PLOS ONE

RESEARCH ARTICLE

Nucleotide composition affects codon usage toward the 3′-end

Fouad Zahdeh[1,2], **Liran Carmel**[1]*

1 Department of Genetics, The Alexander Silberman Institute of Life Sciences, Faculty of Science, The Hebrew University of Jerusalem, Edmond J. Safra Campus, Givat Ram, Jerusalem, Israel, **2** Hereditary Research Lab, Life Sciences Department, Bethlehem University, Bethlehem, Palestinian authority

* liran.carmel@huji.ac.il

Citation: Zahdeh F, Carmel L (2019) Nucleotide composition affects codon usage toward the 3'-end. PLoS ONE 14(12): e0225633. https://doi.org/10.1371/journal.pone.0225633

Editor: Tamir Tuller, Tel Aviv University, ISRAEL

Received: February 7, 2019

Accepted: November 9, 2019

Published: December 4, 2019

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: LC was supported by the Israel Science Foundation (grant No. 1431/13). URL: https://www.isf.org.il. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

The 3'-end of the coding sequence in several species is known to show specific codon usage bias. Several factors have been suggested to underlie this phenomenon, including selection against translation efficiency, selection for translation accuracy, and selection against RNA folding. All are supported by some evidence, but there is no general agreement as to which factors are the main determinants. Nor is it known how universal this phenomenon is, and whether the same factors explain it in different species. To answer these questions, we developed a measure that quantifies the codon usage bias at the gene end, and used it to compute this bias for 91 species that span the three domains of life. In addition, we characterized the codons in each species by features that allow discrimination between the different factors. Combining all these data, we were able to show that there is a universal trend to favor AT-rich codons toward the gene end. Moreover, we suggest that this trend is explained by avoidance from forming RNA secondary structures around the stop codon, which may interfere with normal translation termination.

Introduction

Codon usage bias (CUB) is the name coined for the well-known observation that synonymous codons are used at different frequencies in a genome. The codons that are more frequently used are denoted *optimal codons*. Different organisms may show different CUB, in the sense that they have different optimal codons. It has been observed that optimal codons tend to be over-represented in highly expressed genes [1, 2] and to correspond to high copy number of cognate tRNAs [3–6]. Over the years, there have been many attempts to identify the factors that are at the base of CUB. Of them, two leading explanations suggest selective constraints. The first is termed *translation efficiency* and it asserts that codons with higher density of cognate tRNAs will be translated faster, thus a coding sequence (CDS) made predominantly of optimal codons would be translated more efficiently [1, 7, 8]. The second is termed *translation accuracy* [9–11] and it asserts that less frequent codons are more prone to translational errors due to increased competition from more abundant, near-cognate tRNAs [3, 12, 13]. Both types of selection may be in action in each particular species [14, 15], but their relative contribution

is typically unknown [9]. In some eukaryotes, such as human and drosophila, it had been claimed that the effect of translation efficiency is weak [2, 6, 7, 16, 17]. Similar claim has been made also in some prokaryotes such as *H. pylori*, where translation efficiency seems to have no significant contribution to CUB [18]. In addition to these presumed selective forces, many other factors have been suggested to effect CUB, including biased gene conversion [19], effective population size and evolutionary history [10, 20–22], genome size [7], mutation-selection balance [6, 10, 16], Hill-Robertson effect [23–27], and mRNA secondary structure [28–34].

In addition to the genome-wide CUB, it has been noticed that codon usage shows spatial patterns along the CDS. While CUB at the CDS start (5'-end) has been extensively studied [25, 29, 31–35], less is known about its behavior towards the CDS end. Tuller *et al.* [35] reported that codons at the 3'-end are translationally inefficient (hereinafter, simply inefficient) in many species, especially in eukaryotes. This finding supports previous work by Eyre-walker who found a decrease in the frequency of efficient codons along the last 20 codons of *E. coli* genes [36]. However, this observation does not seem to universally hold. For example, Qin *et al.* measured 3'-end CUB in four prokaryotes and two eukaryotes, and could not find a consistent trend [25].

Translation accuracy, and more specifically selection against nonsense errors, is thought to be a key factor affecting the spatial distribution of CUB [25, 37]. It was suggested that the cost of nonsense errors increases when translation progresses, until it peaks near the 3'-end [29, 38, 39]. Accordingly, it is expected that selection to minimize nonsense errors is stronger near the stop codon [8, 25, 37]. Qin *et al.* [25] tested this by studying spatial CUB along the entire CDS, and concluded that selection against nonsense errors dominates in prokaryotes and yeast. Another version of selection against nonsense errors in eukaryotes was reported by Cusack *et al.* [40]. They claimed that the selection regime along the CDS in eukaryotes is different than in prokaryotes, as the eukaryotic mRNA surveillance mechanism known as nonsense-mediated mRNA decay (NMD) targets for degradation transcripts that harbor premature stop codons. As the main trigger to NMD is the presence of an exon-exon junction downstream to the stop codon [41], all nonsense errors except those in the last exon, do not yield functional transcripts and are therefore not highly deleterious. Cusack *et al.* studied a group of codons that are one substitution away from stop codons, called fragile codons, which should show this unique selection regime more strongly. As expected, fragile codons were shown to be depleted in the last exon of multi-exon genes in human, but not in other exons. It is not discussed in Cusack *et al.*'s paper, but even along the last exon the selection regime on fragile codons is not expected to be even. The reason is that the deleterious effect of nonsense errors is expected to decrease towards the 3'-end, as slightly truncated proteins usually retain most of their functionality.

Another factor that was suggested to be important in shaping spatial CUB is selection against RNA folding. Rocha *et al.* found that there is a tendency toward GC depletion at both the 5' and the 3' gene termini in *B. subtilis* [31]. They suggested that the bias at the 3'-end decreases the propensity to form stable RNA secondary structures, and hence lowers the probability of interference with normal translation termination and the recruitment of release factors. A similar observation was made in an earlier study by Erye-Walker [36], who reported an increase in A-ending codons and a decrease in G-ending codons toward the 3'-end of *E. coli* genes. Because many genes in *E. coli* overlap the CDS or the Shine-Dalgarno sequence of another gene in the opposite strand, Erye-Walker explained this bias by the need to avoid RNA secondary structures near the transcription start site of the opposing gene. However, further studies showed that the bias likely relates to the gene end, rather than to the start of the opposing gene. Katz *et al.* [42] examined the potential to form secondary structures at the 5'

and the 3' termini of genes in *E.coli* and yeast. In *E. coli*, the propensity to form RNA secondary structures was evenly distributed across the CDS, and specifically was about equal at both gene termini. In yeast, however, the propensity to form RNA secondary structures was found to be lower at the 3'-end than at the 5'-end. However, the universality of this factor was questioned as well. Qin *et al*. studied the effect of various factors on CUB near the gene end in yeast and fruit fly, and showed that the spatial variation in CUB is inconsistent with the GC-content variation across the gene [25].

As of today, it is still unclear which of the above models, or a combination thereof, best explains the 3'-end CUB patterns. All models are partly supported, and it may be that different models better describe the CUB in different species. Another difficulty is that the different models have very similar predictions, because of strong dependencies between the relevant codon features. For example, AT-rich codons also tend to be inefficient [34] and fragile [40]. Such dependencies have been accounted for in some works on CUB. For instance, based on the fact that the transcription initiation site in many species experiences a selection against mRNA folding [32, 33], and because inefficient codons were found to be AT-rich, Bentele *et al*. proposed that the reduction in translation efficiency at the gene 5'-end is a side effect of the selection against mRNA folding [34]. However, to the best of our knowledge, no work has compared the different models relevant to the CDS end in a way that accounts for these dependencies. This is the task we wish to accomplish in this work.

Another difficulty in direct comparison of the different models stems from the lack of uniformity in measuring CUB levels. The *codon adaptation index* (CAI) is one of the most popular measures of CUB, that attempts to directly account for the observation that highly expressed genes tend to use more efficient codons [43]. A similar measure is called *tRNA adaptation index* (tAI), and it attempts to normalize codon usage to the background tRNA pool [44]. Other measures, such as the *effective number of codons* (ENC) [45] or the *relative synonymous codon usage* (RSCU) [46], estimate the deviation of codon usage from their expected usage under the null uniform distribution.

A major obstacle in studying spatial CUB is that none of the existing measures is position-dependent, thus all previous works measured an overall, rather than spatial, CUB, even when focusing on a particular region of the CDS. To circumvent this, Qin *et al*. aligned subsets of genes to create super-sequences, and then computed the *ENC* index for each position [25]. Tuller *et al*. used a similar approach, which they termed *local tAI* [35]. In local tAI, genes were aligned either from their 3'-end or from their 5'-end, and the average tAI index was computed at each position. Hockenberry *et al*. [47] studied codon usage bias in *E. coli* using a position-dependent model. They partitioned the sequence into bins of equal number of codons and defined the positional dependency (pD) of each bin as the χ^2 statistics based on the observed frequency of individual codons at the specific bin and their expected frequency derived from a random synonymous shuffling algorithm. Such approaches assume that all positions have the same relative weight in the computation of the index, so it is technically a local measure of the otherwise global index at different gene regions, and not a full fledge position-dependent index.

Here, we develop a position-dependent codon usage index that we dub *relative spatial codon abundance* (RSCA). We used RSCA to characterize the spatial CUB near the 3'-end of 91 species, representing all three domains of life. For each species we tested how much of the 3'-end CUB is explained by each of the three main models: selection against translation efficiency, selection for translation accuracy, and selection against mRNA folding. In all species, we found a strong support to the notion that 3'-end CUB stems from avoidance of RNA secondary structures, suggesting that this is a universal selective force.

Results

Measuring spatial codon usage bias towards the gene end

We developed a measure, called relative spatial codon abundance (RSCA) that measures how over- or under-represented a codon is at each position along the gene. The RSCA is represented by a matrix R^i_α, which computes the relative spatial abundance of codon i at position α, measured as the codon number from the gene end (so that $\alpha = 1$ corresponds to the last codon before the stop codon, $\alpha = 2$ corresponds to the codon just before it, and so on.) $R^i_\alpha = 1$ means that the frequency of codon i at position α is as expected from the overall relative abundance of this codon in the genome. $R^i_\alpha > 1$ means that the codon is over-represented at this position, whereas $R^i_\alpha < 1$ means that the codon is under-represented at that position (see Methods for a complete definition).

We computed RSCA for 91 species representing all three domains of life (S1 Table). Importantly, the list includes three species, *M. pneumonia*, *M. florum and U. parvum*, that use a modified genetic code, allowing to test various connections between the structure of the genetic code and the abundance of codons near the gene end. We noticed that the RSCA in the last position before the stop codon ($\alpha = 1$) usually deviates from its values in the preceding positions. This likely reflects special and strong constraints on the penultimate codon to increase termination efficiency [48]. Consequently, we ignored this position from further analyses, and instead focused on more gradual changes to RSCA values, occurring over at least a few codons.

Multiple strategies for grouping codons

To evaluate the contribution of each of the three models on the RSCA at the CDS 3'-end, we characterized each codon by a series of informative features. Features informative for selection against mRNA folding are the codon GC content and the content of its third nucleotide (wobble position). A feature informative for selection against translation efficiency is the cognate tRNA copy number. Finally, a feature informative for selection for translation accuracy is the codon fragility. Therefore, we used four different classification schemes, dividing all sense codons (excluding those that have no synonymous codons) into mutually exclusive groups based on the value of the relevant features:

1. Wobble position. We divided the sense codons into four groups, based on the identity of their third nucleotide: *A-ending* codons, *T-ending* codons, *C-ending* codons, and *G-ending* codons.
2. GC content. We grouped the codons based on their total GC content, from 0 (the codon harbors neither C nor G) to 3 (the codon harbors only Cs and Gs).
3. Fragility. As explained in the Introduction, fragile codons are codons that are one substitution away from a stop codon. We have further divided fragile codons into *replaceable* codons, which code for an amino acid that is coded by at least one non-fragile codon, and *non-replaceable* codons, which code for an amino acid that is exclusively coded by fragile codons. The reason for this distinction is that if there is a selection against fragile codons, it is much easier to avoid the usage of replaceable codons, but far harder for non-replaceable ones. Overall, this classification scheme divides all sense codons into three groups: non-fragile, replaceable, and non-replaceable.
4. Efficiency. Codons were binned to three groups based on their relative adaptiveness value, which is an estimate of the abundance of the corresponding tRNAs [7]. Higher value of the relative adaptiveness indicates higher abundance of tRNAs that recognize the codon, and

suggests that the codon is more translationally efficient. The three groups are denoted *inefficient* codons, *moderately efficient* codons, and *efficient* codons.

A major obstacle in judging which feature is most explanatory of 3'-end spatial CUB patterns is that the different classification schemes are not independent: Efficiency significantly depends on fragility in 18 species, on wobble position in 51 species, and on GC content in eight species. ($P \leq 0.05$, FDR-corrected χ^2 independence test; S2 Table); fragility significantly depends on wobble position in the standard genetic code ($P = 0.03$) but not in the Mycoplasma genetic code ($P = 0.11$). These results show that different codon features may be strongly correlated, and that these correlations may be species-specific. These connections must be taken into account, and may lay behind part of the disagreements between previous studies.

Measuring RSCA towards the gene end for codon groups

To compute the spatial patterns displayed collectively by a group of codons *S*, we developed a group RSCA measure, denoted R_α^S (see Methods). Like before, $R_\alpha^S = 1$ means that the frequency of the codons in *S* at position α is as expected from their overall relative abundance, $R_\alpha^S > 1$ means that they are over-represented at this position, and $R_\alpha^S < 1$ means that they are under-represented at that position (see Methods).

We tested this measure by applying it to permuted datasets, in which we kept the overall codon usage bias, but erased positional variance. To this end, we picked four genomes (*A. pernix*, *E. coli*, *H. sapiens*, and *M. pneumonia*) with a representative from each of the three domains of life, as well as a representative (*M. pneumonia*) for non-standard genetic code. In each permutation, we replaced each codon by a synonymous codon according to its genomic frequency. For example, Alanine is coded by 'GCA', 'GCC', 'GCG', and 'GCT' codons. Their frequencies in *E. coli* are 0.22, 0.27, 0.34, and 0.17, respectively. We scanned the *E. coli* genome and replaced each Alanine codon by a synonymous codon (including the same codon) with a probability that equals the synonymous codon genomic frequency. At the end of this process, each of these codons appears in roughly the same frequency as in the original genome, but positional bias is erased. We performed this process 100 times for each of the four genomes, and obtained, as expected, group RSCA scores close to 1 along the last 50 codons of the gene in all tested genomes (S1–S4 Figs; S3 Table).

We computed the group RSCA score for each group of codons in each of the classification schemes, and for each of the species, along the last 50 codons of genes. To measure how significantly R_α^S increases or decreases toward the 3'-end, we fit it to a linear model. As the exact position where RSCA starts to deviate from one changes across species and domains (details below), the linear model was evaluated separately at the last 10, 20, and 30 codons (three regions). For each model we computed the adjusted R^2, slope β, and FDR-adjusted p-value *P*. $P<0.05$ and $\beta>0$ at any region implies significant increase in the representation of the codon towards the 3'-end. $P<0.05$ and $\beta<0$ at any region implies significant decrease in the representation of the codon towards the 3'-end (S4 Table). To summaries the linear regression model across multiple species, we scored a codon group in a species by 1 if it shows significant increase, by -1 if it shows significant decrease, and by 0 if the linear model is insignificant in all three regions (S5 Table). This scoring system cancels out the contribution of species which have opposing trends for the same codon group. Finally, each codon group was assigned with a value *f*, which is the sum of scores across all species, divided by the number of species (Fig 1). Codons that are AT-rich, A-ending, and replaceable are the top over-represented groups, while C-ending and GC-rich are the top under-represented codon groups. For further analysis, see below.

A-rich codons are preferred towards the gene end

When classifying codons by the content of their wobble position, we observed that A-ending codons are preferentially used towards the gene end in a striking majority of species (Fig 2; S4 and S5 Tables). The few exceptions include the eukaryote *C. elegans*, and the archaeon *N. martimus*, both showing significant under-representation of A-ending codons. The yeast *S. cerevisiae* and a few prokaryotes show neither under-representation nor over-representation of A-ending codons near the gene end (see Discussion). The extent of this spatial pattern is shared across domains, with A-ending codons showing preferential usage along the last 10–15 codons. Interestingly, only in eukaryotes A-ending codons tend to be highly under-represented also along the preceding region (positions 15/20-50). T-ending codons do not show a consistent trend, although they are markedly under-represented in eukaryotes along most of the last 50 codons of the gene, with the exception of positions 3–5. The spatial patterns of C-ending codons are almost the exact opposite of those for T-ending codons, but the under-representation in prokaryotes is much more pronounced than in eukaryotes. The spatial patterns of G-ending codons show significant contrast between domains. While they seem to be over-represented in positions 3–6 in some prokaryotes, they tend to be under-represented along the last 15–30 codons in eukaryotes.

In summary, spatial patterns of codons ending with a particular nucleotide show different trends between domains, except for A-ending codons that are over-represented in the vast majority of studied species from all the domains.

Similar results are obtained if we group codons based on their GC content (Fig 3; S4 and S5 Tables). Almost all examined species show strong preference to A/T-rich codons at their gene termini, along the last 5–10 codons of the gene. These observations are in agreement with previous works [31, 36, 42], that attributed this trend to a pressure to avoid mRNA folding at the vicinity of the stop codon in order to guarantee efficient translation termination.

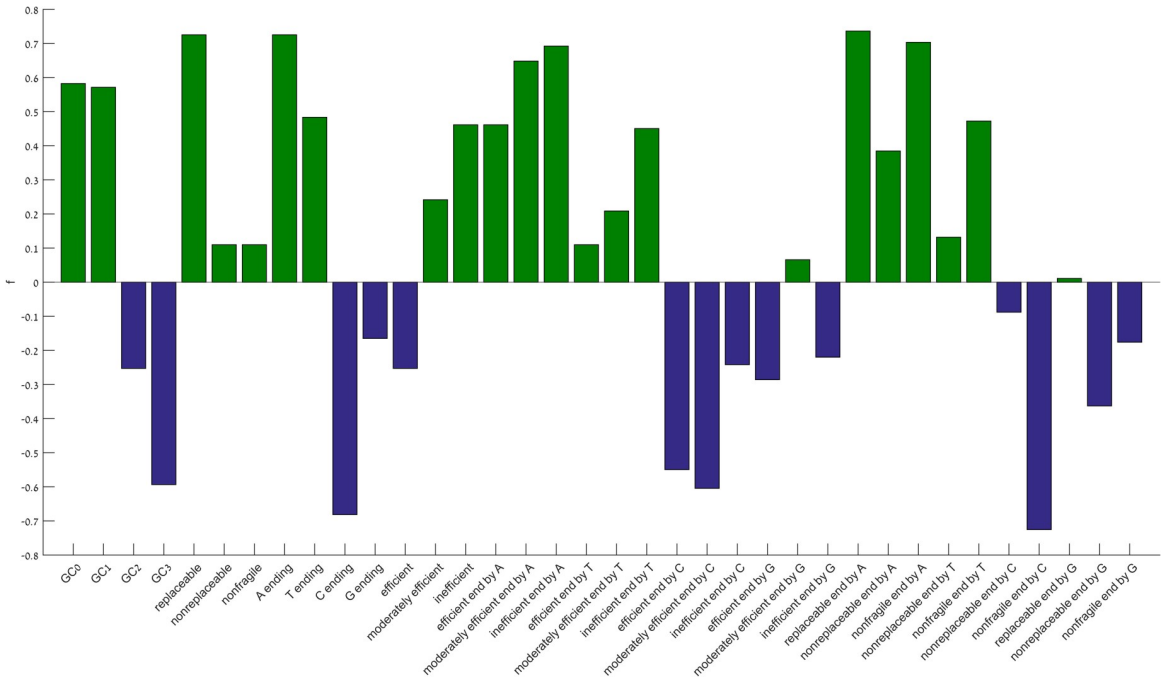

Fig 1. Preference of codon features towards the gene end. For each codon group we calculated the sum of scores (S5 Table) divided by the total number of species, *f*. Green bars mark features in which group RSCA significantly increase, blue bars mark features that are significantly decreased toward 3'-end.

https://doi.org/10.1371/journal.pone.0225633.g001

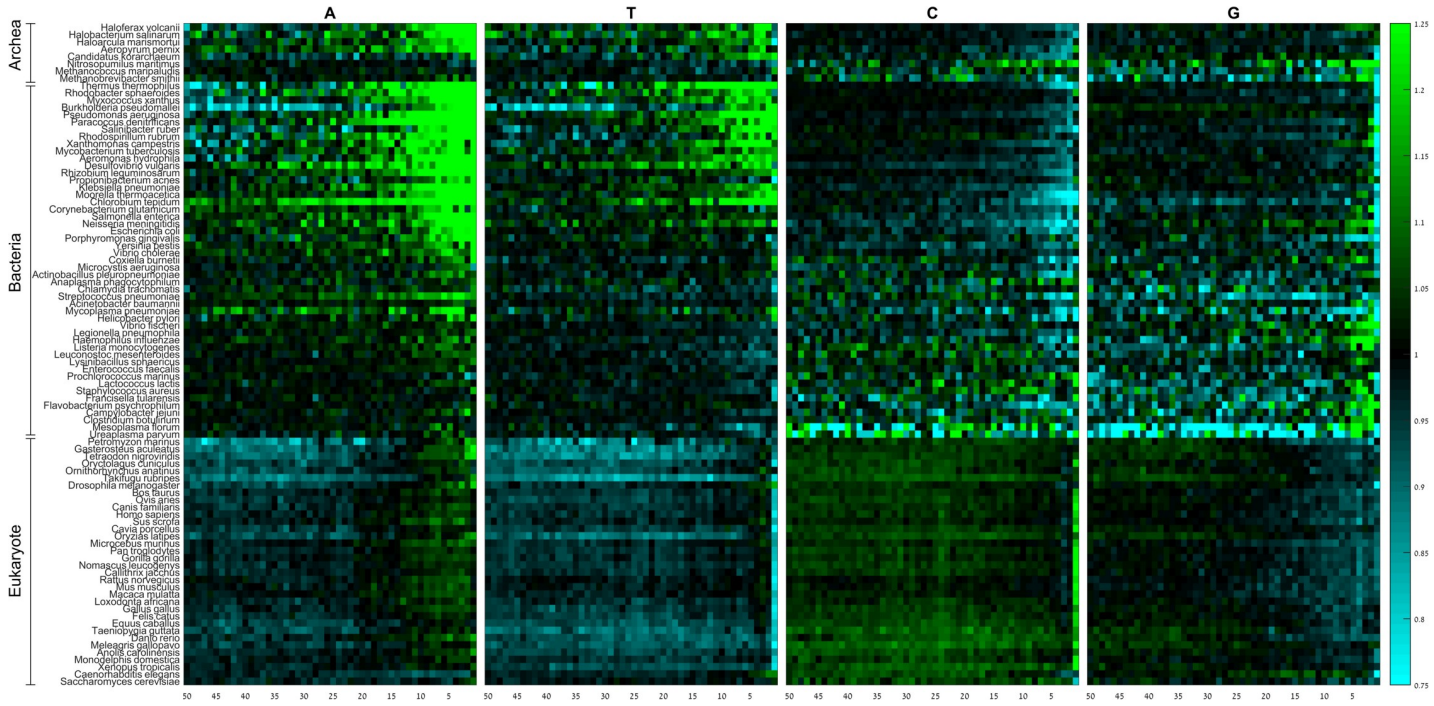

Fig 2. Group RSCA scores of A-ending, T-ending, C-ending, and G-ending codons along the last 50 codons of the gene. Rows denote species, columns denote positions. Species within domains are sorted from high (top) to low (bottom) GC-content.

https://doi.org/10.1371/journal.pone.0225633.g002

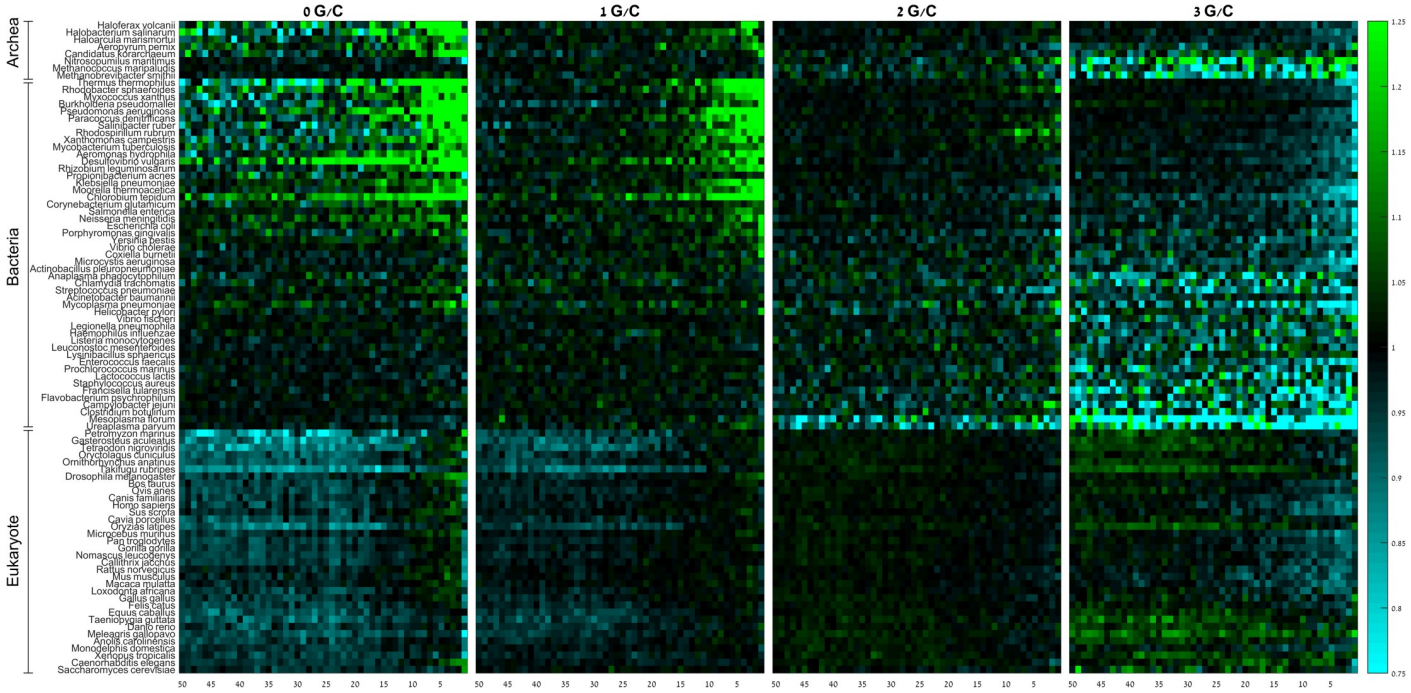

Fig 3. Group RSCA scores of codons containing 0, 1, 2, and 3 G/C nucleotides along the last 50 codons of the gene. Rows denote species, columns denote positions. Species within domains are sorted from high (top) to low (bottom) GC-content.

https://doi.org/10.1371/journal.pone.0225633.g003

We repeated the analyses on triplets that were formed by introducing +1nt and +2nt frameshifts to the coding region. By introducing a +1nt frameshift, namely looking at the content of the first codon position in the original reading frame, we found that RSCA patterns of A, T, C, and G-ending triplets are generally similar to the patterns we observed for the original reading frame: A-ending triplets are preferred while G-ending triplets are avoided toward the 3'-end (S5 Fig). Interestingly, species that deviate from the general preference to A-ending codons, such as *C. elegans* and *N. maritimus* (see Figs 2 and 3), have similar patterns to the rest of species in the +1nt triplets, especially at the last 5–10 codons. However, after introducing +2nt frameshift (i.e., looking at the content of the second codon position in the original reading frame) triplets grouped by wobble position show completely different patterns (S6 Fig). C-ending triplets are under-represented in prokaryotes, especially at the last 15 triplets, and A-ending triplets are slightly over-represented along the last 10–20 codons in many species. G-ending and T-ending triplets show mixed patterns.

A similar pattern is observed when grouping codons by GC richness. For +1nt frameshift, the patterns are similar to the original reading frame, where GC-rich triplets (3/3 codons) are avoided and AT-rich triplets (0/3 codons) are preferred toward the 3'-end (S7 Fig). This pattern is similar, but less pronounced for +2nt frameshift (S8 Fig).

Taken together, these results suggest that there is a strong selection against G/C and a preference to A at the first and third codon positions, but not at the second codon position. Given that any mutation in the second codon position is nonsynonymous, these findings are compatible with the notion that these trends in CUB towards the gene end are independent of selection forces at the amino acid level. This also suggests that the trends we observe generally characterize nucleotide composition in codons, and are not specific to particular position along the codon.

Efficiency is not a major factor determining codon usage bias near the gene end

Next, we wanted to check the effect of codon efficiency on the spatial codon usage bias near the stop codon by computing the group RSCA score of codons based on their adaptiveness values (Fig 4). In the majority of the examined prokaryotes, and in particular in those with high to moderate GC-content, inefficient and moderately efficient codons are preferentially used towards the gene end, consistent with previous reports [35, 36]. These codons tend to be under-represented along the entire region in eukaryotes except for the last 5 codons (Fig 4). Efficient codons generally show the opposite pattern.

As codon efficiency significantly depends on wobble position in more than half of the species (S2 Table), and because inefficient codons are known to be AT rich [34], we wished to test whether the spatial patterns of inefficient codons prevail after we control for the content of their wobble position. To this end, we subdivided the inefficient, moderately efficient, and efficient codon groups into codons that end by A (Fig 5; S4 and S5 Tables), T, C, and G (S9–S11 Figs; S4 and S5 Tables). This analysis shows that the spatial patterns towards the gene ends are predominantly dictated by the wobble position rather than by codon efficiency. The patterns of codons with the same efficiency depend on whether they end by A, T, G, or C, and codons ending with the same nucleotide show similar patterns regardless of their efficiency. To further confirm this we used the (FDR-corrected) Kruskal-Wallis test in the last 15 positions to check whether the distribution of the RSCA values of codons ending by A/T and of codons ending by C/G are affected by the codon efficiency (S6 and S7 Tables). The results show that RSCA values do not depend on the efficiency of the codon, as long as we control for the content of its wobble position. We conclude that selection against translation efficiency plays no significant role in shaping the codon usage bias at the gene 3'-end.

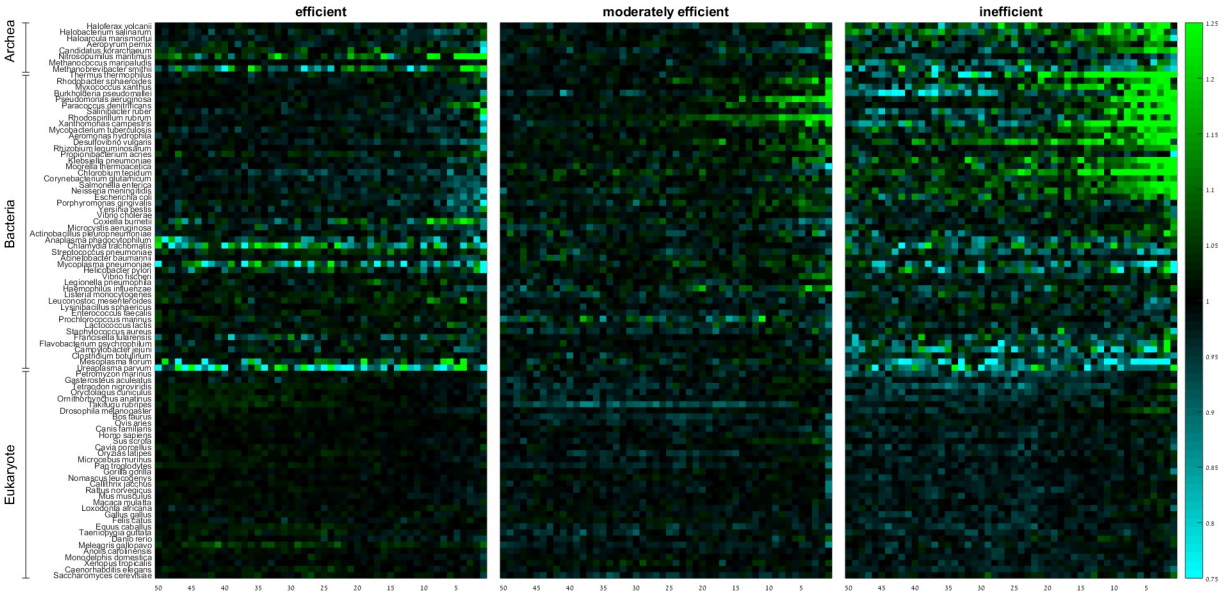

Fig 4. Group RSCA scores of efficient, moderately efficient, and inefficient codons along the last 50 codons of the gene. Rows denote species, columns denote positions. Species within domains are sorted from high (top) to low (bottom) GC-content.

https://doi.org/10.1371/journal.pone.0225633.g004

Fragility is not a major factor determining codon usage bias near the gene end

We computed group RSCA scores for codon fragility (Fig 6; S4 and S5 Tables). The results show a strong preference to replaceable codons along the last 5–10 codons, especially in

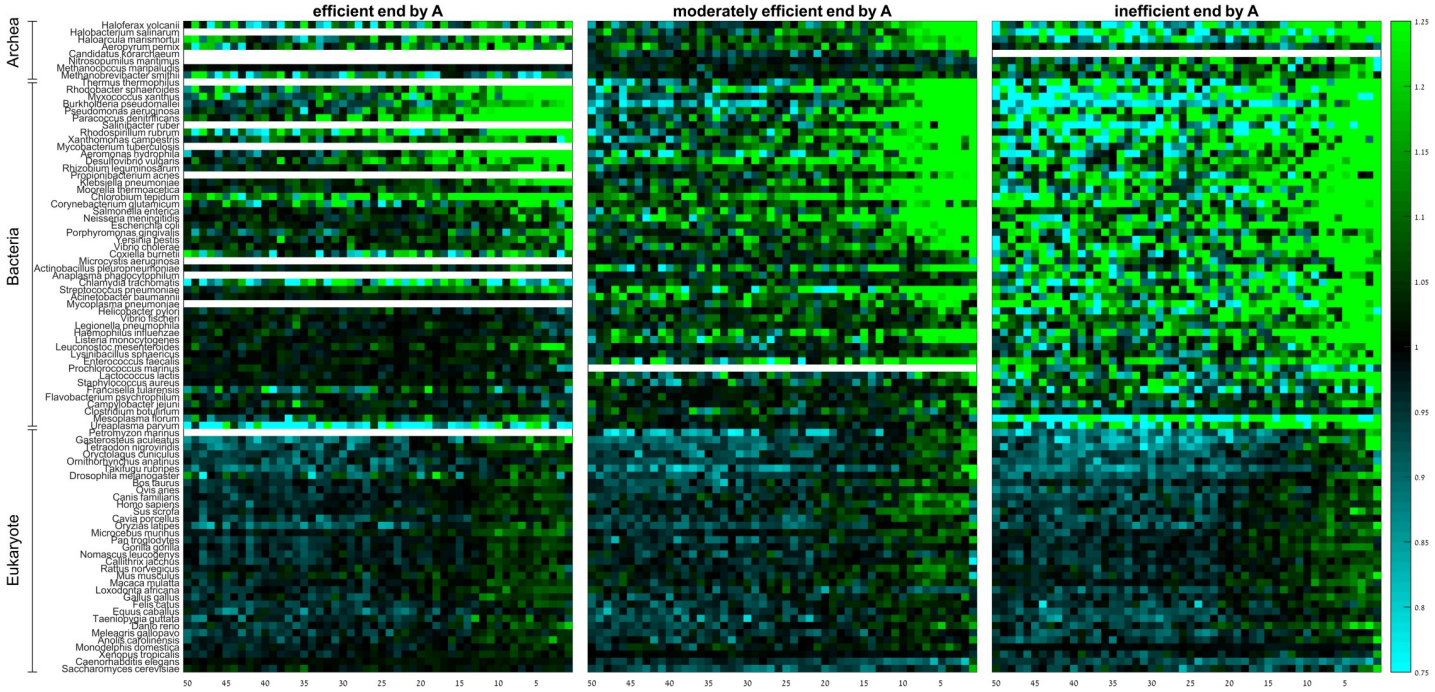

Fig 5. Group RSCA scores of efficient, moderately efficient, and inefficient codons that end by A along the last 50 codons of the gene. Rows denote species, columns denote positions. Species where none of the codons in a particular efficiency group end by A (missing data) are shown as white stripes. Species within domains are sorted from high (top) to low (bottom) GC-content.

https://doi.org/10.1371/journal.pone.0225633.g005

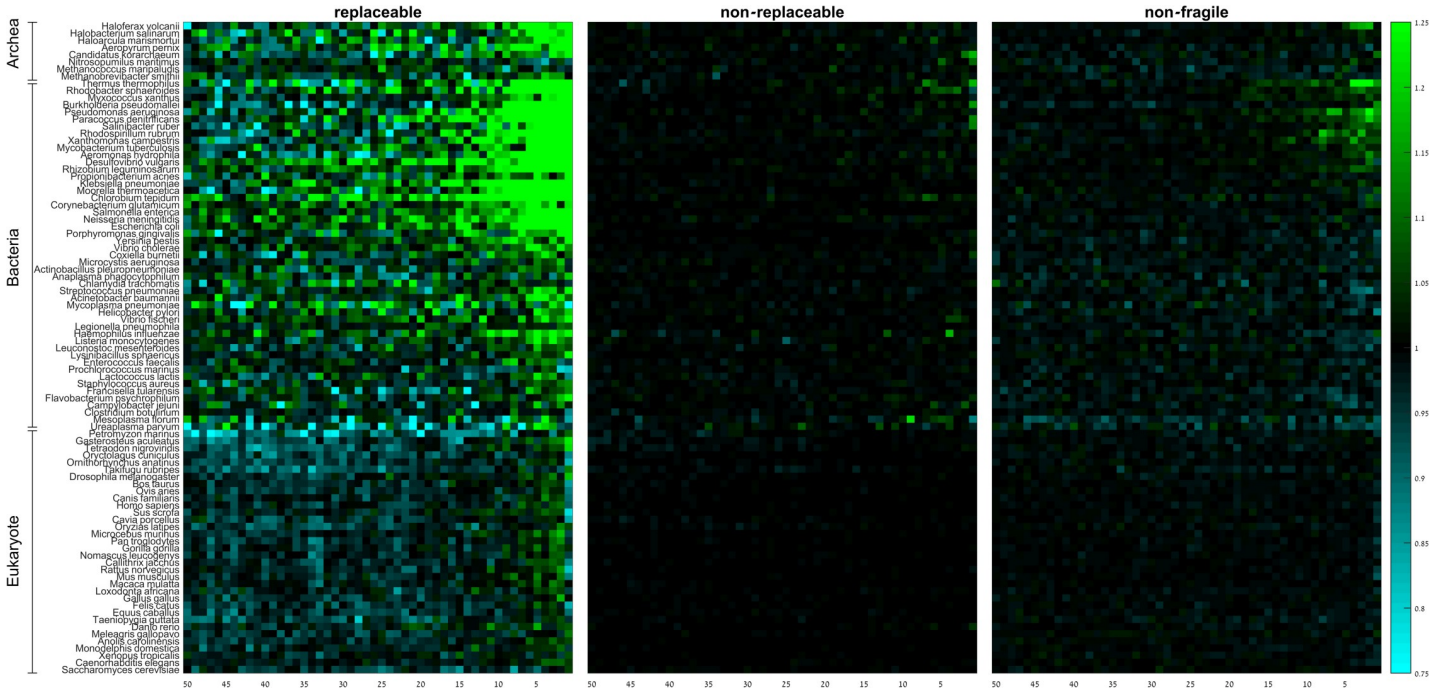

Fig 6. Group RSCA scores of replaceable, non-replaceable, and non-fragile codons along the last 50 codons of the gene. Rows denote species, columns denote positions. Species within domains are sorted from high (top) to low (bottom) GC-content.

https://doi.org/10.1371/journal.pone.0225633.g006

prokaryotes. Non-replaceable codons show negligible spatial preferences. Non-fragile codons exhibit a weak tendency to be over-represented along the last 10–20 codons in high GC-content prokaryotes but a weak tendency to be under-represented along the last 10–20 codons in prokaryotes with low to moderate GC-content.

Replaceable codons are enriched for codons that end by A (71% of the replaceable codons, but only 27% of the non-replaceable and 14% of the non-fragile). Moreover, fragility strongly depends on wobble position in species that utilize the standard genetic code ($P = 0.03$; χ^2 independence test). Therefore, we wanted to test whether their spatial patterns are a simple reflection of the content of their wobble position. Similar to what we did for codon efficiency, we subdivided the replaceable, non-replaceable and non-fragile codons according to the content of their wobble position into A- (Fig 7; S4 and S5 Tables), T-, C-, and G-ending codons (S12–S14 Figs; S4 and S5 Tables), and tested whether fragility affects the distribution of the RSCA values (S8 and S9 Tables). The results demonstrate that replaceable, non-replaceable, and non-fragile codons ending by A behave similarly irrespective to their fragility, and that the distribution of RSCA values is independent on the fragility of the codon. Exceptions include some prokaryotes, especially AT-rich ones, that show weak under-representation of non-replaceable ending by A at the last 5–10 codons. Hence we conclude that it is the content of the wobble position of the codon, and not its fragility, that determines the spatial CUB pattern near the gene end.

A-ending codons are associated with less stable mRNA folding at the gene 3'-end

Next, we wanted to test if the universal over-representation of A-ending codons at the 3'-end lower secondary RNA structures stability compared to what is expected from the general codon bias. To this end, we used RNAfold to compute minimum free energy (MFE) at the last

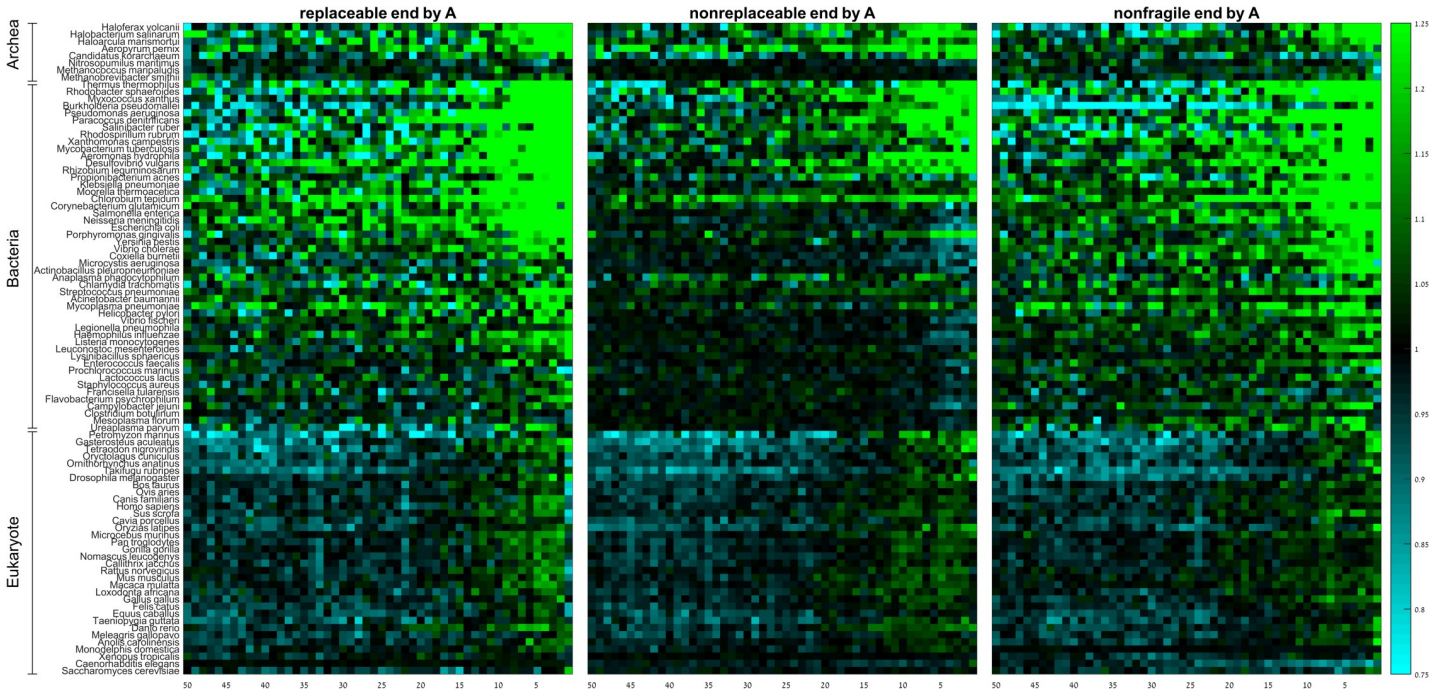

Fig 7. Group RSCA scores of replaceable, non-replaceable, and non-fragile codons that end by A along the last 50 codons of the gene. Rows denote species, columns denote positions. Species within domains are sorted from high (top) to low (bottom) GC-content.

https://doi.org/10.1371/journal.pone.0225633.g007

15 codons (observed MFE), and then replaced codons that end by A with the most frequently used codon for the corresponding amino acid in that species. If that codon also ends by A, we did not replace it. We computed the folding energy of the new sequence after replacing the last 15 codons (expected MFE). The results show (S10 Table) that in all species, the expected MFE was significantly lower than the observed MFE, indicating that the observed secondary structures at the 3'-end are less stable than the secondary structures expected from the general codon bias. These results suggest that the universal preference of A-ending codons at the 3'-end has a direct effect on lowering the stability of mRNA folding at that region.

Highly expressed genes show stronger selection against GC-rich codons at the 3'-end

It has been reported that highly expressed genes have stronger positional dependence of codon usage than lowly expressed genes in regions downstream of the 5'-end [47]. Given that highly expressed genes are assumed to be optimized for fast translation and low error rate, one may expect to see differences in positional CUB across genes with different expression levels. To check if this is the case at the 3'-end, we classified human genes to highly expressed and lowly expressed according to their average expression across 53 human tissues (see Methods; Fig 8). We found that the patterns of RSCA at the last 50 codons of highly expressed and lowly expressed genes behave similarly, and match the overall patterns reported above. There are nevertheless some minor differences. For example, strong depletion of G-ending codons starts earlier (at the 28th codon) in highly expressed genes, while it starts later (at the 15th codon) in lowly expressed genes. This is also the case for codons that are GC-rich (3/3 codons), where depletion starts earlier (at the 34th codon) in highly expressed genes than in lowly expressed ones (at the 12th codon). These results indicate that while the depletion of GC-rich and G-

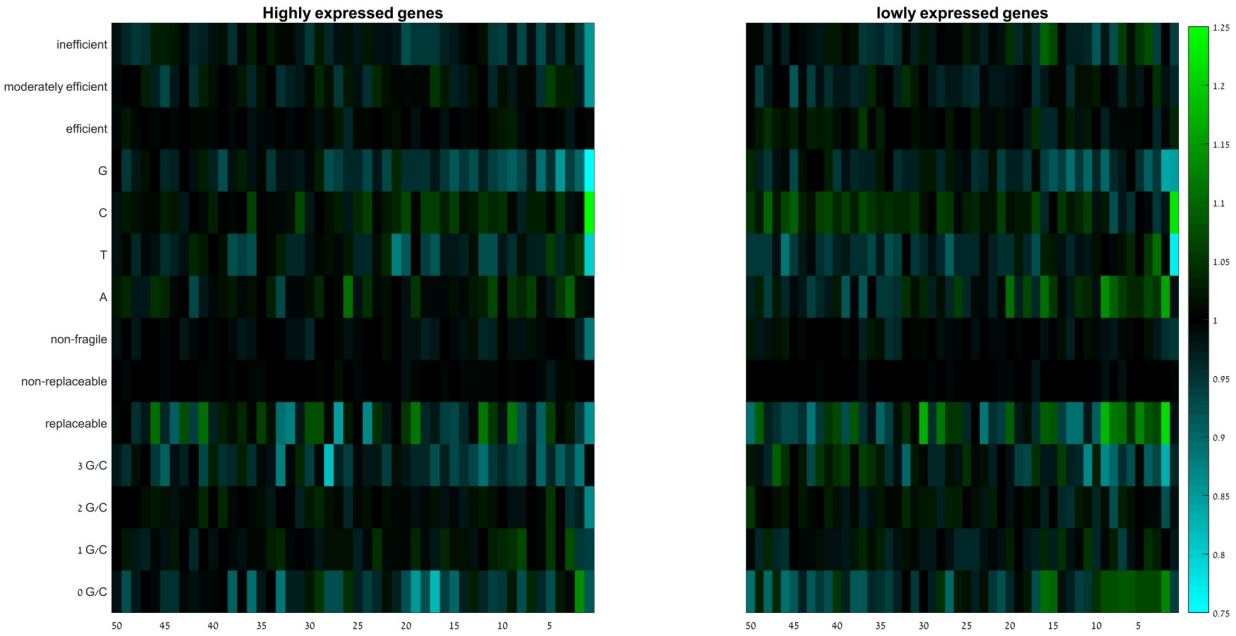

Fig 8. Group RSCA scores of all codons groups along the last 50 codons of human highly expressed and lowly expressed genes. Rows denote codons groups, columns denote positions. Species within domains are sorted from high (top) to low (bottom) GC-content.

https://doi.org/10.1371/journal.pone.0225633.g008

ending codons at the 3'-end is independent of gene expression, the strength of the depletion and the exact starting position might be affected by expression level. Interestingly, inefficient codons are more depleted toward the 3'-end in highly expressed genes, while lowly expressed genes show a mixture of patterns for these codons.

Discussion

Codon usage bias near the gene start has received considerable attention, likely because of its assumed role in translation initiation [35, 47, 49]. In contrast, codon usage bias near the gene end gained less attention. Yet, codons in this region may affect translation termination, and therefore participate in translation regulation. Three main translation-related forces have been suggested to explain the unique nucleotide composition near the gene end, called in this work translation efficiency, translation accuracy, and mRNA folding. The main purpose of our work was to determine which of these three forces predominates codon usage patterns near the gene end.

Although several indices that measure codon usage bias have been developed, none directly accounts for the relative position of the codon along the gene. To fill this gap, we developed a measure called the relative spatial codon abundance (RSCA), which is a position-dependent index for codon usage bias. We focused at the 3'-end of the CDS, and used RSCA to evaluate the spatial CUB patterns of each codon in 91 species, covering all three domains of life. Grouping the codons according to various classification schemes, we were able to determine which of the three models above contribute most to the codon composition at the 3'-end of genes.

We noticed strike differences in codon usage patterns between eukaryotes and prokaryotes. Generally, C-ending and GC-rich codons are over-represented throughout most of the gene end in eukaryotes, whereas the opposite holds for A-ending, T-ending, and AT-rich codons. This observation cannot be attributed to elevated GC-content in eukaryotes, as the studies species from the three domains have similar GC content ($P = 0.75$, Kruskal-Wallis test; S2 Table).

We suggest that the differences in spatial patterns that we see in this work likely reflect the high variability of the GC content along the coding region in eukaryotes, stemming from the fact that splicing sites and recombination hotspots are characterized by increased GC-content compared to other regions [50, 51]. The region roughly 100-200nt from each side of splicing sites is known to exhibit higher GC content compared to farther regions [51], which likely affects nucleotide composition around the TC. For example, in human, the median distance between the TC and the closest upstream splice site is 70nt. Moreover, it was found that codon usage is positively correlated with recombination rate (Marais & Piganeau, 2002), suggesting that Hill Robertson effect is a contributing factor affecting CUB. However, it is possible that this correlation is an artifact due to the mutational bias associated with recombination (Marais & Piganeau, 2002). For instance, since most optimal codons in *D. melanogaster* and *C. elegans* end in G or C, the enrichment of optimal codons observed at recombination hotspots in these species may be a byproduct of the mutational bias associated with high GC content in such regions (Marais, Mouchiroud, & Duret, 2001). Several studies suggest that GC-rich sequences are favored in gene conversion associated with recombination, an observation that was coined "biased gene conversion" (Bill, Duran, Miselis, & Nickoloff, 1998; Brown & Jiricny, 1988). It was claimed that this may explain the correlation between codon usage and recombination rate (Galtier, Piganeau, Mouchiroud, & Duret, 2001). Yet, weak effect of the Hill Robertson effect was found on codons usage after controlling for mutational bias, suggesting that it is a minor contributor to the observed codon usage bias (Marais & Piganeau, 2002). As a result, eukaryotes show high variability in the GC content along the coding region. Prokaryotes, lacking splicing and less affected by GC biased gene conversion and Hill Robertson effect, show lesser variability of the GC content within coding regions. In any case, RSCA is unaffected by overall biases in GC content but is sensitive to the changes in the codon usage across the sequence.

Interestingly, whereas A-ending codons also tend to be generally avoided at gene ends of eukaryotes, they are almost universally preferred along the last 10–15 codons (Fig 2), in agreement with previous studies [31, 36, 42]. However, our results show, for the first time, a variation in this pattern across kingdoms. For example, we show that C-ending codons are avoided near the gene end in prokaryotes, but in eukaryotes it is G-ending codons that are avoided. Moreover, the starting position of the switch in codons usage varies among species, which may reflect some lineage specific factors.

A notable exception to the universal preference of A-ending codons toward 3'-end is *C. elegans*. In C. *elegans* A-ending codons are significantly under-represented, while G-ending codons, and to a lesser extent T-ending ones, tend to be preferred. Yet AT-rich codons are still over-represented, while GC-rich codons are under-represented. The archaeon *N. martimus* deviates from all other species as it shows not only under-representation of A-ending and T-ending codons but also preference for GC-rich codons over AT-rich codons at the last 5 codons. A similar pattern can be seen in other AT-rich bacteria such as *C. jejuni*. Interestingly, Other AT-rich archaea such as M. *maripaludis*, in spite of showing neither over-representation nor under-representation of A-ending codons, still show preference to AT-rich and T-ending codons and avoidance of C-ending and GC-rich codons. Notably,most of the species that deviate from the general preference to A-ending and AT-rich codons near the gene end have small genomes, short genes, and are AT-rich. It is therefore possible that the general abundance of AT-rich codons makes the detection of 3'-end preference harder to detect. It is also possible that such genomes may have evolved different mechanisms to maintain the efficiency of translation termination. We noticed that these species nevertheless share the same trend toward enrichment in A-ending codons towards the gene end when introducing +1nt frameshift to the coding region (S5 and S7 Figs).

Selection for or against translation efficiency has been studied extensively and was shown to significantly contribute to codon usage bias at the gene start. At the gene end, Tuller *et al*., reported an enrichment in inefficient codons in some species, while others show the opposite trend [35]. There are two commonly used approaches to proxy a codon's efficiency. The first computes the relative adaptiveness of a codon from the copy number of its cognate tRNA gene (the tAI measure), thus codons recognized by abundant tRNAs are considered more efficient than those recognized by rare tRNAs. The second computes the relative adaptiveness of a codon from its abundance in highly expressed genes (the CAI measure), thus codons that appear frequently in highly expressed gene are assigned higher efficiency. As the computation of relative adaptiveness from highly expressed genes is challenging in species with poor genomic annotations, and because the enrichment of specific codons in such genes may also be due to factors such as GC content [52], we preferred in this work to use the tAI index. Our work provides no support for appreciable contribution of translation efficiency to the codon composition near the gene end. The trends identified in previous reports disappear almost completely after controlling for the nucleotide content at the wobble position.

While selection against missense errors is believed to be independent on the position along the CDS [25], selection against nonsense errors has a strong positional dependence. It has been claimed that this is the dominant factor shaping spatial codon usage bias at the gene start and end in prokaryotes [25]. Some works make no clear distinction between translation efficiency and accuracy, and thus proxy the accuracy of a codon by its adaptiveness value, relying on the fact that inefficient codons are generally also more prone to translation errors. However, here we used a more direct approach. We looked at fragile codons, which are codons with higher likelihood to become a stop codon following a substitution [40]. We divided fragile codons to two groups: those that have no non-fragile synonymous codons (non-replaceable), and those that have at least one non-fragile synonymous codon (replaceable). We found that replaceable codons are generally over-represented along the last 10 codons in almost all species, while non-replaceable codons are neither over-represented nor under-represented in all species (Fig 6). This pattern may have three different explanations:

(1) Since truncation of the protein may be less harmful toward its very end, there might be a relief in the strength of selection against replaceable codons towards the CDS end.

(2) Replaceable codons are enriched for codons that end with A (71% of the replaceable codons, but only 27% of the non-replaceable and 14% of the non-fragile; S2 Table), thus their enrichment or depletion towards the gene end are largely dictated by the nucleotide composition of the wobble position.

(3) Replaceable codons are enriched in translationally inefficient codons. On average, 20–25% of inefficient codons at the species of every domain are also replaceable, compared to 10% of moderately efficient codons and 5% of efficient codons. Therefore, their enrichment or depletion towards the gene end are largely dictated by the level of codon efficiency.

Our results rule out the third possibility, as we have shown that codon efficiency does not contribute significantly to the spatial CUB near the gene end. Our results also rule out the first explanation, as codons ending with different nucleotides show different spatial patterns (Fig 7, S12–S14 Figs; S4, S5, S8 and S9 Tables). We therefore concluded that it is the content of the wobble position that predominantly determines the spatial CUB at the gene end. Thus, neither selection against nonsense errors nor its relief at the far end of the gene are major contributors to the spatial codon usage bias at the gene 3'-end.

If the nucleotide composition is the main contributor to the codon usage bias at the 3'-end, we expect to see variability in the strength of this bias between highly and lowly expressed genes. Hockeberry *et al*. [47] studied the codon usage bias in *E.coli* from a positional context, and found that highly and lowly expressed genes behave similarly at the 5'-region, but

positional dependency was significantly higher for highly expressed genes at distal parts. Part of the positional dependency was related to the increased usage of A/T rich codons just after the gene start, and part of it, in particular the 4- and 6-fold degenerate codons, was related to gene copy number of cognate tRNA. Because expression data for human is available for many different tissues, we used human as a model to study the effect of gene expression on the RSCA pattern at the 3'-end (Fig 8). Our results show that the avoidance of GC-rich and G-ending codons starts earlier in highly expressed genes and the strength of the depletion is much stronger at the last 5 codons. This suggests that the strength of selection pressure against secondary structures toward the 3'-end depends on the gene expression level. Interestingly, our results also show a strong under-representation of inefficient codons near the 3'-end in highly expressed genes. Highly expressed genes avoid the usage of inefficient codons to maintain proper translation speed specifically at the 5'-end [49]. Our results show that this avoidance is present near the 3'-end.

To sum up, we conclude that neither selection against translation efficiency, nor selection for translation accuracy have a significant role in shaping the spatial codon usage bias at the gene 3'-end. Our work revealed that what mainly governs this pattern is the codon nucleotide content, especially at its wobble position. We showed that A/T-ending codons are preferred at the gene end, while C/G-ending codons are avoided. We have shown this using species from all three domains, which cover a wide range of GC contents and gene lengths (S1 Table). Thus, it is unlikely that species-specific factors such as particular evolutionary history, genome size, and Hill-Robertson effect can explain this observation. Instead, our results are fully compatible with the notion that codon usage patterns towards the gene end are generally a result of selective forces against the formation of RNA secondary structures near the stop codon [31]. It has been shown that GC-rich codons are associated with more stable RNA secondary structures [53]. We showed here that the preference toward A-ending codons lower the stability of mRNA folding at the 3'-end (S10 Table). The presence of mRNA secondary structures along the CDS stalls ribosome elongation [54], and lead to translation abortion [31, 55]. Moreover, translation termination requires the recruitment of release factors RF1, RF2, and RF3. Aberrant, and especially delayed, kinetics of the interactions of these release factors with the stalling ribosome may trigger translation read-through, whereby a near cognate tRNA recognizes the stop codon as a sense codon [31, 56]. We therefore suggest that there is a strong selection against the formation of mRNA secondary structures at the vicinity of the stop codon in order to allow for normal translation termination. We have shown that this nucleotide composition signature is also apparent at the 3'UTR side of the stop codon [57].

Methods

RSCA computation

Let G be a set of genes, and let l_g be the length, measured by the number of codons, of gene g in the set. We are interested in computing the spatial distribution of codons along the last ΔL codons, and will hereinafter assume that all genes in the set G are of minimal length ΔL. Let $C = \Sigma_{g\in G} l_g$ be the total number of codons in our gene set, and let $N_l = \Sigma_{g\in G} 1_{\{l_g \geq l\}}$ be the number of genes whose length is at least l.

Let the position of a codon, α, be its consecutive number from the gene end. For example the, last codon of a gene has $\alpha = 1$. Let $c^i_{g\alpha}$ be 1 if codon type i is at position α in gene g, and 0 otherwise. For example, $c^{CUU}_{g\alpha} = 1$ means that codon CUU is present at position α in gene g. If a position is not present in a gene, for example if we are looking at position 10 for a gene whose length is 9, we set $c^i_{g\alpha}$ to zero for all i. Let $C^i = \Sigma_{g\in G} \Sigma^{l_g}_{\alpha=1} c^i_{g\alpha}$ be the total number of times

that codon of type i appears in our gene set. Similarly, let $C^i_\alpha = \Sigma_{g\in G} c^i_{g\alpha}$ be the total number of codons of type i at position α in our gene set. The total frequency of codon type i is $p^i = C^i/C = C^i/\Sigma_i C^i$. Assuming uniform distribution of each codon both within and between the genes in G, the expected number of times that codon type i would appear at position α in our gene set is $E^i_\alpha = p^i \cdot N_\alpha$ We define the *spatial codon abundance* of codon i at position α as the ratio of the observed to the expected number

$$Q^i_\alpha = \frac{C^i_\alpha}{E^i_\alpha}.$$

$Q^i_\alpha = 1$ describes a codon whose frequency at position α if exactly as expected if it were uniformly distributed along the gene. $Q^i_\alpha > 1$ describes a codon that is over-represented at position α, whereas $Q^i_\alpha < 1$ describes a codon that is under-represented at position α. To measure how significantly Q^i_α deviates from one, we compute its standard error. Since E^i_α is a binomial random variable, its standard deviation is $\Delta E^i_\alpha = \sqrt{p^i(1-p^i)N_\alpha} = \sqrt{(1-p^i)E^i_\alpha}$. By means of error propagation, the standard error of Q^i_α is

$$\Delta Q^i_\alpha = Q^i_\alpha \cdot \frac{\Delta E^i_\alpha}{E^i_\alpha} = Q^i_\alpha \sqrt{\frac{1-p^i}{E^i_\alpha}}.$$

Q^i_α allows the measurement of codon usage bias towards the gene end. However, this bias is built from two contributions: the bias of the coded amino acid, and the bias of the codon itself. For example, assume that the UGG codon is over-represented towards the gene end. However, since it is the only codon for tryptophan, this bias may be entirely due to an over-representation of tryptophan at the gene end. In this work, we are interested only in the contribution of the codon itself to the bias. To take this into account, let $A(i)$ be the set of codons that are synonymous to codon type i (including i itself). Let $A^i_\alpha = \Sigma_{j\in A(i)} \sum_{g\in G} c^j_{g\alpha}$ be the total number of times any codon from the set $A(i)$ appears at position α in our gene set. Let us define the relative usage of codon i as $f^i = C^i/\Sigma_{j\in A(i)} C^j$. Then, the expected number of times that codon i will be used at position α is $H^i_\alpha = f^i A^i_\alpha$. We define the *relative spatial codon abundance* (RSCA) of codon i at position α, after accounting for the amino acid spatial bias, as

$$R^i_\alpha = \frac{C^i_\alpha}{H^i_\alpha}.$$

This is the value that we use throughout this work. Since H^i_α is a binomial random variable, its standard deviation is $\Delta H^i_\alpha = \sqrt{f_i(1-f_i)A^i_\alpha} = \sqrt{(1-f_i)H^i_\alpha}$. By means of error propagation, the standard error of R^i_α is

$$\Delta R^i_\alpha = R^i_\alpha \cdot \frac{\Delta H^i_\alpha}{H^i_\alpha} = R^i_\alpha \sqrt{\frac{1-f_i}{H^i_\alpha}}.$$

With this measure, the codon UGG from the example above would have $R^{UGG}_\alpha = 1$ for all α, as all the bias is entirely explained by the amino acid.

Group RSCA score

Whereas R^i_α measures the spatial abundance of a certain codon, in this work we will mostly be interested in the spatial patterns characterizing a group of n codons, S. We therefore define the group RSCA, R^S_α, as the median of R^i_α scores of S codons which we used in the entire work.

Genomic data

Gene coding sequences of the studied species were downloaded from Ensembl [58]. We randomly sampled half of the eukaryotic species from the ensemble eukaryotic collection at ftp://ftp.ensembl.org/pub/currentfasta. For bacteria and archaea, we used the Enesmbl bacterial collection "bacterial 0 collection". We again randomly sampled half of the species. However, some species, such as *Buchnera aphidicola*, produce many missing values. We filtered out such species, ending up with 34 eukaryotes, 49 bacteria (in which 3 utilize Mycoplasma genetic code), and 8 archaea. If a gene had multiple isoforms, we included only the canonical form (longest CDS) in our analyses. We removed genes whose coding sequence was shorter than 100 codons, as well as genes whose annotation included obvious mistakes such as total coding length not divisible by three, lack of a stop codon, or the presence of internal stop codons.

Relative adaptiveness value

We computed relative adaptiveness values following the derivation of dos Reis *et al.* [7]. For codon type i, the absolute adaptiveness value is

$$W_i = \sum_{j=1}^{n_i} (1 - s_{ij}) C_{ij},$$

where n_i is the number of tRNAs that recognize codon type i, s_{ij} is a selective constraint of the coupling efficiency between the jth tRNA and the ith codon type, and C_{ij} is the copy number of the jth tRNA that recognizes the codon i. The relative adaptiveness value is defined as

$$w_i = \begin{cases} W_i / W_{\max} & W_i \neq 0 \\ W_0 & W_i = 0 \end{cases},$$

where W_0 is the geometric mean of all the W_i values.

Relative adaptiveness value was computed for all codons of each species after retrieving the corresponding tRNA genes copy numbers from genomic tRNA database http://lowelab.ucsc.edu/GtRNAdb. Based on these values, we binned the codons in each species into three groups: Codons with relative adaptiveness at the bottom third are denote inefficient codons; codons with relative adaptiveness at the middle third are denote moderate-efficient codons; and codons with relative adaptiveness at the upper third are denote efficient codons.

Classifying human genes to highly and lowly expressed genes

We obtained from Genotype-Tissue Expression (GTEX) project [59], the expression of 57,073 genes from 53 human tissues. We computed the average expression of each gene in all the tissues, and filtered out genes that have very low expression (less than 0.5) and those that lack coding region. The 15,471 remaining genes were classified as 'highly expressed genes' if their average expression is higher than the 75th percentile (3,817 genes) and as 'lowly expressed genes' if their average expression is below the 25th percentile (3,817 genes).

Supporting information

S1 Fig. Group RSCA scores (R_α^S) of efficient, moderately efficient, and inefficient codons along the last 50 codons of the gene for random codon permutation. Rows denote species, columns denote positions.
(PDF)

S2 Fig. Group RSCA scores (R_α^S) of A-ending, T-ending, C-ending, and G-ending codons along the last 50 codons of the gene for random codon permutation. Rows denote species, columns denote positions.
(PDF)

S3 Fig. Group RSCA scores (R_α^S) of replaceable, non-replaceable, and non-fragile codons along the last 50 codons of the gene for random codon permutation. Rows denote species, columns denote positions.
(PDF)

S4 Fig. Group RSCA scores (R_α^S) of 0, 1, 2, and 3 G/C codons along the last 50 codons of the gene for random codon perumtation. Rows denote species, columns denote positions.
(PDF)

S5 Fig. Group RSCA scores (R_α^S) of A-ending, T-ending, C-ending, and G-ending triplets after introducing +1nt frameshift to the coding region (i.e., examining the first codon position in the original coding sequence) along the last 50 codons of the gene. Rows denote species, columns denote positions. Species within domains are sorted from high (top) to low (bottom) GC-content.
(PDF)

S6 Fig. Group RSCA scores (R_α^S) of A-ending, T-ending, C-ending, and G-ending codons after introducing +2nt frameshift to the coding region (i.e., examining the second codon position in the original coding sequence) along the last 50 codons of the gene. Rows denote species, columns denote positions. Species within domains are sorted from high (top) to low (bottom) GC-content.
(PDF)

S7 Fig. Group RSCA scores (R_α^S) of 0, 1, 2, and 3 G/C nucleotides in codons after introducing +1nt frameshift to the coding region (i.e., examining the first codon position in the original coding sequence) along the last 50 codons of the gene. Rows denote species, columns denote positions. Species within domains are sorted from high (top) to low (bottom) GC-content.
(PDF)

S8 Fig. Group RSCA scores (R_α^S) of 0, 1, 2, and 3 G/C nucleotides in codons after introducing +2nt frameshift to the coding region (i.e., examining the second codon position in the original coding sequence) along the last 50 codons of the gene. Rows denote species, columns denote positions. Species within domains are sorted from high (top) to low (bottom) GC-content.
(PDF)

S9 Fig. Group RSCA scores (R_α^S) of efficient, moderately efficient, and inefficient codons ending by T along the last 50 codons of the gene. Rows denote species, columns denote positions. Species where none of the codons in a particular efficiency group end by T (missing data) are shown as white stripes. Species within domains are sorted from high (top) to low (bottom) GC-content.
(PDF)

S10 Fig. Group RSCA scores (R_α^S) of efficient, moderately efficient, and inefficient codons ending by C along the last 50 codons of the gene. Rows denote species, columns denote positions. Species where none of the codons in a particular efficiency group end by C (missing

data) are shown as white stripes. Species within domains are sorted as in S9 Fig.
(PDF)

S11 Fig. Group RSCA scores (R_α^S) of efficient, moderately efficient, and inefficient codons ending by G along the last 50 codons of the gene. Rows denote species, columns denote positions. Species where none of the codons in a particular efficiency group end by G (missing data) are shown as white stripes. Species within domains are sorted as in S9 Fig.
(PDF)

S12 Fig. Group RSCA scores (R_α^S) of replaceable, non-replaceable, and non-fragile codons ending by T along the last 50 codons of the gene. Rows denote species, columns denote positions. As there are no replaceable codons that end by T, all species are depicted by white stripes (missing data). Species within domains are sorted from high (top) to low (bottom) GC-content.
(PDF)

S13 Fig. Group RSCA scores (R_α^S) of replaceable, non-replaceable, and non-fragile codons ending by C along the last 50 codons of the gene. Rows denote species, columns denote positions. As there are no replaceable codons that end by C, all species are depicted by white stripes (missing data). Species within domains are sorted as in S12 Fig.
(PDF)

S14 Fig. Group RSCA scores (R_α^S) of replaceable, non-replaceable, and non-fragile codons ending by G along the last 50 codons of the gene. Rows denote species, columns denote positions. Species within domains are sorted as in S12 Fig.
(PDF)

S1 Table. List of the species analyzed, and some properties of their genes.
(PDF)

S2 Table. For each species and for every pair of codon classification scheme where efficiency is part of, we computed the (FDR corrected) p-value of the $\chi 2$ independence test.
(PDF)

S3 Table. Linear regression of group RSCA score at the last 10, 20, and 30 codons, for spatially randomized codons. Data for each species is shown in a different excel sheet.
(XLSX)

S4 Table. Linear regression of group RSCA score at the last 10, 20, and 30 codons. Data for each species is shown in a different excel sheet.
(XLSX)

S5 Table. Linear model scores for each codon group and species. The value 1 indicates a significant increase toward the 3'-end in the last 10, 20, or 30 codons. The value -1 indicates a significant decrease, and 0 indicates insignificant linear fit.
(XLSX)

S6 Table. For each position (distance from the stop codon, columns) and for every species, we computed the Kruskal-Wallis (FDR corrected) p-value, measuring the difference in the distribution of RSCA values between efficient, inefficient, and moderately efficient that end by A/T.
(PDF)

S7 Table. For each position (distance from the stop codon, columns) and for every species, we computed the Kruskal-Wallis (FDR corrected) p-value, measuring the difference in the

distribution of RSCA values between efficient, inefficient, and moderately efficient codons that end by C/G.
(PDF)

S8 Table. For each position (distance from the stop codon, columns) and for every species, we computed the Kruskal-Wallis (FDR corrected) p-value, measuring the difference in the distribution of RSCA values between replaceable, non-replaceable, and non-fragile that end by A/T.
(PDF)

S9 Table. For each position (distance from the stop codon, columns) and for every species, we computed the Kruskal-Wallis (FDR corrected) p-value, measuring the difference in the distribution of RSCA values between replaceable, non-replaceable, and non-fragile codons that end by C/G.
(PDF)

S10 Table. Median minimum free energy (MFE) at the 3'-end (observed MFE), median MFE expected by the global (not spatial) codon bias (expected MFE), and p-value for testing whether the median difference between pairs of expected and observed values equals zero (FDR-corrected two-sided sign test).
(PDF)

Author Contributions

Conceptualization: Liran Carmel.

Formal analysis: Fouad Zahdeh.

Methodology: Fouad Zahdeh, Liran Carmel.

Supervision: Liran Carmel.

Writing – original draft: Fouad Zahdeh, Liran Carmel.

Writing – review & editing: Fouad Zahdeh, Liran Carmel.

References

1. Man O, Pilpel Y. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. Nat Genet. 2007; 39(3):415–21. https://doi.org/10.1038/ng1967 PMID: 17277776.
2. Hiraoka Y, Kawamata K, Haraguchi T, Chikashige Y. Codon usage bias is correlated with gene expression levels in the fission yeast Schizosaccharomyces pombe. Genes Cells. 2009; 14(4):499–509. https://doi.org/10.1111/j.1365-2443.2009.01284.x PMID: 19335619.
3. Ikemura T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol. 1981; 151(3):389–409. https://doi.org/10.1016/0022-2836(81)90003-6 PMID: 6175758.
4. Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol. 1985; 2(1):13–34. https://doi.org/10.1093/oxfordjournals.molbev.a040335 PMID: 3916708.
5. Andersson SG, Kurland CG. Codon preferences in free-living microorganisms. Microbiol Rev. 1990; 54(2):198–210. PMID: 2194095; PubMed Central PMCID: PMC372768.
6. Sharp PM, Stenico M, Peden JF, Lloyd AT. Codon usage: mutational bias, translational selection, or both? Biochem Soc Trans. 1993; 21(4):835–41. https://doi.org/10.1042/bst0210835 PMID: 8132077.
7. 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. https://doi.org/10.1093/nar/gkh834 PMID: 15448185; PubMed Central PMCID: PMC521650.

8. Akashi H. Gene expression and molecular evolution. Curr Opin Genet Dev. 2001; 11(6):660–6. https://doi.org/10.1016/s0959-437x(00)00250-1 PMID: 11682310.
9. Wallace EW, Airoldi EM, Drummond DA. Estimating selection on synonymous codon usage from noisy experimental data. Mol Biol Evol. 2013; 30(6):1438–53. https://doi.org/10.1093/molbev/mst051 PMID: 23493257; PubMed Central PMCID: PMC3649678.
10. Bulmer M. The selection-mutation-drift theory of synonymous codon usage. Genetics. 1991; 129 (3):897–907. PMID: 1752426; PubMed Central PMCID: PMC1204756.
11. Akashi H. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics. 1994; 136(3):927–35. PMID: 8005445; PubMed Central PMCID: PMC1205897.
12. Dong H, Nilsson L, Kurland CG. Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. J Mol Biol. 1996; 260(5):649–63. https://doi.org/10.1006/jmbi.1996.0428 PMID: 8709146.
13. Kramer EB, Farabaugh PJ. The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. RNA. 2007; 13(1):87–96. https://doi.org/10.1261/rna.294907 PMID: 17095544; PubMed Central PMCID: PMC1705757.
14. Qian W, Yang JR, Pearson NM, Maclean C, Zhang J. Balanced codon usage optimizes eukaryotic translational efficiency. PLoS Genet. 2012; 8(3):e1002603. https://doi.org/10.1371/journal.pgen.1002603 PMID: 22479199; PubMed Central PMCID: PMC3315465.
15. Hershberg R, Petrov DA. Selection on codon bias. Annu Rev Genet. 2008; 42:287–99. https://doi.org/10.1146/annurev.genet.42.110807.091442 PMID: 18983258.
16. Sharp PM, Averof M, Lloyd AT, Matassi G, Peden JF. DNA sequence evolution: the sounds of silence. Philos Trans R Soc Lond B Biol Sci. 1995; 349(1329):241–7. https://doi.org/10.1098/rstb.1995.0108 PMID: 8577834.
17. Urrutia AO, Hurst LD. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics. 2001; 159(3):1191–9. PMID: 11729162; PubMed Central PMCID: PMC1461876.
18. Lafay B, Atherton JC, Sharp PM. Absence of translationally selected synonymous codon usage bias in Helicobacter pylori. Microbiology. 2000; 146 (Pt 4):851–60. https://doi.org/10.1099/00221287-146-4-851 PMID: 10784043.
19. Galtier N, Piganeau G, Mouchiroud D, Duret L. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics. 2001; 159(2):907–11. PMID: 11693127; PubMed Central PMCID: PMC1461818.
20. Begun DJ. The frequency distribution of nucleotide variation in Drosophila simulans. Mol Biol Evol. 2001; 18(7):1343–52. https://doi.org/10.1093/oxfordjournals.molbev.a003918 PMID: 11420372.
21. Berg OG. Selection intensity for codon bias and the effective population size of Escherichia coli. Genetics. 1996; 142(4):1379–82. PMID: 8846914; PubMed Central PMCID: PMC1207134.
22. Li WH. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol. 1987; 24(4):337–45. https://doi.org/10.1007/bf02134132 PMID: 3110426.
23. Marais G, Mouchiroud D, Duret L. Neutral effect of recombination on base composition in Drosophila. Genet Res. 2003; 81(2):79–87. https://doi.org/10.1017/s0016672302006079 PMID: 12872909.
24. Marais G, Mouchiroud D, Duret L. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc Natl Acad Sci U S A. 2001; 98(10):5688–92. https://doi.org/10.1073/pnas.091427698 PMID: 11320215; PubMed Central PMCID: PMC33274.
25. Qin H, Wu WB, Comeron JM, Kreitman M, Li WH. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. Genetics. 2004; 168(4):2245–60. https://doi.org/10.1534/genetics.104.030866 PMID: 15611189; PubMed Central PMCID: PMC1448744.
26. Comeron JM, Kreitman M. Population, evolutionary and genomic consequences of interference selection. Genetics. 2002; 161(1):389–410. PMID: 12019253; PubMed Central PMCID: PMC1462104.
27. Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genet Res. 1966; 8(3):269–94. PMID: 5980116.
28. Carlini DB, Chen Y, Stephan W. The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr. Genetics. 2001; 159(2):623–33. PMID: 11606539; PubMed Central PMCID: PMC1461829.
29. Eyre-Walker A, Bulmer M. Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Res. 1993; 21(19):4599–603. https://doi.org/10.1093/nar/21.19.4599 PMID: 8233796; PubMed Central PMCID: PMC311196.

30. Chen Y, Carlini DB, Baines JF, Parsch J, Braverman JM, Tanda S, et al. RNA secondary structure and compensatory evolution. Genes Genet Syst. 1999; 74(6):271–86. https://doi.org/10.1266/ggs.74.271 PMID: 10791023.

31. Rocha EP, Danchin A, Viari A. Translation in Bacillus subtilis: roles and trends of initiation and termination, insights from a genome analysis. Nucleic Acids Res. 1999; 27(17):3567–76. https://doi.org/10.1093/nar/27.17.3567 PMID: 10446248; PubMed Central PMCID: PMC148602.

32. 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. https://doi.org/10.1093/gbe/evr129 PMID: 22138151; PubMed Central PMCID: PMC3269970.

33. 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 PMID: 20140241; PubMed Central PMCID: PMC2816680.

34. Bentele K, Saffert P, Rauscher R, Ignatova Z, Bluthgen N. Efficient translation initiation dictates codon usage at gene start. Mol Syst Biol. 2013; 9:675. https://doi.org/10.1038/msb.2013.32 PMID: 23774758; PubMed Central PMCID: PMC3964316.

35. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, et al. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell. 2010; 141(2):344–54. https://doi.org/10.1016/j.cell.2010.03.031 PMID: 20403328.

36. Eyre-Walker A. The close proximity of Escherichia coli genes: consequences for stop codon and synonymous codon use. J Mol Evol. 1996; 42(2):73–8. https://doi.org/10.1007/bf02198830 PMID: 8919857.

37. Eyre-Walker A. Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy? Mol Biol Evol. 1996; 13(6):864–72. PMID: 8754221.

38. Bulmer M. Codon usage and intragenic position. J Theor Biol. 1988; 133(1):67–71. https://doi.org/10.1016/s0022-5193(88)80024-9 PMID: 3066998.

39. Kurland CG. Translational accuracy and the fitness of bacteria. Annu Rev Genet. 1992; 26:29–50. https://doi.org/10.1146/annurev.ge.26.120192.000333 PMID: 1482115.

40. Cusack BP, Arndt PF, Duret L, Roest Crollius H. Preventing dangerous nonsense: selection for robustness to transcriptional error in human genes. PLoS Genet. 2011; 7(10):e1002276. https://doi.org/10.1371/journal.pgen.1002276 PMID: 22022272; PubMed Central PMCID: PMC3192821.

41. Maquat LE. Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol. 2004; 5(2):89–99. https://doi.org/10.1038/nrm1310 PMID: 15040442.

42. Katz L, Burge CB. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res. 2003; 13(9):2042–51. https://doi.org/10.1101/gr.1257503 PMID: 12952875; PubMed Central PMCID: PMC403678.

43. 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. https://doi.org/10.1093/nar/15.3.1281 PMID: 3547335; PubMed Central PMCID: PMC340524.

44. dos Reis M, Wernisch L, Savva R. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. Nucleic Acids Res. 2003; 31(23):6976–85. https://doi.org/10.1093/nar/gkg897 PMID: 14627830; PubMed Central PMCID: PMC290265.

45. Wright F. The 'effective number of codons' used in a gene. Gene. 1990; 87(1):23–9. https://doi.org/10.1016/0378-1119(90)90491-9 PMID: 2110097.

46. Sharp PM, Tuohy TM, Mosurski KR. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 1986; 14(13):5125–43. https://doi.org/10.1093/nar/14.13.5125 PMID: 3526280; PubMed Central PMCID: PMC311530.

47. Hockenberry AJ, Sirer MI, Amaral LA, Jewett MC. Quantifying position-dependent codon usage bias. Mol Biol Evol. 2014; 31(7):1880–93. Epub 2014/04/09. https://doi.org/10.1093/molbev/msu126 PMID: 24710515; PubMed Central PMCID: PMC4069614.

48. Bjornsson A, Mottagui-Tabar S, Isaksson LA. Structure of the C-terminal end of the nascent peptide influences translation termination. EMBO J. 1996; 15(7):1696–704. PMID: 8612594; PubMed Central PMCID: PMC450081.

49. Tuller T, Zur H. Multiple roles of the coding sequence 5' end in gene expression regulation. Nucleic Acids Res. 2015; 43(1):13–28. Epub 2014/12/17. https://doi.org/10.1093/nar/gku1313 PMID: 25505165; PubMed Central PMCID: PMC4288200.

50. Webster MT, Hurst LD. Direct and indirect consequences of meiotic recombination: implications for genome evolution. Trends Genet. 2012; 28(3):101–9. https://doi.org/10.1016/j.tig.2011.11.002 PMID: 22154475.

51. Zhang J, Kuo CC, Chen L. GC content around splice sites affects splicing through pre-mRNA secondary structures. BMC Genomics. 2011; 12:90. https://doi.org/10.1186/1471-2164-12-90 PMID: 21281513; PubMed Central PMCID: PMC3041747.

52. Sabi R, Tuller T. Modelling the efficiency of codon-tRNA interactions based on codon usage bias. DNA Res. 2014; 21(5):511–26. https://doi.org/10.1093/dnares/dsu017 PMID: 24906480; PubMed Central PMCID: PMC4195497.

53. Seffens W, Digby D. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. Nucleic Acids Res. 1999; 27(7):1578–84. https://doi.org/10.1093/nar/27.7.1578 PMID: 10075987; PubMed Central PMCID: PMC148359.

54. Plat T. RNA Structure and function. New York, NY: Cold Spring Harbour Laboratory Press; 1998.

55. Carpousis AJ, Vanzo NF, Raynal LC. mRNA degradation. A tale of poly(A) and multiprotein machines. Trends Genet. 1999; 15(1):24–8. https://doi.org/10.1016/s0168-9525(98)01627-8 PMID: 10087930.

56. Tate WP, Mannering SA. Three, four or more: the translational stop signal at length. Mol Microbiol. 1996; 21(2):213–9. https://doi.org/10.1046/j.1365-2958.1996.6391352.x PMID: 8858577.

57. Zahdeh F, Carmel L. The role of nucleotide composition in premature termination codon recognition. BMC Bioinformatics. 2016; 17(1):519. https://doi.org/10.1186/s12859-016-1384-z PMID: 27927164.

58. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res. 2016; 44(D1):D710–6. https://doi.org/10.1093/nar/gkv1157 PMID: 26687719; PubMed Central PMCID: PMC4702834.

59. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013; 45(6):580–5. Epub 2013/05/30. https://doi.org/10.1038/ng.2653 PMID: 23715323; PubMed Central PMCID: PMC4010069.