



**HAL**  
open science

## Transcriptomic, proteomic and functional consequences of codon usage bias in human cells during heterologous gene expression

Marion a L Picard, Fiona Leblay, Cécile Cassan, Anouk Willemsen, Josquin Daron, Frédérique Bauffe, Mathilde Decourcelle, Antonin Demange, Ignacio G Bravo

### ► To cite this version:

Marion a L Picard, Fiona Leblay, Cécile Cassan, Anouk Willemsen, Josquin Daron, et al.. Transcriptomic, proteomic and functional consequences of codon usage bias in human cells during heterologous gene expression. *Protein Science*, 2023, 10.1002/pro.4576 . hal-03955970

**HAL Id: hal-03955970**

**<https://hal.science/hal-03955970>**

Submitted on 25 Jan 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Transcriptomic, proteomic and functional consequences**  
2 **of codon usage bias in human cells during heterologous gene expression.**

3 **RUNNING TITLE: From codon usage to protein function in human cells.**

4 **AUTHOR LIST:** Marion A.L. Picard<sup>1\*</sup>, Fiona Leblay<sup>1</sup>, Cécile Cassan<sup>1</sup>, Anouk Willemsen<sup>1</sup>, Josquin Daron<sup>1</sup>,  
5 Frédérique Bauffe<sup>1</sup>, Mathilde Decourcelle<sup>2</sup>, Antonin Demange<sup>1</sup>, Ignacio G. Bravo<sup>1\*</sup>

6 **AUTHOR AFFILIATIONS**

7 <sup>1</sup>Laboratory MIVEGEC (CNRS, IRD, University of Montpellier), French National Center for Scientific  
8 Research, Montpellier, France

9 <sup>2</sup>BioCampus Montpellier (University of Montpellier, CNRS, INSERM), Montpellier, France.

10 \*Corresponding authors

11 E-mail: marion.picard@univ-perp.fr (MALP), ignacio.bravo@cnrs.fr (IGB)

12 **AUTHOR CONTRIBUTIONS**

13 Funding Acquisition, Project Administration and Supervision : IGB; Methodology : IGB, AD, MALP;

14 Investigation : MALP, FL, CC, AW, FB, JD, MD, AD; Data Curation : MALP; Formal Analysis : MALP, IGB;

15 Visualization : MALP; Conceptualization and Writing : MALP, IGB.

16 **KEYWORDS**

17 Synonymous codon recoding, heterologous gene expression, translation, genotype to phenotype, mutation-  
18 selection

## 19 **ABSTRACT**

20 Differences in codon frequency between genomes, genes, or positions along a gene, modulate  
21 transcription and translation efficiency, leading to phenotypic and functional differences. Here, we present a  
22 multiscale analysis of the effects of synonymous codon recoding during heterologous gene expression in human  
23 cells, quantifying the phenotypic consequences of codon usage bias at different molecular and cellular levels,  
24 with an emphasis on translation elongation.

25 Six synonymous versions of an antibiotic resistance gene were generated, fused to a fluorescent reporter,  
26 and independently expressed in HEK293 cells. Multiscale phenotype was analysed by means of: quantitative  
27 transcriptome and proteome assessment, as proxies for gene expression; cellular fluorescence, as a proxy for  
28 single-cell level expression; and real-time cell proliferation in absence or presence of antibiotic, as a proxy for  
29 the cell fitness.

30 We show that differences in codon usage bias strongly impact the molecular and cellular phenotype: (i)  
31 they result in large differences in mRNA levels and in protein levels, leading to differences of over fifteen times  
32 in translation efficiency; (ii) they introduce unpredicted splicing events; (iii) they lead to reproducible  
33 phenotypic heterogeneity; and (iv) they lead to a trade-off between the benefit of antibiotic resistance and the  
34 burden of heterologous expression.

35 In human cells in culture, codon usage bias modulates gene expression by modifying mRNA availability  
36 and suitability for translation, leading to differences in protein levels and eventually eliciting functional  
37 phenotypic changes.

## 38 **IMPORTANCE**

39 Synonymous codons encode for the same amino acid, but they are not neutral regarding gene expression  
40 or protein synthesis. Bias between synonymous codons have evolved naturally and are also applied in  
41 biotechnology protein production. We have studied the multilevel impact of codon usage on a human cell

42 system. We show that differences in codon usage lead to transcriptomic, proteomic and functional changes,  
43 modulating gene expression and cellular phenotype.

#### 44 INTRODUCTION

45 The canonical scenario of gene expression posits that DNA sequences are first transcribed into  
46 messenger RNA (mRNA) molecules that are secondly translated into proteins, so that one given nucleotide  
47 sequence encodes one predictable amino acid sequence <sup>1</sup>. The initial version of this scenario did not provide any  
48 explanation on how a unique set of genes could be associated with several cellular phenotypes. Over the last  
49 decades, a large body of studies on gene expression have addressed this question and revealed multi-level  
50 regulation mechanisms increasing the diversity of the proteomic outputs that can be produced from a given  
51 genome. The standard genetic code that establishes a correspondence between the DNA coding units (*i.e.* the  
52 codon, a triplet of nucleotides, 64 in total) and the protein building blocks (*i.e.* the amino acids, 20 in total) is  
53 degenerated: 18 of the amino acids can individually be encoded by two, three, four or six codons, known as  
54 synonymous codons. In a first null hypothesis approach, one would expect synonymous codons to display  
55 similar frequencies. Instead, codon usage bias (CUBias, *i.e.* the uneven representation of synonymous codons <sup>2</sup>)  
56 has been reported in a multiplicity of organisms, and varies not only between species but also within a given  
57 genome or even along positions in a gene <sup>3-8</sup>.

58 The origin and the contribution of the different neutral and/or selective forces shaping CUBias constitute  
59 a classical research subject in evolutionary genetics. The hypothesis of translational selection proposes that  
60 differences in CUBias result in gene expression variations that ultimately lead to phenotypic differences, which  
61 could be subject to natural selection. Indeed, it has been established that variation in CUBias might constitute an  
62 additional layer of gene expression modulation <sup>9-11</sup>. Notably, genetic engineering has extensively resorted to  
63 CUBias recoding for enhancing heterologous protein production, for its use in industrial applications or for  
64 vaccine design <sup>12-15</sup>. The interaction between CUBias and the translation machinery has been well established,  
65 for instance in: (i) the co-variation of genomic CUBias and the tRNA content, from unicellular organisms <sup>4,16,17</sup> to

66 metazoa (*Caenorhabditis elegans*<sup>18</sup>, *Drosophila*<sup>19-21</sup>, or humans<sup>22</sup>; (ii) the correspondence between CUBias and  
67 expression level in bacteria<sup>23</sup> or in yeast<sup>24,25</sup>; (iii) the increase in translation efficiency in bacteria when  
68 supplementing *in trans* with rare tRNAs<sup>26</sup>; or (iv) the changes in tumorigenic phenotype in mice when switching  
69 from rare to common codons in the sequence of a cancer-related GTPase<sup>27</sup>.

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

79 Finally, abundance and chemistry of transfer RNAs (tRNAs) introduces an additional layer of  
80 complexity (and thus an opportunity for regulation). In fast growing unicellular organisms, the tRNA gene  
81 content matches well codon usage preferences of the organism<sup>59</sup>. Heterologous expression can thus be hampered  
82 by the lack or rarity of a tRNA that corresponds to a rare codon in the expression system of choice.  
83 Biotechnology engineering has circumvented this limitation by providing *in trans* the required tRNAs, encoding  
84 them into helper plasmids such as pRIG or pRARE<sup>60</sup>. Further, many genomes actually do not contain dedicated  
85 tRNAs to decode each codon: *e.g.* for all amino acids encoded by two codons ending in C or U (Phe, Asn, Asp  
86 and His) the human genome contains only the tRNAs corresponding to the C-ending codons, which decode also  
87 the U-ending counterparts<sup>61</sup>. Indeed, tRNAs are heavily modified and carry non-canonical nucleotides, which is  
88 often the case for an inosine residue in the first anticodon position<sup>62</sup>. The non Watson-Crick base pairing  
89 interactions available to inosine allow to broaden codon-anticodon recognition<sup>63</sup>, so that in bacteria and  
90 eukaryotes modified tRNAs carrying inosine in the anticodon can decode several synonymous codons for the

91 amino acids Thr, Ala, Pro, Ser, Leu, Ile, Val and Arg <sup>64</sup>. Thus, tRNA modification allows for one-to-many  
92 anticodon-to-codon translation potential, which may have had implications for the evolution of the protein  
93 repertoire in eukaryotes <sup>65</sup>.

94 In this study, we aim at providing an integrated view of the molecular and cellular impact of alternative  
95 CUBias of a heterologous gene expressed in human cells. By combining transcriptomics, proteomics,  
96 fluorescence analysis and cell growth evaluation, we attempt to describe qualitatively, and to quantify as far as  
97 possible, the impact of CUBias and sequence composition of our focal heterologous gene on its own expression.  
98 These are usually called the *cis*-effects of CUBias on gene expression.

## 99 RESULTS

### 100 1. Design of six synonymous gene versions that explore a large sequence space and cover a broad range of 101 sequence composition variables.

102 With the aim of analysing the effects of CUBias on protein synthesis, we have generated six  
103 synonymous variants of the gene encoding for the bleomycin-resistance protein from the bacterium  
104 *Streptoalloteichus hindustanus* (*shble*). We have chosen this heterologous protein as a reporter gene because it  
105 displays a mechanism of action (scission of DNA strands, <sup>66</sup>) which is independent of translation, precisely the  
106 process that we aim to study. For all six *shble* versions we added an AU1 epitope tag in the N-terminus with the  
107 same nucleotide sequence, so that translation initiation will be similar for all *shble* versions and we can focus on  
108 the impact of CUBias on translation elongation. The *shble* ORFs were in-frame coupled via a P2A epitope to an  
109 *egfp* gene that encodes for a fluorescent protein reporter. The nucleotide sequence encoding for the P2A peptide  
110 was identical for all sequences and corresponds to that in the plasmid backbone. The expected heterologous  
111 transcript was a 1,602 base pair (bp) long mRNA encompassing a 1,182bp coding sequence (CDS). The CDS  
112 spanned the *AU1*-tag sequence in 5', the *shble* bleomycin resistance reporter, the *P2A* peptide sequence inducing  
113 ribosomal skipping, and the *egfp* fluorescent reporter (Sup. Fig. 1). The presence of the AU1 epitope allowed us  
114 to use the same antibodies to detect the N-terminus of the SHBLE protein. The presence of the P2A peptide

115 (NPGP motif) induces ribosome skipping <sup>67</sup>, meaning that the ribosome does not perform the Gly-Pro  
116 transpeptidation bond and releases instead the AU1-SHBLE moiety and continues translation of the EGFP  
117 moiety. The synonymous versions of the *shble* were thus the only differences between constructs, and were  
118 characterized by distinct degrees of similarity to the average human CUBias (estimated using the COdon Usage  
119 Similarity INDEX, COUSIN <sup>68</sup>), GC composition at the third nucleotide of codons (GC3), and CpG dinucleotide  
120 frequency (CpG) (Table 1). Modifications in the *shble* sequence also entailed variations on the mRNA folding  
121 energy, calculated using the Vienna RNAfold webserver <sup>69</sup>(Table 1). These four parameters combined allowed  
122 for a good discrimination of all constructs (Sup. Fig. 2), partly reflecting sequence similarities (Sup. Table 1).  
123 The COUSIN index quantifies the match between the CUBias of a focal sequence (in our case the different *shble*  
124 synonymous versions) and the chosen reference (in our case the average CUBias of human genes) <sup>68</sup>. Briefly, the  
125 COUSIN is a normalised index so that a value of 1 corresponds to a focal sequence with similar CUBias to the  
126 reference; values above 1 correspond to similar CUBias to those in the reference, but of larger magnitude; a  
127 value of 0 corresponds to a lack of CUBias; and negative values correspond to CUBias opposite to those in the  
128 reference. The values for the COUSIN index exemplify the large variation in CUBias explored by our construct  
129 repertoire, ranging from “hyper-humanised” versions (namely *shble*#1 and *shble*#2, with COUSIN values above  
130 2) to strongly “de-humanised” versions (namely *shble*#4 and *shble*#6, with negative COUSIN values).

131 **Table 1. Experimental conditions: the different constructs, and their sequence composition variables.** Codon  
132 Usage Similarity index (COUSIN) values have been calculated against the average CUBias of human genes, as  
133 in <sup>68</sup>. Folding energy values for the total mRNA transcripts have been calculated using the RNAfold Webserver <sup>69</sup>.

Condition	Description	COUSIN of the <i>shble</i> sequence <sup>o</sup>	%GC3 of the <i>shble</i> sequence	%CpG of the <i>shble</i> sequence	Folding energy of the total transcript (kcal/mol)
<b>shble#1</b>	The most common codons in the human genome	2.93	93.08	18.46	-649.34
<b>shble#2</b>	The GC-richest among the most common codons	2.982	99.23	22.56	-673.07
<b>shble#3</b>	The AT-richest among the most common codons	-0.414	20.00	4.62	-581.47
<b>shble#4</b>	The rarest codons in the human genome	-1.651	33.85	20.51	-613.49
<b>shble#5</b>	The GC-richest among the rarest codons	0.973	91.54	35.90	-687.76
<b>shble#6</b>	The AT-richest among the rarest codons	-0.924	9.23	0.51	-543.50
<b>#empty</b>	No <i>shble</i> but only <i>EGFP</i> CDS	n.a.	n.a.	n.a.	n.a.

#superempty	Neither <i>shble</i> nor <i>EGFP</i> CDS	n.a.	n.a.	n.a.	n.a.
mock	No plasmid	n.a.	n.a.	n.a.	n.a.

134 **2. Differences in CUBias of the *shble* gene resulted in differences in transcription.**

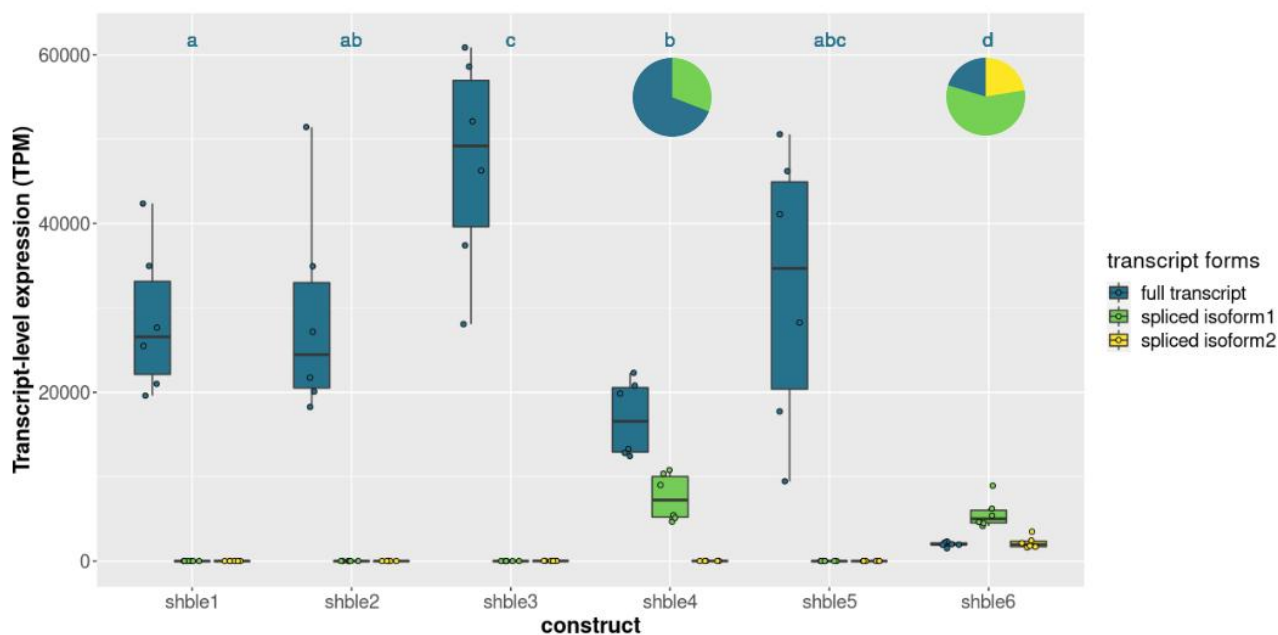
135 After transfection, we monitored DNA levels for transfection efficiency using quantitative PCR (qPCR)  
136 and we monitored mRNA levels for transcription efficiency using retrotranscription followed by qPCR (rt-  
137 qPCR) and RNA-sequencing (RNASeq). Analysis of the RNASeq read distribution revealed the presence of  
138 splicing events within the *shble* CDS for the two constructs with the lowest similarity to the human average  
139 CUBias, namely *shble*#4 (construct using the rarest codon for each amino acid) and *shble*#6 (using rare and AT-  
140 rich codons) (Sup. Fig. 3). The *shble*#6 transcript presented two spliced forms, using the same 5' donor position  
141 and differing in three nucleotides at the 3' acceptor position. The *shble*#4 transcript presented one spliced form,  
142 with donor and acceptor positions in the precise same locations observed for *shble*#6, despite the lack of identity  
143 in the intron-exon boundaries. The spliced intron (either 306 or 309 nucleotides long) was fully comprised  
144 within the 396 bp long *shble* sequence (Sup. Fig. 4), and the event did not involve any frameshift. Thus, *shble*  
145 splicing resulted in the ablation of the SHBLE protein coding potential without affecting the start codon and  
146 without modifying the EGFP coding potential. It is important to state that none of these alternative splicing  
147 events was predicted by the HSF (Human Splicing Finder)<sup>70</sup> nor the SPLM<sup>71</sup> splice detection algorithms used  
148 for sequence scanning during design. Analysis of mRNA abundances showed that the first spliced isoform  
149 (shared by both affected conditions) represented about 30% of the heterologous transcripts for *shble*#4, and 56%  
150 for *shble*#6. The second spliced isoform, exclusively found in condition *shble*#6, corresponded to 22% of the  
151 heterologous transcripts (Figure 1).

152 Full-length mRNA quantification showed differences in transcript levels across conditions, as follows:  
153 (i) the highest values were found in *shble*#3 (using the AT-richest among common codons); (ii) the variance was  
154 largest in *shble*#5 (using the GC-richest among rare codons); and (iii) *shble*#4 and *shble*#6 displayed the lowest  
155 mRNA abundance even when considering the sum of all isoforms (Figure 1, Sup. Table 2). We further verified  
156 that variations in transcript levels were not related to variations in transfection efficiency, by correcting the



157 transcript levels after the plasmid DNA levels in each sample. After this normalisation, the above described  
158 pattern remained unchanged (Sup. Fig. 5). This suggests that variations in mRNA levels are not due to  
159 differences in the DNA level, and may instead be linked to the differentially recoded *shble* sequences.

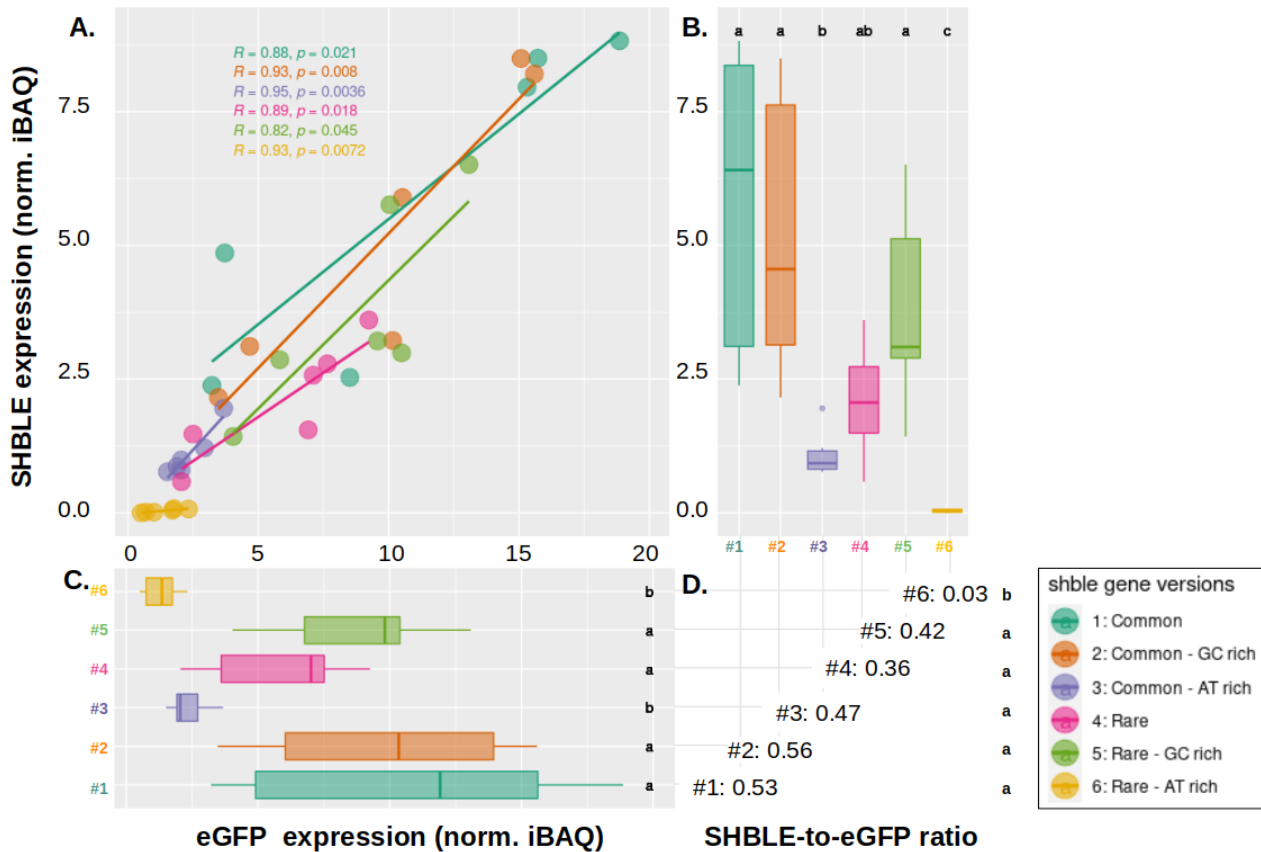
160 In order to allow for further comparison between mRNA and protein levels, while accounting for the  
161 differential splice events, we have taken into account that the SHBLE protein was exclusively encoded by the  
162 full-length mRNA, while the EGFP protein could be translated from any of the three transcript isoforms. Hence,  
163 we used the ratio full-length mRNA over total heterologous transcripts (*i.e.* full-to-total ratio) to estimate the  
164 ratio of SHBLE-encoding over EGFP-encoding transcripts. This ratio was about 69% for *shble*#4, while for  
165 *shble*#6 it was close to 21% (Sup. Table 2). For the rest of the constructs, there was virtually no read  
166 corresponding to spliced transcripts and the ratio was in all cases above 99.96% (Sup. Table 2).



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

### 174 3. Differences in CUBias of the *shble* gene resulted in differences in SHBLE and EGFP protein levels.

175 After transfection, and in paired samples with the mRNA analyses, we quantified SHBLE and EGFP  
176 protein levels by means of western-blot (Sup. Fig. 6, 7 and 8) and of label-free proteomics (Figure 2). Label-free  
177 proteomic analysis allowed to detect EGFP proteins for all constructs, with EGFP abundance in *shble#3* and  
178 *shble#6* being significantly lower than in other conditions (respectively 2.05 and 1.35 normalized iBAQ values,  
179 compared to an overall median of 10.08 for the other constructs) (Figure 2C, Sup. Table 3). The SHBLE protein  
180 was detected in all conditions but, for *shble#6*, it displayed extremely low abundance in five replicates and was  
181 not detected in one replicate (normalized iBAQ value of 0.03) (Figure 2B, Sup. Table 3). Further, the *shble#3*  
182 condition displayed lower SHBLE protein levels than the remaining four other constructs (normalized iBAQ  
183 value of 0.93, compared to an overall median of 3.83) (Figure 2B, Sup. Table 3). Within a given condition,  
184 values for SHBLE and EGFP protein levels displayed a strong, positive correlation, albeit with a particular  
185 expression pattern for version *shble#6* (Pearson's R coefficients ranging from 0.82 to 0.95 depending on the  
186 condition; all p-values < 0.05; Figure 2A). The overall SHBLE-to-EGFP ratio was  $0.46 \pm 0.1$  for all constructs  
187 (ranging between 0.36 and 0.56 for the individual constructs), the exception being *shble#6*, which displayed very  
188 low ratio (0.03), as expected given the very low SHBLE levels (Figure 2D). For this specific construct, we find  
189 good correlation between SHBLE and EGFP levels (Pearson's R = 0.93, p = 0.0072, Figure 2A), but the slope  
190 linking them is ten times lower than for any other construct (Figure 2D). Label-free proteomic quantification  
191 results were validated by image-based western blot quantification (Sup. Fig. 6, 7 and 8).



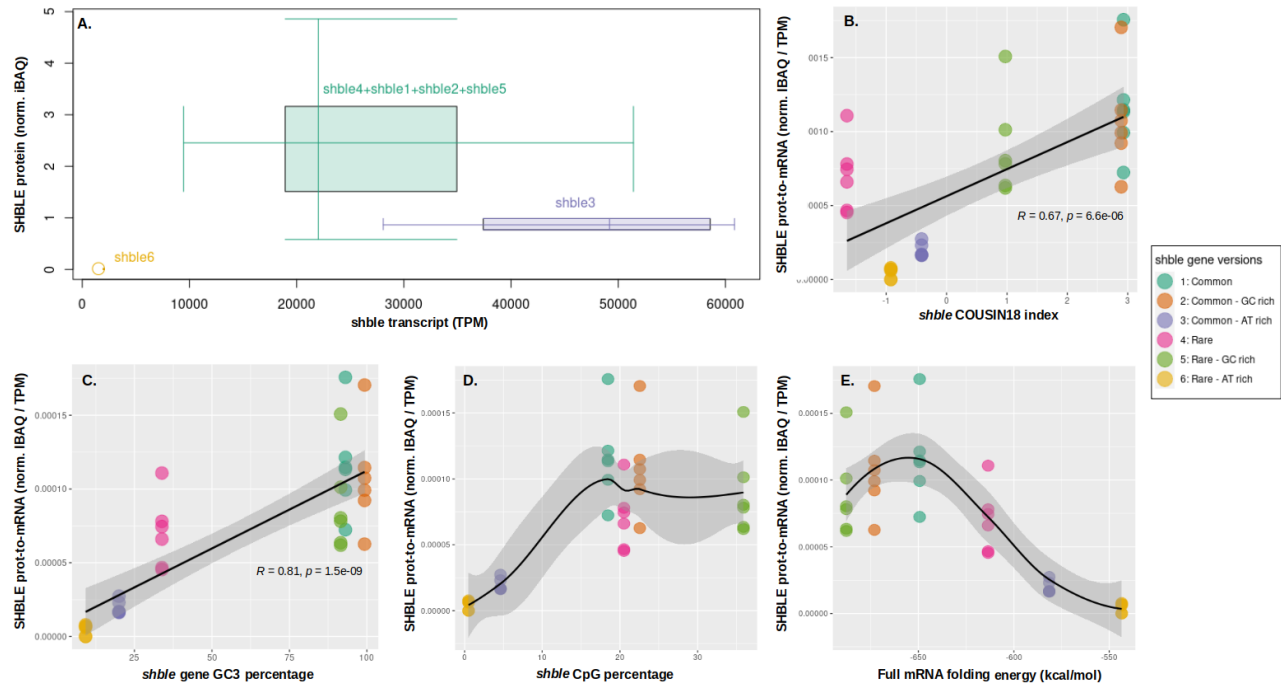
192 **Figure 2. Expression of SHBLE and EGFP at the proteomic level, and relation between them.** Panel A:  
 193 Pearson's correlation between SHBLE (y axis) and EGFP (x axis) protein levels. Six different conditions are  
 194 shown: shble#1 (dark green), shble#2 (orange), shble#3 (purple), shble#4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

201 **4. Differences in CUBias and mRNA physicochemical properties partly explain differences in translation**  
 202 **efficiency.**

203 After separately analysing mRNA and protein levels in cells transfected with the different *shble*  
204 versions, we aimed at establishing a connection between transcription and translation levels. Because SHBLE  
205 and EGFP protein analyses led to similar results, we focus here only on SHBLE. We chose to normalise the  
206 protein levels over the corresponding mRNA levels, and we interpret this protein-to-mRNA ratio as a proxy for  
207 translation efficiency. The median values of the protein-to-mRNA ratio were similar for constructs #1, #2, #4  
208 and #5, whereas conditions #3 and #6 were discordant (Figure 3A and Sup. Table 4): translation efficiency is  
209 over five times lower for condition #3 (which displayed high transcription levels) and over thirteen times lower  
210 for condition #6 (Sup. Table 4). Overall, variation in full-length transcript levels explained 45% of the variation  
211 in SHBLE protein levels (Pearson's  $R=0.45$ ,  $p = 0.0054$ ) (Sup. Fig. 9). This explanatory power of mRNA levels  
212 over protein levels increased to 68% when considering only conditions #1, #2, #4 and #5 (Pearson's  $R=0.68$ ,  $p =$   
213  $0.00025$ ). As discussed below, these values fit well in previous descriptions in the literature about the  
214 explanatory power of variations at the mRNA level to account for variations at the protein level for eukaryotic  
215 cells<sup>72</sup>.

216 In order to understand the differential translation efficiency between constructs, we explored the  
217 explanatory potential of four sequence composition and mRNA physicochemical parameters. We observe first  
218 that the closer the CUBias of the *shble* synonymous versions to the average human CUBias, the higher the  
219 translation efficiency in our human cells in culture (Pearson's  $R=0.67$ ,  $p=6.6e-6$ , Figure 3B). The exception to  
220 this trend was condition *shble#4* which displayed the lowest match to the human CUBias, but a higher protein-  
221 to-transcript ratio than *shble#6* or *shble#3* (Figure 3B). The lower ratio for these two later conditions could be  
222 explained at the light of the three other tested parameters. Indeed, an increase of GC3 content corresponded  
223 monotonically to an increase in the protein-to-transcript ratio (Pearson's  $R=0.81$ ,  $p=1.5e-9$ , Figure 3C), and  
224 *shble#6* and *shble#3* had the lowest GC3 content. Increase in CpG frequency (Figure 3D) resulted in an  
225 increased translation efficiency that reached a plateau for all recoded forms beyond 20% CpG presence, even for  
226 the very CpG-rich form *shble#5*. Finally, variation in mRNA folding energy (Figure 3E), corresponded to a bell-  
227 shaped variation in SHBLE protein-to-transcript ratio so that both low and high values resulted in decreased

228 translation efficiency . Thus, the shble#3 condition combined suboptimal values for all four studied  
 229 characteristics and resulted in poorly efficient translation in spite of the high mRNA levels (see part 2). In  
 230 contrast, shble#1 (recoded using the most used codons), displayed maximum values for each parameter, thus  
 231 resulting in the most efficient translation (highest protein-to-mRNA ratio).



232 **Figure 3. Variation of translation efficiency as a function of CUBias and mRNA physicochemical parameters.**  
 233 **Panel A:** Combined distribution of SHBLE protein level (y axis - normalized iBAQ) and shble transcript level (x  
 234 axis - TPM); individual construct boxes are condensed in a single one when the squares defined by the first and  
 235 third quartiles overlaps (which is the case for shble#1, shble#2, shble#4 and shble#5, shown condensed in dark  
 236 green). For each construct, median values are given in Sup. Table 4. **Panel B:** Pearson's correlation between  
 237 SHBLE protein-to-mRNA ratio and COUSIN index of the shble recoded version. **Panel C:** Pearson's correlation  
 238 between the SHBLE protein-to-mRNA ratio and the GC3 percentage of the shble recoded version. **Panel D:**  
 239 SHBLE protein-to-mRNA ratio variations depending on CpG frequency of the shble recoded version. **Panel E:**  
 240 Correspondence between the SHBLE protein-to-mRNA ratio and the folding energy of the corresponding  
 241 transcript. Curves in panels D and E correspond to a LOWESS (LOcally WEighted Scatter-plot Smoother) local  
 242 polynomial regression to visually display co-variation between the two variables plotted. The results for six full  
 243 biological replicates are shown, each of them with independent RNAseq measurements but pooled by pairs for  
 244 the label-free proteomic analysis.

245 **5. Differences in CUBias lead to differences in EGFP protein expression at population level, but also at**  
 246 **single-cell level.**

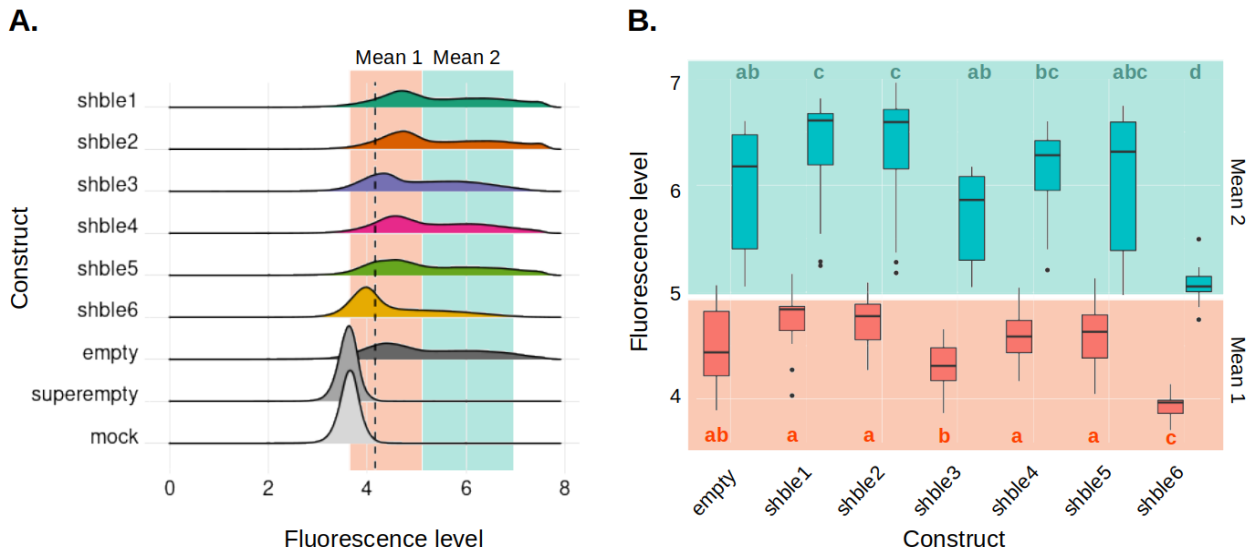
247 We have demonstrated above that variation in SHBLE protein levels were highly correlated to variation  
 248 in the EGFP fluorescent reporter (Figure 2A). On this basis, and in order to further assess the phenotypic  
 249 variation at the single-cell level, we performed an extensive analysis of the cell-based fluorescence values of 16  
 250 transfection replicates by means of fluorescent cytometry analyses. We verified first that the total fluorescence  
 251 signal (*i.e.* the total fluorescence levels in the cell population) was strongly correlated to the EGFP level  
 252 estimated by the label-free proteomics (Pearson's  $R=0.86$ ,  $p=4.8e-15$ , Sup. Fig. 10). We observed then that the  
 253 single-cell distribution of this fluorescence signal was (i) different for all conditions from that obtained with cells  
 254 expressing EGFP alone (*i.e.* "empty" control; individual Anderson-Darling test results shown in Table 2); and (ii)  
 255 multimodal for all the conditions expressing EGFP (Figure 4A, Sup. Fig. 11). We have approximated these  
 256 multimodal cell populations by means of curve deconvolution, and showed that a composite distribution based  
 257 on two underlying Gaussian-like cell populations fitted well the observed distributions (Sup. Fig. 12). We  
 258 conclude thus that synonymous variation of the upstream *shble* sequence modulated and modified the individual  
 259 cell fluorescence phenotype, and that for a given version of the *shble* sequence, cells were differentially  
 260 impacted by the construct expression, overall defining two subpopulations of low or high EGFP expression.

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

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

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

267 For each condition, we describe the fluorescence behaviour of the whole cell population using the  
268 following summary statistics (Table2): (i) the fraction of cells displaying fluorescence above the cell  
269 autofluorescence threshold (discontinuous line in Figure 4A); (ii) the total fluorescence value of the whole  
270 population; (iii) the median fluorescence value of the population; (iv) the mean fluorescence value for each  
271 underlying Gaussian populations. We observed that the median fluorescence value of the population correlated  
272 very well with the overall fluorescence ( $R=0.85$ ,  $p\text{-value} < 2.2e-16$ , Sup. Fig. 13), but that the later allowed for a  
273 better discrimination between conditions. Conditions shble#1 and shble#2 displayed the highest fluorescence  
274 values, while shble#3 displayed ca. 2.5 times lower fluorescence values and shble#6 over seven times lower  
275 fluorescence values (Table 2, Sup. Fig. 13). Differences in total fluorescence levels reflected a reproducible  
276 impact of the synonymous construct expression on the complete cell population, independently of whether  
277 individual cells displayed very high or very low fluorescence: indeed, between each condition, both underlying  
278 Gaussian curves shifted following the same pattern, as illustrated by the variations of their mean values (Figure  
279 4B, Table 2). When combining all our summary statistic variables into a principal component analysis for  
280 describing the cellular population fluorescence we observed that indeed shble#6, and to a lesser extent shble#3,  
281 were the most divergent conditions, characterized by the highest proportion of negative or low-fluorescent cells,  
282 while shble#1 and shble#2 displayed very similar behaviour characterized by high fluorescence values in all  
283 scores (Sup. Fig. 14). These results strengthened the observations obtained by the label-free proteomic  
284 experiments, and underlied the cell-to-cell reproducibility of the impact of synonymous substitutions.



285 **Figure 4. Distribution of the fluorescence signal for the different constructs, and mean values of the two**  
 286 **gaussian curves modeling the fluorescence distribution.** Panel A depicts the density of the green fluorescence  
 287 signal ( $\log_{10}(\text{FITC-A})$ ) considering 480,000 individual cells for each condition: shble#1 (most common codons,  
 288 dark green), shble#2 (common and GC-rich codons, orange), shble#3 (common and AT-rich codons, purple),  
 289 shble#4 (rarest codons, pink), shble#5 (rare and GC-rich codons, orange light green), shble#6 (rare and AT-rich  
 290 codons, yellow). The positive control is "empty" (i.e. transfected cells, expressing EGFP without expressing  
 291 SHBLE, in dark grey); and the negative controls are "superempty" (i.e. transfected cells, not expressing EGFP  
 292 nor SHBLE, in medium grey) and "mock" (i.e. untransfected cells, in light grey). The dashed black line shows  
 293 the threshold for positivity (14,453 green fluorescence units, corresponding to 4.16 in a  $\log_{10}$  scale). Panel B  
 294 represents the first gaussian mean1 (population of lower intensity, in red), and the mean2 (population of higher  
 295 fluorescence intensity, in blue). Values in the y-axis (cell fluorescence) are continuous, and the red and blue  
 296 colours are for representation purposes only. For each category (mean1 and mean 2), the statistical test is a  
 297 pairwise Wilcoxon rank sum test, with Benjamini-Hochberg adjusted p-values on sixteen biological replicates:  
 298 for each colour, conditions associated to a same letter do not display different median values of the  
 299 corresponding variable.

### 300 5. Differences in CUBias of the *shble* gene resulted in differences in cell growth dynamics and antibiotic 301 resistance.

302 Finally, since our *shble* reporter gene confers resistance to the bleomycin antibiotic, we aimed at  
 303 quantifying the functional impact of the different molecular phenotypes described above on cellular fitness. For  
 304 this, we performed a real-time cell growth analysis, both in presence and in absence of antibiotics. For all

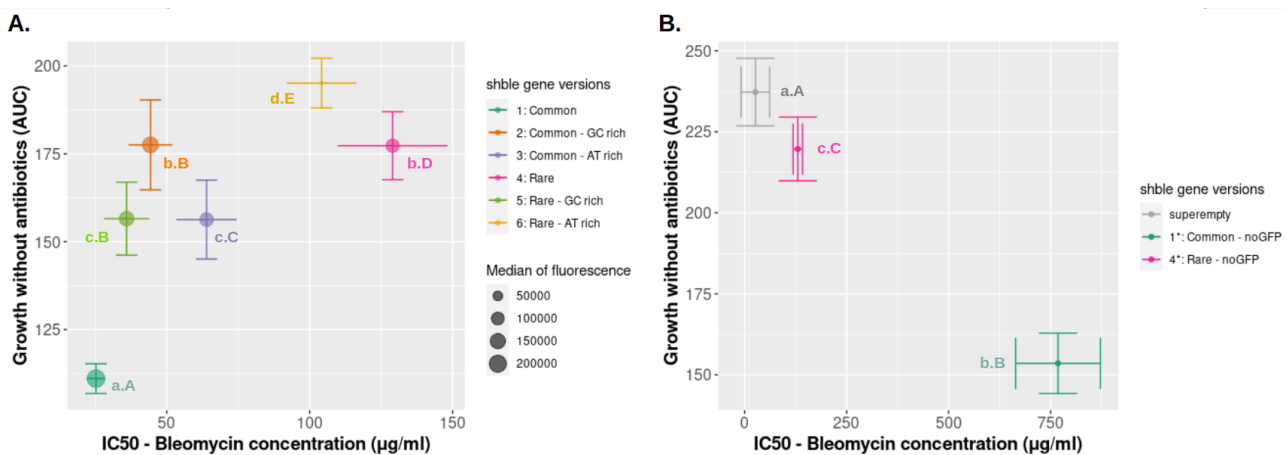


305 conditions, we monitored over time a dimensionless parameter named "Cell Index", that integrates cell density,  
306 adhesion, morphology and viability; and we evaluated the total area below the curve as a proxy for cell growth  
307 (Sup. Method 2.8). We fitted to a Hill's equation the variation of Cell Index values (*i.e.* cell growth) as a function  
308 of the antibiotic concentration, so that we could recover for each condition: (i) the maximum growth in the  
309 absence of antibiotic (Figure 5, variable for the y axis); and (ii) the estimation for the antibiotic concentration  
310 value that inhibited cell growth to half the maximum (IC50; Figure 5, variable for the x axis). Higher values of  
311 the variable "maximum growth in the absence of antibiotics" reflect a lower impact of the heterologous construct  
312 in the cell, while higher values of the IC50 variable reflect a higher resistance potential of the cell to face the  
313 bleomycin antibiotic. The results show that cell populations that grow more in the absence of antibiotics  
314 correspond also to cell populations that resist higher antibiotic concentrations. We interpret that this connection  
315 between growth variables reflects a trade-off between (i) the potential benefit of the antibiotic resistance -  
316 conferred by SHBLE expression, and realised only in the presence of antibiotics-, and (ii) the cost incurred  
317 through heterologous protein overexpression -associated to both SHBLE and EGFP expression and that is  
318 present independently of the presence of the antibiotic. This trade-off results indeed in a non-monotonic  
319 relationship between heterologous protein levels and cell growth: condition shble#6 produces low levels of  
320 heterologous protein and thus allows for the highest growth in the absence of antibiotics, but it does not confer  
321 the highest resistance levels; while conditions shble#1 and shble#2 produce the highest amounts of heterologous  
322 proteins and incur thus in a substantial burden, heavier than the potential benefit of the conferred antibiotic  
323 resistance (Figure 5A, and Sup. Fig 15).

324         We aimed at disentangling the two forces in this trade-off by testing two additional constructs containing  
325 solely versions shble#1 and shble#4 of the *shble* gene, and not linked to the *egfp* reporter (labelled #1\* and #4\*  
326 in Figure 5B). Comparing the growth-related variables for *shble* versions #1 and #4 with or without EGFP, both  
327 versions shble#1\* and shble#4\* displayed a similar increase in maximum growth in the absence of antibiotic  
328 with respect to their EGFP+ relative counterparts (respectively 38% and 24%). However, while the IC50 of  
329 shble#4\* remained similar to shble#4, the antibiotic resistance for version shble#1\* dramatically increased with

330 respect to that of shble#1. Notwithstanding, in the absence of antibiotic shble#4\* still performed better than  
 331 shble#1\*.

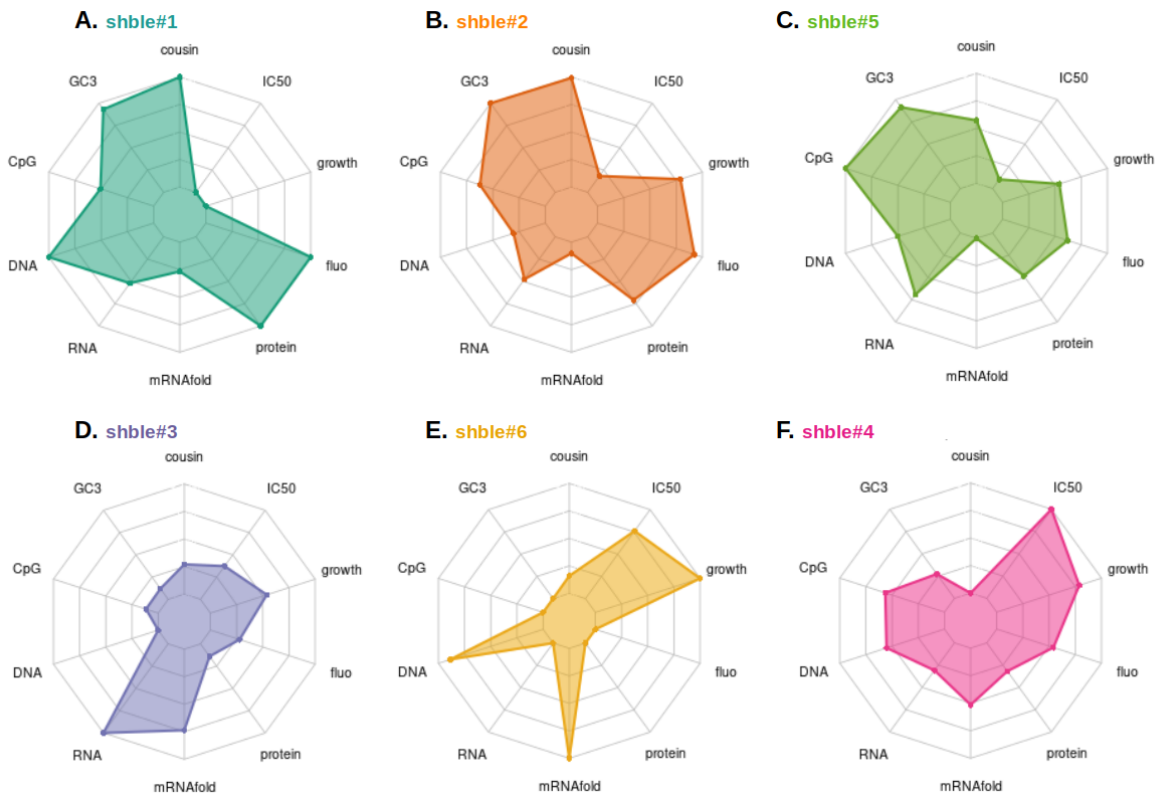
332 Overall, we interpret that (i) in the absence of antibiotic, higher amount of heterologous protein  
 333 (independently of whether they correspond to SHBLE, or to SHBLE and EGFP) had pronounced negative  
 334 impact on cell fitness; and that (ii) in the presence of antibiotic, the optimum between the conferred resistance  
 335 and the cost of protein burden was determined by both, the total amount of heterologous proteins, and the  
 336 abundance of the SHBLE protein, conferring antibiotic resistance.



337 **Figure 5. Variation of cell growth in presence or in absence of antibiotics, for gfp-coupled constructs (A) or**  
 338 **gfp-free constructs (B).** For both panels, the y axis represents maximum cell growth in absence of antibiotics,  
 339 proxied as the area under the curve of the delta Cell Index ("AUC"); and the x axis represents the bleomycin  
 340 concentration that reduces to 50% the corresponding growth ("IC50"). The plotted central values were  
 341 estimated fitting variation of Cell Index data to Hill's equation (pooled data, 3 to 6 biological replicates), and  
 342 bars correspond to the estimated standard error. Statistical tests are Welch modified two-sample t-tests,  
 343 performed for the AUC (small letters, y axis) or the IC50 (big letters, x axis): for each size of letters, conditions  
 344 associated with a same letter do not display different median values of the corresponding variable ( $p < 0.05$  after  
 345 Benjamini-Hochberg correction). As an example for interpretation, orange and pink values are not different in  
 346 the y-axis projection (labelled both with b) but differ on their x-axis projection (labelled respectively with B and  
 347 D). For panel A, the size of the dots is proportional to the corresponding total of fluorescence, which is used as a  
 348 proxy for the level of heterologous proteins. Six different conditions are shown: shble#1 (dark green), shble#2  
 349 (orange), shble#3 (purple), shble#4 (pink), shble#5 (light green), shble#6 (yellow). For panel B, three different  
 350 conditions are shown: superempty control (grey), versions shble#1\* (dark green) and shble#4\* (pink) lacking  
 351 the EGFP reporter gene.

## 352 **DISCUSSION**

353           In the present manuscript we have analysed the multilevel molecular effects of CUBias on gene  
354 expression and have further explored higher-level integration consequences at the cellular level. We have  
355 focused on the effects of CUBias of our focal *shble* gene on its own expression and function, *i.e.* the so-called  
356 *cis*-effects of CUBias. The global *trans*-effects of CUBias of our focal gene on the expression levels of other  
357 cellular genes have been analysed and described in an accompanying manuscript <sup>73</sup>. Our results show that a  
358 combination of synonymous changes results in important multilevel variation in gene expression levels and leads  
359 to dramatic differences in the cellular phenotype. We summarize our observations of these *cis*-effects in Figure 6,  
360 which displays variation in each of the variables that we have monitored, either experimentally or sequence-  
361 dependent. Conditions *shble*#1 and *shble*#2 display a very similar global profile, consistent with the fact that  
362 they are identical in 95.6% of their sequence (Table S1). Further, conditions with CUBias close to human  
363 average one (*shble*#1, *shble*#2 and *shble*#5) cover a similar phenotypic space, very different from the phenotypic  
364 space covered by conditions with CUBias opposite to the human average (*shble*#3, *shble*#4 and *shble*#6).



365 **Figure 6. Summarizing combination of sequence composition parameters and multi-level phenotypes for**  
 366 **each customized version of the *shble* antibiotic resistance gene.** The six versions, designed with the one amino  
 367 acid – one codon strategy, are shown by decreasing similarity to the average genome human CUBias (i.e.  
 368 expressed as their cousin score, (59)): CUBias for *shble*#1 and *shble*#2 are similar to the human CUBias but of  
 369 larger magnitude, #5 CUBias is similar to the human CUBias, and #3, #6, #4 CUBias are opposite to the  
 370 human CUBias. **A. *shble*#1** (most common codons, in dark green), **B. *shble*#2** (common and GC-rich codons, in  
 371 orange), **C. *shble*#5** (rare and GC-rich codons, in light green), **D. *shble*#3** (common and AT-rich codons, in  
 372 purple), **E. *shble*#6** (rare and AT-rich codons, yellow) and **F. *shble*#4** (rarest codons, in pink). The sequence  
 373 characteristics are from the top to the left: "cousin"; "CpG" (the CG dinucleotide frequency in the recoded  
 374 *shble*), "GC3" (the GC content at the third codon position in the recoded *shble*), and "mRNAfold" (the mRNA  
 375 folding energy for the recoded *shble* transcript). The different phenotypic variables, from the bottom to the right  
 376 are: "DNA" (the transfection efficiency, estimated represented by the amount of plasmid after qPCR), "RNA"  
 377 (the amount of SHBLE-coding full mRNA, estimated after rt-qPCR), "protein" (the amount of SHBLE protein,  
 378 estimated by quantitative proteomics), "fluo" (the total fluorescence signal, estimated by flow cytometry),  
 379 "growth" (proxy of the cellular fitness in absence of antibiotics, estimated by real-time cell growth analysis) and  
 380 "IC50" (proxy of the cellular fitness in presence of antibiotics, estimated by real-time cell growth analysis). All  
 381 variables have been independently re-scaled for representation purposes, from lowest (central) to highest  
 382 (periphery) value.

383           **Variation in CUBias modifies alternative splice patterns.** In the heterologous transcripts for the two  
384 versions with the most dissimilar CUBias with respect to the human average (shble#4 and shble#6) we identified  
385 splicing events, located within the *shble* ORF, that were not detected by leading splice site predicting algorithms  
386 <sup>70,71</sup>. Splicing ablated the SHBLE coding potential without modification of the EGFP coding potential. The  
387 spliced transcripts amounted to 20% and 80% of all heterologous transcripts in shble#4 and shble#6 respectively.  
388 Variation in CUBias across intron-exon boundaries has been described in several eukaryotes (*e.g.* human, fishes,  
389 fruit flies, nematodes, plants <sup>11,74,75</sup>); and splicing regulatory motifs that can be disrupted by synonymous  
390 mutations have been described in mammals <sup>9,75-78</sup>. A reduced single nucleotide polymorphism density and a  
391 decreased rate of synonymous substitutions have further been reported in these regulatory regions, which can be  
392 interpreted as a signature for selective pressure <sup>79,80</sup>. Thus, selection against mRNA mis-processing can constitute  
393 an important selective force that results in concomitant selection for a precise local CUBias <sup>81</sup>. This selective  
394 force has even been proposed to outperform translational selection in *Drosophila melanogaster* <sup>82</sup>. It is  
395 interesting to state here that we did not detect any western blot signal in any of our nine biological replicates that  
396 could correspond to the spliced, short SHBLE polypeptides in the shble#4 and shble#6 conditions (see Sup. Figs.  
397 6, 7 and 8). Further, we did not detect in our proteomic analyses any trace of the expected peptides that could  
398 differentiate the spliced SHBLE versions from the full length SHBLE protein. For western-blot detection we  
399 used an AU1 epitope located in the N-terminus of the protein, and that should be present in all SHBLE forms,  
400 spliced or not. Lack of western-blot detection of these N-terminal spliced short SHBLE polypeptides could  
401 simply reflect a technical limitation, as they are barely 54 amino acids-long (or 53 for the minor spliced version  
402 in shble#6). However, in the case of shble#6 the lack of concordance between SHBLE levels and EGFP levels  
403 (see the very low slope in Figure 2A and 2B) suggests rather a genuine very low presence of spliced SHBLE  
404 molecules in our samples. We interpret that our results are rather compatible with a low stability of the spliced,  
405 short SHBLE versions, possibly linked to a faulty folding that could lead to a rapid degradation upon synthesis.  
406 Indeed, these spliced, short SHBLE versions span less than 20 amino acids of the original SHBLE sequence, so

407 that the protein that is expected to fold into the known quaternary structure <sup>83</sup> actually does not exist after  
408 splicing.

409       **Variation in CUBias correlates with differences in mRNA levels.** We observe significant differences  
410 in mRNA levels among the recoded *shble* versions, with *shble#3* showing over five times more full-length  
411 mRNA than *shble#6* (Table S2). Transcript abundance at a given time point is the result of integrating mRNA  
412 synthesis and degradation kinetics. In our experimental setup differential transfection efficiency does not explain  
413 differences in mRNA levels because variation in mRNA levels between conditions was independent of variation  
414 in DNA abundance. We interpret as well that differential ribosomal recruitment is unlikely to explain differences  
415 in mRNA levels, all our synonymous constructs share the same CMV promoter, the 5' untranslated region, and  
416 the AUG context. We interpret therefore that the observed differences in mRNA levels probably result from  
417 differential mRNA stability and decay, rather than from primary transcription regulation. Such an effect has been  
418 described for bacteria (*E. coli* <sup>84</sup>), unicellular eukaryotes (*S. cerevisiae*, *S. pombe* <sup>35</sup>, *N. crassa*, *T. brucei* <sup>85,86</sup>), and  
419 metazoa (fruit fly <sup>87</sup> or zebrafish <sup>88</sup>). In human cells, it has been shown that nucleotide composition and CUBias  
420 have an impact on mRNA stability, so that transcripts with longer half-lives are enriched in GC-rich codons <sup>89</sup>,  
421 possibly through translation-associated decay mechanisms <sup>90</sup>. Nevertheless, in our experimental design  
422 heterologous mRNA levels are not a monotonic function of mRNA composition, as versions *shble#3* and  
423 *shble#6* are the AT-richer ones (respectively 20% and 10% GC3) but display respectively the highest and the  
424 lowest levels of heterologous mRNAs. The effects on version *shble#6* are difficult to address as only 20% of the  
425 total heterologous transcripts contain the customized *shble* sequence. It is thus impossible to disentangle the  
426 effects of sequence composition on the full mRNA level from the consequences of the splicing defect. The very  
427 high transcript levels and very low protein levels for version *shble#3* are interesting in the light of recent  
428 findings on CUBias linked mRNA degradation and/or storage: indeed, AU-rich mRNAs have been found to  
429 locate in P-bodies, potentially leading to accumulation of this transcript in the cell <sup>91</sup>. In addition, the P-body  
430 retention of those transcripts reduce their availability to translation and could further explain the reduced protein  
431 level for this condition (see discussion below).

432 **Variation in CUBias and mRNA structure correlate with differences in translation efficiency.**  
433 Considering all conditions together, our experimental setup allowed us to determine that variation in mRNA  
434 levels explains only around 45% of the variation in protein levels, which fits well previous descriptions in the  
435 literature for a wide diversity of experimental systems <sup>72,92,93</sup>. Such relatively weak explanatory power would not  
436 be expected if all mRNAs were translated at a constant rate, and has thereby motivated studies to elucidate  
437 which factors are involved in the regulation of translation <sup>94</sup>. Indeed, the literature suggests that in general  
438 variations at the mRNA level do not suffice to predict variation at the protein level <sup>95</sup>, and that this lack of  
439 predictive power is worse at the single-cell level than at the cell population level <sup>96</sup>. Here, we provide evidence  
440 that co-variation between mRNA levels and protein levels depends on CUBias of our focal gene. Particularly, the  
441 AT-rich shble#3 version displayed the highest mRNA levels but contrasting low amounts of the corresponding  
442 protein. A possible explanation for this phenomenon could be the selective translation impairment of AT-rich  
443 transcripts. As mentioned above, this can result from P-body retention, which physically sequesters AT-rich  
444 mRNAs in cell granules making them unavailable for translation <sup>91</sup>. Other mechanisms may additionally be  
445 involved in selective translation impairment. For instance, in human cells, the protein Schlafen11 has been  
446 shown to prevent translation of AU-rich transcripts <sup>97,98</sup>. Given that the AT-rich shble#4 version displays only a  
447 moderate translation impairment, we interpret that the dramatic phenotype of shble#3 (high mRNA levels and  
448 low protein levels) arises in fact from the combination of suboptimal variables for which a role in optimizing the  
449 expression of heterologous genes had already been evidenced <sup>11</sup>: (i) similarity to human average CUBias; (ii) the  
450 CpG frequency; and (iii) the mRNA folding energy.

451 (i) gene versions with a better match to the average CUBias result in higher protein-to-mRNA ratios.  
452 This result is in disagreement with previous reports, as well as with descriptions showing the very limited impact  
453 of CUBias on gene expression in mammals, compared to other features <sup>30,99</sup>. Nevertheless, it is complicated to  
454 disentangle the effect of CUBias from other composition characteristics, such as GC and GC3 content. It is even  
455 more difficult to interpret them in terms of neutralist or selectionist origin, as both evolutionary hypotheses could  
456 account for variation in either parameter (10).

457 (ii) Regarding intragenic CpG frequency, we report a negative impact of very low CpG values on  
458 translation efficiency. Such direct effect of low CpG values on translation efficiency had never been reported  
459 before, and CpG frequency had been shown to impact heterologous protein amount through its impact on *de*  
460 *novo* transcription instead <sup>100,101</sup>. More precisely, high CpG depletion was previously associated to low mRNA  
461 levels, that weren't evidenced as a result of changes in nuclear export, alternative splicing or mRNA stability  
462 <sup>100,101</sup>. Indeed, a signature for selection towards decreased values of CpG has been consistently reported <sup>102,103</sup>,  
463 experimentally verified by the detrimental effects of increased CpG levels on protein synthesis <sup>104,105</sup>, and further  
464 corroborated through experimental evolution <sup>106</sup>.

465 (iii) Regarding the total mRNA folding energy, we also report a negative impact on translation of  
466 extreme values. Molecular modelling, along with experimental studies, suggests a prominent role of the  
467 initiation steps, rather than elongation steps, on the translation efficacy <sup>41,107,108</sup>. And indeed, several studies  
468 addressing the impact of mRNA folding on translation, established the importance of the 5' mRNA secondary  
469 structure in translation initiation. A shared trend has been identified in bacteria, yeast, protists, and mammals  
470 <sup>31,55,108-111</sup>: a reduced mRNA stability near the site of translation initiation is correlated to a higher protein  
471 production. In bacteria and yeast, strong folding around the start codon prevents ribosome recruitment <sup>31,108</sup>; and  
472 a "ramp" of rare codons along the 50 to 100 first coding nucleotides has been reported, with the effect of  
473 reducing mRNA folding energy and with the proposed consequences of avoiding ribosome traffic jam <sup>111,112</sup>. A  
474 systematic exploration using 244,000 synthetic DNA sequences on *E. coli* has shown that variation in secondary  
475 mRNA structure stability immediately around the start codon accounts for around 36% of the total variance in  
476 protein synthesis, while variation in downstream mRNA folding energy accounts only for *ca.* 4-5% <sup>113</sup>.  
477 Nonetheless, an important role of translation elongation cannot be ruled out. Particularly, in human transcripts,  
478 de Sousa Abreu and coworkers describe no effect of the initiation rate on translation efficiency <sup>92</sup>. A recent study  
479 in human cell lines, highlights the consequences of the secondary structures along the CDS in the functional half  
480 life of mRNA <sup>110</sup>, which can be related to overall GC and GC3 content as well as to CUBias <sup>89,90</sup>. In our  
481 experimental setup all constructs share by design the nucleotide sequence around the start codon: the 5'UTR  
482 corresponds to the plasmid backbone and the first 24 coding nucleotides are identical (AU1 tag). Thus, there are



483 actually no differences in folding energy when considering only the immediate sequence stretch around the start  
484 codon, but there are instead differences when considering the full mRNA length. We interpret that our  
485 observation of a non-monotonic effect of the full-length mRNA folding energy on the protein-to-mRNA ratio is  
486 related to translation elongation impairment rather than to an effect on translation initiation.

487       **Variation in CUBias modifies intensity and distribution of the fluorescent reporter.** We have  
488 analysed the fluorescence pattern of the cell populations by means of cytometry. We report phenotypic  
489 variability of transfected human cells, observable as multimodal distribution of cellular fluorescence. The  
490 multimodal distribution of cellular fluorescence intensity on the transfected cells could be captured in all cases  
491 by fitting to a combination of two Gaussian curves. This pattern was similar for all constructs expressing *egfp*,  
492 including the empty control and we interpret that it reflects phenotypic plasticity and may be related to transient  
493 cellular states, such as cell division status and/or differential kinetics of recovery from transfection-induced  
494 cellular stress. Similar differences in gene expression have been actually reported when using cytomegalovirus-  
495 based expression vectors <sup>114</sup>, and have been proposed to be related to cell-cycle dependent cellular states. This  
496 bimodal pattern is notwithstanding puzzling, and deserves more attention using a tailored experimental design,  
497 that our setup cannot provide. Beyond the shared bimodal distribution of fluorescence levels, we observe  
498 significant and concerted shifts of both cellular subpopulations towards higher (*e.g.* for the constructs enriched in  
499 the most used codons) or lower (*e.g.* for constructs using AT-rich codons) values of fluorescence intensity. Thus,  
500 differences in overall EGFP-based fluorescence between recoded constructs do not arise from differences in the  
501 number of positive cells expressing a given quantity of EGFP, but rather from differences in EGFP synthesis at  
502 the individual cell level. Our experimental model using human cells shows that CUBias exerts an important  
503 effect on the overall levels but also in the cell-based levels of the heterologous protein produced.

504       **Heterologous gene expression leads to a trade-off between the benefit conferred through antibiotic**  
505 **resistance and the burden imposed by extra protein synthesis.** Strong heterologous expression imposes an  
506 enormous basal burden on the cellular economy <sup>115</sup>. This impact on cellular economy is consistent with the broad  
507 literature about the direct (*cis*) and indirect (*trans*) costs of translation <sup>81</sup>: first, because translation is the per-unit  
508 most expensive step during biological information flow <sup>116</sup>, consuming *ca.* 45% of the whole energy supply in

509 human cells in culture <sup>117</sup>; second, because virtually all ribosomes are bound to mRNA molecules and potentially  
510 engaged in translation <sup>117</sup>, so that highly-transcribed heterologous mRNA increase overall ribosome demand and  
511 cause loss of opportunity for cellular gene translation; and third, because heterologous protein synthesis can lead  
512 to additional downstream costs by protein folding, protein degradation and off-target effects of mistranslated  
513 proteins <sup>45,118–120</sup>. Additionally, the mismatch between the CUBias of the heterologous gene and of the expression  
514 machinery can display strong trans-effects on the cellular homeostasis, by sequestering ribosomes onto mRNAs  
515 that hardly progress over translation but also by creating a competition for the tRNA pools <sup>31,121</sup>. Scarcity of the  
516 less common tRNAs is actually a severe limiting factor for protein synthesis in bacteria <sup>122</sup>, and this pressure  
517 over rare tRNAs can become extreme in conditions of stress, or changes in nutritional status <sup>10,123,124</sup>.

518       The *shble* gene that we have used as a base for synonymous recoding encodes for a small protein that  
519 confers resistance to bleomycin through antibiotic sequestering <sup>125</sup>. This protein-antibiotic binding is  
520 equimolecular and reversible: an SHBLE protein dimer binds two bleomycin molecules <sup>83</sup>. The strength of the  
521 antibiotic resistance conferred is probably a monotonic, function of the SHBLE amount produced. However, our  
522 results suggest that the benefit conferred by SHBLE synthesis in the presence of antibiotic is largely exceeded by  
523 the cost and burden of heterologous protein synthesis. We show an important trade-off between the intensity of  
524 heterologous SHBLE+EGFP protein synthesis and the actual bleomycin resistance levels, consistent with the  
525 strong burden on cell economy imposed by heterologous gene overexpression. This cost can be partly levered  
526 when ablating EGFP synthesis in the heterologous constructs, so that a substantial fraction of the burden is  
527 removed. Overall we conclude that CUBias of the heterologous gene conferring antibiotic resistance  
528 differentially impacts cellular fitness as a function of the differences in heterologous protein synthesis.

## 529 CONCLUSION

530           The main conundrum for scientists approaching CUBias remains the contrast between, on the one hand,  
531 the large and sound body of knowledge showing the strong molecular and cellular impact of gene expression  
532 differences arising from CUBias and, on the other hand, the thin evidence for codon usage selection at the  
533 organismal level. Under the neutral hypothesis, differences in average genome CUBias can be explained by  
534 biochemical biases during DNA synthesis or repair (*e.g.* polymerase bias)<sup>126</sup>; and, in vertebrates, CUBias at the  
535 gene level may be shaped by their relative position to isochores (*e.g.* alternation between GC-rich and AT-rich  
536 stretches along the chromosomes)<sup>127</sup>. In vertebrates, GC-biased gene conversion mechanisms enhance further  
537 such local variations<sup>126,128,129</sup>. The selective explanation, often referred to as "translational selection", proposes  
538 that different codons may lead to differences in gene expression, by changes in alternative splicing patterns,  
539 mRNA localisation or stability, translation efficiency, or protein folding<sup>130</sup>. If such CUBias-induced variation in  
540 gene expression were associated with phenotypic variation that results in fitness differences, it would, by  
541 definition, be subject to natural selection. Nevertheless, differences in fitness associated with individual  
542 synonymous changes seem to be mostly of low magnitude, so that selection may only act effectively in  
543 organisms with large population sizes<sup>131</sup> such as bacteria<sup>7</sup>, yeast<sup>132</sup>, nematodes<sup>133</sup>, but also in fruit flies<sup>19,20,134,135</sup>,  
544 branchiopods<sup>136</sup> and amphibians<sup>137</sup>. In organisms with small population sizes, such as mammals, and  
545 particularly humans, evidences of selection for (or against) certain codons remain nevertheless controversial  
546<sup>22,138</sup>. In the present manuscript, we have intended to contribute to this debate by exploring the multilevel  
547 phenotypic consequences of codon usage differences of heterologous genes in human cells. Our results are  
548 consistent with a scenario in which the potential evolutionary forces at play in shaping human CUBias, select for  
549 a strict control of mRNA processing (*e.g.* splicing, and secondary structure, potentially affecting stability and  
550 decay), and that the resulting mRNA properties *in fine* impact translation elongation. Notwithstanding, the  
551 disparity between predictions and findings encountered in powerful, codon-usage related experimental evolution  
552 approaches highlights the gap in our understanding at connecting phenotype and fitness over different integration  
553 levels: molecules, cells, tissues and organisms. Despite, or thanks to, the immense body of knowledge

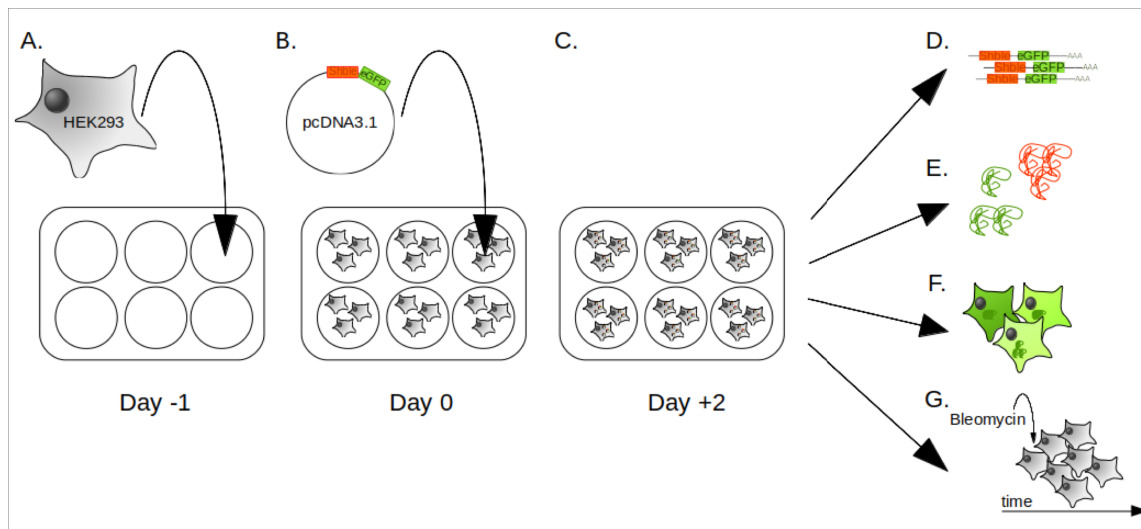
554 accumulated over the last fifty years, the quest for interpreting and integrating the riddle of CUBias over broad  
555 scales of time and biological complexity remains tempting and unsolved.

## 556 MATERIAL AND METHODS

557       **Design of the *shble* synonymous versions and plasmid constructs.** In the present work we have used  
558 as focal gene the bleomycin resistance gene present in the genome of the actinobacterium *Streptoalloteichus*  
559 *hindustanus* (ATCC 31158, GenBank X52869.1). We have chosen to focus on this *shble* gene for a number of  
560 reasons: 1) because of the mechanism of action of the antibiotic: bleomycin is cytotoxic by intercalating and  
561 introducing breaks in the dsDNA<sup>66</sup>. In this experimental setup we are interested in mRNA translation process,  
562 and many antibiotics interfere with protein synthesis at different levels, so that we chose a focal gene with no  
563 impact on the mechanisms that we will be evaluating. 2) because of the mechanism that confers resistance: the  
564 SHBLE protein interacts on an equimolecular fashion with the bleomycin antibiotic, so that the SHBLE  
565 homodimer binds and sequesters two bleomycin molecules<sup>125</sup>, without performing any catabolic activity on  
566 them nor on any other cellular metabolite. The antibiotic resistance level conferred is thus expected to be a  
567 direct, monotonic function of the total amount of SHBLE protein produced. 3) because of the small size of the  
568 protein synthesised: the SHBLE protein is barely 124 amino acids long, thus minimising the total length of the  
569 heterologous mRNA and the impact of translation on the host cell. We did not consider the use of the wild type  
570 *shble* sequence in *S. hindustanus* as a meaningful control in our experimental setting using mammalian cells in  
571 culture, and we have thus focused on recoding strategies to maximise synonymous differences with respect to  
572 our human cell expression system. Six synonymous versions of the *shble* gene were designed applying the "one  
573 amino acid--one codon" approach, *i.e.*, all instances of one amino acid in the *shble* sequence were recoded with  
574 the same codon, as follows (Table 1): shble#1 used the most frequent codons in the human genome; shble#2  
575 used the GC-richest among the two most frequent codons; shble#3 used the AT-richest among the two most  
576 frequent codons; shble#4 used the least frequent codons; shble#5 used the GC-richest among the two less  
577 frequent codons; and shble#6 used the AT-richest among the two less frequent codons. An invariable *AU1*  
578 sequence was added as N-terminal tag (amino acid sequence MDTYRI) to all six versions (Sup. Fig. 1).

579 Nucleotide content between versions are compared in Sup. Table 1. Our recoding strategy succeeds at  
580 maximising differences in nucleotide content and to explore a large sequence space in terms of total GC, GC3  
581 and transcript folding energy (Table 1; Sup. Fig. 2). The normalized COUSIN 18 score (COdon Usage Similarity  
582 Index), which compares the CUBias of a query against a reference, was calculated using the online tool  
583 (<http://cousin.ird.fr>)<sup>68</sup>. A score value below 0 informs that the CUBias of the query sequence is opposite to the  
584 reference CUBias; a value close to 1 informs that the query CUBias is similar to the reference CUBias, and a  
585 value above 1 informs that the query CUBias is similar the reference CUBias, but of larger magnitude<sup>68</sup>. All  
586 *shble* synonymous sequences were chemically synthesised and cloned on the *XhoI* restriction site in the  
587 pcDNA3.1+P2A-EGFP plasmid (InvitroGen), in-frame with the *P2A-EGFP* reporter cassette. In this plasmid,  
588 the expression of the reporter gene is located under the control of the strong human cytomegalovirus (CMV)  
589 promoter and terminated by the bovine growth hormone polyadenylation signal. All constructs encode for a  
590 1,602 bp transcript, encompassing a 1,182 bp *au1-shble-P2A-EGFP* coding sequence (Sup. Fig. 1). The folding  
591 energies of all 1,602 bp transcripts were calculated using the RNAfold Webserver ([http://rna.tbi.univie.ac.at/cgi-](http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi)  
592 [bin/RNAWebSuite/RNAfold.cgi](http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi))<sup>69</sup>, with default parameters (Table 1). During translation, the P2A peptide  
593 (sequence NPGP) induces ribosome skipping<sup>67</sup>, meaning that the ribosome does not perform the Gly-Pro  
594 transpeptidation bond and releases instead the AU1-SHBLE moiety and continues translation of the EGFP  
595 moiety. The HEK293 human cell line used here is proficient at performing ribosome skipping on the P2A  
596 peptide<sup>139</sup> The transcript encodes thus for one single coding sequence but translation results in the production of  
597 two proteins: SHBLE (theoretical molecular mass 17.2 kDa) and EGFP (27.0 kDa). As controls we used two  
598 plasmids: (i) pcDNA3.1+P2A-EGFP (named here "empty"), which encodes for the EGFP protein; (ii)  
599 pcDNA3.1+ (named here "superempty") which does not express any transcript from the CMV promoter (Table  
600 1). In order to explore the burden of EGFP expression we generated two additional constructs by subcloning the  
601 AU1-tagged *shble#1* and *shble#4* coding sequences in the *XhoI* restriction site of the pcDNA3.1+ backbone,  
602 resulting in the constructs *shble#1\** and *shble#4\**, lacking the *P2A-EGFP* sequence.

603 **Transfection and differential cell sampling.** All experiments were carried out on HEK293 cells  
604 (ACCT CRL-1573). Cell culture conditions, transfection methods and related reagents are detailed in Sup.  
605 Methods 2.2. Cells were harvested two days after transfection and submitted to analyses at four levels (Figure 7):  
606 (i) nucleic acid analyses (qPCR and RNAseq); (ii) proteomics (label-free quantitative mass spectrometry  
607 analysis and western blot immuno-assays); (iii) flow cytometry; and (iv) real-time cell growth analysis (RTCA).  
608 Overall, the different experiments were performed on 33 biological replicates, corresponding to a variable  
609 number of repetitions depending on the considered analysis (Sup. Method 1). Transfection efficiency was  
610 evaluated by means of qPCR targeting two invariable regions of the plasmid and revealed no significant  
611 differences between the constructs (Sup. Methods 2.3).



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

618       **RNA sequencing and data analysis.** Transcriptomic analysis was performed on six biological replicates  
619 and eight conditions: shble#1 to shble#6, #empty, and mock (for which the sample is submitted to the exact same  
620 procedures, including the transfection agent, but in absence of plasmid). Paired 150bp Illumina reads were  
621 trimmed (Trimmomatic v0.38) <sup>140</sup> and mapped on eight different genomic references (HISAT2 v2.1.0) <sup>141</sup>,  
622 corresponding to the concatenation of the human reference genome  
623 (GCF\_000001405.38\_GRCh38.p12\_genomic.fna, NCBI database, 7<sup>th</sup> of February 2019) and the corresponding  
624 full sequence of the plasmid. For the mock condition, we considered the human genome and all possible versions  
625 of the plasmid. Virtually no read of those negative controls mapped to the plasmid sequences. For all other  
626 conditions, read distribution patterns along the plasmid sequence were evaluated with IGVtool <sup>142</sup>. In all cases  
627 the *au1-shble-p2a-EGFP* coding sequence displayed highly similar coverage shape for all constructs, except for  
628 shble#4 and shble#6 for which respectively one and two alternative splicing events were observed (Sup. Fig. 3  
629 and 4). None of these splice sites were predicted when the theoretical transcripts were evaluated using *Human*  
630 *Splicing Finder* (HSF, accessed via <https://www.genomnis.com/access-hsf>) <sup>70</sup>, or with *SPLM - Search for human*  
631 *potential splice sites using weight matrices* (accessed via <http://www.softberry.com/>) <sup>71</sup>. When relevant, the three  
632 alternative transcript isoforms identified were further used as reference for read pseudomapping and  
633 quantification with Kallisto (v0.43.1) <sup>143</sup>. Details on RNA preparation and bioinformatic pipeline are provided in  
634 Sup. Methods 2.4 and Sup. Methods 3.

635       **Label-free proteomic analysis.** Label-free proteomic was performed on nine biological replicates (three  
636 of them measured independently, and six pooled by two), and eight different conditions: shble#1 to shble#6,  
637 #empty, and mock. For each sample, 20 to 30 µg of proteins were digested in-gel and the resulting peptides were  
638 analysed online using a Q Exactive HF mass spectrometer coupled with an Ultimate 3000 RSLC system  
639 (Thermo Fisher Scientific). MS/MS analyses were performed using the Maxquant software (v1.5.5.1) <sup>144</sup>. All  
640 MS/MS spectra were searched by the Andromeda search engine <sup>145</sup> against a decoy database consisting in a  
641 combination of *Homo sapiens* entries from Reference Proteome (UP000005640, release 2019\_02,  
642 <https://www.uniprot.org/>), a database with classical contaminants, and the sequences of interest (SHBLE and  
643 EGFP). After excluding the usual contaminants, we obtained a final set of 4,302 proteins detected at least once

644 in one of the samples. Intensity based absolute quantification (iBAQ) values were used to compare protein levels  
645 between samples <sup>146</sup>.

646 **Western blot immunoassays and semi-quantitative analysis.** Western blot immunoassays were  
647 performed on nine replicates and nine conditions: shble#1 to shble#6, #empty, #superempty, and mock. Three  
648 different proteins were targeted:  $\beta$ -TUBULIN, EGFP, and SHBLE (via the invariable AU1 epitope tag).  
649 Analysis from enzyme chemoluminescence data was performed with ImageJ <sup>147</sup> by «plotting lanes» to obtain  
650 relative density plots (Sup. Fig. 7).

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

659 **Real time cell growth analysis (RTCA).** RTCA was carried out on an xCELLigence system for the  
660 mock and the superempty controls, and further eight constructs: the previously analysed shble#1 to shble#6, plus  
661 the shble#1\* and shble#4\* lacking the *EGFP* reporter gene. Cells were grown under different concentrations of  
662 the Bleomycin antibiotic ranging from 0 to 5000  $\mu$ g/mL (Sup. Method 2.8). Three to six biological replicates  
663 were performed, including technical duplicates for each replicate. Cells were grown on microtiter plates with  
664 interdigitated gold electrodes that allow to estimate cell density by means of impedance measurement. Measures  
665 were acquired every 15 minutes, over 70 hours (280 time points). Impedance measurements are reported as "Cell  
666 Index" values, which are compared to the initial baseline values to estimate changes in cellular performance  
667 linked to the expression of the different constructs (Sup. Figure 16) . For each construct we estimated first  
668 cellular fitness by calculating the area below the curve for the delta-Cell index vs time for the cells grown in the



669 absence of antibiotics. We estimated then the ability to resist the antibiotic conferred by each construct through  
670 calculation of IC50 as the bleomycin concentration that reduces the area below the curve to half of the one  
671 estimated in the absence of antibiotics (detailed methods in Sup. Method 2.8).

672 **Data availability.** RNAseq raw reads were deposited on the NCBI-SRA database under the BioProject  
673 number PRJNA753061. The mass spectrometry proteomics data have been deposited to the ProteomeXchange  
674 Consortium via the PRIDE <sup>148</sup> partner repository with the dataset identifier PXD038324.

## 675 ACKNOWLEDGEMENTS

676 This study was supported by the European Union's Horizon 2020 research and innovation program under the  
677 grant agreement CODOVIREVOL (ERC-2014-CoG-647916) to I.G.B. We acknowledge the IRD itrop HPC  
678 (South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research  
679 results reported within this paper. We also acknowledge the facilities of the Functional Proteomics Platform  
680 (FPP) of the Proteomics Pole of Montpellier (PPM, Montpellier France); and the MRI imaging facility, member  
681 of the France-BioImaging national infrastructure supported by the French National Research Agency (ANR-10-  
682 INBS-04, «Investments for the future»).

## 683 CONFLICT OF INTEREST

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

## 685 REFERENCES

- 686 1. Crick, F. Central dogma of molecular biology. *Nat.* 1970 2275258 **227**, 561–563 (1970).
- 687 2. Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pavé, A. Codon catalog usage and the genome  
688 hypothesis. *Nucleic Acids Res.* **8**, 197–197 (1980).
- 689 3. Ikemura, T. Correlation between the abundance of yeast transfer RNAs and the occurrence of the  
690 respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and  
691 *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. *J. Mol. Biol.* **158**, 573–  
692 597 (1982).

- 693 4. Kanaya, S., Yamada, Y., Kudo, Y. & Ikemura, T. Studies of codon usage and tRNA genes of 18 unicellular  
694 organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific  
695 diversity of codon usage based on multivariate analysis. *Gene* **238**, 143–155 (1999).
- 696 5. Novoa, E. M., Jungreis, I., Jaillon, O., Kellis, M. & Leitner, T. Elucidation of codon usage signatures  
697 across the domains of life. *Mol. Biol. Evol.* **36**, 2328–2339 (2019).
- 698 6. Gouy, M. & Gautier, C. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.*  
699 **10**, 7055–7074 (1982).
- 700 7. Sharp, P. M. & Li, W. H. An evolutionary perspective on synonymous codon usage in unicellular  
701 organisms. *J. Mol. Evol.* **24**, 28–38 (1986).
- 702 8. Duret, L. Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* **12**, 640–649  
703 (2002).
- 704 9. Chamary, J. V., Parmley, J. L. & Hurst, L. D. Hearing silence: non-neutral evolution at synonymous sites  
705 in mammals. *Nat. Rev. Genet.* **7**, 98–108 (2006).
- 706 10. Hanson, G. & Collier, J. Codon optimality, bias and usage in translation and mRNA decay. *Nat. Rev. Mol.*  
707 *Cell Biol.* **19**, 20–30 (2017).
- 708 11. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat.*  
709 *Rev. Genet.* **12**, 32–42 (2010).
- 710 12. Mauro, V. P. & Chappell, S. A. A critical analysis of codon optimization in human therapeutics. *Trends*  
711 *Mol. Med.* **20**, 604–613 (2014).
- 712 13. Angov, E., Hillier, C. J., Kincaid, R. L. & Lyon, J. A. Heterologous protein expression is enhanced by  
713 harmonizing the codon usage frequencies of the target gene with those of the expression host. *PLoS One*  
714 **3**, e2189 (2008).
- 715 14. Fath, S. *et al.* Multiparameter RNA and codon optimization: a standardized tool to assess and enhance  
716 autologous mammalian gene expression. *PLoS One* **6**, e17596 (2011).
- 717 15. Martínez, M. A., Jordan-Paiz, A., Franco, S. & Nevot, M. Synonymous virus genome recoding as a tool to  
718 impact viral fitness. *Trends Microbiol.* **24**, 134–147 (2016).
- 719 16. Ikemura, T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of  
720 the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for  
721 the *E. coli* translational system. *J. Mol. Biol.* **151**, 389–409 (1981).
- 722 17. Dong, H., Nilsson, L. & Kurland, C. G. Co-variation of tRNA abundance and codon usage in *Escherichia*  
723 *coli* at different growth rates. *J. Mol. Biol.* **260**, 649–663 (1996).
- 724 18. Duret, L. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal  
725 translation of highly expressed genes. *Trends Genet.* **16**, 287–289 (2000).
- 726 19. Moriyama, E. N. & Powell, J. R. Codon usage bias and tRNA abundance in *Drosophila*. *J. Mol. Evol.* **45**,  
727 514–523 (1997).
- 728 20. Akashi, H. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational  
729 accuracy. *Genetics* **136**, 927–935 (1994).
- 730 21. Powell, J. R. & Moriyama, E. N. Evolution of codon usage bias in *Drosophila*. *Proc. Natl. Acad. Sci.* **94**,  
731 7784–7790 (1997).

- 732 22. Urrutia, A. O. & Hurst, L. D. Codon usage bias covaries with expression breadth and the rate of  
733 synonymous evolution in humans, but this is not evidence for selection. *Genetics* **159**, 1191–1199 (2001).
- 734 23. Lithwick, G. & Margalit, H. Hierarchy of sequence-dependent features associated with prokaryotic  
735 translation. *Genome Res.* **13**, 2665–2673 (2003).
- 736 24. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nat.* 2003 4256959 **425**, 737–741  
737 (2003).
- 738 25. Tuller, T., Kupiec, M. & Ruppin, E. Determinants of protein abundance and translation efficiency in *S.*  
739 *cerevisiae*. *PLOS Comput. Biol.* **3**, e248 (2007).
- 740 26. Burgess-Brown, N. A. *et al.* Codon optimization can improve expression of human genes in *Escherichia*  
741 *coli*: a multi-gene study. *Protein Expr. Purif.* **59**, 94–102 (2008).
- 742 27. Lampson, B. L. *et al.* Rare Codons regulate KRAS oncogenesis. *Curr. Biol.* **23**, 70–75 (2013).
- 743 28. Pop, C. *et al.* Causal signals between codon bias, mRNA structure, and the efficiency of translation and  
744 elongation. *Mol. Syst. Biol.* **10**, 770 (2014).
- 745 29. Li, G. W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein synthesis rates reveals  
746 principles underlying allocation of cellular resources. *Cell* **157**, 624–635 (2014).
- 747 30. Vogel, C. *et al.* Sequence signatures and mRNA concentration can explain two-thirds of protein abundance  
748 variation in a human cell line. *Mol. Syst. Biol.* **6**, 400 (2010).
- 749 31. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of expression in  
750 *Escherichia coli*. *Science*. **324**, 255–258 (2009).
- 751 32. Agashe, D., Martinez-Gomez, N. C., Drummond, D. A. & Marx, C. J. Good codons, bad transcript: large  
752 reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. *Mol. Biol.*  
753 *Evol.* **30**, 549–560 (2013).
- 754 33. Zucchelli, E. *et al.* Codon optimization leads to functional impairment of RD114-TR envelope  
755 glycoprotein. *Mol. Ther. - Methods Clin. Dev.* **4**, 102–114 (2017).
- 756 34. Presnyak, V. *et al.* Codon optimality is a major determinant of mRNA stability. *Cell* **160**, 1111–1124  
757 (2015).
- 758 35. Harigaya, Y. & Parker, R. Analysis of the association between codon optimality and mRNA stability in  
759 *Schizosaccharomyces pombe*. *BMC Genomics* **17**, 1–16 (2016).
- 760 36. Radhakrishnan, A. & Green, R. Connections underlying translation and mRNA stability. *J. Mol. Biol.* **428**,  
761 3558–3564 (2016).
- 762 37. Radhakrishnan, A. *et al.* The DEAD-box protein Dhh1p couples mRNA decay and translation by  
763 monitoring codon optimality. *Cell* **167**, 122–132.e9 (2016).
- 764 38. Chen, S. *et al.* Codon-resolution analysis reveals a direct and context-dependent impact of individual  
765 synonymous mutations on mRNA level. *Mol. Biol. Evol.* **34**, 2944–2958 (2017).
- 766 39. Bettany, A. J. E. *et al.* 5'-Secondary structure formation, in contrast to a short string of non-preferred  
767 codons, inhibits the translation of the pyruvate kinase mRNA in yeast. *Yeast* **5**, 187–198 (1989).
- 768 40. De Smit, M. H. & Van Duin, J. Secondary structure of the ribosome binding site determines translational  
769 efficiency: a quantitative analysis. *Proc. Natl. Acad. Sci.* **87**, 7668–7672 (1990).

- 770 41. Gu, W., Zhou, T. & Wilke, C. O. A universal trend of reduced mRNA stability near the translation-  
771 initiation site in prokaryotes and eukaryotes. *PLOS Comput. Biol.* **6**, e1000664 (2010).
- 772 42. Tuller, T., Waldman, Y. Y., Kupiec, M. & Ruppin, E. Translation efficiency is determined by both codon  
773 bias and folding energy. *Proc. Natl. Acad. Sci.* **107**, 3645–3650 (2010).
- 774 43. Marais, G. & Duret, L. Synonymous codon usage, accuracy of translation, and gene length in  
775 *Caenorhabditis elegans*. *J. Mol. Evol.* **52**, 275–280 (2001).
- 776 44. Kurland, C. G. Translational accuracy and the fitness of bacteria. *Annu. Rev. Genet.* **26**, 29–50 (2003).
- 777 45. Stoletzki, N. & Eyre-Walker, A. Synonymous codon usage in *Escherichia coli*: selection for translational  
778 accuracy. *Mol. Biol. Evol.* **24**, 374–381 (2007).
- 779 46. Johnston, T. C., Borgia, P. T. & Parker, J. Codon specificity of starvation induced misreading. *Mol. Gen.*  
780 *Genet. MGG 1984 1953* **195**, 459–465 (1984).
- 781 47. Johnston, T. C. & Parker, J. Streptomycin-induced, third-position misreading of the genetic code. *J. Mol.*  
782 *Biol.* **181**, 313–315 (1985).
- 783 48. Robinson, M. *et al.* Codon usage can affect efficiency of translation of genes in *Escherichia coli*. *Nucleic*  
784 *Acids Res.* **12**, 6663–6671 (1984).
- 785 49. Sørensen, M. A., Kurland, C. G. & Pedersen, S. Codon usage determines translation rate in *Escherichia*  
786 *coli*. *J. Mol. Biol.* **207**, 365–377 (1989).
- 787 50. Xia, X. A major controversy in codon-anticodon adaptation resolved by a new codon usage index.  
788 *Genetics* **199**, 573–579 (2014).
- 789 51. Sørensen, M. A. & Pedersen, S. Absolute *in vivo* translation rates of individual codons in *Escherichia coli*:  
790 the two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *J. Mol.*  
791 *Biol.* **222**, 265–280 (1991).
- 792 52. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis *in vivo* of  
793 translation with nucleotide resolution using ribosome profiling. *Science.* **324**, 218–223 (2009).
- 794 53. Hussmann, J. A., Patchett, S., Johnson, A., Sawyer, S. & Press, W. H. Understanding biases in ribosome  
795 profiling experiments reveals signatures of translation dynamics in yeast. *PLOS Genet.* **11**, e1005732  
796 (2015).
- 797 54. Gardin, J. *et al.* Measurement of average decoding rates of the 61 sense codons *in vivo*. *Elife* **3**, (2014).
- 798 55. Weinberg, D. E. *et al.* Improved ribosome-footprint and mRNA measurements provide insights into  
799 dynamics and regulation of yeast translation. *Cell Rep.* **14**, 1787–1799 (2016).
- 800 56. Pechmann, S. & Frydman, J. Evolutionary conservation of codon optimality reveals hidden signatures of  
801 cotranslational folding. *Nat. Struct. Mol. Biol.* **20**, 237–243 (2012).
- 802 57. Chaney, J. L. *et al.* Widespread position-specific conservation of synonymous rare codons within coding  
803 sequences. *PLOS Comput. Biol.* **13**, e1005531 (2017).
- 804 58. Zhao, F., Yu, C. H. & Liu, Y. Codon usage regulates protein structure and function by affecting translation  
805 elongation speed in *Drosophila* cells. *Nucleic Acids Res.* **45**, 8484–8492 (2017).
- 806 59. Rocha, E. P. C. Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient  
807 decoding for translation optimization. *Genome Res.* **14**, 2279 (2004).

- 808 60. Baca, A. M. & Hol, W. G. J. Overcoming codon bias: A method for high-level overexpression of  
809 *Plasmodium* and other AT-rich parasite genes in *Escherichia coli*. *Int. J. Parasitol.* **30**, 113–118 (2000).
- 810 61. Chan, P. P. & Lowe, T. M. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in  
811 complete and draft genomes. *Nucleic Acids Res.* **44**, D184–D189 (2016).
- 812 62. Holley, R. W. *et al.* Structure of a ribonucleic acid. *Science* **147**, 1462–1465 (1965).
- 813 63. Crick, F. H. C. Codon--anticodon pairing: the wobble hypothesis. *J. Mol. Biol.* **19**, 548–555 (1966).
- 814 64. Novoa, E. M., Pavon-Eternod, M., Pan, T. & Ribas De Pouplana, L. A role for tRNA modifications in  
815 genome structure and codon usage. *Cell* **149**, 202–213 (2012).
- 816 65. Rafels-Ybern, À. *et al.* The expansion of inosine at the wobble position of tRNAs, and its role in the  
817 evolution of proteomes. *Mol. Biol. Evol.* **36**, 650–662 (2019).
- 818 66. Suzuki, H., Nagai, K., Yamaki, H., Tanaka, N. & Umezawa, H. On the mechanism of action of  
819 Bleomycin : scission of DNA strands in vitro and *in vivo*. *J. Antibiot. (Tokyo)*. **22**, 446–448 (1969).
- 820 67. Ryan, M. D., King, A. M. Q. & Thomas, G. P. Cleavage of foot-and-mouth disease virus polyprotein is  
821 mediated by residues located within a 19 amino acid sequence. *J. Gen. Virol.* **72**, 2727–2732 (1991).
- 822 68. Bourret, J., Alizon, S. & Bravo, I. G. COUSIN (COdon Usage Similarity INdex): a normalized measure of  
823 codon usage preferences. *Genome Biol. Evol.* **11**, 3523–3528 (2019).
- 824 69. Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R. & Hofacker, I. L. The Vienna RNA Websuite.  
825 *Nucleic Acids Res.* **36**, W70–W74 (2008).
- 826 70. Desmet, F. O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals.  
827 *Nucleic Acids Res.* **37**, e67 (2009).
- 828 71. Solovyev, V. Statistical approaches in eukaryotic gene prediction. *Handb. Stat. Genet.* (2004).
- 829 72. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and  
830 transcriptomic analyses. *Nat. Rev. Genet.* **2012 134 13**, 227–232 (2012).
- 831 73. Jallet, A. J. *et al.* Human cellular homeostasis buffers trans-acting translational effects of heterologous  
832 gene expression with very different codon usage bias. *BioRxiv* 2021.12.09.471957 (2021).
- 833 74. Willie, E. & Majewski, J. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. *Trends*  
834 *Genet.* **20**, 534–538 (2004).
- 835 75. Eskesen, S. T., Eskesen, F. N. & Ruvinsky, A. Natural selection affects frequencies of AG and GT  
836 dinucleotides at the 5' and 3' ends of exons. *Genetics* **167**, 543–550 (2004).
- 837 76. Parmley, J. L. & Hurst, L. D. Exonic splicing regulatory elements skew synonymous codon usage near  
838 intron-exon boundaries in mammals. *Mol. Biol. Evol.* **24**, 1600–1603 (2007).
- 839 77. Fairbrother, W. G., Yeh, R. F., Sharp, P. A. & Burge, C. B. Predictive identification of exonic splicing  
840 enhancers in human genes. *Science*. **297**, 1007–1013 (2002).
- 841 78. Louie, E., Ott, J. & Majewski, J. Nucleotide frequency variation across human genes. *Genome Res.* **13**,  
842 2594–2601 (2003).
- 843 79. Chamary, J. V. & Hurst, L. D. Biased codon usage near intron-exon junctions: selection on splicing  
844 enhancers, splice-site recognition or something else? *Trends Genet.* **21**, 256–259 (2005).
- 845 80. Orban, T. & Olah, E. Purifying selection on silent sites – a constraint from splicing regulation? *Trends*  
846 *Genet.* **17**, 252–253 (2001).

- 847 81. Callens, M., Pradier, L., Finnegan, M., Rose, C. & Bedhomme, S. Read between the lines: diversity of  
848 nontranslational selection pressures on local codon usage. *Genome Biol. Evol.* **13**, (2021).
- 849 82. Warnecke, T. & Hurst, L. D. Evidence for a trade-off between translational efficiency and splicing  
850 regulation in determining synonymous codon usage in *Drosophila melanogaster*. *Mol. Biol. Evol.* **24**,  
851 2755–2762 (2007).
- 852 83. Dumas, P., Bergdoll, M., Cagnon, C. & Masson, J. M. Crystal structure and site-directed mutagenesis of a  
853 bleomycin resistance protein and their significance for drug sequestering. *EMBO J.* **13**, 2483 (1994).
- 854 84. Boël, G. *et al.* Codon influence on protein expression in *E. coli* correlates with mRNA levels. *Nat. 2016*  
855 5297586 **529**, 358–363 (2016).
- 856 85. Jeacock, L., Faria, J. & Horn, D. Codon usage bias controls mRNA and protein abundance in  
857 trypanosomatids. *Elife* **7**, (2018).
- 858 86. Nascimento, J. de F., Kelly, S., Sunter, J. & Carrington, M. Codon choice directs constitutive mRNA  
859 levels in trypanosomes. *Elife* **7**, (2018).
- 860 87. Burow, D. A. *et al.* Attenuated codon optimality contributes to neural-specific mRNA decay in  
861 *Drosophila*. *Cell Rep.* **24**, 1704–1712 (2018).
- 862 88. Mishima, Y. & Tomari, Y. Codon Usage and 3' UTR Length determine maternal mRNA stability in  
863 zebrafish. *Mol. Cell* **61**, 874–885 (2016).
- 864 89. Hia, F. *et al.* Codon bias confers stability to human mRNA s. *EMBO Rep.* **20**, (2019).
- 865 90. Wu, Q. *et al.* Translation affects mRNA stability in a codon-dependent manner in human cells. *Elife* **8**,  
866 (2019).
- 867 91. Courel, M. *et al.* Gc content shapes mRNA storage and decay in human cells. *Elife* **8**, (2019).
- 868 92. De Sousa Abreu, R., Penalva, L. O., Marcotte, E. M. & Vogel, C. Global signatures of protein and mRNA  
869 expression levels. *Mol. Biosyst.* **5**, 1512–1526 (2009).
- 870 93. Schwanhüusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–  
871 342 (2011).
- 872 94. Ron Milo & Rob Phillips. *Cell Biology by the Numbers.* (2019).
- 873 95. Liu, Y., Beyer, A. & Aebersold, R. On the dependency of cellular protein levels on mRNA abundance.  
874 *Cell* **165**, 535–550 (2016).
- 875 96. Taniguchi, Y. *et al.* Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in  
876 single cells. *Science* **329**, 533–538 (2010).
- 877 97. Li, M. *et al.* Codon-usage-based inhibition of HIV protein synthesis by human schlafen 11. *Nature* **491**,  
878 125–128 (2012).
- 879 98. Stabell, A. C. *et al.* Non-human primate schlafen11 inhibits production of both host and viral proteins.  
880 *PLoS Pathog.* **12**, (2016).
- 881 99. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E. M. Absolute protein expression profiling estimates the  
882 relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* 2006 251 **25**, 117–  
883 124 (2006).
- 884 100. Bauer, A. P. *et al.* The impact of intragenic CpG content on gene expression. *Nucleic Acids Res.* **38**, 3891–  
885 3908 (2010).

- 886 101. Krinner, S. *et al.* CpG domains downstream of TSSs promote high levels of gene expression. *Nucleic*  
887 *Acids Res.* **42**, 3551–3564 (2014).
- 888 102. Simmonds, P., Xia, W., Baillie, J. K. & McKinnon, K. Modelling mutational and selection pressures on  
889 dinucleotides in eukaryotic phyla -selection against CpG and UpA in cytoplasmically expressed RNA and  
890 in RNA viruses. *BMC Genomics* **14**, 1–16 (2013).
- 891 103. Greenbaum, B. D., Levine, A. J., Bhanot, G. & Rabadan, R. Patterns of evolution and host gene mimicry  
892 in influenza and other RNA viruses. *PLoS Pathog.* **4**, (2008).
- 893 104. Burns, C. C. *et al.* Genetic inactivation of poliovirus infectivity by increasing the frequencies of CpG and  
894 UpA dinucleotides within and across synonymous capsid region codons. *J. Virol.* **83**, 9957–9969 (2009).
- 895 105. Tulloch, F., Atkinson, N. J., Evans, D. J., Ryan, M. D. & Simmonds, P. RNA virus attenuation by codon  
896 pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. *Elife* **3**, e04531  
897 (2014).
- 898 106. Lauring, A. S., Acevedo, A., Cooper, S. B. & Andino, R. Codon usage determines the mutational  
899 robustness, evolutionary capacity, and virulence of an RNA virus. *Cell Host Microbe* **12**, 623–632 (2012).
- 900 107. Riba, A. *et al.* Protein synthesis rates and ribosome occupancies reveal determinants of translation  
901 elongation rates. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 15023–15032 (2019).
- 902 108. Shah, P., Ding, Y., Niemczyk, M., Kudla, G. & Plotkin, J. B. Rate-limiting steps in yeast protein  
903 translation. *Cell* **153**, 1589–1601 (2013).
- 904 109. Wang, S. E., Brooks, A. E. S., Poole, A. M. & Simoes-Barbosa, A. Determinants of translation efficiency  
905 in the evolutionarily-divergent protist *Trichomonas vaginalis*. *BMC Mol. Cell Biol.* **21**, 1–13 (2020).
- 906 110. Mauger, D. M. *et al.* mRNA structure regulates protein expression through changes in functional half-life.  
907 *Proc. Natl. Acad. Sci. U. S. A.* **116**, 24075–24083 (2019).
- 908 111. Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z. & Blüthgen, N. Efficient translation initiation dictates  
909 codon usage at gene start. *Mol. Syst. Biol.* **9**, 675 (2013).
- 910 112. Tuller, T. *et al.* An Evolutionarily conserved mechanism for controlling the efficiency of protein  
911 translation. *Cell* **141**, 344–354 (2010).
- 912 113. Cambray, G., Guimaraes, J. C. & Arkin, A. P. Evaluation of 244,000 synthetic sequences reveals design  
913 principles to optimize translation in *Escherichia coli*. *Nat. Biotechnol.* **36**, 1005 (2018).
- 914 114. Brightwell, G., Poirier, V., Cole, E., Ivins, S. & Brown, K. W. Serum-dependent and cell cycle-dependent  
915 expression from a cytomegalovirus-based mammalian expression vector. *Gene* **194**, 115–123 (1997).
- 916 115. Amorós-Moya, D., Bedhomme, S., Hermann, M. & Bravo, I. G. Evolution in regulatory regions rapidly  
917 compensates the cost of nonoptimal codon usage. *Mol. Biol. Evol.* **27**, 2141–2151 (2010).
- 918 116. Lynch, M. & Marinov, G. K. The bioenergetic costs of a gene. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15690–  
919 15695 (2015).
- 920 117. Princiotta, M. F. *et al.* Quantitating protein synthesis, degradation, and endogenous antigen processing.  
921 *Immunity* **18**, 343–354 (2003).
- 922 118. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins  
923 evolve slowly. *Proc. Natl. Acad. Sci.* **102**, 14338–14343 (2005).

- 924 119. Ribas de Pouplana, L., Santos, M. A. S., Zhu, J. H., Farabaugh, P. J. & Javid, B. Protein mistranslation:  
925 friend or foe? *Trends Biochem. Sci.* **39**, 355–362 (2014).
- 926 120. Walsh, I. M., Bowman, M. A., Soto Santarriaga, I. F., Rodriguez, A. & Clark, P. L. Synonymous codon  
927 substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proc. Natl. Acad. Sci.*  
928 *U. S. A.* **117**, 3528–3534 (2020).
- 929 121. Andersson, S. G. E. & Kurland, C. G. Codon preferences in free-living microorganisms. *Microbiol. Rev.*  
930 **54**, 198–210 (1990).
- 931 122. Frumkin, I. *et al.* Codon usage of highly expressed genes affects proteome-wide translation efficiency.  
932 *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4940–E4949 (2018).
- 933 123. Dittmar, K. A., Sørensen, M. A., Elf, J., Ehrenberg, M. & Pan, T. Selective charging of tRNA isoacceptors  
934 induced by amino-acid starvation. *EMBO Rep.* **6**, 151–157 (2005).
- 935 124. Elf, J., Nilsson, D., Tenson, T. & Ehrenberg, M. Selective charging of tRNA isoacceptors explains patterns  
936 of codon usage. *Science.* **300**, 1718–1722 (2003).
- 937 125. Gatignol, A., Durand, H. & Tiraby, G. Bleomycin resistance conferred by a drug-binding protein. *FEBS*  
938 *Lett.* **230**, 171–175 (1988).
- 939 126. Chen, S. L., Lee, W., Hottes, A. K., Shapiro, L. & McAdams, H. H. Codon usage between genomes is  
940 constrained by genome-wide mutational processes. *Proc. Natl. Acad. Sci.* **101**, 3480–3485 (2004).
- 941 127. Caspersson, T. *et al.* Chemical differentiation along metaphase chromosomes. *Exp. Cell Res.* **49**, 219–222  
942 (1968).
- 943 128. Galtier, N., Piganeau, G., Mouchiroud, D. & Duret, L. GC-content evolution in mammalian genomes: the  
944 biased gene conversion hypothesis. *Genetics* **159**, 907–911 (2001).
- 945 129. Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes.  
946 *Annu. Rev. Genomics Hum. Genet.* **10**, 285–311 (2009).
- 947 130. Chaney, J. L. & Clark, P. L. Roles for synonymous codon usage in protein biogenesis. *Annu. Rev. Biophys*  
948 **44**, 143–166 (2015).
- 949 131. Galtier, N. *et al.* Codon Usage Bias in Animals: Disentangling the effects of natural selection, effective  
950 population size, and GC-biased gene conversion. *Mol. Biol. Evol.* **35**, 1092–1103 (2018).
- 951 132. Sharp, P. M., Tuohy, T. M. F. & Mosurski, K. R. Codon usage in yeast: cluster analysis clearly  
952 differentiates highly and lowly expressed genes. *Nucleic Acids Res.* **14**, 5125–5143 (1986).
- 953 133. Stenico, M., Lloyd, A. T. & Sharp, P. M. Codon usage in *Caenorhabditis elegans*: delineation of  
954 translational selection and mutational biases. *Nucleic Acids Res.* **22**, 2437–2446 (1994).
- 955 134. Shields, D. C., Sharp, P. M., Higgins, D. G. & Wright, F. Silent sites in *Drosophila* genes are not neutral:  
956 evidence of selection among synonymous codons. *Mol. Biol. Evol.* **5**, 704–716 (1988).
- 957 135. Bierne, N. & Eyre-Walker, A. Variation in synonymous codon use and DNA polymorphism within the  
958 *Drosophila* genome. *J. Evol. Biol.* **19**, 1–11 (2006).
- 959 136. Lynch, M. *et al.* Population genomics of *Daphnia pulex*. *Genetics* **206**, 315–332 (2017).
- 960 137. Musto, H., Cruveiller, S., D’Onofrio, G., Romer, H. & Bernardi, G. Translational selection on codon usage  
961 in *Xenopus laevis*. *Mol. Biol. Evol.* **18**, 1703–1707 (2001).



- 962 138. Dhindsa, R. S., Copeland, B. R., Mustoe, A. M. & Goldstein, D. B. Natural selection shapes codon usage  
963 in the human genome. *Am. J. Hum. Genet.* **107**, 83–95 (2020).
- 964 139. Kim, J. H. *et al.* High cleavage efficiency of a 2A peptide derived from porcine teschovirus-1 in human  
965 cell lines, zebrafish and mice. *PLoS One* **6**, (2011).
- 966 140. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data.  
967 *Bioinformatics* **30**, 2114–2120 (2014).
- 968 141. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements.  
969 *Nat. Methods* **12**, 357–360 (2015).
- 970 142. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
- 971 143. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification.  
972 *Nat. Biotechnol.* **34**, 525–527 (2016).
- 973 144. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based  
974 shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).
- 975 145. Cox, J. *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *J.*  
976 *Proteome Res.* **10**, 1794–1805 (2011).
- 977 146. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data.  
978 *Nat. Methods* **13**, 731–740 (2016).
- 979 147. Rueden, C. T. *et al.* ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics*  
980 **18**, 1–26 (2017).
- 981 148. Perez-Riverol, Y. *et al.* The PRIDE database resources in 2022: a hub for mass spectrometry-based  
982 proteomics evidences. *Nucleic Acids Res.* **50**, D543–D552 (2022).