Peeri and Tuller *Genome Biology* (2020) 21:63
https://doi.org/10.1186/s13059-020-01971-y

Genome Biology

RESEARCH Open Access

High-resolution modeling of the selection on local mRNA folding strength in coding sequences across the tree of life

Michael Peeri[1] and Tamir Tuller[1,2*]

Abstract

Background: mRNA can form local secondary structure within the protein-coding sequence, and the strength of this structure is thought to influence gene expression regulation. Previous studies suggest that secondary structure strength may be maintained under selection, but the details of this phenomenon are not well understood.

Results: We perform a comprehensive study of the selection on local mRNA folding strengths considering variation between species across the tree of life. We show for the first time that local folding strength selection tends to follow a conserved characteristic profile in most phyla, with selection for weak folding at the two ends of the coding region and for strong folding elsewhere in the coding sequence, with an additional peak of selection for strong folding located downstream of the start codon. The strength of this pattern varies between species and organism groups, and we highlight contradicting cases.

To better understand the underlying evolutionary process, we show that selection strengths in the different regions are strongly correlated, and report four factors which have a clear predictive effect on local mRNA folding selection within the coding sequence in different species.

Conclusions: The correlations observed between selection for local secondary structure strength in the different regions and with the four genomic and environmental factors suggest that they are shaped by the same evolutionary process throughout the coding sequence, and might be maintained under direct selection related to optimization of gene expression and specifically translation regulation.

Keywords: Protein-coding sequence evolution, mRNA secondary structure, Gene expression regulation, Comparative genomics, Codon usage

Background

There is growing evidence that local mRNA folding (i.e., short-range secondary structure) inside the coding region is often stronger or weaker than expected, but the explanation for this phenomenon is yet to be fully understood. mRNA folding strength affects many central cellular processes, including the transcription rate and termination [1–3], translation initiation [4–14], translation elongation and ribosomal traffic jams [15–18], co-translational folding [19–21], mRNA aggregation [22], mRNA stability [23, 24], and mRNA splicing [10, 25] (reviewed in [26–28]). Many of these effects are mediated by interactions of mRNA within the CDS (protein-coding sequence) with proteins and other RNAs and may include structure-specific or non-structure-specific interactions.

In recent years, several studies showed evidence for selection acting directly to affect mRNA folding strength within the CDS (Fig. 1a). Studies looking at the CDS as a whole found selection for strong mRNA folding in most species [22, 29–32]. Studies focusing on the beginning of the

* Correspondence: tamirtul@post.tau.ac.il
[1]Department of Biomedical Engineering, Tel-Aviv University, Tel-Aviv, Israel
[2]Sagol School of Neuroscience, Tel-Aviv University, Tel-Aviv, Israel

© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

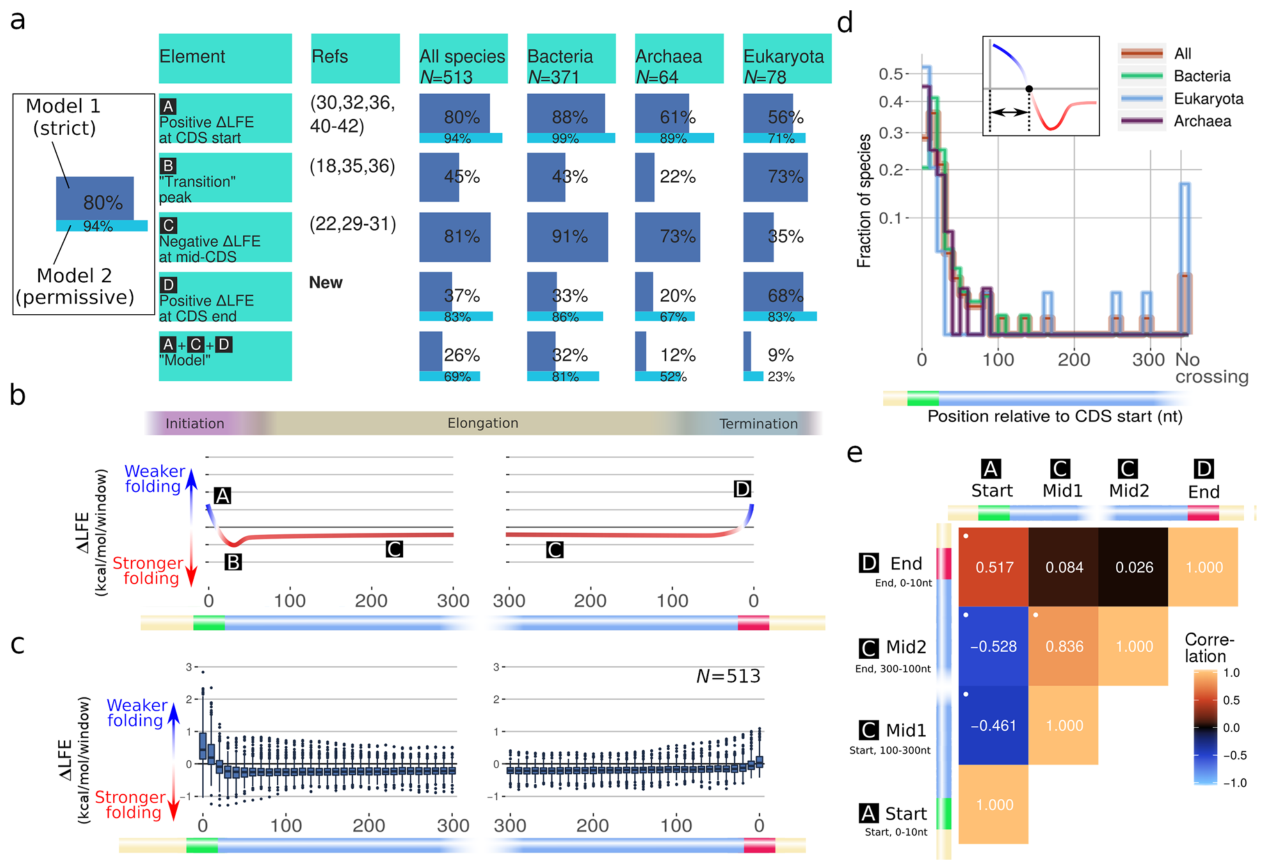


Fig. 1 Common regions of folding bias (ΔLFE) are present across the tree of life, but are not universal. There is correlation between the strengths of these regions in different species, indicating there are factors influencing the bias throughout the coding sequence. **a** Summary of profile features with the fraction of species in which each feature appears in each domain (based on model 1 rules; see "Analysis" under the "Methods" section for details). The results based on the less restrictive model 2 rules (with weaker ΔLFE near the CDS edges not required to be positive; see "Analysis" under the "Methods" section) are shown in bright blue below each bar. References shown here are based on comparison to randomized sequences (i.e., equivalent to ΔLFE). **b** Scheme illustrating profile features reported separately in previous studies within the CDS, showing features [A]–[D] from **a**. **c** Observed distribution of ΔLFE profile values at different positions relative to CDS start (left) and end (right). **d** The distances (in nt) from the start codon where ΔLFE transitions from positive to negative, for species belonging to different domain. The lengths of the initial weak folding region range up to 150 nt in some bacteria. **e** Spearman's correlations between mean ΔLFE profile values in regions [A], [C], and [D]. White dots indicate significant correlation (p value < 0.01)

coding region (i.e., the first 40–50 nucleotides) found evidence for the inverse, with selection acting to weaken mRNA folding in that region [30, 32–34]. In addition, there is some evidence for specifically strong folding in nucleotides 30–70, which may slow down translation elongation near the 5′ end of the mRNA, possibly to prevent ribosomal traffic jams [18, 35, 36]. Finally, it has been suggested that folding is weakened in the region leading to the stop codon [32–34], but not in a way that attributes this weakening to direct selection on folding strength rather than a side effect of some other bias in this region. These results are generally in agreement with available small-scale (e.g., [13, 14]) and large-scale [10–12, 24, 37–39] experimental validation performed in model organisms. Some of these characteristic regions were found to be correlated with genomic GC-content and to be stronger in highly expressed genes [29, 36, 40–42]. However, the previous studies cited did not systematically examine how the selection on folding strength changes along the coding sequence and how this phenomenon varies across the tree of life. In this study, using high-resolution analysis of the folding selection profiles in over 500 organisms from the three domains of life, we examine all data under a common framework and under more stringent controls (including accounting for the evolutionary distances between species), to determine which correlations are likely to stem from causal relationships involved in maintaining mRNA folding. We show that the previously proposed patterns of local selection on mRNA folding are not universal and examine their association with genomic and environmental factors in different taxonomic groups to better understand the underlying evolutionary processes.

Results

To test different hypotheses related to direct selection acting on the local folding energy (LFE) in different regions of the coding sequence, we measured the mean deviation in LFE between the native and randomized sequences (maintaining the amino acid sequence of all CDSs as well as codon and nucleotide composition including the GC-content, see "Analysis" under the "Methods" section for more details). The resulting deviation values, denoted ΔLFE, measure the increase or decrease in local mRNA folding energy relative to what we expect based on the encoded protein and codon frequencies. Any significant deviation from random can be attributed to a specific arrangement of codons that supports increased or decreased base-pairing and folding strength along the mRNA strand (Fig. 2a).

Specifically, if the null hypothesis used to generate the randomized sequences holds for the native sequences at some position, we expect ΔLFE to be 0. Otherwise, a significant deviation from ΔLFE = 0 indicates that the local folding energy values cannot be explained by selection on amino acid content, codon bias, or GC-content alone and serves as evidence for direct selection on local folding energy (Fig. 2a). Positive ΔLFE indicates putative selection for weaker secondary structure, while negative ΔLFE corresponds with selection for stronger secondary structure.

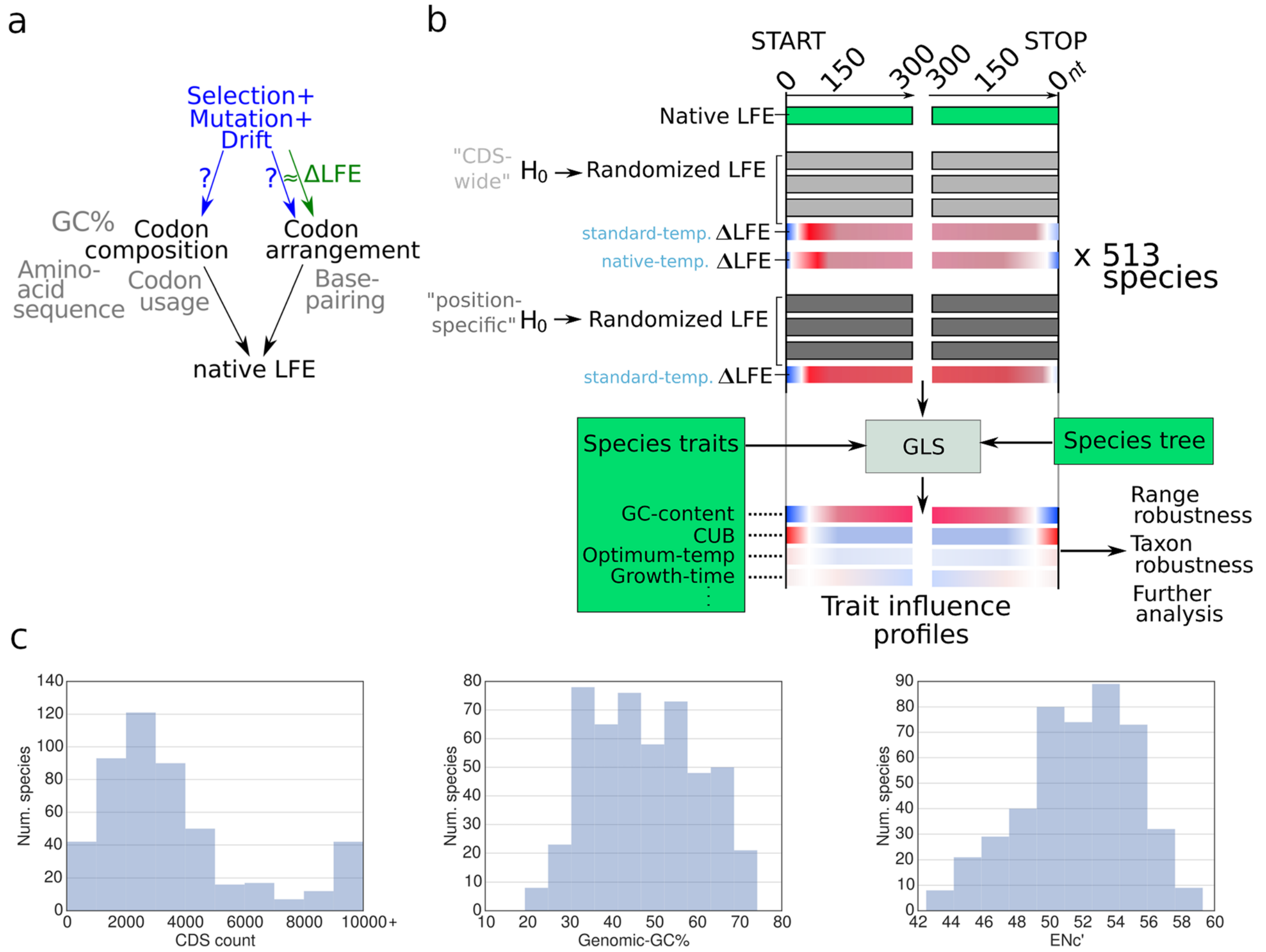


Fig. 2 Overview of the computational analysis to measure ΔLFE while controlling for other factors known to be under selection at different regions of the coding sequence and find factors correlated with it. **a** The variables and concepts involved in determining local folding strength and calculating ΔLFE. The effects of the compositional factors on the left side are removed in order to specifically measure the contribution of codon arrangements to the native folding energy. Blue arrows indicate possible selection forces. **b** The different steps in the computational pipeline used to estimate ΔLFE and the factors affecting it (see "Analysis" under the "Methods" section). For each genome, the CDSs are randomized based on each null model (CDS-wide and position specific), to calculate a mean ΔLFE profile based on that null model. At the next step, based on GLS, correlations between features of the ΔLFE profile and genomic/environmental features are computed. Input data sources (native CDS sequences, species trait values, species tree) are shown in green. **c** The distributions of some genomic properties within the dataset—CDS count, genomic GC-content, and genomic ENc′ (measure of CUB). The dataset was designed to represent a wide range of values (among other considerations; see "Species selection and sequence filtering" under the "Methods" section)

We specifically aimed at finding nearly universal patterns in ΔLFE, as well as groups of organisms and specific organisms with profiles deviating from such patterns. The resulting ΔLFE profiles were subsequently used with the evolutionary tree of the analyzed organisms to detect association between ΔLFE and genomic and environmental traits that cannot be explained by taxonomic relatedness alone and therefore may hint at underlying causal relations. We discuss the influence of genomic features such as codon usage bias (see the "Correlation between codon usage bias and ΔLFE" section) and GC-content (see the "Correlation between GC-content and ΔLFE" section), and of environmental features like intracellular life (see the "Weak ΔLFE in endosymbionts and intracellular organisms" section) and growth temperature (see the "Weak ΔLFE in hyperthermophiles" section).

Conserved regions of folding bias (ΔLFE)

We observed that significant ΔLFE is present in most species and in most regions of the CDS (Fig. 3, Fig. 1a, c). The mean ΔLFE profiles of most species share the same structure (Fig. 3a, Fig. 1b, c), as follows. The region

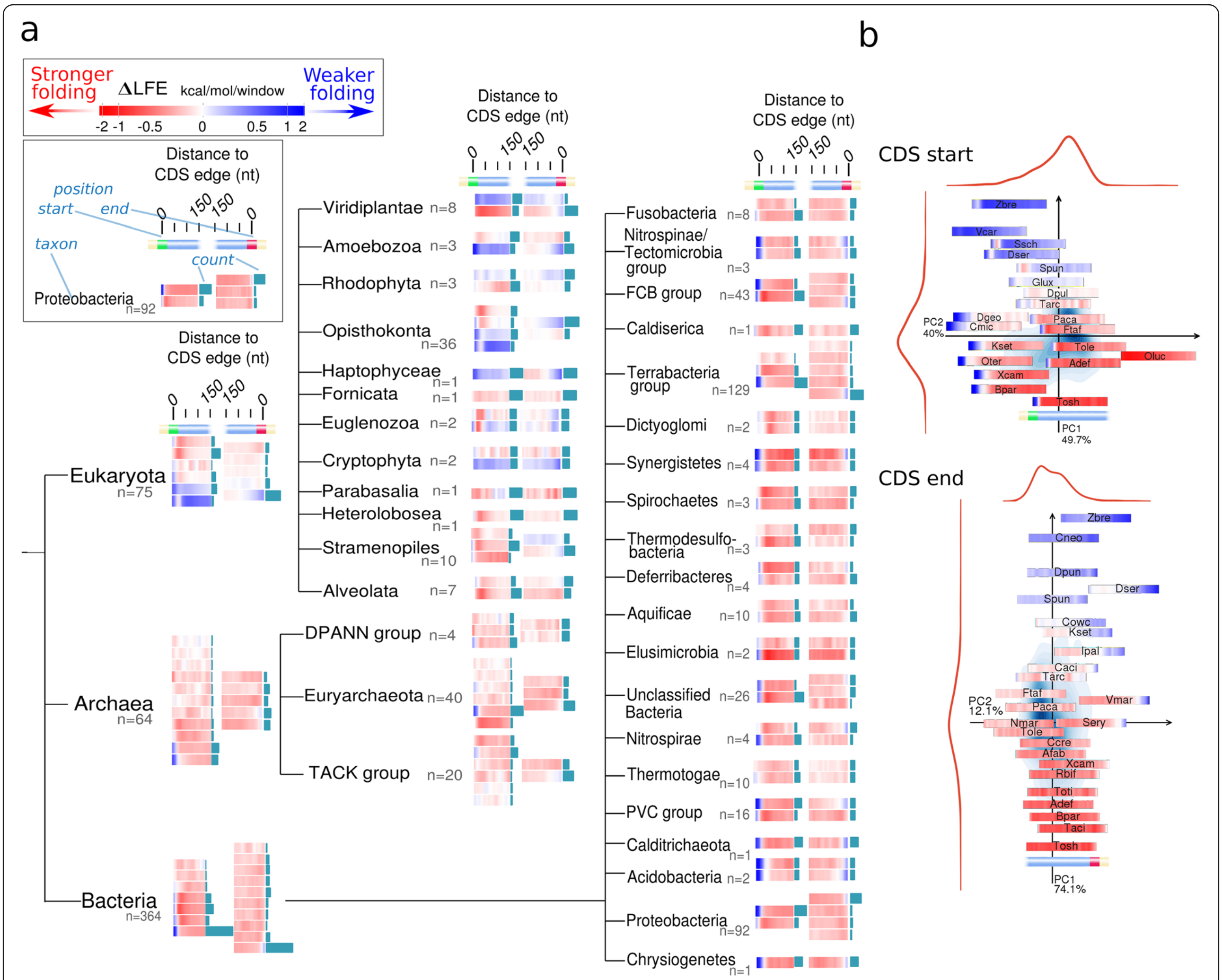


Fig. 3 Two summaries of the ΔLFE profiles demonstrate the consistency and diversity found. **a** Characteristic ΔLFE profiles for species belonging to different taxons. The format of the plots appears in the upper left corner: ΔLFE bias is shown (by color) for windows starting in the range 0–150 nt relative to the CDS start, on the left, and CDS end, on the right; red denotes negative ΔLFE (stronger-than-expected folding) while blue denotes positive ΔLFE (weaker-than-expected folding; see the scale at the upper-right corner). The characteristic profiles for each taxon were calculated using clustering analysis, by grouping similar species according to the correlation between their profiles (see "Visualization" under the "Methods" section for details). The bars (in turquoise) appearing to the right of each characteristic profile indicate the relative number of species it represents. The full ΔLFE profiles for all species appear in Additional file 1: Figure S7. **b**. Summary of ΔLFE profile diversity for all species using dimensionality reduction to 2 dimensions with PCA (see explanations about PCA in the main text), with similar values (profiles) mapped to nearby positions. Background shading (blue) indicates density (see "Visualization" under the "Methods" section for details). This shows most species have similar profiles (located near the center), but different kinds of less typical profiles are also represented. Top, CDS start; bottom, CDS end. Short species names are listed in Additional file 1: Table S3

immediately following the CDS start (typically extending through the windows starting at positions 0–20 nt (Fig. 1a, region A), with a median of 20 nt/10 nt/20 nt in bacteria/archaea/eukaryotes, respectively) has positive mean ΔLFE (evidence of selection for weak folding), usually followed by a transition to negative mean ΔLFE (indicating selection for strong folding) within the first 50 nt and maintained throughout most of the CDS (Fig. 1a region C, Fig. 1c, d). The negative ΔLFE tends to weaken in the area immediately preceding the last codon (typically nucleotides 50–0 nt with median of 50/90/40 nt in bacteria/archaea/eukaryotes, respectively, Fig. 1d) in 83% of the species, and ΔLFE becomes positive there (indicating weaker-than-expected folding) in 37% of the species (including 68% of eukaryotes). This evidence of selection for weak mRNA folding near the stop codon in many organisms across the tree of life is reported here for the first time; two previous studies [18, 32] reported that the local folding energy (LFE) is weak near the start codon in three organisms and without showing that it cannot be explained by direct selection on the amino acid sequence (e.g., using computation of ΔLFE as was done here).

To measure how frequently these elements appear together within the same species, we tested them against a model, based on two variants. The stricter variant, model 1, counts species in which the regions of weak folding at the beginning and end of the CDS have, on average, weaker than expected folding, i.e., significantly positive ΔLFE. The less restrictive model 2 requires folding in these regions to be significantly weaker than in the middle of the CDS, but not necessarily significantly weaker than random (see "Analysis" under the "Methods" section for details). Since the models are applied to the mean ΔLFE of a population of genes which may vary greatly in their individual values, both estimates of the adherence to the model are informative. The combined models (composed of the three regions described) are found in 23% (model 1) and 69% (model 2) of the species analyzed (Fig. 1a), appearing very frequently in bacteria but also commonly in archaea and eukaryotes. The conservation of the ΔLFE profile structure in species across the tree of life is evidence of its biological significance.

GC-content and LFE both change during evolution, and it is worthwhile to compare their level of conservation in related species. LFE is to a large degree determined by GC-content (as evident by the almost perfect correlations found between GC-content and native or randomized LFE, Additional file 1: Figure S1), so one might argue the observed ΔLFE is a side effect of selection acting on GC-content. However, we found that the ΔLFE profile is more conserved than genomic GC-content at any phylogenetic distance within the same domain (Additional file 1: Figure S2). We also found that the profile does not consistently correlate with local variation in CUB (Additional file 1: Figure S3), demonstrating that the results reported here are not side effects of selection on codon bias (e.g., due to adaptation to the tRNA pool).

Additional tests also support direct selection acting to maintain folding strength. ΔLFE profile features are also preserved when calculated using a null distribution that maintains the codon distribution at any position in the CDS relative to the CDS start; thus, local (position-specific) genomic amino acid or codon distributions are not enough to explain the ΔLFE profile (Additional file 1: Figure S4). These features appear in many cases to be stronger in highly expressed genes, genes coding for highly abundant proteins, and genes with a strong codon adaptation to translation elongation, I_TE [43] (see Additional file 1: Figure S5). Finally, these results remain after controlling for the strength of the Shine-Dalgarno binding in the 5′-UTR [44] (Bahiri Elitzur S, Cohen-Kupiec R, Fine L, Yacobi D, Apt B, Diament A, et al.: Prokaryotic rRNA-mRNA interactions are involved in all translation steps and shape bacterial transcripts, Manuscript submitted for publication 2020) and for genes with short or overlapping 5′-UTRs (see, for example, [45]). Together, these results show that the ΔLFE profiles are unlikely to be explained as side effects of selection for a genomic or CDS position-dependent compositional bias in nucleotide, codon, or amino acids acting alone, although many such biases have been reported and are believed to have important biological effects [36].

Note that the randomized LFE profiles also are not always flat, revealing some residual influence on LFE, caused by the amino acid frequencies at different regions, remains even after randomization. ΔLFE controls for this by separately measuring the folding energy biases found in each position.

The different elements making up the model profile structure have functions associated with them. The weak folding region at the beginning of the coding region may improve access to the regulatory signals in this region (e.g., the start codon) [5, 36]. The region of positive ΔLFE preceding the CDS end may help recognition of the stop codon and ribosomal dissociation from the mRNA and prevent ribosomal read-trough. Strong folding in the middle of the coding sequence may assist co-translational folding [19–21] by slowing down translation in specific positions to allow protein folding or other co-translational processes to take place, as well as regulate mRNA stability [23] or prevent mRNA aggregation [22].

The division of the profile into the three regions described here is also apparent when the data is analyzed in an unsupervised manner via principal component analysis (PCA) [46] (Fig. 3b and Additional file 1: Figure S6). This arranges species on a two-dimensional plane according to their ΔLFE profiles, so species with more similar ΔLFE profiles are placed closer together. The resulting plots (for the beginning and end of the coding sequence) show the majority of species have similar ΔLFE profiles (located very close to each other near the center of the plot), with positive ΔLFE near the ends of the coding sequence and negative ΔLFE in the middle of the coding sequence. Groups of species containing other types of profiles are arranged around them on the plots. At either end of the coding sequence, 2 variables (principal components) are sufficient to describe at least 85% of the variability between all ΔLFE profiles, supporting the division of the ΔLFE into three regions (since the mid-CDS region appears in both analyses, see Fig. 1e).

In 45% of the organisms, we found an additional feature: a peak of selection for strong mRNA folding around 30–70 nt downstream of the start codon (Fig. 1a, region B). It was suggested ([34, 35], based solely on evidence in *Eschericia coli* and *Saccharomyces cerevisiae*) that this peak is responsible for increasing translation throughput, by minimizing ribosomal traffic jams occurring because of uneven translation elongation rates throughout the CDS. There is also some evidence [4, 9] that strong secondary structure downstream of the start codon can enhance translation. Whatever the mechanism responsible for it, the results here show that this feature is common across the tree of life. This feature was also shown previously to be stronger in highly expressed genes in 3 species [45], and our results extend this claim (see Additional file 1: Figure S5).

The ΔLFE profiles of eukaryotes are much more diverse than those found in prokaryotes. One striking observation is that significant *positive* ΔLFE throughout the mid-CDS region, present in 13% of the eukaryotes tested, is not observed in any of the 371 bacterial species tested except in *Deinococcus puniceus* (Additional file1: Figure S8, see also Fig. 1a). This seemingly universal rule hints at a constraint on bacterial CDSs not obeyed in eukaryotes and is one of two major differences observed between the domains (along with the correlation with genomic-GC, see the "Correlation between GC-content and ΔLFE" section).

Despite these general trends, there is also significant variation in the ΔLFE profiles across and within taxonomic groups. In the subsequent sections, we discuss genomic and environmental factors that explain some of the variation between mean ΔLFE profiles in different species.

Correlations between ΔLFE regions

The strengths of the three major regions of the ΔLFE profile described above are strongly correlated (Fig. 1e): organisms with relatively stronger ΔLFE (in absolute value) in one model region appear to also have stronger ΔLFE in other regions. For example, the 0–20-nt region has a strong negative correlation with the 150–300-nt region (Spearman's $\rho = -0.46$; p value < 1e−8). This correlation remains highly significant for different ranges and when testing using GLS (Additional file 1: Fig. S9). The two mid-CDS regions (relative to CDS start and end) are positively correlated ($\rho = 0.84$, p value < 1e−8), as are the CDS start and end regions ($\rho = 0.52$, p value < 1e−8). These correlations indicate ΔLFE profiles of different species can generally be ordered by magnitude from species having strong (positive or negative) ΔLFE features *throughout* the CDS to those showing weak or no ΔLFE. In eukaryotes, the negative correlation between the CDS start and mid-CDS regions is not present (results not shown), but in this case, neither do the ΔLFE profiles generally follow the structure of positive start ΔLFE and negative mid-CDS ΔLFE and the profile values may continue to change farther away from the CDS edges.

Together, these results suggest that the different elements making up the typical profile structure are influenced at the genome level by a factor or combination of factors acting jointly on all regions and strengthening or weakening |ΔLFE|, as well as distinct factors acting on each region differently. Some factors contributing to this scaling effect are discussed in the following sections.

Correlation between codon usage bias and ΔLFE

Codon usage bias is generally correlated with adaptation to translation efficiency [47–50]. If ΔLFE is also related to selection for translation efficiency, it is reasonable to expect it would correlate with CUB. To test this hypothesis, we used ENc′ (ENc prime, [51, 52]), a measure of codon usage bias (CUB) that compensates for the influence of extreme GC-content values that skew standard ENc (effective number of codons) scores. Indeed, such a correlation is found (Fig. 4, Additional file 1: Figure S10b)—ΔLFE tends to be stronger (in absolute value) in species having strong CUB (low ENc′), and this holds both near the CDS edges and in the mid-CDS regions. Similar results were obtained when using other measures of CUB (CAI [53] and DCBS [49], Additional file 1: Figure S11), and these correlations persist within many individual taxons (Fig. 9, Additional file 1: Figure S10b). In addition, species with strong CUB tend to have ΔLFE profiles that closely match the model elements (Fig. 4b, c), and further analysis shows the correlation of CUB with the ΔLFE profiles is due to correlation with the magnitude of the profiles and not due to specific profile regions (Additional file 1: Figure S12). Since ΔLFE is computed while controlling for the CUB of each sequence, the reported results suggest that organisms with higher selection on CUB also have, "independently" from a statistical point of view, higher selection on ΔLFE.

Using genomic CUB as measure of optimization for efficient translation elongation, we found that it is also a good predictor of the strength of ΔLFE. One interpretation of this is that the genomic variation in ΔLFE can largely be explained not by different species having

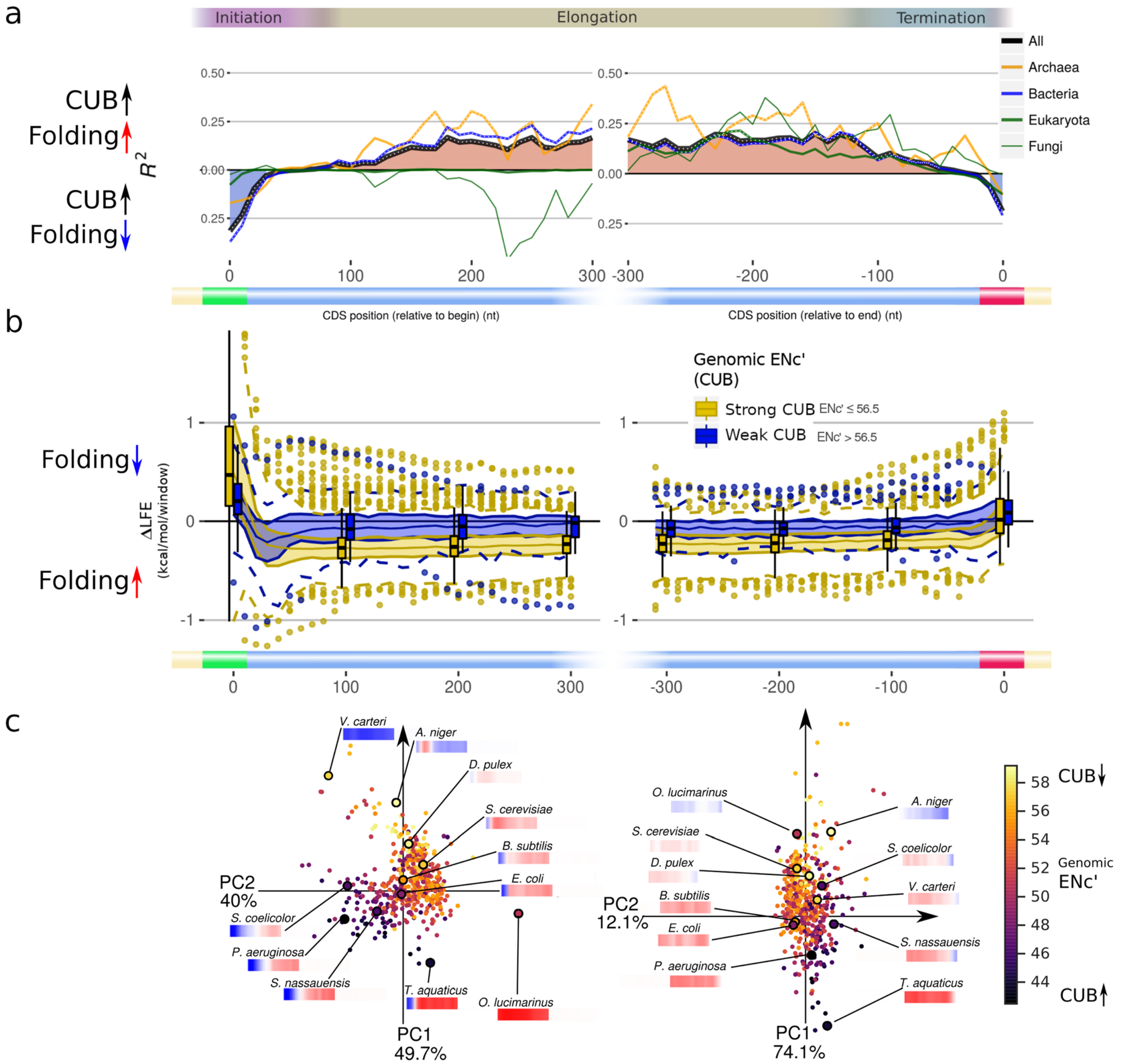


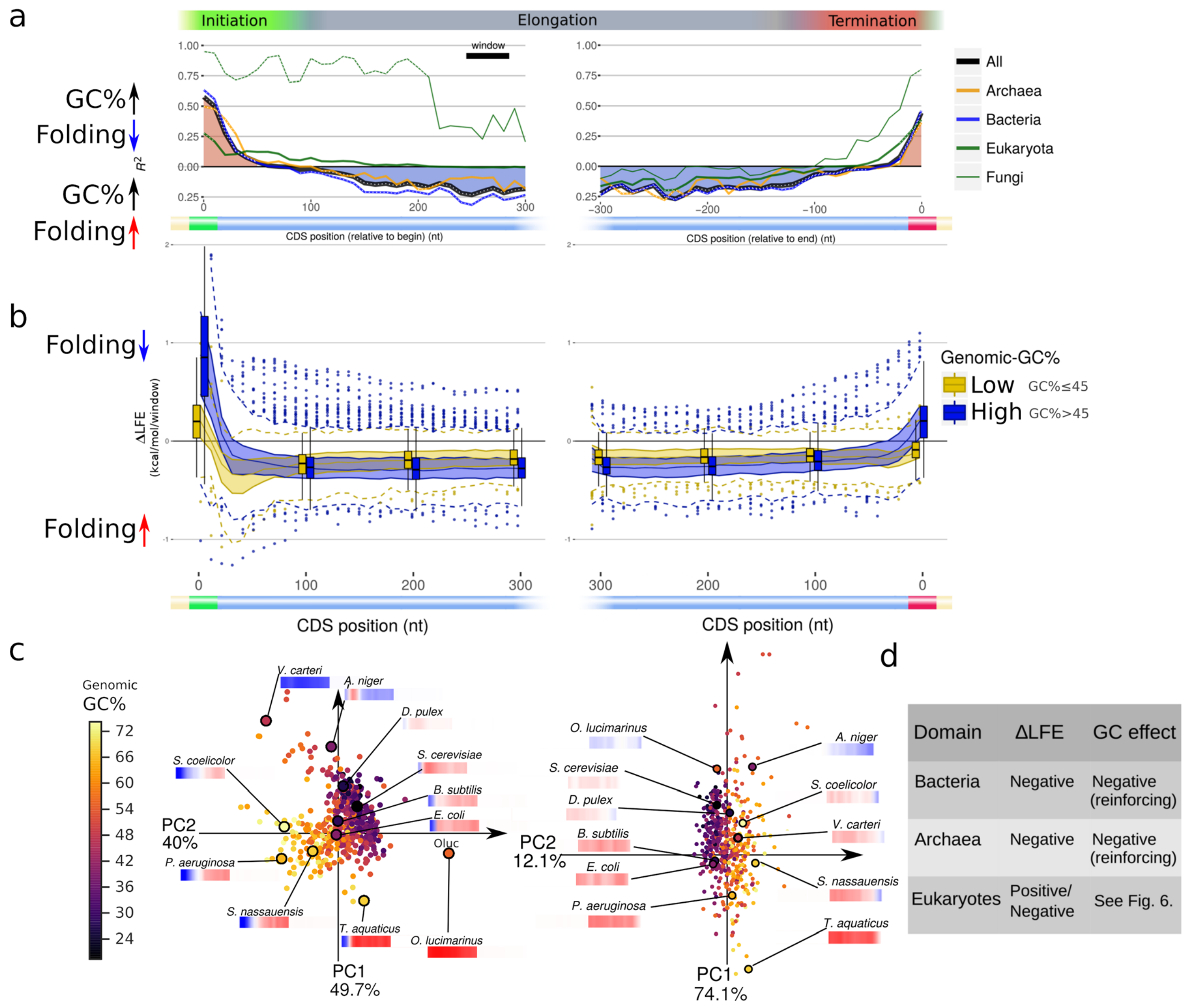
Fig. 4 Folding bias (ΔLFE) is positively correlated with genomic CUB (measured as ENc′) throughout the CDS. This correlation indicates stronger folding bias exists in species with stronger CUB at all regions of the CDS. **a** Correlation strength (R^2, measured using GLS regression) between genomic ENc′ and ΔLFE at different positions relative to the CDS start (left) and end (right). R^2 values below the *X*-axis indicate negative regression slope (i.e., negative correlation with ΔLFE). The regression slope generally has the opposite sign as ΔLFE, indicating strong ΔLFE is correlated with strong codon bias throughout the CDS. Major taxonomic groups are plotted as different colored lines. White dots indicate regression *p* value < 0.01. **b** Comparison of ΔLFE profile values in species with strong vs. weak CUB. Species with strong CUB (yellow, ENc′ ≤ 56.5) tend to have more extreme ΔLFE and show the conserved ΔLFE regions more clearly, while species with weak CUB (blue, ENc′ > 56.6) tend to also have weak ΔLFE. **c** Genomic ENc′ values plotted using coordinates determined by ΔLFE profiles. Species with strong CUB (left plot, lower left quadrant and right plot, right side) have stronger ΔLFE profiles that more strongly adhere to the conserved ΔLFE regions. Coordinates are based on PCA for profile positions 0–300 nt relative to CDS start (left) and end (right). The PCA coordinates are the same as those in Fig. 3b

distinct "target" ΔLFE levels, but by different species having varying "abilities" to maintain ΔLFE in the presence of mutations and drift because the selection pressure is insufficient under their effective population size (either because the selection pressure is low or because the effective population size is low).

Correlation between GC-content and ΔLFE

GC-content is a fundamental genomic feature and is correlated with many other genomic traits and environmental aspects [54, 55]. It might be a trait maintained under direct selection, or merely a statistical measure of the genome that other traits evolve in response to because of its biological and thermodynamic consequences. GC-content is also the strongest factor determining the native LFE (Additional file 1: Figure S1a), since G-C base pairs are more stable than A-T pairs (due to the increase in the number of hydrogen bonds and more stable base stacking). Selection on folding strength (measured by ΔLFE) also influences folding strength, and we would like to measure the correlation between these two factors that influence the folding strength (namely, GC-content and ΔLFE). This is made possible since ΔLFE is calculated relative to the baseline maintaining the GC-content of the original coding regions in the randomized ones (see “Randomization procedures” under the “Methods” section for a description of the null models). This controls for the direct effect of GC-content, allowing us to directly study the interaction between ΔLFE and GC-content (see also Additional file 1: Figure S1a).

The correlations (expressed as R^2) between genomic GC-content and ΔLFE at different points near the CDS



Domain	ΔLFE	GC effect
Bacteria	Negative	Negative (reinforcing)
Archaea	Negative	Negative (reinforcing)
Eukaryotes	Positive/ Negative	See Fig. 6.

Fig. 5 Folding bias (ΔLFE) is positively correlated with genomic GC-content throughout the CDS. **a** The effect of genomic-GC on ΔLFE at each position along the CDS start (left) and end (right), measured using GLS regression R^2 values. R^2 values above the *X*-axis indicate positive regression slope (indicating moderating effect of GC-content); R^2 values below the *X*-axis indicate negative regression slope (i.e., reinforcing effect of GC-content). Near the CDS edges (where ΔLFE is usually positive), genomic-GC generally has a moderating effect on ΔLFE. In the mid-CDS region (where ΔLFE is usually negative), genomic-GC generally has a reinforcing effect on ΔLFE. Major taxonomic groups are plotted as different colored lines. White dots indicate regression *p* value < 0.01. **b** Comparison of ΔLFE profile values in species with high vs. low genomic GC-content. Species with high GC-content (blue, genomic-GC > 45%) tend to have more extreme ΔLFE and show the conserved ΔLFE regions more clearly, while species with low GC-content (yellow, genomic-GC ≤ 45%) tend to also have weak ΔLFE. **c**. Genomic GC-content for all species plotted on the PCA coordinates of their ΔLFE profiles (same coordinates as in Fig. 3b. *N* = 513) for CDS start (left) and end (right). Low-GC species are generally clustered in a small region, indicating they have similar ΔLFE profiles, and that region is characterized by weak ΔLFE. **d** Qualitative summary of ΔLFE in relation to GC-content in the mid-CDS

start and end are shown in Fig. 5a. This dependence shows a similar pattern to that seen in the ΔLFE profiles themselves (Fig. 1c, Fig. 5a, and for the correlation with CUB, see the "Correlation between codon usage bias and ΔLFE" section), with significant correlations appearing in roughly the same CDS regions described for the ΔLFE profiles. The correlation takes the opposite directions in the CDS edges than that maintained throughout the inner CDS region, which means GC-content is positively correlated with the strength of ΔLFE (in absolute value) throughout the CDS (like CUB is).

Near the CDS start, positive correlation (indicating a moderating effect) exists in the windows starting at 0–60 nt (Fig. 5a, Additional file 1: Figure S10a). This effect appears in almost all taxons analyzed, with R^2 values between 0.2 and 0.9 and significant *p* values in most taxons, and may be explained as counteracting the strengthening influence of GC-content on secondary structures to prevent them from hindering the translation initiation process.

The opposite effect exists in the mid-CDS: negative (reinforcing) dependence on genomic GC-content appears in the region at 70–300 nt after CDS start in most bacterial and archaeal taxons (Fig. 5a–c, Fig. 9, Additional file 1: Figure S10a) and is generally maintained throughout the length of the CDS (excluding the edge regions). As mentioned above, selection for strong mRNA folding and mRNA structures inside the coding may be related to transcription elongation [2], co-translational folding [19–21, 26], and mRNA stability [23]. The observed ΔLFE in this region is indeed negative in nearly all bacterial and archaeal species; it is possible that the folding is further reinforced in species higher GC-content since they are under stronger selection for these processes. Note that the effects of genomic GC-content and CUB (see the "Correlation between codon usage bias and ΔLFE" section) are somewhat overlapping, but each factor significantly contributes to the total observed effect (Additional file 1: Figure S13).

In eukaryotes, we observed a wider variation in mid-CDS ΔLFEs (which is not found in other groups), from strongly positive to strongly negative, with a non-linear dependence on genomic-GC (Fig. 6, Fig. 9). Low-GC eukaryotes tend to have weak ΔLFE in the mid-CDS region, while high-GC eukaryotes tend to have strong *positive* or *negative* ΔLFE in the same region. To evaluate this relation, which is not linear, we used maximal information coefficient (MIC) [56, 57], a measure that can capture any statistical dependence including non-linear dependencies. We found that this relation is quite significant (MIC = 0.54, *p* value ≤ 2e−5; see "Analysis" under the "Methods" section). Fungi, however, show a strong positive (moderating) correlation between genomic-GC and ΔLFE (Fig. 5a, Fig. 6a; *Eremothecium gossypii*, GC% = 51.7, is the only observed fungus with GC% > 45 and negative ΔLFE in the mid-CDS region). There are also clear internal disparities in ΔLFE among fungi families (Additional file 1: Figure S7). Note that in some species (e.g., *Zymoseptoria tritici*), the positive ΔLFE seems to extend throughout the CDS. In other species,

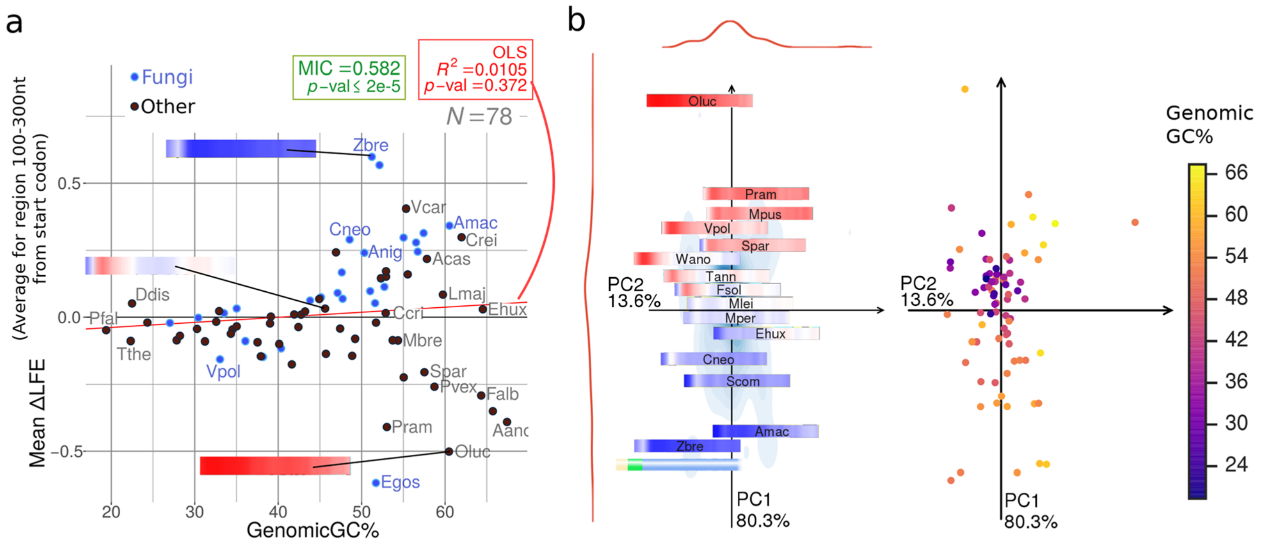


Fig. 6 Genomic-GC effect on ΔLFE in eukaryotes shows divergence in high GC-content species that is not observed in other domains, while low GC-content species have weak ΔLFE. **a** mean ΔLFE values for eukaryotes in the range 100–300 nt from CDS start, plotted against genomic-GC. Fungi are highlighted in blue. There is no linear relation between the variables ($R^2 = 0.01$), but there is strong statistical dependence nevertheless (MIC = 0.582, *p* value < 2e−5, *N* = 78); see some explanation on MIC in the main text. **b** PCA plot for the same species shows the same two classes of ΔLFE profiles in eukaryotes having high genomic-GC (top and bottom regions; see "Visualization" under the "Methods" section for details). On the left, ΔLFE profiles are plotted in the positions given by their first 2 PCA components. On the right, genomic-GC values for the profiles plotted at the same coordinates. Short species names are listed in Additional file 1: Table S3

there is a transition to negative ΔLFE further downstream (as much as 500 nt from CDS start, results not shown).

The group of fungi and other eukaryotes having strong selection for weak local mRNA folding in the mid-CDS region (all of which have high genomic GC-content) runs counter to the general trend in prokaryotes. It is possible that these species are under selection for higher translation elongation speeds, which tend to be hindered by stronger mRNA folding [15–18]; however, it is not clear why such cases are not observed in other groups like bacteria. The correlation with GC-content reported here may also be partially explained by the fact that both GC-content and ΔLFE are affected by common factors such as the ability to maintain the selected sequences under the effective population size. The wide range of ΔLFE values for eukaryotic species and the absence of linear correlation with GC-content (in general) reveal additional factors are involved in this aspect of gene expression.

Weak ΔLFE in endosymbionts and intracellular organisms

Many endosymbionts and other species with intracellular life stages have low effective population sizes, because their life cycle includes recurring population bottlenecks [58, 59] or has lower selective pressure due to reliance on the host [60]. These species generally have weaker ΔLFE compared to their relatives, as can be clearly seen from their ΔLFE profiles (Fig. 7, also see Additional file 1: Figure S7, e.g., *Richelia intracellularis*, *Blattabacterium* sp.). The apparent disparity between endosymbionts and their relatives is strongest near the CDS start. Taken as a whole, the difference in ΔLFE is small (Fig. 7a), but when comparing within smaller taxons, the difference is much more noticeable (e.g., gammaproteobacteria in Fig. 7b–d). Endosymbionts also tend to have lower GC-content and CUB [60], but the results are still generally significant after considering this at least in proteobacteria, where we have a sufficient sample size (Additional file 1: Figure S14). The dichotomic grouping of species as endosymbionts is an oversimplification

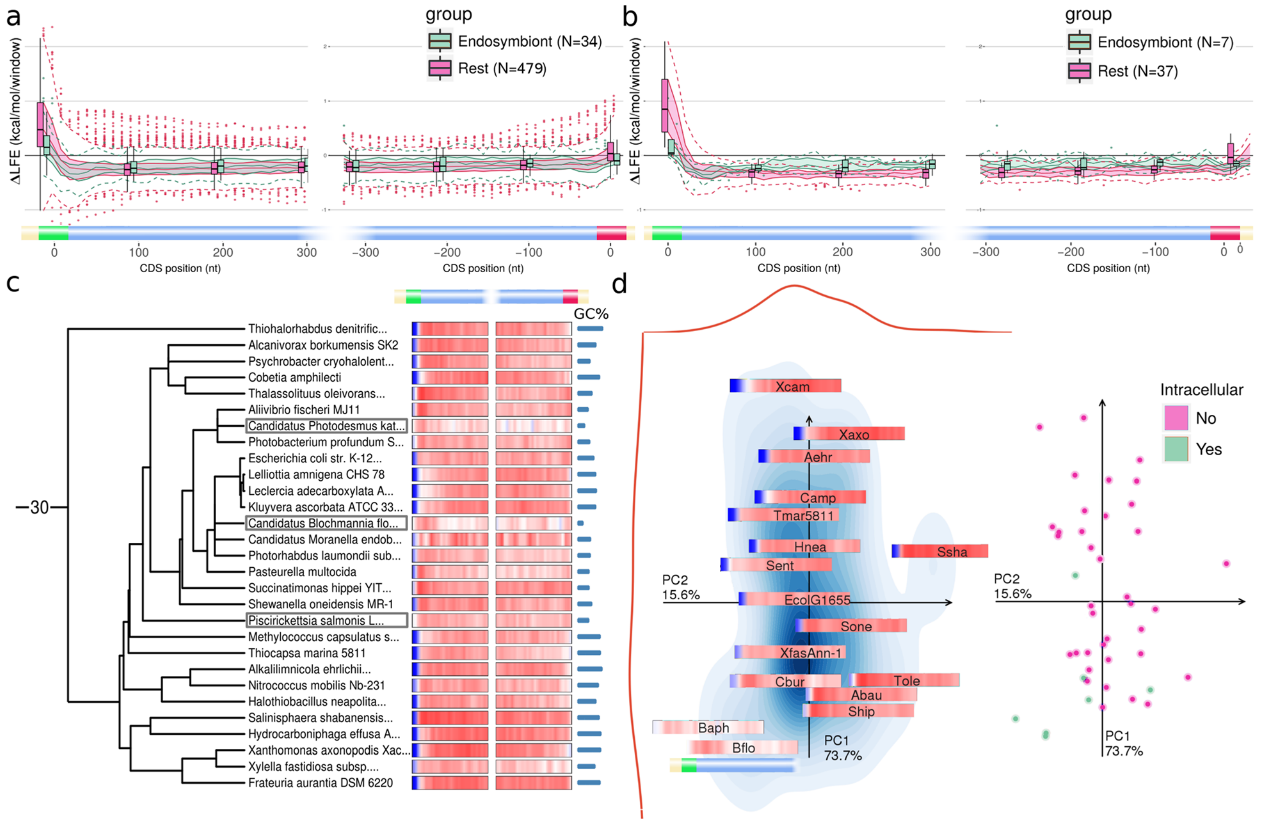


Fig. 7 Endosymbionts and other intracellular species have generally weak ΔLFE. **a** Comparison of ΔLFE values at different CDS positions between endosymbionts (green) and other species (pink). The ΔLFE values are less extreme in endosymbionts, indicating lower selection on local folding strength. **b** Comparison of ΔLFE distributions at different CDS positions between endosymbionts (green) and other species (pink) within gammaproteobacteria (*N* = 44). **c** ΔLFE for species included in the tree within gammaproteobacteria; the endosymbionts and intracellular species (marked) have weaker ΔLFE bias compared to their relatives. **d** PCA plot for ΔLFE profiles (left, see "Visualization" under the "Methods" section) and the intracellular classification (right) for the species in gammaproteobacteria (*N* = 44). For clarity, overlapping profiles are hidden on the left (as in all PCA plots for ΔLFE profiles); all species are plotted on the right. Short species names in the PCA plot on the left panel are listed in Additional file 1: Table S3

and ignores the variety of species with intracellular stages, including obligate and facultative intracellular parasites (and our annotation of species as endosymbionts, based on the literature, may not be complete). Indeed, some species we classify as endosymbionts (e.g., *Halobacteriovorax marinus SJ*) nevertheless have low genomic ENc′ and strong ΔLFE.

Weak ΔLFE in hyperthermophiles

In temperatures approaching the RNA melting temperature, base-pairing is destabilized and it is likely that codon arrangement and ΔLFE can no longer significantly affect the secondary structure. We found hyperthermophilic archaea and bacteria to have weaker (closer to 0) ΔLFE in the mid-CDS region (Fig. 8). This effect is not apparent at lower temperatures (below 65 °C) or across all temperatures, with temperature having no significant correlation with ΔLFE (Fig. 8e, Fig. 9) when controlling for species relatedness. Our results are consistent with [40], which argued for negative correlation with growth temperature, but that paper only analyzed the beginning of the coding region and did not control for the evolutionary relations among organisms. Based on our analysis, the linear relation between

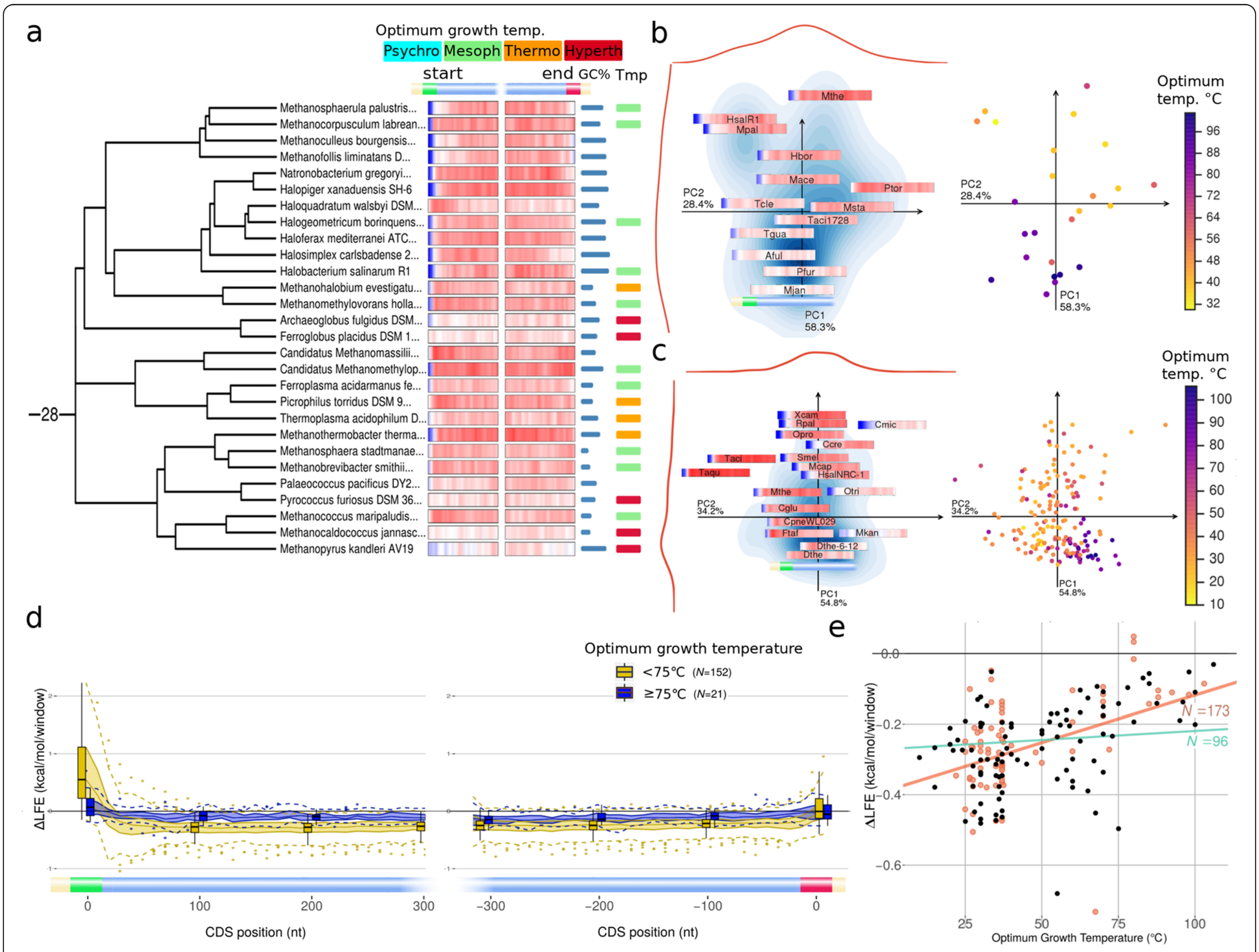


Fig. 8 Hyperthermophiles have weak ΔLFE. **a** ΔLFE profiles (for CDS beginning and end) for members of euryarchaeota covered by the phylogenetic tree (*N* = 28) and their annotated optimum growth temperature classification (mesophile—green, moderate thermophile—orange, hyperthermophile—red) and genomic GC-contents. Hyperthermophiles have weak ΔLFE that cannot be explained by the tree topology or their genomic GC-contents. **b** ΔLFE profiles (left) and optimum growth temperatures (right) for all members of euryarchaeota having annotated optimum growth temperatures (*N* = 25), plotted using their PCA coordinates (see "Visualization" under the "Methods" section). Hyperthermophiles seem to be clustered in a small region characterized by weak ΔLFE. **c** ΔLFE profiles (left) and optimum growth temperature (right) for all species having annotated optimum growth temperature (*N* = 173), plotted using their PCA coordinates (see "Visualization" under the "Methods" section). Short species names from PCA plots are listed in Additional file 1: Table S3. **d** Comparison of ΔLFE values for species having optimum temperature above (blue) or below 75 °C (yellow), for positions relative to CDS start (left) or end (right). **e** Regression for optimum growth temperature vs. mean ΔLFE (average for positions 100–300 nt after CDS start) using GLS (green regression line, *N* = 96, R^2 = 0.004, *p* value = 0.6) and OLS (red regression line, *N* = 173, R^2 = 0.45). The apparent linear relation is no longer significant when controlling for the phylogenetic relationships. Points plotted in red are included only in OLS

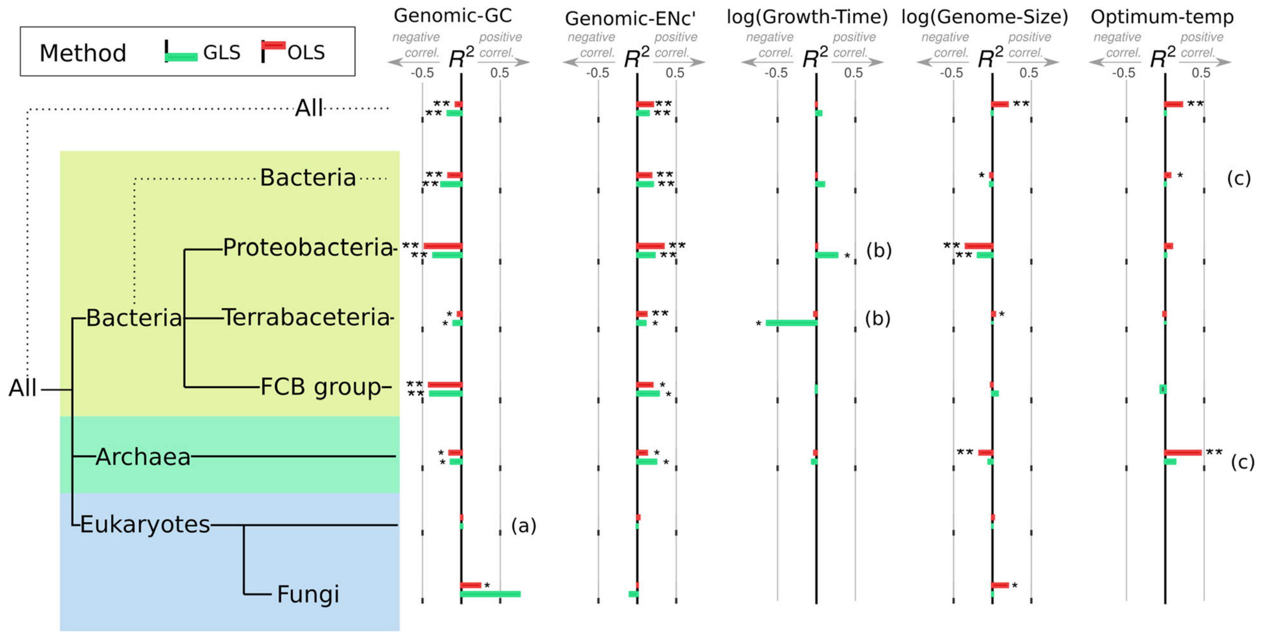


Fig. 9 Summary of trait correlations with ΔLFE in the mid-CDS region for different taxonomic groups. Many of these correlations are discussed in the previous sections. For each taxonomic group and trait combination, correlations are measured using R^2 with GLS (phylogenetically corrected, green bars) and OLS (uncorrected linear relationship, red bars). Significant correlations are marked with "*" (p value < 0.05) or "**" (p value< 0.001). Correlations with genomic-GC% and genomic-ENc′ are robust in prokaryotes, whereas other traits do not have consistent linear relationships. All correlations are for the region 100–300 nt after CDS start. (a) No linear dependence, but a significant relationship does exist (see Fig. 6). (b) Linear dependence appears in GLS but not in OLS. Small sample size exists in some taxons. (c) No significant linear relationship found over the entire range of values, but hyperthermophiles have significantly lower ΔLFE (see the "Weak ΔLFE in hyperthermophiles" section)

temperature and ΔLFE is not generally supported by GLS (Fig. 8e, Fig. 9, Additional file 1: Figure S10c); however, since species tend to have similar temperature requirements as their close relatives, it is hard to conclusively decide if any similarity in ΔLFE is derived from association with temperature or the evolutionary relationship without having considerably more data. In hyperthermophiles (species with optimum growth temperature above 75 °C), however, there is a significant decrease in ΔLFE (even when the folding strengths are predicted at room temperature, Additional file 1: Figure S15). These results suggest that mRNA folding is not effective in higher temperatures (in general), and consequently, ΔLFE is not preserved. In moderate thermophiles, ΔLFE may follow the precedence of genomic GC-content, which previous studied concluded is not an adaptation to high temperatures at the genomic level, but may still be part of such an adaptation at specific rRNA and tRNA sites where secondary RNA structure is particularly important [61, 62].

Discussion

The results we presented here provide a wide integrated view on the way evolution shapes local mRNA secondary structures in the coding regions of organism across the tree of life. In addition, the results include novel attempts to tie this phenomenon to genomic, evolutionary, and environmental variables in the hope of further clarifying the processes involved. In this section, we will summarize and discuss key results.

First, we show that selection on mRNA folding strength in most (but not all) species follows a conserved structure with three distinct regions (Fig. 1)—decreased local folding strength at the beginning and end of the coding region and increased folding strength in mid-CDS. The fact that this structure is more conserved than other genomic traits like GC-content (Additional file 1: Figure S2), as well as its alignment to the coding regions, suggests these features are related, at least in part, to translation regulation. Our statistical tests demonstrate that these features cannot be merely side effects of factors known to be under selection like codon usage bias and amino acid composition.

In general, the model features for the beginning and mid-CDS appear much more frequently in the analyzed organisms (appearing in around 80% of the organisms), while selection for weak folding near the stop codon, first demonstrated here, is comparatively rare (it appears in around 37% of the organisms). This may suggest that generally, the first two features tend to be under stronger selection (possibly since they tend to contribute more significantly to organism fitness).

Conformance to different model elements varies significantly between the three domains: weak folding at the beginning of the coding regions appears in the great majority of bacterial species (88%) but only in 56%/60% of eukaryotes/archaea, respectively (Fig. 1a, Fig. 3a). These differences may be related to polycistronic gene expression (see Additional file 1: Figure S16) or to generally higher effective population sizes and selection for high growth rate in bacteria; they may also indicate complementary constraints imposed by eukaryotic gene expression mechanisms (e.g., Cap-dependent translation initiation) and unique environmental constrains in archaea. On the other hand, selection for weak mRNA folding at the end of coding region (first conclusively shown here) is much more frequent in eukaryotes (appearing in 68% of the analyzed organism) than in the prokaryotes (20% in archaea and 33% in bacteria). This may be related to alternative mechanisms for efficient translation termination fidelity in prokaryotes (including mRNA folding outside the boundaries of the CDS) and/or to translation of polycistronic transcripts (see [63] for related observations in the 3′-UTR).

Second, we found that in some eukaryotes (in 13% of the analyzed eukaryotes and in one bacteria: *D. puniceus*), there is significant *positive* ΔLFE throughout the mid-CDS region (i.e., opposite to the general trend in prokaryotes, Fig. 1a, Fig. 6, Additional file 1: Figure S8). This phenomenon, more widespread than previously reported, may be related to selection improving elongation speed [18]. It is currently not clear why this type of selection appears only in these eukaryotes and is extremely rare in the other domains.

Third, we show that the "transition peak," a region of selection for strong mRNA folding beginning around 30–70 nt downstream of the start codon that was reported elsewhere to be associated with translation efficiency [18, 35, 36, 45], appears frequently (45%) in the analyzed organisms, indicating this mechanism is common (Fig. 1a, c). This feature appears much more frequently in eukaryotes (73%) than in prokaryotes (22% in archaea and 43% in bacteria). Here, too, it is possible the lower frequency in prokaryotes hints at a complementary mechanism for translation initiation and elongation efficiency and fidelity in prokaryotes.

Forth, despite these differences, we found strong correlation between the strengths of three profile elements (found at the beginning, middle, and end of the coding regions, Fig. 1e) across the analyzed organisms. This supports the conjecture that much of the variation in their strength among organisms is caused by common factors acting jointly on the level of ΔLFE at all regions of the CDS.

Fifth, we discussed several variables that correlate with ΔLFE (and account for much of the variation mentioned above). The variables showing the strongest correlation are genomic GC-content (despite being explicitly controlled for by our randomizations as explained above, Fig. 5) and CUB (measured using ENc′, Fig. 4). Strong CUB and higher GC-content tend to be associated with more efficient selection on translation efficiency (see, for example, [64, 65]), and the fact that ΔLFE is correlated with them suggests the same underlying mechanism (or mechanisms) contributes to their selection.

The influence on ΔLFE of all traits analyzed in the mid-CDS region can be compared in Fig. 9. Other genomic and environmental traits analyzed (including genome size and growth time) were not found to have significant linear interaction with ΔLFE at the domain level. In many cases, there appears to be potential interaction with ΔLFE in smaller taxons (which may or may not be due to real interactions specific to those taxons, Additional file 1: Figure S10).

Sixth, we proposed four specific characteristics of species having weak ΔLFE (separately and together), demonstrating the conditions in which ΔLFE cannot be effectively maintained (or does not yield sufficient benefit to be maintained). The first two characteristics are based on the correlated traits described above: low GC-content and low CUB. Another characteristic is optimum growth temperature, since in higher temperatures, base-pairing is weakened, and consequently, the influence of codon arrangement and composition must also be reduced, and so is any possible effect of ΔLFE. The last disrupting factor, an intracellular life phase, stems from the fact that such organisms generally have lower effective population size (due to recurring population bottlenecks) and lower selection pressure on gene expression (because they partly rely on the host, [58, 59]). A binary classification model based on these four features has precision 0.66 and recall 0.82 in classification of ΔLFE strength (see "Analysis" under the "Methods" section and Fig. 10). Note that this binary classification discriminates species with very weak ΔLFE and has weak predictive value for ΔLFE strength in species where none of the factors hold, giving $R^2 = 0.2$ (p value = 5e−25, OLS, all species) against mean |ΔLFE| in the 150–300-nt region relative to CDS start. These conditions support the proposed mechanism of ΔLFE being the result of selection on secondary structure strength related to gene expression regulation and efficiency.

Our results point to cases where evolutionarily close organisms exhibit very different ΔLFE patterns and selection levels. For example, in fungi, members of *Pezizomycotina* (such as *Aspergilus niger* or *Zymoseptoria brevis*) have much more positive ΔLFE compared to members of *Saccharomycotina* (including *Eremothecium gossypii* and *Candida albicans*). Notably, a few eukaryotic species (e.g., the unrelated species *Fonticula alba* and *Saprolegnia parasitica*) have a ΔLFE profile that looks typical for bacteria (Additional file 1: Figure S7). This highlights the variety of gene expression mechanisms in eukaryotes, as well

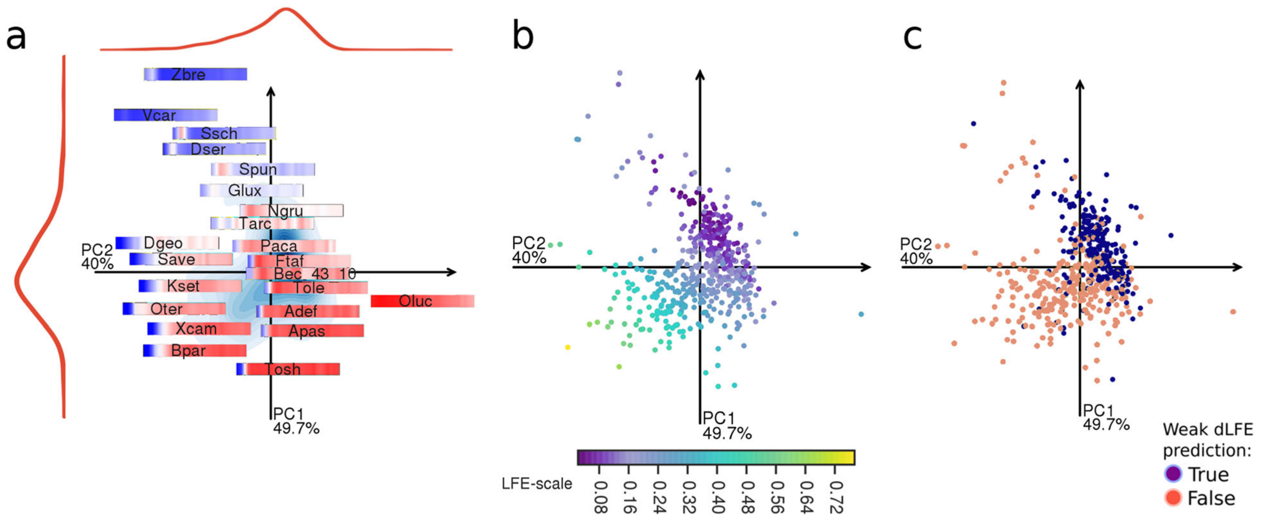


Fig. 10 Classification model for weak ΔLFE based on four species traits. **a** PCA plot of ΔLFE profiles relative to CDS start (see "Visualization" under the "Methods" section). Short species names are listed in Additional file 1: Table S3. **b** ΔLFE profile strength, measured using standard deviation, for profile positions 0–300 nt relative to CDS start. **c** Predicted ΔLFE strength for each species using binary model for weak ΔLFE (precision = 0.66, recall = 0.82, *N* = 513, see "Binary classifier for ΔLFE strength" under the "Methods" section)

as the risk in generalizing about disparate groups based on observations on model organisms.

We would also like to emphasize the fact that ΔLFE has been considered a direct result of selection by previous studies cited here; we believe our results further support this hypothesis, for example, by showing ΔLFE is more conserved than genomic GC-content and demonstrating biologically reasonable trait interactions that may indicate a (direct or indirect) causal link. We should note however that our methodology does not assume any specific evolutionary process at work to produce the measured ΔLFE and this is an additional topic for further research.

Finally, we should note our analysis is based on average values over entire genomes. This provides important statistical power and reduces the random effects of other factors on specific genes. It is important to remember, however, that some of the gene-level factors filtered this way are nevertheless important and there is considerable variation between genes. This means that the reported features should be further analyzed in higher resolution, as well as validated experimentally to understand their origin. For example, ΔLFE in the mid-CDS region was suggested to be influence by both global factors like mRNA aggregation and local factors, like co-translational folding [29, 31], which may cause non-uniform selection pressure across the CDS. These differences may allow the effect of each factor to be experimentally validated separately. In addition, in future studies, it will be helpful and challenging to study the relation between ΔLFE and the position of genes in the operon (see [63]), and the influence of ΔLFE on the outcomes of translation initiation, termination, and splicing.

Conclusions

1. The previously proposed regions of selection on local mRNA folding strength are widespread and appear in many species across domains. For two such regions (strong folding downstream of the beginning of the CDS and weak folding near the CDS end), this is first conclusively demonstrated here. However, none of these regions is universal and exceptions, which sometimes run opposite to the common trend, are quite common. Nevertheless, the CDS in most species does contain consistent regions of tendency for increased or decreased secondary structure strength. These regions coincide with parts of the CDS involved in different gene expression processes and in particular different stages of mRNA translation (initiation, elongation, and termination), supporting the conjecture that mRNA folding strength has a role in these stages of mRNA translation. In addition, stark differences in the prevalence of the regions suggest interactions with domain-specific regulatory mechanisms: For example, the selection for weak folding at the end of the coding region seems to be more common in eukaryotes while the selection for weak folding at the beginning of the coding region appears more commonly in prokaryotes.
2. The tendencies for increased or decreased secondary structure strength in different parts of the coding sequence are correlated among species across the tree of life, indicating common factors

are affecting them throughout the coding sequence. We present four factors that predict the strength of local mRNA folding selection within the coding sequence—GC-content, CUB, intracellular life stage, and a hyperthermophilic environment. These factors are characteristic of species with strong optimization for gene expression efficiency or fidelity, suggesting mRNA folding strength also contributes to this optimization.

3. A “transition peak” of selection for strong mRNA folding around 30–70 nt downstream of the start codon appears in ~ 50% of the analyzed organisms, showing this phenomenon (suspected of being linked to optimization of translation elongation) is widespread.
4. The statistical framework we proposed for studying position-specific selection effects on traits like local mRNA folding across taxonomic groups, while controlling for confounding factors such as amino acid bias, codon, and evolutionary distance, enables inferring factors that may directly affect these traits.

Methods

Analysis

Species selection and sequence filtering

The set of species included in the dataset (Additional file 1: Table S1, Additional file 2) was chosen to maximize taxonomic coverage, include closely related species which differ in GC-contents and other traits (Fig. 2c), and take advantage of the limited overlap between available annotated genomes, NCBI environmental traits data, and the phylogenetic tree (see below). To prevent under-representation of taxons in the dataset, included species were tabulated by phylum and species from missing phyla and classes were added if possible (Additional file 1: Table S2). Over-representation of closely related species is controlled by GLS (see below).

CDS sequences and gene annotations for all species were obtained from Ensembl genomes [66], NCBI [67], JGI [68], and SGD [69] (Additional file 1: Table S3). CDS sequences were matched with their GFF3 annotations to filter suspect sequences, as follows. The dataset excludes CDSs marked as pseudo-genes or suspected pseudogenes, incomplete CDSs, and those with sequencing ambiguities, as well as CDSs of length < 150 nt. If multiple isoforms were available, only the primary (or first) transcript was included. Genes annotated as belonging to organelle genomes were also excluded. Genomic GC-content, optimum growth temperatures, and translation tables were extracted from NCBI Entrez automatically, using a combination of Entrez and E-utilities requests (Additional file 1: Table S3). A few general characteristics of the included CDSs are shown in Fig. 2c.

The taxonomic hierarchy and classifications used to analyze and present the data were obtained from NCBI Taxonomy. Endosymbionts were annotated using a literature survey (Additional file 1: Table S3). Growth rates were extracted from [52] (Supplementary Table A1).

Randomization procedures

To test different hypotheses regarding local folding energy (LFE), native sequences were compared against randomized sequences preserving attributes as defined by each null hypothesis, as follows (Fig. 2a, b):

To test the hypothesis that the native *arrangement* of synonymous codons causes a significant bias in LFE, synonymous codons were randomly permuted within each CDS (i.e., all codons encoding for the same amino acid within a given CDS are randomly rearranged). This “CDS-wide” randomization preserves the encoded protein sequence, nucleotide frequencies (including GC-content), and codon frequencies of each CDS (but generally disrupts longer-range dependencies). Synonymous codons were determined according to the nuclear genetic code annotated for each species in NCBI genomes.

To test the contribution of position-specific biases in amino acid composition, nucleotide frequencies, and codon frequencies including CUB (factors that are equalized at the CDS level by the CDS-wide randomization) on the observed LFE, a second “position-specific” randomization was used. In this randomization, synonymous codons were randomly permuted between codons found at the same position (relative to the CDS start) across all CDSs in each genome. This randomization preserves the amino acid sequence of each CDS, while nucleotide (including GC-content) and codon frequencies are preserved at each *position* across a genome.

LFE profile calculation

Local folding energy (LFE) profiles were created by calculating the folding energy of all 40-nt-long windows, at 10-nt intervals, relative to the CDS start and end, on each native and randomized sequence. This measure estimates local secondary structure strength (ignoring the specific structures) and reflects (among other considerations) the structure of mRNA during translation, which prevents long-range structures but allows formation of local secondary structure and generally agrees with existing large-scale experimental validation results [37]. Previous studies (e.g., [35]) showed that this measure is robust to changes in the window size. The coordinates shown always refer to the window start position relative to the CDS start (e.g., window 0 includes the first 40 nt in the CDS) or to the window end position relative to the CDS end. Estimated folding energies were calculated for each window using *RNAfold* from the *ViennaRNA* package 2.3.0 [70], with the default settings. All folding energies were estimated at 37 °C so as to compare equivalent quantities between all genomes (but see below under native-temperature profiles). The ΔLFE profile for each protein, defined as the estimated excess local folding energy caused by the arrangement of synonymous codons at any CDS position, was created by subtracting the average profile of 20 randomized sequences for that protein from the native LFE profile:

$$\Delta\mathrm{LFE}_i = \mathrm{nativeLFE}_i - \frac{1}{N}\sum_{1\leq n\leq N}\mathrm{randomizedLFE}_i(n)$$

(i—CDS position, N—number of randomized sequences)

The mean ΔLFE profile for each species was created by averaging each position i over all proteins of sufficient length (so a different number of sequences may be averaged at each position). Note that while the native LFE of different CDSs within each genome varies considerably, the LFE of each native CDS is compared to *its own* set of randomized sequences.

To determine if the mean ΔLFE for a species in position i (relative to CDS start or end) is significantly different than 0, the differences $d_i(p, n)$ between LFE of the native and randomized sequences for each CDS p at position i were collected:

$$d_i(p,n) = \mathrm{nativeLFE}_i(p) - \mathrm{randomizedLFE}_i(p,n)$$

(p—CDS index, $1 \leq n \leq N = 20$—number of randomized sequences, i—CDS position)

The Wilcoxon signed-rank test was used on all values $d_i(p, n)$ (with the null hypothesis implying that the distribution is symmetrical).

Native-temperature profiles

The predicted folding energy calculations for native and randomized sequences for a sample of $N = 71$ bacterial and archaeal species were repeated using the same procedure but with folding predicted at the optimal growth temperature specified for that species (instead of 37 °C).

Phylogenetic tree preparation

To study the relation between ΔLFE profiles and other traits, the profiles were analyzed using a phylogenetic tree as follows. The phylogenetic tree is based on [71] (Supplementary Dataset 2 and Supplementary Table 1) and contains species from our dataset across the three domains of life. Since there are slight discrepancies in some node identifiers between the tree ([71] Supplementary Dataset 2) and accession table ([71] Supplementary Table 1), species names were matched by hand. Tree nodes and profiles were then matched by NCBI tax-ID at the species or lower level between the available genomes and phylogenetic tree nodes (e.g., when the tree species a species, and there is only one genome available for a specific strain of this species). The tree distances were converted to approximate relative ultrametric distances using *PATHd8* [72] version 1.9.8 with the default settings. Finally, the tree was pruned to the set of leaf nodes found in the dataset (or a subset of them which has data for both variables being correlated), by removing unused inner and leaf nodes and merging single-child inner nodes by summing distances. The resulting ultrametric tree (Additional file 3) was used to create a covariance matrix using a Brownian process (to reflect the null hypothesis that a trait is not under selection), using the *ape* package [73] in *R*.

Phylogenetically controlled regression

To test for correlations between traits among species while controlling for the similarity expected to exist between related species even in the absence of selection on either trait, generalized least-squared (GLS) regression was performed [74, 75] with the *nlme* package [76] in *R* and using REML optimization. Each regression included the subset of species for which data for both correlated traits was available, and which were also included in the tree. Regression p values are based on the null hypothesis that the slope of the explanatory variable is 0 (i.e., that the variables are independent), and estimated using the t test. Coefficient of determination (R^2) values were calculated according to [75, 77]:

$$R^2 = 1 - \frac{\hat{u}' V^{-1}\hat{u}}{\left(Y-\overline{Y}e\right)' V^{-1}\left(Y-\overline{Y}e\right)}$$

$\hat{u}$—residuals, V—variance-covariance matrix, Y—observations, $\overline{Y}$—intercept of equivalent intercept-only model, and e—first column of design matrix.

For continuous traits, regression formulas included an intercept term. Discrete traits were represented by ordered or unordered factors, and the intercept term was omitted from the regression formula. For discrete traits, values of the explained variable (such as ΔLFE) were centered to have mean 0 (so regression is based on a null hypothesis that all levels have the same mean).

Regression robustness verification

To test the robustness of a correlation between traits at different CDS regions, the regression was repeated at all profile positions starting between 0 and 300 nt (relative to CDS start and end) and all contiguous subranges (using the mean ΔLFE value in each range) and reported only if consistent over the relevant range of positions (Additional file 1: Figure S17).

To test for specific trait correlations in individual taxons, the regression procedure was repeated for each taxonomic group (at any rank) containing at least 9 species (Additional file 1: Figure S10). For each taxonomic group, the value shown is the median R^2 value for positions within the relevant range. The significance p value threshold was determined by applying FDR correction according to the number of taxonomic groups (treating them as independent to get a "worst-case" result).

Model element definition rules

Elements of the ΔLFE profile model were formalized as follows to allow estimation of their prevalence (Fig. 1a). Significance for all rules is defined using the Wilcoxon signed-rank test (see above) having p value < 0.05 at all positions within the range specified.

Model 1 (positive ends)

A. Positive start: ΔLFE value at positions 0–10 nt relative to CDS start is positive and significant.
B. Transition peak: the position of the minimum ΔLFE value in the range 0–300 nt, i^*, is located in the range 20–80 nt relative to CDS start, and is significantly lower compared to all points in the ranges 0–10 nt and 100–200 nt relative to CDS start.
   To determine if the mean ΔLFE for a species in a given position i is significantly higher than the minimum (i^*), the differences $w_i(p, n)$ between ΔLFE at the peak position and ΔLFE at the tested position were collected:

$$w_i(p, n) = d_{i*}(p, n) – d_i(p, n)$$

   (p—CDS index, $N \leq 20$—number of randomized sequences, i—position in CDS relative to start)

The Wilcoxon signed-rank test was used on all values $w_i(p, n)$.

C. Negative mid: ΔLFE values at each position in the range 200–300 nt relative to CDS start and in the range 300–200 nt relative to CDS end are all negative and significant.
D. Positive end: ΔLFE value at positions 10–0 nt relative to CDS end is positive and significant.
E. Model structure: A + C + D

Model 2 (weak ends)

A. Weak start: ΔLFE value at position 0 nt relative to CDS start is significantly higher than at positions 200–300 nt.
B. Same as in model 1.
C. Same as in model 1.
D. Weak end: ΔLFE value at position 0 nt relative to CDS end is significantly higher than at positions 200–300 nt.
E. Model structure: A + C + D

Binary classifier for ΔLFE strength

To measure the performances of several criteria in predicting ΔLFE strength, the following simple model was used. ΔLFE values for all species were divided into weak and strong groups based on the standard deviation of the mean ΔLFE at positions 0–300 nt. Species with standard deviation < 0.14 were included in the "weak ΔLFE" group. The binary classification of each species is based on 4 species traits as inputs, using the following rule (optimized using grid search):

$$\begin{aligned}\text{PredictedWeakLFE} = {} & (\text{Endosymbiont} = \text{True}) \text{ or} \\ & (\text{Genomic GC} < 38\%) \text{ or} \\ & (\text{Genomic ENc}' > 56.5) \text{ or} \\ & (\text{Optimum temp} > 58^\circ\text{C})\end{aligned}$$

Maximal information coefficient

Maximal information coefficient (MIC, [56, 57]) is a statistical measure of general (not necessarily linear) dependence between two variables. Informally, it is a generalization of R^2 and also has values in the range 0.0–1.0, with high values indicating knowing the value of one variable allows inferring the value of the other. MIC was calculated using the *minerva* [78] package in *R*. p values were estimated using 10,000 random samples.

Correlogram plot

Correlogram plot (Additional file 1: Figure S2) was prepared using the *phylosignal* package in *R*.

Codon-bias metrics

Codon-bias metrics (CAI, CBI, Nc, Fop) were calculated for each genome using *codonW* [79] version 1.4.4. ENc′ [80] was calculated using *ENCprime* (github user jnovembre, commit 0ead568, October 2016) using the default settings. I_TE [43] was calculated using DAMBE7 [81], based on the included codon frequency tables for each species. DCBS was calculated according to [49].

Shine-Dalgarno binding strength

The Shine-Dalgarno (SD) strength for each gene was calculated according to, based on the minimal anti-SD hybridization energy found in the 20-nt region upstream of the start codon.

Visualization

Taxon characteristic profile chart

The mean ΔLFE profiles for CDS positions 0–300 nt relative to the CDS start and end within each taxon were summarized (Fig. 3a) by grouping species with similar profiles and plotting one profile representing each group.

The grouping was achieved by clustering the ΔLFE profiles (as vectors of length 31) using *K*-nearest neighbors agglomerative clustering with correlation distances, using *SciKit Learn* [82]. The profile plotted to represent each group is the centroid (mean) of each cluster. To allow easy viewing of the region of interest, only positions 0–150 nt are shown for each cluster. *K*, the number of clusters for each taxon, was chosen (separately for the start and end profiles) to be the smallest value for which the maximum distance of any profile to the centroid cluster mean (i.e., the profile shown) was smaller than 0.8 for the start-referenced profiles and 1.3 for the end-referenced profiles. The full ΔLFE profiles for all species appear in Additional file 1: Figure S7.

PCA display for ΔLFE profiles

To summarize ΔLFE profiles and show how different values related to different profile types, we used PCA to obtain a two-dimensional arrangement in which similar ΔLFE profiles are mapped to nearby positions (see, for example, Fig. 3b). Also shown are the amounts of variance explained by each of the first two principal components.

PCA for the ΔLFE profiles (treated as vectors of length 31) was performed using *SciKit Learn* [82]. Analysis was limited to the first 3 components, and only the first two components are displayed (Additional file 1: Figure S6a,b). To verify the robustness of the PCA results, they were repeated using 500 samples with replacement from the same PCA input vectors and of the same size, and the angles between the component were verified to be approximately equal (Additional file 1: Figure S6c). To reduce clutter, overlapping profiles are hidden and the relative density at each position is shown in the background as blue shading (estimated as bivariate KDE with bandwidth determined by Scott's rule using *seaborn* [83]) and also plotted on the axes.

Evolutionary and taxonomic trees were plotted using the *ETE toolkit* [84].

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-01971-y.

Additional file 1: Table S1. List of species. **Table S2**. Phyla representation. **Table S3**. Genomic and environmental properties. **Figure S1.** Correlations of traits with ΔLFE are not present in its individual components. **Figure S2.** The ΔLFE profile is more conserved than other genomic traits. **Figure S3.** Local CUB vs. Local ΔLFE. **Figure S4.** Comparison between ΔLFE calculated using CDS-wide and position-specific ("vertical") randomizations. **Figure S5.** ΔLFE is stronger in highly expressed genes and genes encoding for highly abundant proteins. **Figure S6.** Unsupervised discovery of profile regions. **Figure S7.** ΔLFE profiles for all species. **Figure S8.** Comparison between ΔLFE profiles in different domains. **Figure S9.** Autocorrelation between ΔLFE profile regions. **Figure S10.** Trait correlations in taxonomic subgroups. **Figure S11.** Correlation of ΔLFE with different genomic measures of CUB is consistent. **Figure S12.** ENc' correlates with ΔLFE magnitude, not shape. **Figure S13.** Genomic-GC and genomic-ENc' both predict ΔLFE. **Figure S14.** Endosymbionts have weaker ΔLFE. **Figure S15.** Range robustness for GLS regressions between ΔLFE and related traits. **Figure S16.** Additional controls for phenomenon related to translation initiation. **Figure S17**. Dependence of ΔLFE profiles on temperature.

Additional file 2. Species ΔLFE profiles and additional data used for GLS regression analysis.

Additional file 3. Processed ultrametric phylogenetic tree used for GLS regression analysis.

Additional file 4. Review history.

Acknowledgements

We thank Dr. Zohar Zafrir for helpful comments. We thank Prof. Uri Gophna for helpful discussions.

Peer review information

Yixin Yao was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional file 4.

Authors' contributions

MP and TT conceived and designed the study. MP and TT analyzed the data. TT supervised the overall study. MP and TT wrote the paper. All authors read and approved the final manuscript.

Funding

MP gratefully acknowledges the support of the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University.

Availability of data and materials

All data reused in this study is publicly available from the sources specified in the methods. The annotated genomes used are available from the source specified in Additional file 1: Table S1. The dataset used for analysis is included in Additional file 2. The processed tree used for GLS analysis is included in Additional file 3. Software versions are specified in the methods. Python and R source code used for analysis is available from github repository https://github.com/michaelpeeri/rnafold-public [85]. All source code is licensed under the GNU General Public License (GPL) v3.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests, but have filed provisional patents overlapping the content of this paper.

Received: 14 October 2019 Accepted: 22 February 2020
Published online: 09 March 2020

References

1. Trotta E. Selection on codon bias in yeast: a transcriptional hypothesis. Nucleic Acids Res. 2013;41(20):9382–95.
2. Zamft B, Bintu L, Ishibashi T, Bustamante C. Nascent RNA structure modulates the transcriptional dynamics of RNA polymerases. Proc Natl Acad Sci. 2012;109(23):8948–53.
3. Ray-Soni A, Bellecourt MJ, Landick R. Mechanisms of bacterial transcription termination: all good things must end. Annu Rev Biochem. 2016;85(1):319–47.
4. Ben-Yehezkel T, Atar S, Zur H, Diament A, Goz E, Marx T, et al. Rationally designed, heterologous S. cerevisiaetranscripts expose novel expression determinants. RNA Biol. 2015;12(9):972–84.
5. Kozak M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene. 2005;361:13–37.
6. Gilbert WV, Zhou K, Butler TK, Doudna JA. Cap-independent translation is required for starvation-induced differentiation in yeast. Science. 2007; 317(5842):1224–7.

7. Xia X, Holcik M. Strong eukaryotic IRESs have weak secondary structure. PLoS One. 2009;4(1):e4136.
8. Zid BM, Rogers AN, Katewa SD, Vargas MA, Kolipinski MC, Lu TA, et al. 4E-BP extends lifespan upon dietary restriction by enhancing mitochondrial activity in Drosophila. Cell. 2009;139(1):149–60.
9. Jagodnik J, Chiaruttini C, Guillier M. Stem-loop structures within mRNA coding sequences activate translation initiation and mediate control by small regulatory RNAs. Mol Cell. 2017;68(1):158–70 e3.
10. Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM. *In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature. 2014;505(7485):696–700.
11. Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A, et al. Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast. Proc Natl Acad Sci. 2013;110(30):E2792–801.
12. Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, et al. Genome-wide measurement of RNA secondary structure in yeast. Nature. 2010; 467(7311):103–7.
13. Bhattacharyya S, Jacobs WM, Adkar BV, Yan J, Zhang W, Shakhnovich EI. Accessibility of the Shine-Dalgarno sequence dictates N-terminal codon bias in *E. coli*. Mol Cell. 2018;70(5):894–905 e5.
14. Behloul N, Wei W, Baha S, Liu Z, Wen J, Meng J. Effects of mRNA secondary structure on the expression of HEV ORF2 proteins in Escherichia coli. Microb Cell Factories. 2017;16(1):200.
15. Wu B, Zhang H, Sun R, Peng S, Cooperman BS, Goldman YE, et al. Translocation kinetics and structural dynamics of ribosomes are modulated by the conformational plasticity of downstream pseudoknots. Nucleic Acids Res. 2018;46(18):9736–48.
16. Wen J-D, Lancaster L, Hodges C, Zeri A-C, Yoshimura SH, Noller HF, et al. Following translation by single ribosomes one codon at a time. Nature. 2008 Apr;452(7187):598–603.
17. Qu X, Wen J-D, Lancaster L, Noller HF, Bustamante C, Tinoco I. The ribosome uses two active mechanisms to unwind messenger RNA during translation. Nature. 2011;475(7354):118–21.
18. Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, Ziv-Ukelson M. Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol. 2011;12(11):R110.
19. Komar AA. A pause for thought along the co-translational folding pathway. Trends Biochem Sci. 2009;34(1):16–24.
20. Park C, Chen XS, Yang JR, Zhang JZ. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A. 2013;110(8):E678–86.
21. Zhang G, Hubalewska M, Ignatova Z. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nat Struct Mol Biol. 2009;16(3):274–80.
22. Zur H, Tuller T. Strong association between mRNA folding strength and protein abundance in S. cerevisiae. EMBO Rep. 2012;13(3):272–7.
23. Lenz G, Doron-Faigenboim A, Ron EZ, Tuller T, Gophna U. Sequence features of *E. coli* mRNAs affect their degradation. PLOS ONE. 2011;6(12): e28544.
24. Wan Y, Qu K, Ouyang Z, Kertesz M, Li J, Tibshirani R, et al. Genome-wide measurement of RNA folding energies. Mol Cell. 2012;48(2):169–81.
25. Zafrir Z, Zur H, Tuller T. Selection for reduced translation costs at the intronic 5′ end in fungi. DNA Res. 2016;23(4):377–94.
26. Mortimer SA, Kidwell MA, Doudna JA. Insights into RNA structure and function from genome-wide studies. Nat Rev Genet. 2014;15(7):469–79.
27. Mauger DM, Siegfried NA, Weeks KM. The genetic code as expressed through relationships between mRNA structure and protein function. FEBS Lett. 2013;587(8):1180–8.
28. Jacobs E, Mills JD, Janitz M. The role of RNA structure in posttranscriptional regulation of gene expression. J Genet Genomics. 2012;39(10):535–43.
29. Faure G, Ogurtsov AY, Shabalina SA, Koonin EV. Role of mRNA structure in the control of protein folding. Nucleic Acids Res. 2016;44(22):10898–911.
30. Itzkovitz S, Hodis E, Segal E. Overlapping codes within protein-coding sequences. Genome Res. 2010;20:1582–9. Available from: https://doi.org/10.1101/gr.105072.110.
31. Katz L, Burge CB. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res. 2003;13(9): 2042–51.
32. Shabalina SA, Ogurtsov AY, Spiridonov NA. A periodic pattern of mRNA secondary structure created by the genetic code. Nucleic Acids Res. 2006; 34(8):2428–37.
33. Xia X. DAMBE6: new tools for microbial genomics, phylogenetics, and molecular evolution. J Hered. 2017;108(4):431–7.
34. Xia X. Bioinformatics and the cell: modern computational approaches in genomics. Proteomics and Transcriptomics: Springer; 2018. p. 494.
35. Mao Y, Wang W, Cheng N, Li Q, Tao S. Universally increased mRNA stability downstream of the translation initiation site in eukaryotes and prokaryotes. Gene. 2013;517(2):230–5.
36. Tuller T, Zur H. Multiple roles of the coding sequence 5′ end in gene expression regulation. Nucleic Acids Res. 2015;43(1):13–28.
37. Del Campo C, Bartholomäus A, Fedyunin I, Ignatova Z. Secondary structure across the bacterial transcriptome reveals versatile roles in mRNA regulation and function. PLoS Genet. 2015;11(10):e1005613. https://doi.org/10.1371/journal.pgen.1005613.
38. Kozak M. Influence of mRNA secondary structure on binding and migration of 40S ribosomal subunits. Cell. 1980;19(1):79–90.
39. Osterman IA, Evfratov SA, Sergiev PV, Dontsova OA. Comparison of mRNA features affecting translation initiation and reinitiation. Nucleic Acids Res. 2013;41(1):474–86.
40. Gu W, Zhou T, Wilke CO. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comput Biol. 2010;6(2):e1000664. https://doi.org/10.1371/journal.pcbi.1000664.
41. Keller TE, Mis SD, Jia KE, Wilke CO. Reduced mRNA secondary-structure stability near the start codon indicates functional genes in prokaryotes. Genome Biol Evol. 2012;4(2):80–8.
42. Tuller T, Waldman YY, Kupiec M, Ruppin E. Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci U S A. 2010;107(8):3645–50.
43. Xia X. A major controversy in codon-anticodon adaptation resolved by a new codon usage index. Genetics. 2015;199(2):573–9.
44. Wei Y, Xia X. Unique Shine–Dalgarno sequences in cyanobacteria and chloroplasts reveal evolutionary differences in their translation initiation. Genome Biol Evol. 2019;11(11):3194–206.
45. Xia X. Optimizing phage translation initiation. OBM Genet. 2019;3(4):1–1.
46. Dunteman GH. Principal components analysis. Newbury Park: SAGE Publication, Inc; 1989. https://uk.sagepub.com/en-gb/mst/principal-components-analysis/book2504.
47. Bennetzen JL, Hall BD. Codon selection in yeast. J Biol Chem. 1982;257(6): 3026–31.
48. Grosjean H, Fiers W. Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene. 1982;18(3):199–209.
49. Sabi R, Tuller T. Modelling the efficiency of codon–tRNA interactions based on codon usage bias. DNA Res. 2014;21(5):511–26.
50. Wright F. The "effective number of codons" used in a gene. Gene. 1990; 87(1):23–9.
51. Rocha EPC. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res. 2004;14(11):2279–86.
52. Vieira-Silva S, Rocha EPC. The systemic imprint of growth and its uses in ecological (meta)genomics. PLoS Genet. 2010;6(1):e1000808.
53. Sharp PM, Li WH. The codon adaptation index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 1987;15(3):1281–95.
54. Hildebrand F, Meyer A, Eyre-Walker A. Evidence of selection upon genomic GC-content in bacteria. PLoS Genet. 2010;6(9):e1001107.
55. Lee KY, Wahl R, Barbu E. Contenu en bases puriques et pyrimidiques des acides désoxyribonucléiques des bactéries. Ann Inst Pasteur (Paris). 1956; 91(2):212-24.
56. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting novel associations in large data sets. Science. 2011; 334(6062):1518–24.
57. Shaham G, Tuller T. Most associations between transcript features and gene expression are monotonic. Mol BioSyst. 2014;10(6):1426–40.
58. Andersson SGE, Kurland CG. Reductive evolution of resident genomes. Trends Microbiol. 1998;6(7):263–8.
59. Woolfit M. Effective population size and the rate and pattern of nucleotide substitutions. Biol Lett. 2009;5(3):417–20.
60. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol. 2012;10(1):13–26.
61. Hickey DA, Singer GA. Genomic and proteomic adaptations to growth at high temperature. Genome Biol. 2004;5(10):117.

62. Hurst LD, Merchant AR. High guanine–cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc R Soc Lond B Biol Sci. 2001;268(1466):493–7.
63. Chemla Y, Peeri M, Heltberg ML, Eichler J, Jensen MH, Tuller T, et al. mRNA secondary structure stability regulates bacterial translation insulation and re-initiation. BioRxiv. 2020; biorxiv.org. Available from: https://doi.org/10.1101/2020.02.10.941153.
64. dos Reis M, Wernisch L. Estimating translational selection in eukaryotic genomes. Mol Biol Evol. 2009;26(2):451–61.
65. dos Reis M, Savva R, Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 2004;32(17):5036–44.
66. Kersey PJ, Allen JE, Allot A, Barba M, Boddu S, Bolt BJ, et al. Ensembl genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic Acids Res. 2018;46(D1):D802–8.
67. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2018;46(D1):D8-D13. https://doi.org/10.1093/nar/gkx1095.
68. Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, et al. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res. 2014;42(Database issue):D26–31.
69. Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, et al. The reference genome sequence of *Saccharomyces cerevisiae*: then and now. G3 GenesGenomesGenetics. 2013;4(3):389–98.
70. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6(1):26.
71. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat Microbiol. 2016;1:16048.
72. Britton T, Anderson CL, Jacquet D, Lundqvist S, Bremer K, Anderson F. Estimating divergence times in large phylogenetic trees. Syst Biol. 2007;56(5):741–52.
73. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20:289–90.
74. Aitken AC. IV.—On least squares and linear combination of observations. Proc R Soc Edinb. 1936;55:42–8.
75. Paradis E. Analysis of macroevolution with phylogenies. Anal Phylogenetics Evol R. 2012:203–312.
76. Pinheiro J, Bates D, DebRoy S, Sarkar D, Heisterkamp S, Van Willigen B. nlme: linear and nonlinear mixed effects models. R Package 3rd Edn. 2017;1–336.
77. Buse A. Goodness of fit in generalized least squares estimation. Am Stat. 1973;27(3):106–8.
78. Albanese D, Filosi M, Visintainer R, Riccadonna S, Jurman G, Furlanello C. Minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers. Bioinformatics. 2013;29(3):407-8. https://doi.org/10.1093/bioinformatics/bts707. Epub 2012 Dec 14.
79. Peden JF. Analysis of codon usage. PhD dissertation. Nottingham: University of Nottingham; 1999. Available from: http://codonw.sourceforge.net/.
80. Novembre JA. Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol. 2002;19(8):1390–4.
81. Xia X. DAMBE7: new and improved tools for data analysis in molecular biology and evolution. Mol Biol Evol. 2018;35(6):1550–2.
82. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
83. Waskom M. Seaborn: statistical data visualization, version 0.9.0. 2019. Available from: https://seaborn.pydata.org/ . Accessed 22 Apr 2019.
84. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol. 2016;33(6):1635–8.
85. Peeri M, Tuller T. High resolution modeling of the selection on local mRNA folding strength in coding sequences across the tree of life. Source code. 2020. Available from: github https://github.com/michaelpeeri/rnafold-public/. Accessed 25 Feb 2020.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC