The multi-level phenotypic impact of synonymous substitutions: heterologous gene expression in human cells.

SHORT TITLE: Multi-level impact of synonymous substitutions in human cells.

AUTHOR LIST: Marion A.L. Picard[1,*], Fiona Leblay[1], Cécile Cassan[1], Anouk Willemsen[1], Josquin Daron[1], Frédérique Bauffe[1], Mathilde Decourcelle[2], Antonin Demange[1], Ignacio G. Bravo[1,*]

AUTHOR AFFILIATIONS

[1] Laboratory MIVEGEC (CNRS, IRD, University of Montpellier), French National Center for Scientific Research, Montpellier, France

[2] Institut de Génomique Fonctionnelle (BCM, University of Montpellier, CNRS, INSERM), Montpellier, France

*Corresponding authors

E-mail: marionpicard@hotmail.com (MALP), ignacio.bravo@cnrs.fr (IGB)

AUTHOR CONTRIBUTIONS

Funding Acquisition, Project Administration and Supervision : IGB; Methodology : IGB, AD, MALP; Investigation : MALP, FL, CC, AW, FB, JD, MD, AD; Data Curation : MALP; Formal Analysis : MALP, IGB; Visualization : MALP; Conceptualization and Writing : MALP, IGB.

ABSTRACT

Redundancy in the genetic code allows for differences in transcription and/or translation efficiency between sequences carrying synonymous polymorphisms, potentially leading to phenotypic changes. It is commonly admitted that the evolution codon usage bias ("CUB", the over-representation of certain codons in a genome, a gene or in positions along a gene) are driven by a combination of neutral and selective processes, but their relative contribution is a matter of debate, especially in mammals. Particularly, integrative studies quantifying the phenotypic impact of CUB at different molecular and cellular levels are lacking. Here we report a multiscale analysis of the effects of synonymous codon recoding during heterologous gene expression in human cells.

Six synonymous versions of the *shble* antibiotic resistance gene were generated, fused to a fluorescent reporter, and independently expressed in HEK293 cells. Multiscale phenotype was assessed by means of: i) mRNA-to-DNA and protein-to-mRNA ratios for each *shble* version; ii) cellular fluorescence, using flow cytometry, as a proxy for single-cell level expression; and iii) real-time cell proliferation in absence or presence of antibiotic, as a proxy for the cell fitness.

We show that differences in CUB strongly impact the molecular and the cellular phenotype: i) they result in large differences in mRNA and in protein levels, as well in mRNA-to-protein ratio; ii) they introduce splicing events not predicted by current algorithms; iii) they lead to reproducible phenotypic heterogeneity; iv) they lead to a trade-off between the benefit of antibiotic resistance and the burden of heterologous expression.

We interpret that CUB modulate mRNA availability and suitability for translation in human cells, leading to differences in protein levels and eventually eliciting phenotypic differences.

AUTHOR SUMMARY

The genetic code is redundant, with several codons encoding for the same amino acid. These synonymous codons are not used with equal frequencies. Instead, codon usage bias (CUB) varies between species, genes, and even positions along a gene. At each of these levels, CUB are shaped by the overall

balance between mutational biases and selection forces. To shed light on the molecular mechanisms underlying molecular and organism phenotypes, integrative studies quantifying the phenotypic impact of CUB at different levels of biological integration are necessary. Here, we monitored the multiscale changes induced by six synonymous versions of an antibiotic resistance gene independently expressed in a human cell line. We show that: 1. both mRNA levels, protein levels, and their ratios are affected by CUB; 2. potential effects on mRNA (splicing and availability for translation) seem to constitute the main level of action; 3. cell fitness is severely impacted by a trade-off between the burden of heterologous expression and the benefits of antibiotic resistance. This integrative study provides new insights on translation regulation and the associated phenotypic impact in human cells, associated to CU.

INTRODUCTION

The canonical scenario of gene expression posits that a DNA sequence is first transcribed into messenger RNAs (mRNAs) that are secondly translated into proteins, such as one given sequence of nucleotides encodes one predictable sequence of amino acids (1). The initial version of this scenario did not provide any explanation on how a unique set of genes could be associated with several cellular phenotypes. But, through the last decades, a large body of studies on gene expression have addressed this question and revealed multi-level regulation mechanisms increasing the diversity of the proteomic output from a given genome. The genetic code which establishes a correspondence between the DNA coding units (*i.e.* the codon, a triplet of nucleotides, 64 in total) and the protein building blocks (*i.e.* the amino acids, 20 in total) is degenerated: 18 of the amino acids can individually be encoded by two, three, four or six triplets, known as synonymous codons. In a first null hypothesis approach, one would expect synonymous codons to display similar frequencies. Instead, CUB (*i.e.* the uneven representation of synonymous codons (2)) have been reported in a multiplicity of organisms, and vary not only between species but also within a given genome or even along positions in a gene (3–8).

The origin and the contribution of the different neutral and/or selective forces shaping CUB constitute a classical research subject in evolutionary genetics. The hypothesis of translational selection proposes that differences in CUB result in gene expression variations that ultimately lead to phenotypic

differences, which could be subject to natural selection. And indeed, it has been established that variation in CUB might constitute an additional layer of gene expression modulation (9–11). Notably, genetic engineering has extensively resorted to CUB recoding for enhancing heterologous protein production, for its use in industrial applications or for vaccine design (12–15). Besides the plethora of successful gene recoding strategies, the interaction between CUB and the translation machinery has been well established, for instance in: i) the co-variation of genomic CUB and the tRNA content, from unicellular organisms (4,16,17) to metazoa (*Caenorhabditis elegans* (18), *Drosophila* (19–21), or humans (22); ii) the correspondence between CUB and expression level in bacteria (23) or in yeast (24,25); iii) the increase in translation efficiency in bacteria when supplementing *in trans* with rare tRNAs (26); iv) the changes in tumorigenic phenotype in mice when switching from rare to common codons in the sequence of a cancer-related GTPase (27).

In contrast, a number of studies have communicated the lack of covariation between CUB and gene expression (in bacteria, yeast, or human) (28–31); or even a negative impact of a presupposed "optimization", which may in fact decrease the expression or the activity of the protein product (32,33). To address these conflicting results, it is important to tease apart the underlying mechanisms through which CUB can impact the molecular, cellular and/or organismal phenotype. It has hitherto been established that CUB can impact: 1. mRNA localisation, stability and decay (34–38), 2. translation initiation (31,39–41), 3. translation efficiency (20,42–55); 4. co-translational protein folding (56–58). But, fueling the controversy, the respective contribution of each mechanism, if any, depends on the studied system (*e.g.* in which organism, whether the expressed gene is autologous or heterologous gene, whether it has been recoded or not).

In this study, we aim at providing an integrated view of the molecular and cellular impact of alternative CUB of a heterologous gene in human cells. We designed six synonymous version of the *shble* antibiotic resistance gene with distinct CUB coupled them to a *egfp* reporter, and transfected them into cultured cells. By combining transcriptomics, proteomics, fluorescence analysis and cell growth evaluation, we attempt to describe qualitatively, and to quantify as far as possible, the impact of CUB, and associated sequence composition, on the molecular and cellular phenotype of human cells in culture.

RESULTS

1. Alternative CUB of the *shble* gene resulted in differences in mRNA abundance, and splicing.

The expected transcript was a 1,602 base pair (bp) long mRNA encompassing a 1,182bp coding sequence (CDS). The CDS spanned an *AU1*-tag sequence in 5', a *shble* CDS, a *P2A* peptide sequence inducing ribosomal skipping, and an *EGFP* reporter CDS (Sup. Fig. 1). Only the *shble* CDS differed between constructs, and was characterized by distinct degrees of similarity to the average human CUB (estimated using the COdon Usage Similarity INdex, i.e. COUSIN) (59), GC composition at the third nucleotide of codons (GC3), and CpG dinucleotide frequency (CpG) (Table 1). Modifications in the *shble* sequence also entailed variations on the mRNA folding energy (Table 1). All these four parameters allowed for a good discrimination of all constructs (Sup. Fig. 2), partly reflecting sequence similarities (Sup. Table 1).

Table 1. Experimental conditions: the different constructs, and their sequence composition variables.

Condition	**Description**	**COUSIN of the *shble* sequence**	**%GC3 of the *shble* sequence**	**%CpG of the *shble* sequence**	**Folding energy of the total transcript (kcal/mol)**
shble#1	The most common codons in the human genome	2.93	93.08	18.46	-649.34
shble#2	The GC-richest among the two most common codons	2.982	99.23	22.56	-673.07
shble#3	The AT-richest among the two most common codons	-0.414	20.00	4.62	-581.47
shble#4	The rarest codons in the human genome	-1.651	33.85	20.51	-613.49
shble#5	The GC-richest among the two rarest codons	0.973	91.54	35.90	-687.76
shble#6	The AT-richest among the two rarest codons	-0.924	9.23	0.51	-543.50
#empty	No *shble* but only *EGFP* CDS	n.a.	n.a.	n.a.	n.a.
#superempty	Neither *shble* nor *EGFP* CDS	n.a.	n.a.	n.a.	n.a.
mock	No plasmid	n.a.	n.a.	n.a.	n.a.

Transcriptomic analysis (RNA-seq), through the observation of the read distribution along the plasmid sequence, revealed the presence of splicing events for the two constructs with the lowest similarity to the human average CUB, namely shble#4 (construct using the rarest codon for each amino acid) and

shble#6 (using rare and AT-rich codons) (Sup. Fig. 3). The shble#6 transcript presented two spliced forms, using the same 5' donor position and differing in three nucleotides at the 3' acceptor position. The shble#4 transcript presented one spliced form, with donor and acceptor positions in the precise same location than observed for shble#6, despite the lack of identity in the intron-exon boundaries. In all cases the spliced intron (either 306 or 309 nucleotides long) was fully comprised within the 396 bp long *shble* sequence (Sup. Fig. 4), and the event did not involve any frameshift. Thus, *shble* splicing resulted in the ablation of the SHBLE protein coding potential without affecting the EGFP coding potential. It is important to state that none of these alternative splicing events was predicted by the HSF (Human Splicing Finder) (60) nor the SPLM (61) splice detection algorithms used for sequence scanning during design.

The mRNA abundances, expressed as transcript per millions (TPM), showed that the spliced isoform 1 (shared by both affected conditions) represented about 30% of the heterologous transcripts for shble#4, and 56% for shble#6. The spliced isoform 2, exclusively found in condition shble#6, corresponded to 22% of the heterologous transcripts (Figure 1). The full-length mRNA, albeit present in all conditions, was differentially represented, as follows: (i) the highest values were found in shble#3 (using the AT-richest among common codons); (ii) the variance was largest in shble#5 (using the GC-richest among rare codons); and (iii) shble#4 and shble#6 displayed the lowest mRNA abundance even when considering the sum of all isoforms (Figure 1, Sup. Table 2). We further verified that variations in transcript levels were not related to variations in transfection efficiency, by correcting the TPM values with the plasmid DNA levels in each sample as estimated by qPCR. After this normalisation, the above described pattern remained unchanged (Sup. Fig. 5). This suggests that variations in mRNA levels are not due to differences in the DNA level, and may instead be linked to the differentially recoded *shble* sequences.

In order to allow further comparison between mRNA and protein levels, while accounting for the differential splice events, we have taken into account that the SHBLE protein was exclusively encoded by the full-length mRNA, while the EGFP protein could be translated from any of the three transcript isoforms. Hence, we used the ratio full-length mRNA over total transcripts (*i.e.* full-to-total ratio) to estimate the ratio of SHBLE-encoding over EGFP-encoding transcripts. This ratio was about 69% shble#4, while for shble#6 it

was close to 21% (Sup. Table 2). For the rest of the constructs, there was virtually no read corresponding to spliced transcripts and the ratio was in all cases above 99.96% (Sup. Table 2).

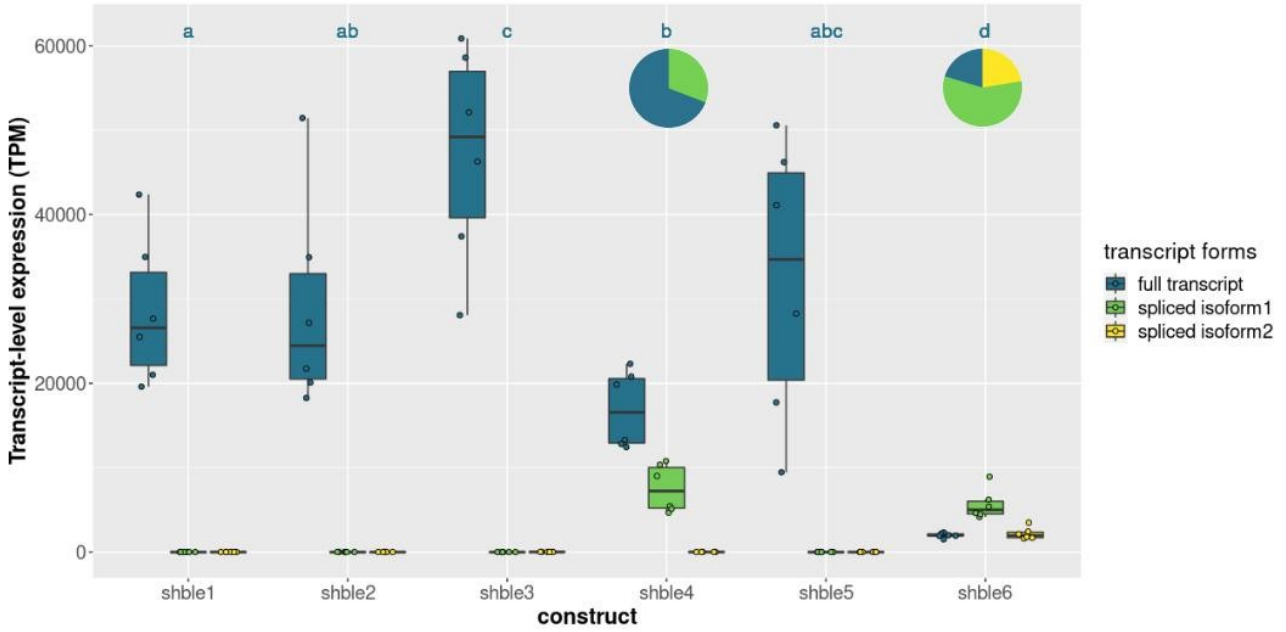

Figure 1. Transcript abundance after transfection with the different shble gene versions. *mRNA-levels are expressed as transcripts per million values (TPM) for the full form (in dark blue) as well as for the two spliced forms (in green and yellow). Median values are given in Sup. Table 2. Pie charts illustrate the proportions of the spliced forms detected in shble#4 and shble#6 conditions. The experiment was performed on six biological replicates. Dark blue letters above the different bars refer to the results of a Wilcoxon rank sum test. Conditions associated with a same letter do not display different median TPM values for the full mRNA ($p<0.05$ after Benjamini-Hochberg correction).*

2. Alternative CUB of the *shble* gene impacted SHBLE and EGFP protein levels.

Label-free proteomic analysis allowed to detect EGFP proteins for all constructs, with EGFP abundance in shble#3 and shble#6 being significantly lower than in other conditions (respectively 2.05 and 1.35 normalized iBAQ values, compared to an overall median of 10.08 for the other constructs) (Figure 2C, Sup. Table 3). The SHBLE protein was detected in all conditions but, for shble#6, it displayed extremely low abundance in five replicates and was not detected in one replicate (overall normalized iBAQ value of 0.03) (Figure 2B, Sup. Table 3). Further, the shble#3 condition displayed lower SHBLE protein levels than the remaining four other constructs (normalized iBAQ value of 0.93, compared to an overall median of 3.83) (Figure 2B, Sup. Table 3). Within a given condition, values for SHBLE and EGFP protein levels displayed a

strong, positive correlation (Pearson R coefficients ranging from 0.82 to 0.95 depending on the condition; all p-values < 0.05; Figure 2A). The overall SHBLE-to-EGFP ratio was 0.46±0.1 for all constructs (ranging between 0.36 and 0.56 for the individual constructs), the exception being shble#6, which displayed a ratio close to zero, linked to the very low SHBLE levels (Figure 2D). Label-free proteomic quantification results were validated by semi-quantitative western blot experiments (Sup. Fig. 6, 7 and 8).

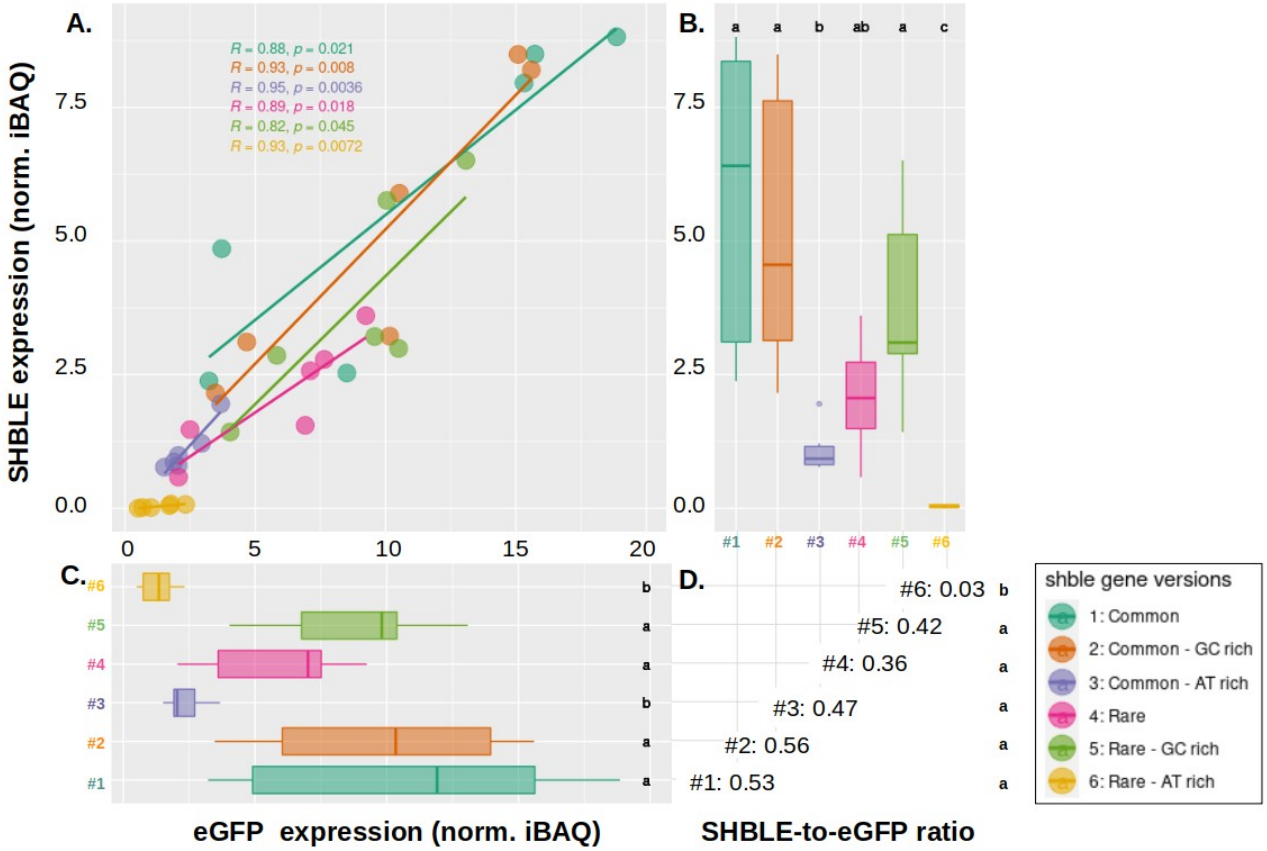

Figure 2. Expression of SHBLE and EGFP at the proteomic level, and relation between them. *Panel A: Pearson's correlation between SHBLE (y axis) and EGFP (x axis) protein levels. Six different conditions are shown: shble#1 (dark green), shble#2 (orange), shble#3 (purple), shble#4 (pink), shble#5 (light green) and shble#6 (yellow). Marginal boxplots (panels B and C) respectively show SHBLE and EGFP protein levels expressed as normalized iBAQ values. Median values are given in Sup. Table 3. The SHBLE-to-EGFP ratio for each of the six conditions (median of the ratios for each replicate) are given in panel D. Six replicates are shown (with three of them corresponding to two pooled biological replicates). Letters in the different panels refer to the results of a pairwise Wilcoxon rank sum test. Within each panel, conditions associated with a same letter do not display different median values of the corresponding variable (p<0.05 after Benjamini-Hochberg correction).*

3. Transcriptomic and proteomic phenotypes matched to different extent, and a combination several composition variables could explain the differences.

After analysing separately mRNA and protein levels in cells transfected with the different *shble* versions, we aimed at establishing a connection between those transcriptomic and proteomic phenotypes. Because SHBLE and EGFP protein analyses led to similar results, we focus here only on SHBLE. Variation in full-length transcripts levels explained 45% of the variation in the protein levels (Figure 3A). Interestingly, the shble#3 condition behaved differently from the rest and rendered similar SHBLE protein values for all replicates, independently of the variation in transcript levels (Figure 3A). When removing this condition from the correlation analysis, variation in full-length transcripts levels explained 83% of the variation in the SHBLE protein levels (Figure 3B).

In order to understand the differential matches between transcriptomic and proteomic phenotypes, we explored the explanatory potential of four sequence composition and mRNA physicochemical parameters. First, an increase of the match between the *shble* sequence and the human average CUB (i.e. COUSIN score) corresponded to an elevation in the protein-to-transcript ratio (Pearson's R=0.67, p=6.6e-6, Figure 3C) - with the exception of the shble#4 condition which displayed the lowest match to the human CUB, but a higher protein-to-transcript ratio than shble#6 and #3 (Figure 3C). In fact, the lower ratio for these two later conditions could be explained at the light of the three other tested parameters. Indeed, an increase of the GC3 content corresponded monotonically to an augmentation in the protein-to-transcript ratio (Pearson's R=0.81, p=1.5e-9, Figure 3D), and shble#6 and #3 had the lowest GC3 content. Then, variations in CpG frequency (Figure 3E), and in mRNA folding energy (Figure 3F), corresponded to a bell-shaped variation in SHBLE protein-to-transcript ratio so that both low and high values resulted in decreased protein-to-transcript ratio (Figure 3E and 3F): shble#6 and #3 had the lowest CpG frequency, and the highest mRNA folding energy. Thus, the shble#3 condition combined suboptimal values for all four studied characteristics and resulted in poorly efficient translation in spite of the high mRNA levels (see part1). In contrast, shble#1 (made of the most used codons), displayed an optimal value for all parameters and resulted in the most efficient translation (highest protein-to-mRNA ratio).

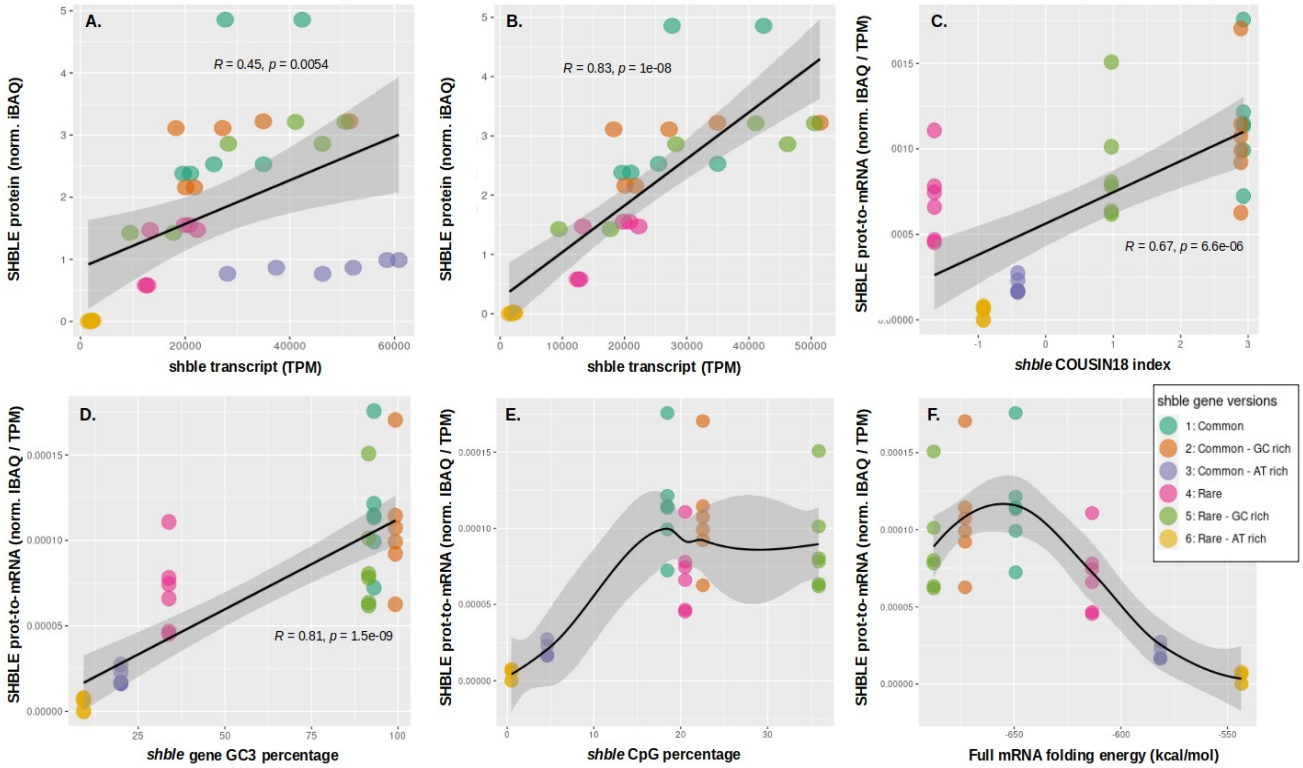

Figure 3. Relation between the transcriptomic and the proteomic phenotypes, and potential explicative parameters for variations in SHBLE protein levels. *Six different conditions are shown, using the colour code: shble#1 (dark green), shble#2 (orange), shble#3 (purple), shble#4 (pink), shble#5 (light green) and shble#6 (yellow).* ***Panel A:*** *Pearson's correlation of SHBLE protein level and* shble *transcript level taking into account all six constructs.* ***Panel B:*** *The construct #3, which displayed a discordant pattern from the others (see panel A), was excluded from the Pearson's correlation.* ***Panel C:*** *Pearson's correlation between SHBLE protein-to-mRNA ratio and COUSIN index of the shble recoded version.* ***Panel D:*** *Pearson's correlation between the SHBLE protein-to-mRNA ratio and the GC3 percentage of the shble recoded version.* ***Panel E:*** *SHBLE protein-to-mRNA ratio variations depending on CpG frequency of the shble recoded version.* ***Panel F:*** *Correspondence between the SHBLE protein-to-mRNA ratio and the folding energy of the corresponding transcript. The results for six biological replicates are shown, each of them with independent RNAseq measurements but pooled by pairs for the label-free proteomic analysis.*

4. Single-cell EGFP protein expression varied within each condition, but CUB variations induced shifts of the whole population.

We have demonstrated above that the EGFP reporter was a relevant proxy for SHBLE abundance, as their iBAQ values were highly correlated (Figure 2A). On this basis, and in order to further assess the

phenotypic variation at the single-cell level, we performed an extensive analyse on the cell-based fluorescence values of 16 transfection replicates. We verified first that the total fluorescence signal (i.e. adding all cells) was strongly correlated to the EGFP level estimated by the label-free proteomics (Pearson's R=0.86, p=4.8e-15, Sup. Fig. 9). We observed then that the distribution of this fluorescence signal was 1. for all conditions, different from that obtained with cells expressing EGFP alone (i.e. "empty" control; individual Anderson-Darling test results shown in Table 2); and 2. multimodal for all the conditions expressing EGFP (Figure 4A, Sup. Fig. 10). We described these multimodal populations by means of curve deconvolution, and showed that an approximation based on two underlying Gaussian populations fitted well the observed distributions (Sup. Fig. 11). Thus, 1. the construct expression changed the fluorescence phenotype, and synonymous variations of the upstream *shble* sequence modulated it; 2. for a given version of the *shble* sequence, cells were differentially impacted by the construct expression, overall defining two subpopulations of low or high EGFP expression.

Table 2. Quantitative parameters of green fluorescence signal distribution per condition.

Condition	Distribution similarity to #empty (AD score and associated p-value)*		Percentage of fluorescent cells	Total fluorescence value for the whole population[$]		Mean fluorescence value for the underlying first Gaussian subpopulation (log10)	Mean fluorescence value for the underlying second Gaussian subpopulation (log10)
#shble1	1580	0	89.56%	105.269 e9	bc	4.84	6.61
#shble2	1480	0	90.17%	98.311 e9	b	4.78	6.59
#shble3	497	4.637 e-272	79.37%	39.384 e9	d	4.31	5.86
#shble4	463	7.325 e-254	88.00%	63.395 e9	ac	4.58	6.28
#shble5	108	4.244 e-59	83.85%	70.719 e9	abc	4.63	6.32
#shble6	11600	0	51.78%	13.990 e9	e	3.97	5.05
#empty	0	1	82.26%	57.692 e9	a	4.44	6.18
#superempty	64100	0	0.45%	135.449 e6	*n.a.*	*n.a.*	*n.a.*
mock	62500	0	1.00%	141.163 e6	*n.a.*	*n.a.*	*n.a.*

**"AD", results of an Anderson-Darling test for distribution similarity, comparing each curve distribution in Figure 4A against that obtained for the "empty" condition (the null hypothesis being that the samples compared could have been drawn from a common population). [$]The statistical test is a pairwise Wilcoxon rank sum test. Conditions associated with a same letter do not display different median values for the corresponding variable (p<0.05 after Benjamini-Hochberg correction).*

For each condition, we summarized the fluorescence behaviour of the whole cell population by describing the following statistics (Table2): (i) the fraction of cells displaying fluorescence over 99th

242 percentile of the "mock" fluorescence distribution (*i.e.* 14,453 'FITC-A' fluorescence units, which
243 corresponded to the cell autofluorescence, as the mock did not carry any plasmid); (ii) the total fluorescence
244 value of the whole population; (iii) the median fluorescence value of the population; (iv) the mean
245 fluorescence value for each underlying Gaussian populations. We observed that the central fluorescence
246 value (i.e. the median) of the population correlated very well with the overall fluorescence (R=0.85, p-
247 value<2.2e-16, Sup. Fig. 12), but that the later allowed a better discrimination between conditions. Shble#1
248 and #2 thereby appeared more fluorescent than the #empty control, and shble#6 displayed a lower
249 fluorescence than all the other conditions, as did shble#3 in a lower magnitude (Table 2, Sup. Fig. 12). Those
250 differences in the total signal reflected in fact an reproductible impact of the synonymous construct
251 expression on all cells, independently of their affiliation to the low or high fluorescent populations: indeed,
252 between each condition, both underlying Gaussian curves shifted following the same pattern, as illustrated
253 by the variations of their mean values (Figure 4B, Table 2). When combining all our summary statistic
254 variables for describing the population cellular fluorescence we observed that indeed shble#6, and to a lesser
255 extent shble#3, were the most divergent conditions, characterized by the highest proportion of negative or
256 low-fluorescent cells, while shble#1 and shble#2 displayed very similar behaviour characterized by high
257 fluorescence values in all scores (Sup. Fig. 13). Those results strengthened the observations obtained by the
258 label-free proteomic experiments, and underlied the cell-to-cell reproductibilty of the impact of synonymous
259 substitutions.

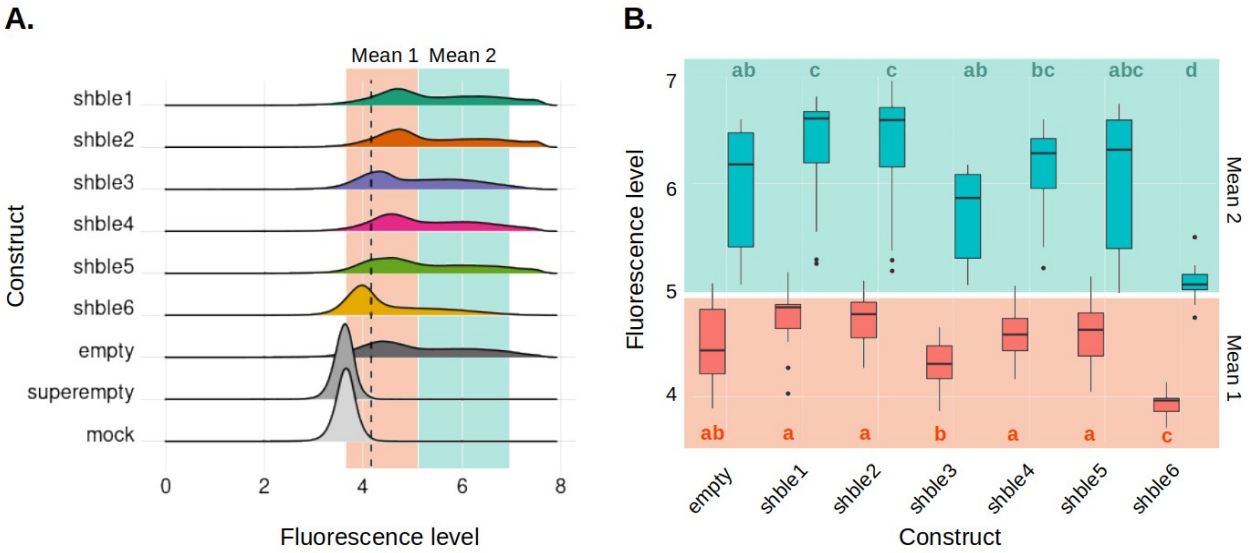

Figure 4. Distribution of the fluorescence signal for the different constructs, and mean values of the two gaussian curves modeling the fluorescence distribution. *Panel A depicts the density of the green fluorescence signal (log10(FITC-A)) considering 480,000 individual cells for each condition: shble#1 (most common codons, dark green), shble#2 (common and GC-rich codons, orange), shble#3 (common and AT-rich codons, purple), shble#4 (rarest codons, pink), shble#5 (rare and GC-rich codons, orange light green), shble#6 (rare and AT-rich codons, yellow). The positive control is "empty" (i.e. transfected cells, expressing EGFP without expressing SHBLE, in dark grey); and the negative controls are "superempty" (i.e. transfected cells, not expressing EGFP nor SHBLE, in medium grey) and "mock" (i.e. untransfected cells, in light grey). The dashed black line shows the threshold for positivity (14,453 green fluorescence units, corresponding to 4.16 in a log10 scale). Panel B represents the first gaussian mean1 (population of lower intensity, in red), and the mean2 (population of higher fluorescence intensity, in blue). For each category of mean, the statistical test is a pairwise wilcoxon rank sum test, with Benjamini-Hochberg adjusted p-values on sixteen biological replicates: for each color, conditions associated to a same letter do not display different median values of the corresponding variable.*

5. Alternative CUB of the *shble* gene resulted in different cell growth dynamics.

To assess the functional impact of the molecular phenotypes described above, we performed a real-time cell growth analysis, both in presence and in absence of antibiotics. We anticipated a trade-off between a potential benefit confered by the antibiotic resistance, and a potential cost through protein overexpression and its associated burden. Thus, to disentangle the effects linked to the total expression of heterologous proteins (SHBLE + EGFP), and the effect of the antibiotic resistance gene alone, we tested two additional constructs solely containing versions shble#1 and shble#4 of the *shble* gene, not linked to the *EGFP* reporter

(labelled shble#1* and shble#4* in Figure 5 and Sup. Fig. 14). For all conditions, we monitored over time a dimensionless parameter named "Cell Index", that integrates cell density, adhesion, morphology and viability; and we evaluated the total area below the curve as a proxy for cell growth (Sup. Method 2.8, Sup. Fig. 14). We fitted to a Hill's equation these values of cell growth as a function of the antibiotic concentration to recover, for each condition: 1. the maximum growth in the absence of antibiotic (Figure 5, y axis); and 2. the estimation for the antibiotic concentration value that inhibited cell growth to half the maximum (IC50; Figure 5, x axis).

First, we observed that transfection with an empty vector (i.e. the "superempty" control, not expressing EGFP nor SHBLE) resulted in a drop of about 50% in max. growth in the absence of antibiotic (Figure 5, y axis), and in a drop of about 85% in IC50 value (Figure 5, x axis) compared to the mock. This meant that, independently of heterologous gene expression, the transfection alone had an impact on cell fitness.

Second, all cell populations transfected with any of the constructs displayed a lower max. growth in the absence of antibiotics than the "superempty" control (Figure 5, y axis). Only shble#4 and #6, expressing SHBLE and EGFP at low level, and shble#1* and shble#4* lacking EGFP, had a higher IC50 value (Figure 5, x axis). Further, all transfected cells resisted less the presence of antibiotics than the mock independently of the construct used (Figure 5, x axis), at the exception of shble#1* (high level of SHBLE and no EGFP). This meant that the burden induced by the expression of heterologous proteins was stronger than the potential benefit of the antibiotic resistance.

Focusing on the comparison of the *shble* versions #1 and #4 with or without EGFP, both shble#1* and shble#4* versions displayed a similar increase in max. growth in the absence of antibiotic with respect to their EGFP+ relative counterparts (respectively 38% and 24%, shown as coloured arrows on Figure 5, y axis). However, while the IC50 of shble#4* remained similar to shble#4, the antibiotic resistance for version shble#1* dramatically increased with respect to that of shble#1 (green arrow on Figure 5, x axis). As mentioned before, shble#1* was the only condition in which resistance to the antibiotic was actually better than for the untransfected cells, in spite of a remaining substantial negative impact on maximum growth on the absence of antibiotics (Figure 5, y axis). This meant that 1. in absence of antibiotics, higher amount of heterologous protein (whatever if it correspond to SHBLE or EGFP) had a more pronounced negative impact

on cell fitness; 2. in presence of antibiotic, the optimum between the confered resistance and the cost of protein burden was determined by both, the total amount of the two heterologous proteins (including EGFP, which was unecessary to the cell fitness), and the abundance of the protein confering the resistance itself (SHBLE).

Overall, even if no significant correlation could be established because of the limited number of experimental conditions, a trend appeared: variation in cell fitness in absence of antibiotics seemed to be inversely related to variation in total amount of heterologous proteins (using fluorescence as a proxy, showed as dot size in Figure 5), so that conditions displaying strong cellular fluorescence (*e.g.* shble#1) grew less in the absence of antibiotics, and resisted worse the presence of antibiotics, than conditions displaying lower fluorescence (*e.g.* shble#6) (Figure 5 , Sup. Fig. 14). Our results suggested thus: first, the existence of an important stress related to plasmid transfection; and second, the establishment of a trade-off between the benefit of heterologous protein expression conferring resistance and the additional burden of fluorescent protein expression coupled to the resistance.

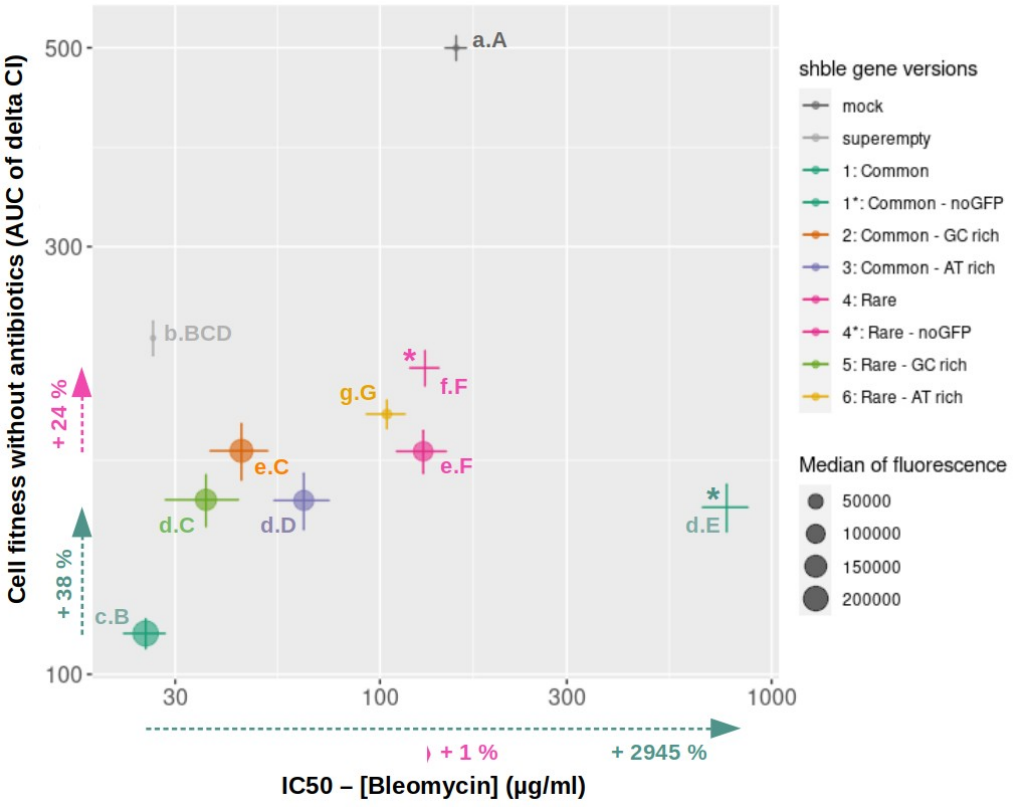

Figure 5. Variation of cell growth in presence or in absence of antibiotics. *The y axis represents maximum cell growth in absence of antibiotics, proxied as the area under the curve of the delta Cell Index (AUC, log scale). The x axis represents the bleomycin concentration reducing to 50% the corresponding maximum growth (e.g. IC50; log10 scale). Represented central values were estimated fitting Cell Index data to Hill's equation (pooled data, 3 to 6 biological replicates), and bars correspond to the standard error (left standard error for superempty IC50 was out of the graph limit and is not plotted – but see Sup. Fig. 15 for representation on linear axes). Statistical tests are Welch modified two-sample t-tests, performed for the AUC (small letters, y axis) or the IC50 (big letters, x axis): for each size of letters, conditions associated with a same letter do not display different median values of the corresponding variable ($p<0.05$ after Benjamini-Hochberg correction). The size of the dots is proportional to the corresponding median of fluorescence, which is used as a proxy for the level of heterologous proteins. Nine different conditions are shown: mock control (dark grey), superempty control (light grey), shble#1 (dark green), shble#2 (orange), shble#3 (purple), shble#4 (pink), shble#5 (light green), shble#6 (yellow) and versions shble#1* (dark green) and shble#4* (pink) lacking the EGFP reporter gene. Arrows on the margins represent the shift of values (expressed as percentage of the initial value) for shble#1* and shble#4* against shble#1 and shble#4 respectively.*

DISCUSSION

In the present manuscript we have analysed the multilevel molecular *cis*-effects of codon usage preferences on gene expression, and have further explored higher-level integration consequences at the cellular level. The global *trans*-effects of codon usage preferences of our focal gene on the expression levels of other cellular genes have been analysed and described in an accompanying paper (62). We summarize our observations of these *cis*-effects in Figure 6, which displays variation in the each of the composition and phenotypic variables monitored for the different genotypes analysed. This representation highlights that a combination of synonymous changes results in important multilevel variation in gene expression levels and leads to dramatic differences in the cellular phenotype.

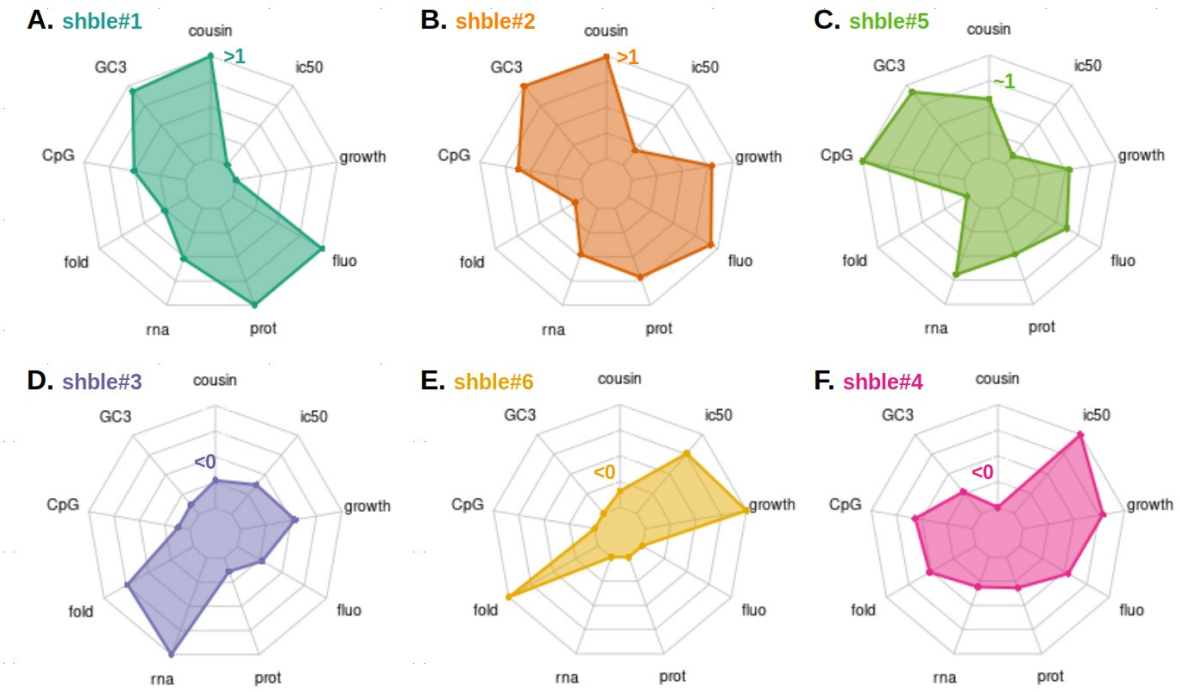

Figure 6. Summarizing combination of sequence composition parameters and multi-level phenotypes for each customized version of the* shble *antibiotic resistance gene. *The six versions, designed with the one amino acid – one codon strategy, are shown by decreasing similarity to the human CUB (i.e. cousin score). They are defined as follow:* ***A****. shble#1 (most common codons, cousin > 1, in dark green),* ***B****. shble#2 (common and GC-rich codons, cousin > 1, in orange),* ***C****. shble#5 (rare and GC-rich codons, cousin ~ 1, in light green),* ***D****. shble#3 (common and AT-rich codons, cousin < 0, in purple),* ***E****. shble#6 (rare and AT-rich codons, cousin < 0, yellow) and* ***F****. shble#4 (rarest codons, cousin < 0, in pink). The sequence characteristics are from the top to the left: "cousin" (expressing the similarity to the average genome human CUB: a score < 0 is opposite; ~1 is similar; >1 is similar human CUB, but of larger magnitude (59)), "CpG" (the CG dinucleotide frequency), "GC3" (the GC content at the third base of the codons), and "fold" (the mRNA folding energy). The different phenotypes, from the bottom to the right: "rna" (the SHBLE coding full mRNA amount), "prot" (the SHBLE protein amount), "fluo" (the total fluorescence signal), "growth" (proxy of the cellular fitness in absence of antibiotics) and "ic50" (proxy of the cellular fitness in presence of antibiotics).*

Variation in codon usage preferences modifies alternative splice patterns and leads to differences in mRNA levels. The two versions having the most dissimilar CUB with respect to the human average (shble#4 and shble#6) were characterized by splicing events, which were not detected by leading splice site predicting algorithms (60,61), and which reduced the coding potential of the resulting mRNA by 30 to 80%. CUB variations across intron-exon boundaries have been described in several eukaryotes (*e.g.*

372 human, fishes, fruit flies, nematodes, plants (11,63,64)); and splicing regulatory motifs, that can be disrupted
373 by synonymous mutations, have been described in mammals (9,64–67). As signature for selective pressure, a
374 reduced SNP density and decreased rate of synonymous substitutions have further been reported in these
375 regulatory regions (68,69). Thus, selection against mRNA mis-processing can constitute an important
376 selective force that results in concomitant selection for a precise local CUB (70), and this selective force has
377 even been proposed to outperform translational selection in *D. melanogaster* (71).

378 In our experimental setup, variation in mRNA levels between conditions was independent of
379 variation in DNA abundance, ruling out a possible effect of differential transfection efficacy. The two
380 versions with the largest deviation in mRNA levels were the AT-richest ones (shble#3 and #6, respectively
381 the highest and the lowest level of mRNA). Transcript abundance at a given time point is the result of
382 integrating mRNA synthesis and degradation kinetics. As all versions share the same CMV promoter, and 5'
383 untranslated region, we hypothesize that the observed differences in mRNA levels may most probably result
384 from differential mRNA stability and decay, rather than primary transcription regulation. Such an effect has
385 been described for bacteria (*E. coli* (72)), unicellular eukaryotes (*S. cerevisae, S. pombe* (35), *N. crassa, T.*
386 *brucei (73,74)),* and metazoa (fruit fly (75) or zebrafish (76)). The effects on version shble#6 are difficult to
387 address as only 20% of the total transcript contain the customized *shble* sequence; it is thus impossible to
388 disantangle the effect of sequence composition on the full mRNA level from the consequences of the splicing
389 defect. The very high transcript levels of version shble#3 are interesting in the light of recent findings on
390 CUB linked mRNA degradation and/or storage: indeed, AU-rich mRNAs have been found to accumulate in
391 P-bodies, which can lead to an accumulation of those transcript in the cell (77). In addition, the P-body
392 retention of those transcripts reduce their availability to translation and could further explain the reduced
393 protein level for this condition (see below).

394

395 **Variation in mRNA nucleotide composition, codon usage preference and structure lead to**
396 **differences in translation efficiency.** Considering all conditions together, our experimental setup allowed us
397 to determine that the mRNA levels explain only around 40% of the variation in protein levels, which fits well
398 previous descriptions in the literature for a wide diversity of experimental systems (78–80). Such weak
399 explanatory power would not be expected if mRNAs were translated at a constant rate, and has thereby

motivated studies to elucidate which factors are involved in the regulation of translation (81). Here, we evidenced that this discrepancy between mRNA and protein level was unequal between CUB. Particularly, the version using the AT-richer codons among the two most common ones in the average human genome (shble#3) displayed the highest mRNA levels but contrasting low amount of the corresponding protein. A possible explanation for this phenomenon could be the selective translation impairment of AT-rich transcripts. As proposed before, this can result from P-body retention (77), but other mechanisms may alternatively, but not exclusively, be involved. For instance, in human cells, the protein Schlafen11 has been shown to prevent translation of viral transcript (known to be AU-enriched), in a codon usage dependant manner (82,83). Noticing that the AT-rich condition #4 displays a moderated impairment of translation, we interpret that condition #3 dramatic phenotype arises in fact from the combination of suboptimal variables for which a role in optimizing the expression of heterologous genes had already been evidenced (11) : (i) similarity to human average CUB, (ii) the CpG frequency, and (iii) the mRNA folding energy.

(i) Regarding similarity in codon bias between the focal gene and the expression system (i.e. average human CUB), gene versions with a better match resulted in higher protein-to-mRNA ratios. This result is in disagreement with previous reports, as well as with descriptions showing the very limited impact of CUB on gene expression in mammals, compared to other features (30,84). Nevertheless, it is complicated to disentangle the effect of CUB from other composition characteristics, such as GC and GC3 content. It is even more difficult to interpret them in terms of neutralist or selectionist origin, as both evolutionary hypotheses could account for variation in either parameter (10).

(ii) Regarding intragenic CpG frequency, we report a negative impact of extreme values (either too high or too low) on translation. Such direct effect on translation efficiency had never been reported before, and CpG frequency had been shown to impact heterologous protein amount through its impact on *de novo* transcription instead (85,86). More precisely, high CpG depletion was previously associated to low mRNA level, that weren't evidenced as a result of changes in nuclear export, alternative splicing or mRNA stability (85,86). Indeed, a signature for selection towards decreased values of CpG has been consistently reported (87,88), experimentally verified by the detrimental effects of increased CpG levels on protein synthesis (89,90), and further corroborated through experimental evolution (91).

(iii) Regarding the total mRNA folding energy, we also report a negative impact of extreme values (either too high or too low) on the translation. Molecular modeling, along with experimental studies, suggest a prominent role of the initiation, rather than elongation, on the efficacy of translation (41,92,93). And indeed, several studies adressing the impact of mRNA folding on translation, established the importance of the 5' mRNA secondary structure in translation initiation. A shared trend was identified in bacteria, yeast, protists, and mammals (31,55,93–96): a reduced mRNA stability near the site of translation initiation is correlated to a higher protein production. In bacteria and yeast, strong folding around the start codon prevents ribosome recruitment (31,93); and a "ramp" of rare codon along the 50 to 100 first coding nucleotides has been reported, with the effect of reducing mRNA folding energy and with the proposed function of avoiding ribosome traffic jam (96,97). A systematic exploration using 244,000 synthetic DNA sequences on *E. coli* shows that the strength of the secondary structures predicted 60 nucleotides around the start codon is capturing around 36% of the total variance in protein synthesis, while variation in downstream mRNA folding energy accounts only for *ca.* 4-5% (98). Nonetheless, an important role of translation elongation cannot be ruled out. Particularly, in human transcripts, de Sousa Abreu *et al.* describe no effect of the initiation rate on translation efficiency (78). A recent study, in human cell lines, highlight the consequences of the secondary structures along the CDS in the functional half life of mRNA (95). In our experimental setup all constructs share by design the nucleotide sequence around the start codon: the 5'UTR corresponds to the plasmid backbone and the first 24 coding nucleotides are identical (AU1 tag). Thus, there are actually no differences in folding energy when considering only the immediate sequence stretch around the start codon, but there are instead differences when considering the full mRNA length. We interpret thus that our observation of a non-monotonic effect of the full-length mRNA folding energy on the protein-to-mRNA ratio is related to translation elongation impairment rather than to an effect on translation initiation.

Transfected human cells differentially express heterologous genes, independently of their sequence; but sequence recoding impact the individual cell expression in a reproductible way. We report phenotypic variability of transfected human cells, observable as multimodal distribution of cellular fluorescence. The multimodal distribution of cellular fluorescence intensity on the transfected cells could be captured in all cases by fitting to a combination of two Gaussian curves. This pattern was similar for all

455 constructs expressing *egfp*, including the empty control, which contains only the *egfp* sequence, and was thus
456 not influenced by codon usage preferences of the recoded *shble* sequence. We interpret that phenotypic
457 variability in cell-based fluorescence levels reflects phenotypic plasticity and may be related to transient
458 cellular states, such as cell division status and/or differential kinetics of recovery from transfection-induced
459 cellular stress. Cell cycle-dependent differences in gene expression have been actually reported when using
460 cytomegalovirus-based expression vectors (99). Beyond the shared bimodal distribution of fluorescence
461 levels, we observe significant and concerted shifts of both cellular subpopulations towards higher (*e.g.* for
462 the constructs enriched in the most used codons) or lower (*e.g.* for constructs using AT-rich codons) values of
463 fluorescence intensity. Thus, differences in overall EGFP-based fluorescence between recoded constructs do
464 not arise from differences in the number of positive cells expressing a given quantity of EGFP, but rather
465 from differences in EGFP synthesis at the individual cell level. Our experimental model using human cells
466 shows that CUB exerts an important effect on the overall levels but also in the cell-based levels of the
467 heterologous protein produced.
468
469 **We showcase a trade-off between the cellular burden imposed by extra protein synthesis and**
470 **the benefit conferred through antibiotic resistance.** In the absence of antibiotic all transfected cells grow
471 less than the parental cells: *ca.* 2 to 4.8 times less, even for cells transfected with a control plasmid
472 containing an empty expression cassette (i.e. not expressing neither *shble* nor *egfp).* In addition to this basal
473 cost, related to transfection alone, strong heterologous expression imposes an enormous basal burden on the
474 cellular economy. A similar effect has been described in bacteria (100) and the interpretation is consistent
475 with the broad literature about the direct (-cis) and indirect (-trans) costs of translation (70): first, translation
476 is the per-unit most expensive step during biological information flow (101), consuming ca. 45% of the
477 whole energy supply in human cells in culture (102); second, and virtually all ribosomes are bound to mRNA
478 molecules and potentially engaged in translation (102), so that highly-transcribed heterologous mRNA
479 increase overall ribosome demand and cause loss of opportunity for cellular gene translation; and third,
480 heterologous protein synthesis can lead to additional downstream costs by protein folding, protein
481 degradation and off-target effects of mistranslated proteins (45,103,104). Additionally, mismatch between the
482 CUB of the heterologous gene can display strong trans-effects on the cellular homeostasis, by sequestering

ribosomes onto mRNAs that hardly progress over translation but also by creating a competition for the tRNA pools (31,105). Scarcity of the less common tRNAs is actually a severe limiting factor for protein synthesis in bacteria (106), and this pressure over rare tRNAs can become extreme in conditions of stress, or changes in nutritional status (10,107,108).

The *shble* gene that we have used as a base for synonymous recoding encodes for a small protein that confers resistance to bleomycin through antibiotic sequestering (109). This protein-antibiotic binding is equimolecular and reversible: a SHBLE protein dimer binds two bleomycin molecules (110). It can be thus hypothesised that the strength of the antibiotic resistance conferred is a direct, probably monotonic, function of the SHBLE amount produced. Our results, however, show that the benefit conferred by SHBLE synthesis in the presence of antibiotic is largely exceeded by the cost and burden of heterologous protein synthesis. Thus, as described above for the fitness in the absence of antibiotic, we state an important trade-off between the intensity of heterologous SHBLE+EGFP protein synthesis and the actual bleomycin resistance levels; and an important rescue of the antibiotic resistance for cells expressing SHBLE but not EGFP. Altogether, we conclude that: (i) transfection alone introduces an important cellular stress that impairs cellular growth; (ii) heterologous gene overexpression imposes a strong burden on the cell economy, sufficient to severely affect cell growth; (iii) CUB of heterologous gene differentially impact cellular fitness as a function of the differences in heterologous protein synthesis.

CONCLUSION

The main conundrum for scientists approaching codon usage bias remains the contrast between, on the one hand, the large and sound body of knowledge showing the strong molecular and cellular impact of gene expression differences arising from codon usage preferences and, on the other hand, the thin evidence for codon usage selection at the organismal level. Under the neutral hypothesis, differences in average genome CUB can be explained by biochemical biases during DNA synthesis or repair (e.g. polymerase bias) (111); and, in vertebrates, CUB at the gene level may be shaped by their relative position to isochores (e.g. alternation between GC-rich and AT-rich stretches along the chromosomes) (112). Further, GC-biased gene conversion mechanisms can enhance those local variations (111,113,114). The selective explanation, often referred to as "translational selection", proposes that different codons may led to differences in gene

expression, by changes in alternative splicing patterns, mRNA localisation or stability, translation efficiency, or protein folding (115). If such codon-bias induced variation in gene expression were associated with phenotypic variation that results in fitness differences, it may, by definition, be subject to natural selection. Nevertheless, differences in fitness associated with individual synonymous changes seem to be mostly of low magnitude, so that selection may only act effectively in organisms with large population sizes (116) such as bacteria (7), yeast (117), nematodes (118), but also in fruit flies (19,20,119,120), branchiopods (121) and amphibians (122). In organisms with small population sizes, such as mammals, and particularly humans, evidences of selection for (or against) certain codons remain nevertheless controversial (22). In the present manuscript, we have intended to contribute to this debate by exploring the phenotypic consequences of codon usage differences of heterologous genes in human cells. We claim that the potential evolutionary forces at play in shaping human CUB, select for a strict control of mRNA processing: splicing, and secondary structure (potentially affecting stability and decay); and that the resulting mRNA properties *in fine* impact translation elongation. Notwithstanding, the disparity between predictions and findings encountered in powerful, codon-usage related experimental evolution approaches highlights the gap in our understanding at connecting phenotype and fitness over different integration levels: molecules, cells, tissues and organisms. Despite, or thanks to, the immense body of knowledge accumulated over the last fifty years, the quest for interpreting and integrating the riddle of codon usage preferences over broad scales of time and biological complexity remains tempting and unsolved.

MATERIAL AND METHODS

Design of the *shble* synonymous versions and plasmid constructs. Six synonymous versions of the *shble* gene were designed applying the "one amino acid - one codon" approach, *i.e.*, all instances of one amino acid in the *shble* sequence were recoded with the same codon, depending on their frequency in the human genome (Table 1): shble#1 used the most frequent codons in the human genome; shble#2 used the GC-richest among the two most frequent codons; shble#3 used the AT-richest among the two most frequent codons; shble#4 used the least frequent codons; shble#5 used the GC-richest among the two less frequent codons; and shble#6 used the AT-richest among the two less frequent codons. An invariable *AU1* sequence

was added as N-terminal tag (amino acid sequence MDTYRI) to all six versions. Nucleotide contents between versions are compared in Sup. Table 1. The normalized COUSIN 18 score (COdon Usage Similarity Index), which compares the CUB of a query against a reference, was calculated on the online tool (http://cousin.ird.fr) (59). A score value below 0 informs that the CUB of the query sequence is opposite to the reference CUB; a value close to 1 informs that the query CUB is similar to the reference CUB , and a value above 1 informs that the query CUB is similar the reference CUB , but of larger magnitude (59). All *shble* synonymous sequences were chemically synthesised and cloned on the *XhoI* restriction site in the pcDNA3.1+P2A-EGFP plasmid (InvitroGen), in-frame with the *P2A-EGFP* reporter cassette. In this plasmid, the expression of the reporter gene is located under the control of the strong human cytomegalovirus (CMV) promoter and terminated by the bovine growth hormone polyadenylation signal. All constructs encode for a 1,602 bp transcript, encompassing a 1,182 bp *au1-shble-P2A-EGFP* coding sequence (Sup. Fig. 1). The folding energy of the 1,602 bp transcripts was calculated on the RNAfold Webserver (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi), with default parameters (Table 1). During translation, the P2A peptide (sequence NPGP) induces ribosome skipping (123), meaning that the ribosome does not perform the transpeptidation bond and releases instead the AU1-SHBLE moiety and continues translation of the EGFP moiety. The HEK293 human cell line used here is proficient at performing ribosome skipping on the P2A peptide (124) The transcript encodes thus for one single coding sequence but translation results in the production of two proteins: SHBLE (theoretical molecular mass 17.2 kDa) and EGFP (27.0 kDa). As controls we used two plasmids: (i) pcDNA3.1+P2A-EGFP (named here "empty"), which encodes for the EGFP protein; (ii) pcDNA3.1+ (named here "superempty") which does not express any transcript from the CMV promoter (Table 1). In order to explore the burden of EGFP expression we generated two additional constructs by subcloning the AU1-tagged shble#1 and shble#4 coding sequences in the XhoI restriction site of the pcDNA3.1+ backbone, resulting in the constructs shble#1* and shble#4*, lacking the *P2A-EGFP* sequence.

Transfection and differential cell sampling. As mentioned above, all experiments were carried out on HEK293 cells. Cell culture conditions, transfection methods and related reagents are detailed in Sup. Methods 2.2. Cells were harvested two days after transfection and submitted to analyses at four levels

(Figure 6): (i) nucleic acid analyses (qPCR and RNAseq); (ii) proteomics (label-free quantitative mass spectrometry analysis and western blot immuno-assays); (iii) flow cytometry; and (iv) real-time cell growth analysis (RTCA). Overall, the different experiments were performed on 33 biological replicates, corresponding to a variable number of repetitions depending on the considered analysis (Sup. Method 1). Transfection efficiency was evaluated by means of qPCR targeting two invariable regions of the plasmid and revealed no significant differences between the constructs (Sup. Methods 2.3).

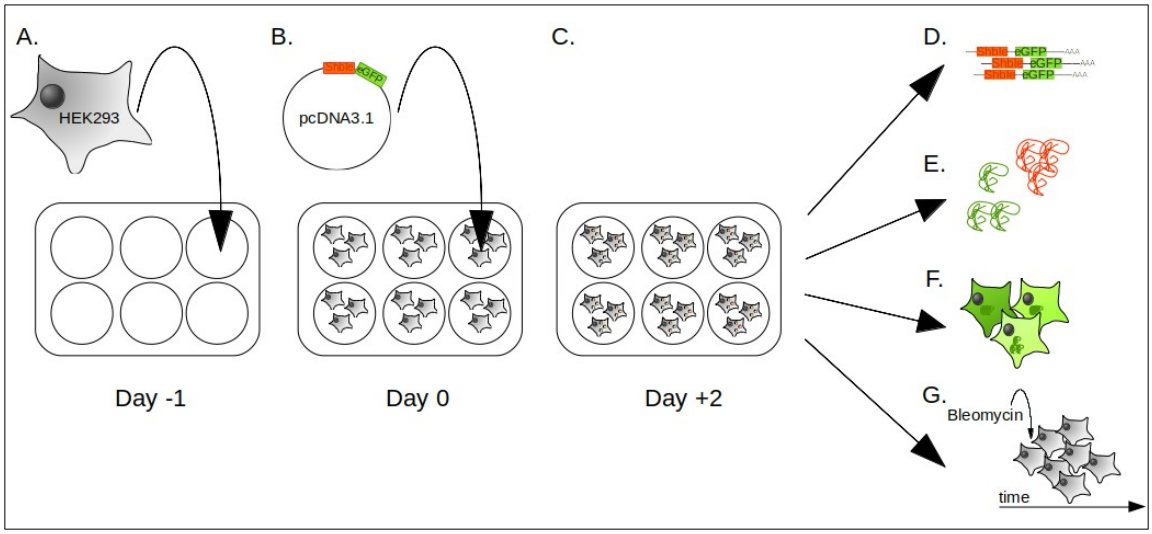

Figure 7. Overview of the sampling protocol and the measured phenotypes. HEK293 cells were seeded on 6-well plates (A) one day before transfection with the customized pcDNA3.1 plasmids (B). Transfected cells were harvested two days later (C). mRNA levels were assessed by RNAseq (D), protein levels were measured by label-free proteomics (E), EGFP fluorescence was assessed at the single cell level by flow cytometry (F) and cell growth was assessed by xCELLigence RTCA (Real Time Cell growth Analysis) in presence of different concentrations of the bleomycin antibiotic (G).

RNA sequencing and data analysis. The transcriptomic analysis was performed on six biological replicates and eight conditions: shble#1 to shble#6, #empty, and mock (for which the sample is submitted to the exact same procedures, including the transfection agent, but in absence of plasmid). Paired 150bp Illumina reads were trimmed (Trimmomatic v0.38) (125) and mapped on eight different genomic references (HISAT2 v2.1.0) (126), corresponding to the concatenation of the human reference genome

(GCF_000001405.38_GRCh38.p12_genomic.fna, NCBI database, 7th of February 2019) and the corresponding full sequence of the plasmid. For the mock condition, we considered the human genome and all possible versions of the plasmid. Virtually no read of those negative controls mapped to the plasmid sequences. For all other conditions, read distribution patterns along the plasmid sequence were evaluated with IGVtool (127). In all cases the *au1-shble-p2a-EGFP* coding sequence displayed highly similar coverage shape for all constructs, except for shble#4 and shble#6 for which respectively one and two alternative splicing events were observed (Sup. Fig. 3 and 4). None of these splice sites were predicted when the theoretical transcripts were evaluated using *Human Splicing Finder* (HSF, accessed via https://www.genomnis.com/access-hsf) (60), or with *SPLM - Search for human potential splice sites using weight matrices* (accessed via http://www.softberry.com/) (61). When relevant, the three alternative transcript isoforms identified were further used as reference for read pseudomapping and quantification with Kallisto (v0.43.1) (128). Details on RNA preparation and bioinformatic pipeline are provided in Sup. Methods 2.4 and Sup. Methods 3.

Label-free proteomic analysis. The label-free proteomic was performed on nine biological replicates (three of them measured independently, and six pooled by two), and eight different conditions: shble#1 to shble#6, #empty, and mock. 20 to 30 µg of proteins were in-gel digested and resulting peptides were analysed online using a Q Exactive HF mass spectrometer coupled with an Ultimate 3000 RSLC system (Thermo Fisher Scientific). MS/MS analyses were performed using the Maxquant software (v1.5.5.1) (129). All MS/MS spectra were searched by the Andromeda search engine (130) against a decoy database consisting in a combination of *Homo sapiens* entries from Reference Proteome (UP000005640, release 2019_02, https://www.uniprot.org/), a database with classical contaminants, and the sequences of interest (SHBLE and EGFP). After excluding the usual contaminants, we obtained a final set of 4,302 proteins detected at least once in one of the samples. Intensity based absolute quantification (iBAQ) was used to compare protein levels between samples (131).

Western blot immunoassays and semi-quantitative analysis. Western blot immunoassays were performed on nine replicates and nine conditions: shble#1 to shble#6, #empty, #superempty, and mock. Three different proteins were targetted: β-TUBULIN, EGFP, and SHBLE (*via* the invariable AU1 epitope

612 tag). Semi-quantitative analysis from enzyme chemoluminiscence data was performed with ImageJ (132) by
613 «plotting lanes» to obtain relative density plots (Sup. Fig. 7).

614 **Flow cytometry analysis.** Flow cytometry experiments were performed on a NovoCyte flow
615 cytometer system (ACEA biosciences). 50,000 ungated events were acquired with the NovoExpress
616 software, and further filtering of debris and doublets was performed in R with an in-house script (filtering
617 strategy is detailed in Sup. Method 2.7). For subsequent analysis, 30,000 events were randomly picked up
618 from each sample. Seven samples had less than 30,000 events and, in order to ensure the same sample size
619 for all conditions, the four corresponding replicates were excluded. After a first visualization of the data, two
620 replicates were ruled out because they displayed a typical pattern of failed transfection for the condition
621 shble#1 (Sup. Method 2.7), resulting in 16 final replicates being fully examined.

622 **Real time cell growth analysis (RTCA).** RTCA was carried out on an xCELLigence system for the
623 mock and the superempty controls, and further eight constructs: the previously analysed shble#1 to shble#6,
624 plus the shble#1* and shble#4* lacking the *EGFP* reporter gene. Cells were grown under different
625 concentrations of the Bleomycin antibiotic ranging from 0 to 5000 µg/mL (Sup. Method 2.8). Three to six
626 biological replicates were performed, including technical duplicates for each replicate. Cells were grown on
627 microtiter plates with interdigitated gold electrodes that allow to estimate cell density by means of
628 impedance measurement. Measures were acquired every 15 minutes, over 70 hours (280 time points).
629 Impedance measurements are reported as "Cell Index" values, which are compared to the initial baseline
630 values to estimate changes in cellular performance linked to the expression of the different constructs. For
631 each construct we estimated first cellular fitness by calculating the area below the curve for the delta-Cell
632 index *vs* time for the cells grown in the absence of antibiotics. We estimated then the ability to resist the
633 antibiotic conferred by each construct through calculation of IC50 as the bleomycin concentration that
634 reduces the area below the curve to half of the one estimated in the absence of antibiotics (detailled methods
635 in Sup. Method 2.8).

636 **Data availability.** RNAseq raw reads were deposited on the NCBI-SRA database under the
637 BioProject number PRJNA753061. R scripts and input files are available at
638 https://github.com/malpicard/synonymous-but-not-neutral.git.

AKNOWLEDGEMENTS

We acknowledge the IRD itrop HPC (South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. We also acknowledge the facilities of the Functional Proteomics Platform (FPP) of the Proteomics Pole of Montpellier (PPM, Montpellier France); and the MRI imaging facility, member of the France-BioImaging national infrastructure supported by the French National Research Agency (ANR-10-INBS-04, «Investments for the future»).

REFERENCES

1. Crick F. Central Dogma of Molecular Biology. Nature. 1970;227(5258):561–3.
2. Grantham R, Gautier C, Gouy M, Mercier R, Pavé A. Codon catalog usage and the genome hypothesis. Nucleic Acids Res. 1980 Jan 11;8(1):197–197.
3. Ikemura T. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol. 1982 Jul 15;158(4):573–97.
4. Kanaya S, Yamada Y, Kudo Y, Ikemura T. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene. 1999 Sep 30;238(1):143–55.
5. Novoa EM, Jungreis I, Jaillon O, Kellis M, Leitner T. Elucidation of Codon Usage Signatures across the Domains of Life. Mol Biol Evol. 2019 Oct 1;36(10):2328–39.
6. Gouy M, Gautier C. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 1982 Nov 25;10(22):7055–74.
7. Sharp PM, Li WH. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol. 1986 Dec;24(1):28–38.
8. Duret L. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev. 2002 Dec 1;12(6):640–9.
9. Chamary J V., Parmley JL, Hurst LD. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 2006 72. 2006 Feb;7(2):98–108.
10. Hanson G, Coller J. Codon optimality, bias and usage in translation and mRNA decay. Nat Rev Mol Cell Biol 2017 191. 2017 Oct 11;19(1):20–30.
11. Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 2011 121. 2010 Nov 23;12(1):32–42.
12. Mauro VP, Chappell SA. A critical analysis of codon optimization in human therapeutics. Trends Mol Med. 2014 Nov 1;20(11):604–13.
13. Angov E, Hillier CJ, Kincaid RL, Lyon JA. Heterologous Protein Expression Is Enhanced by Harmonizing the Codon Usage Frequencies of the Target Gene with those of the Expression Host. PLoS One. 2008 May 14;3(5):e2189.

14. Fath S, Bauer AP, Liss M, Spriestersbach A, Maertens B, Hahn P, et al. Multiparameter RNA and Codon Optimization: A Standardized Tool to Assess and Enhance Autologous Mammalian Gene Expression. PLoS One. 2011;6(3):e17596.

15. Martínez MA, Jordan-Paiz A, Franco S, Nevot M. Synonymous Virus Genome Recoding as a Tool to Impact Viral Fitness. Trends Microbiol. 2016 Feb 1 ;24(2):134–47.

16. Ikemura T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol. 1981 Sep 25;151(3):389–409.

17. Dong H, Nilsson L, Kurland CG. Co-variation of tRNA Abundance and Codon Usage inEscherichia coliat Different Growth Rates. J Mol Biol. 1996 Aug 2;260(5):649–63.

18. Duret L. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. Trends Genet. 2000 Jul 1;16(7):287–9.

19. Moriyama EN, Powell JR. Codon usage bias and tRNA abundance in Drosophila. J Mol Evol. 1997;45(5):514–23.

20. Akashi H. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics. 1994 Mar 1;136(3):927–35.

21. Powell JR, Moriyama EN. Evolution of codon usage bias in Drosophila. Proc Natl Acad Sci. 1997 Jul 22;94(15):7784–90.

22. Urrutia AO, Hurst LD. Codon Usage Bias Covaries With Expression Breadth and the Rate of Synonymous Evolution in Humans, but This Is Not Evidence for Selection. Genetics. 2001 Nov 1;159(3):1191–9.

23. Lithwick G, Margalit H. Hierarchy of Sequence-Dependent Features Associated With Prokaryotic Translation. Genome Res. 2003 Dec 1;13(12):2665–73.

24. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, et al. Global analysis of protein expression in yeast. Nature. 2003 Oct 16;425(6959):737–41.

25. Tuller T, Kupiec M, Ruppin E. Determinants of Protein Abundance and Translation Efficiency in S. cerevisiae. PLOS Comput Biol. 2007 Dec;3(12):e248.

26. Burgess-Brown NA, Sharma S, Sobott F, Loenarz C, Oppermann U, Gileadi O. Codon optimization can improve expression of human genes in Escherichia coli: A multi-gene study. Protein Expr Purif. 2008 May 1;59(1):94–102.

27. Lampson BL, Pershing NLK, Prinz JA, Lacsina JR, Marzluff WF, Nicchitta C V., et al. Rare Codons Regulate KRas Oncogenesis. Curr Biol. 2013 Jan 7;23(1):70–5.

28. Pop C, Rouskin S, Ingolia NT, Han L, Phizicky EM, Weissman JS, et al. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. Mol Syst Biol. 2014 Dec 1;10(12):770.

29. Li GW, Burkhardt D, Gross C, Weissman JS. Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. Cell. 2014 Apr 24;157(3):624–35.

30. Vogel C, De Sousa Abreu R, Ko D, Le SY, Shapiro BA, Burns SC, et al. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Mol Syst Biol. 2010 Jan 1;6(1):400.

31. Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of expression in escherichia coli. Science. 2009 Apr 10;324(5924):255–8.

32. Agashe D, Martinez-Gomez NC, Drummond DA, Marx CJ. Good Codons, Bad Transcript: Large Reductions in Gene Expression and Fitness Arising from Synonymous Mutations in a Key Enzyme. Mol Biol Evol. 2013 Mar 1;30(3):549–60.

33. Zucchelli E, Pema M, Stornaiuolo A, Piovan C, Scavullo C, Giuliani E, et al. Codon Optimization Leads to Functional Impairment of RD114-TR Envelope Glycoprotein. Mol Ther - Methods Clin Dev. 2017 Mar 17;4:102–14.

34. Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, et al. Codon Optimality Is a Major Determinant of mRNA Stability. Cell. 2015 Mar 12;160(6):1111–24.

35. Harigaya Y, Parker R. Analysis of the association between codon optimality and mRNA stability in Schizosaccharomyces pombe. BMC Genomics. 2016 Nov 8;17(1):1–16.

36. Radhakrishnan A, Green R. Connections Underlying Translation and mRNA Stability. J Mol Biol. 2016 Sep 11;428(18):3558–64.

37. Radhakrishnan A, Chen YH, Martin S, Alhusaini N, Green R, Coller J. The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. Cell. 2016 Sep 22;167(1):122-132.e9.

38. Chen S, Li K, Cao W, Wang J, Zhao T, Huan Q, et al. Codon-Resolution Analysis Reveals a Direct and Context-Dependent Impact of Individual Synonymous Mutations on mRNA Level. Mol Biol Evol. 2017 Nov 1;34(11):2944–58.

39. Bettany AJE, Moore PA, Cafferkey R, Bell LD, Goodey AR, Carter BLA, et al. 5′-Secondary structure formation, in constrast to a short string of non-preferred codons, inhibits the translation of the pyruvate kinase mRNA in yeast. Yeast. 1989 May 1;5(3):187–98.

40. De Smit MH, Van Duin J. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. Proc Natl Acad Sci. 1990 Oct 1;87(19):7668–72.

41. Gu W, Zhou T, Wilke CO. A Universal Trend of Reduced mRNA Stability near the Translation-Initiation Site in Prokaryotes and Eukaryotes. PLOS Comput Biol. 2010 Feb;6(2):e1000664.

42. Tuller T, Waldman YY, Kupiec M, Ruppin E. Translation efficiency is determined by both codon bias and folding energy. Proc Natl Acad Sci. 2010 Feb 23;107(8):3645–50.

43. Marais G, Duret L. Synonymous Codon Usage, Accuracy of Translation, and Gene Length in Caenorhabditis elegans. J Mol Evol. 2001;52(3):275–80.

44. Kurland CG. Translational Accuracy and the Fitness of Bacteria. Annu Rev Genet. 2003 Nov 28;26:29–50.

45. Stoletzki N, Eyre-Walker A. Synonymous Codon Usage in Escherichia coli: Selection for Translational Accuracy. Mol Biol Evol. 2007 Feb 1;24(2):374–81.

46. Johnston TC, Borgia PT, Parker J. Codon specificity of starvation induced misreading. Mol Gen Genet. 1984 Jul;195(3):459–65.

47. Johnston TC, Parker J. Streptomycin-induced, third-position misreading of the genetic code. J Mol Biol. 1985 Jan 20;181(2):313–5.

48. Robinson M, Lilley R, Little S, Emtage JS, Yarranton G, Stephens P, et al. Codon usage can affect efficiency of translation of genes in Escherichia coli. Nucleic Acids Res. 1984 Sep 11;12(17):6663–71.

49. Sørensen MA, Kurland CG, Pedersen S. Codon usage determines translation rate in Escherichia coli. J Mol Biol. 1989 May 20;207(2):365–77.

50. Xia X. A major controversy in codon-Anticodon adaptation resolved by a new codon usage index. Genetics. 2014 Feb 1;199(2):573–9.

51. Sørensen MA, Pedersen S. Absolute in vivo translation rates of individual codons in Escherichia coli: The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. J Mol Biol. 1991 Nov 20;222(2):265–80.

52. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009 Apr 10;324(5924):218–23.

53. Hussmann JA, Patchett S, Johnson A, Sawyer S, Press WH. Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. PLOS Genet. 2015;11(12):e1005732.

54. Gardin J, Yeasmin R, Yurovsky A, Cai Y, Skiena S, Futcher B. Measurement of average decoding rates of the 61 sense codons in vivo. Elife. 2014;3.

55. Weinberg DE, Shah P, Eichhorn SW, Hussmann JA, Plotkin JB, Bartel DP. Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. Cell Rep. 2016 Feb 23;14(7):1787–99.

56. Pechmann S, Frydman J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat Struct Mol Biol. 2012 Dec 23;20(2):237–43.

57. Chaney JL, Steele A, Carmichael R, Rodriguez A, Specht AT, Ngo K, et al. Widespread position-specific conservation of synonymous rare codons within coding sequences. PLOS Comput Biol. 2017 May 1;13(5):e1005531.

58. Zhao F, Yu CH, Liu Y. Codon usage regulates protein structure and function by affecting translation elongation speed in Drosophila cells. Nucleic Acids Res. 2017 Aug 21;45(14):8484–92.

59. Bourret J, Alizon S, Bravo IG. COUSIN (COdon Usage Similarity INdex): A Normalized Measure of Codon Usage Preferences. Genome Biol Evol. 2019 Dec 1;11(12):3523–8.

60. Desmet FO, Hamroun D, Lalande M, Collod-Bëroud G, Claustres M, Béroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. 2009;37(9):e67.

61. Solovyev V. Statistical Approaches in Eukaryotic Gene Prediction. Handb Stat Genet. 2004 Jul 15

62. Jallet AJ, Demange A, Leblay F, Decourcelle M, Koulali K El, Picard M AL, et al. Human cellular homeostasis buffers trans-acting translational effects of heterologous gene expression with very different codon usage bias. bioRxiv. 2021 Dec 10;2021.12.09.471957.

63. Willie E, Majewski J. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. Trends Genet. 2004 Nov 1;20(11):534–8.

64. Eskesen ST, Eskesen FN, Ruvinsky A. Natural Selection Affects Frequencies of AG and GT Dinucleotides at the 5′ and 3′ Ends of Exons. Genetics. 2004 May 1;167(1):543–50.

65. Parmley JL, Hurst LD. Exonic Splicing Regulatory Elements Skew Synonymous Codon Usage near Intron-exon Boundaries in Mammals. Mol Biol Evol. 2007 Aug 1;24(8):1600–3.

66. Fairbrother WG, Yeh RF, Sharp PA, Burge CB. Predictive identification of exonic splicing enhancers in human genes. Science. 2002 Aug 9;297(5583):1007–13.

67. Louie E, Ott J, Majewski J. Nucleotide Frequency Variation Across Human Genes. Genome Res. 2003 Dec 1;13(12):2594–601.

68. Chamary JV, Hurst LD. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? Trends Genet. 2005 May 1;21(5):256–9.

69. Orban T, Olah E. Purifying selection on silent sites – a constraint from splicing regulation? Trends Genet. 2001 May 1;17(5):252–3.

70. Callens M, Pradier L, Finnegan M, Rose C, Bedhomme S. Read between the Lines: Diversity of Nontranslational Selection Pressures on Local Codon Usage. Genome Biol Evol. 2021 Sep 1;13(9).

71. Warnecke T, Hurst LD. Evidence for a Trade-Off between Translational Efficiency and Splicing Regulation in Determining Synonymous Codon Usage in Drosophila melanogaster. Mol Biol Evol. 2007 Dec 1;24(12):2755–62.

72. Boël G, Letso R, Neely H, Price WN, Wong KH, Su M, et al. Codon influence on protein expression in E. coli correlates with mRNA levels. Nat 2016 5297586. 2016 Jan 13;529(7586):358–63.

73. Jeacock L, Faria J, Horn D. Codon usage bias controls mRNA and protein abundance in trypanosomatids. Elife. 2018 Mar 15;7.

74. Nascimento J de F, Kelly S, Sunter J, Carrington M. Codon choice directs constitutive mRNA levels in trypanosomes. Elife. 2018 Mar 15;7.

75. Burow DA, Martin S, Quail JF, Alhusaini N, Coller J, Cleary MD. Attenuated Codon Optimality Contributes to Neural-Specific mRNA Decay in Drosophila. Cell Rep. 2018 Aug 14;24(7):1704–12.

76. Mishima Y, Tomari Y. Codon Usage and 3′ UTR Length Determine Maternal mRNA Stability in Zebrafish. Mol Cell. 2016 Mar 17;61(6):874–85.

77. Courel M, Clément Y, Bossevain C, Foretek D, Cruchez OV, Yi Z, et al. Gc content shapes mRNA storage and decay in human cells. Elife. 2019 Dec 1;8.

78. De Sousa Abreu R, Penalva LO, Marcotte EM, Vogel C. Global signatures of protein and mRNA expression levels. Mol Biosyst. 2009 Nov 12;5(12):1512–26.

79. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet 2012 134. 2012 Mar 13;13(4):227–32. Available from: https://www.nature.com/articles/nrg3185

80. Schwanhüusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. Nature. 2011 May 19 ;473(7347):337–42.

81. Ron Milo, Rob Phillips. Cell Biology by the Numbers. Garland Science. 2019.

82. Li M, Kao E, Gao X, Sandig H, Limmer K, Pavon-Eternod M, et al. Codon-usage-based inhibition of HIV protein synthesis by human schlafen 11. Nature. 2012 Nov 1 ;491(7422):125–8.

83. Stabell AC, Hawkins J, Li M, Gao X, David M, Press WH, et al. Non-human Primate Schlafen11 Inhibits Production of Both Host and Viral Proteins. PLoS Pathog. 2016 Dec 27 ;12(12).

84. Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat Biotechnol. 2006 Dec 24;25(1):117–24.

85. Bauer AP, Leikam D, Krinner S, Notka F, Ludwig C, Längst G, et al. The impact of intragenic CpG content on gene expression. Nucleic Acids Res. 2010 Jul 1;38(12):3891–908.

86. Krinner S, Heitzer AP, Diermeier SD, Obermeier I, Längst G, Wagner R. CpG domains downstream of TSSs promote high levels of gene expression. Nucleic Acids Res. 2014 ;42(6):3551–64.

87. Simmonds P, Xia W, Baillie JK, McKinnon K. Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla -selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. BMC Genomics. 2013 Sep 10;14(1):1–16.

88. Greenbaum BD, Levine AJ, Bhanot G, Rabadan R. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog. 2008 Jun;4(6).

89. Burns CC, Campagnoli R, Shaw J, Vincent A, Jorba J, Kew O. Genetic Inactivation of Poliovirus Infectivity by Increasing the Frequencies of CpG and UpA Dinucleotides within and across Synonymous Capsid Region Codons. J Virol. 2009;83(19):9957–69.

90. Tulloch F, Atkinson NJ, Evans DJ, Ryan MD, Simmonds P. RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. Elife. 2014;3:e04531.

91. Lauring AS, Acevedo A, Cooper SB, Andino R. Codon Usage Determines the Mutational Robustness, Evolutionary Capacity, and Virulence of an RNA Virus. Cell Host Microbe. 2012 Nov 15;12(5):623–32.

92. Riba A, Nanni N Di, Mittal N, Arhné E, Schmidt A, Zavolan M. Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates. Proc Natl Acad Sci U S A. 2019 Jul 23;116(30):15023–32.

93. Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. Rate-Limiting Steps in Yeast Protein Translation. Cell. 2013 Jun 20;153(7):1589–601.

94. Wang SE, Brooks AES, Poole AM, Simoes-Barbosa A. Determinants of translation efficiency in the evolutionarily-divergent protist Trichomonas vaginalis. BMC Mol Cell Biol. 2020 Jul 20;21(1):1–13.

95. Mauger DM, Joseph Cabral B, Presnyak V, Su S V., Reid DW, Goodman B, et al. mRNA structure regulates protein expression through changes in functional half-life. Proc Natl Acad Sci U S A. 2019 Nov 26;116(48):24075–83.

96. Bentele K, Saffert P, Rauscher R, Ignatova Z, Blüthgen N. Efficient translation initiation dictates codon usage at gene start. Mol Syst Biol. 2013 Jan 1;9(1):675.

97. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, et al. An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. Cell. 2010 Apr 16;141(2):344–54.

98. Cambray G, Guimaraes JC, Arkin AP. Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli. Nat Biotechnol. 2018 Nov 1;36(10):1005.

99. Brightwell G, Poirier V, Cole E, Ivins S, Brown KW. Serum-dependent and cell cycle-dependent expression from a cytomegalovirus-based mammalian expression vector. Gene. 1997 Jul 18;194(1):115–23.

100. Amorós-Moya D, Bedhomme S, Hermann M, Bravo IG. Evolution in regulatory regions rapidly compensates the cost of nonoptimal codon usage. Mol Biol Evol. 2010 Sep;27(9):2141–51.

101. Lynch M, Marinov GK. The bioenergetic costs of a gene. Proc Natl Acad Sci U S A. 2015 Dec 22;112(51):15690–5.

102. Princiotta MF, Finzi D, Qian SB, Gibbs J, Schuchmann S, Buttgereit F, et al. Quantitating Protein Synthesis, Degradation, and Endogenous Antigen Processing. Immunity. 2003 Mar 1;18(3):343–54.

103. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci. 2005 Oct 4;102(40):14338–43.

104. Ribas de Pouplana L, Santos MAS, Zhu JH, Farabaugh PJ, Javid B. Protein mistranslation: friend or foe? Trends Biochem Sci. 2014 Aug 1;39(8):355–62.

105. Andersson SGE, Kurland CG. Codon preferences in free-living microorganisms. Microbiol Rev. 1990 Jun;54(2):198–210.

106. Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM, Pilpel Y. Codon usage of highly expressed genes affects proteome-wide translation efficiency. Proc Natl Acad Sci U S A. 2018 May 22;115(21):E4940–9.

107. Dittmar KA, Sørensen MA, Elf J, Ehrenberg M, Pan T. Selective charging of tRNA isoacceptors induced by amino-acid starvation. EMBO Rep. 2005 Feb 1;6(2):151–7.

108. Elf J, Nilsson D, Tenson T, Ehrenberg M. Selective charging of tRNA isoacceptors explains patterns of codon usage. Science. 2003 Jun 13;300(5626):1718–22.

109. Gatignol A, Durand H, Tiraby G. Bleomycin resistance conferred by a drug-binding protein. FEBS Lett. 1988 Mar 28;230(1–2):171–5.

110. Dumas P, Bergdoll M, Cagnon C, Masson JM. Crystal structure and site-directed mutagenesis of a bleomycin resistance protein and their significance for drug sequestering. EMBO J. 1994;13(11):2483.

111. Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci. 2004 Mar 9;101(10):3480–5.

112. Caspersson T, Farber S, Foley GE, Kudynowski J, Modest EJ, Simonsson E, et al. Chemical differentiation along metaphase chromosomes. Exp Cell Res. 1968 Jan 1;49(1):219–22.

113. Galtier N, Piganeau G, Mouchiroud D, Duret L. GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. Genetics. 2001 Oct 1;159(2):907–11.

114. Duret L, Galtier N. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. Annual review of genomics and human genetics. 2009 Aug 28;10:285–311.

115. Chaney JL, Clark PL. Roles for Synonymous Codon Usage in Protein Biogenesis. Annu Rev Biophys. 2015;44:143–66.

116. Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, et al. Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. Mol Biol Evol. 2018 May 1;35(5):1092–103.

117. Sharp PM, Tuohy TMF, Mosurski KR. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 1986 Jul 11;14(13):5125–43.

118. Stenico M, Lloyd AT, Sharp PM. Codon usage in Caenorhabditis elegans : delineation of translational selection and mutational biases. Nucleic Acids Res. 1994 Jul 11;22(13):2437–46.

119. Shields DC, Sharp PM, Higgins DG, Wright F. Silent sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol. 1988 Jul 1;5(6):704–16.

120. Bierne N, Eyre-Walker A. Variation in synonymous codon use and DNA polymorphism within the Drosophila genome. J Evol Biol. 2006 Jan 1;19(1):1–11.

121. Lynch M, Gutenkunst R, Ackerman M, Spitze K, Ye Z, Maruki T, et al. Population genomics of Daphnia pulex. Genetics. 2017 May 1;206(1):315–32.

122. Musto H, Cruveiller S, D'Onofrio G, Romer H, Bernardi G. Translational Selection on Codon Usage in Xenopus laevis. Mol Biol Evol. 2001 Sep 1;18(9):1703–7.

123. Ryan MD, King AMQ, Thomas GP. Cleavage of foot-and-mouth disease virus polyprotein is mediated by residues located within a 19 amino acid sequence. J Gen Virol. 1991 Nov 1;72(11):2727–32.

124. Kim JH, Lee SR, Li LH, Park HJ, Park JH, Lee KY, et al. High Cleavage Efficiency of a 2A Peptide Derived from PorcineTeschovirus-1 in Human Cell Lines, Zebrafish and Mice. PLoS One. 2011;6(4).

125. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–20.

126. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015 Mar 9;12(4):357–60.

127. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011 Jan 1;29(1):24–6.

128. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016 Apr 4;34(5):525–7.

129. Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc. 2016 Dec 27;11(12):2301–19.

130. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen J V., Mann M. Andromeda: A peptide search engine integrated into the MaxQuant environment. J Proteome Res. 2011 Apr 1;10(4):1794–805.

131. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods. 2016 Jun 27;13(9):731–40.

132. Rueden CT, Schindelin J, Hiner MC, DeZonia BE, Walter AE, Arena ET, et al. ImageJ2: ImageJ for the next generation of scientific image data. BMC Bioinformatics. 2017 Nov 29;18(1):1–26.