The non-neutral, multi-level, phenotypic impact of synonymous mutations revealed through heterologous gene expression in human cells.

Marion AL Picard[1,*], Fiona Leblay[1], Cécile Cassan[1], Anouk Willemsen[1], Frédérique Bauffe[1], Mathilde Decourcelle[2], Antonin Demange[1], Ignacio G Bravo[1,*]

AUTHOR AFFILIATIONS

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

[2]Institut de Génomique Fonctionelle (BCM, Univ. Montpellier, CNRS, INSERM,) Montpellier, France

*Correspondence should be adressed to: marion.picard@obs-banyuls.fr, ignacio.bravo@cnrs.fr

AUTHOR CONTRIBUTIONS

IGB conceived the study and obtained funding; MALP, FL, CC, AW, FB, MD, and AD performed experiments and processed primary data; MALP and IGB analyzed the data, MALP and IGB conceptualized the results and wrote the manuscript.

KEY WORDS

codon usage bias, evolution, diversity, noise, fitness, transcriptomic, proteomic, heterogeneity, trade-off.

ABSTRACT

Redundancy in the genetic code allows for differences in transcription and/or translation efficiency between mRNA molecules carrying synonymous polymorphisms, with potential phenotypic impact at the molecular and at the organismal level. A combination of neutral and selective processes determines the global genome codon usage preferences, as well as local differences between genes within a genome and between positions along a single gene. The relative contribution of evolutionary forces at shaping codon usage bias in eukaryotes is a matter of debate, especially in mammals. The main riddle remains understanding the sharp contrast between the strong molecular impact of gene expression differences arising from codon usage preferences and the thin evidence for codon usage selection at the organismal level. Here we report a multiscale analysis of the consequences of alternative codon usage on heterologous gene expression in human cells.

We generated synonymous versions of the *shble* antibiotic resistance gene, fused to a fluorescent reporter, and expressed independently them in human HEK293 cells. We analysed: 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 construct expression; and iii) real-time cell proliferation in absence or presence of antibiotic, as a proxy for the cellular fitness.

Our results show that differences in codon usage preferences in our focal gene strongly impacted the molecular and the cellular phenotype: i) they elicited large differences in mRNA and in protein levels, as well in mRNA-to-protein ratio; ii) they introduced splicing events not predicted by current algorithms; iii) they lead to reproducible phenotypic heterogeneity as different multimodal distributions of cellular fluorescence EGFP; iv) they resulted in a trade-off between burden of heterologous expression and antibiotic resistance. While certain codon usage-related variables monotonically correlated with protein expression, other variables (*e.g.* CpG content or mRNA folding energy) displayed a bell-like behaviour. We interpret that codon usage preferences strongly shape the molecular and cellular phenotype in human cells through a direct impact on gene expression.

INTRODUCTION

The cellular phenotype is the integrated result of deterministic, statistical and random molecular processes. The canonical scenario for gene expression, the “sequence hypothesis”, 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 (Crick, 1970). 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 plethora of studies on gene expression have addressed this question through the last decades, and revealed multi-level regulation mechanisms increasing the diversity of the proteomic landscape available for expression from a given genome. Genetic information flow relies on the genetic code, which establishes a chemical correspondence between the DNA coding informative 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). The genetic code is degenerate as 18 amino acids can individually be encoded by a group of two, three, four or six codons, known as synonymous codons. In a first null hypothesis approach, one would expect synonymous codons to display similar frequencies. Instead, codon usage biases (*i.e.* the uneven representation of synonymous codons (Grantham et al., 1980) 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 (Duret, 2002; Gouy & Gautier, 1982; Ikemura, 1982; Kanaya et al., 1999; Novoa et al., 2019; Sharp & Li, 1986).

The evolutionary origin and the identification of the role and contribution of different forces contributing to codon usage preferences constitute a classical research subject in evolutionary genetics. The scientific debate is centered in identifying the differential explanatory power of neutral mechanisms and of natural selection at having shaped global and local codon usage preferences. An extensive body of knowledge has established that variation in codon usage preferences can indeed be under selection, as it might constitute an additional layer of information allowing for differences in gene expression (J. V. Chamary et al., 2006; Hanson & Coller, 2017; Plotkin & Kudla, 2010). And indeed, in parallel to the scientific controversy, genetic engineering and the revolution of affordable gene synthesis have extensively resorted to codon usage recoding for enhancing heterologous protein production, for its use in industrial applications or for vaccine design (Angov et al., 2008; Fath et al., 2011; Mauro & Chappell, 2014) The hypothesis of translational selection proposes that differences in codon usage preferences result in gene expression differences that ultimately lead to phenotypic differences, which could be subject to natural selection. Besides the plethora of succesful gene recoding strategies, the differential interaction between codon usage preferences and the translation machinery has been well established, for instance in: i) the evolutionary co-variation of genomic codon usage and the tRNA content, from unicellular organisms (Dong et al., 1996; Ikemura, 1981; Kanaya et al., 1999) to metazoa (*Caenorhabditis elegans* (Duret, 2000), *Drosophila* (Akashi, 1994; Moriyama & Powell, 1997; Powell & Moriyama, 1997), or humans (Urrutia & Hurst, 2001); ii) the correspondence between codon usage preferences and expression level in bacteria (Lithwick & Margalit, 2003) or in yeast (Ghaemmaghami et al., 2003; Tuller et al., 2007); iii) the increase in translation efficiency iun bacteria when supplementing *in trans* with rare tRNAs (Burgess-Brown et al., 2008); iv) the changes in tumorigenic phenotype in mice when switching from rare to common codons in the sequence of a cancer-related GTPase (Lampson et al., 2013).

One may argue that only successful gene recoding strategies are communicated, thus introducing an important literature and knowledge bias that overstates the importance of codon usage preferences in gene expression. Actually, a number of studies have communicated the lack of covariation between codon usage and gene expression (in bacteria, yeast, or human) (Kudla et al., 2009; Li et al., 2014; Pop et al., 2014; Vogel et al., 2010); or even a negative impact of a presupposed "optimization", which may in fact decrease the expression or the activity of the protein product (Agashe et al., 2013; Zucchelli et al., 2017). To address these conflicting results, it is important to tease apart the underlying mechanisms, as it is allowed by combining genomics, system biology, and predictive models. It has hitherto been established that codon usage preferences can impact the molecular, cellular and organismal phenotype by modifying: 1. mRNA localisation, stability and decay (S. Chen et al., 2017; Harigaya & Parker, 2016; Presnyak et al., 2015; Radhakrishnan et al., 2016; Radhakrishnan & Green, 2016), 2. translation initiation (Bettany et al., 1989; De Smit & Van Duin, 1990; Gu et al., 2010; Kudla et al., 2009), 3. translation efficiency (Akashi, 1994; Gardin et al., 2014; Hussmann et al., 2015; Ingolia et al., 2009; Johnston et al., 1984; Johnston & Parker, 1985; Kurland, 2003; Marais & Duret, 2001; M. Robinson et al., 1984; Sørensen et al., 1989; Sørensen & Pedersen, 1991; Stoletzki & Eyre-Walker, 2007; Tuller, Waldman, et al., 2010; Weinberg et al., 2016; Xia, 2014); 4. co-translational protein folding (Chaney et al., 2017; Pechmann & Frydman, 2012; Zhao et al., 2017). The respective contribution of each mechanism, if any, depends on the specific expression system (*e.g.* in which organism, whether the expressed gene is autologous or heterologous gene, whether it has been recoded or not). To date, no universal rules have been identified, and the explanatory power of our interpretations remains limited, even when using large-scale heuristic approaches in tractable experimental systems (Cambray et al., 2018).

In this study, we aim at providing an integrated view of the molecular and cellular impact of alternative codon usage (and the associated nucleotide composition) of a heterologous gene in human cells. We designed six synonymous version of the *shble* antibiotic resistance gene with distinct codon usage preferences, coupled them to a *egfp* reporter gene, that allows for further single-cell assessment of the gene expression 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 codon usage bias and sequence composition on the molecular and cellular phenotype of human cells in culture.

RESULTS

1. Codon usage preferences of the *shble* heterologous gene resulted in differences in mRNA abundance, and alternative splicing profile.

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 codon usage (estimated using the COdon Usage Similarity Index, COUSIN) (Bourret et al., 2019), GC composition at the third nucleotide of codons (GC3), and CG dinucleotide percentage (CpG). 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), prtly reflecting sequence similarity (Sup. Table 1).

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

Condition	Description	COUSIN index	%GC3	%CpG	Transcript folding energy (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 codon usage, 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 splice alternatives, using the same 5' donor position and differing in three nucleotides at the 3' acceptor position. The shble#4 transcript presented one splice alternative, with donor and acceptor positions in the precise same location than 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 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) (Desmet et al., 2009) nor the SPLM (Solovyev, 2004) splice detection algorithms used for sequence scanning during design.

The mRNA abundances, expressed as transcript per millions (TPM), showed that the spliced isoform 1 represented about 30% of the heterologous transcripts in the condition 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 depending on the construct version, 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 after 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 considered that the SHBLE protein could be translated exclusively from 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. For shble#4 this ratio was about 69 %, 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).

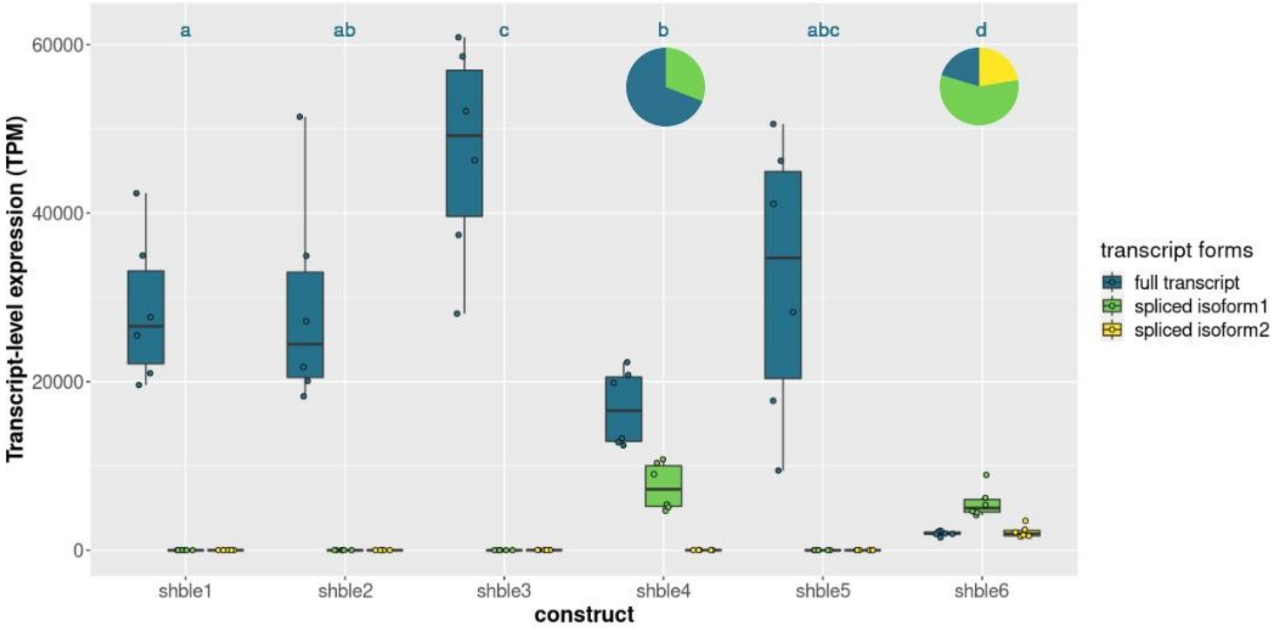


Figure 1. Transcript abundance after transfection with the different *shble* gene versions. mRNA-levels are expressed as transcripts per million values (TPM) for the full form (in dark blue) as well as for the two spliced forms (in green and yellow). Median values are given in Sup. Table 2. Pie charts illustrate the proportions of the spliced transcript 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. Codon usage preferences of the *shble* heterologous gene modulate 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 for shble#6) (Figure 2B, Sup. Table 3). Further, the shble#3 condition displayed lower SHBLE protein levels than the remaining four other constructs (normalized iBAQ value 0.93 for shble#3, 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 on nine biological replicates (Sup. Fig. 6, 7 and 8).

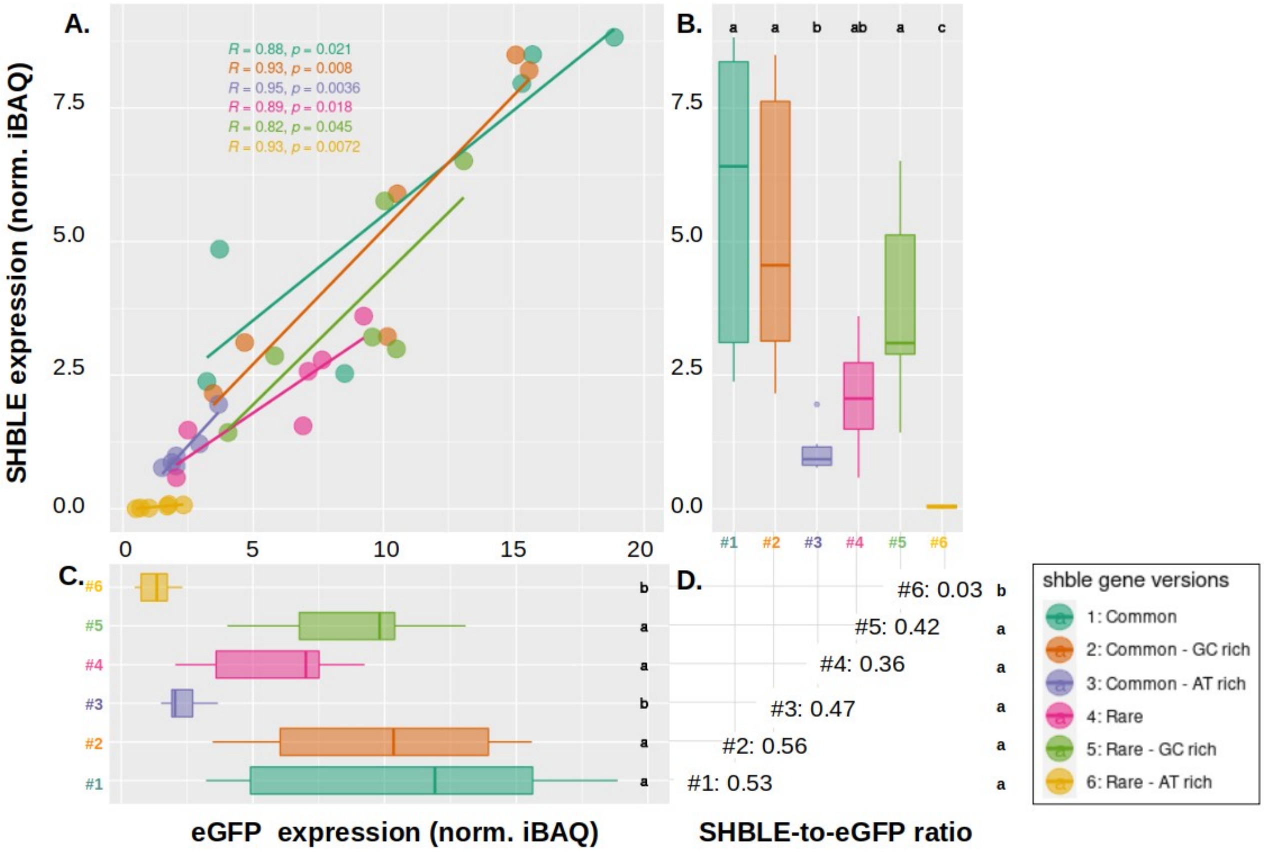


Figure 2. Expression of SHBLE and EGFP at the proteomic level, and relation between them. Panel A: Pearson's correlation between SHBLE (y axis) and EGFP (x axis) protein levels. Six different conditions are shown: shble#1 (dark green), shble#2 (orange), shble#3 (purple), shble#4 (pink), shble#5 (light 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. Codon usage preferences of the *shble* heterologous gene modulate the match between transcriptomic and proteomic phenotypes.

After analysing separately mRNA and protein levels in cells transfected with the different *shble* versions, we aimed at establishing a connection between the transcriptomic and the proteomic phenotypes. Figure 3A presents the positive and significant correlation between variation in the full-to-total ratio of heterologous transcripts, and variation in the SHBLE-to-EGFP ratio of protein levels. This correlation is dominated by (i) the very low SHBLE protein levels detected in shble#6, as described above, and also by (ii) the match in mRNA and protein variation in shble#4. Indeed, for shble#4, the SHBLE-to-EGFP protein ratio (0.36, Figure 2D) corresponded to 73% of the median of the other four remaining constructs (0.50, calculated from Sup. Table 3), which is close to the 69% fraction of the full-to-total transcripts (the full-length mRNA being the only one with SHBLE coding potential) (Sup. Table 2). When considering the SHBLE protein alone, variation in full-length transcripts levels explained 45% of the variation in the SHBLE protein levels (Sup. Fig. 9). 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 3B, Sup. Fig. 9). When removing this shble#3 condition from the correlation analysis, variation in full-length transcripts levels explained 83% of the variation in the SHBLE protein levels (Sup. Fig. 9).

We explored subsequently the explanatory potential of sequence composition and mRNA physicochemical parameters in order to understand the differential matches between transcriptomic and proteomic phenotypes. Regarding nucleotide composition, an increase of the match between the *shble* condition and the human average codon usage (as evaluated using the COUSIN index), and of the GC3 content, corresponded monotonically to an increase in the SHBLE protein-to-transcript ratio (respectively Pearson's R=0.67, p=6.6e-6, Figure 3C; and Pearson's R=0.81, p=1.5e-9, Figure 3D). Such shared behaviour is not unexpected, as COUSIN and GC3 values are highly correlated (Pearson's R=0.88, p=0.02, data not shown). However, variation in CpG dinucleotides abundance in the recoded *shble* coding sequence corresponded to a bell-shaped variation in SHBLE protein-to-transcript ratio, so that high CpG values (as in shble#5) resulted in decreased protein-to-transcript ratio (Figure 3E). Interestingly, the maximum value was observed for the shble#1 version, which uses exclusively the most common codon for each amino acid. A similar bell-shape was observed when displaying variation in SHBLE protein-to-transcript ratio as a function of the full mRNA folding energies: very strong folding energies (as in shble#5) or very weak folding energies (as in shble#3 or shble#6) resulted in a decreased protein-to-transcript ratio (Figure 3F). Once again, the shble#1 version displayed the highest values of the dependent variable. Interestingly, the shble#3 condition (using the AT-richest among the most used codons) combined suboptimal values for all the studied characteristics (low COUSIN, GC3 and CpG content values, and low folding energy). We interpret that codon usage bias and the associated mRNA chemistry could explain the loss of concordance between high mRNA levels and low protein abundance (Figure 3B). A similar reasoning would explain for shble#4 (using the rarest codons) the good protein-to-mRNA ratio, as despite the low COUSIN value, CpG content is close to the optimum (Figure 3E) and mRNA folding energy displays intermediate values (Figure 3F).

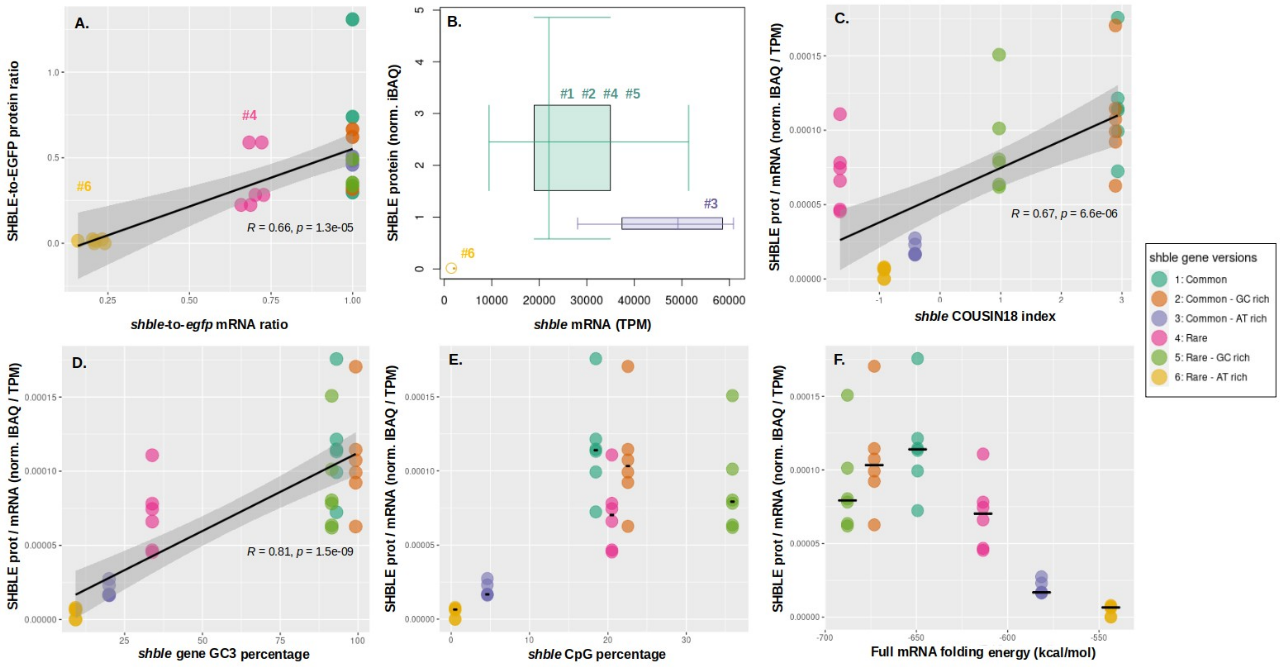


Figure 3. Relation between the transcriptomic and the proteomic phenotypes, and potential explicative parameters for variations in SHBLE protein levels. Six different conditions are shown, using the colour code: shble#1 (dark green), shble#2 (orange), shble#3 (purple), shble#4 (pink), shble#5 (light green) and shble#6 (yellow). **Panel A:** Pearson's correlation of the SHBLE-to-EGFP protein ratio and the full-to-total transcript level (a proxy for *shble*-to-*EGFP* mRNA ratio). **Panel B:** Combined distribution of SHBLE protein level (y axis - normalized iBAQ) and *shble* transcript level (x axis - TPM); individual construct boxes are condensed in a single one when the squares defined by the first and third quartiles overlaps (which is the case for shble#1, shble#2, shble#4 and shble#5, shown condensed in dark green). Correlations per condition are shown in Sup. Fig. 9. **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 percentage of the *shble* recoded version. Black bars represent the median for each condition. **Panel F:** Correspondence between the SHBLE protein-to-mRNA ratio and the folding energy of the corresponding transcript. Black bars represent the median for each condition. 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. Codon usage of the *shble* heterologous gene modifies cellular fluorescence intensity.

We have demonstrated above that the EGFP reporter was a relevant proxy for SHBLE abundance, as their iBAQ values were highly correlated, and have also shown that differences in SHBLE-to-EGFP ratios were largely attributable to splice events at the mRNA level. On these basis we performed analyses based on the cell-based fluorescence values to further characterise our experimental model.

We aimed at assessing the phenotypic variation at single-cell level by performing an extensive fluorescence analysis on 16 transfection replicates, overall corresponding to 480,000 cells per condition. We verified first that the total fluorescence signal was strongly correlated to the EGFP level, as estimated by label-free proteomic, strengthening the results reported in the previous sections (Pearson's R=0.86, p=4.8e-

15, Sup. Fig. 10). We observed then that for all conditions expressing EGFP the distribution of fluorescence intensity was multimodal (Figure 4A, Sup. Fig. 11), and that the underlying distribution for each condition is different from that obtained with cells expressing EGFP alone (*i.e.* the "#empty" control; individual Anderson-Darling test results are shown in Table 2). We tried to describe 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. 12). We chose further to summarise the fluorescence behaviour of the full cellular population by describing for each condition the following summary statistics: (i) the fraction of cells displaying fluorescence over 99th percentile of the "mock" fluorescence distribution (*i.e.* 14,453 fluorescence units, which corresponds to cellular autofluorescence, as the mock does not carry any plasmid); (ii) the total fluorescence value for all cells in the population; (iii) the median fluorescence value for the cellular population; (iv) the mean fluorescence value for each underlying Gaussian populations; and (v) the fraction of the cellular population displaying fluorescence values stratified into five log10-width intervals, between the fluorescence positivity threshold and the highest fluorescence value detected by our cytometer (respectively labeled as "neg" and "i4", Figure 4A) (Table2).

The analysis of the results showed that the central fluorescence value of the population correlates very well with the overall fluorescence (R=0.85, p-value<2.2e-16, Sup. Fig. 13). Further, condition shble#6 (and, to a lesser extent, condition shble#3 as well) displayed significantly lower global fluorescence values (Table 2, Sup. Fig. 13). The global lower fluorescence for these two conditions is related to a population-level shift, as both underlying Gaussian curves displayed also lower mean fluorescence values (Table 2, Sup. Fig. 14). Consistently, condition shble#6 displayed the highest fraction of non-fluorescent cells, and shble#3 and shble#6 presented the lowest fractions of strongly positive cells (Figure 4B). When combining all our summary statistic variables for describing the population cellular fluorescence we observed that indeed shble#6, and to a lesser extent shble#3, were the most divergent conditions, characterised by the highest proportion of negative or low-fluorescent cells, while shble#1 and shble#2 displayed very similar behaviour characterised by high fluorescence values in all scores (Sup. Fig. 15).

Table 2. Quantitative parameters of green fluorescence signal distribution per condition. "AD", results of an Anderson-Darling test for distribution similarity, comparing each curve distribution in Figure 4 against that obtained for the "empty" condition (the null hypothesis being that the samples compared could have been drawn from a common population). For the three last parameters, 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).

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	**a**	6.61	**c**
#shble2	1480	0	90.17 %	98.311 e9	**b**	4.78	**a**	6.59	**c**
#shble3	497	4.637 e-272	79.37 %	39.384 e9	**d**	4.31	**b**	5.86	**ab**
#shble4	463	7.325 e-254	88.00 %	63.395 e9	**ac**	4.58	**a**	6.28	**bc**
#shble5	108	4.244 e-59	83.85 %	70.719 e9	**abc**	4.63	**a**	6.32	**abc**

#shble6	11600	0	51.78 %	13.990 e9	**e**	3.97	**c**	5.05	**d**
#empty	0	1	82.26 %	57.692 e9	**a**	4.44	**ab**	6.18	**ab**
#superempty	64100	0	0.45 %	135.449 e6	na	na	na	na	na
mock	62500	0	1.00 %	141.163 e6	na	na	na	na	na

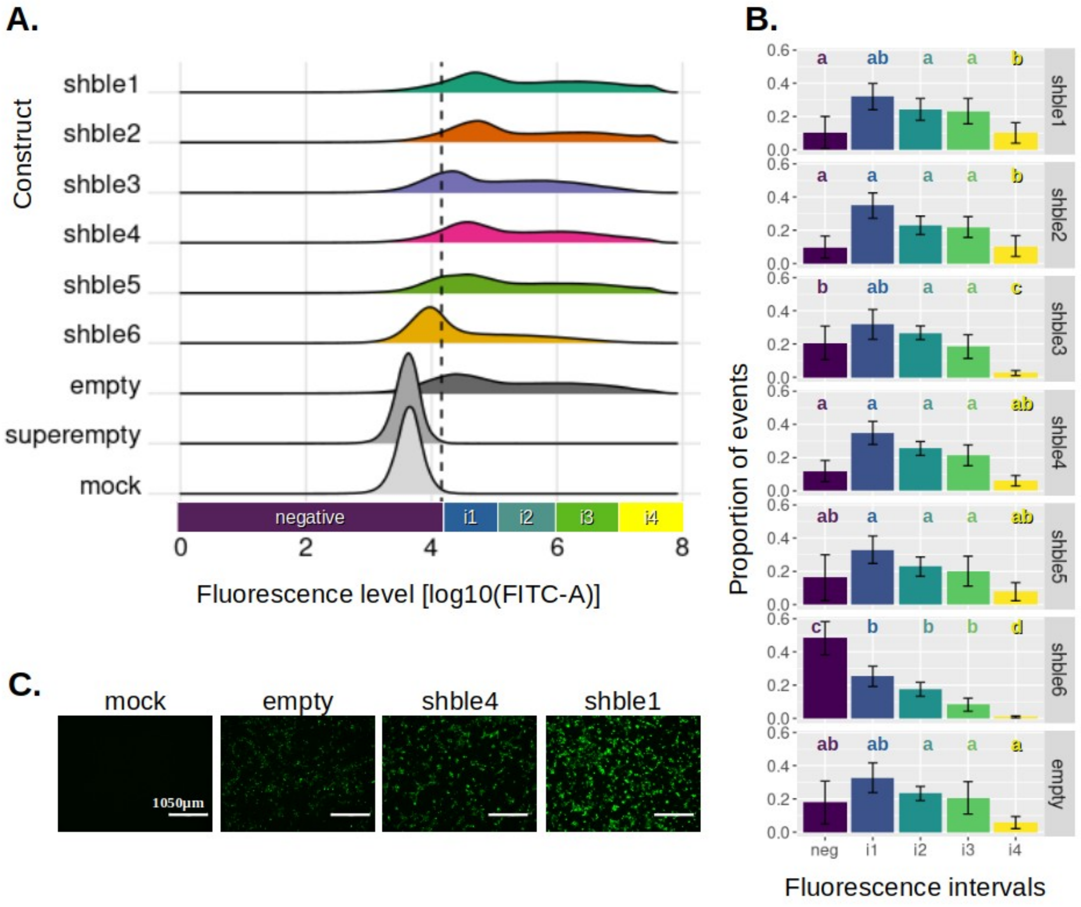


Figure 4. Distribution of the fluorescence signal for the different constructs. Panel A depicts the density of the green fluorescence signal (log10(FITC-A)) considering the 480,000 studied events (*i.e.* individual cells) for each condition: shble#1 (most common codons, dark green), shble#2 (common and 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). The coloured squares along the x axis correspond to the five intervals of fluorescence used for further analyses (see panel B). Panel B represents the proportion of events after stratification into different arbitrary fluorescence intervals: negative cells in dark purple, “i1” in dark blue, “i2” in dark green, “i3” in light green and “i4” in yellow. Letters in panel B refer to the results of a pairwise Wilcoxon rank sum test performed vertically, for each fluorescence intensity category; conditions associated with a same letter do not display different median values of the corresponding variable (p<0.05 after Benjamini-Hochberg correction). Panel C illustrates cell fluorescence as observed in microscopy with identical capture settings immediately before sampling, for four representative conditions: the negative control « mock », the positive control « empty », the rarest version #4, and the most common version #1 (indicated scale of 1050µm). Two other replicates are shown as Sup. Fig. 16.

5. Codon usage preferences of the *shble* heterologous gene resulted in different cell growth dynamics.

To assess the functional impact of the different molecular phenotypes described above, we performed a real-time cell growth analysis as a proxy for cellular fitness, both in presence and in absence of antibiotics. We anticipated a trade-off between a potential benefit through antibiotic resistance (*i.e.* resistance to bleomycin conferred by the SHBLE protein), and a potential cost through protein overexpression and the associated burden. We monitored over time a functional parameter named “Cell Index”, that integrates cell density, adhesion intensity, morphology and viability, and evaluated the total area below the curve as a proxy for cellular growth (Sup. Method 2.8, Sup. Fig. 17). We fitted to a Hill’s equation the values of cellular growth as a function of the antibiotic concentration to recover, for each condition, the estimation for the maximum growth in the absence of antibiotic as well as the estimation for the antibiotic concentration value that inhibited cellular growth to half the maximum (IC50).

We observed that simple transfection with an empty vector not expressing EGFP nor SHBLE (*i.e.* the “superempty” control) resulted in a drop of about 50% in maximum cellular growth in the absence of antibiotic (Figure 5, y axis) and in a drop of about 85% in IC50 value (Figure 5, x axis), with respect to the mock. All cellular populations transfected with any of the *shble* constructs displayed further lower maximum growth values in the absence of antibiotics than the “superempty” control (Figure 5, y axis). Further, variance in SHBLE protein levels resulted in different degrees of bleomycin resistance although, surprisingly, all transfected cells resisted less the presence of antibiotics than the mock, untransfected control, independently of the construct used (as evaluated using IC50, Figure 5, x axis). Even if no significant correlation could be established because of the limited number of experimental conditions, a very interesting 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. 17). 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.

To disentangle the effects linked to the total expression of heterologous proteins (SHBLE + EGFP), and the effect of the antibiotic resistance gene alone, we further synthesised and 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. 17). Very interestingly, both versions displayed a similar increase in growth in absence of antibiotics with respect to their shble#1 and shble#4 relative counterparts (respectively 38% and 24%, shown as coloured arrows on Figure 5, y axis). However, while the IC50 of shble#4* and shble#4 remained similar, antibiotic resistance for version shble#1* dramatically increased with respect to that of shble#1 (around 300 times increase, shown as green arrow on Figure 5, x axis). Indeed, shble#1* condition is the only one in which resistance to the antibiotic is actually better than for the untransfected cells, in spite of a remaining substantial negative impact on maximum growth on the absence of antibiotics.

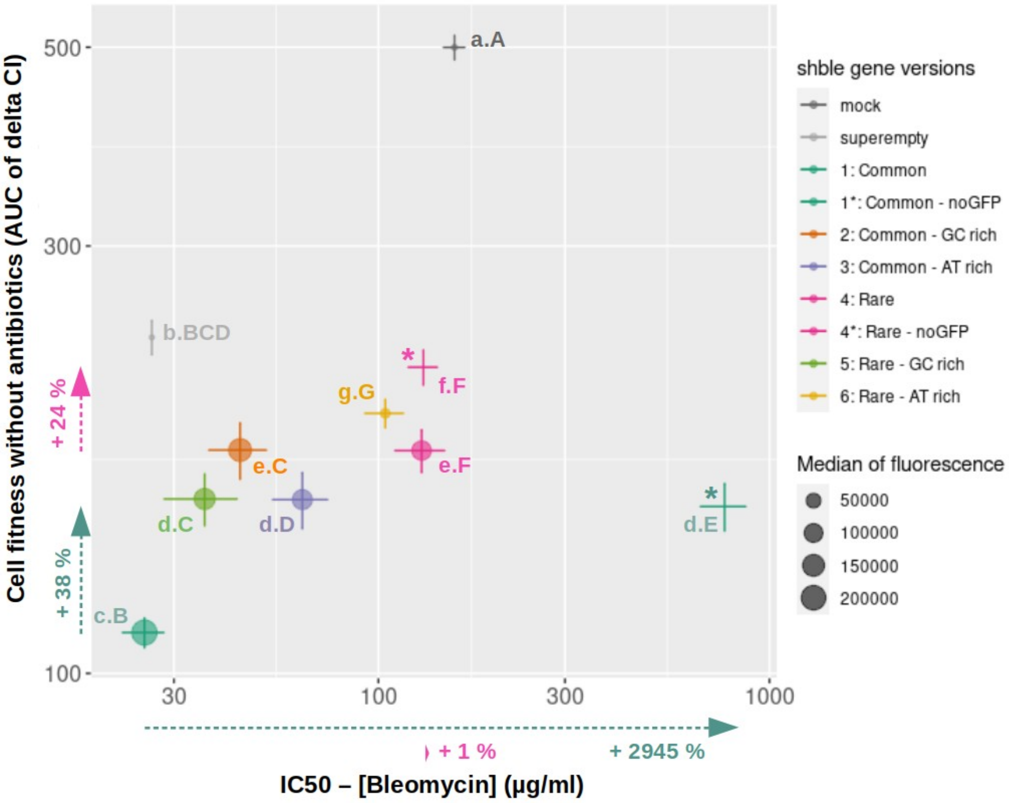


Figure 5. Variation of cell growth in presence or in absence of antibiotics. The y axis represents maximum cellular growth in absence of antibiotics, proxied as the area under the curve of the delta Cell Index (AUC, log scale). The x axis represents the bleomycin concentration reducing to 50% the corresponding 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. 18 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

The origin and meaning (if any) of differences in codon usage preferences between species and between genes within a genome are classical questions of evolutionary genetics. Two main non-exclusive hypotheses are usually presented aiming at explaining the evolution of codon usage bias, respectively based on neutralist and selectionist arguments (Bulmer, 1991; Duret, 2002; Duret & Galtier, 2009; Nicolas Galtier et al., 2018; Hershberg & Petrov, 2008). The neutral hypothesis posits that differences in average genome codon usage bias are linked to non-selective processes, such as biochemical biases during DNA synthesis or repair (*e.g.* polymerase bias) (S. L. Chen et al., 2004). Additionally, local composition bias such as the alternation between GC-rich and AT-rich stretches in vertebrates chromosomes, known as isochores (Caspersson et al., 1968), strongly shape the codon usage preferences of the genes therein residing, further enhanced by GC-biased gene conversion mechanisms (N. Galtier et al., 2001). 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. 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. Indeed, biotechnology engineering approaches show that codon recoding is a powerful tool for ameliorating heterologous gene expression (Mauro & Chappell, 2014), viral codon recoding is currently applied on a regular basis for most vaccine development (Garmory et al., 2003), and a number of medical conditions in humans have been mapped to synonymous polymorphisms (Sauna & Kimchi-Sarfaty, 2011). 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 (Nicolas Galtier et al., 2018) such as bacteria (*E. coli* (Sharp & Li, 1986)), yeast (*S. cerevisae* (Sharp et al., 1986)), nematodes (*C. elegans* (Stenico et al., 1994)), but also in fruit flies (Akashi, 1994; Bierne & Eyre-Walker, 2006; Moriyama & Powell, 1997; Shields et al., 1988), branchiopods (*Daphnia pulex* (Lynch et al., 2017)) and amphibians (*Xenopus laevi* (Musto et al., 2001)). In mammals, and particularly in humans, evidences of selection for (or against) certain codons remain nevertheless controversial (Urrutia & Hurst, 2001).

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. 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 have analysed here 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 (Jallet et al., 2021). 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, even if minor, as between shble#1 and shble#2, results in important multilevel changes in gene expression levels and leads to dramatic differences in the cellular phenotype.

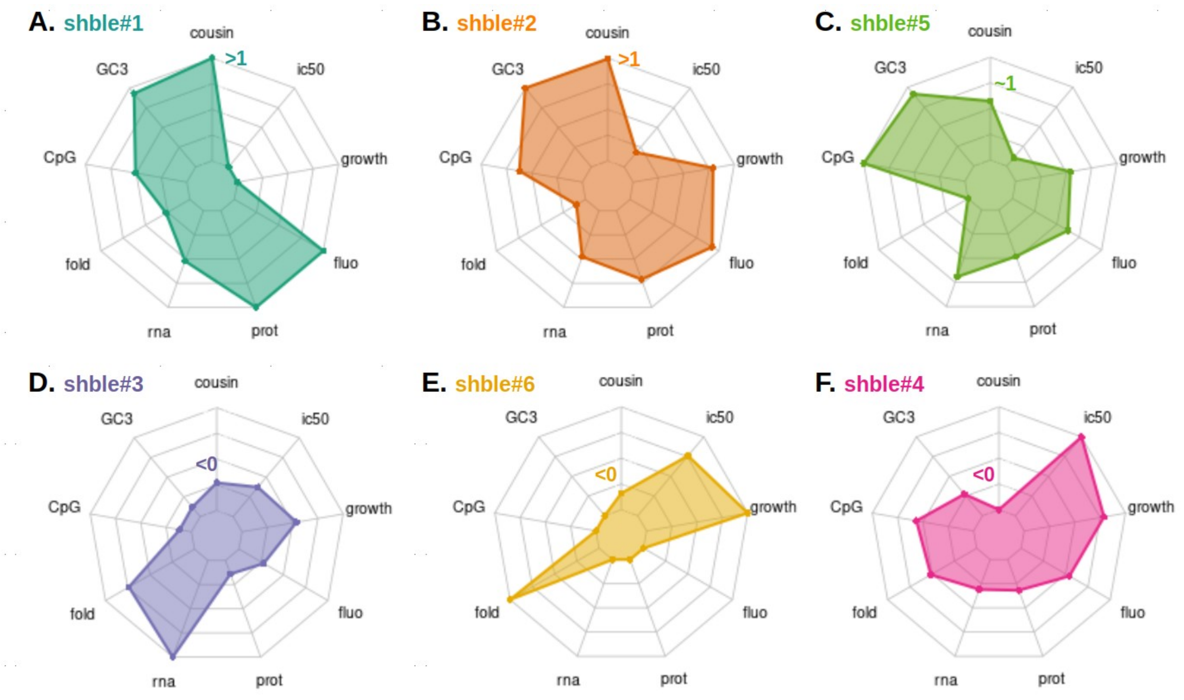


Figure 6. Summarizing combination of sequence composition parameters and multi-level phenotypes for each customized version of the *shble* antibiotic resistance gene. The six versions, designed with the one amino acid – one codon strategy, are showed by decreasing similarity to the human codon usage (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 human genome codon bias), "CpG" (the CG dinucleotide proportion), "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).

Codon usage preferences modify mRNA levels and modify alternative splice patterns. The behaviour of the two *shble* versions recoded with the most dissimilar codon usage preferences with respect to the human average (shble#4 and shble#6) was characterized by splicing events, which reduced the coding potential of the resulting mRNA by 30 to 80 % and ablated synthesis of our focal protein, SHBLE, in the spliced transcripts. Further, splicing efficiency was largely dependent on the precise codon recoding around these novel splice sites, with *ca.* 20% of the total shble#4 transcripts being spliced compared to the *ca.* 80% for shble#6. It should be emphasised that none of these spliced events was detected by leading splice site predicting algorithms (Desmet et al., 2009; Solovyev, 2004). Such potential impact at modifying the exon-intron nucleotide context, which serves as the basis for directing transcript splicing, is classically recognised as one of the potential effects of synonymous mutations (Callens et al., 2021). Codon usage variation across intron-exon boundaries has indeed been described in several eukaryotes (*e.g.* human, fishes, fruit flies, nematodes, plants (Eskesen et al., 2004; Plotkin & Kudla, 2010; Willie & Majewski, 2004)), albeit the pattern of codon distribution varied between species. Additional splicing regulatory motifs that can be disrupted by synonymous mutations have also been described close to the intron–exon boundary in mammals (J. V. Chamary et al., 2006; Eskesen et al., 2004; Fairbrother et al., 2002; Louie et al., 2003; Parmley & Hurst, 2007). A reduced SNP density and decreased rate of synonymous substitutions have been reported in these regulatory regions, which is a signature for selective pressure (J. V. Chamary & Hurst, 2005; Orban & Olah, 2001). In humans, splicing defects can have a dramatic impact on the phenotype and cause disease (Faustino & Cooper, 2003). Thus, selection against mRNA mis-processing can constitute an important selective force that results in concomitant selection for a precise local codon usage (Callens et al., 2021), and this selective force has even been propose to outperform translational selection in *D. melanogaster* (Warnecke & Hurst, 2007). Overall, our results highlight thus the direct impact of local codon usage preferences at introducing diversity during transcription.

We describe here how variation in codon usage preferences leads to differences in mRNA levels that do not necessarily translate into differences in protein levels. In our experiments, variation in mRNA abundance between conditions was independent variation in DNA abundance, ruling out a possible effect of differential transfection efficacy. We interpret instead that specific compositional properties of the different mRNA transcripts, arising from differences during codon recoding, may lead to differential mRNA stability, as has been described for bacteria (*E. coli* (Boël et al., 2016)), unicellular eukaryotes (*S. cerevisae, S. pombe* (Harigaya & Parker, 2016), *N. crassa, T. brucei (Jeacock et al., 2018; Nascimento et al., 2018)),* and metazoa (fruit fly (Burow et al., 2018) or zebrafish (Mishima & Tomari, 2016)). Beyond differences in mRNA levels between conditions, in our experimental setup variations in the mRNA levels explain only around 40% of the variation in protein levels, which fits well previous descriptions in the literature for a wide diversity of experimental systems (De Sousa Abreu et al., 2009; Vogel & Marcotte, 2012). This weak explanatory power would not be expected if mRNAs were translated at a constant rate, and has motivated studies to elucidate which explanatory factors are involved in the regulation of translation ((De Sousa Abreu et al., 2009) for review). We also evidenced that this discrepancy between mRNA and protein level was unequal between conditions: particularly, the version using the AT-rich codons among the two most common (shble#3) displayed the highest mRNA levels but contrasting low amount of protein. We interpret that this phenotype arises from the combination of suboptimal variables directly or inderectly linked to codon usage preferences (similarity to human average codon usage, GC3 and CpG content, and mRNA folding energy), which we have shown to be good predictors of the match between mRNA and protein levels. As reviewed by Plotkin and Kudla (Plotkin & Kudla, 2010), a role in optimizing the expression of heterologous genes had already been evidenced for those four parameters. Regarding similarity in codon preferences between the focal gene and the expression system, gene versions with a better match to the average human codon usage bias 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 codon usage preferences on gene expression in mammals, compared to other features (Lu et al., 2006; Vogel et al., 2010). Nevertheless, it is complicated to disentangle the effect of codon usage preferences 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 (Hanson & Coller, 2017). While variation in GC3 was monotonically related to variation in protein expression, this was not the case for variation in CpG dinucleotides. We report instead a bell-shaped correlation with the protein-to-mRNA ratio, defined by extreme values of CpG (either too high or too low) resulting in lower translation levels, and with the maximum response corresponding to the gene recoded version close to the human preferences. In eukaryotic genes, selection for presence or absence of CpG dinucleotides usually focuses on the upstream regulatory region and is usually explained in terms of epigenetic control of gene expression (Callens et al., 2021). The consequences of CpG dinucleotide content is usually more assessed at the transcriptomic than at the translational level. Nonetheless, it has been shown to impact heterologous protein amount, but through its impact on *de novo* transcription rather than on translation efficiency (Bauer et al., 2010). For the mRNA folding energy, we report a bell-shaped correlation with the protein-to-mRNA ratio, defined by extreme values of folding energy predicting a sub-optimal translation. The impact of the secondary structures along the transcript has rarely been adressed, but recent studies highlighted its role in the functional half life of mRNA (Mauger et al., 2019). Besides, several studies focusing on the 5' region established the importance of the mRNA secondary structure in translation initiation, and highlighted a shared trend (bacteria, yeast, protists, and mammals (Kudla et al., 2009; Mauger et al., 2019; Shah et al., 2013; Wang et al., 2020; Weinberg et al., 2016)): a reduced mRNA stability near the site of translation initiation, correlated to a higher protein production. Indeed, in bacteria and yeast, strong folding around the start codon prevents ribosome recruitment (Kudla et al., 2009; Shah et al., 2013); and an analysis of more than 400 bacteria genomes highlighted that codons reducing the mRNA folding are overrepresented at the beginning of the genes, independently of their representation in the rest of the genome (Bentele et al., 2013). Molecular modeling, along with experimental studies, suggest an higher impact of translation initiation than elongation (Gu et al., 2010; Riba et al., 2019; Shah et al., 2013). Nonetheless, de Sousa Abreu et al, 2009 that described no effect of the initiation rate on translation efficiency in human transcripts. In addition, it has been described a "ramp" of rare codon along the 50 to 100 first nucleotides of the genes, of mRNA which is thought to reduce the mRNA folding, and to avoid ribosome stack (Bentele et al., 2013; Tuller, Carmi, et al., 2010). We propose here that a complex secondary structure of the mRNA (and not of the upstream sequence only) can have an impact on the translation; but, contrarily to what has been recently described (Mauger et al., 2019; Tuller, Waldman, et al., 2010), this role won't be monotonous, but rather display an 'optimal' state. Anyhow, a complex combination of parameters seem to be at play, and recent study interesingly has shown that an optimized sequence (i.e. leading to the higher level of protein protduction) would actually be mosaic, including rare codon at specific positions, rather than a sequence fully composed of frequent codons (Perach et al., 2021). This is important for further strategy of codon optimization and actually meet previous suggestion of randomization or harmonization strategies (Angov et al., 2008; Menzella, 2011).

The cell by cell fluorescence analysis revealed that, each transfected cell population was in fact formed by at least two subpopulations expressing the EGFP reporter at different level, but that variations in codon usage affected them equally. This can be explained in light of previous published observations suggesting a cell cycle-dependent regulation of transcription under the CMV promoter/enhancer (Brightwell et al., 1997). When comparing conditions, the concerted shift of both subpopulations towards higher (e.g. for the most used and/or GC-rich codons) or lower (e.g. for AT-rich codons) values of fluorescence intensity suggests that the codon usage version of *shble* impacts gene expression whatever the cell stage. Nonetheless, our model is not refined enough to adress the question of cell-cycle dependent codon usage oscillation that was reported before (Frenkel-Morgenstern et al., 2012).

We highlighted a physiological impact linked to the expression of heterologous proteins which seems "stronger" than the conferred antibiotic resistance. And, when decreasing the quantity of heterologous proteins (by removing EGFP), the antibiotic resistance is effective (also meaning that the produced protein are functional). This observation can be discussed in the light of previous reports evidencing the competition for the transcriptional machinery. First, Kudla *et al.* (Kudla et al., 2009) reported that rare codons in an over-expressed heterologous gene decreased cellular fitness, because of the ribosome sequestration along the non-adapted mRNA (also consistent with Andersson 1990 (Andersson & Kurland, 1990)). Indeed, considering that almost all the ribosomes are engaged in translational process at any moment, they are a limitating component (Princiotta et al., 2003). They alternatively propose that common codons would reduce the errors in protein conformations that are deleterious for the cell (Drummond et al., 2005; Stoletzki & Eyre-Walker, 2007). A further study precised that the limitating factor was in fact the tRNA matching rare codons, because by supplying the system with those codons, they recovered the fitness (Frumkin et al., 2018). The proposed selective mechanism beyond these observation is that selection act in fact at the genome wide scale and that highly expressed genes and lowly expressed gene use different codon in order to allow the homeostasy of the cell. This would be particularly true in conditions of stress, or changes in nutritional status (Hanson & Coller, 2017). For instance in bacteria, genes that are essential during amino acid starvation (e.g. amino acid biosynthetic enzymes) preferentially use rare codons that do not match the typical pool of tRNAs, but instead match starvation-induced tRNA pools (Dittmar et al., 2005; Elf et al., 2003).

To conclude, the present study highlighted that most of the potential evolutionary forces at play in shaping human codon usage (and related nucleotide content), select for a strict control of mRNA processing: splicing, secondary structure and decay. In contrast, we could not evidence proper translational selection, but more investigation of the secondary structure along the coding sequence remain to explore. Whether these predictors are relevant in the light of the evolutionary theories can not be stated, but further studies involving experimental evolution may shed light on these mechanisms.

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 codon usage preference of a query against a reference, was calculated on the online tool (http://cousin.ird.fr) (Bourret et al., 2019). A score value below 0 informs that the codon usage preferences (CUPrefs) of the query sequence is opposite to the reference CUPrefs; a value close to 1 informs that the query CUPrefs is similar to the reference CUPrefs, and a value above 1 informs that the query CUPrefs is similar the reference CUPrefs, but of larger magnitude (Bourret et al., 2019). 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 (Ryan et al., 1991), 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 (J. H. Kim et al., 2011) 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).

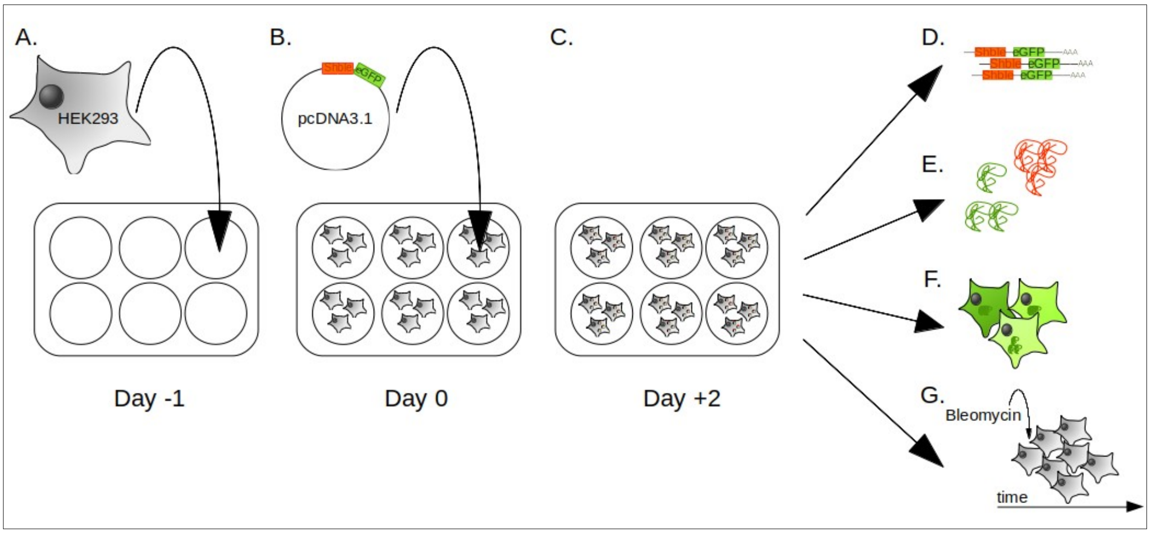


Figure 7. Overview of the sampling protocol and the measured phenotypes. HEK293 cells were seeded on 6-well plates (A) one day before transfection with the customized pcDNA3.1 plasmids (B). Transfected cells were harvested two days later (C). mRNA levels were assessed by RNAseq (D), protein levels were 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) (Bolger et al., 2014) and mapped on eight different genomic references (HISAT2 v2.1.0) (D. Kim et al., 2015), 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 (J. T. Robinson et al., 2011). 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) (Desmet et al., 2009), or with *SPLM - Search for human potential splice sites using weight matrices* (accessed via http://www.softberry.com/) (Solovyev, 2004). When relevant, the three alternative transcript isoforms identified were further used as reference for read pseudomapping and quantification with Kallisto (v0.43.1) (Bray et al., 2016). 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 analyzed 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) (Tyanova, Temu, & Cox, 2016). All MS/MS spectra were searched by the Andromeda search engine (Cox et al., 2011) 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 (Tyanova, Temu, Sinitcyn, et al., 2016).

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 tag). Semi-quantitative analysis from enzyme chemoluminiscence data was performed with ImageJ (Rueden et al., 2017) by «plotting lanes» to obtain relative density plots (Sup. Fig. 7).

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

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

DATA AVAILABILITY

RNAseq raw reads were deposited on the NCBI-SRA database under the BioProject number PRJNA753061. R scripts and input files are available at 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»).

SOURCE OF FUNDING

This study was supported by the European Union’s Horizon 2020 research and innovation program under the grant agreement CODOVIREVOL (ERC-2014-CoG-647916) to IGB.

LIST OF FIGURES

LIST OF TABLES

REFERENCES

Agashe, D., Martinez-Gomez, N. C., Drummond, D. A., & Marx, C. J. (2013). Good Codons, Bad Transcript: Large Reductions in Gene Expression and Fitness Arising from Synonymous Mutations in a Key Enzyme. Molecular Biology and Evolution, 30(3), 549–560.

Akashi, H. (1994). Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics, 136(3), 927–935.

Andersson, S. G. E., & Kurland, C. G. (1990). Codon preferences in free-living microorganisms. Microbiological Reviews, 54(2), 198–210.

Angov, E., Hillier, C. J., Kincaid, R. L., & Lyon, J. A. (2008). Heterologous Protein Expression Is Enhanced by Harmonizing the Codon Usage Frequencies of the Target Gene with those of the Expression Host. PLOS ONE, 3(5), e2189.

Bauer, A. P., Leikam, D., Krinner, S., Notka, F., Ludwig, C., Längst, G., & Wagner, R. (2010). The impact of intragenic CpG content on gene expression. Nucleic Acids Research, 38(12), 3891–3908.

Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z., & Blüthgen, N. (2013). Efficient translation initiation dictates codon usage at gene start. Molecular Systems Biology, 9(1), 675.

Bettany, A. J. E., Moore, P. A., Cafferkey, R., Bell, L. D., Goodey, A. R., Carter, B. L. A., & Brown, A. J. P. (1989). 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, 5(3), 187–198.

Bierne, N., & Eyre-Walker, A. (2006). Variation in synonymous codon use and DNA polymorphism within the Drosophila genome. Journal of Evolutionary Biology, 19(1), 1–11.

Boël, G., Letso, R., Neely, H., Price, W. N., Wong, K. H., Su, M., Luff, J. D., Valecha, M., Everett, J. K., Acton, T. B., Xiao, R., Montelione, G. T., Aalberts, D. P., & Hunt, J. F. (2016). Codon influence on protein expression in E. coli correlates with mRNA levels. Nature 2016 529:7586, 529(7586), 358–363.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114–2120.

Bourret, J., Alizon, S., & Bravo, I. G. (2019). COUSIN (COdon Usage Similarity INdex): A Normalized Measure of Codon Usage Preferences. Genome Biology and Evolution, 11(12), 3523–3528.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology 2016 34:5, 34(5), 525–527.

Brightwell, G., Poirier, V., Cole, E., Ivins, S., & Brown, K. W. (1997). Serum-dependent and cell cycle-dependent expression from a cytomegalovirus-based mammalian expression vector. Gene, 194(1), 115–123.

Bulmer, M. (1991). The Selection-Mutation-Drift Theory of Synonymous Codon Usage. Genetics, 129(3), 897.

Burgess-Brown, N. A., Sharma, S., Sobott, F., Loenarz, C., Oppermann, U., & Gileadi, O. (2008). Codon optimization can improve expression of human genes in Escherichia coli: A multi-gene study. Protein Expression and Purification, 59(1), 94–102.

Burow, D. A., Martin, S., Quail, J. F., Alhusaini, N., Coller, J., & Cleary, M. D. (2018). Attenuated Codon Optimality Contributes to Neural-Specific mRNA Decay in Drosophila. Cell Reports, 24(7), 1704–1712.

Callens, M., Pradier, L., Finnegan, M., Rose, C., & Bedhomme, S. (2021). Read between the Lines: Diversity of Nontranslational Selection Pressures on Local Codon Usage. Genome Biology and Evolution, 13(9).

Cambray, G., Guimaraes, J. C., & Arkin, A. P. (2018). Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli. Nature Biotechnology, 36(10), 1005.

Caspersson, T., Farber, S., Foley, G. E., Kudynowski, J., Modest, E. J., Simonsson, E., Wagh, U., & Zech, L. (1968). Chemical differentiation along metaphase chromosomes. Experimental Cell Research, 49(1), 219–222.

Chamary, J. V., & Hurst, L. D. (2005). Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? Trends in Genetics, 21(5), 256–259.

Chamary, J. V., Parmley, J. L., & Hurst, L. D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature Reviews Genetics 2006 7:2, 7(2), 98–108.

Chaney, J. L., Steele, A., Carmichael, R., Rodriguez, A., Specht, A. T., Ngo, K., Li, J., Emrich, S., & Clark, P. L. (2017). Widespread position-specific conservation of synonymous rare codons within coding sequences. PLOS Computational Biology, 13(5), e1005531.

Chen, S. L., Lee, W., Hottes, A. K., Shapiro, L., & McAdams, H. H. (2004). Codon usage between genomes is constrained by genome-wide mutational processes. Proceedings of the National Academy of Sciences, 101(10), 3480–3485.

Chen, S., Li, K., Cao, W., Wang, J., Zhao, T., Huan, Q., Yang, Y. F., Wu, S., & Qian, W. (2017). Codon-Resolution Analysis Reveals a Direct and Context-Dependent Impact of Individual Synonymous Mutations on mRNA Level. Molecular Biology and Evolution, 34(11), 2944–2958.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., & Mann, M. (2011). Andromeda: A peptide search engine integrated into the MaxQuant environment. Journal of Proteome Research, 10(4), 1794–1805. https://doi.org/10.1021/PR101065J/SUPPL_FILE/PR101065J_SI_002.ZIP

Crick, F. (1970). Central Dogma of Molecular Biology. Nature 1970 227:5258, 227(5258), 561–563.

De Smit, M. H., & Van Duin, J. (1990). Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. Proceedings of the National Academy of Sciences, 87(19), 7668–7672.

De Sousa Abreu, R., Penalva, L. O., Marcotte, E. M., & Vogel, C. (2009). Global signatures of protein and mRNA expression levels. Molecular BioSystems, 5(12), 1512–1526.

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

Dittmar, K. A., Sørensen, M. A., Elf, J., Ehrenberg, M., & Pan, T. (2005). Selective charging of tRNA isoacceptors induced by amino-acid starvation. EMBO Reports, 6(2), 151–157.

Dong, H., Nilsson, L., & Kurland, C. G. (1996). Co-variation of tRNA Abundance and Codon Usage inEscherichia coliat Different Growth Rates. Journal of Molecular Biology, 260(5), 649–663.

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., & Arnold, F. H. (2005). Why highly expressed proteins evolve slowly. Proceedings of the National Academy of Sciences, 102(40), 14338–14343.

Duret, L. (2000). tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. Trends in Genetics, 16(7), 287–289.

Duret, L. (2002). Evolution of synonymous codon usage in metazoans. Current Opinion in Genetics & Development, 12(6), 640–649.

Duret, L., & Galtier, N. (2009). Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. Annual review of genomics and human genetics, 10, 285-311.

Elf, J., Nilsson, D., Tenson, T., & Ehrenberg, M. (2003). Selective charging of tRNA isoacceptors explains patterns of codon usage. Science, 300(5626), 1718–1722.

Eskesen, S. T., Eskesen, F. N., & Ruvinsky, A. (2004). Natural Selection Affects Frequencies of AG and GT Dinucleotides at the 5′ and 3′ Ends of Exons. Genetics, 167(1), 543–550.

Fairbrother, W. G., Yeh, R. F., Sharp, P. A., & Burge, C. B. (2002). Predictive identification of exonic splicing enhancers in human genes. Science, 297(5583), 1007–1013.

Fath, S., Bauer, A. P., Liss, M., Spriestersbach, A., Maertens, B., Hahn, P., Ludwig, C., Schäfer, F., Graf, M., & Wagner, R. (2011). Multiparameter RNA and Codon Optimization: A Standardized Tool to Assess and Enhance Autologous Mammalian Gene Expression. PLOS ONE, 6(3), e17596.

Faustino, N. A., & Cooper, T. A. (2003). Pre-mRNA splicing and human disease. Genes & Development, 17(4), 419–437.

Frenkel-Morgenstern, M., Danon, T., Christian, T., Igarashi, T., Cohen, L., Hou, Y. M., & Jensen, L. J. (2012). Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels. Molecular Systems Biology, 8(1), 572.

Frumkin, I., Lajoie, M. J., Gregg, C. J., Hornung, G., Church, G. M., & Pilpel, Y. (2018). Codon usage of highly expressed genes affects proteome-wide translation efficiency. Proceedings of the National Academy of Sciences of the United States of America, 115(21), E4940–E4949.

Galtier, N., Piganeau, G., Mouchiroud, D., & Duret, L. (2001). GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. Genetics, 159(2), 907–911.

Galtier, Nicolas, Roux, C., Rousselle, M., Romiguier, J., Figuet, E., Glémin, S., Bierne, N., & Duret, L. (2018). Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. Molecular Biology and Evolution, 35(5), 1092–1103.

Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S., & Futcher, B. (2014). Measurement of average decoding rates of the 61 sense codons in vivo. ELife, 3.

Garmory, H. S., Brown, K. A., & Titball, R. W. (2003). DNA vaccines: Improving expression of antigens. Genetic Vaccines and Therapy, 1(1), 1–5.

Ghaemmaghami, S., Huh, W. K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K., & Weissman, J. S. (2003). Global analysis of protein expression in yeast. Nature 2003 425:6959, 425(6959), 737–741.

Gouy, M., & Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Research, 10(22), 7055–7074.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., & Pavé, A. (1980). Codon catalog usage and the genome hypothesis. Nucleic Acids Research, 8(1), 197–197.

Gu, W., Zhou, T., & Wilke, C. O. (2010). A Universal Trend of Reduced mRNA Stability near the Translation-Initiation Site in Prokaryotes and Eukaryotes. PLOS Computational Biology, 6(2), e1000664.

Hanson, G., & Coller, J. (2017). Codon optimality, bias and usage in translation and mRNA decay. Nature Reviews Molecular Cell Biology 2017 19:1, 19(1), 20–30.

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

Hershberg, R., & Petrov, D. A. (2008). Selection on Codon Bias. Annual review of genetics, 42, pp.287-299.

Hussmann, J. A., Patchett, S., Johnson, A., Sawyer, S., & Press, W. H. (2015). Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. PLOS Genetics, 11(12), e1005732.

Ikemura, T. (1981). 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. Journal of Molecular Biology, 151(3), 389–409.

Ikemura, T. (1982). 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. Journal of Molecular Biology, 158(4), 573–597.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science, 324(5924), 218–223.

Jallet, A. J., Demange, A., Leblay, F., Decourcelle, M., Koulali, K. El, Picard, M. AL, & Bravo, I. G. (2021). Human cellular homeostasis buffers trans-acting translational effects of heterologous gene expression with very different codon usage bias. BioRxiv, 2021.12.09.471957.

Jeacock, L., Faria, J., & Horn, D. (2018). Codon usage bias controls mRNA and protein abundance in trypanosomatids. ELife, 7.

Johnston, T. C., Borgia, P. T., & Parker, J. (1984). Codon specificity of starvation induced misreading. Molecular and General Genetics MGG 1984 195:3, 195(3), 459–465.

Johnston, T. C., & Parker, J. (1985). Streptomycin-induced, third-position misreading of the genetic code. Journal of Molecular Biology, 181(2), 313–315.

Kanaya, S., Yamada, Y., Kudo, Y., & Ikemura, T. (1999). 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, 238(1), 143–155.

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nature Methods 2015 12:4, 12(4), 357–360.

Kim, J. H., Lee, S. R., Li, L. H., Park, H. J., Park, J. H., Lee, K. Y., Kim, M. K., Shin, B. A., & Choi, S. Y. (2011). High Cleavage Efficiency of a 2A Peptide Derived from PorcineTeschovirus-1 in Human Cell Lines, Zebrafish and Mice. PLoS ONE, 6(4).

Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of expression in escherichia coli. Science, 324(5924), 255–258.

Kurland, C. G. (1992). Translational accuracy and the fitness of bacteria. Annual review of genetics, 26(1), 29-50.

Lampson, B. L., Pershing, N. L. K., Prinz, J. A., Lacsina, J. R., Marzluff, W. F., Nicchitta, C. V., MacAlpine, D. M., & Counter, C. M. (2013). Rare Codons Regulate KRas Oncogenesis. Current Biology, 23(1), 70–75.

Li, G. W., Burkhardt, D., Gross, C., & Weissman, J. S. (2014). Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. Cell, 157(3), 624–635.

Lithwick, G., & Margalit, H. (2003). Hierarchy of Sequence-Dependent Features Associated With Prokaryotic Translation. Genome Research, 13(12), 2665–2673.

Louie, E., Ott, J., & Majewski, J. (2003). Nucleotide Frequency Variation Across Human Genes. Genome Research, 13(12), 2594–2601.

Lu, P., Vogel, C., Wang, R., Yao, X., & Marcotte, E. M. (2006). Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nature Biotechnology 2006 25:1, 25(1), 117–124.

Lynch, M., Gutenkunst, R., Ackerman, M., Spitze, K., Ye, Z., Maruki, T., & Jia, Z. (2017). Population genomics of Daphnia pulex. Genetics, 206(1), 315–332.

Marais, G., & Duret, L. (2001). Synonymous Codon Usage, Accuracy of Translation, and Gene Length in Caenorhabditis elegans. Journal of Molecular Evolution 2001 52:3, 52(3), 275–280.

Mauger, D. M., Joseph Cabral, B., Presnyak, V., Su, S. V., Reid, D. W., Goodman, B., Link, K., Khatwani, N., Reynders, J., Moore, M. J., & McFadyen, I. J. (2019). mRNA structure regulates protein expression through changes in functional half-life. Proceedings of the National Academy of Sciences of the United States of America, 116(48), 24075–24083.

Mauro, V. P., & Chappell, S. A. (2014). A critical analysis of codon optimization in human therapeutics. Trends in Molecular Medicine, 20(11), 604–613.

Menzella, H. G. (2011). Comparison of two codon optimization strategies to enhance recombinant protein production in Escherichia coli. Microbial Cell Factories, 10(1), 1–8.

Mishima, Y., & Tomari, Y. (2016). Codon Usage and 3′ UTR Length Determine Maternal mRNA Stability in Zebrafish. Molecular Cell, 61(6), 874–885.

Moriyama, E. N., & Powell, J. R. (1997). Codon usage bias and tRNA abundance in Drosophila. Journal of Molecular Evolution, 45(5), 514–523.

Musto, H., Cruveiller, S., D'Onofrio, G., Romer, H., & Bernardi, G. (2001). Translational Selection on Codon Usage in Xenopus laevis. Molecular Biology and Evolution, 18(9), 1703–1707.

Nascimento, J. de F., Kelly, S., Sunter, J., & Carrington, M. (2018). Codon choice directs constitutive mRNA levels in trypanosomes. ELife, 7.

Novoa, E. M., Jungreis, I., Jaillon, O., Kellis, M., & Leitner, T. (2019). Elucidation of Codon Usage Signatures across the Domains of Life. Molecular Biology and Evolution, 36(10), 2328–2339.

Orban, T., & Olah, E. (2001). Purifying selection on silent sites – a constraint from splicing regulation? Trends in Genetics, 17(5), 252–253.

Parmley, J. L., & Hurst, L. D. (2007). Exonic Splicing Regulatory Elements Skew Synonymous Codon Usage near Intron-exon Boundaries in Mammals. Molecular Biology and Evolution, 24(8), 1600–1603.

Pechmann, S., & Frydman, J. (2012). Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nature Structural & Molecular Biology 2012 20:2, 20(2), 237–243.

Perach, M., Zafrir, Z., Tuller, T., & Lewinson, O. (2021). Identification of conserved slow codons that are important for protein expression and function. RNA Biology, 18(12), 2296–2307.

Plotkin, J. B., & Kudla, G. (2010). Synonymous but not the same: the causes and consequences of codon bias. Nature Reviews Genetics 2011 12:1, 12(1), 32–42.

Pop, C., Rouskin, S., Ingolia, N. T., Han, L., Phizicky, E. M., Weissman, J. S., & Koller, D. (2014). Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. Molecular Systems Biology, 10(12), 770.

Powell, J. R., & Moriyama, E. N. (1997). Evolution of codon usage bias in Drosophila. Proceedings of the National Academy of Sciences, 94(15), 7784–7790.

Presnyak, V., Alhusaini, N., Chen, Y. H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K. E., Graveley, B. R., & Coller, J. (2015). Codon Optimality Is a Major Determinant of mRNA Stability. Cell, 160(6), 1111–1124.

Princiotta, M. F., Finzi, D., Qian, S. B., Gibbs, J., Schuchmann, S., Buttgereit, F., Bennink, J. R., & Yewdell, J. W. (2003). Quantitating Protein Synthesis, Degradation, and Endogenous Antigen Processing. Immunity, 18(3), 343–354.

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

Radhakrishnan, A., & Green, R. (2016). Connections Underlying Translation and mRNA Stability. Journal of Molecular Biology, 428(18), 3558–3564.

Riba, A., Nanni, N. Di, Mittal, N., Arhné, E., Schmidt, A., & Zavolan, M. (2019). Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates. Proceedings of the National Academy of Sciences of the United States of America, 116(30), 15023–15032.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. Nature Biotechnology 2011 29:1, 29(1), 24–26.

Robinson, M., Lilley, R., Little, S., Emtage, J. S., Yarranton, G., Stephens, P., Millican, A., Eaton, M., & Humphreys, G. (1984). Codon usage can affect efficiency of translation of genes in Escherichia coli. Nucleic Acids Research, 12(17), 6663–6671.

Rueden, C. T., Schindelin, J., Hiner, M. C., DeZonia, B. E., Walter, A. E., Arena, E. T., & Eliceiri, K. W. (2017). ImageJ2: ImageJ for the next generation of scientific image data. BMC Bioinformatics, 18(1), 1–26.

Ryan, M. D., King, A. M. Q., & Thomas, G. P. (1991). Cleavage of foot-and-mouth disease virus polyprotein is mediated by residues located within a 19 amino acid sequence. Journal of General Virology, 72(11), 2727–2732.

Sauna, Z. E., & Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. Nature Reviews. Genetics, 12(10), 683–691.

Shah, P., Ding, Y., Niemczyk, M., Kudla, G., & Plotkin, J. B. (2013). Rate-Limiting Steps in Yeast Protein Translation. Cell, 153(7), 1589–1601.

Sharp, P. M., & Li, W. H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. Journal of Molecular Evolution 1986 24:1, 24(1), 28–38.

Sharp, P. M., Tuohy, T. M. F., & Mosurski, K. R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Research, 14(13), 5125–5143.

Shields, D. C., Sharp, P. M., Higgins, D. G., & Wright, F. (1988). Silent sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Molecular Biology and Evolution, 5(6), 704–716.

Solovyev, V. (2004). Statistical Approaches in Eukaryotic Gene Prediction. Handbook of Statistical Genetics.

Sørensen, M. A., Kurland, C. G., & Pedersen, S. (1989). Codon usage determines translation rate in Escherichia coli. Journal of Molecular Biology, 207(2), 365–377.

Sørensen, M. A., & Pedersen, S. (1991). 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. Journal of Molecular Biology, 222(2), 265–280.

Stenico, M., Lloyd, A. T., & Sharp, P. M. (1994). Codon usage in Caenorhabditis elegans : delineation of translational selection and mutational biases. Nucleic Acids Research, 22(13), 2437–2446.

Stoletzki, N., & Eyre-Walker, A. (2007). Synonymous Codon Usage in Escherichia coli: Selection for Translational Accuracy. Molecular Biology and Evolution, 24(2), 374–381.

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(2), 344–354.

Tuller, T., Kupiec, M., & Ruppin, E. (2007). Determinants of Protein Abundance and Translation Efficiency in S. cerevisiae. PLOS Computational Biology, 3(12), e248.

Tuller, T., Waldman, Y. Y., Kupiec, M., & Ruppin, E. (2010). Translation efficiency is determined by both codon bias and folding energy. Proceedings of the National Academy of Sciences, 107(8), 3645–3650.

Tyanova, S., Temu, T., & Cox, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nature Protocols, 11(12), 2301–2319.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., & Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nature Methods 2016 13:9, 13(9), 731–740.

Urrutia, A. O., & Hurst, L. D. (2001). Codon Usage Bias Covaries With Expression Breadth and the Rate of Synonymous Evolution in Humans, but This Is Not Evidence for Selection. Genetics, 159(3), 1191–1199.

Vogel, C., De Sousa Abreu, R., Ko, D., Le, S. Y., Shapiro, B. A., Burns, S. C., Sandhu, D., Boutz, D. R., Marcotte, E. M., & Penalva, L. O. (2010). Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Molecular Systems Biology, 6(1), 400.

Vogel, C., & Marcotte, E. M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nature Reviews Genetics 2012 13:4, 13(4), 227–232.

Wang, S. E., Brooks, A. E. S., Poole, A. M., & Simoes-Barbosa, A. (2020). Determinants of translation efficiency in the evolutionarily-divergent protist Trichomonas vaginalis. BMC Molecular and Cell Biology, 21(1), 1–13.

Warnecke, T., & Hurst, L. D. (2007). Evidence for a Trade-Off between Translational Efficiency and Splicing Regulation in Determining Synonymous Codon Usage in Drosophila melanogaster. Molecular Biology and Evolution, 24(12), 2755–2762.

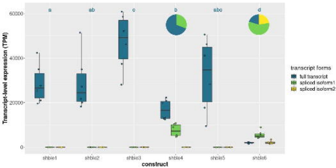
Weinberg, D. E., Shah, P., Eichhorn, S. W., Hussmann, J. A., Plotkin, J. B., & Bartel, D. P. (2016). Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. Cell Reports, 14(7), 1787–1799.

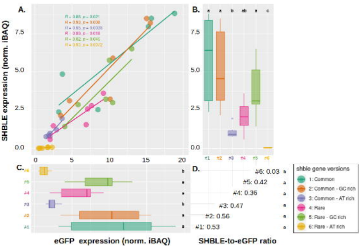
Willie, E., & Majewski, J. (2004). Evidence for codon bias selection at the pre-mRNA level in eukaryotes. Trends in Genetics, 20(11), 534–538.

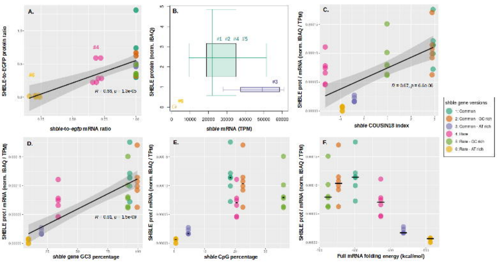
Xia, X. (2014). A major controversy in codon-Anticodon adaptation resolved by a new codon usage index. Genetics, 199(2), 573–579.

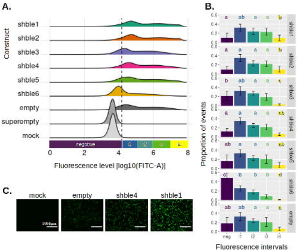
Zhao, F., Yu, C. H., & Liu, Y. (2017). Codon usage regulates protein structure and function by affecting translation elongation speed in Drosophila cells. Nucleic Acids Research, 45(14), 8484–8492.

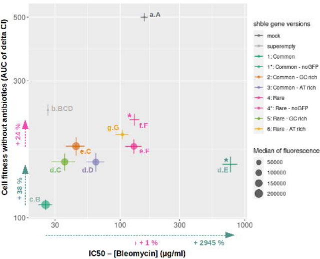
Zucchelli, E., Pema, M., Stornaiuolo, A., Piovan, C., Scavullo, C., Giuliani, E., Bossi, S., Corna, S., Asperti, C., Bordignon, C., Rizzardi, G. P., & Bovolenta, C. (2017). Codon Optimization Leads to Functional Impairment of RD114-TR Envelope Glycoprotein. Molecular Therapy - Methods & Clinical Development, 4, 102–114.

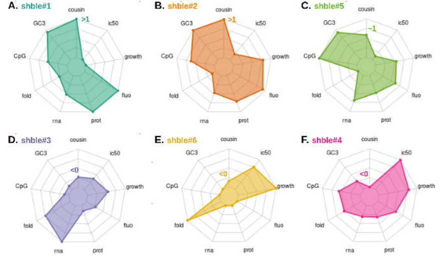

Transcript-level expression (TPM)
60000
40000
20000
0
a
ab
c
b
abc
d
shble1
shble2
shble3
shble4
shble5
shble6
construct
transcript forms
full transcript
spliced isoform1
spliced isoform2

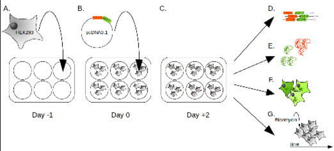

A.
R = 0.88, p = 0.021
R = 0.93, p = 0.008
R = 0.95, p = 0.0038
R = 0.89, p = 0.018
R = 0.82, p = 0.046
R = 0.93, p = 0.0072
SHBLE expression (norm. iBAQ)
7.5
5.0
2.5
0.0
0
5
10
15
20
B.
a
a
b
ab
a
c
7.5
5.0
2.5
0.0
#1
#2
#3
#4
#5
#6
C.
#6
#5
#4
#3
#2
#1
b
a
a
b
a
a
eGFP expression (norm. iBAQ)
D.
#6: 0.03
#5: 0.42
#4: 0.36
#3: 0.47
#2: 0.56
#1: 0.53
b
a
a
a
a
a
SHBLE-to-eGFP ratio
shble gene versions
1: Common
2: Common - GC rich
3: Common - AT rich
4: Rare
5: Rare - GC rich
6: Rare - AT rich


A.
SHBLE-to-EGFP protein ratio
1.0
0.5
0.0
#4
#6
R = 0.88, p = 1.3e-05
0.25
0.50
0.75
1.00
shble-to-egfp mRNA ratio
B.
SHBLE protein (norm. IBAQ)
5
4
3
2
1
0
#1 #2 #4 #5
#3
#6
0
10000
20000
30000
40000
50000
60000
shble mRNA (TPM)
C.
SHBLE prot / mRNA (norm. IBAQ / TPM)
R = 0.67, p = 6.6e-06
-1
0
1
2
3
shble COUSIN18 index
shble gene versions
1: Common
2: Common - GC rich
3: Common - AT rich
4: Rare
5: Rare - GC rich
6: Rare - AT rich
D.
SHBLE prot / mRNA (norm. IBAQ / TPM)
R = 0.91, p = 1.5e-09
25
50
75
100
shble gene GC3 percentage
E.
SHBLE prot / mRNA (norm. IBAQ / TPM)
0
10
20
30
shble CpG percentage
F.
SHBLE prot / mRNA (norm. IBAQ / TPM)
Full mRNA folding energy (kcal/mol)


A.
Construct
shble1
shble2
shble3
shble4
shble5
shble6
empty
superempty
mock
negative
i1
i2
i3
i4
0
2
4
6
8
Fluorescence level [log10(FITC-A)]
B.
Proportion of events
shble1: a ab a a b
shble2: a a a a b
shble3: b ab a a c
shble4: a a a a ab
shble5: ab a a a ab
shble6: c b b b d
empty: ab ab a a a
0.0
0.2
0.4
0.6
neg
i1
i2
i3
i4
Fluorescence intervals
C.
mock
empty
shble4
shble1
1000µm


a.A
b.BCD
g.G
f.F
e.F
e.C
d.C
d.D
d.E
c.B
*
*
Cell fitness without antibiotics (AUC of delta CI)
500
300
100
+ 24 %
+ 38 %
30
100
300
1000
+ 1 %
+ 2945 %
IC50 – [Bleomycin] (µg/ml)
shble gene versions
mock
superempty
1: Common
1*: Common - noGFP
2: Common - GC rich
3: Common - AT rich
4: Rare
4*: Rare - noGFP
5: Rare - GC rich
6: Rare - AT rich
Median of fluorescence
50000
100000
150000
200000


A. shble#1
cousin
>1
ic50
GC3
growth
CpG
fold
fluo
rna
prot
B. shble#2
cousin
>1
ic50
GC3
growth
CpG
fold
fluo
rna
prot
C. shble#5
cousin
~1
ic50
GC3
growth
CpG
fold
fluo
rna
prot
D. shble#3
cousin
<0
ic50
GC3
growth
CpG
fold
fluo
rna
prot
E. shble#6
cousin
<0
ic50
GC3
growth
CpG
fold
fluo
rna
prot
F. shble#4
cousin
<0
ic50
GC3
growth
CpG
fold
fluo
rna
prot


A.
HEK293
B.
pcDNA3.1
C.
Day -1
Day 0
Day +2
D.
eGFP
eGFP
eGFP
E.
F.
G.
Bleomycin
time