain} and u_{loss} are the rates of site gain and loss, respectively, and K is the (constant) number of suitable ancestral contexts for the shift gain. The solution is:

$$F(t) = \left[F(0) - \frac{Ku_{gain}}{(u_{gain} + u_{loss})}\right] \times exp[-(u_{gain} + u_{loss})t] + \frac{Ku_{gain}}{(u_{gain} + u_{loss})}$$

To simplify, let $A = Ku_{\text{gain}}/(u_{\text{gain}} + u_{\text{loss}})$, $b = Ku_{\text{gain}}$ and F(0) be the current number of frameshift sites at $t = 0$, then:

$$F(t) = [F(0) - A] \times exp[-tb/A] + A \rightarrow A\text{, as } t \rightarrow \infty$$

Selection Inference. We calculated the balance between selection and drift in the form of Kimura's scaled selection coefficient S defined as $S = 4sN_e$, where s is the selection coefficient and N_e is the effective population size.

Next, we derive the expression to infer S from the ratio of FS gain and loss probabilities. The observed mutation rate under mutation, selection, and drift is (85):

$$u = 2N_e\mu[1 - exp(-2s)]/[1 - exp(-4N_e s)]$$

where μ is the mutation rate.

The selection coefficient s can be presumed small, as events of frameshift site gain are observed. Thus, for $s \to 0$, $1 - \exp(-2s) = 2s$, and

$$u = 4N_e\mu s/[1 - \exp(-4N_e s)] = \mu S/[1 - \exp(-S)]$$

As we are dealing with rare evolutionary events, i.e., indels and relatively small evolutionary times, the mutation rate may be presumed to equal the mutation probability: $u = P$.

Thus, the probability of FS gain is:

$$P_{\text{gain}} = P(S) = \frac{S\mu_{\text{gain}}}{[1 - \exp(-S)]}$$

And, as the fitness effect of a frameshift site loss is equal to the negative fitness effect of a frameshift site gain, the probability of a frameshift site loss is:

$$P_{\text{loss}} = P(-S) = (-S)\mu_{\text{loss}}/[1-exp(S)]$$

The analysis of thymine indels in our data (*SI Appendix*, Fig. S8) suggests that $\mu_{\text{gain}} = \mu_{\text{loss}}$. Thus

$$P_{\text{gain}}/\ P_{\text{loss}} = -[1-exp(S)]/[1-exp(-S)]$$

Or, alternatively:

$$S = \ln\left(P_{\text{gain}}/\ P_{\text{loss}}\right)$$

To obtain the lag load of the total body of frameshift sites, we first calculated the expected total number of FSs at equilibrium by normalizing the numbers of genes in our samples containing FSs to the total number of genes estimated for *E. focardii* and *E. octocarinatus* (121) (*SI Appendix*, Fig. S6). Next, by presuming independent fitness effects of each frameshift, we calculated the net lag load of frameshift sites as the product of fitness effects of all sites.

Equilibrium Frameshift Site Numbers. The expected per-coding genome counts of FSs at mutational equilibrium were estimated from the calculated equilibrium numbers of FSs in our purified transcript sets and the total numbers of genes or the total coding bp numbers of *E. focardii* and *E. octocarinatus* estimated in ref. 121. The per-genome counts were obtained either as products of the equilibrium FS frequencies and coding genome lengths or as products of frequencies of genes containing FS at equilibrium and the total numbers of genes. The per-genome equilibrium counts were consistent between the approaches and between the two considered species (*SI Appendix*, Table S2).

Data, Materials, and Software Availability. The RNAseq and RiboSeq data generated in the present study and the RiboSeq data generated in the previous study (13) have been submitted to SRA as bioproject PRJNA896607 (URL https://www.ncbi.nlm.nih.gov/bioproject/PRJNA896607). All scripts and data analysis protocols along with the constructed transcriptomes and tabulated frameshift sites are available online at https://github.com/sofyagdk/euplotes. All study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. P.V.B. wishes to acknowledge financial support from SFI-HRB-Wellcome Trust Biomedical Research Partnership (Investigator in Science award 210692/Z/18/). M.A.M. was supported by the Russian Foundation for Basic Research (RFBR), grant 20-34-90033. S.A.G. and M.S.G. were supported by the RFBR grant 20-54-14005. J.F.A. was personally supported by the Irish Research Council Advanced Laureate grant IRCLA/2019/74. A.V. thanks G. Di Giuseppe and P. Luporini for providing the *Euplotes* species used in this study.

Author affiliations: [a]Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow 199911, Russia; [b]A. A. Kharkevich Institute for Information Transmission Problems RAS, Moscow 127051, Russia; [c]Laboratory of Eukaryotic Microbiology and Animal Biology, School of Biosciences and Veterinary Medicine, University of Camerino, Camerino 62032, Italy; [d]School of Biochemistry and Cell Biology, University College Cork, Cork T12 XF62, Ireland; and [e]Department of Human Genetics, University of Utah, Salt Lake City, UT 84112

Author contributions: M.A.M., J.F.A., M.S.G., and P.V.B. designed research; S.A.G., M.A.M., S.M.H., M.S.G., and P.V.B. performed research; S.A.G. and A.V. contributed new reagents/analytic tools; S.A.G., M.A.M., M.S.G., and P.V.B. analyzed data; and S.A.G., M.A.M., M.S.G., and P.V.B. wrote the paper.

1. F. H. C. Crick, L. Barnett, S. Brenner, R. J. Watts-Tobin, General nature of the genetic code for proteins. *Nature* **192**, 1227–1232 (1961).
2. P. J. Farabaugh, G. R. Björk, How translational accuracy influences reading frame maintenance. *EMBO J.* **18**, 1427–1434 (1999).
3. P. J. Farabaugh, "Chapter 3: Errors during elongation can cause translational frameshifting" in *Programmed Alternative Reading of the Genetic Code*, (Springer, 1997), pp. 29–32.
4. O. L. Gurvich *et al.*, Sequences that direct significant levels of frameshifting are frequent in coding regions of *Escherichia coli*. *EMBO J.* **22**, 5941–5950 (2003).
5. A. A. Shah *et al.*, Computational identification of putative programmed translational frameshift sites. *Bioinformatics* **18**, 1046–1053 (2002).
6. J. F. Atkins, G. Loughran, P. R. Bhatt, A. E. Firth, P. V. Baranov, Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. *Nucleic Acids Res.* **44**, 7007–7078 (2016).
7. A. E. Firth, I. Brierley, Non-canonical translation in RNA viruses. *J. Gen. Virol.* **93**, 1385–1409 (2012).
8. X. Gao, E. R. Havecker, P. V. Baranov, J. F. Atkins, D. F. Voytas, Translational recoding signals between gag and pol in diverse LTR retrotransposons. *RNA* **9**, 1422–1430 (2003).
9. I. Brierley, P. Digard, S. C. Inglis, Characterization of an efficient coronavirus ribosomal frameshifting signal: Requirement for an RNA pseudoknot. *Cell* **57**, 537–547 (1989).
10. S. Matsufuji *et al.*, Autoregulatory frameshifting in decoding mammalian ornithine decarboxylase antizyme. *Cell* **80**, 51–60 (1995).
11. P. Gupta, K. Kannan, A. S. Mankin, N. Vázquez-Laslop, Regulation of gene expression by macrolide-induced ribosomal frameshifting. *Mol. Cell* **52**, 629–642 (2013).
12. M. M. Yordanova, C. Wu, D. E. Andreev, M. S. Sachs, J. F. Atkins, A nascent peptide signal responsive to endogenous levels of polyamines acts to stimulate regulatory frameshifting on antizyme mRNA. *J. Biol. Chem.* **290**, 17863–17878 (2015).
13. A. V. Lobanov *et al.*, Position dependent termination and widespread obligatory frameshifting in *Euplotes* translation. *Nat. Struct. Mol. Biol.* **24**, 61–68 (2017).
14. L. A. Klobutcher, Sequencing of random *Euplotes crassus* macronuclear genes supports a high frequency of +1 translational frameshifting. *Eukaryot Cell* **4**, 2098–2105 (2005).
15. L. A. Klobutcher, P. J. Farabaugh, Shifty ciliates: Frequent programmed translational frameshifting in euplotids. *Cell* **111**, 763–766 (2002).
16. R. Wang, J. Xiong, W. Wang, W. Miao, A. Liang, High frequency of +1 programmed ribosomal frameshifting in *Euplotes octocarinatus*. *Sci. Rep.* **6**, 21139 (2016).
17. R. Wang, Z. Zhang, J. Du, Y. Fu, A. Liang, Large-scale mass spectrometry-based analysis of *Euplotes octocarinatus* supports the high frequency of +1 programmed ribosomal frameshift. *Sci. Rep.* **6**, 33020 (2016).
18. E. C. Swart, V. Serra, G. Petroni, M. Nowacki, Genetic codes with no dedicated stop codon: Context-dependent translation termination. *Cell* **166**, 691–702 (2016).
19. K. Záhonová, A. Y. Kostygov, T. Ševčíková, V. Yurchenko, M. Eliáš, An unprecedented non-canonical nuclear genetic code with all three termination codons reassigned as sense codons. *Curr. Biol.* **26**, 2364–2369 (2016).
20. S. M. Heaphy, M. Mariotti, V. N. Gladyshev, J. F. Atkins, P. V. Baranov, 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 (2016).
21. T. R. Bachvaroff, A precedented nuclear genetic code with all three termination codons reassigned as sense codons in the syndinean *Amoebophrya sp.* ex *Karlodinium veneficum*. *PLoS One* **14**, e0212912 (2019).
22. A. Kachale *et al.*, Short tRNA anticodon stem and mutant eRF1 allow stop codon reassignment. *Nature* **613**, 751–758 (2023).
23. B. K. B. Seah, A. Singh, E. C. Swart, Karyorelict ciliates use an ambiguous genetic code with context-dependent stop/sense codons. *Peer Community J.* **2**, e42 (2022).
24. B. Zinshteyn, R. Green, When stop makes sense. *Science* **354**, 1106 (2016).
25. P. V. Baranov, J. F. Atkins, M. M. Yordanova, Augmented genetic decoding: Global, local and temporal alterations of decoding processes and codon meaning. *Nat. Rev. Genet.* **16**, 517–529 (2015).
26. B. G. Barrell, A. T. Bankier, J. Drouin, A different genetic code in human mitochondria. *Nature* **282**, 189–194 (1979).
27. R. D. Knight, S. J. Freeland, L. F. Landweber, Rewiring the keyboard: Evolvability of the genetic code. *Nat. Rev. Genet.* **2**, 49–58 (2001).
28. Y. Shulgina, S. R. Eddy, A computational screen for alternative genetic codes in over 250,000 genomes. *ELife* **10**, e71402 (2021).
29. M. J. Berry, L. Banu, J. W. Harney, P. R. Larsen, Functional characterization of the eukaryotic SECIS elements which direct selenocysteine insertion at UGA codons. *EMBO J.* **12**, 3315–3322 (1993).
30. G. V. Kryukov *et al.*, Characterization of mammalian selenoproteomes. *Science* **300**, 1439–1443 (2003).
31. J. Baclaocos *et al.*, Processive recoding and metazoan evolution of Selenoprotein P: Up to 132 UGAs in molluscs. *J. Mol. Biol.* **431**, 4381–4407 (2019).
32. J. F. Atkins, P. V. Baranov, The distinction between recoding and codon reassignment. *Genetics* **185**, 1535–1536 (2010).
33. R. F. Gesteland, J. F. Atkins, RECODING: Dynamic reprogramming of translation. *Annu. Rev. Biochem.* **65**, 741–768 (1996).
34. P. V. Baranov, R. F. Gesteland, J. F. Atkins, Recoding: Translational bifurcations in gene expression. *Gene* **286**, 187–201 (2002).
35. O. Namy, J.-P. Rousset, S. Napthine, I. Brierley, Reprogrammed genetic decoding in cellular gene expression. *Mol. Cell* **13**, 157–168 (2004).
36. I. P. Ivanov, J. F. Atkins, 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).
37. A. M. Dinan, J. F. Atkins, A. E. Firth, ASXL gain-of-function truncation mutants: Defective and dysregulated forms of a natural ribosomal frameshifting product? *Biol. Direct* **12**, 24 (2017).
38. K. Shigemoto *et al.*, Identification and characterisation of a developmentally regulated mammalian gene that utilizes -1 programmed ribosomal frameshifting. *Nucleic Acids Res.* **29**, 4079–4088 (2001).
39. E. Manktelow, K. Shigemoto, I. Brierley, Characterization of the frameshift signal of Edr, a mammalian example of programmed −1 ribosomal frameshifting. *Nucleic Acids Res.* **33**, 1553–1563 (2005).
40. N. M. Wills, B. Moore, A. Hammer, R. F. Gesteland, J. F. Atkins, A functional -1 ribosomal frameshift signal in the human paraneoplastic Ma3 gene. *J. Biol. Chem.* **281**, 7082–7088 (2006).
41. A. T. Belew *et al.*, Ribosomal frameshifting in the CCR5 mRNA is regulated by miRNAs and the NMD pathway. *Nature* **512**, 265–269 (2014).
42. S. Meydan *et al.*, Programmed ribosomal frameshifting generates a copper transporter and a copper chaperone from the same gene. *Mol. Cell* **65**, 207–219 (2017).
43. Y. A. Khan *et al.*, Evaluating ribosomal frameshifting in CCR5 mRNA decoding. *Nature* **604**, E16–E23 (2022).
44. G. Loughran, A. D. Fedorova, Y. A. Khan, J. F. Atkins, P. V. Baranov, Lack of evidence for ribosomal frameshifting in ATP7B mRNA decoding. *Mol. Cell* **82**, 3745–3749.e2 (2022).
45. S. Meydan *et al.*, Response to: Lack of evidence for ribosomal frameshifting in ATP7B mRNA decoding. *Mol. Cell* **82**, 3523 (2022).

46. W. J. Craigen, C. T. Caskey, Expression of peptide chain release factor 2 requires high-efficiency frameshift. *Nature* **322**, 273–275 (1986).
47. M. Bekaert, J. F. Atkins, P. V. Baranov, ARFA: A program for annotating bacterial release factor genes, including prediction of programmed ribosomal frameshifting. *Bioinformatics* **22**, 2463–2465 (2006).
48. P. V. Baranov, R. F. Gesteland, J. F. Atkins, Release factor 2 frameshifting sites in different bacteria. *EMBO Rep.* **3**, 373–377 (2002).
49. A. L. Blinkowa, J. R. Walker, 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 (1990).
50. A. M. Flower, C. S. McHenry, 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 (1990).
51. Z. Tsuchihashi, A. Kornberg, Translational frameshifting generates the gamma subunit of DNA polymerase III holoenzyme. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 2516–2520 (1990).
52. V. Sharma *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).
53. M. M. Yordanova, P. V. Baranov, A frameshift in time. *ELife* **11**, e78373 (2022).
54. Y. Feng *et al.*, "Pseudo-pseudogenes" in bacterial genomes: Proteogenomics reveals a wide but low protein expression of pseudogenes in *Salmonella enterica*. *Nucleic Acids Res.* **50**, 5158–5170 (2022).
55. L. L. Prieto-Godino *et al.*, Olfactory receptor pseudo-pseudogenes. *Nature* **539**, 93–97 (2016).
56. S. Aigner *et al.*, *Euplotes* telomerase contains an La motif protein produced by apparent translational frameshifting. *EMBO J.* **19**, 6230–6239 (2000).
57. M. Tan, K. Heckmann, C. Brünen-Nieweler, Analysis of micronuclear, macronuclear and cDNA sequences encoding the regulatory subunit of cAMP-dependent protein kinase of *Euplotes octocarinatus*: Evidence for a ribosomal frameshift. *J. Eukaryot. Microbiol.* **48**, 80–87 (2001).
58. M. Tan, A. Liang, C. Brünen-Nieweler, K. Heckmann, Programmed translational frameshifting is likely required for expressions of genes encoding putative nuclear protein kinases of the ciliate *Euplotes octocarinatus*. *J. Eukaryot. Microbiol.* **48**, 575–582 (2001).
59. L. Wang, S. R. Dean, D. E. Shippen, Oligomerization of the telomerase reverse transcriptase from *Euplotes crassus*. *Nucleic Acids Res.* **30**, 4032–4039 (2002).
60. D. P. Giedroc, P. V. Cornish, Frameshifting RNA pseudoknots: Structure and mechanism. *Virus Res.* **139**, 193–208 (2009).
61. B.Y.-W. Chung, A. E. Firth, J. F. Atkins, Frameshifting in alphaviruses: A diversity of 3′ stimulatory structures. *J. Mol. Biol.* **397**, 448–456 (2010).
62. R. B. Weiss, D. M. Dunn, A. E. Dahlberg, J. F. Atkins, R. F. Gesteland, Reading frame switch caused by base-pair formation between the 3′ end of 16S rRNA and the mRNA during elongation of protein synthesis in *Escherichia coli*. *EMBO J.* **7**, 1503–1507 (1988).
63. J. F. Atkins *et al.*, Overriding standard decoding: Implications of recoding for ribosome function and enrichment of gene expression. *Cold Spring Harb. Symp. Quant. Biol.* **66**, 217–232 (2001).
64. O. L. Gurvich, S. J. Näsvall, P. V. Baranov, G. R. Björk, J. F. Atkins, Two groups of phenylalanine biosynthetic operon leader peptides genes: A high level of apparently incidental frameshifting in decoding *Escherichia coli* pheL. *Nucleic Acids Res.* **39**, 3079–3092 (2011).
65. S. Napthine *et al.*, A novel role for poly(C) binding proteins in programmed ribosomal frameshifting. *Nucleic Acids Res.* **44**, 5491–5503 (2016).
66. S. Napthine, S. Bell, C. H. Hill, I. Brierley, A. E. Firth, Characterization of the stimulators of protein-directed ribosomal frameshifting in Theiler's murine encephalomyelitis virus. *Nucleic Acids Res.* **47**, 8207–8223 (2019).
67. S. Napthine *et al.*, Protein-directed ribosomal frameshifting temporally regulates gene expression. *Nat. Commun.* **8**, 15582 (2017).
68. M. F. Belcourt, P. J. Farabaugh, Ribosomal frameshifting in the yeast retrotransposon Ty: tRNAs induce slippage on a 7 nucleotide minimal site. *Cell* **62**, 339–352 (1990).
69. P. V. Baranov, R. F. Gesteland, J. F. Atkins, P-site tRNA is a crucial initiator of ribosomal frameshifting. *RNA* **10**, 221–230 (2004).
70. D. C. Hoffman, R. C. Anderson, M. L. DuBois, D. M. Prescott, Macronuclear gene-sized molecules of hypotrichs. *Nucleic Acids Res.* **23**, 1279–1283 (1995).
71. D. A. Mangus, M. C. Evans, A. Jacobson, Poly(A)-binding proteins: Multifunctional scaffolds for the post-transcriptional control of gene expression. *Genome Biol.* **4**, 223 (2003).
72. A. Ivanov *et al.*, PABP enhances release factor recruitment and stop codon recognition during translation termination. *Nucleic Acids Res.* **44**, 7766–7776 (2016).
73. A. Ivanov *et al.*, Polyadenylate-binding protein-interacting proteins PAIP1 and PAIP2 affect translation termination. *J. Biol. Chem.* **294**, 8630–8639 (2019).
74. K. Mangkalaphiban *et al.*, Transcriptome-wide investigation of stop codon readthrough in *Saccharomyces cerevisiae*. *PLoS Genet.* **17**, e1009538 (2021).
75. F. M. Adamski, B. C. Donly, W. P. Tate, Competition between frameshifting, termination and suppression at the frameshift site in the *Escherichia coli* release factor-2 mRNA. *Nucleic Acids Res.* **21**, 5074–5078 (1993).
76. J. Salas-Marco *et al.*, Distinct paths to stop codon reassignment by the variant-code organisms *Tetrahymena* and *Euplotes*. *Mol. Cell. Biol.* **26**, 438–447 (2006).
77. E. C. Swart *et al.*, The *Oxytricha trifallax* macronuclear genome: A complex eukaryotic genome with 16,000 tiny chromosomes. *PLoS Biol.* **11**, e1001473 (2013).
78. L. C. Wong, L. F. Landweber, Evolution of programmed DNA rearrangements in a scrambled gene. *Mol. Biol. Evol.* **23**, 756–763 (2006).
79. D. H. Ardell, C. A. Lozupone, L. F. Landweber, Polymorphism, recombination and alternative unscrambling in the DNA polymerase alpha gene of the ciliate *Stylonychia lemnae* (Alveolata; class Spirotrichea). *Genetics* **165**, 1761–1777 (2003).
80. V. S. Bondarenko, M. S. Gelfand, Evolution of the exon-intron structure in ciliate genomes. *PLoS One* **11**, e0161476 (2016).
81. O. Jaillon *et al.*, Translational control of intron splicing in eukaryotes. *Nature* **451**, 359–362 (2008).
82. M. M. Slabodnick *et al.*, The macronuclear genome of *Stentor coeruleus* reveals tiny introns in a giant cell. *Curr. Biol.* **27**, 569–575 (2017).
83. E. Alkalaeva, T. Mikhailova, Reassigning stop codons via translation termination: How a few eukaryotes broke the dogma. *Bioessays* **39**, 1600213 (2017).
84. W. P. Tate *et al.*, The translational stop signal: Codon with a context, or extended factor recognition element? *Biochimie* **78**, 945–952 (1996).
85. M. Kimura, "Chapter 3: The neutral mutation-random drift hypothesis as an evolutionary paradigm, Neutral and nearly neutral mutations" in *The Neutral Theory of Molecular Evolution*, (Cambridge University Press, 1983), pp. 43–46.
86. W. Chen *et al.*, The hidden genomic diversity of ciliated protists revealed by single-cell genome sequencing. *BMC Biol.* **19**, 264 (2021).
87. S. R. Browning, B. L. Browning, Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am. J. Hum. Genet.* **97**, 404–418 (2015).
88. D. M. Prescott, The DNA of ciliated protozoa. *Microbiol. Rev.* **58**, 233–267 (1994).
89. S. J. Freeland, L. D. Hurst, The genetic code is one in a million. *J. Mol. Evol.* **47**, 238–248 (1998).
90. S. E. Massey, A neutral origin for error minimization in the genetic code. *J. Mol. Evol.* **67**, 510–516 (2008).
91. A. S. Novozhilov, Y. I. Wolf, E. V. Koonin, Evolution of the genetic code: Partial optimization of a random code for robustness to translation error in a rugged fitness landscape. *Biol. Direct* **2**, 24 (2007).
92. E. Janzen *et al.*, Emergent properties as by-products of prebiotic evolution of aminoacylation ribozymes. *Nat. Commun.* **13**, 3631 (2022).
93. M. Di Giulio, The error minimization of the genetic code would have been determined by natural selection and not by a neutral evolution. *Biosystems* **224**, 104838 (2023).
94. F. H. C. Crick, The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379 (1968).
95. E. V. Koonin, Frozen accident pushing 50: Stereochemistry, expansion, and chance in the evolution of the genetic code. *Life* **7**, 22 (2017).
96. Y. Zhao, Z. Yi, A. Warren, W. Song, Species delimitation for the molecular taxonomy and ecology of the widely distributed microbial eukaryote genus *Euplotes* (Alveolata, Ciliophora). *Proc. Biol. Sci.* **285**, 20172159 (2018).
97. A. Valbonesi, G. Di Giuseppe, A. Vallesi, P. Luporini, Two new species of *Euplotes* with cirrotype-9, *Euplotes foissneri* sp. nov. and *Euplotes warreni* sp. nov. (Ciliophora, Spirotrichea, Euplotida), from the coasts of Patagonia: Implications from their distant, early and late branching in the *Euplotes* phylogenetic tree. *Int. J. Syst. Evol. Microbiol.* **71**, 004568 (2021).
98. A. Valbonesi, C. Ortenzi, P. Luporini, An integrated study of the species problem in the *Euplotes crassus-minuta-vannus* Group1. *J. Protozool.* **35**, 38–45 (1988).
99. D. Méndez-Sánchez, R. Mayén-Estrada, X. Hu, *Euplotes octocarinatus* Carter, 1972 (Ciliophora, Spirotrichea, Euplotidae): Considerations on its morphology, phylogeny, and biogeography. *Eur. J. Protistol.* **74**, 125667 (2020).
100. C. Vannini, G. Petroni, F. Verni, G. Rosati, Polynucleobacter bacteria in the brackish-water species *Euplotes harpa* (Ciliata Hypotrichia). *J. Eukaryot. Microbiol.* **52**, 116–122 (2005).
101. A. Valbonesi, P. Luporini, Biology of *Euplotes focardii*, an Antarctic ciliate. *Polar Biol.* **13**, 489–493 (1993).
102. G. Di Giuseppe *et al.*, Improved description of the bipolar ciliate, *Euplotes petzi*, and definition of its basal position in the *Euplotes* phylogenetic tree. *Eur. J. Protistol.* **50**, 402–411 (2014).
103. G. Giuseppe, F. Dini, A. Vallesi, P. Luporini, Genetic relationships in bipolar species of the protist ciliate, *Euplotes*. *Hydrobiologia* **761**, 71–83 (2015).
104. M. J. Syberg-Olsen *et al.*, Biogeography and character evolution of the ciliate genus *Euplotes* (Spirotrichea, Euplotia), with description of *Euplotes curdsi sp. nov*. *PLoS One* **11**, e0165442 (2016).
105. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
106. M. G. Grabherr *et al.*, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
107. N. L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
108. M. Lechner *et al.*, Proteinortho: Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics* **12**, 124 (2011).
109. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
110. F. Abascal, R. Zardoya, M. J. Telford, TranslatorX: Multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **38**, W7–13 (2010).
111. T. F. Smith, M. S. Waterman, Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
112. M. Nei, T. Gojobori, Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
113. R. Nielsen, Z. Yang, Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936 (1998).
114. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
115. R. L. Harrison, Introduction to Monte Carlo simulation. *AIP Conf. Proc.* **1204**, 17–21 (2010).
116. D. H. Lynn, "Chapter 4: Phylum ciliophora" in *The Ciliated Protozoa* (Springer, 2008), pp. 89–90.
117. V. Boscaro, F. Husnik, C. Vannini, P. J. Keeling, Symbionts of the ciliate *Euplotes*: Diversity, patterns and potential as models for bacteria–eukaryote endosymbioses. *Proc. R. Soc. B Biol. Sci.* **286**, 20190693 (2019).
118. M. Kolisko, V. Boscaro, F. Burki, D. H. Lynn, P. J. Keeling, Single-cell transcriptomics for microbial eukaryotes. *Curr. Biol.* **24**, R1081–R1082 (2014).
119. W. Zheng *et al.*, Insights into an extensively fragmented eukaryotic genome: De novo genome sequencing of the multinuclear ciliate *Uroleptopsis citrina*. *Genome Biol. Evol.* **10**, 883–894 (2018).
120. E. Lasek-Nesselquist, M. D. Johnson, A phylogenomic approach to clarifying the relationship of *Mesodinium* within the ciliophora: A case study in the complexity of mixed-species transcriptome analyses. *Genome Biol. Evol.* **11**, 3218–3232 (2019).
121. M. Mozzicafreddo *et al.*, The macronuclear genome of the Antarctic psychrophilic marine ciliate *Euplotes focardii* reveals new insights on molecular cold adaptation. *Sci. Rep.* **11**, 18782 (2021).
122. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
123. S. Guindon *et al.*, New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
124. V. Lefort, J.-E. Longueville, O. Gascuel, SMS: Smart model selection in PhyML. *Mol. Biol. Evol.* **34**, 2422–2424 (2017).
125. W. M. Fitch, Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.* **20**, 406–416 (1971).
126. J. Farris, A. Kluge, M. EcKardt, A numerical approach to phylogenetic systematics. *Syst. Zool.* **19**, 172–191 (1970).