THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Translation comes first

Citation for published version:
González-Serrano, F, Abreu-Goodger, C & Delaye, L 2022, 'Translation comes first: Ancient and convergent selection of codon usage bias across prokaryotic genomes', *Journal of molecular evolution*. https://doi.org/10.1007/s00239-022-10074-0

Digital Object Identifier (DOI):
10.1007/s00239-022-10074-0

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Journal of molecular evolution

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 27. Oct. 2022

Translation comes first: ancient and convergent selection of codon usage bias across prokaryotic genomes

Francisco González-Serrano[1,2], Cei Abreu-Goodger[3], Luis Delaye[1*]

1. Genetic Engineering Department, CINVESTAV Irapuato, Guanajuato, México
2. Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico
3. Institute of Ecology and Evolution, University of Edinburgh, Edinburgh, UK.

*Corresponding author. E-mail: luis.delaye@cinvestav.mx

Abstract

Codon usage is the outcome of different evolutionary processes and can inform us about the conditions in which organisms live and evolve. Here we present *R_ENC'*, which is an improvement to the original *S* index developed by dos Reis et al. (2004). Our index is less sensitive to G+C content, which greatly affects synonymous codon usage in prokaryotes, making it better suited to detect selection acting on codon usage. We used *R_ENC'* to estimate the extent of selected codon usage bias in 1800 genomes representing 26 prokaryotic phyla. We found that Gammaproteobacteria, Betaproteobacteria, Actinobacteria, and Firmicutes are the phyla/subphyla showing more genomes with selected codon usage bias. In particular, we found that several lineages within Gammaproteobacteria and Firmicutes show a similar set of functional terms enriched in genes under selected codon usage bias, indicating convergent evolution. We also show that selected codon usage bias tends to evolve in genes coding for the translation machinery before other functional GO terms. Finally, we discuss the possibility to use *R_ENC'* to predict whether lineages evolved in copiotrophic or oligotrophic environments.

Keywords: translational selection, genome evolution, codon usage, prokaryotes.

Introduction

The genetic code consists of a set of molecular interactions used by all cells to translate information encoded in nucleotide sequences into proteins. The principle of these interactions lies in the assignment of 61 codons to the main 20 amino acids that make up proteins. The code is degenerate, meaning that the same amino acid can be specified by more than one synonymous codon. The relative frequency of these synonymous codons varies between genes and genomes. This uneven use of codons is known as codon usage bias (Grantham et al. 1981; Ikemura 1981a, b, 1985; Plotkin and Kudla 2011; Parvathy et al. 2022).

The codon usage of a gene is the result of the combined action of mutational bias, genetic drift, and natural selection (Bulmer 1991). When natural selection is not strong enough, the codon usage of a gene is mostly the result of genetic drift sampling on codons originated by mutation. This is particularly notable in the genomes of endosymbiotic bacteria of insects that tend to become A+T rich by nonadaptive processes (Martínez-Cano et al. 2015). In these cases, it is typical to find the majority (if not all) of the genes in the genome showing a bias towards A+T rich codons.

There are other occasions when natural selection is strong enough to modify the codon usage of genes. For instance, in highly expressed genes (like those encoding ribosomal proteins and translation factors), selection favors a codon usage that facilitates translation. This is, in highly expressed genes, selection favors codons that best match the abundance of tRNA species in the cytoplasm. This effect is more pronounced in species that have short generation times (Rocha 2004; Sharp et al. 2005; Vieira-Silva and Rocha 2010). Genes showing this kind of selected

codon usage bias, which is also called translational selection, are translated faster and released earlier from the ribosome, making their translation more efficient (Quax et al. 2015; Gustafsson et al. 2004; Ran and Higgs 2012; Frumkin et al. 2018). In genomes where selection affects codon usage bias, highly expressed genes typically show codon usage that correlates with tRNA abundance; while for genes with lower levels of expression codon usage is determined by mutational bias.

Selected codon usage bias has been related to the lifestyle of bacteria and archaea. Prokaryotes showing selected codon usage bias tend to live in a wider range of habitats (like pathogenic bacteria that can live outside or within their host); while prokaryotes with a more specialized lifestyle (like thermophiles, mutualistic endosymbiotic bacteria, or obligate intracellular parasites) tend to show no selected codon usage bias (Botzman and Margalit 2011). According to these observations, selected codon usage bias allows bacteria to grow faster and to face metabolic diverse environments where competition is strongest.

Organisms that are able to grow in environments where nutrients are scarce are known as oligotrophs. By contrast, organisms that grow in environments with larger nutrient availability are known as copiotrophs (Koch 2001). Oligotrophs tend to grow slowly and make efficient use of resources, while copiotrophs are capable of growing faster when nutrients are available (Roller and Schmidt 2015). An association between selected codon usage bias and nutrient availability (oligotrophic/copiotrophic) has been recently demonstrated. Copiotrophic environments favor bacteria showing translational selection in their ribosomal protein-coding genes, while genomes of bacteria in oligotrophic environments suggest codon usage bias selection playing a relaxed role (Okie et al. 2020).

There have been several attempts to measure the strength of selected codon usage bias. One of the first was developed by Wright (1990) and then modified by Fuglsang (2006). In this work, we propose an improvement to a previous method developed by dos Reis et al. (2004). We then use this improved method to revisit a) the correlation between selected codon usage bias and minimal generation time taking into account the phylogenetic structure of the data; b) the pattern of selected codon usage bias in different functional GO terms across diverse lineages of bacteria and archaea; and c) the evolutionary precedence of selected codon usage bias of the genes coding for the translation apparatus relative to other functional GO terms.

Materials and Methods

Improving selected codon usage bias estimation

To estimate the strength of selected codon usage bias we relied on the method proposed by dos Reis et al. (2004). In summary, this method works as follows. For a given genome: first 1) measure the tRNA adaptation index (*tAI*); then 2) measure the difference between the expected and observed number of codons (*dNc*); and 3) calculate a Pearson correlation coefficient between *tAI* and *dNc*. This correlation is named *S* by dos Reis et al. (2004). The higher the correlation coefficient, the stronger natural selection is predicted to have shaped codon usage bias on a whole-genome basis.

There are two important things to note about *S*. The first one is that *S* measures the coadaptation between the codon usage and the tRNA gene pool (this pool is taken as a proxy of the abundance of cytoplasmic tRNA). The second one is that *S* measures the effects of selection on codon usage bias irrespectively of a specific set of protein-coding genes. Therefore, *S* does not assume a priori that there is a set of genes (like those coding for ribosomal proteins) on which selected codon

usage will be strongest. In this last respect, *S* differs from other methods that rely on ribosomal protein-coding genes as gold standards to measure the strength of selected codon bias.

We improved *S* by making *dNc* less sensitive to G+C content as follows. First, *dNc* is the difference between the expected number of codons due to G+C content at third codon positions (Nc_e) and the observed number of codons (Nc) in a given gene [Equation 1]. It measures how far the effective number of codons depart from the neutral expectancy due to the G+C content of 3rd codon positions.

[Equation 1]

$$dNc = Nc_e - Nc$$

To calculate the expected number of codons under the hypothesis of no selection (Nc_e), dos Reis et al. (2004) used [Equation 2]:

[Equation 2]

$$Nc_e = f_1(x_g) = a + x_g + (b/(x^2_g + (c - x_g)^2))$$

Where x_g is the GC content of a gene *g,* and *a*, *b,* and *c* are constants whose values were estimated empirically.

To calculate the observed number of codons (Nc), dos Reis et al. (2004) used the original formula developed by Wright (1990) [Equation 3]:

[Equation 3]

$$Nc = 2 + (9/F_2) + (1/F_3) + (5/F_5) + (3/F_6)$$

Where:

[Equation 4]

$F_a = (n_a \sum_{i=1}^{k} p^2_i - 1)/(n_a - 1)$

And n_a is the observed number of codons for the *a* amino acid; p_i is the frequency of the *i*th codon, and *k* is the number of synonymous codons for the amino acid of interest.

Unfortunately, *Nc* is affected by G+C content since it assumes equal usage of all codons. Here we used instead, an improved statistic (*Nc'*) to calculate the expected effective number of codons that better accounts for background nucleotide composition (Novembre 2002). In this new formulation, Novembre (2002) used the Pearson X^2 statistic "*to quantify the departure of each codon's usage (p_i) from some expected usage (e_i) for each amino acid*" [Equation 5]:

[Equation 5]

$X^2_a = \sum_{i=1}^{k} (n_a(p_i - e_i)^2/e_i)$

And using the X^2_a values to calculate F'_a:

[Equation 6]

$F'_a = (X^2_a + n_a - k)/k(n_a - 1)$

And then the F'_a are used to calculate *Nc'* with Equation 7 which is equivalent to Equation 3 (see Novembre 2002 for details).

[Equation 7]

$$Nc' = 2 + (9/F'_2) + (1/F'_3) + (5/F'_5) + (3/F'_6)$$

Here we estimate selected codon usage bias by measuring the correlation between *tAI* and *Nc'* and multiplying Pearson's correlation coefficient by -1 (to make it comparable to *S*). We call this statistic *R_ENC'* to differentiate it from the *S* of dos Reis et al. (2004).

Estimation of selected codon usage bias from prokaryotic genomes

Nc, *dNc*, *Nc'* and *tAI* were calculated on all protein coding regions encoding a minimum of 100 amino acids from 1800 genomes downloaded from Genome list NCBI browser searching only for complete prokaryotic genomes (https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/). These genomes are representative of 26 bacterial and archaeal phyla (Supplementary Table S1). The strategy for these calculations was as follows. *Nc* was calculated using CodonW (J Peden, version 1.4.2 http://codonw.sourceforge.net/) taking as input the CDS nucleotide sequences from each genome. *dNc* was calculated by using an in-house R script on the CodonW output. CodonM was used to calculate the codon appearance in CDS (dos Reis et al. 2004). For the calculation of *tAI*, all tRNAs were annotated *de novo* with tRNAscan-SE (Lowe & Eddy, version 2.0, http://lowelab.ucsc.edu/tRNAscan-SE/). And then, *tAI* was calculated in R using the *tAI* function from dos Reis et al. (2004) and the CodonM output. *Nc'* was calculated by the ENCprime software (Novembre, 2002).

Estimating tAI from RNA-seq data

We also introduced a modified version of *tAI*. As mentioned above, *tAI* indicates how well adapted a given gene is to the tRNA pool based on the diversity and abundance of coded tRNA and taking into account wobble base pair interactions (see dos Reis et al. 2004 for details).

Instead of using tRNA coding genes, when data were available, we estimated *tAI* from tRNA abundances as measured by RNA-seq experiments. We call this value *tAI'*. For *tAI'*, we calculated transcripts per million (*tpm*) from RNA-seq experiments conducted during log phase or exponential growth that were fetched from GEO DataSets (https://www.ncbi.nlm.nih.gov/gds/) and processed by (Wei et al. 2019). We estimated *tAI'* on the following experiments (genomes): SRX020805 (*Bacteroides thetaiotaomicron*); SRX515181 (*Bacillus subtilis*); SRX515174 (*Escherichia coli*); SRX2448246 (*Leptospira interrogans*); SRX1372108 (*Mycobacterium tuberculosis*); SRX1638989 (*Salmonella enterica*); SRX347145 (*Synechocystis sp.* PCC 6803).

Other measures of selected codon usage bias

For the sake of comparison, we also estimated selected codon usage bias by using the approach followed by Vieira-Silva and Rocha (2010). These authors proposed to estimate translational selection by an index named *ΔENC'*. To estimate *ΔENC'* we first estimated translational selection by averaging *Nc'* values from all ribosomal protein-coding genes (*ENC'r*) [Equation 8]:

[Equation 8]

$$ENC'r = \sum_{i=1}^{k} Nc'_i / k$$

k = number of ribosomal protein coding genes

Nc'_i = *Nc'* from the ith ribosomal protein coding gene

And used the above value to calculate *ΔENC'*, which is a standardization of the mean *Nc'* from all coding genes in the genome with respect to the mean of the ribosomal protein coding genes *ENC'r* values [Equation 9].

[Equation 9]

$$\Delta ENC' = Nc' - ENCr'/Nc'$$

We also calculated the average *tAI* for ribosomal protein coding genes (*tAIr*) [Equation 10] and fitted it to a Z distribution with mean 0 and standard distribution 1 named here as *tAIrz*.

[Equation 10]

$$tAIr = \sum_{i=1}^{k} tAI_i/k$$

k =number of ribosomal protein coding genes

tAI_i= *tAI* from the ith ribosomal protein coding gene

It was necessary to normalize *tAIr* to a Z distribution to be able to compare this value between genomes. Additionally, we calculated *ΔtAI* [Equation 11]

[Equation 11]

$$\Delta tAI = tAI - tAIr/tAI$$

Phylogenomic inference, phylogenetic signal and phylogenetic contrasts

Data on minimal generation times (*d*) from diverse prokaryotes were gathered and published by Vieira-Silva and Rocha (2010) (Supplementary Table S2). A phylogenomic tree of these 210 organisms was inferred using a concatenation of five protein sequences: valine--tRNA ligase (*valS*); elongation factor G (*fusA*); ATP-dependent zinc metalloprotease FtsH (*ftsH*); DNA-directed RNA polymerase subunit beta (*rpoC*); and elongation factor Tu (*tufA*). These phylogenetic markers were detected and aligned using PhyloPhlAn (Segata et al. 2014). For phylogenetic inference, the best model (LG;+G) was predicted by SMS (Lefort et al. 2017) and

the tree was reconstructed by PhyML (Guindon et al. 2010) with non-parametric branch support based on the Shimodaira-Hasegawa-like procedure (SH-like).

Phylogenetic signals for selected codon usage bias and minimal generation times were estimated with Pagel's lambda (λ) as implemented in the Phytools R package (Revell 2012).

Correlations between minimal generation time and: *R_ENC'*, *ENCr'*, *ΔENC'*, tRNA copy number, *tAIrz* and *ΔtAI* were evaluated with a Spearman's rank-order correlation and taking into account phylogenetic inertia by performing phylogenetic contrasts with the PGLS method as implemented in the Caper R package (Orme 2018).

GO terms enriched in genes showing selected codon usage bias

We also wanted to know if there were GO terms enriched in genes showing selected codon usage bias. First, for each one of the genomes in Table S2, we normalized *Nc'* and *tAI* values to mean 0 and standard deviation 1. Next, we performed two different enrichment analyses. For the first one, we plotted all genes from each genome from Table S2 in a *tAI* versus *Nc'* chart (see Figure 4). We then asked whether the set of ribosomal protein coding genes is enriched in the upper left quadrant by using a Kolmogorov–Smirnov test (p-value < 0.05). The results of these analysis are shown in Figure 4 and Supplementary Table S7. We also tested the above association by using a logistic regression.

For the second enrichment analysis, we grouped genes according to their Gene Ontology terms (Ashburner et al. 2000). Gene ontology annotations were downloaded from the UniProt database (UniProt Consortium, 2019; https://www.uniprot.org/). In order to assess statistical significance, a Kolmogorov–Smirnov test was implemented for each metric (*Nc'* and *tAI*) by using the topGO

R package (Alexa and Rahnenfuhrer 2021). We obtained a total of 85 non-redundant terms automatically by using REVIGO (Supek et al. 2011) and manually inspecting the GO hierarchy. We only consider a term as statistically significant if it has an FDR adjusted p-value < 0.05 in *Nc'* and *tAI*. The results of these analyses are shown in Figure 5.

We also wanted to know if minimal generation times were significantly different between organisms differing in having (or not) a given GO term enriched in genes showing selected codon usage bias. For this, Wilcoxon one-side signed-rank tests were performed for GO terms under inspection (Supplementary Table S3). The same procedure was done using *tAI'* for the model organisms (*Bacteroides thetaiotaomicron, Bacillus subtilis, Escherichia coli, Leptospira interrogans, Mycobacterium tuberculosis, Salmonella entérica, and Synechocystis sp. PCC 6803*). And as above, a given GO term is considered enriched only if it is significantly enriched at the same time in the metrics *tAI* and *Nc'*.

Ancestral state reconstruction

To study the evolutionary precedence of selected codon usage bias in the translation machinery with respect to other GO terms, we applied the following procedure on the set of 210 genomes. First, we defined two kinds of events: "T" and "G". The "T" event was defined to occur when the "Translation" term (GO:0006412) appeared as significantly enriched in the two metrics (*Nc'* and *tAI*, FDR < 0.05); and the "G" event was defined to occur when a given GO term, other than "Translation", appeared as significantly enriched in the same two metrics. The frequency of these events per genome and GO term is shown in Supplementary Table S4. GO terms without "T" or "G" events were discarded from further analysis. Other GO terms directly linked to "Translation" were removed manually by looking at the GO hierarchy.

Next, events "T" and "G" were converted to binary [0,1] resulting in two vectors: One vector for "T" and another for "G" events. In these vectors, the number "1" represents that the event ("T" or "G") occurred. In the case of having more than one "G" event in different GO terms, we simply assigned a number "1" to the vector of "G" events. Next, we used these vectors as character states at the tips of the phylogenetic tree to infer the ancestral character states at the internal nodes. For this, we applied a maximum likelihood method from the phytools R package that was developed by Yang et al. (1995). This method uses the rerooting function with a symmetric Q matrix (Revell 2012); and a parsimony method by using the function asr_max_parsimony from the castor R package (Louca and Doebeli 2018).

All our scripts are available at https://github.com/PacoMax/CUBs_max.

Results

R_ENC' is less influenced by G+C content than S

We revisited the unified framework proposed by dos Reis et al. (2004) to estimate selected codon usage bias. In particular, we replaced *dNc* by *Nc'* in the Pearson correlation calculations (see Materials and Methods). We used *Nc'* because it is less biased by G+C content and performs better when evaluating codon usage bias (Liu et al. 2018). As expected, (Figure 1A and B), our index *R_ENC'* is less influenced by G+C content than the index (*S*) proposed by dos Reis et al. (2004). Since G+C content is a major determinant of codon usage in the absence of selection (Sharp et al. 2010), we consider the improvement presented here to be significant. For larger values, *R_ENC'* and *S* tend to converge to similar values (Figure 1 C).

As proposed by dos Reis et al. (2004), the genomic landscape is defined by the tRNA gene copy number and genome size (dos Reis et al. 2004) and can be used to study the effect of the above variables on selected codon usage bias. To gain insights on the behavior of the two metrics (*R_ENC'* and *S*) in the context of the prokaryotic genomic landscape, we built simple additive models (corrected and log scaled). As described in Supplementary Figure S1, both models show similar fit to the genomic landscape (*lm* R^2 = 0.246 *p*-value < 2.2e-16 for *S*; and *lm* R^2 = 0.264 *p*-value < 2.2e-16 for *R_ENC'*). However, a more sophisticated analysis like that of dos Reis et al. (2004) and the inclusion of eucaryotes which are out of the scope of this work, would be required to properly account for the correlation between tRNA gene copy number and genome size.

Prokaryotic lineages showing strongest selected codon usage bias

We then explored the distribution of *R_ENC'* values across 1800 genomes from 26 bacterial and archaeal phyla. Four lineages: Gammaproteobacteria, Firmicutes, Actinobacteria, and Betaproteobacteria, showed significantly stronger genome selected codon usage bias than the rest of the taxa (Figure 2, Table S5a, FDR <= 0.05). It is important to mention that these four phyla (Gammaproteobacteria, Firmicutes, Actinobacteria, Betaprotebacteria) together with Alphaproteobacteria, Bacteroidetes/Chlorobi and Delta/epsilon are the most represented in the sample. To investigate whether the above result is due to sampling bias, we repeated the analysis 10 times by randomly selecting 30 genomes from each phylum. The results show that in most repetitions, the same four phyla show significantly stronger genome selected codon usage bias than the rest of the taxa (Table S5b, FDR <= 0.05).

It is worth mentioning that *Escherichia coli* stands as the species showing the strongest *R_ENC'* value, closely followed by the Gammaproteobacteria: *Tolumonas auensis*, *Ferrimonas balearica,* and *Citrobacter freundii*. The first two species are facultative anaerobes while *C. freundii* is an opportunistic pathogen. Within Firmicutes, the species showing the strongest selected codon usage bias is the heterofermentative lactic acid bacterium *Weissella cibaria*; for Betaproteobacteria it is *Neisseria sicca*, an opportunistic pathogen; and *Glutamicibacter arilaitensis* for Actinobacteria, which grows on the surface of Reblochon cheese from France. *Vibrio* stands as the genus most represented among the Gammaproteobacteria showing strong *R_ENC'*; *Neisseria* among Betaproteobacteria; *Lactobacillus* among Firmicutes; and most notably *Corynebacterium* among Actinobacteria. On the other extreme of the distribution, the species showing the weakest *R_ENC'* value was *Blochmannia* endosymbiont of *Camponatus* ants, followed by *Candidatus Kinetoplastibacterium blastocrithidii*, a trypanosomatid endosymbiont.

R_ENC' is correlated to minimal generation time

Minimal generation time (*d*) was shown to correlate with selected codon usage bias in prokaryotes (Rocha 2004; Vieira-Silva and Rocha 2010; Thiele et al. 2011). Here we revisited this result by exploring if *R_ENC'* also correlates with the minimal generation time. We found that *R_ENC'* shows a moderate but statistically significant correlation, even when considering the phylogenetic structure of the data (PGLS' r = 0.32 p-value = 1.9e-6) (Figure 3A).

In addition to *R_ENC'*, we also studied the correlation between (*d*) and other metrics used to estimate selected codon usage bias. These were: *tAIrz*, *ΔtAI ΔENC'*, *ENCr', S dos Reis* and tRNA gene copy number (Table S6). We found that *ENCr'* (Figure 3B) and *ΔENC'* had the

highest correlation with (*d*) (PGLS' r = 0.4, p-value = 6.8e-10; and PGLS' r = 0.4, p-value = 1.2e-9, respectively, while the *S dos Reis* has a value of PGLS' r = 0.048, p-value = 1.38e-3. This is in agreement with Vieira-Silva and Rocha (2010) who reported this by using conventional Spearman correlation. These results confirm previous observations indicating that selected codon usage bias is correlated to minimal generation time. This correlation is stronger when selected codon bias is measured with indexes that use ribosomal protein-coding genes as the reference to estimate selected codon usage bias, like *ENCr'* and *ΔENC'*.

Ribosomal protein-coding genes often show signals of translational selection when R_ENC' indicates genome-wide selected codon usage bias

Ribosomal protein-coding genes tend to be highly expressed and their codon usage usually shows selected codon usage bias. Because of this, they are used by several metrics as gold standards to measure the effects of selection on codon usage. We first asked how often ribosomal protein-coding genes show selected codon usage bias when *R_ENC'* indicates genome-wide selected codon usage bias. This can be done by measuring the frequency with which ribosomal protein-coding genes have a codon usage that is highly adapted to the tRNA gene pool (show large *tAI* values) and using a few different synonymous codons for each amino acid (low *Nc'* values). A paradigmatic example is that of *Escherichia coli* where selection has clearly shaped the codon usage of ribosomal protein-coding genes and a few others like the gene coding for the chaperon GroEL (Figure 4A). There are other unusual cases where ribosomal protein-coding genes show the opposite effect, appearing towards the bottom right of the graph, such as in *Nitrosomonas europaea* and *Syntrophus aciditrophicus* (Kolmogorov Smirnov *Nc'* lower tail p-value < 0.05; *tAI* upper tail p-value < 0.05) (Figure 4B). When selection has not optimized codon

usage bias of ribosomal protein-coding genes these tend to appear scattered around the graph. This is the case of *Buchnera aphidicola* genomes (Figure 4C). We found that among the set of 210 genomes, 147 had their ribosomal protein-coding genes in the upper left quadrant (Kolmogorov Smirnov *Nc'* upper tail p-value < 0.05; *tAI* lower tail p-value < 0.05, Supplementary Table S7).

We also tested for the above association by using logistic regression. The logistic regression showed a moderate association between having the ribosomal protein-coding genes in the upper left quadrant and a strong *R_ENC'* (Mc Fadden's R squared = 0.12, Supplementary Figure S2). In addition, we also found a statistically significant association by using a Pearson correlation between translational selection (*ENC'r*) and *R_ENC'* (r^2 = 0.58, p-value < 2.2e-16). And the same result by using GroEL protein-coding genes and *R_ENC'* (r^2 = 0.43 , p-value < 2.2e-16). These results show that in general terms, ribosomal protein-coding genes show selected codon usage bias when *R_ENC'* indicates selected codon usage bias at the genome level.

Gammaproteobacteria and Firmicutes convergently evolved similar GO terms with selected codon usage bias

We then asked which functional terms were enriched in genes showing selected codon usage bias. For this, we considered a term to be enriched if a statistically significant portion of its genes mapped to the upper left quadrant defined by the *tAI* versus *Nc'* plot. Not surprisingly, translation (GO:0006412) was the term enriched in most genomes (Figure 5). This term is followed by others that are functionally related to translation (like tRNA aminoacylation or gene expression) or by terms related to ATP metabolism. This is not unexpected, since terms related to translation are crucial for fast-growing prokaryotes (Karlin 2001; Klumpp 2013).

Interestingly, we found that in general, Gammaproteobacteria and Firmicutes had more enriched terms than the rest of the taxa, many of which are shared (Figure 5). We interpret the above pattern as an adaptive convergence due to natural selection. Since the same GO terms exists in other phyla but show less enrichment (orange, yellow or gray colors in Figure 5), the alternative explanation (divergence from common ancestry followed by secondary losses of enrichment) seems less likely.

In fact, a heatmap analysis of the matrix shown in Figure 5 shows that several species of Gammaproteobacteria and Firmicutes cluster together in three different but related branches in the dendrogram, thus supporting our interpretation (Figure S4). Of course, not all species from Gammaproteobacteria and Firmicutes showing GO terms enriched in genes showing selected codon usage bias, cluster together. Note that the clade containing *Escherichia coli*, *Shewanella oneidensis* MR-1, *Salmonella enterica* and *Yersinia pestis* cluster with the clades at the bottom of the figure.

On the side of the GO terms, there is a clear cluster formed by GO:0006412 translation, GO:0046034 ATP metabolic process, GO:0006520 cellular amino acid metabolic process, GO:0010467 gene expression, GO:0006418 tRNA aminoacylation for protein translation, GO:0006096 glycolytic process, GO:0046031 ADP metabolic processes. These terms tend to be enriched together and tend to be present in the three clusters described above. However, this association should be interpreted carefully since genes can have more than one GO term, including sharing the ones mentioned above. For instance, in E. coli there are 150 genes annotated with both GO:0006412 translation and GO:0010467 gene expression; 28 sharing

translation and GO: 0006520 cellular amino acid metabolic process; and 26 with translation as well as GO:0006418 tRNA aminoacylation for protein translation.

It is also interesting to see that there is another set of procaryotes clustering together. This includes those species not showing GO terms enriched in genes showing selected codon usage bias. Species within this cluster includes epibionts like *Nanoarchaeum equitans* Kin4-M, ensosymbionts like *Buchnera aphidicola* and *Wigglesworthia glossinidia*, parasites such as *Mycoplasma genitalium*, as well as free living bacteria like *Prochlorococcus marinus* among many others.

We also found anecdotal associations between enriched GO terms and the lifestyle of organisms. For instance, photosynthesis and methanogenesis were enriched in cyanobacteria and methanogenic archaea respectively (Figure 5). And in agreement with previous studies (Martínez-Cano et al. 2015), endosymbiotic bacteria like *Candidatus Blochmannia floridanus*, *Buchnera aphidicola*, *Wigglesworthia glossinidia*, showed a general lack of gene terms enriched for selected codon usage bias.

Selected codon usage bias tends to evolve first in ribosomal protein-coding genes and then in other processes

One of the patterns shown in Figure 5 is that the translation GO:0006412 term is almost always enriched in genes showing selected codon usage bias whenever other GO terms are enriched. Examples of these other GO terms include those related to obtaining energy (glycolytic process GO:0006096; tricarboxylic acid cycle GO:0006099; citrate metabolic process GO:0006101; and ATP metabolic process GO:0046034), as well as niche-specific processes like methanogenesis GO:0015948 in methanogenic archaea and photosynthesis GO:0015979 in cyanobacteria. This

suggests that selected codon usage bias evolved first in genes coding for the translation machinery and subsequently in genes involved in other cellular processes.

To test the above hypothesis, we inferred when selected codon usage bias evolved along the phylogenetic tree for diverse GO terms. For this we defined two kinds of events: "T" and "G". The "T" event was defined to occur when the translation GO term (GO:0006412) appeared as significantly enriched in the two metrics (*Nc'* and *tAI*, FDR < 0.05); and the "G" event was defined to occur when any GO term other than translation, appeared as significantly enriched in the same two metrics as before (see Methods). By this, we ended up with two vectors, one for the "T" event and the other for the "G" event. We then inferred by maximum likelihood and parsimony when in the phylogeny these events evolved. In general, we found that "T" events are more widespread than "G" events. This implies that translation evolved selected codon usage bias earlier than other GO terms ("G" events) (Figure 6). The strong phylogenetic signal of selected codon usage bias (Pagel's $\lambda = 0.99$, p-value < 0.001) supports the above analysis. More interestingly, we found that other GO terms (represented by "G" events) appear to have evolved independently in different lineages. This is particularly conspicuous for Gammaproteobacteria and Firmicutes.

In addition, we compared the *R_ENC'* values, between genomes showing enrichment in the translation GO term and those not showing enrichment in this term (Supplementary Figure S3A). The Wilcoxon test showed us the distributions are not the same (p-value 6.21e-13) and the logistic regression showed a considerable association (Mc Fadden's R squared = 0.34) (Supplementary Figure S3B). This suggests that in genomes where the selected codon bias is intense, the efficiency of the translation machinery is improved.

GO terms associated with short generation times

Microorganisms showing short generation times typically exhibit upregulation of proteins involved in translation, gene expression, and protein synthesis (Scott et al. 2010; Molenaar et al. 2009; Peebo et al. 2015; Zavřel et al. 2019; Mori et al. 2017). As shown above, the GO term of "Translation" is enriched in genes showing codon usage bias in organisms showing short generation times. We, therefore, asked if there are significant differences in the distribution of minimal generation times between microorganisms showing enrichment of selected codon usage bias in the "Translation" and in other GO terms.

GO Terms related to translation, protein folding, gene expression, and processes related to tricarboxylic acid and photosynthesis show significant differences (Supplementary File S3 and Table S3). Minimal generation times tend to be short in organisms having these enriched terms except in the case of photosynthesis where the opposite occurs (p-value < 0.01). This suggests that in some species, terms other than "Translation" contribute to fast cell division.

Using tRNA expression data

As mentioned above, the metrics we used only measure approximately the selected codon usage bias because the abundance of cytoplasmic tRNA is estimated from tRNA gene copy number. To overcome this, Wei used *tpm* (transcripts per million) from RNA-seq experiments instead of tRNA gene copy number in seven organisms (Wei et al. 2019). We used their data to attempt a better estimation of selected codon usage bias for these organisms. On the whole, our analysis showed similar results between *R_ENC'* estimated from *tAI* or *tAI'*. Enrichment of GO terms showed similar patterns using *tAI* and *tAI'* (Figure 7). Yet there were some exceptions: *Synechocystis* and *L. interrogans* showed enrichment in translation related GO terms that were not previously identified using *tAI*.

Discussion

Here we present *R_ENC'*, which is an improvement to the original *S* index developed by dos Reis (2004). Our index is less biased by G+C content than *S*. Similar to other measures of selected codon usage bias, *R_ENC'* correlates with minimal generation time and with selected codon usage bias of ribosomal protein-coding genes.

We used *R_ENC'* to explore the frequency of selected codon usage in 1800 complete genomes from diverse Archaea and Bacteria. We found that genomes from four phyla/subphyla: Gammaproteobacteria, Firmicutes, Actinobacteria, and Betaproteobacteria showed stronger selected codon usage bias than the rest of the lineages. Nonetheless, it is important to consider ascertainment bias, since there are not enough representative genomes from many other phylogenetic lineages to discard that this same effect might be present elsewhere. In addition, we found that several species within Gammaproteobacteria and Firmicutes show a similar set of GO terms enriched in genes under selected codon usage bias, indicating convergent evolution.

Conditions for the evolution of translational efficiency

It has been argued that highly adapted codons are selected because they improve the efficiency and the accuracy of translation (Plotkin and Kudla 2010). Some authors suggest that improving the efficiency of translation is more relevant than improving its accuracy, particularly among fast-growing bacteria (Ran and Higgs 2012). The correlation between minimal duplication time and selected codon usage bias of highly expressed genes coding for ribosomal proteins supports this view (Sharp et al. 2010). Accordingly, highly adapted codons contribute to faster ribosome recycling which in turn increases cellular fitness (Kudla et al. 2009). As mentioned before, this is particularly relevant among fast-growing bacteria.

Yet under which environmental conditions will selection favor short generation times and its associated trait: translation efficiency? A recent study showed that copiotrophic environments are associated with species showing selected codon usage bias as well as other genomic traits like abundant tRNA and rRNA genes, larger genomes, and higher G+C content (Okie et al. 2020). In contrast, organisms growing in oligotrophic environments tend to show the opposite features. Accordingly, selection favors codon usage bias (and the other associated genomic traits) as an adaptation for growing fast when resources are abundant (at least periodically); and favors slow growth and efficient use of resources when resources are constantly scarce. The correlation between minimal generation time and codon usage bias as measured by *R_ENC'* (and other indexes) can be attributed to the opposing effects selection has on genomic traits depending on the growth strategy of cells. And the growth strategy of cells are adaptations to the different kinds of environments that prokaryotes evolve in.

Therefore, *R_ENC'* (and other measures of selected codon usage bias) can serve as gross indicators of the environment on which different lineages of prokaryotes have evolved, an idea that was already proposed by Carbone et al. (2004) and explored more recently by Botzman and Margalit (2011) and Arella et al. (2021).

There are several lines of evidence correlating the lifestyle of organisms to selected codon usage bias. Initially, Rocha (2004) discovered the correlation between selected codon usage bias of highly expressed genes and growth rate. Later on, Sharp (2005) suggested that bacteria living in variable environments would tend to show more often selected codon usage in highly expressed genes than parasitic bacteria. A more in-depth statistical analysis between environment and selected codon usage bias was provided by Botzman and Margalit (2011). These authors showed

that variable "environments" favor selected codon usage of highly expressed genes. Other statistical analyses also showed a correlation between growth rate and the ability of bacteria to survive in different environments (Freilich et al. 2009). This is, fast growers tend to inhabit diverse environments while slow growers inhabit more specialized niches.

Based on the degree of coadaptation between codon usage of highly expressed genes and tRNA genes, as well as other features of the genome like the number of rRNA genes and genome size (components of the genomic landscape) we should be able to predict, with some degree of accuracy, minimal duplication times and the kind of environment on which prokaryotes thrive. This is an idea that has been explored for small genomic sequences derived from metagenomic samples (Vieira-Silva and Rocha 2010). The correlation between selected codon usage bias and growth rate should inform us about the kind of environment (regarding nutrient availability) on which lineages have evolved (Sharp et al. 2010). Because of that, we infer that most lineages of Gammaproteobacteria and Firmicutes studied here adapted convergently to growth in copiotrophic environments.

Evolutionary precedence of translation efficiency

Here we provide phylogenetic evidence indicating that selected codon usage bias tends to evolve first in the genes coding for the translation machinery and afterward in genes coding for other cellular functions. To our knowledge, this is the first time that such phenomena has been reported in the literature. This result is consistent with the hypothesis that the main advantage of highly adapted codons is to free the ribosome from highly expressed genes (under rapid growth conditions) so they can be used to translate more mRNAs. Theoretically, improving the

translation efficiency of genes involved in the translation machinery itself would have a major effect on the translation efficiency of any other gene in the genome.

Concluding Remarks

Darwinian theory predicts that there should be a correlation between heritable phenotypes and environments. However, it is clear that more studies are required to better understand the association between genome features (like codon usage bias) and the physiology, lifestyle, and environment of prokaryotes (Iriarte et al. 2021). Here we provide another step towards better understanding the association between genomic traits and environments. In particular, the potential of selected codon usage bias to inform us about life-history traits such as growth rate and feeding strategy.

Acknowledgments

We are indebted to CONACYT for supporting Francisco González during his master's degree in Integrative Biology in Cinvestav Irapuato (CVU: 856429). We also thank UGA/LANGEBIO for giving access to its HPC (High-Performance Computing Cluster) “MAZORKA” and to LAICBIO for providing computer facilities.

Statements and Declarations

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Francisco González-Serrano. Luis Delaye directed the analyses and Cei Abreu-Goodger helped to evaluate the results. The first draft of the manuscript was

written by Francisco González-Serrano and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

The authors declare no conflicts of interest.

References

Alexa A, Rahnenfuhrer J (2021) topGO: Enrichment Analysis for Gene Ontology. R package version 2.46.0. doi: 10.18129/B9.bioc.topGO

Arella D, Dilucca M and Giansanti A. (2021) Codon usage bias and environmental adaptation in microbial organisms. Mol Genet Genomics 296(3): 751-762. https://doi.org/10.1007/s00438-021-01771-4

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. (2000) Gene ontology: tool for the unification of biology. Nat Genet 25(1): 25-29. doi: 10.1038/75556

Botzman M and Margalit H (2011) Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. Genome Biol 12(10), R109. doi: https://doi.org/10.1186/gb-2011-12-10-r109

Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129(3): 897-907. doi: 10.1093/genetics/129.3.897

Carbone A and Madden R (2005) Insights on the evolution of metabolic networks of unicellular translationally biased organisms from transcriptomic data and sequence analysis. J Mol Evol 61(4): 456-469. doi: 10.1007/s00239-004-0317-z

dos Reis MD, Savva R and Wernisch L (2004) Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 32(17): 5036-5044. doi: 10.1093/nar/gkh834

Freilich S, Kreimer A, Borenstein E, Yosef N, Sharan R, Gophna U and Ruppin E (2009) Metabolic-network-driven analysis of bacterial ecological strategies. Genome Biol 10(6): 1-8. doi: https://doi.org/10.1186/gb-2009-10-6-r61

Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM and Pilpel Y (2018) Codon usage of highly expressed genes affects proteome-wide translation efficiency. Proc Natl Acad Sci U S A 115(21): E4940-E4949. doi: https://doi.org/10.1073/pnas.1719375115

Fuglsang A (2006) Estimating the “Effective Number of Codons”: The Wright Way of Determining Codon Homozygosity Leads to Superior Estimates. Genetics 172(2):1301–1307. https://doi.org/10.1534/genetics.105.049643

Grantham R, Gautier C, Gouy M, Jacobzone M and Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9(1): 213-213. doi: 10.1093/nar/9.1.213-b

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W and Gascuel, O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology 59(3): 307-321. doi: 10.1093/sysbio/syq010

Gustafsson C, Govindarajan S and Minshull J (2004) Codon bias and heterologous protein expression. Trends in Biotechnol 22(7): 346-353. doi: https://doi.org/10.1016/j.tibtech.2004.04.006

Ikemura T (1981a) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol 146(1): 1-21. doi: 10.1016/0022-2836(81)90363-6

Ikemura T (1981b) 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 151(3): 389-409. doi: 10.1016/0022-2836(81)90003-6

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2(1): 13-34. doi: 10.1093/oxfordjournals.molbev.a040335

Iriarte A, Lamolle G, and Musto H (2021) Codon Usage Bias: An Endless Tale. J Mol Evol 89: 589–593. https://doi.org/10.1007/s00239-021-10027-z

Karlin S, Mrázek J, Campbell A and Kaiser D (2001) Characterizations of highly expressed genes of four fast-growing bacteria. J Bacteriol 183(17): 5025-5040.

Klumpp S, Scott M, Pedersen S and Hwa T (2013) Molecular crowding limits translation and cell growth. Proc Natl Acad Sci U S A 110(42): 16754-16759. doi: 10.1073/pnas.1310377110

Koch AL (2001) Oligotrophs versus copiotrophs. Bioessays 23(7): 657-661. doi: https://doi.org/10.1002/bies.1091

Kudla G, Murray AW, Tollervey D and Plotkin JB (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. Science 324:255-258. doi: 10.1126/science.1170160

Lefort V, Longueville JE and Gascuel O (2017) SMS: smart model selection in PhyML. Mol Biol Evol 34(9): 2422-2424. doi: https://doi.org/10.1093/molbev/msx149

Liu SS, Hockenberry AJ, Jewett MC and Amaral LA (2018) A novel framework for evaluating the performance of codon usage bias metrics. J R Soc Interface 15(138): 20170667. doi: https://doi.org/10.1098/rsif.2017.0667

Louca S and Doebeli M (2018) Efficient comparative phylogenetics on large trees. Bioinformatics 34(6): 1053-1055. doi: 10.1093/bioinformatics/btx701

Martínez-Cano DJ, Bor G, Moya A and Delaye L (2018) Testing the domino theory of gene loss in Buchnera aphidicola: the relevance of epistatic interactions. Life (Basel) 8(2), 17. doi: https://doi.org/10.3390/life8020017

Molenaar D, van Berlo R, de Ridder D and Teusink B (2009) Shifts in growth strategies reflect tradeoffs in cellular economics. Mol Syst Biol 5:323. doi: 10.1038/msb.2009.82

Mori M, Schink S, Erickson DW, Gerland U and Hwa T (2017) Quantifying the benefit of a proteome reserve in fluctuating environments. Nat Commun 8(1), 1225. doi: https://doi.org/10.1038/s41467-017-01242-8

Novembre JA (2002) Accounting for background nucleotide composition when measuring codon usage bias. Mol Biol Evol 19(8): 1390-1394. doi: https://doi.org/10.1093/oxfordjournals.molbev.a004201

Okie JG, Poret-Peterson AT, Lee ZM, Richter A, Alcaraz LD, Eguiarte LE, et al. (2020). Genomic adaptations in information processing underpin trophic strategy in a whole-ecosystem nutrient enrichment experiment. eLife 9, e49816. doi: 10.7554/eLife.49816

Orme D (2018) The Caper Package: Comparative Analysis of Phylogenetics and Evolution in R (version 1.0. 1). https://cran.r-project.org/web/packages/caper

Parvathy ST, Udayasuriyan V and Bhadana V (2022) Codon usage bias. Mol Biol Rep. 2022 49(1): 539–565. doi: 10.1007/s11033-021-06749-4

Peebo K, Valgepea K, Maser A, Nahku R, Adamberg K and Vilu R (2015) Proteome reallocation in Escherichia coli with increasing specific growth rate. Mol BioSyst 11(4): 1184-1193. doi: https://doi.org/10.1039/C4MB00721B

Plotkin JB and Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 12(1), 32. doi: https://doi.org/10.1038/nrg2899

Quax TE, Claassens NJ, Söll D and van der Oost, J (2015) Codon bias as a means to fine-tune gene expression. Mol Cell 59(2): 149-161. doi: https://doi.org/10.1016/j.molcel.2015.05.035

Ran W and Higgs PG (2012) Contributions of speed and accuracy to translational selection in bacteria. PloS One 7(12), e51652. doi: 10.1371/journal.pone.0051652

Revell LJ (2012) phytools: An R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3: 217-223. doi: https://doi.org/10.1111/j.2041-210X.2011.00169.x

Rocha EP (2004) Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res 14(11): 2279-2286. doi: 10.1101/gr.2896904

Roller BR and Schmidt TM (2015) The physiology and ecological implications of efficient growth. ISME J 9(7): 1481-1487. doi: https://doi.org/10.1038/ismej.2014.235

Scott M, Gunderson CW, Mateescu EM, Zhang Z and Hwa T (2010) Interdependence of cell growth and gene expression: origins and consequences. Science 330: 1099 – 1102. doi: 10.1126/science.1192588

Segata N, Börnigen D, Morgan XC and Huttenhower C (2013) PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat Commun 4, 2304. doi: 10.1038/ncomms3304

Sharp PM, Bailes E, Grocock RJ, Peden JF and Sockett RE (2005) Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res 33(4): 1141-1153. doi: https://doi.org/10.1093/nar/gki242

Sharp PM, Emery LR and Zeng K (2010) Forces that influence the evolution of codon bias. Philos Trans R Soc Lond B Biol Sci 365(1544): 1203-1212. https://doi.org/10.1098/rstb.2009.0305

Supek F, Bošnjak M, Škunca N and Šmuc T (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. PloS One 6(7), e21800. doi:10.1371/journal.pone.0021800

Thiele I, Fleming R, Que R, Bordbar A and Palsson B (2011) A systems biology approach to the evolution of codon use pattern. Nat Prec. https://doi.org/10.1038/npre.2011.6312.1

UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research 47(D1), D506-D515. doi: https://doi.org/10.1093/nar/gky1049

Vieira-Silva S and Rocha EP (2010) The systemic imprint of growth and its uses in ecological (meta) genomics. PLoS Genet 6(1), e1000808. doi: https://doi.org/10.1371/journal.pgen.1000808

Yang Z, Goldman N and Friday A (1995) Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Systematic Biology 44(3): 384-399. doi: https://doi.org/10.1093/sysbio/44.3.384

Wei Y, Silke JR and Xia X (2019) An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria. Sci Rep 9(1), 3184. doi: https://doi.org/10.1038/s41598-019-39369-x

Wright F (1990) The 'effective number of codons' used in a gene. Gene 87(1): 23-29. doi: https://doi.org/10.1016/0378-1119(90)90491-9

Zavřel T, Faizi M, Loureiro C, Poschmann G, Stühler K, Sinetova M, et al. (2019) Quantitative insights into the cyanobacterial cell economy. eLife 8, e42508. doi: https://doi.org/10.7554/eLife.42508.001

Figure and Table captions

Figure 1. **Spearman correlation between G+C content and selected codon usage bias across 1800 bacterial genomes.** (A) *S* index proposed by dos Reis et al. (2004); (B) *R_ENC'*, the index proposed here. GC/AT% represents the percentage of G+C or A+T, depending which is larger; (C) correlation between *R_ENC'* and *S*. Each dot represents a genome.

Figure 2. **Bacterial phyla showing strongest selected codon usage bias.** Numbers at the top of each boxplot represent the number of genomes per taxa. Stars indicate taxa enriched with lineages showing strongest selected codon usage bias (p-value < 0.05, Wilcoxon tests).

Figure 3. **Spearman correlation between codon usage bias and growth rates.** (A) *R_ENC'* versus minimal generation time *d*(h); and (B) *ENC'r* versus *d*(h).

Figure 4. **Correlation between tAI and NC' showing the evolutionary landscape of selected codon usage bias and the position of the ribosomal proteins.** Axes are normalized to mean 0 and standard deviation 1. Coding genes for ribosomal proteins are in red, GroEL proteins in blue and the rest of the genes in grey. (A) *Escherichia coli str. K-12_substr. MG1655,* (B) *Syntrophus aciditrophicus SB,* (C) *Buchnera aphidicola* str. Bp.

Figure 5. **Phylogenomic tree and heatmap displaying significant GO terms per OTU.** From left to right: Phylogenomic tree; OTU colors indicate different taxa. d(h), green triangle size is proportional to minimal generation time in hours; and purple triangle size is proportional to *R_ENC'*. Heat map: The red color on the heat map indicate GO terms showing significance in both metrics (*tAI* and *Nc'*); orange indicate that only *tAI* was significant; yellow where only *Nc'* was significant; gray indicate none was significant; and white indicate that the GO term is absent.

Figure 6. **Translation evolved selected codon usage bias earlier than other GO terms.** (A) Selected codon usage bias of internal nodes was inferred by maximum parsimony. Nodes where selected codon usage bias was inferred to be present are shown in red and those were not, in gray. Left-side tree: GO term of Translation; and right-side for any other GO term. For clarity, pie charts in internal nodes are larger than those of tips; (B) Frequency of selected codon usage bias versus root to node distance. The distance is measured in number of nodes to the root and

the probability of selected codon usage bias is averaged over all nodes having the same distance to the root; (C) Number of internal nodes showing selected codon usage bias (red) and not showing in (gray) inferred with maximum parsimony and with the likelihood re-rooting method.

Figure 7. **Comparison between tAI and tAI'.** Selected codon usage bias was estimated with *tAI* and *tAI'* for five organisms having RNA-seq data. Brown color indicates that both (*tAI* and *tAI'*) are significant (p-value < 0.05); yellow that only *tAI'* is significant; red that only *tAI* is significant; pink none are significant; and grey indicates there are not genes associated to the corresponding GO term.

Supplementary material

Supplementary figure S1. ***R_ENC'* resembles S dos Reis in the Genomic Landscape.** Correlation between (A) the genomic landscape (tRNA copy number and genome size) and S dos Reis; (B) the genomic landscape and *R_ENC'*; (C) Statistics related to the multicollinearity of the models.

Supplementary figure S2. **Logistic correlation between codon usage bias detected in ribosomal proteins and *R_ENC'* per genome**. Selected codon usage bias of ribosomal proteins were defined as "1", if ribosomal protein coding genes map to the upper left quadrant.

Supplementary figure S3. **Selected codon usage bias in translation and *R_ENC'*.** (A) Box plots of Translation Gene Ontology enriched using different metrics and *R_ENC'*. (B) Logistic correlation between selected codon usage bias translation GO term enrichment and *R_ENC'*.

Supplementary figure S4. **Heatmap analysis of GO terms showing enrichment in genes with selected codon usage bias**. The matrix shown in Figure 5 is transformed by the following rules: all red colors are transformed to 1 (indicating enrichment in genes showing selected codon usage

bias); and all other colors (orange, yellow and brown) are transformed to 0 (indicating lack of enrichment in genes showing selected codon usage bias). White colors are treated like NA. The heatmap() R function is used to make the heatmap analysis on the transformed matrix.

Supplementary table S1. **RefSeq IDs, codon usage bias metrics, and genomic features (G+C content, tRNA copy number, genome size) from 1800 representative prokaryotic genomes.** *FTP*, RefSeq IDs; *Name*, organism name; *Taxa*, taxonomy; *Size*, Genome size; *GC%*, GC content; *CDS*, number of protein coding sequences; *link*, link to download; *S*, S dos Reis; *S p-value*, p-value of S dos Reis calculated using Montecarlo method; *R_ENC'*, R ENC'; *R_ECN' p-value*, p-value of R ECN' calculated using Montecarlo method; *mean_tAI*, arithmetic mean of tAI; *mean_dNc*, arithmetic mean of dNc; *mean_Nc*, arithmetic mean of Nc; *mean_Nc'*, arithmetic mean of Nc'; *sd_tAI*, standard deviation of tAI; *sd_dNc*, standard deviation of dNc; *sd_Nc*, standard deviation of Nc; *sd_Nc'*, standard deviation of Nc'; *trnacn*, number of tRNAs.

Supplementary table S2. **RefSeq IDs, codon usage bias metrics, and genomic features (G+C content, tRNA copy number, genome size), results of gene set enrichment analysis, and minimal generation times from 210 genomes obtained from Vieira-Silva, S. & Rocha, E. P. work (2010).** *FTP*, RefSeq IDs; *Name*, organism name; *Taxa*, taxonomy; *Size*, Genome size; *GC%*, GC content; *CDS*, number of total protein coding sequences; *link*, link to download; *S*, S dos Reis; *S p-value*, p-value of S dos Reis calculated using Montecarlo method; *R_ENC'*, *R_ENC'*; *R_ECN' p-value*, p-value of *R_ECN'* calculated using Montecarlo method; *mean_tAI*, arithmetic mean of *tAI*; *mean_dNc*, arithmetic mean of *dNc*; *mean_Nc*, arithmetic mean of *Nc*; *mean_Nc'*, arithmetic mean of *Nc'*; *sd_tAI*, standard deviation of *tAI*; *sd_dNc*, standard deviation of *dNc*; *sd_Nc*, standard deviation of *Nc*; *sd_Nc'*, standard deviation of *Nc'*; *ENCr'*, Nc' mean of ribosomal genes; *tAIrz*, *tAI* mean of ribosomal genes after z normalization; *tAIr*, *tAI* mean of ribosomal genes; *trnacn*, number of tRNAs; list of GO terms: 1 for significant *tAI* value, 2 for

significant *Nc'* value, 3 for both significant and 0 for none; *timeg*, minimal generation time d(h); *ref*, reference of the minimal generation time obtained from Vieira-Silva & Rocha, 2010.

Supplementary table S3. **Wilcoxon signed-rank test results from distributions of minimal generation times per GO term.** Columns: *GO*, GO term; *Num*, total number of organisms with the genes associated to the GO term; *Num_CUB*, number of organisms which have significant selected codon usage bias in the GO term; *Num_non*, number of organisms which don't have significant selected codon usage bias in the GO term; *wilcox_greater*, p-value of the wilcoxon right tail test; *wilcox_lower*, p-value of the wilcoxon left tail test; *FDR_greater*, p-value adjusted by FDR method of the wilcoxon right tail test; *FDR_lower*, p-value adjusted by FDR method of the wilcoxon left tail test.

Supplementary table S4. **Relative frequencies, dependent and independent relative frequencies of the "T" and "G" events**. See materials and methods for definition of "T" and "G" events. Columns: *GO*, GO term; *Num Org*, Total of organisms with the GO term. Observed frequencies of the different events: *freq(0t)*, no occurrence of "T" event; *freq(1t)*, occurrence of "T" event; *freq(0g)*, no occurrence of "G" event; *freq(1g)*, occurrence of "G" event. Observed dependent frequencies of the different events: *freq(0t0g)*, dependent no occurrence of "G" given the no occurrence of "T"; *freq(0t1g)*, dependent occurrence of "G" given the no occurrence of "T"; *freq(1t0g)*, dependent no occurrence of "G" given the occurrence of "T"; *freq(1t1g)*, dependent occurrence of "G" given the occurrence of "T"; *freq(0g0t)*, dependent no occurrence of "T" given the no occurrence of "G"; *freq(0g1t)*, dependent occurrence of "T" given the no occurrence of "G"; *freq(1g0t)*, dependent no occurrence of "T" given the occurrence of "G"; *freq(1g1t)*, dependent occurrence of "T" given the occurrence of "G"; *freq(T)*freq(GO)*, estimated frequency if the events are independent; *freq(T n GO)*, estimated frequency if the events are dependent.

Supplementary table S5. **Wilcoxon signed-rank test results from the distributions of R_ENC' per taxa.** Columns: *Taxa*, phylum/subphylum level; *p-value*; *FDR*, FDR correction; *Bonferroni*, Bonferroni correction.

Supplementary table S6. **Spearman and PGLS correlation between different codon usage bias metrics and growth rates.** Columns: *ρ2*, rho squared of the Spearman correlation; *(r2) pgls*, R^2 of the PGLS.

Supplementary table S7. **Comparisons of selected codon usage bias in ribosomal protein genes and *R_ENC'*.** *Name*, name of the organism; *total_genes*, number of total protein coding sequences following the criteria for selected codon usage bias estimation (see materials and methods); *total_RP_genes*, number of total of ribosomal protein genes; *ks_greater_RP_ncp*, p-value from the right tail of kolmogorov smirnov test of Nc'; *ks_lower_RP_ncp*, p-value from the left tail of kolmogorov smirnov test of Nc'; *ks_greater_RP_tai*, p-value from the right tail of kolmogorov smirnov test of tAI; *ks_lower_RP_tai*, p-value from the left tail of kolmogorov smirnov test of *tAI*; *R_ENC' p-value from Montecarlo test*, *R_ENC' significant*: * if R_ENC' p-value from Montecarlo test is < 0.05; *RP_genes_upper_left_quadrant significant*: * if *ks_lower_RP_ncp* and *ks_greater_RP_tai* are < 0.05; *R_ENC' & RP's CUBs significant*: * if *ks_lower_RP_ncp*, *ks_greater_RP_tai* and *R_ENC' p-value from Montecarlo test* are < 0.05.

Supplementary File 1. Correlations between *tAI* and *NC'* showing the evolutionary landscape of selected codon usage bias and the position of the ribosomal proteins of the 210 prokaryotic genomes.

Supplementary File 2. GSEA raw results per organism.

Supplementary File 3. Distributions of minimal generation times between significant and non-significant GO terms.

Supplementary Files 1, 2 and 3 are available at: https://github.com/PacoMax/CUBs_max