The distribution, conservation and correlation with translational efficiency of the first five codons in *Escherichia coli* genes

Mariana H. Moreira[1#], Géssica C. Barros[1#], Rodrigo D. Requião[1#], Silvana Rossetto[2], Tatiana Domitrovic[3] and Fernando L. Palhano[1*]

[1] Programa de Biologia Estrutural, Instituto de Bioquímica Médica Leopoldo de Meis, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, 21941-902, Brazil.

[2] Departamento de Ciência da Computação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, 21941-902, Brazil.

[3] Departamento de Virologia, Instituto de Microbiologia Paulo de Góes, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 21941-902, Brazil.

* To whom correspondence should be addressed: Tel.: +55 21 3938-6761; Email: palhano@bioqmed.ufrj.br

# These authors contributed equally.

Keywords: translational ramp, translational efficiency, bacteria, ribosome profiling

ABSTRACT

Translation initiation is a critical step in the regulation of protein synthesis, and it is subjected to different control mechanisms, such as 5' UTR secondary structure and initiation codon context, that can influence the rates at which initiation and consequentially translation occurs. For some genes, translation elongation also affects the protein synthesis rate. Recently, it was proposed that the identity of codons three to five, called short translational ramp, have a strong influence on translation elongation and protein expression. By the use of a GFP library where nearly all combinations of nucleotides at these positions were created, it was demonstrated that some of nucleotides combinations increased GFP expression up to four orders of magnitude by enhancing their translation efficiency (TE). While it is clear that the short ramp can influence protein expression levels of artificial constructs, its impact on physiological proteins is still unknown. In this work, we aimed to investigate the relevance of the short translational ramp on a physiological context. Through bioinformatics analysis, we identified the nucleotide combinations from the GFP library on *Escherichia coli* genes and examined their correlation with TE. We observed that *E. coli* genes were enriched with nucleotide compositions that enhanced protein expression on the GFP library, but, surprisingly, it seems to affect the TE only marginally. Nevertheless, our data indicate that different enterobacteria present similar nucleotide composition enrichment as *E. coli*, suggesting an evolutionary pressure towards the conservation of the short translational ramp.

INTRODUCTION

Translation rate is generally determined by translation initiation rate[1]. Translation efficiency (TE), that is the relative number of protein molecules synthesized per equally transcribed mRNAs, is mainly governed by translation initiation, that in turn depends primarily on the sequence of mRNA 5' UTR[2]. However, it was shown that the expression levels of some proteins are a resultant of both translation initiation and translation elongation rates[3]. Elongation plays a role in TE when the initiation is very fast at mRNAs carrying specific features, e.g., excess of non-optimal codons[2,3]. Previous reports showed that non-optimal codons are concentrated after the start codon in several genes of several species, including *Saccharomyces cerevisiae* and *Escherichia coli*[4,5]. This was called the low-efficiency ramp, or translational ramp, and it was suggested that this ramp was necessary to keep low rates of translation elongation at the beginning of protein synthesis, which would avoid ribosome jams and minimize the cost of protein translation[4]. This hypothesis, however, has been challenged by several subsequent papers. One strong criticism of the work is the use of translation inhibitors, which is now known to distort ribosome profiling data, especially at the N-terminal[6,7,8,9,10]. By ribosome profiling analyses of *S. cerevisiae* performed in the absence of such inhibitors, we have already demonstrated that only a small number of genes with sequences containing an extremely high concentration of non-optimal codons possess lower elongation rates[11]. Other studies claim that translation elongation rates can only enhance protein expression to a small extent, and that, even though it is located inside the open reading frame, the low-efficiency ramp affects translation initiation rates primarily[12,13]. Despite the many examples of connections between the role of translation elongation and protein expression levels, there is currently no clear understanding of the mechanisms by which translation elongation exerts translational control[1,14,15].

Very recently, Djuranovic's group reported the effect of the five first codons on translation efficiency in *E. coli*, which was called short translational ramp[16]. A GFP library was built where the fluorescence levels of different GFP proteins, with virtually all possible nucleotide compositions from codons 3 to 5 (262,144), were analyzed, giving each GFP molecule a score (based on its fluorescence level). Interestingly, some nucleotide compositions enhance the magnitude of protein abundance up to 4 orders, without interfering with their mRNA levels; in other words, the nucleotide composition of the beginning of the mRNA coding sequence drastically affects the TE. Interestingly, no correlation was found between the tRNA adaptation index, that reflects the optimality of codons[16], and the GFP score. It means that, at least in the GPF reporter, the presence of non-optimal codons close to the ATG does not improve TE. The mechanism behind this enormous variation in TE relies on the pausing of the ribosome followed by the abortion of protein synthesis. This process was affected by nascent amino-acid-sequence composition coded at specific mRNA positions (codons 3 to 5). Based on the GPF library scores, the authors found that the motifs AAVATT and AADTAT (V = A, G or C, and D = G, A or T), when present specifically at codons 3 and 4, increased the protein abundance of several reporters. The authors showed that this phenomenon was not restricted to GFP and that it is present in other protein reporters, independent of the expression vector used, or *in vitro* and *in vivo* expression conditions.

However, all experiments were done by use of artificial reporters, meaning that the biological significance and conservation of this short translational ramp in *E. coli* genome were not determined.

Our main aim was to verify the biological relevance of the GFP score obtained with the GFP library. Through bioinformatics, we searched for each nucleotide composition measured by the GFP library in the *E. coli* genome, attributing the respective GFP score to each gene. We observed that there is a bias towards nucleotide compositions at codons 3 to 5 with higher scores when compared to a scrambled genome. Interestingly, 91% of *E. coli* genes have a unique nucleotide composition at these positions. The correlation matrix between the score obtained herein vs. different parameters, such as translation efficiency, protein abundance, mRNA abundance, and mRNA stability showed no correlation. However, when just a group of genes with short 5'UTR (<25 nucleotides) was analyzed, minimizing the influence of possible regulatory mechanisms at ribosome binding sites, we observed a Spearman correlation ranging from -0.16 to 0.33, depending on the parameters analyzed. Using the TE data measured by ribosome profiling from three independent experiments, we observed that genes with high GFP scores possess higher TE values when compared with genes with low scores. The ribosome occupancy at codon positions 3 to 5 was lower when compared to the genome average, suggesting that this short ramp modulates the translation speed. Finally, the evolutionary conservation of the short translational ramp was investigated using three different enterobacteria: *E. coli*, *Klebsiella oxytoca*, and *Enterobacter asburiae*. All enterobacteria analyzed possess a bias for genes with high scores at codon positions 3 to 5. The top scored genes of each of the three enterobacteria have a higher frequency of orthology than that expected randomly. We discuss the distribution and the impact on translation efficiency of the short translational ramp of endogenous bacterial genes.

RESULTS AND DISCUSSION

Biological distribution of the short translational ramp.

Previously, a library of the eGFP genes with the insertion of 9 random nucleotides after the second codon was created by Djuranovic's group[16]. The global expression of the eGFP library was FACS sorted into five bins, normalized by DNA amount and sequenced. A high GFP score indicates sequences with high GFP expression and vice versa. A GFP score of 259,134 unique sequences was calculated, which covers 98.8% of all possible nucleotide combinations at codon positions 3-5 (4^9 = 262,144). A cut-off of 10 reads was used, and a Pearson correlation of 0.79 between the replicates was found[16]. For our analyses, the sequences with stop codons were eliminated, and just sequences present in both replicates were used (182,289 sequences, Pearson correlation = 0.70, Figure 1A). A linear fit equation with Q = 1% was used to eliminate the outliers (29,945 sequences), and the geometrical mean of the replicates was used to score 152,344 unique nucleotide sequences (Figure 1A). The frequency of distribution of those sequences (GFP library) shows a non-Gaussian distribution with a main peak and median score of 3.10 (Figure 1B, green line). We investigated whether this observation could be expanded to endogenous *E. coli* genes. Using the GFP score dataset described before (Figure 1A), we

developed a program that was able to screen the sequence of a given *E. coli* gene and calculate the score at different codons positions. First, we analyzed the score at codons 3-5, the same positions used in the GFP library. Using this approach, we attributed a score for 2,591 genes, which covers 64% of the *E. coli* genome (Figure 1B, red line). Similar results were obtained by Djuranovic's group (personal communication). Different from the GFP library, the density histogram of the *E. coli* genome contains two main peaks and a median score of 3.23, showing a bias towards higher scores (Figure 1B). We created a scrambled version of each mRNA sequence maintaining the proportions of all 61 amino acid codons (genomic codon bias), but randomizing the codon choice for each transcript, and no score bias was observed (Figure 1B, black line). The score bias was position dependent since it was also absent when the analyses were made using the codon positions 9-11 (Figure 1B, orange line).

To address the effect of mRNA composition on score bias observed for the *E. coli* genome, we performed the same analyses described before, but shifting the window of analyzed nucleotides (+1, +2 and +3, comprising nucleotides 8-16, 9-17 and 10-18 respectively). We observed the reduction, but not complete elimination, of bias for sequences, analyzed out of frame (Figure 1C, +1 and +2, lines orange and black, respectively). When the frame was restored (+3, nucleotides 10-18), the bias was partially restored (Figure 1C, green line). It suggests that mRNA composition plays a role in score bias of the *E. coli* genome. Then, we analyzed the influence of amino acids composition on score bias. We maintained the proportions of all 61 amino acid codons as well as the final protein sequence but randomized the codon choice for each transcript of *E. coli* (Figure 1C, blue line). We observed that the scoring bias exists using this scramble genome, suggesting that the amino acids composition also contributes to *E. coli* score bias. We conclude that *E. coli* genes have a bias towards nucleotide compositions at codons 3-5 that in the GFP library leads to higher expression of the GFP reporter. The bias is position dependent, and both amino acid sequence composition and local mRNA sequence are important forces acting to keep this bias.

The GFP library from Djuranovic's group analyzed 259,134 different nucleotide compositions, which is about sixty-five times the number of *E. coli* proteins. This means that each *E. coli* gene can have a different nucleotide composition at codons 3-5, and repeated compositions would only exist if some kind of pressure were acting on its selection. Based on this, we asked ourselves if this observed high score bias was caused by the enrichment of a few conserved high scoring nucleotide sequences or by a multitude of unique random sequences. We observed that the vast majority (91%) of *E. coli* genes have a unique nucleotide composition at codon positions 3-5 (Figure 2A). Only the remaining 9% of genes possess a repeated nucleotide sequence, where the same sequence is present in at least two different genes (occurrence > 1). These genes (occurrence > 1) are enriched in adenine at nucleotide positions 7 and 8 (Figure 2B). The most common sequence shared by *E. coli* genes is GTC ATC TGG; this sequence was found in 10 orthologous genes known as transposase InsH. The average score of repeated nucleotide sequences was higher when compared to unique sequences (Figure 2C). In order to see if the higher scores of genes with repeated sequences could reflect a higher efficiency on protein synthesis, we took advantage of ribosome profiling experiments datasets[17]. This methodology is based on the deep sequencing of ribosome-protected mRNA fragments and reflects, hence, how often a particular mRNA is translated[17]. The ratio of ribosome-protected footprint read counts to total mRNA

read counts for each gene is a measure of the relative rate of translation, defined as the TE for that gene[17]. When the TE of genes with repeated nucleotide sequences was compared to the genome, no statistical difference was observed (Figure 2D). We conclude that regardless of the *E. coli* genome bias towards high score nucleotide composition, only a few genes share the same nucleotide composition. The repeated nucleotide sequences seem to have no particular biological advantage regarding TE (Figure 2D).

Biological consequences of the short translational ramp score in *E. coli* protein translation.

Next, we compared the score of each *E. coli* gene calculated at codon positions 3-5 (SCORE GENOME) with independent metrics used to estimate TE[18-20], protein abundance[19,21], mRNA abundance[20,22-24] and mRNA secondary structure (Gini index)[20]. As a control, we used an *E. coli* scramble genome that maintained the proportions of all 61 amino acid codons but randomized the codon choice for each transcript (SCORE SCRAMBLE). Three main clusters were found. The cluster number 1 is formed by the Gini index, which has a negative correlation with the other parameters (Figure 3A). This was already expected, as a negative correlation between the Gini index and translation efficiency has been recently shown[20]. The second cluster was formed by TE, protein abundance, and mRNA levels that have a high positive correlation between themselves (Figure 3A). SCORE GENOME and SCORE SCRAMBLE form the cluster 3, with little to no correlation with the other parameters analyzed (Spearman correlation ranging from -0.08 to 0.13) (Figure 3A). This lack of correlation, however, might be due to diverse control mechanisms that may be present, such as the binding of regulatory proteins to the 5' UTR of the mRNA molecules. It is well known that translation initiation is an important parameter that controls the translation efficiency[25-27]. In bacteria, ribosome binding sites and other regulatory RNA sequences are active control elements for translation initiation[28]. The optimal position and length of the Shine-Dalgarno sequence appeared to be essential for efficient initiation of translation. For example, sequestration of ribosome binding site by mRNA secondary structure negatively affects the initiation of translation[17,28]. To minimize the influence of these possible regulatory mechanisms, we repeated the analyses described in panel A of Figure 3 with genes with short 5' UTR (<25 nucleotides) [29]. We observed a Spearman correlation ranging from -0.16 to 0.33, depending on the parameters analyzed, just for SCORE GENOME (Figure 3B). Interestingly, the best positive correlations observed for SCORE GENOME were with TE and protein abundance from different datasets (Figure 3B).

We wondered if a less restrictive analysis could be used to study the effect of the short translational ramp on translation efficiency. The GFP score bias observed in the *E. coli* genome starts with genes with scores higher than 4.2 (Figure 4A, green area). So, we decided to split the *E. coli* genes into two groups, one formed for genes with scores higher than 4.2 and other with scores lower than 4.2. The translation efficiency of these two groups of genes was compared using three different datasets[18-20]. Regardless of the TE dataset utilized, the group of genes with higher scores have higher TE when compared to genes with lower scores (Figure 4B-D). The same was true when the protein abundance was compared between low score vs. high score genes (Figure 4E). Gene ontology analysis revealed

that genes with higher scores were enriched in cellular component classes related to cytoplasmic part (data not shown). Sequences logos for GFP reporters and *E. coli* genes, both with scores higher than 4.2, showed a similar nucleotide composition between them. The preference for adenine at nucleotide positions 7 and 8 were more evident in *E. coli* genes (Figure 4F).

Using a motif-scanning approach, Djuranovic's group found two motifs in GFP variants with scores over 4[16]. The motifs AAVATT (V = A, G or C) or AADTAT (D = G, A or T) when present in several reports at codon positions 3-4 displayed higher protein expression levels than wild type constructs. We screened the *E. coli* genome searching for these two motifs and found 41 genes (Figure 5A). As a control, we performed the same screening, but this time, the codon positions 4-5 were used as a criterion, and 34 genes were found (Figure 5A). The 41 genes with one of these two motifs have higher scores when compared to the *E. coli* genome (Figure 5B). Interestingly, genes with these motifs at codon positions 3-4 have higher TE when compared to the genome, but the scenario is different when the motifs are positioned at codons 4-5 (Figure 5C). These data agree with the data obtained by the use of reporters and western blot measurements from Djuranovic's group[16].

To verify if ribosome accumulation is implicated in the ramp formation, we compared the average ribosome footprint count per nucleotide of two groups of genes, genes with scores higher than 4.2 versus scores lower than 4.2, using ribosome profiling data. Some protocols, especially those that include antibiotics, can introduce artifacts in the ribosome profiling data[6-9] and the best practices for generating ribosome profiling libraries in bacteria are still in debate[30]. So, we decided to use ribosome profiling data obtained by two distinct protocols, with flash frozen or filtered cells; both protocols were avoiding the use of antibiotics[30]. We normalized the coverage within the same transcript by dividing each nucleotide read by the sum of the total number of reads for each gene. Independent of the protocol used, the group of genes with higher scores (>4.2) showed less ribosome accumulation at codons 3-5 (nucleotides 7-15) when compared to the genes with lower scores (<4.2) (Figure 6A-B). These data suggest that the translation of the stretch of codons with high score occurs with high efficiency.

Evolutionary conservation of the short translational ramp.

Finally, the evolutionary conservation of the short translational ramp was investigated. Three enterobacteria were analyzed, namely, *Escherichia coli*, *Klebsiella oxytoca*, and *Enterobacter asburiae*. We used the *Homo sapiens* genome as a control since it is a phylogenetically distant species from bacteria. Figure 7A shows that, while the three bacteria analyzed have a bias to high score genes, the density histogram of *H. sapiens* was identical to the GFP library, without any bias. Each one of these bacteria shares in common 1,595 orthologous genes, comprising approximately 40% of each genome. To exemplify the calculation used herein to address the conservation of short ribosomal ramp, we illustrate three groups of orthologous, genes α, β and γ, representing L-threonine 3-dehydrogenase, sensory histidine kinase QseC and RNA-binding protein YhbY genes, respectively (Figure 7B). The scores at codon positions 3-5 and 9-11 of all 1,595 orthologous genes were determined, and a

correlation matrix of the scores of each gene was built (Figure 7C). The best correlations were found when the score calculation was performed at codon positions 3-5 (Figure 7C). Also, we measured the standard deviation value of scores from each group of three orthologous genes and found less variation when the score was calculated at codon positions 3-5 (Figure 7D). For example, 231 groups of orthologous genes have precisely the same score (standard deviation = 0) at codon positions 3-5, while for codon positions 9-10 the number of genes drops to 151 (Figure 7D, dotted square).

To investigate if orthologous genes are enriched with high GFP scoring sequences, we ranked, the genes of *E. coli*, *K. oxytoca*, and *E. asburiae* in three distinct lists, according to the score obtained for codon positions 3-5. The top 500 genes with higher scores of each list of each bacteria (1500 genes in total per analysis) were evaluated in OrthoVenn2 sever, which computes the orthologous gene clusters among multiple species[31]. The number of orthologous genes found was 351 (23%), a number 5.3 fold higher than the scramble list (Figure 7E, compare the first hatched bar with the first red bar). As a control, a similar strategy was followed, but, this time, the score used to rank them was obtained for codons at positions 9-11. In this scenario, the number of orthologous genes dropped to 210 (14%).

CONCLUSIONS

Our data suggest an evolutionary pressure selecting some nucleotide compositions at codon positions 3 to 5 since it affects the TE and protein abundance in a group of enterobacterial genes. On the other hand, our data do not point to a clear correlation between these nucleotide compositions and different cellular parameters, meaning that, at least in a physiological condition, other features of mRNA, such as cis-regulatory elements at 5' and 3' UTR, codon usage, mRNA secondary structure, among others, play a more relevant role on translation efficiency. We do not underscore the potential applications driven by Djuranovic's discovery[16]. The strong effect of some motifs at codon positions 3-5 might have an enormous biotechnology impact and could be used to improve the protein syntheses in recombinant systems. Moreover, the mechanism behind the effect of mRNA composition and protein nascent chain at the beginning of translation impacting ribosome processivity is an open question raised by Djuranovic's study. It will also be interesting to test in the future the impact of codon composition at positions 3-5 on protein synthesis of some endogenous bacterial genes.

MATERIALS AND METHODS

Data	Source	Identifier
GFP score	Verma *et al*., 2019	
Translation efficiency	Morgan *et al*., 2018	
Translation efficiency	Li *et al*., 2014	
Translation efficiency	Burkahardt *et al*., 2017	
Gini index	Burkahardt *et al*., 2017	
Protein average	PaxDb: Protein Abundance Database	
Protein abundance (APEX)	Arike *et al*., 2012	
Protein abundance (iBAQ)	Arike *et al*., 2012	
Protein abundance (em PAI)	Arike *et al*., 2012	
Protein abundance (APEX)	Lu *et al*., 2007	
Protein synthesis rate	Li *et al*., 2014	
mRNA abundance	Burkahardt *et al*., 2017	
mRNA abundance	Blattner *et al*., 2003	
mRNA abundance	Covert *et al*., 2004	
mRNA abundance	Corbin *et al*., 2003	
Ribosome profiling frozen/$MgCl_2$	Mohammad *et al*., 2019	SRR7759812
Ribosome profiling filter/$MgCl_2$	Mohammad *et al*., 2019	SRR7759808
Software and algorithms		
Motifs finder	This paper	https://github.com/mhoyerm/Codons_finder
GFP score assigner	This paper	https://github.com/mhoyerm/Score
Genome mixer	This paper	https://github.com/mhoyerm/Mixer
Codon shuffler	This paper	https://github.com/mhoyerm/Shuffler

Data sources

Coding sequences and annotation of *Escherichia coli*, *Homo sapiens*, *Klebsiella oxytoca*, and *Enterobacter asburiae* were obtained from the Ensemble Genomes (http://ensemblgenomes.org/). The list of *E. coli* genes with short 5' UTR was obtained at RegulonDBdatabse (http://regulondb.ccg.unam.mx/)[29].

Statistical analyses, correlations, and raw data

The raw data used to create Figures 1-4 and for the statistical analyses, including sample size, p-value, D'Agostino & Pearson normality test calculations, Kolmogorov-Smirnov test, when used, are presented in the Supplementary Tables (1-5). For Figure 1A, the best fit equation found was line equation; the outliers were identified with Q=1% with a confidence level of 95%. All statistical analyses were performed with GraphPad Prism 7 software.

Gene ontology

The gene ontology analyses were performed in the Gene Ontology Consortium (http://geneontology.org/).

Motifs finder

In the codon positions 3 and 4 or 4 and 5 of an open reading frame genomic sequence, the algorithm searches for sequences of AADTAT or AAVATT, where D is any nucleotide except cytosine, and V is any nucleotide except thymine. When the target sequence is found, the respective gene name and the six nucleotides sequence is written in the output file.

GFP score assigner

In a defined window codon of an open reading frame from a genomic sequence, the algorithm finds the respective GFP score[16] for the sequence of nine nucleotides, and generates an output file containing the gene name, the sequence found, and its respective GFP score.

Shuffled genome

Data from the original genome were compared to that from a scrambled genome. For that purpose, we designed and implemented two new algorithms that exchanged codons with their synonymous codons.

Genome mixer: An algorithm that handles the input files by keeping the first codon (ATG), mixing the positions of the following codons within a genomic sequence, keeping the ratio of nucleotide bases. The random genome generated by this algorithm preserves the number of copies of each codon.

Codon shuffler: An algorithm that changes codons in a genomic sequence, maintaining the correspondent amino acid and ratio of nucleotide bases. The random genome generated by this algorithm preserves the number of copies of each codon as well as the amino acid sequence of the original genome.

Ribosome profiling data

E. coli ribosome profiling data were treated as described previously[7,17,18]. The data were analyzed as described by Ingolia and collaborators[17] except that the program used here was Geneious R11 (Biomatter Ltd., New Zealand) instead of the CASAVA 1.8 pipeline. The data were downloaded from GEO, and the adaptor sequence (CTGTAGGCACCATCAAT) was trimmed. The trimmed FASTA sequences were aligned to *E. coli* ribosomal and noncoding RNA sequences to remove rRNA reads. The unaligned reads were aligned to the *E. coli* genome. First, we removed any reads that mapped to multiple locations. Then, the reads were aligned to the *E. coli* coding sequence database allowing two mismatches per read. We normalized the coverage within the same transcript by dividing each nucleotide read by the sum of the number of reads for each gene.

Clustering Analysis

All clustering analyses were performed by the Euclidean distance using the Orange 3 software[32].

Orthologous Analysis

The list of 1595 orthologous genes of *Escherichia coli*, *Klebsiella oxytoca*, and *Enterobacter asburiae* was obtained from OrthoVenn2[31].

DATA AVAILABILITY

The algorithms used in this work are available in the GitHub repository (https://github.com/mhoyerm).

ACKNOWLEDGEMENT

We thank Rafael S. Rocha, Claudio A. Masuda e Maite Vaslin for helpful discussions. We thank Francisco de Oliveira for critically reading the manuscript. We also thank Sergej Djuranovic for helpful discussions and for sharing unpublished data.

FUNDING

This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo a Pesquisa do Estado do Rio de Janeiro (FAPERJ) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest with the contents of this article.

References

1. Choi J, Grosely R, Prabhakar A, Lapointe CP, Wang J, Puglisi JD. (2018). How Messenger RNA and Nascent Chain Sequences Regulate Translation Elongation. Annu Rev Biochem. 87:421–49.
2. Evfratov SA, Osterman IA, Komarova ES, Pogorelskaya AM, Rubtsova MP, Zatsepin TS, Semashko TA, Kostryukova ES, Mironov AA, Burnaev E, *et al*. (2016). Application of sorting and next generation sequencing to study 5´-UTR influence on translation efficiency in *Escherichia coli*. Nucleic Acids Research. 45:3487–502.
3. Chu D, Kazana E, Bellanger N, Singh T, Tuite MF, Haar von der T. (2014). Translation elongation can control translation initiation on eukaryotic mRNAs. EMBO J. 33:21–34.
4. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. (2010). An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. Cell. 141:344–54.
5. Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, Ziv-Ukelson M. (2011). Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol. 12:R110.
6. Requião,R.D., Fernandes,L., de Souza,H.J.A., Rossetto,S., Domitrovic,T. and Palhano,F.L. (2017) Protein charge distribution in proteomes and its impact on translation. PLoS Comput. Biol., 13, e1005549.

7. Requião,R.D., de Souza,H.J.A., Rossetto,S., Domitrovic,T. and Palhano,F.L. (2016) Increased ribosome density associated to positively charged residues is evident in ribosome profiling experiments performed in the absence of translation inhibitors. RNA Biol., 13, 561–568.
8. Weinberg,D.E., Shah,P., Eichhorn,S.W., Hussmann,J.A., Plotkin,J.B. and Bartel,D.P. (2016) Improved Ribosome-Footprint and mRNA measurements provide insights into dynamics and regulation of Yeast translation. Cell Rep., 14, 1787–1799.
9. Gerashchenko,M.V. and Gladyshev,V.N. (2014) Translation inhibitors cause abnormalities in ribosome profiling experiments. Nucleic Acids Res., 42, e134.
10. Hussmann,J.A., Patchett,S., Johnson,A., Sawyer,S. and Press,W.H. (2015) Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in Yeast. PLoS Genet., 11, e1005732.
11. Carneiro RL, Requião RD, Rossetto S, Domitrovic T, Palhano FL. (2019). Codon stabilization coefficient as a metric to gain insights into mRNA stability and codon bias and their relationships with translation. Nucleic Acids Research. 47:2216–28.
12. Goodman DB, Church GM, Kosuri S. (2013). Causes and effects of N-terminal codon bias in bacterial genes. Science;342(6157):475-9.
13. Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. (2013). Rate-limiting steps in yeast protein translation. Cell. 153(7):1589-601.
14. Goodman DB, Church GM, Kosuri S. (2013). Causes and effects of N-terminal codon bias in bacterial genes. Science. 342:475–9.
15. Kudla G, Murray AW, Tollervey D, Plotkin JB. (2009). Coding-sequence determinants of gene expression in Escherichia coli. Science. 324:255–8.
16. Verma M, Choi J, Cottrell KA, Lavagnino Z, Thomas EN, Pavlovic-Djuranovic S, Szczesny P, Piston DW, Zaher H, Puglisi JD, *et al*. (2019). Short translational ramp determines efficiency of protein synthesis. 2019; bioRxiv 571059.
17. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nature Protocols. 7:1534–50.
18. Morgan GJ, Burkhardt DH, Kelly JW, Powers ET. (2018). Translation efficiency is maintained at elevated temperature in Escherichia coli. J Biol Chem. 293:777–93.
19. Li G-W, Burkhardt D, Gross C, Weissman JS. (2014). Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. Cell. 157:624–35.
20. Burkhardt DH, Rouskin S, Zhang Y, Li G-W, Weissman JS, Gross CA. (2017). Operon mRNAs are organized into ORF-centric structures that predict translation efficiency. Elife. 6:811.
21. Arike L, Valgepea K, Peil L, Nahku R, Adamberg K, Vilu R. (2012). Comparison and applications of label-free absolute proteome quantification methods on *Escherichia coli*. Journal of Proteomics. 75:5437–48.
22. Allen TE, Herrgård MJ, Liu M, Qiu Y, Glasner JD, Blattner FR, Palsson BØ. (2003). Genome-scale analysis of the uses of the *Escherichia coli* genome: model-driven analysis of heterogeneous data sets. J Bacteriol. 185:6392–9.
23. Covert MW, Knight EM, Reed JL, Herrgård MJ, Palsson BØ. (2004). Integrating high-throughput and computational data elucidates bacterial networks. Nature. 429:92–6.
24. Corbin RW, Paliy O, Yang F, Shabanowitz J, Platt M, Lyons CE, Root K, McAuliffe J, Jordan MI, Kustu S, *et al*. (2003). Toward a protein profile of *Escherichia coli*: comparison to its transcription profile. Proceedings of the National Academy of Sciences. 100:9232–7.
25. Mustoe AM, Busan S, Rice GM, Hajdin CE, Peterson BK, Ruda VM, Kubica N, Nutiu R, Baryza JL, Weeks KM. (2018). Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing. Cell. 173:181–195.
26. Cho B-K, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BØ. (2009). The transcription unit architecture of the *Escherichia coli* genome. Nature Biotechnology. 27:1043–9.
27. Guimaraes JC, Rocha M, Arkin AP. (2014). Transcript level and sequence determinants of protein abundance and noise in *Escherichia coli*. Nucleic Acids Research. 42:4791–9.
28. Kelsic ED, Chung H, Cohen N, Park J, Wang HH, Kishony R. (2016). RNA Structural Determinants of Optimal Codons Revealed by MAGE-Seq. Cell Systems. 3:563–6.
29. Santos-Zavaleta A, Salgado H, Gama-Castro S, Sánchez-Pérez M, Gómez-Romero L, Ledezma-Tejeida D, García-Sotelo JS, Alquicira-Hernández K, Muñiz-Rascado LJ, Peña-Loredo P, *et al*. (2019). RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12. Nucleic Acids Research. 47:D212–20.

30. Mohammad F, Green R, Buskirk AR. (2019). A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. Elife. 8:8324.
31. Xu L, Dong Z, Fang L, Luo Y, Wei Z, Guo H, Zhang G, Gu YQ, Coleman-Derr D, Xia Q, *et al*. (2019). OrthoVenn2: a web server for whole-genome comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Research. 47:W52–8.
32. Demšar J, Curk T, Erjavec A, Machine ČGTJO, 2013. Orange: data mining toolbox in Python. Jmlrorg
33. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Research 2004; 14:1188–90.

FIGURES

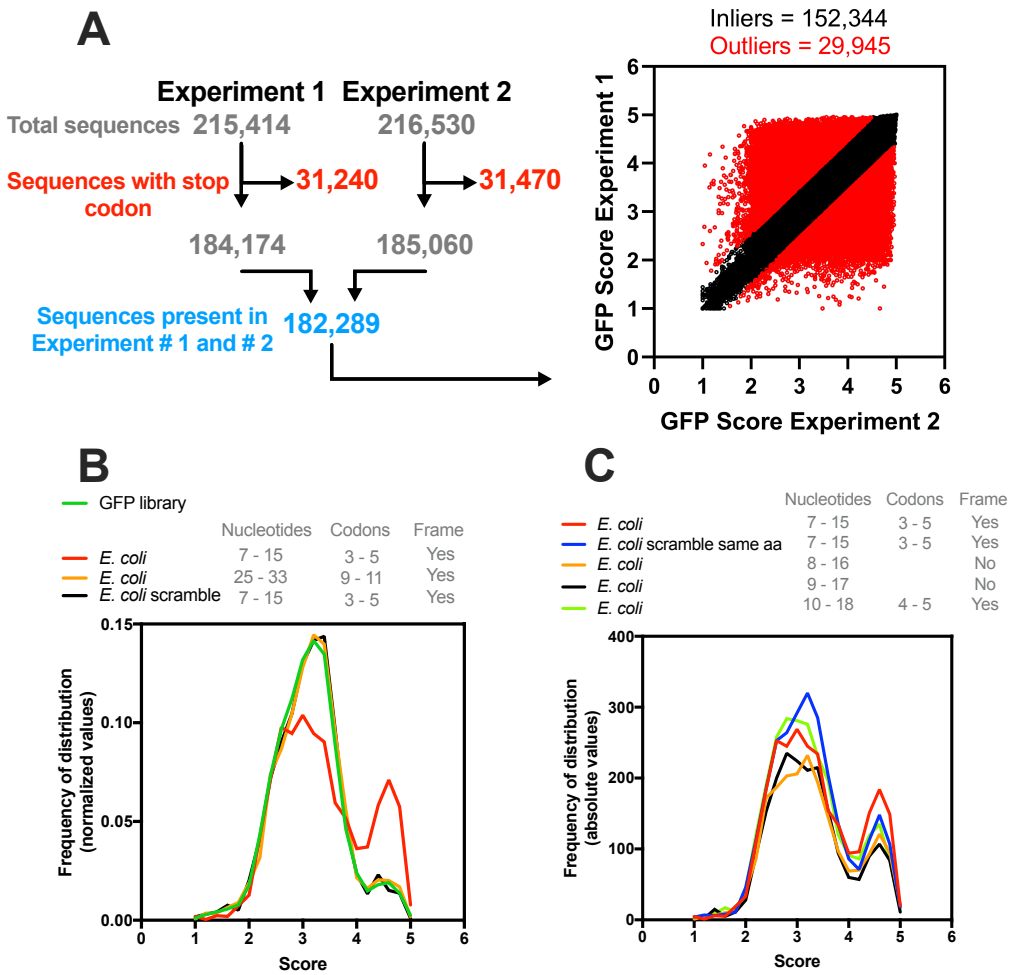

Figure 1. *E. coli* genome has a bias towards nucleotide composition at codons 3-5. (A) Djuranovic's group performed a high throughput screening varying the nucleotide composition at codons 3-5 in GFP. In experiment #1 and #2, 215,414 and 261,530 different compositions were analyzed, respectively, regarding the GFP fluorescence levels. The sequences with a stop codon (31,240 and 31,470 for experiments #1 and #2, respectively) were removed, and only sequences present in both experiments were used (182,289). The outliers (29,945) were defined as setting Q = 1% in linear regression. We then calculated the average GFP score of the inliers (152,344) from experiments #1 and #2. This list was used in all subsequent bioinformatics experiments. (B) Density histogram of GFP scores for genes identified in *E. coli*. The nucleotide composition at codon positions 3-5 or 9-11 was analyzed. As a control, we used a scrambled genome where the codons proportion was maintained, but their position was randomly changed. Note that just the codon positions 3-5 in the real genome possess a bias towards the high GFP scores. (C) The effect of amino acids composition and mRNA sequence on GFP score bias was analyzed. As a control, we used a scrambled genome where the codons were randomly changed, keeping the codons proportion and amino acids sequence of each gene (*E. coli* scramble same aa).

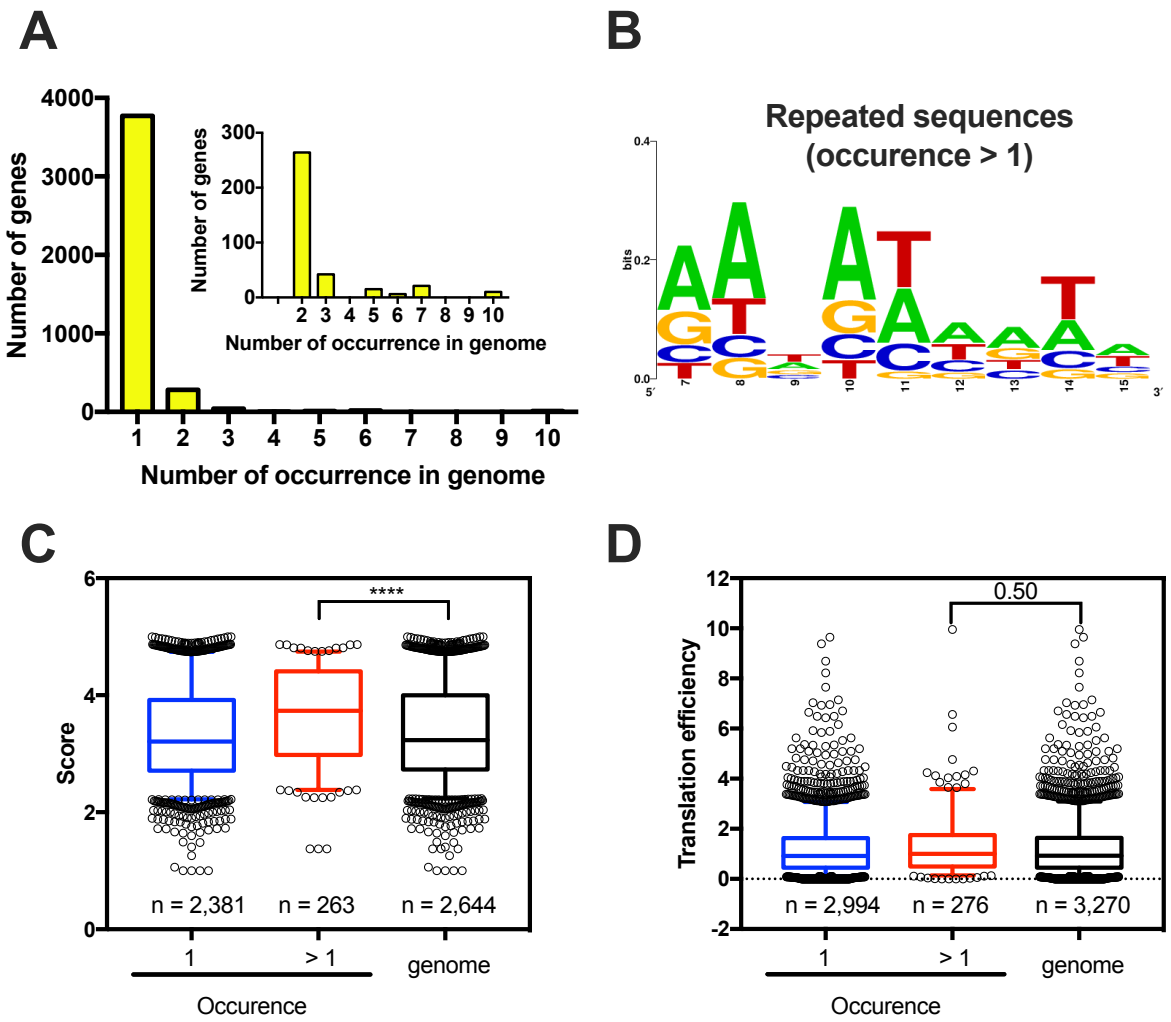

Figure 2. Most *E. coli* genes possess a unique nucleotide composition at codon positions 3-5. (A) The frequency of distribution of nucleotide composition at codon positions 3-5 shows that 91% of *E. coli* genes have a unique nucleotide composition (occurrence = 1). The other 9% share at least one nucleotide composition with another gene (occurrence > 1). (B) Web logo[33] of genes that share the same nucleotide composition at codon positions 3-5. GFP score average (C) and translation efficiency (D) comparison of genes with a unique nucleotide composition at codon positions 3-5 vs. genes with repeated nucleotide compositions. The TE data was obtained from Morgan *et al.*, 2018[18]. Kruskal-Wallis nonparametric test ****<0.0001, 0.5 = non-significative.

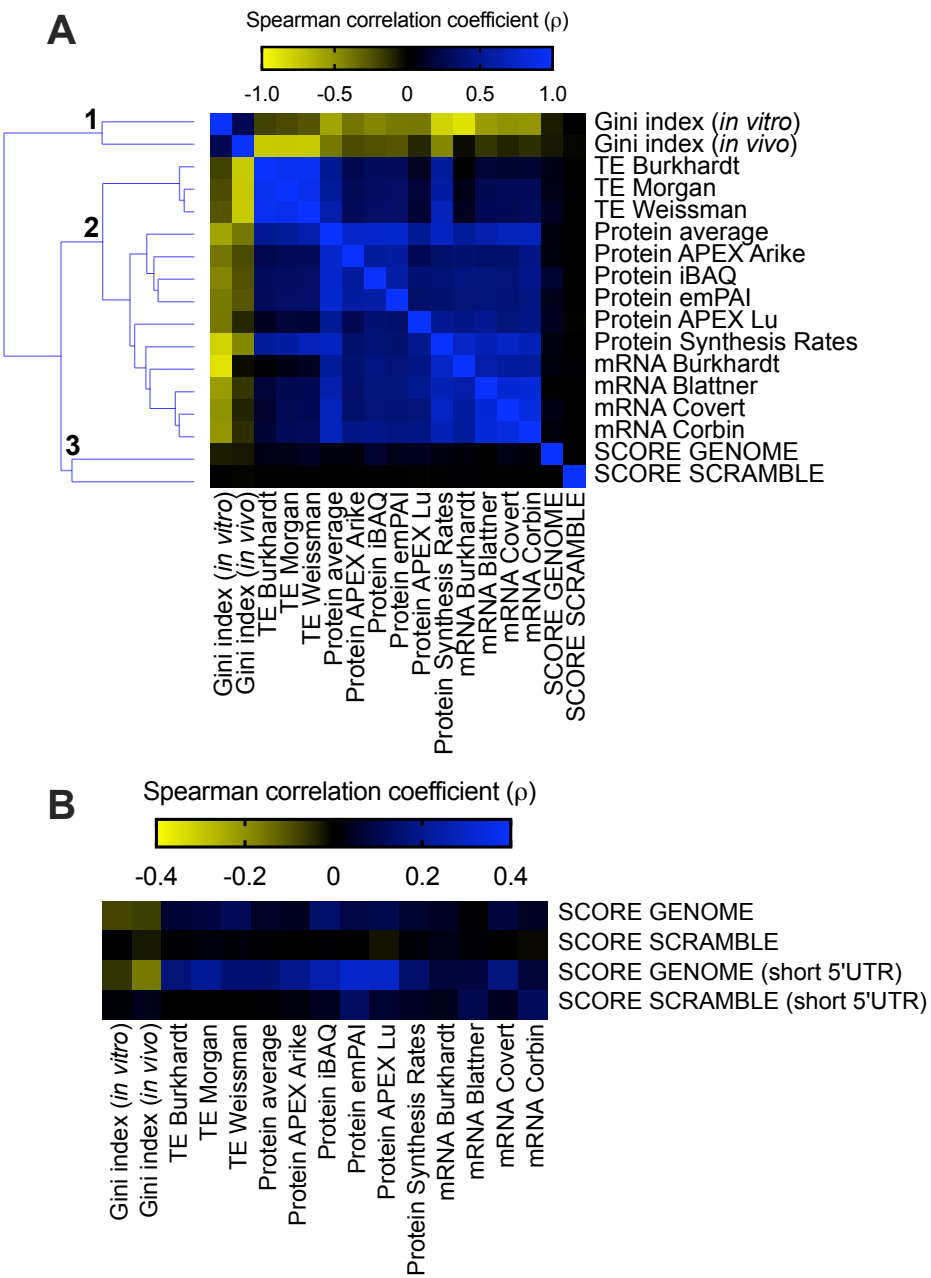

Figure 3. *E. coli* individual gene score calculation and its relationship with different parameters involved with gene expression. (A) Spearman's correlation between the *E. coli* score derived from GFP library dataset (SCORE GENOME) with different cellular parameters: Gini index, TE, protein abundance, protein synthesis rate, and mRNA abundance. The heat map shows Spearman's correlation coefficient (ρ) values ranging from -0.88 (negative correlation, yellow panels) to 1.0 (positive correlation, blue panels). The Spearman's correlation coefficient (ρ) values of GFP score with other parameters ranging from -0.06 to 0.11. As a control, we used an *E. coli* scrambled genome to calculate the GFP score (SCORE SCRAMBLE). (B) The same analysis described in panel A was performed with a group of genes with short 5' UTR (< 25 nucleotides).

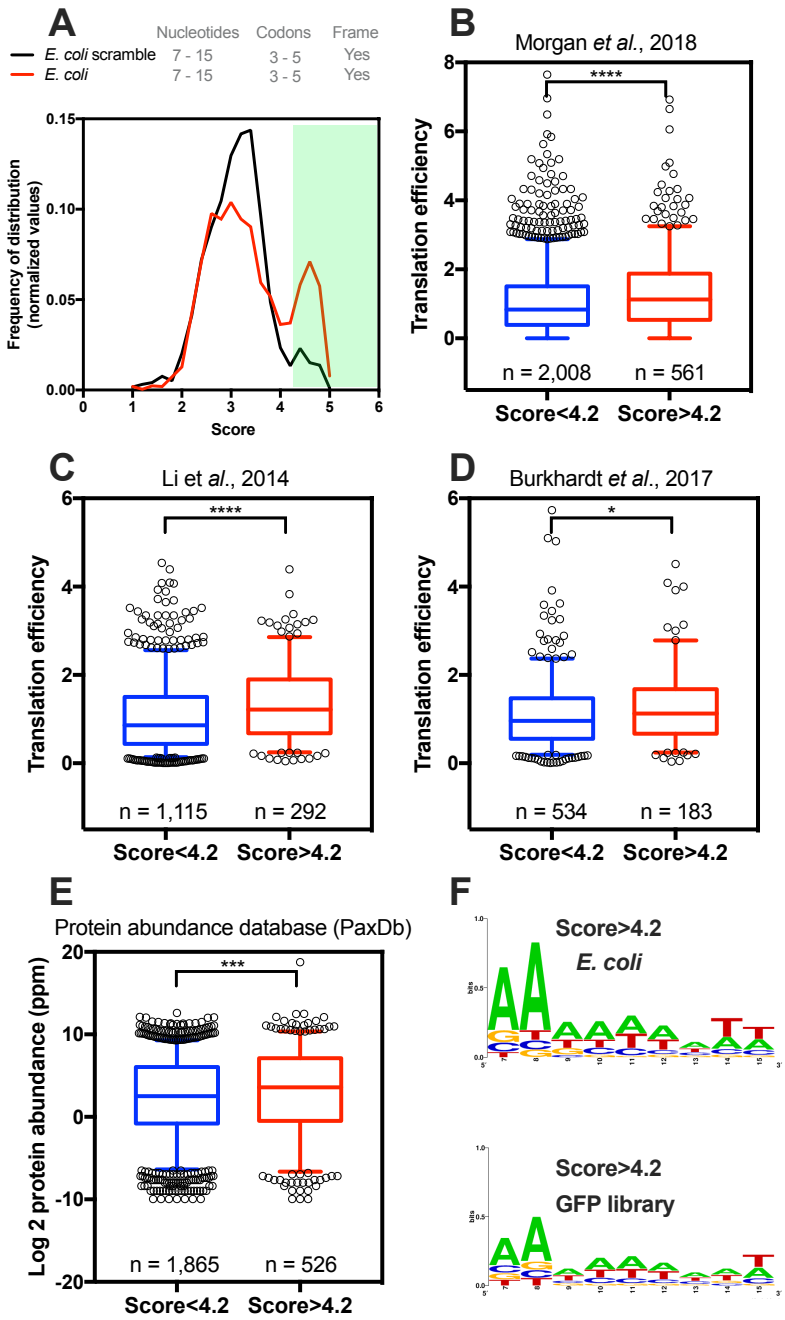

Figure 4. *E. coli* genes with GFP score higher than 4.2 have higher translation efficiency and protein abundance than other genes. (A) Density histogram of GFP scores for genes identified in *E. coli* for nucleotide composition at codon positions 3-5. As a control, we used a scrambled genome. Based on GFP score, the *E. coli* genes were divided into two groups: genes with a score higher than 4.2 and score lower than 4.2. Translation efficiency measured by three independent studies, Morgan *et al.*, 2018 [18](B), Li *et al.*, 2014 [19](C) and Burkahardt *et al.*, 2017 [20](D), were used. (E) Protein abundance of genes with a score higher than 4.2 and score lower than 4.2. Kolmogorov-Smirnov non-parametric t-test, ****<0.0001, ***0.0004, *0.028. (F) Web Logo of nucleotide composition at codon positions 3-5 of *E. coli* genes or GFP library with a score higher than 4.2.

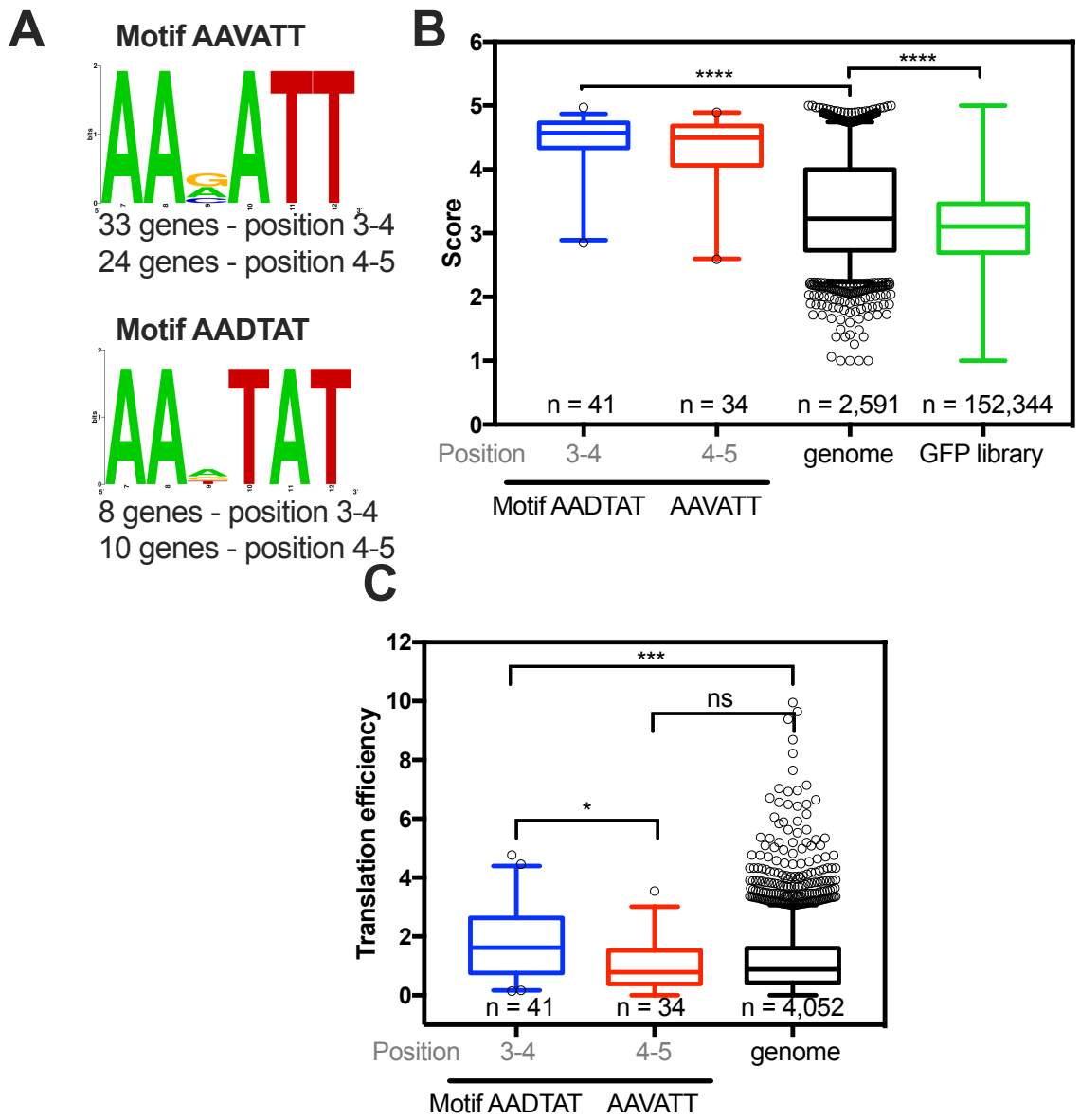

Figure 5. *E. coli* genes with the motifs AAVATT and AADTAT at codon positions 3 and 4 are more efficiently translated than other genes. (A) The motifs AAVATT or AADTAT (V = A, G or C and D = G, A or T) were identified in *E. coli* genome at codon positions 3 and 4 (33 genes with AAVATT and 8 genes with AADTAT) or at codon positions 4 and 5 (24 genes with AAVATT and 10 genes with AADTAT). The genes with motifs AAVATT or AADTAT at codon positions 3 to 4 have higher GFP score (B) and translation efficiency[18](C) than the *E. coli* genome. Kruskal-Wallis nonparametric test *** 0.0005 and ****<0.0001. The TE data was obtained from Morgan *et al.*, 2018.

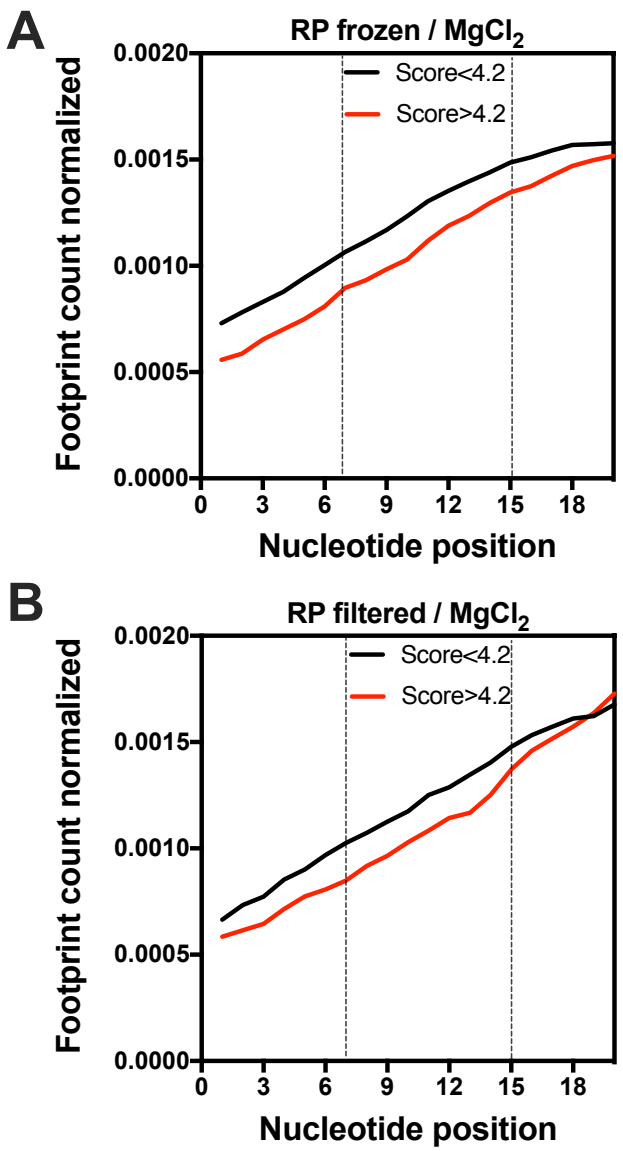

Figure 6. Ribosome occupancy at the first 20 nucleotides of genes with GFP scores higher than 4.2 at the codon positions 3-5 (nucleotides 7-15) are lower when compared to other genes. The average ribosome footprint counts of each group were obtained from libraries of ribosome profiling (RP) of differently treated samples: frozen/$MgCl_2$ (A) or filtered/$MgCl_2$ (B)[30]. For each gene, the number of reads per base was normalized to the total number of reads.

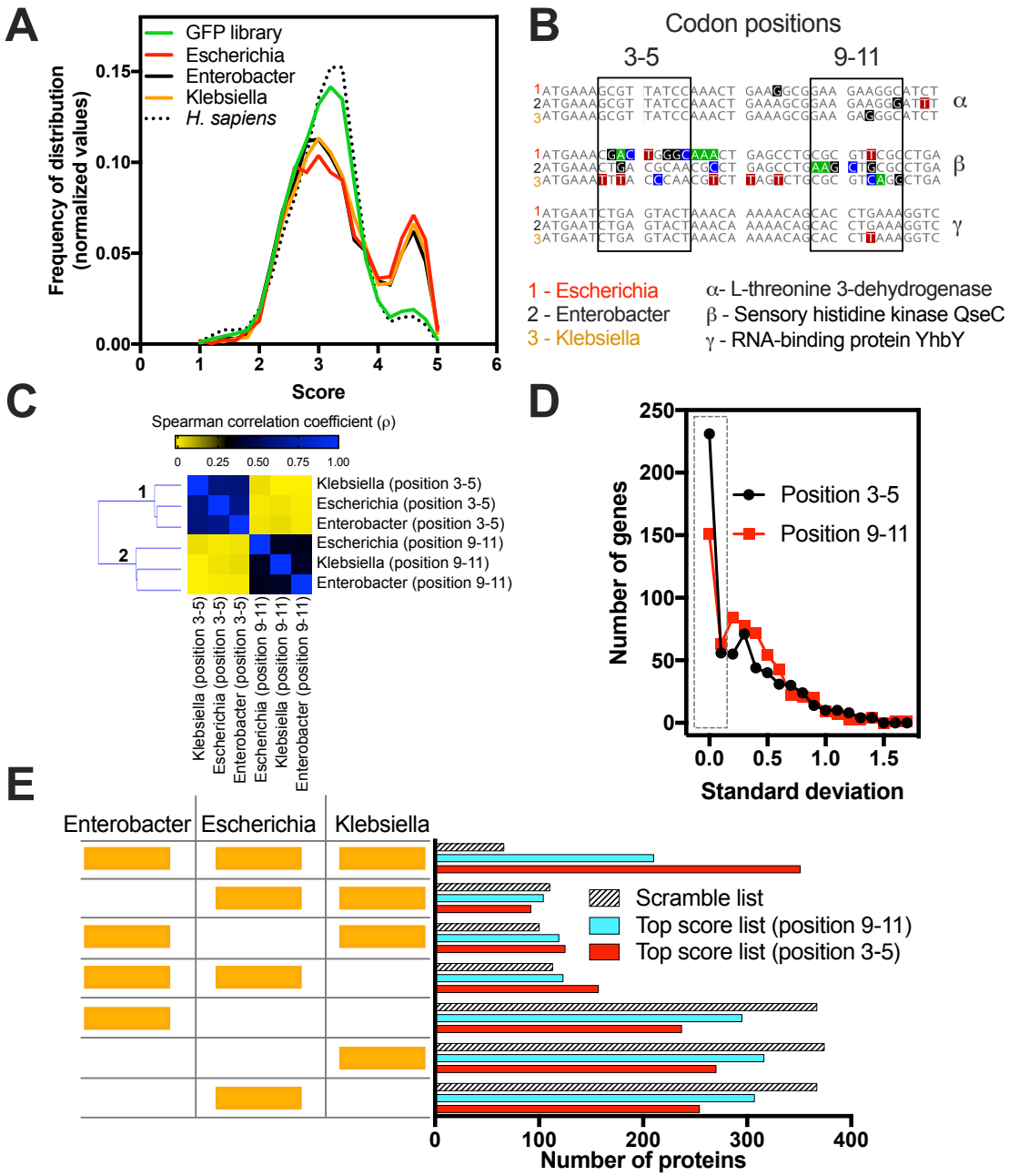

Figure 7. Evolutionary conservation of the short ribosomal ramp. (A) Density histogram of GFP scores of genes identified in *E. coli* (Escherichia), *E. asburiae* (Enterobacter), *K. oxytoca* (Klebsiella) and *Homo sapiens* for nucleotide compositions at codon positions 3-5. As a control, it was plotted the GPF library score. (B) A list of 1,595 orthologous genes of *E. coli*, *E. asburiae*, and *K. oxytoca* was analyzed regarding the score of each gene at codon positions 3-5 or 9-11. As an example, three genes (α, β, and γ) are shown. (C) Correlation matrix of the score of orthologous genes of *E. coli*, *E. asburiae*, and *K. oxytoca*. (D) The standard deviation score of each set of three orthologous genes was calculated and plotted as a frequency of distribution. (E) The top 500 genes with the best GFP score of *E. coli*, *E. asburiae,* and *K. oxytoca* were analyzed regarding their orthology. The codon positions 3-5 or 9-11 were used to calculate the GFP score. As a control, a scramble list with 500 genes of each bacteria was used.